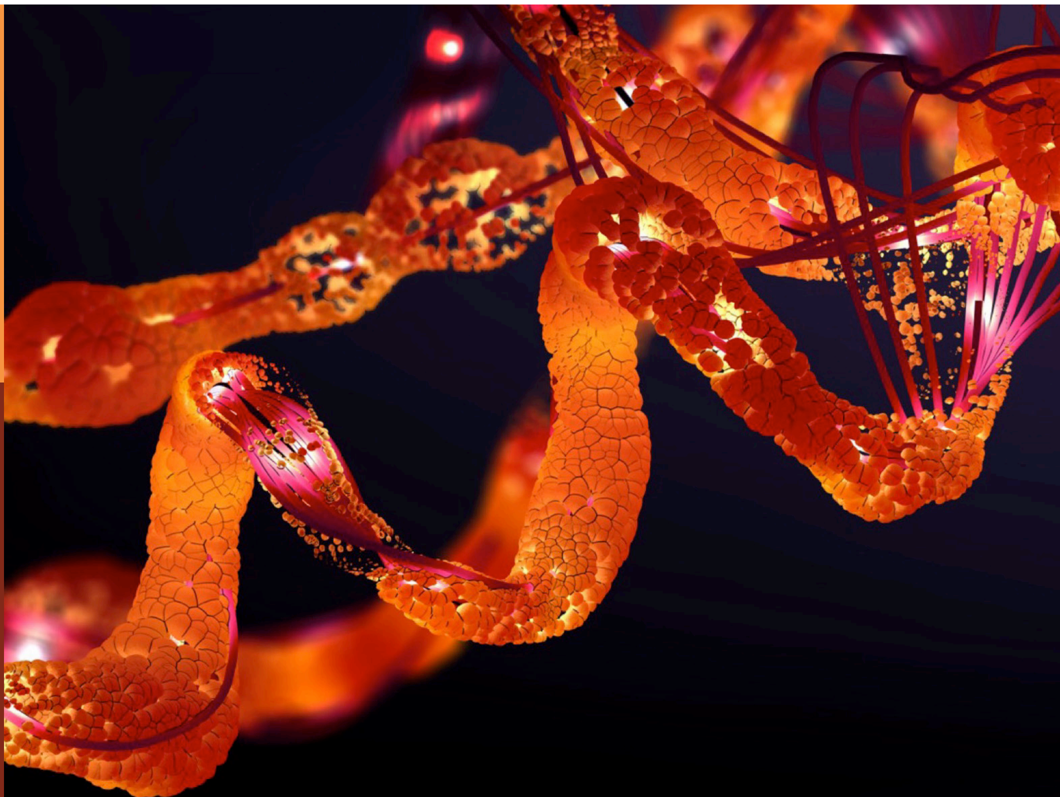
Research Topics

Insights in metabolomics 2021

Edited by
Wolfram Weckwerth

Published in
Frontiers in Molecular Biosciences



ISSN 1664-8714
ISBN 978-2-8325-3761-9
DOI 10.3389/978-2-8325-3761-9

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Insights in metabolomics: 2021

Topic editor
Wolfram Weckwerth — University of Vienna, Austria

Citation
Weckwerth, W., ed. (2023). *Insights in metabolomics: 2021*.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3761-9

Table of contents

frontiers
in Molecular Biosciences

REVIEW
published: 07 February 2022
doi: 10.3389/fmolb.2022.789889

Metabolomics and its Applications in Cancer Cachexia

*Pengfei Cui[1], Xiaoyi Li[2], Caihua Huang[3], Qinxi Li[4] and Donghai Lin[5]**

[1]College of Food and Pharmacy, Xuchang University, Xuchang, China, [2]Xuchang Central Hospital, Xuchang, China, [3]Department of Physical Education, Xiamen University of Technology, Xiamen, China, [4]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen, China, [5]Key Laboratory for Chemical Biology of Fujian Province, MOE Key Laboratory of Spectrochemical Analysis and Instrumentation, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen, China

Cancer cachexia (CC) is a complicated metabolic derangement and muscle wasting syndrome, affecting 50–80% cancer patients. So far, molecular mechanisms underlying CC remain elusive. Metabolomics techniques have been used to study metabolic shifts including changes of metabolite concentrations and disturbed metabolic pathways in the progression of CC, and expand further fundamental understanding of muscle loss. In this article, we aim to review the research progress and applications of metabolomics on CC in the past decade, and provide a theoretical basis for the study of prediction, early diagnosis, and therapy of CC.

Keywords: cancer cachexia, metabolomics, metabolic alterations, progress, biomarker

OPEN ACCESS

Edited by:
Wolfram Weckwerth,
University of Vienna, Austria

Reviewed by:
Cheng Guo,
Zhejiang University, China
Jian-Bo Wan,
University of Macau, China

****Correspondence:***
Donghai Lin
dhlin@xmu.edu.cn

Specialty section:
This article was submitted to Metabolomics,
a section of the journal
Frontiers in Molecular Biosciences

Received: *05 October 2021*
Accepted: *17 January 2022*
Published: *07 February 2022*

Citation:
Cui P, Li X, Huang C, Li Q and Lin D (2022) Metabolomics and its Applications in Cancer Cachexia. Front. Mol. Biosci. 9:789889. doi: 10.3389/fmolb.2022.789889

INTRODUCTION

Cancer cachexia (CC) is a multifactorial syndrome, which is characterized by disturbed metabolism, declined body weight, depleted muscle mass, and reduced food intake (Evans et al., 2008; Fearon et al., 2011). Overall, CC affects approximately 50–80% of cancer patients and leads to around 30% of mortality, with the highest incidence reported in gastrointestinal and pancreatic cancers (Loberg et al., 2007; Kumar et al., 2010). Lately, four stages of CC have been proposed to define the guidelines (Bozzetti and Mariani, 2009; Blum et al., 2010). Initially, CC begins in a pre-cachexia stage with unwitting body weight loss, along with a more severe and noninvertible fat tissues and skeletal muscles loss, followed by disturbances in metabolic pathway and immune system, ultimately resulting in death (Hamerman, 2002; Deans et al., 2009).

Declined body weight primarily arise from skeletal muscle loss, which is recognized as the major feature of CC. Muscle loss makes routine activities difficult and results in tiredness, in addition to the tremendous damage to quality of life and poor response to surgery or chemotherapy (Lok, 2015). Study showed that treatment on skeletal muscle loss could not only attenuate the symptoms of CC, but also remarkably prolongs lifespan (Zhou et al., 2010). Previous studies have found that CC is linked to various factors including fasting hormones, pro-inflammatory cytokines, such as interleukin 1 (IL-1), tumor necrosis factor-alpha (TNF-α), interferon-gamma (IFN-γ) (Nagaya et al., 2006; Gupta et al., 2011; Argiles et al., 2013; Argiles and Stemmler, 2013). The two main cell proteolysis pathways including ubiquitin-proteasome pathway and autophagy pathway regulate protein turnover in muscle tissues (Bonaldo and Sandri, 2013; Halle et al., 2020; Lim et al., 2020; Yang et al., 2020). In addition, several major signaling pathways including IGF1-Akt-FoxO pathway, TGFβ-myostatin pathway, NF-κB signaling, and glucocorticoids pathway have all been implicated in muscle atrophy of CC (Bodine et al., 2001; Musaro et al., 2001; Lee, 2004; Sandri et al., 2004; Waddell et al., 2008; Peterson et al., 2011). Identification of signaling pathways associated with CC and muscle atrophy has achieved great progress in recent decades. Given that CC is a typical metabolic syndrome, metabolomic techniques can be applied to explore biomarkers for early diagnosis of CC, address metabolic characteristics for mechanistic understanding of the pathogenesis of CC, and develop therapeutics strategies for treatments of CC.

As an omics technology developing after genomics, transcriptomics and proteomics, metabolomics has rapid developments at present, which can simultaneously analyze all of metabolites with small molecular weights in a biological system (Newgard, 2017). Compared to genomics, transcriptomics and proteomics, metabolomics is based on extensively used detection equipment including either mass spectrometry (MS) or nuclear magnetic resonance spectroscopy (NMR), which has the features of high sensitivity, high precision, good resolution, and small sample volume (Beckonert et al., 2007; Ma et al., 2018). In the last 2 decades, metabolomic techniques have been extensively used to exploring various diseases such as cancer (Wishart, 2016; Schmidt et al., 2021), type 2 diabetes (Newgard et al., 2009; Ma et al., 2018), fatty liver (Gao et al., 2009; Lallukka and Yki-Jarvinen, 2016), and cardiovascular diseases (Shah et al., 2012a; Shah et al., 2012b).

Metabolomic techniques create ideas and clues for scholars to predict and screen CC at early stage. Recently, researchers have applied metabolomic analysis to perform global and in-depth studies for identifying metabolic signatures in patients or animal models or cell models with CC, and also for identifying potential biomarkers and crucial metabolic pathways to mechanistically understand the pathogenesis of CC. We searched for articles from PubMed, Scopus and Google Scholar relevant to cancer cachexia by using the keywords "cancer cachexia and metabolomics," "cancer cachexia and metabonomics," "cancer cachexia and metabolic," "muscle atrophy and metabolomics," "muscle loss and metabolomics" and so on. We have only included the studies based on the animal models or clinical samples related to CC by using metabolomics methodologies. So far, only two reviews of omics studies on CC have been reported (Gallagher et al., 2016; Twelkmeyer et al., 2017). However, these reviews did not pay much attention to the field of metabolomics. To widely expand the knowledge of CC and give inspirations for the cachexia studies from the view of biomarkers, signatures and therapeutic targets, we focus on the progress made in the past decade, novel developments, and latest discoveries in the study of CC using metabolomic techniques, and look forward to its future developments. To the best of our knowledge, this article presents the first review on the progress of metabolomic applications in CC.

METABOLOMIC RESEARCH METHODOLOGIES AND TECHNIQUES

Followed by genomics, transcriptomics, and proteomics, metabolomics is a promising subject that has promptly developed in recent years. It can be qualitatively and quantitatively employed to analyze various sample sources, which include cells or tissues extract, bio-fluids, and microorganisms caused by genetically engineered or drug treatment. Metabolomic analyses usually focus on small molecular metabolites such as amino acids, lipids, small molecular peptides, and organic acids with a relative molecular weight of less than 1,000 Da (Nicholson et al., 1999; Fiehn et al., 2000; Xia and Wan, 2021). Generally, metabolomics includes two tools: non-targeted and targeted metabolomics. Non-targeted metabolomics is most widely used in CC studies to explore biomarkers (Yang et al., 2018), signatures (Cui et al., 2019b), and therapeutic targets (Fukawa et al., 2016). The process of metabolomic analysis in CC is depicted in **Figure 1**, which contains sample sources, analytical platforms, data collection and analysis, biomarkers identification, metabolic pathways exploration, and biological significance elucidation.

Sample Sources

The most commonly used type of samples for metabolomic studies in CC are serum/plasma, urine, tumor tissues, liver and skeletal muscle, and other tissues. The collected blood samples are further processed with cell separation to obtain sera and plasma at 4°C before analysis. This step might be one of the major factors of pre-analysis errors in blood metabolomics research (Beckonert et al., 2007; Nikolic et al., 2014). It is generally suggested that the interval between blood sample collection and cell separation should be finished in 35 min to avoid the increased lactate levels. In addition, repeated freezing and thawing steps should be avoided in the whole experiment (Yin et al., 2015). Compared with blood samples, the biological composition of urine samples is relatively simple, the protein content is low, and additional metabolite extraction steps are not usually required. The commonly used pretreatment method for skeletal muscle and tumor tissues in CC studies is liquid-liquid extraction. In general, tissue samples are initially extracted with cold solutions which contains chloroform, methanol and water in a certain ratio to generate a two-phase system. The polar and non-polar metabolites are separated, lyophilized and dissolved in corresponding solvents, respectively (Jonsson et al., 2005; Beckonert et al., 2007; Legido-Quigley et al., 2010).

Data Collection Techniques

Metabolomic techniques are applied to measure the number, type, condition, and level of metabolites and to explore metabolic profiles (Beckonert et al., 2007). Compared to other detection techniques, NMR and MS are the two mostly used techniques for metabolomic analysis (Emwas, 2015).

NMR technology is a spectroscopic technology that uses different atomic nuclei to absorb ratio-frequency radiation with different resonance frequencies, which are converted into molecular chemistry and structural information related to environments of the nuclei (Bothwell and Griffin, 2011). With the development of NMR technology, researchers can directly analyze intact gastrocnemius muscle without any pretreatment of samples by using high-resolution magic angle rotation (HRMAS-NMR) spectroscopy in a CC mouse model (Yang et al., 2015). Overall, NMR spectroscopy has many advantages such as simple sample preparation, non-invasive and unbiased measurement of the sample, good objectivity and reproducibility (Li et al., 2015). However, signal overlap and low sensitivity are two obvious shortcomings in complicated ^{1}H-NMR spectra.

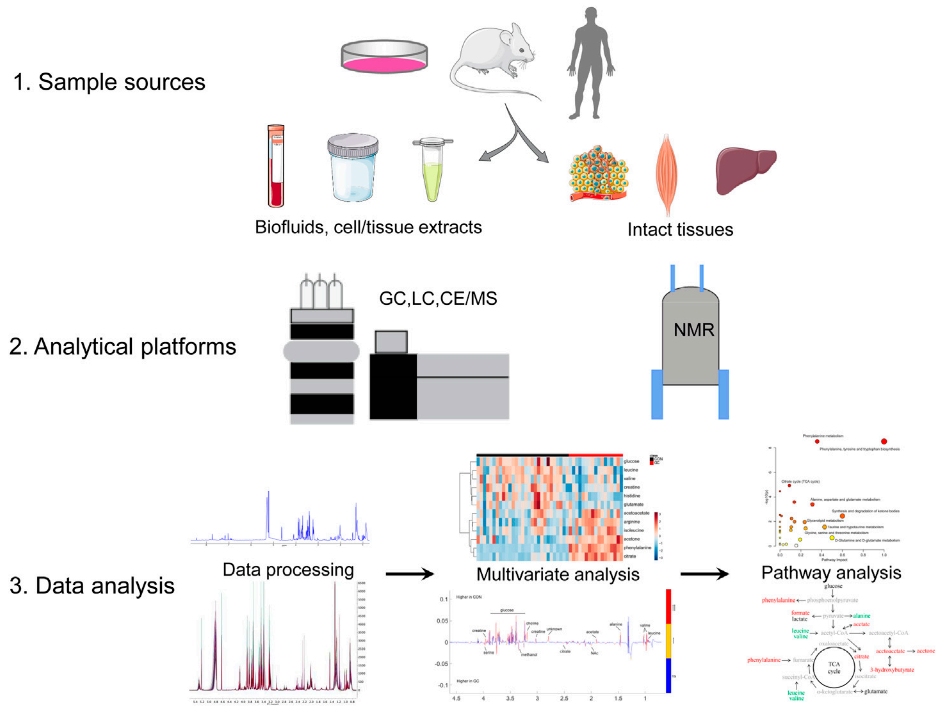


FIGURE 1 | Metabolomics analysis workflow. Abbreviations: NMR, nuclear magnetic resonance; CE, capillary electrophoresis; GC, gas chromatography; LC, liquid chromatography; MS, mass spectrometry.

TABLE 1 | Summarization of advantages and drawbacks of MS and NMR detections.

Features	NMR	MS
Sample preparation	Simple	Complex
Sample measurement	Simple	Complex, various chromatography methods
Sample recovery	Good, non-invasive	Destructive
Selectivity and targeted analysis capabilities	General, mostly in untargeted analysis	Good, untargeted and targeted analysis
Sensibility	Low, <100 metabolites per test	High, >1,000 metabolites per test
Resolution	General	General
Repeatability	High	low

MS spectroscopy uses electric and magnetic fields to separate moving ions and detect them according to the mass-to-charge ratio (m/z) (Tsiropoulou et al., 2017). At present, MS combined with chromatography are divided into three types including capillary electrophoresis-mass (CE-MS), gas chromatography-mass (GC-MS), and liquid chromatography-mass (LC-MS). CE-MS has high performance for polar and ionic compounds with high resolution and sensitivity rather than uncharged compounds (Ramautar et al., 2019; Stolz et al., 2019). Compared to GC and LC, CE has a superiority over them for the resolution of charged molecules along with the isomers due to the excellent separation. GC-MS is generally conducted to analyze non-polar, low-boiling and volatile molecules, and samples usually need to be derivatized. LC-MS has relatively high sensitivity and strong detection ability for polar and thermally unstable compounds, by which a wider range of metabolites with low detection limit can be analyzed. It can be used for trace analysis and is more suitable for metabolomic analysis of complex biological samples (Römisch-Margl et al., 2011; Xiao et al., 2012). Cala and colleagues performed a combination of 3 types of MS (GC-MS, CE-MS, and LC-MS) to obtain plasma metabolite fingerprinting in a CC clinical study (Cala et al., 2018).

Compared with NMR spectroscopy, MS has several advantages such as high sensitivity and resolution, which could detect thousands of metabolites in a large dynamic range at the same time. However, MS also has its own shortcomings such as complicated sample preparation and low reproducibility. The advantages and drawbacks of MS and NMR detections are listed in **Table 1**. To promote the entire performance of metabolomics studies, Pin and colleagues combined MS and NMR to investigate differences between CC and chemotherapy induced cachexia (Pin et al., 2019).

Data Preprocessing and Analysis

Data analysis includes data preprocessing, multivariate statistical analysis, model establishment and verification, and selection of difference variables, etc. Prior to obtaining metabolomics data for statistical analysis, it is necessary to preprocess the data, which mainly includes baseline correction, peak screening (peak identification, peak alignment and correction), noise filtering, missing value processing, normalization and scaling (Dunn et al., 2011). Thereafter, multivariate statistical analysis is conducted to decrease the dimensionality of acquired data and extract information, including principal component analysis (PCA), clustering analysis, partial least square analysis (PLS), PLS-discriminant analysis (PLS-DA), orthogonal PLS (OPLS)-DA and random forests (RF) (Idborg-Bjorkman et al., 2003; Wiklund et al., 2008; Sugimoto et al., 2012; Schwammle et al., 2015). PCA, PLS-DA, OPLS loading plot and heatmap analysis were most commonly used in CC metabolomics studies. Besides, general statistical analyses including analysis of variance (ANOVA) and Student's t-test are also applied to quantitatively analyze the abundance of metabolites between different groups. When performing the multiple comparisons, the familywise error rate (FWER) might cause false-positive detection, which could be diminished by the procedures of false discovery rate (FDR) with Holm, Bonferroni and Benjamini-Hochberg corrections in metabolomic analysis (Sugimoto et al., 2012; Muroya et al., 2020). Overall, the combination of multivariate statistical analysis and classical statistical analysis can improve the reliability of the data analysis.

Data Elucidation

After multivariate statistical analysis, one can uncover and illustrate metabolic signatures based on several databases, including significantly altered concentrations of metabolites and certain disturbed metabolic pathways corresponding to external metabolic stimuli. These databases include HMDB (http://www.hmdb.ca/), METLIN (https://metlin.scripps.edu/), SMPDB (https://smpdb.ca), MassBank (http://www.massbank.jp/), The Kyoto Encyclopedia of Genes and Genomes (KEGG; https://www.genome.jp/kegg/), and software such as MetaboAnalyst 5.0 (https://www.metaboanalyst.ca/) (Pang et al., 2021). A growing number of studies have been using MetaboAnalyst website to conduct the pathway analysis and ROC analysis in CC studies (Yang et al., 2018; Cui et al., 2019b; Sadek et al., 2021).

ADVANCES IN THE PATHOGENESIS OF CANCER CACHEXIA BASED ON METABOLOMICS

Skeletal muscle loss might occur in the early stage, which might be masked by dysfunction and symptoms of other tissues. Methods used to assess muscle loss involve diagnostic imaging techniques, including computed tomography (CT), dual energy X-ray absorptiometry (DXA), and magnetic resonance imaging (MRI). However, these methods are associated with several shortcomings such as time consuming, expensive, complicated, and invasive when clinicians wish to screen the early or slow muscle loss (Heymsfield et al., 1997; Mitsiopoulos et al., 1998; Shen et al., 2004; Mourtzakis et al., 2008). In order to exploit the progress of muscle loss dynamically, several methods have been developed to detect CC syndromes and shorten the period for early prevention (Evans et al., 2008). Recently, metabolomic analysis is widely being applied to uncover novel biomarkers, explore certain metabolic pathways associated with the pathogenesis of various diseases including CC, and ultimately exploit potential therapeutic strategies in the future (Twelkmeyer et al., 2017). Applications of metabolomics analysis in CC are depicted in **Figure 2**, which cover biomarkers, signatures and therapeutic targets.

Biomarkers

Metabolomics can be applied to detect hundreds of small metabolites simultaneously for providing better elucidation of metabolic pathways related to the pathological mechanisms of CC, ultimately identifying reliable biomarkers for diagnosis and monitoring of cachexia.

Metabolomics studies in CC began in 2008 relied on the classical colon-26 (C26) mouse model. Connell and colleagues demonstrated that metabolomic analysis has the ability to diagnose and discover the surrogate serum biomarkers in CC for the first time (O'Connell et al., 2008). They conducted NMR-based metabolomic analysis on serum samples, and observed significant metabolic alterations including elevated amounts of very low-density lipoprotein (VLDL) and low-density lipoprotein (LDL) related to aberrant glycosylation of β-Dystroglycan (O'Connell et al., 2008). In a recent study based on the same C26 model, Lautaoja and colleagues identified free phenylalanine in sera and muscle tissues as a promising biomarker of cachectic muscle atrophy by using GC-MS-based metabolomic analysis (Lautaoja et al., 2019).

Furthermore, Kunz and colleagues performed untargeted LC-MS-based metabolomic analysis of plasma and skeletal muscle in a Lewis lung carcinoma (LLC) mouse model. They detected increased levels of asymmetric dimethylarginine, and NG-monomethyl-L-arginine in LLC group relative to normal group. In order to further explore the function of these two methylarginines in muscle turnover, the researchers treated the cultured myotubes with these two metabolites and found impaired muscle protein synthesis *in vitro* study. Surprisingly, increased levels of asymmetric dimethylarginine were also observed in muscle tissues from clinic patients. This study not only discovered two novel potential biomarkers, but also provided therapeutic ideas for CC (Kunz et al., 2020).

In addition, Yang and colleagues revealed dynamically changing metabolic profiles in sera and intact muscle of CC in the C26 mouse model from pre-cachexia to the refractory cachexia period. They identified five unique metabolic features including declined levels of serum glucose and BCAAs, increased levels of ketone bodies, neutral amino acids and 3-methylhistidine (Yang et al., 2015). Using HRMAS-NMR spectroscopy, they performed metabolic profiling of cachectic gastrocnemius muscle for the first time. To further validate the metabolic features identified from the mouse model, recently, Yang and colleagues recruited 33 pre-cachectic, 84 cachectic and 105 cancer patients with stable body weights and 74 healthy controls, according to the international definition and

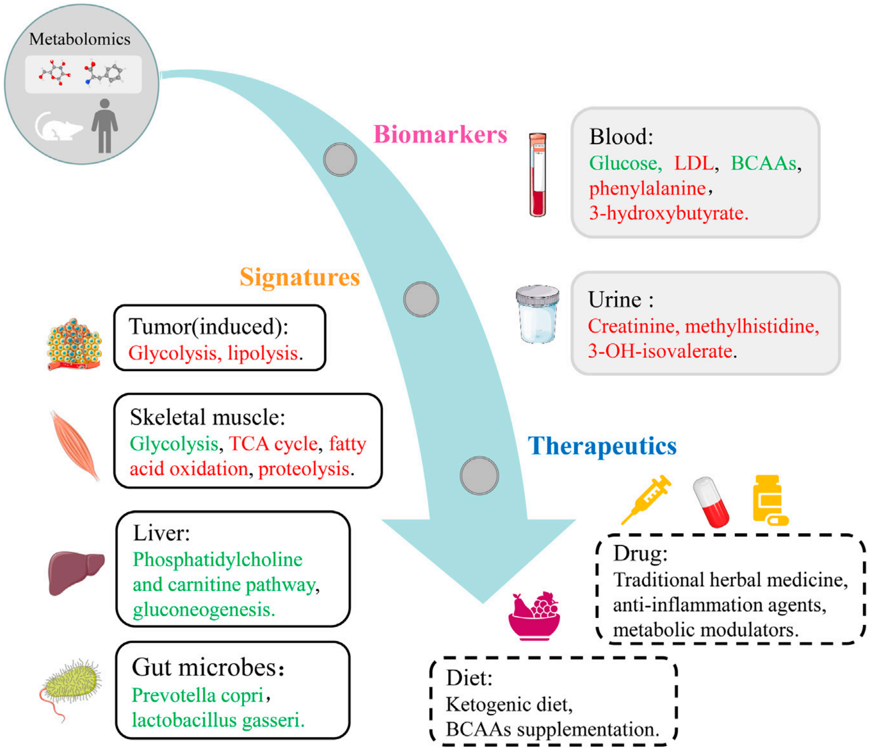


FIGURE 2 | Metabolomics applications covering biomarkers, signatures and therapeutic targets in CC. Red, upregulated metabolites and pathways in CC group. Green, downregulated metabolites, microbes and pathways in CC group.

classification of CC (Fearon et al., 2011). They conducted NMR metabolomic analyses on sera and urine of CC patients to reveal the metabolic profile of CC, and identified 15 metabolites for discriminating different disease states (Yang et al., 2018). Based on three identified metabolites (carnosine, leucine and phenyl acetate), they established a diagnostic model for predicting the presence of cachexia with high accuracy.

In a previous study, Fujiwara and colleagues enrolled 21 advanced pancreatic cancer patients with or without cachexia, collected serum samples at different time point, and performed GC-MS-based metabolomic analysis (Fujiwara et al., 2014). They observed intraday differences in serum metabolite concentration, which were observably altered in the evening but basically identical in the daytime. Specifically, abundance of paraxanthine was significantly decreased in CC patients compared to those without cachexia all day long, which was potentially associated with cachexia. Additionally, another study performed NMR-based metabolomics analysis on 170 patients with head and neck squamous cell carcinoma cancer (HNSCC). These patients experienced radical treatments with radio-/chemo-radiotherapy (RT/CHRT) (Boguszewicz et al., 2019). Boguszewicz and colleagues indicated that serum metabolic alterations primarily related to high 3-hydroxybutyrate levels could be detected at an early stage of the treatment experienced by HNSCC patients. Thus, 3-hydroxybutyrate could be exploited as a fast and sensitive biomarker of malnutrition or cachexia.

Similarly, Miller and colleagues conducted LC-MS-based metabolomic analysis and identified potential biomarkers related to weight loss in patients with upper gastrointestinal cancer, which could be applied for the assessment of therapeutic intervention (Miller et al., 2019). Cancer patients with ≥5% weight loss displayed plasma metabolic profiles distinguished from those with <5% weight loss. Totally, six metabolites were highly discriminative of body weight loss, including lysoPC18.2 and 16:1, hexadecanoic acid, octadecanoic acid, phenylalanine.

Metabolites in urine samples have also been investigated to discover novel biomarkers for CC. Eisner and colleagues did the first attempt to use single time-point urinary metabolite profiles to diagnose muscle wasting occurring in CC humans (Eisner et al., 2011). After analyzing 93 random urine samples from cancer patients, the researchers found that some metabolites such as creatinine and methylhistidine arising from muscle proteolysis were particularly released into urine. This study provides an inspiration that it might also be convenient, cheap and safe to detect muscle wasting based on ^{1}H-NMR urine metabolomic analysis. Overall, these results obtained from previous studies on biomarkers for CC mostly depend on the samples derived from animal models and human, and also on tumor type, bio-fluids and analytical platforms.

Metabolic Signatures and Metabolic Pathways

Although numerous researches have been exploring molecular mechanisms underlying muscle wasting in CC, the effect of muscle wasting on muscle function and metabolic signatures remains unclear (Diffee et al., 2002). Metabolic impairments in the skeletal muscle are related to its physiological dysfunction. Thus, metabolic derangements might be involved in molecular mechanisms underlying protein synthesis and breakdown (Tisdale, 2003; Santarpia et al., 2011).

Yang and colleagues indicated that serum metabolic disturbances associated with promoted tricarboxylic acid (TCA) cycle and amino acid metabolism were the major features of CC in C26 mouse model. Amino acids, ketone bodies and metabolites involved in TCA cycle were recognized as potential biomarkers related to the corresponding metabolic pathways (Quanjun et al., 2013). Furthermore, Torossian and colleagues performed GC-MS-based and LC-MS-based metabolomic analyses to reveal metabolic distinctions between cachectic gastrocnemius muscles and control muscles in the C26 mouse model (Der-Torossian et al., 2013b). They showed predominant effects of CC including: enhanced oxidative stress, impaired redox homeostasis, altered metabolite concentrations in glycolysis and declined carbon flow through TCA cycle. This study found the tumor Warburg-like metabolic pattern in skeletal muscle of CC for the first time, which is considered as novel metabolic signature in CC research.

Compared to malabsorption, fasting, age-induced muscle loss, and sarcopenia, CC has its own metabolic features (Fearon, 1992; Tisdale, 2009; Evans, 2010; Rolland et al., 2011). Consistently, Torossian and colleagues performed a NMR-based metabolomic analysis of sera, and indicated that metabolic alterations including hyperlipidemia, hyperglycemia and reduced BCAAs distinguish cachexia from effects of starvation (Der-Torossian et al., 2013a). Another previous study explored metabolic differences between sarcopenia and CC in senile cancer animals. The researchers conducted NMR-based metabolomic analysis dynamically on sera derived from adult and ageing rats. The metabolic alterations mostly focused in several metabolic pathways, including amino acid biosynthesis which was upregulated in the aging group and downregulated in the tumor groups (Viana et al., 2020). Recently, they also performed NMR-based metabolomic analysis on gastrocnemius derived from weanling and young adult rats, aiming to explore metabolic alterations in cachectic hosts during the whole lifespan (Chiocchetti et al., 2021). They indicated that the most significant variations of metabolites such as glutamate, glutamine, glycine, and methylhistidine, might be associated with the early muscle catabolism and declined energy generation in cachectic muscles.

Chemotherapy is widely used to cancer patients in the clinical treatment, however, growing evidences have shown that several chemotherapeutic drugs could also lead to the occurrence of cachexia and deterioration of muscle mass. To date, only one metabolomics investigation has been done in chemotherapy-induced cachexia. Based on the C26 mouse model, Pin and colleagues revealed significant differences in amino acid catabolism, TCA cycle, and β-oxidation between CC and chemotherapy-induced cachexia by a combination of NMR-based metabolomics with targeted MS analysis (Pin et al., 2019).

Although skeletal muscles are the main tissue impaired dramatically in CC, other tissues such as liver and gut may also be affected and involved in the pathophysiology of this complex syndrome (Rohm et al., 2019). As an essential metabolic organ, liver regulates body energy metabolism and maintains its homeostasis. Dysfunction of liver metabolism are prone to cause promoted energy consumption in CC. Furthermore, gut microbial species play key roles in nutrients supplementation, cytokines and gut hormones regulation, and gut barrier function improvement. Based on these beneficial effects, scholars are exploring if these micrograms could act as novel therapeutic targets for CC (Valdes et al., 2018). Based on the C26 mouse model, Pötgens and colleagues explored the crosstalk in four different samples including caecal, portal vein, vena cava and liver by a combination of NMR-based metabolomics with gut gene sequencing and hepatic transcriptomics. Their results showed depressed glycolysis and gluconeogenesis, activated hexosamine pathway and phosphatidylcholine pathway, reduced abundances of hepatic carnitine and caecal acetate and butyrate, and decreased levels of aromatic amino acids (Potgens et al., 2021). Given that CC also induces anorexia and reduced food intake, Uzu and colleagues focused on studying metabolic signatures of brains and conducted a CE-MS-based metabolomic analysis on brain samples derived from a CC mouse model. They observed activated purine metabolism and increased xanthine oxidase activity in brains of cachexic mice relative to controls (Uzu et al., 2019).

Ni and colleagues conducted a comprehensive analysis on 31 patients with lung cancer by a combination of plasma metabolomics and gut microbiota metagenomics (Ni et al., 2021). For the first time, they explored gut microbiota functions in the clinical CC study, and observed remarkably decreased levels of BCAAs, methylhistamine, and vitamins in CC blood. They further discovered that increased levels of BCAAs and 3-oxocholic acid in non-CC blood were closely related to gut microbiota especially *Prevotella copri* and *Lactobacillus gasseri*, respectively. These results shed lights on molecular mechanisms underlying host-microbiota crosstalk in CC, and provided new strategies for preventing or treating CC through regulating gut microbiota in the future nutritional supplements.

Previously, preclinical mouse models (mainly C26 and LLC) were established by using subcutaneous implantation methods to conduct CC studies. Few murine models of CC with orthotopic implantation have been employed. Thus, our group established two orthotopic models including glioma cachexia and gastric cancer cachexia to mimic clinical characteristics of CC. In the first study, we conducted NMR-based metabolomic analysis to explore metabolic profiles in cachectic muscle based on a glioma induced cachexia murine model (Cui et al., 2019b). Our results indicated that significantly impaired pathways including energy metabolism, muscle protein breakdown and synthesis, and profoundly increased amino acids involved in TCA cycle anaplerotic. After that, we established a gastric CC murine model and performed NMR-based metabolomic analysis of gastric tissues (tumor), blood and skeletal muscle. Cachectic mice exhibited impaired glucose and nucleic acid metabolisms in tumor, hyperlipidemia and hypoglycemia in blood, and disturbed carbohydrate and amino acid metabolism in gastrocnemius (Cui et al., 2019a). Besides, we further explored the role of α-ketoglutarate in muscle protein turnover, and found α-ketoglutarate can alleviate the myotubes atrophy induced by glucose deprivation.

At present, only one study was performed using a combination of three metabolomics techniques (GC-MS, CE-MS, and LC-MS) to access a markedly different metabolic pattern in human plasma (Cala et al., 2018). Cala and colleagues collected two groups of plasma samples from 8 cachectic and 7 non-cachectic patients (Cala et al., 2018). Their results exhibited significantly decreased levels of amino acids and glycerophospholipids, and increased cortisol levels associated with cachexia. The disturbed metabolic pathways in CC included amino acid metabolism, aminoacyl-tRNA biosynthesis, fatty acid elongation, and TCA cycle. In another study, Stretch and colleagues investigated metabolic profiles of urine and plasma derived from 55 weight-losing patients by conducting NMR-based and direct injection MS-based metabolomics analyses. Their results indicated that large amounts of glycerophospholipids variations can be used to discover sarcopenia in cancer patients (Stretch et al., 2012). This study addressed one main issue that the variability of tissue mass might impact metabolic profiles, and thus could provide hints for the field of nutrition and metabolism studies. Overall, numerous researchers have investigated the metabolic signatures for CC from various aspects such as starvation, sarcopenia, chemotherapy, gut microbes, orthotopic implantation and analytical platforms, in order to give clues and inspirations to better elucidate the pathogenesis of CC and therapeutic targets discovery.

Therapeutic Strategies by Using Metabolomics

As discussed above, metabolomics is being extensively used to uncover biomarkers, metabolic signatures and metabolic pathways of CC, and to exploit novel drug targets. In this section, we discuss studies on how metabolomics contributes to the discovery of new targets for therapy.

Gut microbiota could depress inflammation response and tumor development (Bindels et al., 2012). Some researchers have been exploring the roles of gut microbiome in CC and addressing certain metabolic signatures in the last section. Bindels and colleagues performed a further study on gut microbiota with the expectation of finding novel interventions for CC treatments. They integrated gene sequencing and metabolomics as well as molecular profiling of the host, so as to obtain a comprehensive view on the pathophysiology of CC (Bindels et al., 2016). The portal metabolome reflected significantly decreased glucose and lipoproteins levels, increased creatine and lactate levels. These data demonstrated that gut microbiota can impact intestinal homeostasis, confer benefits to the host, prolong survival and attenuate cachexia.

Increased expressions of inducible nitric oxide synthase (iNOS) have been observed in muscle tissues of cancer, AIDS, chronic heart failure, and COPD cachexia patients, suggesting that iNOS may be involved in the onset of cachexia under various conditions (Adams et al., 2003; Agusti et al., 2004; Ramamoorthy et al., 2009). Sadek and colleagues identified a signature of amino acids that were altered by iNOS activity in muscle by performing LC-MS-based and GC-MS-based metabolomic analyses based on the C26 murine model (Sadek et al., 2021). Notably, iNOS could significantly increase levels of arginine, lysine, tryptophan and methylhistidine, which could be decreased by inhibiting iNOS. Furthermore, they also found iNOS-induced significant decreases in levels of pyruvate, α-ketoglutarate and succinate, which were restored by KO iNOS. These results demonstrated that drug blockade or gene knockout of iNOS could rescue muscle loss and improve metabolic disorders in CC. This study provided the idea on how to use metabolomic techniques to identify potential targeted metabolic pathways. Initially, the researchers clarified metabolic alterations in animal models by conducting metabolomics analysis, thereafter they conducted genetic or pharmacological inhibition of iNOS on certain metabolic pathways including glycolysis, TCA cycle and fatty acid oxidation, which were all related to the energy production. Ultimately, they clearly elucidated the role of the iNOS/NO pathway in promoting energy crisis during cachexia-induced muscle wasting.

In addition, Ballarò and colleagues found that abnormal muscle mitochondrial function is correlated with excessive proteolysis, autophagy and mitophagy in the established CC model (Penna et al., 2019). They conducted NMR-based metabolomic analyses of skeletal muscle, liver and plasma. They identified significantly altered energy and protein metabolism such as decreased muscle NADH, increased glutamine, BCAAs and phenylalanine in tumor hosts. Partially, mitochondria-targeted compound SS-31 could modulate both skeletal muscle metabolome and liver metabolome, restore levels of alanine and ATP, as well as liver glycogen and glutathione. This study suggested that targeting mitochondrial function might be an efficient therapeutic approach for CC (Ballaro et al., 2021).

Researches have illustrated that intervening targeted metabolic pathways could attenuate CC symptoms and prevent muscle loss. Yang and colleagues investigated metabolic signatures of CC and the contribution of formoterol to serum metabolites in the C26 mouse model with NMR-based metabolomics approach. They identified several potential biomarkers including amino acids, ketone bodies and citrate cycle metabolites, which well reflected the effects of formoterol treatment (Quanjun et al., 2013). In a later study, this group conducted NMR-based metabolomic analysis based on the C26 mouse model. They exhibited that primary disturbed metabolic pathways in CC were biosynthesis of the BCAAs and glycine, serine, and threonine metabolism. Significantly, treatment with curcumin changed glycolysis with declined levels of lactate, alanine and glucose (Quan-Jun et al., 2015). In addition, Ohbuchi and colleagues exploited molecular mechanisms under the effects of rikkunshito (RKT) acting as a Japanese traditional herbal medicine (Kampo) for the treatment of CC. The researchers performed GC-MS-based plasma metabolomic analysis based on a rat model, and indicated that increased plasma glucarate following the RKT administration could delay body weight loss, reduce muscle wasting and ascites content (Ohbuchi et al., 2015). These studies shed lights on applications of traditional medicines for alleviating the progression of CC.

TABLE 2 | Overview of metabolic characteristics of CC.

References	Study object	Sample information	Analytical technology	Metabolic characteristics
O'Connell et al. (2008)	Mice	Serum	NMR	UP: VLDL/LDL; DOWN: glucose.
Eisner et al. (2011)	Patients	Urine	NMR	UP: creatine, creatinine, 3-OH-isovalerate.
Stretch et al. (2012)	Patients	Urine, Plasma	NMR, MS	Glycerophospholipids and metabolites associated with amino acid metabolism.
Quanjun et al. (2013)	Mice	Serum	NMR	Enhanced citrate cycle and amino acid metabolism.
Der-Torossian et al. (2013a)	Mice	Serum	NMR	Hyperlipidemia, hyperglycemia; DOWN: BCAAs.
Der-Torossian et al. (2013b)	Mice	Muscle	LC-MS	Enhanced Warburg effect; Disrupted TCA cycle, promoted oxidative stress, impaired redox homeostasis.
Fujiwara et al. (2014)	Patients	Serum	GC-MS	Down: paraxanthine.
Ohbuchi et al. (2015)	Rat	Plasma	GC-MS	DOWN: glucarate;
Yang et al. (2015)	Mice	Serum, Muscle	NMR	UP: neutral amino acids, creatine, ketone bodies, 3-methylhistidine; DOWN: BCAAs, glucose.
Quan-Jun et al. (2015)	Mice	Serum	NMR	UP: phenylalanine; DOWN: BCAA, acetoacetate.
Tseng et al. (2015)	Mice	Muscle	LC-MS	Impaired glycolysis, glycogen synthesis; protein degradation.
Viana et al. (2016)	Rat	Serum, Tumor	NMR	UP: tryptophan, lactate, ketone bodies.
Bindels et al. (2016)	Mice	Portal plasma	NMR	UP: creatine, lactate; DOWN: glucose, lipoproteins.
Fukawa et al. (2016)	Mice	Muscle, cell	LC-MS	Excessive fatty acid oxidation, enhanced oxidative stress.
Yang et al. (2018)	Patients	Serum, urine	NMR	UP: Carnosine, phenylacetate; Down: leucine.
Cala et al. (2018)	Patients	Plasma	LC-MS, GC-MS, CE-MS	UP: cortisol; DOWN: Glycerophospholipids, Sphingolipids.
Lautaoja et al. (2019)	Mice	Serum, Muscle	GC-MS	UP: phenylalanine.
Boguszewicz et al. (2019)	Patients	Serum	NMR	UP: 3-hydroxybutyrate.
Miller et al. (2019)	Patients	Plasma	LC-MS	UP: lysoPC 18.2, L-proline, hexadecanoic acid, octadecanoic acid, phenylalanine and lysoPC 16:1.
Pin et al. (2019)	Mice	Plasma, Muscle, Liver	NMR, MS	UP: low-density lipoprotein particles; DOWN: circulating glucose, liver glucose and glycogen.
Uzu et al. (2019)	Mice	Brain	CE-MS	Activated purine metabolism, Enhanced xanthine oxidase activity.

Cui et al. (2019a)	Mice	Tumor,
		Serum,
		Muscle
			NMR	UP (tumor): pyruvate and lactate;
				DOWN (tumor): hypoxanthine, inosine, inosinate;
				UP (serum): lactate and glycerol;
				DOWN (serum): glucose;
				UP (muscle): α-ketoglutarate;
				DOWN (muscle): glucose.

Cui et al. (2019b)	Mice	Muscle	NMR	UP: glutamate, arginine, BCAAs; DOWN: glucose, glycerol, 3-hydroxybutyrate.
Kunz et al. (2020)	Mice	Plasma, Muscle	LC-MS	UP: asymmetric dimethylarginine; and NG-monomethyl-L-arginine.
Viana et al. (2020)	Rat	Serum	NMR	Promoted amino acid biosynthesis and metabolism.
Miyaguti et al. (2020)	Rat	Serum, Muscle	NMR	UP: tryptophan, phenylalanine, histidine, glutamine.
Chiocchetti et al. (2021)	Rat	Muscle	NMR	Increased amino acid levels and disordered energetic metabolism.
Potgens et al. (2021)	Mice	Caecal, portal vein, liver, vena cava	NMR	Suppressed glycolysis and gluconeogenesis, hepatic carnitine and phosphatidylcholine pathway activity; activated hexosamine pathway.
Ni et al. (2021)	Patients	Plasma, Gut	LC-MS	DOWN: methylhistamine, BCAAs, vitamins.
Ballaro et al. (2021)	Mice	Muscle, Liver, Plasma	NMR	UP: glutamine, isoleucine, leucine, valine and phenylalanine; DOWN: NADH and succinate.
Sadek et al. (2021)	Mice	Muscle	LC-MS GC-MS	UP: arginine, lysine, tryptophan, and methylhistidine; DOWN: pyruvate, α-ketoglutarate and succinate.
Zhou et al. (2021)	Mice	Muscle	NMR	Enhanced muscular proteolysis, suppressed glycolysis and ketone body oxidation.

On the other hand, Tseng and colleagues performed in-depth assessments of anti-cachectic activities of a novel histone deacetylase inhibitor AR-42 in C26 and LLC mouse models (Tseng et al., 2015). The LC-MS-based metabolomic analysis displayed that impaired glycolysis, glycogen synthesis and protein turnover in cachectic muscle tissues, could be improved by AR-42 and maintain the homeostatic metabolism relative to controls. Furthermore, Fukawa and colleagues conducted LC-MS-based metabolomic analysis integrated with transcriptomic analysis in muscles in a subcutaneous kidney murine model. They found that tumor-secreted factors induced excessive fatty acid oxidation, leading to muscle tissue dysfunction and activated p38 pathway. Afterwards, they indicated that drug inhibition of fatty acid oxidation could ameliorate human myotubes atrophy *in vitro*, and further restore muscle mass and body weight of mice *in vivo* (Fukawa et al., 2016). This work provided the inspiration on how to use non-targeted metabolomic techniques to explore new therapeutic targets. Initially, the researchers elucidated metabolic alterations mainly related to excessive fatty acid oxidation in animal models by performing metabolomics analysis. Then, they conducted pharmacological inhibition of fatty acid oxidation based on *in vivo* and *in vitro* models. Ultimately, they successfully rescued body weight loss and muscle atrophy in CC mice.

Recently, our group performed integrative NMR-based metabolomic and transcriptomic analyses of gastrocnemius in two murine models of CC (CT26 and LLC), and evaluated the beneficial effects of amiloride for CC treatments (Zhou et al., 2021). We identified significantly impaired metabolic pathways including enhanced muscular proteolysis, suppressed glycolysis and ketone body oxidation in cachectic gastrocnemius. Our results indicated that amiloride can alleviate muscle loss and the progression of CC through blocking exosome release originated from cancer cells. Our study suggests that tumor-released exosome can be a potential target to attenuate muscle wasting during the progression of CC in the future.

In addition to the drug prevention, nutritional supplementation of metabolites such as BCAAs has been applied to improve impaired skeletal muscle metabolisms in diseases like AIDS and diabetes (Viana and Gomes-Marcondes, 2013). Furthermore, previous studies have demonstrated that leucine supplementation can promote nitrogen balance and restore muscle mass (Ventrucci et al., 2004; Salomao and Gomes-Marcondes, 2012). This team assessed if a leucine-rich diet could affect metabolic profiles of sera and tumor tissues in a rat model. The results exhibited down-regulated levels of tryptophan and lactate associated with a suppressed hypermetabolic state, and up-regulated levels of β-hydroxybutyrate and acetoacetate, which might indirectly contribute to the prevention of CC (Viana et al., 2016). Recently, this group conducted metabolomic analyses of sera and gastrocnemius derived from rats with leucine supplementation. The tumor-bearing rats displayed distinctly altered metabolic pathways including protein biosynthesis, glycine, serine and threonine metabolism, and ammonia recycling. Significantly, the leucine-rich diet rats showed attenuated Warburg effect and improved lipid metabolism (Miyaguti et al., 2020).

Ketone body supplementation might also contribute to regulation of glucose and lipid metabolism and prevent body weight loss (Kennedy et al., 2007). Shukla and colleagues addressed anti-cancerous and anti-cachectic properties of a ketogenic diet *in vitro*, and assessed the effects of ketone bodies on tumor mass and CC symptoms of mice by conducting NMR-based metabolomic analysis *in vivo* (Shukla et al., 2014). They observed reduced glycolytic flux and diminished glutamine uptake, decreased overall ATP content in tumor cells. These results suggest that treatment with ketone bodies could prevent cachexia phenotype. Collectively, we anticipate that exploitation of the global metabolome with metabolomics techniques can achieve more comprehensive knowledge of CC and discover effective therapeutic strategies.

CONCLUSION

Even though metabolomics is relatively less used compared with other omics approaches, it is able to provide key information for further exploration of CC, including mechanistic understanding, potential biomarkers, metabolic signatures, and therapeutic strategies. With the rapid development and wide application of metabolomics analysis in the field of CC research, in-depth understandings of CC have been broadly expanded and systemized. Researchers propose novel hypotheses and develop approaches using metabolomic techniques, to exploit the features of CC and therapeutic targets for the treatments of CC. Metabolomics can be employed to identify potential biomarkers for screening early symptoms and monitoring the progression of CC, through measuring alterations in concentrations of hundreds of endogenous metabolites in bio-fluids and tissues derived from animal and human beings. In addition, numerous studies have shown that targeting specific metabolic pathways could regulate abnormal metabolisms induced by CC and ultimately alleviate syndromes of CC. **Table 2** displays the summary of the metabolomics studies on CC in the past decade with novel discoveries.

Current studies indicate that the metabolites of carbohydrates, lipids and amino acids are closely linked to the development and progression of CC (**Figure 2**). Carbohydrates related to CC primarily include glucose and lactate, TCA cycle metabolites such as citrate, succinate, and α-ketoglutarate. Lipids relevant to CC mainly include glycerophospholipids, LDL and lipid derivatives. Amino acids participating in the pathogenesis of CC mostly include BCAAs, phenylalanine and their metabolites. In addition, three kinds of ketone bodies and methylhistidine and its metabolites are also important substances involved in the pathological mechanisms of CC.

Significantly impaired metabolic pathways are associated with the pathogenesis of CC, including two main types: energy metabolism and amino acid metabolism (**Figure 2**). The metabolism of amino acids is usually disordered in CC mainly due to muscle turnover imbalance, such as BCAAs metabolism, arginine metabolism, glutamate and glutamine metabolism, phenylalanine and tyrosine metabolism. Glycolysis, fatty acid oxidation, and TCA cycle are mostly disturbed because of the shifted energy needs. Additionally, impaired metabolisms of carbohydrates and lipids contribute to the progression of CC via a series of metabolic pathways.

FUTURE PERSPECTIVES

Although the metabolic signatures of cachectic muscle are being investigated in the past decade, we still need to know how to elucidate the molecular mechanisms based on the results obtained from metabolomic analyses. The chemical complexity and large number of metabolites might be one of the challenges associated with metabolomic analyses. For example, metabolite compositions of sera, plasma and urine are manifestations of tumor, liver, muscle, and functions of gut microbes, type of diets, clinical cancer treatment, and other tumor-derived factors like exosomes and cytokines. Compared with other omics approaches, metabolomics has many advantages and also some drawbacks. No techniques are really flawless as a fact. Expectedly, metabolomic analyses should be integrated with other omics approaches, bioinformatics, biophysical techniques and signaling pathway analysis, which would provide comprehensive views on the complicated pathogenesis of CC, and expand our knowledge of fundamental mechanisms underlying metabolic disorder and muscle wasting. As unified workflows, inexpensive equipment, and humanized acquisition software and high throughput measurements as well as powerful computational analysis become more broadly available, metabolomics will play increasingly vital roles in the studies of molecular biosciences.

AUTHOR CONTRIBUTIONS

The conception and design were performed by PC, CH, and DL. The manuscript was drafted by PC. The manuscript was discussed and revised by PC, XL, QL, CH, and DL. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 31971357) and the Open Research Fund of State Key Laboratory of Cellular Stress Biology, Xiamen University (SKLCSB2020KF002).

REFERENCES

Adams, V., Späte, U., Kränkel, N., Schulze, P. C., Linke, A., Schuler, G., et al. (2003). Nuclear Factor-Kappa B Activation in Skeletal Muscle of Patients with Chronic Heart Failure: Correlation with the Expression of Inducible Nitric Oxide Synthase. *Eur. J. Cardiovasc. Prev. Rehabil.* 10 (4), 273–277. doi:10.1097/00149831-200308000-00009

Agusti, A., Morla, M., Sauleda, J., Saus, C., and Busquets, X. (2004). NF- B Activation and iNOS Upregulation in Skeletal Muscle of Patients with COPD and Low Body Weight. *Thorax* 59 (6), 483–487. doi:10.1136/thx.2003.017640

Argilés, J. M., López-Soriano, F. J., and Busquets, S. (2013). Mechanisms and Treatment of Cancer Cachexia. *Nutr. Metab. Cardiovasc. Dis.* 23 (Suppl. 1), S19–S24. doi:10.1016/j.numecd.2012.04.011

Argilés, J. M., and Stemmler, B. (2013). The Potential of Ghrelin in the Treatment of Cancer Cachexia. *Expert Opin. Biol. Ther.* 13 (1), 67–76. doi:10.1517/14712598.2013.727390

Ballarò, R., Lopalco, P., Audrito, V., Beltrà, M., Pin, F., Angelini, R., et al. (2021). Targeting Mitochondria by SS-31 Ameliorates the Whole Body Energy Status in Cancer- and Chemotherapy-Induced Cachexia. *Cancers* 13 (4), 850. doi:10.3390/cancers13040850

Beckonert, O., Keun, H. C., Ebbels, T. M. D., Bundy, J., Holmes, E., Lindon, J. C., et al. (2007). Metabolic Profiling, Metabolomic and Metabonomic Procedures for NMR Spectroscopy of Urine, Plasma, Serum and Tissue Extracts. *Nat. Protoc.* 2 (11), 2692–2703. doi:10.1038/nprot.2007.376

B. Heymsfield, S., Wang, Z., Baumgartner, R. N., and Ross, R. (1997). Human Body Composition: Advances in Models and Methods. *Annu. Rev. Nutr.* 17, 527–558. doi:10.1146/annurev.nutr.17.1.527

Bindels, L. B., Beck, R., Schakman, O., Martin, J. C., De Backer, F., Sohet, F. M., et al. (2012). Restoring Specific Lactobacilli Levels Decreases Inflammation and Muscle Atrophy Markers in an Acute Leukemia Mouse Model. *PLoS One* 7 (6), e37971. doi:10.1371/journal.pone.0037971

Bindels, L. B., Neyrinck, A. M., Claus, S. P., Le Roy, C. I., Grangette, C., Pot, B., et al. (2016). Synbiotic Approach Restores Intestinal Homeostasis and Prolongs Survival in Leukaemic Mice with Cachexia. *ISME J.* 10 (6), 1456–1470. doi:10.1038/ismej.2015.209

Blum, D., Omlin, A., Omlin, A., Fearon, K., Baracos, V., Radbruch, L., et al. (2010). Evolving Classification Systems for Cancer Cachexia: Ready for Clinical Practice? *Support Care Cancer* 18 (3), 273–279. doi:10.1007/s00520-009-0800-6

Bodine, S. C., Stitt, T. N., Gonzalez, M., Kline, W. O., Stover, G. L., Bauerlein, R., et al. (2001). Akt/mTOR Pathway Is a Crucial Regulator of Skeletal Muscle Hypertrophy and Can Prevent Muscle Atrophy *In Vivo*. *Nat. Cel Biol* 3 (11), 1014–1019. doi:10.1038/ncb1101-1014

Boguszewicz, Ł., Bieleń, A., Mrochem-Kwarciak, J., Skorupa, A., Ciszek, M., Heyda, A., et al. (2019). NMR-based Metabolomics in Real-Time Monitoring of Treatment Induced Toxicity and Cachexia in Head and Neck Cancer: a Method for Early Detection of High Risk Patients. *Metabolomics* 15 (8), 110. doi:10.1007/s11306-019-1576-4

Bonaldo, P., and Sandri, M. (2013). Cellular and Molecular Mechanisms of Muscle Atrophy. *Dis. Model. Mech.* 6 (1), 25–39. doi:10.1242/dmm.010389

Bothwell, J. H. F., and Griffin, J. L. (2011). An Introduction to Biological Nuclear Magnetic Resonance Spectroscopy. *Biol. Rev. Camb Philos. Soc.* 86 (2), 493–510. doi:10.1111/j.1469-185X.2010.00157.x

Bozzetti, F., and Mariani, L. (2009). Defining and Classifying Cancer Cachexia: a Proposal by the SCRINIO Working Group. *JPEN J. Parenter. Enteral Nutr.* 33 (4), 361–367. doi:10.1177/0148607108325076

Cala, M. P., Agulló-Ortuño, M. T., Prieto-García, E., González-Riano, C., Parrilla-Rubio, L., Barbas, C., et al. (2018). Multiplatform Plasma Fingerprinting in Cancer Cachexia: a Pilot Observational and Translational Study. *J. Cachexia, Sarcopenia Muscle* 9 (2), 348–357. doi:10.1002/jcsm.12270

Chiocchetti, G. d. M. e., Lopes-Aguiar, L., Miyaguti, N. A. d. S., Viana, L. R., Salgado, C. d. M., Orvoën, O. O., et al. (2021). A Time-Course Comparison of Skeletal Muscle Metabolomic Alterations in Walker-256 Tumour-Bearing Rats at Different Stages of Life. *Metabolites* 11 (6), 404. doi:10.3390/metabo11060404

Cui, P., Huang, C., Guo, J., Wang, Q., Liu, Z., Zhuo, H., et al. (2019a). Metabolic Profiling of Tumors, Sera, and Skeletal Muscles from an Orthotopic Murine Model of Gastric Cancer Associated-Cachexia. *J. Proteome Res.* 18 (4), 1880–1892. doi:10.1021/acs.jproteome.9b00088

Cui, P., Shao, W., Huang, C., Wu, C.-J., Jiang, B., and Lin, D. (2019b). Metabolic Derangements of Skeletal Muscle from a Murine Model of Glioma Cachexia. *Skeletal Muscle* 9 (1), 3. doi:10.1186/s13395-018-0188-4

Deans, D. A. C., Tan, B. H., Wigmore, S. J., Ross, J. A., de Beaux, A. C., Paterson-Brown, S., et al. (2009). The Influence of Systemic Inflammation, Dietary Intake and Stage of Disease on Rate of Weight Loss in Patients with Gastro-Oesophageal Cancer. *Br. J. Cancer* 100 (1), 63–69. doi:10.1038/sj.bjc.6604828

Der-Torossian, H., Asher, S. A., Winnike, J. H., Wysong, A., Yin, X., Willis, M. S., et al. (2013a). Cancer Cachexia's Metabolic Signature in a Murine Model Confirms a Distinct Entity. *Metabolomics* 9 (3), 730–739. doi:10.1007/s11306-012-0485-6

Der-Torossian, H., Wysong, A., Shadfar, S., Willis, M. S., McDunn, J., and Couch, M. E. (2013b). Metabolic Derangements in the Gastrocnemius and the Effect of Compound A Therapy in a Murine Model of Cancer Cachexia. *J. Cachexia Sarcopenia Muscle* 4 (2), 145–155. doi:10.1007/s13539-012-0101-7

Diffee, G. M., Kalfas, K., Al-Majid, S., and McCarthy, D. O. (2002). Altered Expression of Skeletal Muscle Myosin Isoforms in Cancer Cachexia. *Am. J. Physiology-Cell Physiol.* 283 (5), C1376–C1382. doi:10.1152/ajpcell.00154.2002

Dunn, W. B., Broadhurst, D. I., Atherton, H. J., Goodacre, R., and Griffin, J. L. (2011). Systems Level Studies of Mammalian Metabolomes: the Roles of Mass Spectrometry and Nuclear Magnetic Resonance Spectroscopy. *Chem. Soc. Rev.* 40 (1), 387–426. doi:10.1039/b906712b

Eisner, R., Stretch, C., Eastman, T., Xia, J., Hau, D., Damaraju, S., et al. (2011). Learning to Predict Cancer-Associated Skeletal Muscle Wasting from 1H-NMR Profiles of Urinary Metabolites. *Metabolomics* 7 (1), 25–34. doi:10.1007/s11306-010-0232-9

Emwas, A.-H. M. (2015). The Strengths and Weaknesses of NMR Spectroscopy and Mass Spectrometry with Particular Focus on Metabolomics Research. *Methods Mol. Biol.* 1277, 161–193. doi:10.1007/978-1-4939-2377-9_13

Evans, W. J., Morley, J. E., Argilés, J., Bales, C., Baracos, V., Guttridge, D., et al. (2008). Cachexia: A New Definition. *Clin. Nutr.* 27 (6), 793–799. doi:10.1016/j.clnu.2008.06.013

Evans, W. J. (2010). Skeletal Muscle Loss: Cachexia, Sarcopenia, and Inactivity. *Am. J. Clin. Nutr.* 91 (4), 1123S–1127S. doi:10.3945/ajcn.2010.28608A

Fearon, K. C. H. (1992). The Mechanisms and Treatment of Weight Loss in Cancer. *Proc. Nutr. Soc.* 51 (2), 251–265. doi:10.1079/pns19920036

Fearon, K., Strasser, F., Anker, S. D., Bosaeus, I., Bruera, E., Fainsinger, R. L., et al. (2011). Definition and Classification of Cancer Cachexia: an International Consensus. *Lancet Oncol.* 12 (5), 489–495. doi:10.1016/S1470-2045(10)70218-7

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R. N., and Willmitzer, L. (2000). Metabolite Profiling for Plant Functional Genomics. *Nat. Biotechnol.* 18 (11), 1157–1161. doi:10.1038/81137

Fujiwara, Y., Kobayashi, T., Chayahara, N., Imamura, Y., Toyoda, M., Kiyota, N., et al. (2014). Metabolomics Evaluation of Serum Markers for Cachexia and Their Intra-day Variation in Patients with Advanced Pancreatic Cancer. *PLoS One* 9 (11), e113259. doi:10.1371/journal.pone.0113259

Fukawa, T., Yan-Jiang, B. C., Min-Wen, J. C., Jun-Hao, E. T., Huang, D., Qian, C.-N., et al. (2016). Excessive Fatty Acid Oxidation Induces Muscle Atrophy in Cancer Cachexia. *Nat. Med.* 22 (6), 666–671. doi:10.1038/nm.4093

Gallagher, I. J., Jacobi, C., Tardif, N., Rooyackers, O., and Fearon, K. (2016). Omics/systems Biology and Cancer Cachexia. *Semin. Cel Develop. Biol.* 54, 92–103. doi:10.1016/j.semcdb.2015.12.022

Gao, H., Lu, Q., Liu, X., Cong, H., Zhao, L., Wang, H., et al. (2009). Application of1H NMR-Based Metabonomics in the Study of Metabolic Profiling of Human Hepatocellular Carcinoma and Liver Cirrhosis. *Cancer Sci.* 100 (4), 782–785. doi:10.1111/j.1349-7006.2009.01086.x

Gupta, S. C., Kim, J. H., Kannappan, R., Reuter, S., Dougherty, P. M., and Aggarwal, B. B. (2011). Role of Nuclear Factor-Kb-Mediated Inflammatory Pathways in Cancer-Related Symptoms and Their Regulation by Nutritional Agents. *Exp. Biol. Med. (Maywood)* 236 (6), 658–671. doi:10.1258/ebm.2011.011028

Halle, J. L., Counts, B. R., and Carson, J. A. (2020). Exercise as a Therapy for Cancer-Induced Muscle Wasting. *Sports Med. Health Sci.* 2 (4), 186–194. doi:10.1016/j.smhs.2020.11.004

Hamerman, D. (2002). Molecular-based Therapeutic Approaches in Treatment of Anorexia of Aging and Cancer Cachexia. *Journals Gerontol. Ser. A: Biol. Sci. Med. Sci.* 57 (8), M511–M518. doi:10.1093/gerona/57.8.m511

Idborg-Björkman, H., Edlund, P.-O., Kvalheim, O. M., Schuppe-Koistinen, I., and Jacobsson, S. P. (2003). Screening of Biomarkers in Rat Urine Using LC/electrospray Ionization-MS and Two-Way Data Analysis. *Anal. Chem.* 75 (18), 4784–4792. doi:10.1021/ac0341618

Jonsson, P., Bruce, S. J., Moritz, T., Trygg, J., Sjöström, M., Plumb, R., et al. (2005). Extraction, Interpretation and Validation of Information for Comparing Samples in Metabolic LC/MS Data Sets. *Analyst* 130 (5), 701–707. doi:10.1039/b501890k

Kennedy, A. R., Pissios, P., Otu, H., Xue, B., Asakura, K., Furukawa, N., et al. (2007). A High-Fat, Ketogenic Diet Induces a Unique Metabolic State in Mice. *Am. J. Physiology-Endocrinology Metab.* 292 (6), E1724–E1739. doi:10.1152/ajpendo.00717.2006

Kumar, N. B., Kazi, A., Smith, T., Crocker, T., Yu, D., Reich, R. R., et al. (2010). Cancer Cachexia: Traditional Therapies and Novel Molecular Mechanism-Based Approaches to Treatment. *Curr. Treat. Options. Oncol.* 11 (3-4), 107–117. doi:10.1007/s11864-010-0127-z

Kunz, H. E., Dorschner, J. M., Berent, T. E., Meyer, T., Wang, X., Jatoi, A., et al. (2020). Methylarginine Metabolites Are Associated with Attenuated Muscle Protein Synthesis in Cancer-Associated Muscle Wasting. *J. Biol. Chem.* 295 (51), 17441–17459. doi:10.1074/jbc.RA120.014884

Lallukka, S., and Yki-Järvinen, H. (2016). Non-alcoholic Fatty Liver Disease and Risk of Type 2 Diabetes. *Best Pract. Res. Clin. Endocrinol. Metab.* 30 (3), 385–395. doi:10.1016/j.beem.2016.06.006

Lautaoja, J. H., Lalowski, M., Nissinen, T. A., Hentilä, J., Shi, Y., Ritvos, O., et al. (2019). Muscle and Serum Metabolomes Are Dysregulated in colon-26 Tumor-Bearing Mice Despite Amelioration of Cachexia with Activin Receptor Type 2B Ligand Blockade. *Am. J. Physiology-Endocrinology Metab.* 316 (5), E852–E865. doi:10.1152/ajpendo.00526.2018

Lee, S.-J. (2004). Regulation of Muscle Mass by Myostatin. *Annu. Rev. Cel Dev. Biol.* 20, 61–86. doi:10.1146/annurev.cellbio.20.012103.135836

Legido-Quigley, C., Stella, C., Perez-Jimenez, F., Lopez-Miranda, J., Ordovas, J., Powell, J., et al. (2010). Liquid Chromatography-Mass Spectrometry Methods for Urinary Biomarker Detection in Metabonomic Studies with Application to Nutritional Studies. *Biomed. Chromatogr.* 24 (7), 737–743. doi:10.1002/bmc.1357

Li, A.-P., Li, Z.-Y., Sun, H.-F., Li, K., Qin, X.-M., and Du, G.-H. (2015). Comparison of Two Different Astragali Radix by a 1H NMR-Based Metabolomic Approach. *J. Proteome Res.* 14 (5), 2005–2016. doi:10.1021/pr501167u

Lim, S., Brown, J. L., Washington, T. A., and Greene, N. P. (2020). Development and Progression of Cancer Cachexia: Perspectives from Bench to Bedside. *Sports Med. Health Sci.* 2 (4), 177–185. doi:10.1016/j.smhs.2020.10.003

Loberg, R. D., Bradley, D. A., Tomlins, S. A., Chinnaiyan, A. M., and Pienta, K. J. (2007). The Lethal Phenotype of Cancer: The Molecular Basis of Death Due to Malignancy. *CA: A Cancer J. Clinicians* 57 (4), 225–241. doi:10.3322/canjclin.57.4.225

Lok, C. (2015). Cachexia: The Last Illness. *Nature* 528 (7581), 182–183. doi:10.1038/528182a

Ma, Q., Li, Y., Wang, M., Tang, Z., Wang, T., Liu, C., et al. (2018). Progress in Metabonomics of Type 2 Diabetes Mellitus. *Molecules* 23 (7), 1834. doi:10.3390/molecules23071834

Miller, J., Alshehri, A., Ramage, M. I., Stephens, N. A., Mullen, A. B., Boyd, M., et al. (2019). Plasma Metabolomics Identifies Lipid and Amino Acid Markers of Weight Loss in Patients with Upper Gastrointestinal Cancer. *Cancers* 11 (10), 1594. doi:10.3390/cancers11101594

Mitsiopoulos, N., Baumgartner, R. N., Heymsfield, S. B., Lyons, W., Gallagher, D., and Ross, R. (19981985). Cadaver Validation of Skeletal Muscle Measurement by Magnetic Resonance Imaging and Computerized Tomography. *J. Appl. Physiol.* 85 (1), 115–122. doi:10.1152/jappl.1998.85.1.115

Miyaguti, N. A. d. S., Stanisic, D., Oliveira, S. C. P. d., Dos Santos, G. S., Manhe, B. S., Tasic, L., et al. (2020). Serum and Muscle 1H NMR-Based Metabolomics Profiles Reveal Metabolic Changes Influenced by a Maternal Leucine-Rich Diet in Tumor-Bearing Adult Offspring Rats. *Nutrients* 12 (7), 2106. doi:10.3390/nu12072106

Mourtzakis, M., Prado, C. M. M., Lieffers, J. R., Reiman, T., McCargar, L. J., and Baracos, V. E. (2008). A Practical and Precise Approach to Quantification of Body Composition in Cancer Patients Using Computed Tomography Images Acquired during Routine Care. *Appl. Physiol. Nutr. Metab.* 33 (5), 997–1006. doi:10.1139/H08-075

Muroya, S., Ueda, S., Komatsu, T., Miyakawa, T., and Ertbjerg, P. (2020). MEATabolomics: Muscle and Meat Metabolomics in Domestic Animals. *Metabolites* 10 (5), 188. doi:10.3390/metabo10050188

Musarò, A., McCullagh, K., Paul, A., Houghton, L., Dobrowolny, G., Molinaro, M., et al. (2001). Localized Igf-1 Transgene Expression Sustains Hypertrophy and Regeneration in Senescent Skeletal Muscle. *Nat. Genet.* 27 (2), 195–200. doi:10.1038/84839

Nagaya, N., Kojima, M., and Kangawa, K. (2006). Ghrelin, a Novel Growth Hormonereleasing Peptide, in the Treatment of Cardiopulmonaryassociated Cachexia. *Intern. Med.* 45 (3), 127–134. doi:10.2169/internalmedicine.45.1402

Newgard, C. B., An, J., Bain, J. R., Muehlbauer, M. J., Stevens, R. D., Lien, L. F., et al. (2009). A Branched-Chain Amino Acid-Related Metabolic Signature that Differentiates Obese and Lean Humans and Contributes to Insulin Resistance. *Cel Metab.* 9 (4), 311–326. doi:10.1016/j.cmet.2009.02.002

Newgard, C. B. (2017). Metabolomics and Metabolic Diseases: Where Do We Stand? *Cel Metab.* 25 (1), 43–56. doi:10.1016/j.cmet.2016.09.018

Ni, Y., Lohinai, Z., Heshiki, Y., Dome, B., Moldvay, J., Dulka, E., et al. (2021). Distinct Composition and Metabolic Functions of Human Gut Microbiota Are Associated with Cachexia in Lung Cancer Patients. *ISME J.* 15, 3207–3220. doi:10.1038/s41396-021-00998-8

Nicholson, J. K., Lindon, J. C., and Holmes, E. (1999). 'Metabonomics': Understanding the Metabolic Responses of Living Systems to Pathophysiological Stimuli via Multivariate Statistical Analysis of Biological NMR Spectroscopic Data. *Xenobiotica* 29 (11), 1181–1189. doi:10.1080/004982599238047

Nikolic, S. B., Sharman, J. E., Adams, M. J., and Edwards, L. M. (2014). Metabolomics in Hypertension. *J. Hypertens.* 32 (6), 1159–1169. doi:10.1097/HJH.0000000000000168

O'Connell, T. M., Ardeshirpour, F., Asher, S. A., Winnike, J. H., Yin, X., George, J., et al. (2008). Metabolomic Analysis of Cancer Cachexia Reveals Distinct Lipid and Glucose Alterations. *Metabolomics* 4 (3), 216–225. doi:10.1007/s11306-008-0113-7

Ohbuchi, K., Nishiumi, S., Fujitsuka, N., Hattori, T., Yamamoto, M., Inui, A., et al. (2015). Rikkunshito Ameliorates Cancer Cachexia Partly through Elevation of Glucarate in Plasma. *Evidence-Based Complement. Altern. Med.* 2015, 1–11. doi:10.1155/2015/871832

Pang, Z., Chong, J., Zhou, G., de Lima Morais, D. A., Chang, L., Barrette, M., et al. (2021). MetaboAnalyst 5.0: Narrowing the gap between Raw Spectra and Functional Insights. *Nucleic Acids Res.* 49 (W1), W388–W396. doi:10.1093/nar/gkab382

Penna, F., Ballarò, R., Martinez-Cristobal, P., Sala, D., Sebastian, D., Busquets, S., et al. (2019). Autophagy Exacerbates Muscle Wasting in Cancer Cachexia and Impairs Mitochondrial Function. *J. Mol. Biol.* 431 (15), 2674–2686. doi:10.1016/j.jmb.2019.05.032

Peterson, J. M., Bakkar, N., and Guttridge, D. C. (2011). NF-κB Signaling in Skeletal Muscle Health and Disease. *Curr. Top. Dev. Biol.* 96, 85–119. doi:10.1016/B978-0-12-385940-2.00004-8

Pin, F., Barreto, R., Couch, M. E., Bonetto, A., and O'Connell, T. M. (2019). Cachexia Induced by Cancer and Chemotherapy Yield Distinct Perturbations to Energy Metabolism. *J. Cachexia, Sarcopenia Muscle* 10 (1), 140–154. doi:10.1002/jcsm.12360

Pötgens, S. A., Thibaut, M. M., Joudiou, N., Sboarina, M., Neyrinck, A. M., Cani, P. D., et al. (2021). Multi-compartment Metabolomics and Metagenomics Reveal Major Hepatic and Intestinal Disturbances in Cancer Cachectic Mice. *J. Cachexia, Sarcopenia Muscle* 12 (2), 456–475. doi:10.1002/jcsm.12684

Quan-Jun, Y., Jun, B., Li-Li, W., Yong-Long, H., Bin, L., Qi, Y., et al. (2015). NMR-based Metabolomics Reveals Distinct Pathways Mediated by Curcumin in Cachexia Mice Bearing CT26 Tumor. *RSC Adv.* 5 (16), 11766–11775. doi:10.1039/c4ra14128h

Quanjun, Y., Genjin, Y., Lili, W., Bin, L., Jin, L., Qi, Y., et al. (2013). Serum Metabolic Profiles Reveal the Effect of Formoterol on Cachexia in Tumor-Bearing Mice. *Mol. Biosyst.* 9 (12), 3015–3025. doi:10.1039/c3mb70134d

QuanJun, Y., GenJin, Y., LiLi, W., Yan, H., YongLong, H., Jin, L., et al. (2015). Integrated Analysis of Serum and Intact Muscle Metabonomics Identify Metabolic Profiles of Cancer Cachexia in a Dynamic Mouse Model. *RSC Adv.* 5 (112), 92438–92448. doi:10.1039/c5ra19004e

Ramamoorthy, S., Donohue, M., and Buck, M. (2009). Decreased Jun-D and Myogenin Expression in Muscle Wasting of Human Cachexia. *Am. J. Physiology-Endocrinology Metab.* 297 (2), E392–E401. doi:10.1152/ajpendo.90529.2008

Ramautar, R., Somsen, G. W., and de Jong, G. J. (2019). CE-MS for Metabolomics: Developments and Applications in the Period 2016-2018. *Electrophoresis* 40 (1), 165–179. doi:10.1002/elps.201800323

Rohm, M., Zeigerer, A., Machado, J., and Herzig, S. (2019). Energy Metabolism in Cachexia. *EMBO Rep.* 20 (4). doi:10.15252/embr.201847258

Rolland, Y., Van Kan, G. A., Gillette-Guyonnet, S., and Vellas, B. (2011). Cachexia versus Sarcopenia. *Curr. Opin. Clin. Nutr. Metab. Care* 14 (1), 15–21. doi:10.1097/MCO.0b013e328340c2c2

Römisch-Margl, W., Prehn, C., Bogumil, R., Röhring, C., Suhre, K., and Adamski, J. (2011). Procedure for Tissue Sample Preparation and Metabolite Extraction for High-Throughput Targeted Metabolomics. *Metabolomics* 8 (1), 133–142. doi:10.1007/s11306-011-0293-4

Sadek, J., Hall, D. T., Colalillo, B., Omer, A., Tremblay, A. M. K., Sanguin-Gendreau, V., et al. (2021). Pharmacological or Genetic Inhibition of iNOS Prevents Cachexia-mediated Muscle Wasting and its Associated Metabolism Defects. *EMBO Mol. Med.* 13 (7), e13591. doi:10.15252/emmm.202013591

Salomão, E. M., and Gomes-Marcondes, M. C. C. (2012). Light Aerobic Physical Exercise in Combination with Leucine And/or Glutamine-Rich Diet Can Improve the Body Composition and Muscle Protein Metabolism in Young Tumor-Bearing Rats. *J. Physiol. Biochem.* 68 (4), 493–501. doi:10.1007/s13105-012-0164-0

Sandri, M., Sandri, C., Gilbert, A., Skurk, C., Calabria, E., Picard, A., et al. (2004). Foxo Transcription Factors Induce the Atrophy-Related Ubiquitin Ligase Atrogin-1 and Cause Skeletal Muscle Atrophy. *Cell* 117 (3), 399–412. doi:10.1016/s0092-8674(04)00400-3

Santarpia, L., Contaldo, F., and Pasanisi, F. (2011). Nutritional Screening and Early Treatment of Malnutrition in Cancer Patients. *J. Cachexia Sarcopenia Muscle* 2 (1), 27–35. doi:10.1007/s13539-011-0022-x

Schmidt, D. R., Patel, R., Kirsch, D. G., Lewis, C. A., Vander Heiden, M. G., and Locasale, J. W. (2021). Metabolomics in Cancer Research and Emerging Applications in Clinical Oncology. *CA A. Cancer J. Clin.* 71 (4), 333–358. doi:10.3322/caac.21670

Schwämmle, V., Verano-Braga, T., and Roepstorff, P. (2015). Computational and Statistical Methods for High-Throughput Analysis of post-translational Modifications of Proteins. *J. Proteomics* 129, 3–15. doi:10.1016/j.jprot.2015.07.016

Shah, S. H., Kraus, W. E., and Newgard, C. B. (2012a). Metabolomic Profiling for the Identification of Novel Biomarkers and Mechanisms Related to Common Cardiovascular Diseases. *Circulation* 126 (9), 1110–1120. doi:10.1161/CIRCULATIONAHA.111.060368

Shah, S. H., Sun, J.-L., Stevens, R. D., Bain, J. R., Muehlbauer, M. J., Pieper, K. S., et al. (2012b). Baseline Metabolomic Profiles Predict Cardiovascular Events in Patients at Risk for Coronary Artery Disease. *Am. Heart J.* 163 (5), 844–850. e841. doi:10.1016/j.ahj.2012.02.005

Shen, W., Punyanitya, M., Wang, Z., Gallagher, D., St.-Onge, M.-P., Albu, J., et al. (20041985). Total Body Skeletal Muscle and Adipose Tissue Volumes: Estimation from a Single Abdominal Cross-Sectional Image. *J. Appl. Physiol.* 97 (6), 2333–2338. doi:10.1152/japplphysiol.00744.2004

Shukla, S. K., Gebregiworgis, T., Purohit, V., Chaika, N. V., Gunda, V., Radhakrishnan, P., et al. (2014). Metabolic Reprogramming Induced by Ketone Bodies Diminishes Pancreatic Cancer Cachexia. *Cancer Metab.* 2, 18. doi:10.1186/2049-3002-2-18

Stolz, A., Jooss, K., Höcker, O., Römer, J., Schlecht, J., and Neusüß, C. (2019). Recent Advances in Capillary Electrophoresis-Mass Spectrometry: Instrumentation, Methodology and Applications. *Electrophoresis* 40 (1), 79–112. doi:10.1002/elps.201800331

Stretch, C., Eastman, T., Mandal, R., Eisner, R., Wishart, D. S., Mourtzakis, M., et al. (2012). Prediction of Skeletal Muscle and Fat Mass in Patients with Advanced Cancer Using a Metabolomic Approach. *J. Nutr.* 142 (1), 14–21. doi:10.3945/jn.111.147751

Sugimoto, M., Kawakami, M., Robert, M., Soga, T., and Tomita, M. (2012). Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. *Cbio* 7 (1), 96–108. doi:10.2174/157489312799304431

Tisdale, M. J. (2003). Pathogenesis of Cancer Cachexia. *J. Support. Oncol.* 1 (3), 159–168.

Tisdale, M. J. (2009). Mechanisms of Cancer Cachexia. *Physiol. Rev.* 89 (2), 381–410. doi:10.1152/physrev.00016.2008

Tseng, Y.-C., Kulp, S. K., Lai, I.-L., Hsu, E.-C., He, W. A., Frankhouser, D. E., et al. (2015). Preclinical Investigation of the Novel Histone Deacetylase Inhibitor AR-42 in the Treatment of Cancer-Induced Cachexia. *JNCI.J* 107 (12), djv274. doi:10.1093/jnci/djv274

Tsiropoulou, S., McBride, M., and Padmanabhan, S. (2017). Urine Metabolomics in Hypertension Research. *Methods Mol. Biol.* 1527, 61–68. doi:10.1007/978-1-4939-6625-7_5

Twelkmeyer, B., Tardif, N., and Rooyackers, O. (2017). Omics and Cachexia. *Curr. Opin. Clin. Nutr. Metab. Care* 20 (3), 181–185. doi:10.1097/Mco.0000000000000363

Uzu, M., Nonaka, M., Miyano, K., Sato, H., Kurebayashi, N., Yanagihara, K., et al. (2019). A Novel Strategy for Treatment of Cancer Cachexia Targeting Xanthine Oxidase in the Brain. *J. Pharmacol. Sci.* 140 (1), 109–112. doi:10.1016/j.jphs.2019.04.005

Valdes, A. M., Walter, J., Segal, E., and Spector, T. D. (2018). Role of the Gut Microbiota in Nutrition and Health. *BMJ* 361, k2179. doi:10.1136/bmj.k2179

Ventrucci, G., Ramos Silva, L. G., Roston Mello, M. A., and Gomes Marcondes, M. C. C. (2004). Effects of a Leucine-Rich Diet on Body Composition during Nutritional Recovery in Rats. *Nutrition* 20 (2), 213–217. doi:10.1016/j.nut.2003.10.014

Viana, L. R., Canevarolo, R., Luiz, A. C. P., Soares, R. F., Lubaczeuski, C., Zeri, A. C. d. M., et al. (2016). Leucine-rich Diet Alters the 1H-NMR Based Metabolomic Profile without Changing the Walker-256 Tumour Mass in Rats. *BMC Cancer* 16 (1), 764. doi:10.1186/s12885-016-2811-2

Viana, L. R., and Gomes-Marcondes, M. C. C. (2013). Leucine-Rich Diet Improves the Serum Amino Acid Profile and Body Composition of Fetuses from Tumor-Bearing Pregnant Mice1. *Biol. Reprod.* 88 (5), 121. doi:10.1095/biolreprod.112.107276

Viana, L. R., Lopes-Aguiar, L., Rossi Rosolen, R., Willians Dos Santos, R., and Cintra Gomes-Marcondes, M. C. (2020). 1H-NMR Based Serum Metabolomics Identifies Different Profile between Sarcopenia and Cancer Cachexia in Ageing Walker 256 Tumour-Bearing RatsH-NMR Based Serum Metabolomics Identifies Different Profile between Sarcopenia and Cancer Cachexia in Ageing Walker 256 Tumour-Bearing Rats. *Metabolites* 10 (4), 161. doi:10.3390/metabo10040161

Waddell, D. S., Baehr, L. M., van den Brandt, J., Johnsen, S. A., Reichardt, H. M., Furlow, J. D., et al. (2008). The Glucocorticoid Receptor and FOXO1 Synergistically Activate the Skeletal Muscle Atrophy-Associated MuRF1 Gene. *Am. J. Physiology-Endocrinology Metab.* 295 (4), E785–E797. doi:10.1152/ajpendo.00646.2007

Wiklund, S., Johansson, E., Sjöström, L., Mellerowicz, E. J., Edlund, U., Shockcor, J. P., et al. (2008). Visualization of GC/TOF-MS-based Metabolomics Data for Identification of Biochemically Interesting Compounds Using OPLS Class Models. *Anal. Chem.* 80 (1), 115–122. doi:10.1021/ac0713510

Wishart, D. S. (2016). Emerging Applications of Metabolomics in Drug Discovery and Precision Medicine. *Nat. Rev. Drug Discov.* 15 (7), 473–484. doi:10.1038/nrd.2016.32

Xia, F., and Wan, J. B. (2021). Chemical Derivatization Strategy for Mass Spectrometry-based Lipidomics. *Mass. Spec. Rev.*, e21729. doi:10.1002/mas.21729

Xiao, J. F., Zhou, B., and Ressom, H. W. (2012). Metabolite Identification and Quantitation in LC-MS/MS-based Metabolomics. *Trac Trends Anal. Chem.* 32, 1–14. doi:10.1016/j.trac.2011.08.009

Yang, Q.-J., Zhao, J.-R., Hao, J., Li, B., Huo, Y., Han, Y.-L., et al. (2018). Serum and Urine Metabolomics Study Reveals a Distinct Diagnostic Model for Cancer Cachexia. *J. Cachexia, Sarcopenia Muscle* 9 (1), 71–85. doi:10.1002/jcsm.12246

Yang, W., Huang, J., Wu, H., Wang, Y., Du, Z., Ling, Y., et al. (2020). Molecular Mechanisms of Cancer Cachexia-induced M-uscle A-trophy (Review). *Mol. Med. Rep.* 22 (6), 4967–4980. doi:10.3892/mmr.2020.11608

Yin, P., Lehmann, R., and Xu, G. (2015). Effects of Pre-analytical Processes on Blood Samples Used in Metabolomics Studies. *Anal. Bioanal. Chem.* 407 (17), 4879–4892. doi:10.1007/s00216-015-8565-x

Zhou, L., Zhang, T., Shao, W., Lu, R., Wang, L., Liu, H., et al. (2021). Amiloride Ameliorates Muscle Wasting in Cancer Cachexia through Inhibiting Tumor-Derived Exosome Release. *Skeletal Muscle* 11 (1), 17. doi:10.1186/s13395-021-00274-5

Zhou, X., Wang, J. L., Lu, J., Song, Y., Kwak, K. S., Jiao, Q., et al. (2010). Reversal of Cancer Cachexia and Muscle Wasting by ActRIIB Antagonism Leads to Prolonged Survival. *Cell* 142 (4), 531–543. doi:10.1016/j.cell.2010.07.011

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 07 February 2022
doi: 10.3389/fmolb.2022.781619

Integrated Transcriptomics and Metabolomics Analyses of Stress-Induced Murine Hair Follicle Growth Inhibition

Xuewen Wang[1,2†], *Changqing Cai*[3†], *Qichang Liang*[1,2†], *Meng Xia*[4], *Lihua Lai*[4], *Xia Wu*[1,2], *Xiaoyun Jiang*[1,2], *Hao Cheng*[1,2]*, *Yinjing Song*[1,2]* and *Qiang Zhou*[1,2]*

[1]*Department of Dermatology and Venereology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China,* [2]*Hair Research Center, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China,* [3]*Yonghe Medical Group Co. Ltd., Beijing, China,* [4]*Institute of Immunology, Zhejiang University School of Medicine, Hangzhou, China*

OPEN ACCESS

Edited by:
Wolfram Weckwerth,
University of Vienna, Austria

Reviewed by:
Francisco Solano,
University of Murcia, Spain
Fidele Tugizimana,
Omnia, South Africa

***Correspondence:**
Hao Cheng
chenghao1@zju.edu.cn
Yinjing Song
3315023@zju.edu.cn
Qiang Zhou
3197035@zju.edu.cn

†*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

Received: *23 September 2021*
Accepted: *13 January 2022*
Published: *07 February 2022*

Citation:
Wang X, Cai C, Liang Q, Xia M, Lai L, Wu X, Jiang X, Cheng H, Song Y and Zhou Q (2022) Integrated Transcriptomics and Metabolomics Analyses of Stress-Induced Murine Hair Follicle Growth Inhibition. Front. Mol. Biosci. 9:781619. doi: 10.3389/fmolb.2022.781619

Psychological stress plays an important role in hair loss, but the underlying mechanisms are not well-understood, and the effective therapies available to regrow hair are rare. In this study, we established a chronic restraint stress (CRS)-induced hair growth inhibition mouse model and performed a comprehensive analysis of metabolomics and transcriptomics. Metabolomics data analysis showed that the primary and secondary metabolic pathways, such as carbohydrate metabolism, amino acid metabolism, and lipid metabolism were significantly altered in skin tissue of CRS group. Transcriptomics analysis also showed significant changes of genes expression profiles involved in regulation of metabolic processes including arachidonic acid metabolism, glutathione metabolism, glycolysis gluconeogenesis, nicotinate and nicotinamide metabolism, purine metabolism, retinol metabolism and cholesterol metabolism. Furthermore, RNA-Seq analyses also found that numerous genes associated with metabolism were significantly changed, such as Hk-1, in CRS-induced hair growth inhibition. Overall, our study supplied new insights into the hair growth inhibition induced by CRS from the perspective of integrated metabolomics and transcriptomics analyses.

Keywords: metabolomics, transcriptomics, hair growth inhibition, chronic restraint stress, hexokinase-1

INTRODUCTION

As one of the most common skin diseases, hair loss has negative effects on patient's psychological well-being and reduces their life quality (Williamson et al., 2001). Previous studies indicate that psychoemotional stress plays a pivotal role in triggering and aggravating hair loss, such as alopecia areata (AA), telogen effluvium and androgenetic alopecia (Hadshiew et al., 2004; Peters et al., 2006; Alexopoulos and Chrousos, 2016; Dainichi and Kabashima, 2017). Numerous studies reveal that hair loss is highly related to hair follicle (HF) pathophysiological changes (Cotsarelis and Millar, 2001; Pratt et al., 2017). HFs go through successive cycles of anagen (growth), catagen (regression), and telogen (rest) phases (Millar, 2002; Rishikaysh et al., 2014). The hair follicle cycling is modulated by various signals which control quiescence and activation of hair follicle stem cells (HFSCs) (Chai et al., 2019; Feng et al., 2020).

Psychological stress has been reported to alter the hair cycle via neuroendocrine or neuroimmunological signaling pathways (Paus et al., 2008; Ito, 2010; Paus et al., 2014). The generation of HFs can be affected by numerous neuromediators which regulate HF growth, pigmentation, remodeling, immune status, stem cell biology, and energy metabolism (Peters et al., 2006; Paus et al., 2014; Choi et al., 2021). Numerous studies demonstrate that stress increases apoptotic cells, inhibits hair bulge stem cells and hair bulb keratinocytes proliferation, promotes mast cell degranulation, and induces premature catagen and neurogenic inflammation. In addition, it has been reported that chronic restraint stress (CRS) induces the delay of hair cycle via autophagy (L. Wang et al., 2015). But other researchers find that supplementation of a metabolite a-ketobutyrate (a-KB) in old mice can increase longevity and prevent alopecia by inducing autophagy (Chai et al., 2019). Thus, the mechanisms of CRS on hair growth remains to be further investigated.

Metabolomics is an omics category focused on simultaneous qualitative and quantitative analyses of low molecular-weight metabolites within an organism or cell during a specific physiological period (Christodoulou et al., 2020; Huang et al., 2021). Changes in metabolites play a critical role in various diseases, including hair loss. Clinical investigations have suggested that androgenic alopecia (AGA) patients showed significant abnormal lipid profiles (Antoni and Dhabhar, 2019; Kim et al., 2017). Lipid-modulatory therapies have been reported to alter hair growth (Lattouf et al., 2015; Cervantes et al., 2018; Shin et al., 2021). Moreover, cholesterol is involved in proliferation and differentiation of HF cell population. It is reported that primary cicatricial alopecia is also related to cholesterol metabolism disorders (Palmer et al., 2020).

Intriguingly, psychological stress can trigger metabolic changes and psychological stress is also associated with many metabolic-related diseases including diabetes, cardiovascular disease, as well as cancers (Gu et al., 2012; Hackett and Steptoe, 2017; Antoni and Dhabhar, 2019). CRS also drastically increases the expression of genes related to fatty acid/lipid/sterol metabolism in the liver of mice (Ha et al., 2003). Some researchers also report that HFSCs maintain a dormant metabolic state and could utilize glycolytic metabolism, thus producing more lactate than other cells in the epidermis (Flores et al., 2017). Small molecules that activate autophagy could initiate anagen and stimulate hair growth, including some metabolites associated with carbohydrate metabolism, a-ketoglutarate (a-KG), and a-ketobutyrate (a-KB) (Chai et al., 2019). Obesity-induced stress, such as that induced by a high-fat diet accelerates hair loss mainly through depletion of HFSCs, which indicates metabolic changes may affect hair growth via stem cell inflammatory signals (Morinaga et al., 2021). However, the metabolic pathways and molecules involved in the mechanisms of psychological stress effects on hair growth are still unclear.

Therefore, in order to elucidate the pathogenesis and explore potential therapeutic strategies, our study investigated important biological metabolites, genes and signaling pathways that were related to CRS-induced hair growth inhibition. Our results not only provided a validated and comprehensive understanding of integrated transcriptomics and metabolism analyses in CRS-induced hair growth inhibition but also found some genes including Hk-1 which might be new targets for the treatment of CRS-induced hair growth inhibition.

MATERIALS AND METHODS

Mice

All experiments were approved by the center of experiment animal, Zhejiang University (China). C57BL/6 male mice were obtained at 6–8 weeks of age from Shanghai SLAC Laboratory Animal Co., Ltd. All mice were acclimated for 7 days before the onset of studies at Experimental Animal Center of Zhejiang University (China). The standard conditions of animal facility were maintained as following: temperature 21–24°C; 12 h light/dark cycle (lights on 06:00–18:00); humidity 50–60%. Sterilized water and food were provided ad libitum during this period. The study was approved by the Ethics Committee of Sir Run Run Shaw Hospital of Zhejiang University School of Medicine (Approval no. SRRSH2021401).

Stress Application and Anagen Induction

The procedure of CRS was conducted as previously reported method and lasted for 20 days (Q. Wang et al., 2019). Mice were placed into 50 ml conical tubes, without physically compressed for 6 h (10:00–16:00) each day (Liu et al., 2013; Zhao et al., 2013). During the period of stress application, control mice were kept undisturbed in their original cages, and all groups of mice were not provided with food and water. On day 8 of the experiment, wax/rosin mixture (1:1 on weight) was applied to the dorsal skin (from neck to tail) of mice to induce anagen. Then we peeled off the mixture and removed all hair shafts to induce synchronization of hair cycle, as evidenced by the homogeneously pink skin color in the back, which indicated all hair follicles in telogen (Müller-Röver et al., 2001). Mice were not exposed to CRS on the day of depilation.

Assessment of Hair Cycle

Assessment of Hair cycle were based on the appearance of skin pigmentation and hair shaft which were monitored by pictures, as previously described (Stenn and Paus, 2001). To quantify the stage of the hair follicles, skin pigmentation score values from 0 to 100 were calculated based on skin pigmentation levels and hair shaft density, with 0 indicating no hair growth (and no pigmentation) and a higher number corresponding to darker skin and larger areas of dense hair growth (Chai et al., 2019). Briefly, skin pigmentation scored 50 refers to 50 percent of full-length hair shaft on back skin or 100 percent of skin pigmentation without visible hair growth. Skin pigmentation scored 70 refers to 70 percent of full-length hair shaft in back skin or 100 percent of skin pigmentation with 40 percent of full-length hair shaft. Skin pigmentation scored 100 refers to 100 percent of full-length hair shaft on back skin (Feng et al., 2020).

Tissue Preparation and Immunohistochemistry Staining

C57BL/6J mouse dorsal skin specimens were harvested about 2 × 4 cm on day 21 of the experiment before being collected for histological and molecular analyses. Full-thickness skin tissues (measured thickness 400–700 um) were then fixed in 4% formalin and dehydrated for embedding in paraffin. 5 mm paraffin sections were subjected to hematoxylin and eosin (H&E) staining and immunohistochemistry. The Ki67 (ab15580) antibody was purchased from Abcam. Images were captured using an Olympus microscope (IX73) at X40 and X400 magnification. The remaining skin specimens were immediately snap frozen in liquid nitrogen and stored at −80°C for subsequent use.

Library Construction, RNA Sequencing and Primary Analysis

Three replicate samples of control and CRS C57BL/6 mice dorsal skin specimens were used for library construction and RNA sequencing, respectively. Total RNA was isolated from skin tissues and purified using TRIzol reagent (Invitrogen, Carlsbad, CA, United States) following the manufacturer's procedure. The NanoDrop ND-1000 (NanoDrop, Wilmington, DE, United States) was used to quantify the amount of RNA and purity of each sample. The Bioanalyzer 2,100 (Agilent, CA, United States) with RIN number >7.0 was used to assess the integrity of RNA, which was also confirmed by electrophoresis with denaturing agarose gel. Poly (A) RNA was purified from 1 μg total RNA by Dynabeads Oligo (dT)25–61,005 (Thermo Fisher, CA, United States) and was fragmented into small pieces using Magnesium RNA Fragmentation Module (NEB, cat. e6150, United States) under 94°C 5–7 min. The SuperScript™ II Reverse Transcriptase (Invitrogen, cat. 1896649, United States) was used to reverse-transcribe the cleaved RNA fragments to create the cDNA, which were then transform to the U-labeled second-stranded DNAs with E. coli DNA polymerase I (NEB, cat. m0209, United States), RNase H (NEB, cat. m0297, United States) and dUTP Solution (Thermo Fisher, cat. R0133, United States). An A-base was then added to the blunt ends of each strand, preparing them for ligation to the sequencing adapters. Subsequently, the ligated products were amplified with PCR amplification. At last, the Illumina Novaseq™ 6,000 (LC-Bio Technology CO., Ltd., Hangzhou, China) was used to perform the 2 × 150bp paired-end sequencing (PE150). The differentially expressed mRNAs were selected with fold change >2 or fold change <0.5 and *p* value <0.05 by R package edgeR (https://bioconductor.org/packages/release/bioc/html/edgeR.html).

Metabolite Extraction and LC-MS Analysis

Metabolomics sample collection, preparation, and metabolome profiling were carried out as previously described (Ruiying et al., 2020). The back skin tissues from mice treated with CRS or control were thawed on ice, and metabolites were extracted from 20 μL of each sample using 120 μL of precooled 50% methanol buffer (methanol and distilled water were mixed in a 1:1 ratio). Then the mixture of metabolites was vortexed for 1 min and incubated for 10 min at room temperature, and stored at −20°C overnight. The mixture was centrifugated at 4,000 g for 20 min, subsequently the supernatant was transferred to 96-well plates. The samples were stored at −80 °C prior to the LC-MS analysis. Pooled quality control (QC) samples were also prepared by combining 10 μL of each extraction mixture. All samples were detected by a Triple TOF 5600 Plus high-resolution tandem mass spectrometer (SCIEX, Warrington, United Kingdom) with both positive and negative ion modes. Chromatographic separation was performed using an ultraperformance liquid chromatography (UPLC) system (SCIEX, United Kingdom). The data acquisition mode was DDA.

Data Processing and Annotation

The XCMS software was used to acquire the LC-MS pretreatment data including peak picking, peak grouping, retention time correction, second peak grouping, and annotation of isotopes and adducts. Raw data files were transformed into mzXML format and then processed by the XCMS, CAMERA and metaX toolbox included in R software. The comprehensive information of retention time and m/z data was identified for each ion, recorded the intensity of each peak, generated a three-dimensional matrix containing arbitrarily assigned peak indices (retention time-m/z pairs), sample names (observations) and ion intensity information (variables), and matched to the in-house and public database. The metabolites by matching the exact molecular mass data (m/z) to those from the database within a threshold of 10 ppm was annotated by the open access databases, Kyoto Encyclopedia of Genes and Genomes (KEGG) and Human Metabolome Database (HMDB). The metaX was used to further preprocess the peak intensity data. Those features that were detected less than 50 percent of QC samples or 80 percent of test samples were removed, and values for missing peaks were imputed with the k-nearest neighbor algorithm to improve the quality of data. Principal component analysis (PCA) was used to identify outliers and batch effects using the pre-processed dataset. To minimize signal intensity drift over time, QC-based robust LOESS signal correction was used to fit to the QC data. Besides, the relevant standard deviations of the metabolic features were calculated across all QC samples, and those with standard deviations >30 percent were removed. All the annotated secondary metabolites and their Metabolomics Standard Initiative (MSI) level are showed in **Supplementary Table S1**.

The group datasets were normalized before analysis was performed. Data normalization was carried out using the probabilistic quotient normalization algorithm. Differential enrichment of metabolite features between CRS and control groups was analyzed by Student's t-test FDR-adjusted *p*-value less than 0.05. Then, QC-robust spline batch correction was performed using QC samples. Supervised partial least-squares discriminant analysis (PLS-DA) was conducted through metaX to discriminate the different variables between the groups. The Variable Important for the Projection (VIP) cut-off value of 1.0 was set to select important features.

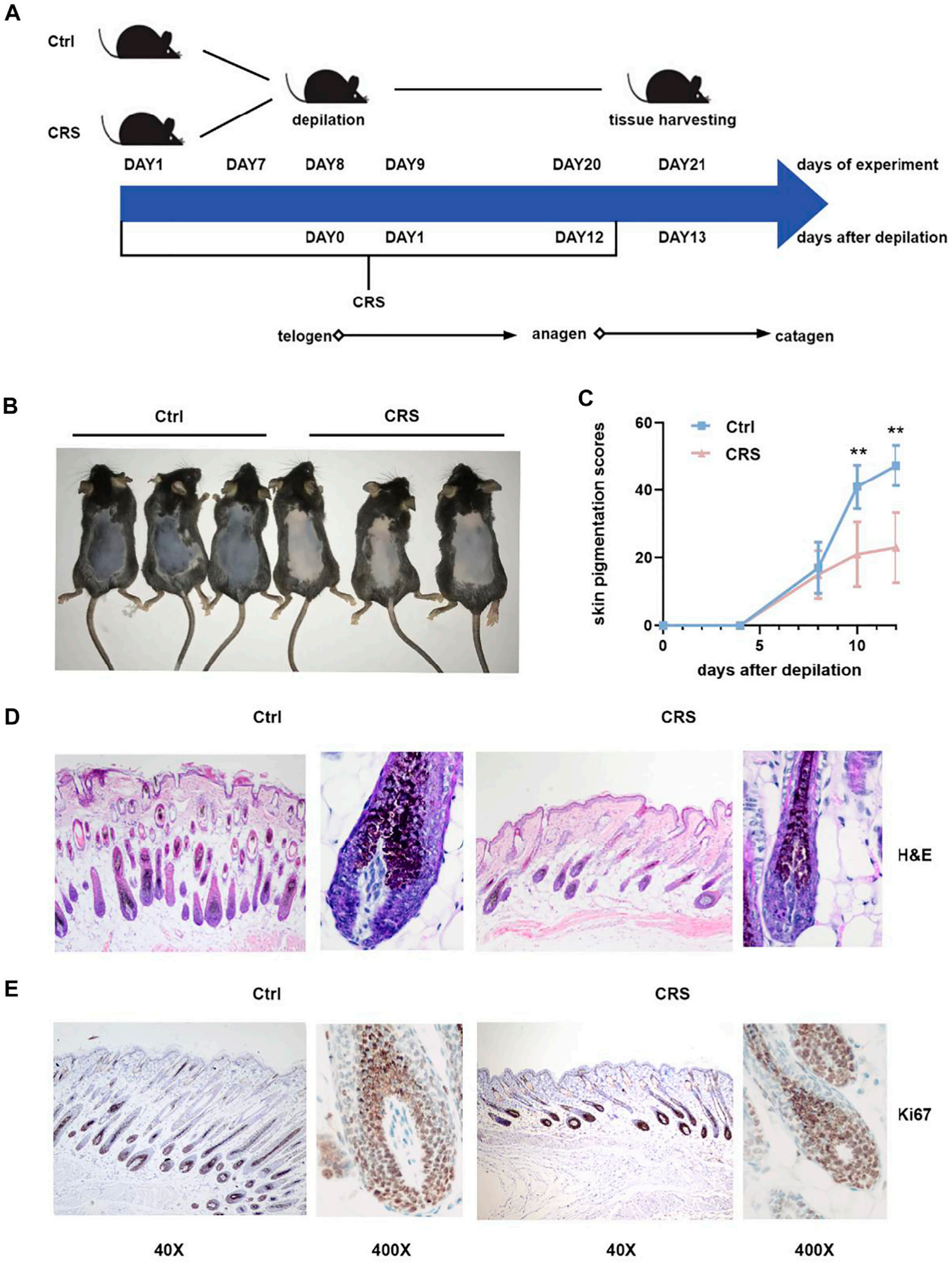


FIGURE 1 | CRS significantly suppresses hair growth. **(A)** Flow chart of the animal experiment. The CRS treatment was applied from day1 and lasted for 20 days. The depilation of mice was on day 8 of the experiment, without applying CRS. On day 21 of the experiment, all animals were sacrificed for histological and metabolomics and transcriptomics analysis. **(B)** Photograph shown was taken on day 21 of experiment in (A), by which time mice treated with CRS exhibited hair growth inhibition versus control. **(C)** Skin melanin pigmentation scores (described in Materials and Methods) of murine dorsal skins treated with CRS versus control. Data are represented as mean ± SD. n = 5 mice in each group. *p* values are determined by Student's t-test. $^{**}p < 0.01$ compared with control group. **(D)** H&E staining showed the morphological changes in hair follicle. Magnification: 40×on the left; 400× on the right. **(E)** Immunohistochemistry for Ki-67 in back skin sections obtained from CRS group and control group. Magnification: 40× on the left; 400× on the right.

Joint Analysis of Metabolites and Genes

Metabolites and genes in the same pathways were always dysregulated together, so we utilized a pathway-based approach and integrated different levels of omics in the biological process. Enriched differential genes and metabolites were used in the joint pathway module for integrative analysis in MetaboAnalyst5.0. After uploaded our differential metabolites on MetaboAnalyst (https://www.metaboanalyst.ca/), the metabolites were then mapped to KEGG metabolic pathways for enrichment analysis.

Statistical Analyses

The statistical analyses were expressed as the mean ± SD and performed using GraphPad Prism software (v.8.0). Statistical significance between two groups was determined by Student's t-test. All experiments are repeated three times independently. Asterisk coding is indicated in Figure legends as **, $p < 0.01$.

RESULTS

CRS Significantly Suppresses Hair Growth

To confirm the inhibition of hair growth induced by CRS, we established the inhibition of hair growth affected by CRS model on C57BL/6 mice (**Figure 1A**). The dorsal skin color of the mice was pink in the telogen phase on the day of depilation and gradually became black, as the melanogenic activity of follicular melanocytes is related to the anagen stage of the hair cycle (Q. Wang et al., 2019). As shown in **Figure 1B**, on day 12 after depilation, no pigmentation or only a few scattered pigmented spots were visible on the dorsal skin of mice in the CRS group. In contrast, skin pigmentation was apparent in the control group, and some of the hair shafts were visible (**Figure 1B**). Statistical analyses also showed the skin pigmentation scores of murine dorsal skins in CRS group are significantly less than the control group on 10 days after depilation (day 18 of experiment) ($p <$ 0.01) (**Figure 1C**).

Next, we took advantage of H&E staining and immunohistochemistry to detect the formation and proliferation of hair follicles. Compared to controls, CRS dramatically decreased the number of hair follicles, the length of hair shafts, and the thickness of dermal layers (**Figure 1D**). The expression of proliferation marker Ki-67 (Magerl et al., 2001) was lower in hair follicles of the CRS group than that of control group (**Figure 1E**). Our results demonstrated that CRS significantly suppressed the hair growth of dorsal skin in mice.

CRS Significantly Regulates Metabolic Profile of the Skin Tissue

To systematically analyze the metabolic changes affected by CRS in hair growth, we performed the metabolomic analysis of dorsal skin between CRS group and control group. Compared to the control group, 158 features were significantly down-regulated and 138 features were significantly up-regulated in skin tissues of CRS group (**Supplementary Figure S1A**). PCA based on metabolite analysis showed that skin tissues of CRS-treated group were distinct from control group (**Supplementary Figure S1B**). PLS-DA was used to supervise the data analysis, and the permutation test was used to prevent PLS-DA model overfitting (**Supplementary Figure S1C**). In this study, the CRS group and control group were easily distinguished and the PLS-DA model was reliable (**Supplementary Figure S1D**). The aligned total ion chromatograms (TICs) and retention time width of all the groups in negative mode were shown in **Supplementary Figure S2A**, and those in positive modes were shown in **Supplementary Figure S2B**. Analysis of other checking parameters, including average m/z distribution, metabolite intensity distribution and coefficient of variation distribution, indicated effective sample preparation and high-quality raw data (**Supplementary Figures S3A,B**).

As shown in **Supplementary Figure S4**, we identified the primary metabolites with positive and negative ion modes by HMDB database. Among these features, the largest group was “lipids and lipid-like molecules”. The amino acids and the carbohydrates that we identified were belong to “Organic acids and derivatives” HMDB super class and “Organic oxygen compounds” HMDB super class, respectively. To facilitate the observation of metabolic changes, we normalized significantly differential metabolites and created a heatmap of all the secondary metabolites (**Supplementary Figure S5**).

CRS Significantly Affects the Profiles of Primary Metabolites in Skin Tissues

Metabolic pathways identified by MetaboAnalyst 5.0 for primary metabolites differentially identified by positive and negative polarity ionization in the skin tissue of CRS-treated mice compared to those in control mice are shown in **Figure 2A**. Among the relevant pathways identified, galactose metabolism (C00095, C00031, C00124, C00159, C00984, C00267, C00137, C00446, C00103, C00668, andC01097), fructose and mannose metabolism (C00095, C00267, C00159, C01094, C05345, C00275, and C00636), amino sugar and nucleotide sugar metabolism (C00984, C00267, C02336, C00159, C00085, C00446, C00103, C00668, C05345, C00275, and C00636), starch and sucrose metabolism (C00095, C00031, C00092, C00085, C00103), phenylalanine, tyrosine and tryptophan biosynthesis (C00082), glutamine and glutamate metabolism (C00217, C00025) were found to be the most important significant metabolic pathways (**Figures 2B,C**). These pathways were mainly involved in carbohydrate metabolism (fructose and mannose metabolism, galactose metabolism, amino sugar and nucleotide sugar metabolism, starch and sucrose metabolism) (**Figure 2B**) and amino acid metabolism (phenylalanine, tyrosine and tryptophan biosynthesis, glutamine and glutamate metabolism) (**Figure 2C**). It has been reported glutamine and glutamate metabolism play important roles in the epidermis and stem cells metabolism (C. S. Kim et al., 2020; Simsek et al., 2010; Takubo et al., 2013). Details

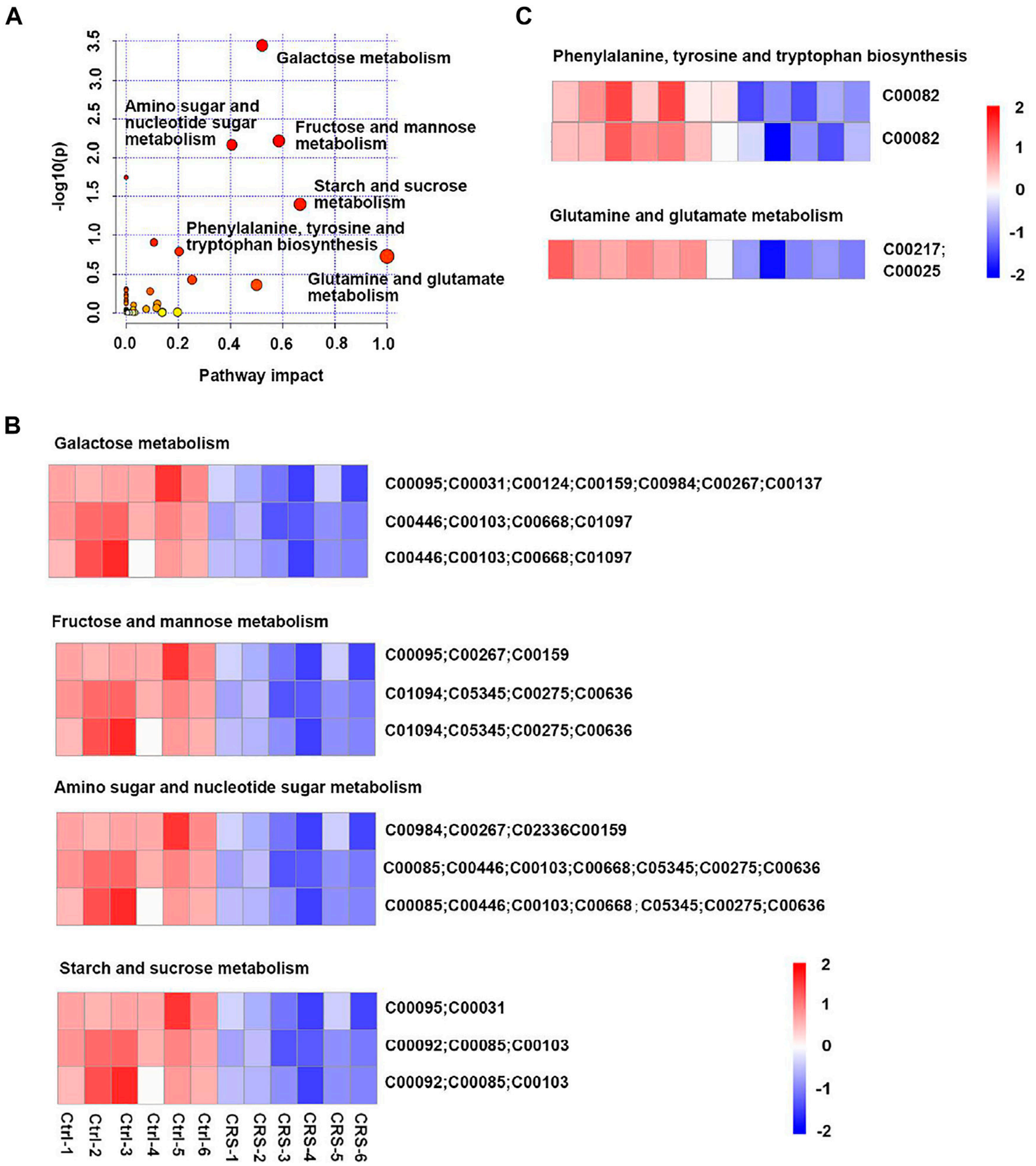


FIGURE 2 | CRS significantly affects the profiles of primary metabolites in skin tissues. **(A)** Scatterplot of enriched KEGG pathways in primary metabolites when comparing CRS group with control group using the MetaboAnalyst 5.0 pathway analysis module. Color shift indicates level of significance, size of dots correlates with the number of differential metabolites. The darker the color and the larger the dot, the stronger is the significance. Top enriched metabolic pathways were labeled. **(B)** Heatmap analysis showed differential metabolites in galactose metabolism (C00095, C00031, C00124, C00159, C00984, C00267, C00137, C00446, C00103, C00668, and C01097), fructose and mannose metabolism (C00095, C00267, C00159, C01094, C05345, C00275, and C00636), amino sugar and nucleotide sugar metabolism (C00984, C00267, C02336, C00159, C00085, C00446, C00103, C00668, C05345, C00275, and C00636), starch and sucrose metabolism (C00095, C00031, C00092, C00085, and C00103) between CRS group and control group. The ordinate was the KEGG ID matched to database. **(C)** Heatmap analysis showed differential metabolites in phenylalanine, tyrosine and tryptophan biosynthesis (C00082), glutamine and glutamate metabolism (C00217, C00025) pathways between CRS group and control group. The ordinate was the KEGG ID matched to database. The color blocks represented the relative expression of metabolites, red represented up-regulation, and blue represented down-regulation.

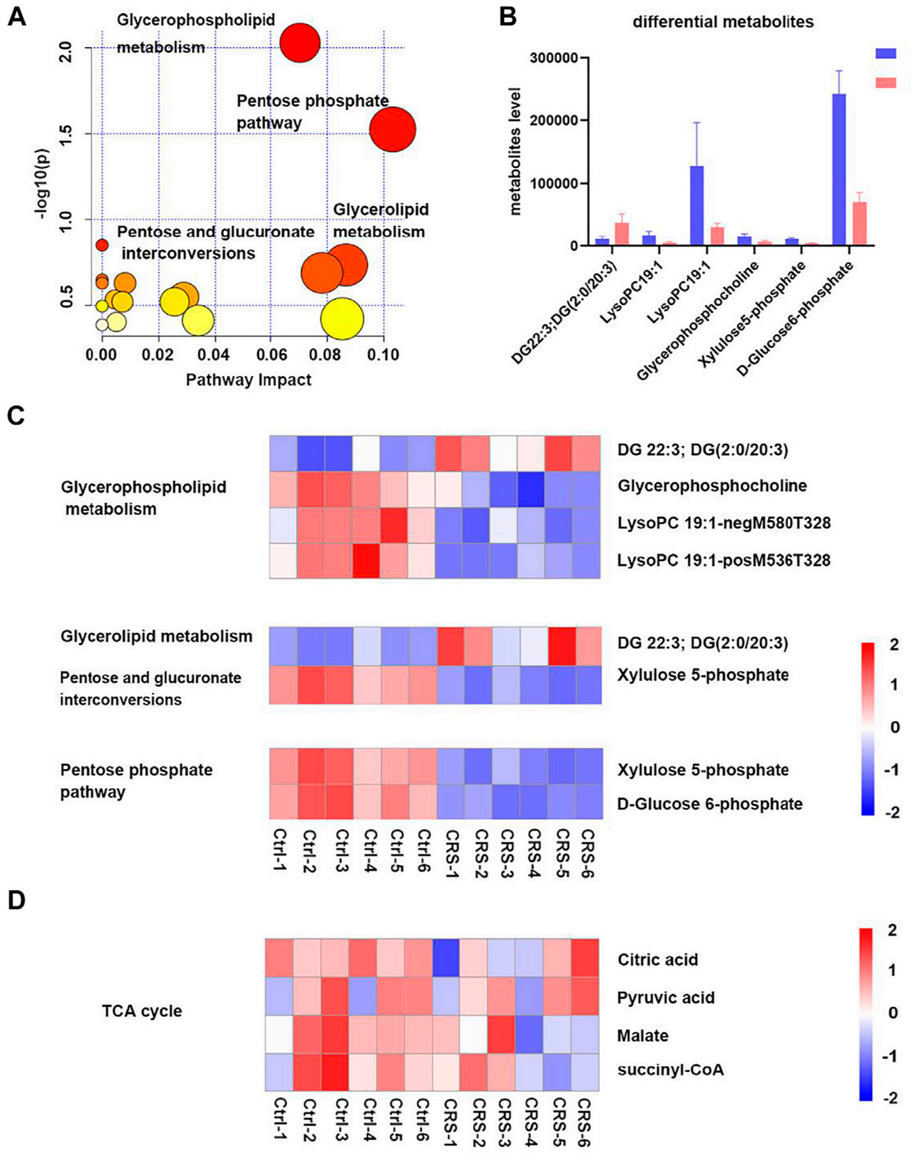


FIGURE 3 | CRS significantly affects the profiles of secondary metabolites in skin tissues. **(A)** Scatterplot of enriched KEGG pathways in secondary metabolites when comparing CRS group with control group using the MetaboAnalyst 5.0 pathway analysis module. Top enriched metabolic pathways were labeled. **(B)** Comparison of secondary metabolites level between control group and CRS group in the top 4 significant pathways. **(C)** Heatmap analysis showed differential metabolites in glycerophospholipid metabolism, pentose phosphate pathway, glycerolipid metabolism, pentose and glucuronate interconversions pathways. **(D)** Heatmap analysis showed metabolites changes in TCA cycle.

of differential metabolites in skin tissues are shown in **Supplementary Table S2**.

CRS Significantly Affects the Profiles of Secondary Metabolites in Skin Tissues

The significant differential secondary metabolites were subjected for KEGG pathway analysis. As shown in **Figure 3A**, the most significantly changed pathway was the glycerophospholipid metabolism. Pentose phosphate pathway, glycerolipid metabolism, pentose and glucuronate interconversions were also significantly altered after CRS treatment. A total of six significant differential secondary metabolites were identified, including DG 22:3; DG (2:0/20:3) which was significantly upregulated, LysoPC 19:1-neg-M580T328, LysoPC 19:1-pos-M536T328, glycerophosphocholine, xylulose 5-phosphate and D-Glucose 6-phosphate which were significantly downregulated (**Figures 3B,C**). Among them, D-Glucose 6-phosphate was the most drastically reduced metabolite. In

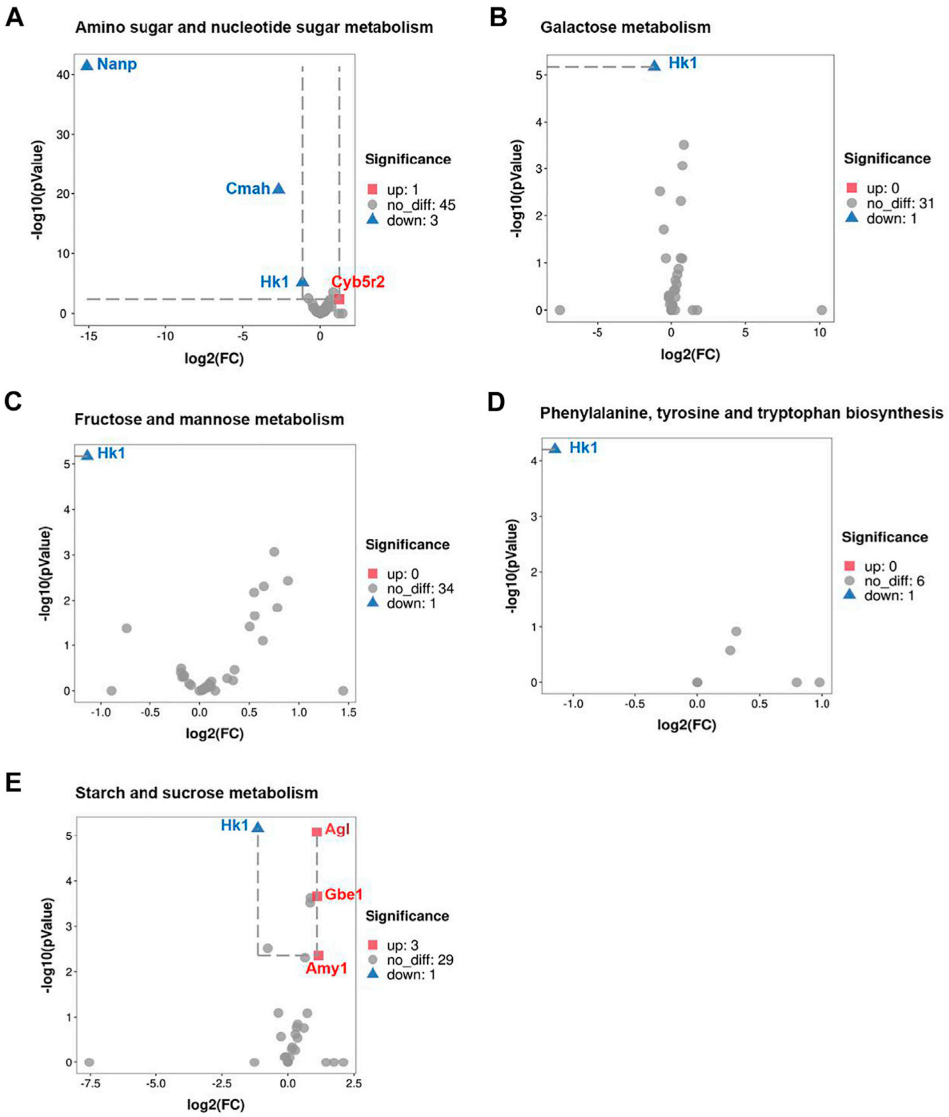


FIGURE 4 | CRS significantly affects genes expression associated with primary metabolites. **(A)** Volcano plot showed regulation of genes expression in amino sugar and nucleotide sugar metabolism. Significantly DEGs were labeled. FC is for gene expression fold change in CRS group compared to control group. The DEGs were selected with fold change >2 or fold change <0.5 and *p* value <0.05. **(B)** Volcano plot showed regulation of genes expression in galactose metabolism. **(C)** Volcano plot showed regulation of genes expression in fructose and mannose metabolism. **(D)** Volcano plot showed regulation of genes expression in phenylalanine, tyrosine and tryptophan biosynthesis. **(E)** Volcano plot showed regulation of genes expression in starch and sucrose metabolism.

addition, D-Fructose 6-phosphate was also significantly downregulated in CRS group in primary metabolites. Both D-Glucose-6-phosphate and D-Fructose 6-phosphate are involved in glycolytic metabolism pathways, which indicated that glycolytic metabolism might play a critical role in the inhibition of hair growth induced by CRS. Conversely, metabolites in TCA cycle were not significantly changed between the CRS group and control group (**Figure 3D**). Collectively these results suggested that although skin tissue use the TCA cycle to generate energy, CRS could not regulate TCA metabolism to inhibit hair growth.

Furthermore, KEGG pathway enrichment analyses were conducted to further analyze the metabolic profiles in skin tissues of CRS-induced hair growth inhibition. The top 20 KEGG pathways were shown in **Supplementary Figure S6**. These results demonstrated that biosynthesis of amino acids, central carbon metabolism in cancer, protein digestion and

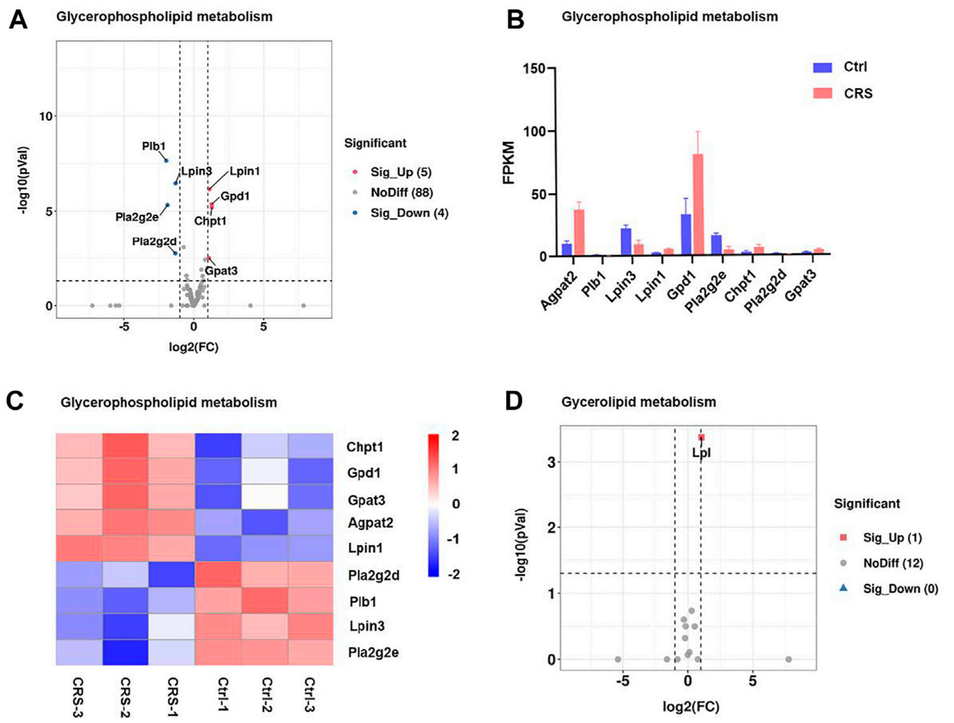


FIGURE 5 | CRS significantly affects genes expression associated with secondary metabolites. **(A)** Volcano plot showed regulation of genes expression in glycerophospholipid metabolism. Significantly DEGs were labeled. FC is for gene expression fold change in CRS group compared to control group. The DEGs were selected with fold change >2 or fold change <0.5 and *p* value <0.05. **(B)** The fragments per kilobase of transcript per million mapped reads (FPKM) values of DEGs in glycerophospholipid metabolism. **(C)** Heatmap analysis showed regulation of genes expression in glycerophospholipid metabolism. The color blocks represented the relative genes expression, red represented up-regulation, and blue represented down-regulation. Volcano plot showed regulation of gene expression in glycerolipid metabolism.

absorption, ABC transporters, glycerophospholipid metabolism, carbon metabolism pathways were the most significantly altered pathways.

CRS Significantly Affects Genes Expression Associated With Primary Metabolites

Among these pathways above, we identified several differentially expressed genes (DEGs) based on KEGG pathway analysis in transcriptomic. Volcano plot analysis indicated that a total of 5 pathways were matched to significantly DEGs for primary metabolites, including galactose metabolism, fructose and mannose metabolism, amino sugar and nucleotide sugar metabolism, phenylalanine, tyrosine and tryptophan biosynthesis, starch and sucrose metabolism. As shown in **Figure 4**, Hk-1 was significantly downregulated in all 5 metabolic pathways, suggested that Hk-1 played a very important role in relative biological process of CRS inhibiting hair growth. Furthermore, some other genes expression was also significantly changed. In amino sugar and nucleotide sugar metabolism, *Nanp,* and *Cmah* expression were significantly down-regulated and *Cyb5r2* was significantly up-regulated (**Figure 4A**). The *Nanp* gene is related to synthesis of substrates in N-glycan biosynthesis and metabolism of proteins pathways, *Cmah* encodes cytidine monophosphate-N-acetylneuraminic acid hydroxylase, an enzyme responsible for Neu5Gc biosynthesis (Burzyńska et al., 2021), and *Cyb5r2* encodes Cytochrome B5 Reductase 2 which participates in many processes including cholesterol biosynthesis, fatty acid desaturation and elongation. *Agl, Gbe1, and Amy1* were significantly up-regulated which related to starch and sucrose metabolism (**Figure 4E**). *Agl* encodes the glycogen debrancher enzyme that is involved in glycogen degradation, *Gbe1* encodes the glycogen branching which is important to increase the solubility of the glycogen molecule and, consequently, reducing the osmotic pressure within cells (Malinska et al., 2020). *Amy1* encodes amylase alpha produced by the salivary gland. Amylases catalyze the first step in digestion of dietary starch and glycogen. Previous researches have repeatedly demonstrated the activation of salivary alpha-amylase induced by psychosocial stress (Skosnik et al., 2000; Rohleder et al., 2004), which may explain the up-regulation of *Amy1* in starch and sucrose

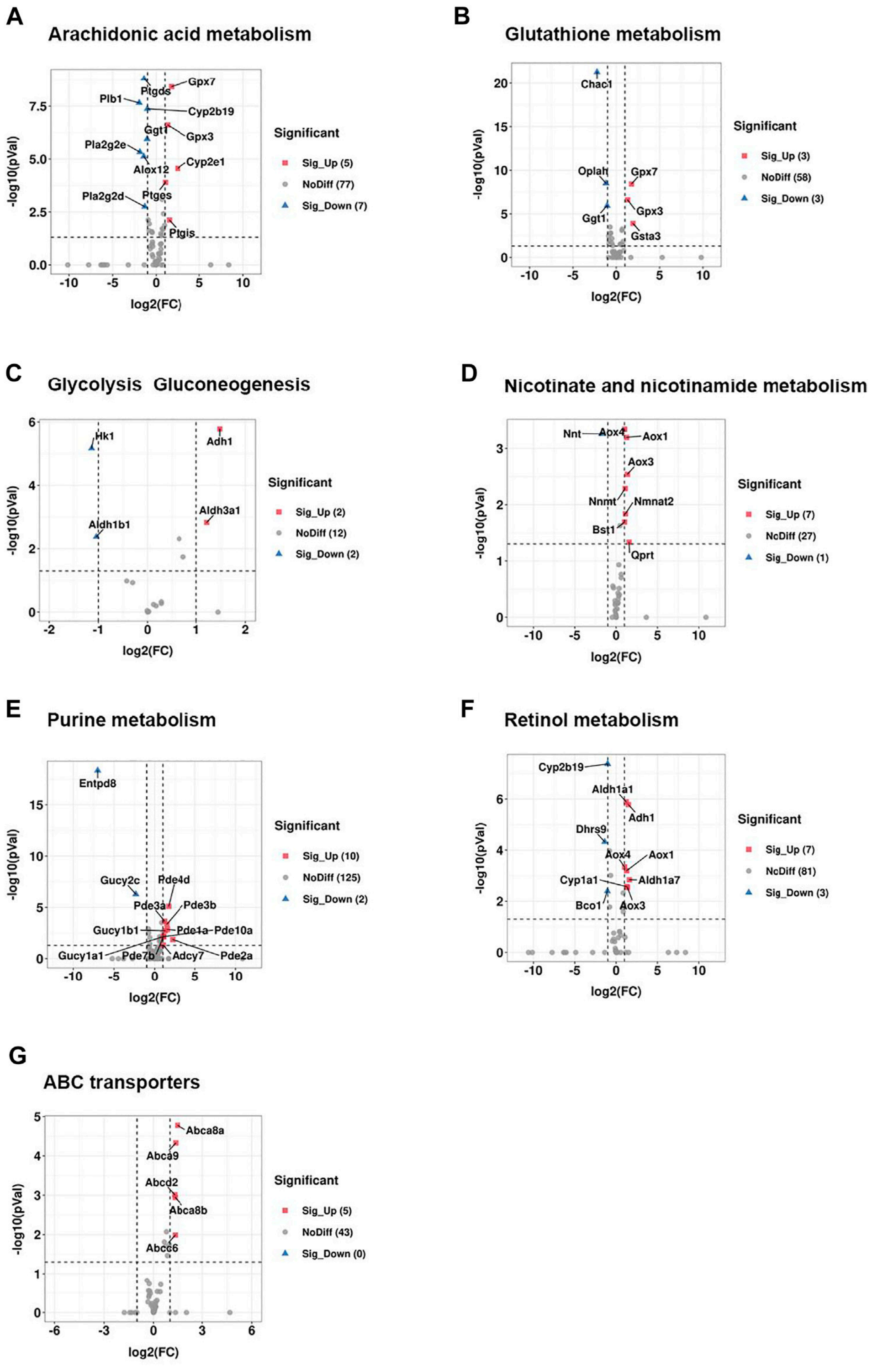


FIGURE 6 | CRS significantly alters genes expression related to metabolism pathways based on RNA-seq. **(A)** Volcano plot showed regulation of genes expression in arachidonic acid metabolism between CRS group and control group. Significantly DEGs were labeled. FC is for gene expression fold change in CRS group compared to control group. The DEGs were selected with fold change >2 or fold change <0.5 and *p* value <0.05. **(B)** Volcano plot showed regulation of genes expression in glutathione metabolism between CRS group and control group. **(C)** Volcano plot showed regulation of genes expression in glycolysis gluconeogenesis between CRS group and control group. **(D)** Volcano plot showed regulation of genes expression in nicotinate and nicotinamide metabolism between CRS group and control group. **(E)** Volcano plot showed regulation of genes expression in purine metabolism between CRS group and control group. **(F)** Volcano plot showed regulation of genes expression in retinol metabolism between CRS group and control group. **(G)** Volcano plot showed regulation of genes expression in ABC transporters between CRS group and control group.

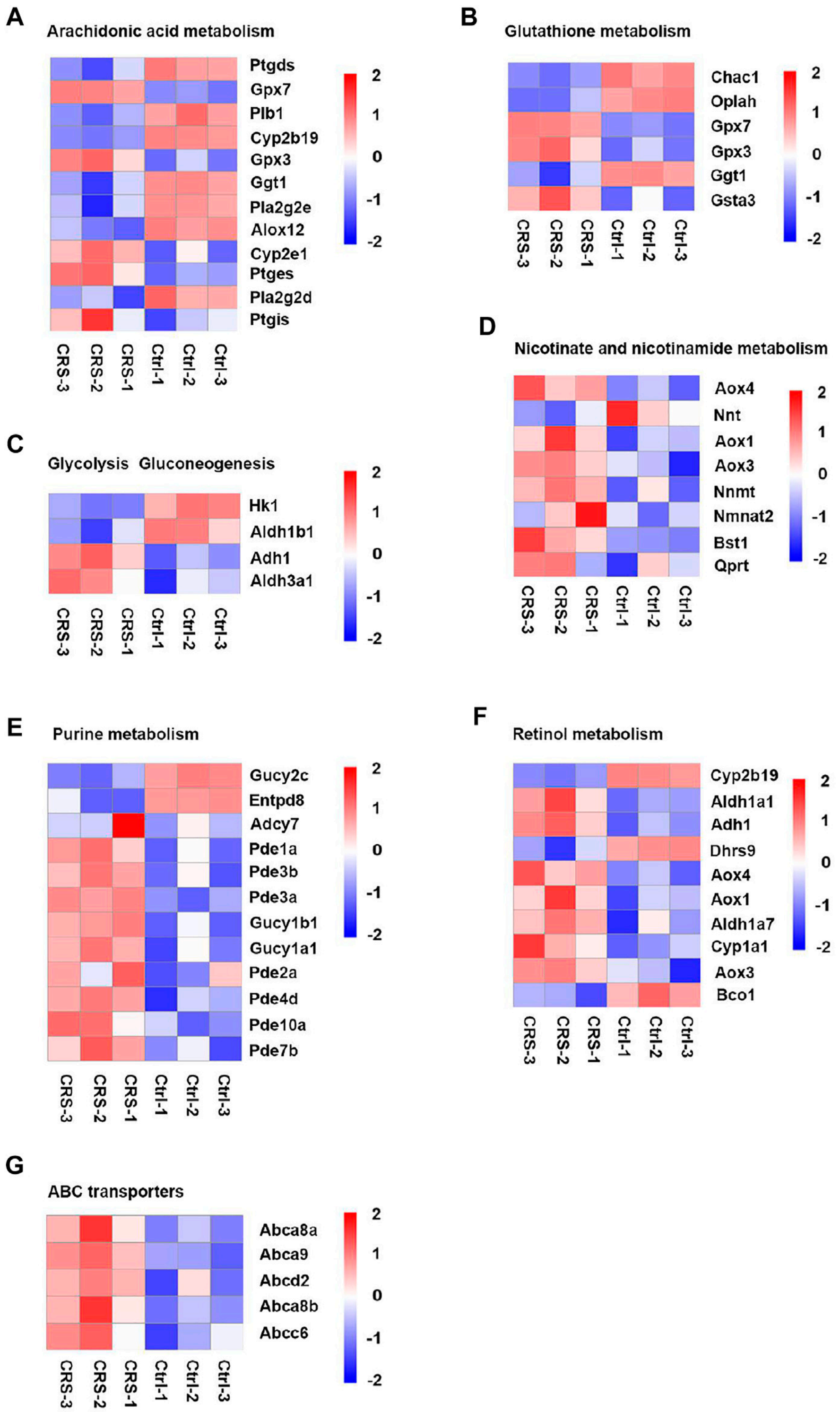


FIGURE 7 | CRS significantly alters genes expression related to metabolism pathways based on RNA-seq. D. Heatmap analysis showed regulation of genes expression in arachidonic acid metabolism between CRS group and control group. **(A)** Heatmap analysis showed regulation of genes expression in glutathione metabolism between CRS group and control group. **(B)** Heatmap analysis showed regulation of genes expression in glycolysis gluconeogenesis between CRS group and control group. **(C)** Heatmap analysis showed regulation of genes expression in nicotinate and nicotinamide metabolism between CRS group and control group. **(D)** Heatmap analysis showed regulation of genes expression in purine metabolism between CRS group and control group. **(E)** Heatmap analysis showed regulation of genes expression in retinol metabolism between CRS group and control group. **(F)** Heatmap analysis showed regulation of genes expression in ABC transporters between CRS group and control group.

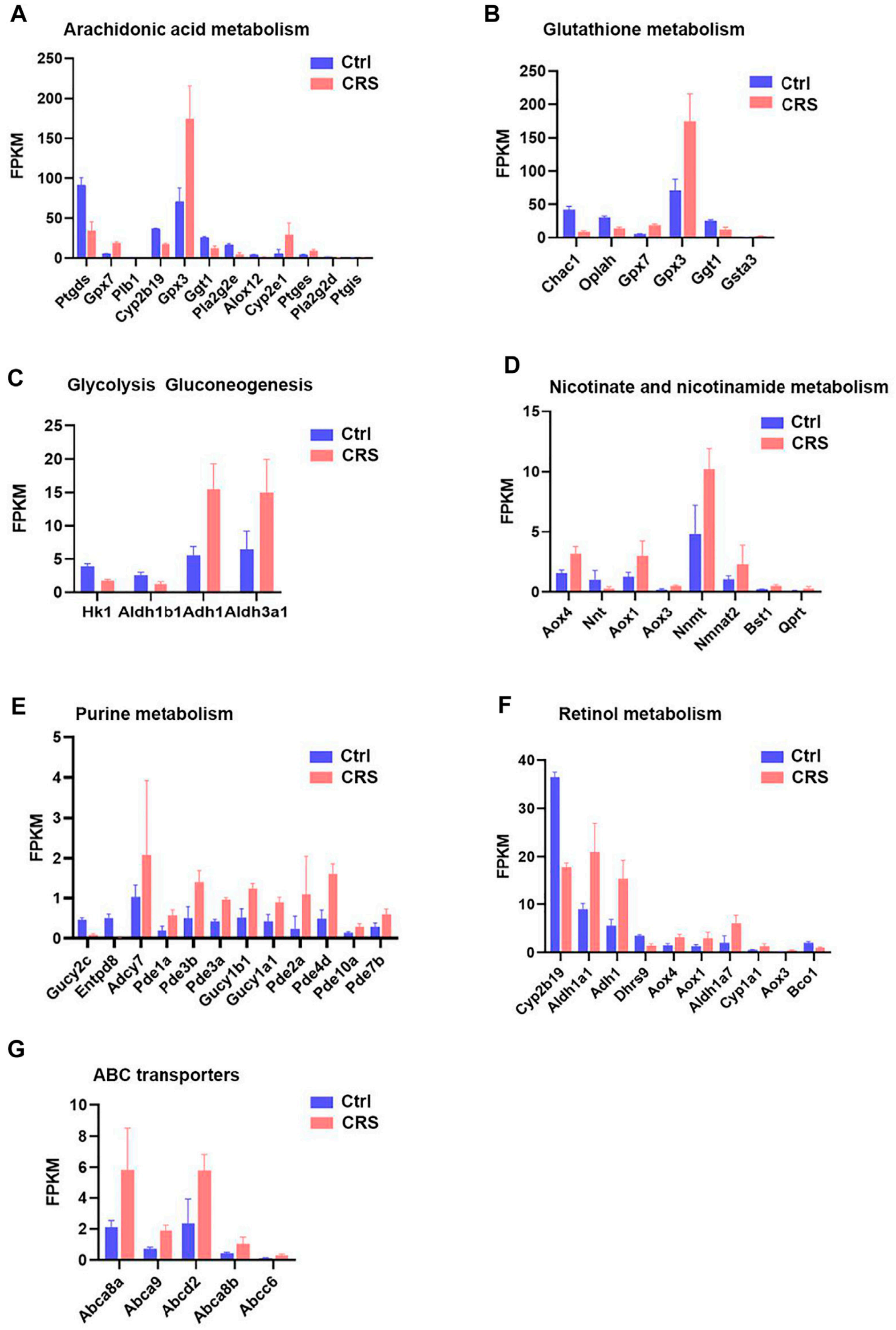


FIGURE 8 | CRS significantly alters genes expression related to metabolism pathways based on RNA-seq. **(A)** The FPKM values of DEGs associated with arachidonic acid metabolism between CRS group and control group. **(B)** The FPKM values of DEGs associated with glutathione metabolism between CRS group and control group. **(C)** The FPKM values of DEGs associated with glycolysis gluconeogenesis between CRS group and control group. **(D)** The FPKM values of DEGs associated with nicotinate and nicotinamide metabolism between CRS group and control group. **(E)** The FPKM values of DEGs associated with purine metabolism between CRS group and control group. **(F)** The FPKM values of DEGs associated with retinol metabolism between CRS group and control group. **(G)** The FPKM values of DEGs associated with ABC transporters between CRS group and control group.

metabolism after treated with CRS. The FPKM values of differentially expressed genes associated with primary metabolites were shown in **Supplementary Table S3**.

CRS Significantly Affects Genes Expression Associated With Secondary Metabolites

Consistent with the metabolomics analysis for secondary metabolites, the expression of genes associated with glycerophospholipid metabolism and glycerolipid metabolism pathways were also significantly changed. Among a total of 97 genes involved in the pathway of glycerophospholipid metabolism, 9 genes (*Agpat2, plb1, Gpd1, Pla2g2e, Pla2g2d, Gapt3, Lpin3, Lpin1, and Chpt1*) were significantly changed (**Figure 5A**). Among these genes, *Agpat2* and *Gpd1* were markedly up-regulated (**Figures 5B,C**). *Gpd1* plays a critical role in carbohydrate and lipid metabolism. *Agpat2* converts lysophosphatidic acid to phosphatidic acid, the second step in *de novo* phospholipid biosynthesis. In addition, *Lpl* which encodes lipoprotein lipase was involved in the glycerolipid metabolism, and the gene expression was also drastically increased (**Figure 5D**). The FPKM values of differentially expressed genes associated with secondary metabolites were shown in **Supplementary Table S4**. The remarkable regulation of these genes suggested the impact of CRS on hair growth is strongly linked to lipid metabolism.

CRS Significantly Alters Genes Expression Related to Metabolism Pathways Based on RNA-Seq

To further analyze broad metabolic pathways of CRS participation in hair growth inhibition, KEGG pathway enrichment analyses were conducted and revealed that CRS also affected gene expression in numerous other metabolic pathways, including arachidonic acid metabolism, glutathione metabolism, glycolysis gluconeogenesis, nicotinate and nicotinamide metabolism, purine metabolism, retinol metabolism and ABC transporters. A total of 97 genes associated with metabolism were significantly differentially expressed in the skin of CRS group compared to that of control group (**Supplementary Figure S7**).

Among a total of 89 genes involved in the pathway of arachidonic acid (AA) metabolism, 12 genes were significantly changed between CRS group and control group, including 5 genes that were upregulated and 7 genes that were downregulated. The upregulated genes included *Gpx7, Gpx3, Cyp2e1, Ptges, Ptgis* while *Ptgds, Plb1, Cyp2b19, Ggt1, Pla2g2e, Pla2g2d, Alox12* were dramatically downregulated (**Figures 6A, 7A, 8A**). Arachidonic acid is one of the major polyunsaturated fatty acids in mammals (Yu and Wang, 2021). In consideration of the significant decrease of AA-residue-enriched LPCs (e.g., LysoPC 19:1-neg-M580T328, LysoPC 19:1- pos-M536T328etc.), we speculated that arachidonic acid (AA) metabolism was critical for CRS-induced hair growth inhibition.

CRS also affected the expression of genes clustered into glutathione metabolism, including 3 downregulated genes (including *Chac1, Oplah, Ggt1*) and 3 upregulated genes (including *Gpx7, Gpx3, Gsta3*) (**Figures 6B**, **7B**, **8B**). For glycolysis gluconeogenesis, CRS significantly suppressed the expression of 2 genes (*Hk1, Aldh1b1*) and increased the expression of 2 genes (*Adh1, Aldh3a1*) (**Figures 6C**, **7C**, **8C**). For nicotinate and nicotinamide metabolism, 7 genes exhibited increased expression in CRS group (*Aox4, Aox1, Aox3, Nnmt, Nmnat2, Bst1, Qprt*) and 1 gene exhibited decreased expression (*Nnt*) (**Figures 6D**, **7D**,**8D**). In addition, there were 12 DEGs between CRS group and control group that clustered to the purine metabolism pathway of which expression was decreased for 2 genes, including *Gucy2c* and *Entpd8*, and increased for 10 genes, including *Adcy7, Pde1a, Pde3b, Pde3a, Gucy1b1, Gucy1a1, Pde2a, Pde4d, Pde10a, Pde7b* (**Figures 6E**, **7E**, **8E**). A total of 10 DEGs clustered to the pathway of retinol metabolism after CRS treatment, including 3 genes with decreased expression (*Cyp2b19, Dhrs9* and *Bco1*), and 7 genes with increased expression, including *Aldh1a1, Adh1*, *Aox4, Aox1, Aldh1a7, Cyp1a1, Aox3* (**Figures 6F**, **7F**, **8F**). CRS also significantly upregulated the expression of genes important for ABC transporters including *Abca8a, Abca9, Abcd2, Abca8b, Abcc6* (**Figures 6G**, **7G**, **8G**). The FPKM values of differentially expressed genes associated with arachidonic acid metabolism, glutathione metabolism, glycolysis gluconeogenesis, nicotinate and nicotinamide metabolism, purine metabolism, retinol metabolism, ABC transporters were shown in **Supplementary Tables S5–11**. These results showed that the gene expression profiles of multiple metabolic pathways were significantly different between CRS group and control group.

DISCUSSION

Previous studies have reported that CRS influences hair growth via various hormones, neuropeptides, and neurotransmitters, but little is known about its regulation from the perspective of metabolic signals (Paus et al., 2014). Substance P(SP), Calcitonin gene-related peptide (CGRP) and nerve growth factors (NGF) have been regarded as the critical mediators in stress-induced hair loss (Arck et al., 2005; Samuelov et al., 2012). It has also been demonstrated that the existence of a "brain-hair follicle axis" (BHA), and some neuropeptides such as CGRP, SP and NGF could induce apoptosis of murine follicular keratinocytes and stimulate mast cell degranulation, thus inhibiting hair growth (Arck et al., 2001; Arck, Handjiski, Peters, Hagen, et al., 2003; Arck, Handjiski, Peters, Peter, et al., 2003).

However, psychological stress could also affect metabolic levels, but the underlying molecular mediators are poorly defined (Noerman et al., 2020). Our results suggested that CRS significantly suppressed hair growth and showed significant changes in metabolites between CRS and control group. In this study, we found that some metabolites related to lipid metabolism were significantly changed. Notably, DG 22:3; DG (2:0/20:3) was increased in both glycerophospholipid metabolism and glycerolipid metabolism pathways, while lysophosphatidylcholine

(LysoPC 19:1) was decreased robustly. Diacylglycerol (DG), as one of the primary lipid sub-groups in living systems and a second messenger in multiple cell activities, serve as a critical role in hastening the β-oxidation of fatty acids, as well as influence the expression of lipid metabolism-linked genes (Almena and Mérida, 2011; Eichmann and Lass, 2015). Importantly, it has been shown that chronic stress alters the levels of DG in stress-susceptible brain regions (Patel et al., 2009; Oliveira et al., 2016). The high level of DG, such as DG 22:3; DG (2:0/20:3), in the dorsal skin of CRS group might be linked to signal transduction and structural components of epidermis under chronic stress (Lee, 2011). It is also reported that lysophosphatidylcholine (LPC) levels in the prefrontal cortex in brain are directly correlated with blood corticosterone levels (Oliveira et al., 2016). Another research shows that excessive expression of LPCs is correlated with high oxidative stress (Hung et al., 2020). Typically, stress is characterized by activation of the sympathetic nervous system and hypothalamic–pituitary–adrenal axis, resulting in release of glucocorticoids (Marin et al., 2007). Chronic stress increases the levels of corticosterone to extend HFSC quiescence and inhibit hair growth in mice, which indicates that LPC may be related to hair growth under stress (Choi et al., 2021).

Lipid metabolism is thought to play an essential role in maintaining normal physiological cellular functions and involving in hair development and function (W. S. Lee, 2011; Palmer et al., 2020). Thus, our current investigation took advantage of metabolomics and transcriptomics analysis to further verify the metabolomics results and showed mRNA levels of the relevant glycerophospholipid metabolism were significantly increased, such as *Agpat2, Gpd1, Gapt3, Chpt1, and Lpin1;* while others such as *plb1, Pla2g2e, Pla2g2d, and Lpin3* were decreased. The expression of *Lpl* involved in the glycerolipid metabolism was also significantly increased. Changes in these genes might suggest the underlying connection between CRS inhibit hair growth and lipid metabolism. We also found that genes related to ATP-binding cassette (ABC) transporter, such as *Abca8a, Abca9, Abcd2, Abca8b, and Abcc6* are all significantly up-regulated. ABC transporters mediate the transport of lipids. In particular, the ABCA family is involved in both cholesterol efflux and intracellular transport (Quazi and Molday, 2011; Tarling et al., 2013). It has been revealed that cholesterol modulates HF cycling by regulating bone morphogenic protein (BMP) family members, Wnt/β-catenin and Notch pathways (Cooper et al., 2003; Lee and Tumbar, 2012; Mathay et al., 2011; Sheng et al., 2014), which indicates the importance of cholesterol homeostasis in stress inhibit hair growth.

Other metabolic pathways involved in carbohydrate metabolism (fructose and mannose metabolism, galactose metabolism, amino sugar and nucleotide sugar metabolism, starch and sucrose metabolism, pentose phosphate pathway, pentose and glucuronate interconversions) are also changed significantly after CRS treatment during the hair growth. Compared to control group, expression of Hk-1 decreased markedly in skin tissues of CRS group in all 5 primary metabolic pathways. Hk-1 encodes a ubiquitous form of hexokinase which localizes on the outer membrane of mitochondria. Hexokinases catalyzes the conversion of glucose to glucose-6-phosphate in the first step of glycolytic metabolism. Then glucose-6-phosphate convert to fructose-6-phosphate catalyzed by glucose-6-phosphate isomerase (GPI) in glycolysis. As mentioned above, D-glucose 6-phosphate was the most significantly down-regulated metabolite in secondary metabolites. This is usually attributed to Hk-1 activity decrease or to G6P dehydrogenase (G6PD) activity increase (Rodriguez-Rodriguez et al., 2013). G6PD catalyzes the oxidation of glucose-6-phosphate to 6-phosphogluconate (Kotaka et al., 2005). This transformation is a rate-limiting step of the pentose-phosphate pathway, which represents a route for the dissimilation of carbohydrates besides glycolysis (Kirsch et al., 2009). However, the two genes we identified relating to G6P Dehydrogenase, *G6pdx and G6pd2* exhibited no significant difference between CRS group and control group. Thus, we presume that Hk-1 may contribute to the decrease of D-glucose 6-phosphate, which was another piece of evidence proving the disturbances in the glycolytic metabolism. However, the function of Hk-1 in CRS induced hair growth inhibition is still unknown. Therefore, we will assess the hexokinase activity and construct the conditional knockout Hk-1 mice in hair follicle stem cell to research the mechanisms and function of Hk-1 *in vivo* and *vitro*.

However, most metabolites in TCA cycle were not significantly changed between the CRS group and control group, which suggested that glycolytic metabolism rather than TCA cycle might play an important role in hair growth under CRS. It has been reported that HFSC likely have relatively higher levels of glycolysis compared to the rest of the epidermis, which indicates the changes of glycolysis metabolite may be linked to metabolic status of HFSC(Flores et al., 2017). HFSCs quickly in response to the barrage of cues which is dependent on increasing glycolytic rate that orchestrates the onset of a new hair cycle, whereas chronic stress may prolong HFSC quiescence and maintain hair follicles in an extended resting phase (Choi et al., 2021). Glutamine metabolism also regulates hair follicle stem cell progenitor state (C. S. Kim et al., 2020). In our study, we found the metabolic features such as glutamine and glutamate were downregulated in skin tissues of CRS group. The rapidly proliferating stem cells required ATP as well as nucleotides, aerobic glycolysis, which could also explain the alter of genes involved in purine metabolism (Ahmed et al., 2018). Although, there are numerous researchers take advantage of mouse model to explore the mechanisms of hair loss and the regulation of hair follicle cycling. For example, it has been reported that JAK inhibition regulates the activation of key hair follicle populations to promote the hair growth in both mouse and human by topical treatment (Harel et al., 2015). Furthermore, there has been reported that the retinoid metabolism is altered in human and mouse cicatricial alopecia (Everts et al., 2013). However, the function of these pathways still needs to be further investigated and the differences between the model of murine hair growth and human scalp hair growth should be concerned. Moreover, we will further explore the metabolism changes in hair-loss patients caused by stress and compare with the mouse model to find the similar metabolic changes.

CONCLUSION

In this study, we discovered that CRS suppressed hair growth, found the metabolism pathways including carbohydrate metabolism, amino acid metabolism, lipid metabolism were significantly changed, and revealed the metabolism associated DEGs such as Hk-1 by transcriptomics and metabolomics analyses in skin tissues of C57BL/6 mice. Our results provided new insights into the molecular mechanisms of CRS-induced hair growth inhibition and indicated that targeting to specific metabolic pathways might be useful for therapy of CRS inhibit hair growth.

DATA AVAILABILITY STATEMENT

Based on our identifications, metabolite information was submitted to MetaboLights public repository (www.ebi.ac.uk/metabolights/MTBLS4085). The transcriptomics data presented in the study are deposited in the Sequence Read Archive repository, accession number (PRJNA763049). The datasets can be found in online repositories (https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA763049).

ETHICS STATEMENT

The animal study was reviewed and approved by the Ethics Committee of Sir Run Run Shaw Hospital of Zhejiang University School of Medicine.

AUTHOR CONTRIBUTIONS

CC, HC, YS, and QZ conceived and designed the study. XW, XJ and QL performed the experiments. XW, MX, and LL analyzed the data. XW, YS wrote the manuscript, HC, YS, and QZ reviewed the manuscript. QZ and HC have overall responsibility for this manuscript. HC and YS were responsible for supervision and funding acquisition.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (82173439, 81573057, and 81703135), and the Natural Science Foundation of Zhejiang Province (LY19H190006).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.781619/full#supplementary-material

REFERENCES

Ahmed, N., Escalona, R., Leung, D., Chan, E., and Kannourakis, G. (2018). Tumour Microenvironment and Metabolic Plasticity in Cancer and Cancer Stem Cells: Perspectives on Metabolic and Immune Regulatory Signatures in Chemoresistant Ovarian Cancer Stem Cells. *Semin. Cancer Biol.* 53, 265–281. doi:10.1016/j.semcancer.2018.10.002

Alexopoulos, A., and Chrousos, G. P. (2016). Stress-related Skin Disorders. *Rev. Endocr. Metab. Disord.* 17 (3), 295–304. doi:10.1007/s11154-016-9367-y

Almena, M., and Mérida, I. (2011). Shaping up the Membrane: Diacylglycerol Coordinates Spatial Orientation of Signaling. *Trends Biochem. Sci.* 36 (11), 593–603. doi:10.1016/j.tibs.2011.06.005

Antoni, M. H., and Dhabhar, F. S. (2019). The Impact of Psychosocial Stress and Stress Management on Immune Responses in Patients with Cancer. *Cancer* 125 (9), 1417–1431. doi:10.1002/cncr.31943

Arck, P. C., Handjiski, B., Hagen, E., Joachim, R., Klapp, B. F., and Paus, R. (2001). Indications for a Brain-hair Follicle axis: Inhibition of Keratinocyte Proliferation and Up-regulation of Keratinocyte Apoptosis in Telogen Hair Follicles by Stress and Substance P. *FASEB j.* 15 (13), 2536–2538. doi:10.1096/fj.00-0699fje

Arck, P. C., Handjiski, B., Kuhlmei, A., Peters, E. M. J., Knackstedt, M., Peter, A., et al. (2005). Mast Cell Deficient and Neurokinin-1 Receptor Knockout Mice Are Protected from Stress-Induced Hair Growth Inhibition. *J. Mol. Med.* 83 (5), 386–396. doi:10.1007/s00109-004-0627-z

Arck, P. C., Handjiski, B., Peters, E. M. J., Hagen, E., Klapp, B. F., and Paus, R. (2003). Topical Minoxidil Counteracts Stress-Induced Hair Growth Inhibition in Mice. *Exp. Dermatol.* 12 (5), 580–590. doi:10.1034/j.1600-0625.2003.00028.x

Arck, P. C., Handjiski, B., Peters, E. M. J., Peter, A. S., Hagen, E., Fischer, A., et al. (2003). Stress Inhibits Hair Growth in Mice by Induction of Premature Catagen Development and Deleterious Perifollicular Inflammatory Events via Neuropeptide Substance P-dependent Pathways. *Am. J. Pathol.* 162 (3), 803–814. doi:10.1016/s0002-9440(10)63877-1

Burzyńska, P., Sobala, Ł., Mikołajczyk, K., Jodłowska, M., and Jaśkiewicz, E. (2021). Sialic Acids as Receptors for Pathogens. *Biomolecules* 11 (6), 831. doi:10.3390/biom11060831

Cervantes, J., Jimenez, J. J., DelCanto, G. M., and Tosti, A. (2018). Treatment of Alopecia Areata with Simvastatin/Ezetimibe. *J. Invest. Dermatol. Symp. Proc.* 19 (1), S25–s31. doi:10.1016/j.jisp.2017.10.013

Chai, M., Jiang, M., Vergnes, L., Fu, X., de Barros, S. C., Doan, N. B., et al. (2019). Stimulation of Hair Growth by Small Molecules that Activate Autophagy. *Cel Rep.* 27 (12), 3413–3421. e3413. doi:10.1016/j.celrep.2019.05.070

Choi, S., Zhang, B., Ma, S., Gonzalez-Celeiro, M., Stein, D., Jin, X., et al. (2021). Corticosterone Inhibits GAS6 to Govern Hair Follicle Stem-Cell Quiescence. *Nature* 592, 428–432. doi:10.1038/s41586-021-03417-2

Christodoulou, C. C., Zachariou, M., Tomazou, M., Karatzas, E., Demetriou, C. A., Zamba-Papanicolaou, E., et al. (2020). Investigating the Transition of Pre-symptomatic to Symptomatic Huntington's Disease Status Based on Omics Data. *Int. J. Mol. Sci.* 21 (19), 7414. doi:10.3390/ijms21197414

Cooper, M. K., Wassif, C. A., Krakowiak, P. A., Taipale, J., Gong, R., Kelley, R. I., et al. (2003). A Defective Response to Hedgehog Signaling in Disorders of Cholesterol Biosynthesis. *Nat. Genet.* 33 (4), 508–513. doi:10.1038/ng1134

Cotsarelis, G., and Millar, S. E. (2001). Towards a Molecular Understanding of Hair Loss and its Treatment. *Trends Mol. Med.* 7 (7), 293–301. doi:10.1016/s1471-4914(01)02027-5

Dainichi, T., and Kabashima, K. (2017). Alopecia Areata: What's New in Epidemiology, Pathogenesis, Diagnosis, and Therapeutic Options? *J. Dermatol. Sci.* 86 (1), 3–12. doi:10.1016/j.jdermsci.2016.10.004

Eichmann, T. O., and Lass, A. (2015). DAG Tales: the Multiple Faces of Diacylglycerol-Stereochemistry, Metabolism, and Signaling. *Cell. Mol. Life Sci.* 72 (20), 3931–3952. doi:10.1007/s00018-015-1982-3

Everts, H. B., Silva, K. A., Montgomery, S., Suo, L., Menser, M., Valet, A. S., et al. (2013). Retinoid Metabolism Is Altered in Human and Mouse Cicatricial Alopecia. *J. Invest. Dermatol.* 133 (2), 325–333. doi:10.1038/jid.2012.393

Feng, S., Wu, J., Qiu, W.-L., Yang, L., Deng, X., Zhou, Y., et al. (2020). Large-scale Generation of Functional and Transplantable Hepatocytes and Cholangiocytes from Human Endoderm Stem Cells. *Cel Rep.* 33 (10), 108455. doi:10.1016/j.celrep.2020.108455

Flores, A., Schell, J., Krall, A. S., Jelinek, D., Miranda, M., Grigorian, M., et al. (2017). Lactate Dehydrogenase Activity Drives Hair Follicle Stem Cell Activation. *Nat. Cel Biol* 19 (9), 1017–1026. doi:10.1038/ncb3575

Gu, H.-f., Tang, C.-k., and Yang, Y.-z. (2012). Psychological Stress, Immune Response, and Atherosclerosis. *Atherosclerosis* 223 (1), 69–77. doi:10.1016/j.atherosclerosis.2012.01.021

Ha, H., Kim, K. S., Yeom, Y. I., Lee, J. K., and Han, P. L. (2003). Chronic Restraint Stress Massively Alters the Expression of Genes Important for Lipid Metabolism and Detoxification in Liver. *Toxicol. Lett.* 146 (1), 49–63. doi:10.1016/j.toxlet.2003.09.006

Hackett, R. A., and Steptoe, A. (2017). Type 2 Diabetes Mellitus and Psychological Stress - a Modifiable Risk Factor. *Nat. Rev. Endocrinol.* 13 (9), 547–560. doi:10.1038/nrendo.2017.64

Hadshiew, I. M., Foitzik, K., Arck, P. C., and Paus, R. (2004). Burden of Hair Loss: Stress and the Underestimated Psychosocial Impact of Telogen Effluvium and Androgenetic Alopecia. *J. Invest. Dermatol.* 123 (3), 455–457. doi:10.1111/j.0022-202X.2004.23237.x

Harel, S., Higgins, C. A., Cerise, J. E., Dai, Z., Chen, J. C., Clynes, R., et al. (2015). Pharmacologic Inhibition of JAK-STAT Signaling Promotes Hair Growth. *Sci. Adv.* 1 (9), e1500973. doi:10.1126/sciadv.1500973

Huang, X., Tang, W., Lin, C., Sa, Z., Xu, M., Liu, J., et al. (2021). Protective Mechanism of Astragalus Polysaccharides against Cantharidin-induced Liver Injury Determined *In Vivo* by Liquid Chromatography/mass Spectrometry Metabolomics. *Basic Clin. Pharmacol. Toxicol.* 129, 61–71. doi:10.1111/bcpt.13585

Hung, C.-H., Lee, C.-H., Tsai, M.-H., Chen, C.-H., Lin, H.-F., Hsu, C.-Y., et al. (2020). Activation of Acid-Sensing Ion Channel 3 by Lysophosphatidylcholine 16:0 Mediates Psychological Stress-Induced Fibromyalgia-like Pain. *Ann. Rheum. Dis.* 79 (12), 1644–1656. doi:10.1136/annrheumdis-2020-218329

Ito, T. (2010). Hair Follicle Is a Target of Stress Hormone and Autoimmune Reactions. *J. Dermatol. Sci.* 60 (2), 67–73. doi:10.1016/j.jdermsci.2010.09.006

Kim, C. S., Ding, X., Allmeroth, K., Biggs, L. C., Kolenc, O. I., L'Hoest, N., et al. (2020). Glutamine Metabolism Controls Stem Cell Fate Reversibility and Long-Term Maintenance in the Hair Follicle. *Cel Metab.* 32 (4), 629–642. e628. doi:10.1016/j.cmet.2020.08.011

Kim, M. W., Shin, I. S., Yoon, H. S., Cho, S., and Park, H. S. (2017). Lipid Profile in Patients with Androgenetic Alopecia: a Meta-Analysis. *J. Eur. Acad. Dermatol. Venereol.* 31 (6), 942–951. doi:10.1111/jdv.14000

Kirsch, M., Talbiersky, P., Polkowska, J., Bastkowski, F., Schaller, T., de Groot, H., et al. (2009). A Mechanism of Efficient G6PD Inhibition by a Molecular Clip. *Angew. Chem. Int. Edition* 48 (16), 2886–2890. doi:10.1002/anie.200806175

Kotaka, M., Gover, S., Vandeputte-Rutten, L., Au, S. W. N., Lam, V. M. S., and Adams, M. J. (2005). Structural Studies of Glucose-6-Phosphate and NADP+binding to Human Glucose-6-Phosphate Dehydrogenase. *Acta Crystallogr. D Biol. Cryst.* 61 (Pt 5), 495–504. doi:10.1107/s0907444905002350

Lattouf, C., Jimenez, J. J., Tosti, A., Miteva, M., Wikramanayake, T. C., Kittles, C., et al. (2015). Treatment of Alopecia Areata with Simvastatin/ezetimibe. *J. Am. Acad. Dermatol.* 72 (2), 359–361. doi:10.1016/j.jaad.2014.11.006

Lee, J., and Tumbar, T. (2012). Hairy Tale of Signaling in Hair Follicle Development and Cycling. *Semin. Cel Develop. Biol.* 23 (8), 906–916. doi:10.1016/j.semcdb.2012.08.003

Lee, W.-S. (2011). Integral Hair Lipid in Human Hair Follicle. *J. Dermatol. Sci.* 64 (3), 153–158. doi:10.1016/j.jdermsci.2011.08.004

Liu, N., Wang, L.-H., Guo, L.-L., Wang, G.-Q., Zhou, X.-P., Jiang, Y., et al. (2013). Chronic Restraint Stress Inhibits Hair Growth via Substance P Mediated by Reactive Oxygen Species in Mice. *PLoS One* 8 (4), e61574. doi:10.1371/journal.pone.0061574

Magerl, M., Tobin, D. J., Müller-Röver, S., Hagen, E., Lindner, G., McKay, I. A., et al. (2001). Patterns of Proliferation and Apoptosis during Murine Hair Follicle Morphogenesis. *J. Invest. Dermatol.* 116 (6), 947–955. doi:10.1046/j.0022-202x.2001.01368.x

Malinska, D., Testoni, G., Duran, J., Brudnicka, A., Guinovart, J. J., and Duszynski, J. (2020). Hallmarks of Oxidative Stress in the Livers of Aged Mice with Mild Glycogen Branching Enzyme Deficiency. *Arch. Biochem. Biophys.* 695, 108626. doi:10.1016/j.abb.2020.108626

Marin, M. T., Cruz, F. C., and Planeta, C. S. (2007). Chronic Restraint or Variable Stresses Differently Affect the Behavior, Corticosterone Secretion and Body Weight in Rats. *Physiol. Behav.* 90 (1), 29–35. doi:10.1016/j.physbeh.2006.08.021

Mathay, C., Pierre, M., Pittelkow, M. R., Depiereux, E., Nikkels, A. F., Colige, A., et al. (2011). Transcriptional Profiling after Lipid Raft Disruption in Keratinocytes Identifies Critical Mediators of Atopic Dermatitis Pathways. *J. Invest. Dermatol.* 131 (1), 46–58. doi:10.1038/jid.2010.272

Millar, S. E. (2002). Molecular Mechanisms Regulating Hair Follicle Development. *J. Invest. Dermatol.* 118 (2), 216–225. doi:10.1046/j.0022-202x.2001.01670.x

Morinaga, H., Mohri, Y., Grachtchouk, M., Asakawa, K., Matsumura, H., Oshima, M., et al. (2021). Obesity Accelerates Hair Thinning by Stem Cell-Centric Converging Mechanisms. *Nature* 595, 266–271. doi:10.1038/s41586-021-03624-x

Müller-Röver, S., Foitzik, K., Paus, R., Handjiski, B., van der Veen, C., Eichmüller, S., et al. (2001). A Comprehensive Guide for the Accurate Classification of Murine Hair Follicles in Distinct Hair Cycle Stages. *J. Invest. Dermatol.* 117 (1), 3–15. doi:10.1046/j.0022-202x.2001.01377.x

Noerman, S., Klåvus, A., Järvelä-Reijonen, E., Karhunen, L., Auriola, S., Korpela, R., et al. (2020). Plasma Lipid Profile Associates with the Improvement of Psychological Well-Being in Individuals with Perceived Stress Symptoms. *Sci. Rep.* 10 (1), 2143. doi:10.1038/s41598-020-59051-x

Oliveira, T. G., Chan, R. B., Bravo, F. V., Miranda, A., Silva, R. R., Zhou, B., et al. (2016). The Impact of Chronic Stress on the Rat Brain Lipidome. *Mol. Psychiatry* 21 (1), 80–88. doi:10.1038/mp.2015.14

Palmer, M. A., Blakeborough, L., Harries, M., and Haslam, I. S. (2020). Cholesterol Homeostasis: Links to Hair Follicle Biology and Hair Disorders. *Exp. Dermatol.* 29 (3), 299–311. doi:10.1111/exd.13993

Patel, S., Kingsley, P. J., Mackie, K., Marnett, L. J., and Winder, D. G. (2009). Repeated Homotypic Stress Elevates 2-arachidonoylglycerol Levels and Enhances Short-Term Endocannabinoid Signaling at Inhibitory Synapses in Basolateral Amygdala. *Neuropsychopharmacol* 34 (13), 2699–2709. doi:10.1038/npp.2009.101

Paus, R., Arck, P., and Tiede, S. (2008). (Neuro-)endocrinology of Epithelial Hair Follicle Stem Cells. *Mol. Cell Endocrinol.* 288 (1-2), 38–51. doi:10.1016/j.mce.2008.02.023

Paus, R., Langan, E. A., Vidali, S., Ramot, Y., and Andersen, B. (2014). Neuroendocrinology of the Hair Follicle: Principles and Clinical Perspectives. *Trends Mol. Med.* 20 (10), 559–570. doi:10.1016/j.molmed.2014.06.002

Peters, E. M. J., Arck, P. C., and Paus, R. (2006). Hair Growth Inhibition by Psychoemotional Stress: a Mouse Model for Neural Mechanisms in Hair Growth Control. *Exp. Dermatol.* 15 (1), 1–13. doi:10.1111/j.0906-6705.2005.00372.x

Pratt, C. H., King, L. E., Jr., Messenger, A. G., Christiano, A. M., and Sundberg, J. P. (2017). Alopecia Areata. *Nat. Rev. Dis. Primers* 3, 17011. doi:10.1038/nrdp.2017.11

Quazi, F., and Molday, R. S. (2011). Lipid Transport by Mammalian ABC Proteins. *Essays Biochem.* 50 (1), 265–290. doi:10.1042/bse0500265

Rishikaysh, P., Dev, K., Diaz, D., Qureshi, W., Filip, S., and Mokry, J. (2014). Signaling Involved in Hair Follicle Morphogenesis and Development. *Int. J. Mol. Sci.* 15 (1), 1647–1670. doi:10.3390/ijms15011647

Rodriguez-Rodriguez, P., Fernandez, E., and Bolaños, J. P. (2013). Underestimation of the Pentose-Phosphate Pathway in Intact Primary Neurons as Revealed by Metabolic Flux Analysis. *J. Cereb. Blood Flow Metab.* 33 (12), 1843–1845. doi:10.1038/jcbfm.2013.168

Rohleder, N., Nater, U. M., Wolf, J. M., Ehlert, U., and Kirschbaum, C. (2004). Psychosocial Stress-Induced Activation of Salivary Alpha-Amylase: an

Indicator of Sympathetic Activity? *Ann. N. Y Acad. Sci.* 1032, 258–263. doi:10.1196/annals.1314.033

Ruiying, C., Zeyun, L., Yongliang, Y., Zijia, Z., Ji, Z., Xin, T., et al. (2020). A Comprehensive Analysis of Metabolomics and Transcriptomics in Non-small Cell Lung Cancer. *PLoS One* 15 (5), e0232272. doi:10.1371/journal.pone.0232272

Samuelov, L., Kinori, M., Bertolini, M., and Paus, R. (2012). Neural Controls of Human Hair Growth: Calcitonin Gene-Related Peptide (CGRP) Induces Catagen. *J. Dermatol. Sci.* 67 (2), 153–155. doi:10.1016/j.jdermsci.2012.04.006

Sheng, R., Kim, H., Lee, H., Xin, Y., Chen, Y., Tian, W., et al. (2014). Cholesterol Selectively Activates Canonical Wnt Signalling over Non-canonical Wnt Signalling. *Nat. Commun.* 5, 4393. doi:10.1038/ncomms5393

Shin, J.-M., Jung, K.-E., Yim, S.-H., Rao, B., Hong, D., Seo, Y.-J., et al. (2021). Putative Therapeutic Mechanisms of Simvastatin in the Treatment of Alopecia Areata. *J. Am. Acad. Dermatol.* 84 (3), 782–784. doi:10.1016/j.jaad.2020.03.102

Simsek, T., Kocabas, F., Zheng, J., Deberardinis, R. J., Mahmoud, A. I., Olson, E. N., et al. (2010). The Distinct Metabolic Profile of Hematopoietic Stem Cells Reflects Their Location in a Hypoxic Niche. *Cell Stem Cell* 7 (3), 380–390. doi:10.1016/j.stem.2010.07.011

Skosnik, P. D., Chatterton, R. T., Jr., Swisher, T., and Park, S. (2000). Modulation of Attentional Inhibition by Norepinephrine and Cortisol after Psychological Stress. *Int. J. Psychophysiology* 36 (1), 59–68. doi:10.1016/s0167-8760(99)00100-2

Stenn, K. S., and Paus, R. (2001). Controls of Hair Follicle Cycling. *Physiol. Rev.* 81 (1), 449–494. doi:10.1152/physrev.2001.81.1.449

Takubo, K., Nagamatsu, G., Kobayashi, C. I., Nakamura-Ishizu, A., Kobayashi, H., Ikeda, E., et al. (2013). Regulation of Glycolysis by Pdk Functions as a Metabolic Checkpoint for Cell Cycle Quiescence in Hematopoietic Stem Cells. *Cell Stem Cell* 12 (1), 49–61. doi:10.1016/j.stem.2012.10.011

Tarling, E. J., Vallim, T. Q. d. A., and Edwards, P. A. (2013). Role of ABC Transporters in Lipid Transport and Human Disease. *Trends Endocrinol. Metab.* 24 (7), 342–350. doi:10.1016/j.tem.2013.01.006

Wang, L., Guo, L.-L., Wang, L.-H., Zhang, G.-X., Shang, J., Murao, K., et al. (2015). Oxidative Stress and Substance P Mediate Psychological Stress-Induced Autophagy and Delay of Hair Growth in Mice. *Arch. Dermatol. Res.* 307 (2), 171–181. doi:10.1007/s00403-014-1521-3

Wang, Q., Wang, Y., Pang, S., Zhou, J., Cai, J., and Shang, J. (2019). Alcohol Extract from Vernonia Anthelmintica Willd (L.) Seed Counteracts Stress-Induced Murine Hair Follicle Growth Inhibition. *BMC Complement. Altern. Med.* 19 (1), 372. doi:10.1186/s12906-019-2744-9

Williamson, D., Gonzalez, M., and Finlay, A. (2001). The Effect of Hair Loss on Quality of Life. *J. Eur. Acad. Dermatol. Venerol* 15 (2), 137–139. doi:10.1046/j.1468-3083.2001.00229.x

Yu, B., and Wang, J. (2021). Lipidomics Identified Lyso-Phosphatidylcholine and Phosphatidylethanolamine as Potential Biomarkers for Diagnosis of Laryngeal Cancer. *Front. Oncol.* 11, 646779. doi:10.3389/fonc.2021.646779

Zhao, X., Seese, R. R., Yun, K., Peng, T., and Wang, Z. (2013). The Role of Galanin System in Modulating Depression, Anxiety, and Addiction-like Behaviors after Chronic Restraint Stress. *Neuroscience* 246, 82–93. doi:10.1016/j.neuroscience.2013.04.046

Conflict of Interest: CC was employed by the company Yonghe Medical Group Co. Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Molecular Biosciences

MINI REVIEW
published: 16 February 2022
doi: 10.3389/fmolb.2022.841223

Gut Microbiome and the Role of Metabolites in the Study of Graves' Disease

Haihua Liu [1,2], *Huiying Liu* [1,2], *Chang Liu* [1,2], *Mengxue Shang* [1,2], *Tianfu Wei* [1,2] *and Peiyuan Yin* [1]*

[1] Clinical Laboratory of Integrative Medicine, First Affiliated Hospital of Dalian Medical University, Dalian, China, [2] Institute of Integrative Medicine, Dalian Medical University, Dalian, China

Graves' disease (GD) is an autoimmune thyroid disease (AITD), which is one of the most common organ-specific autoimmune disorders with an increasing prevalence worldwide. But the etiology of GD is still unclear. A growing number of studies show correlations between gut microbiota and GD. The dysbiosis of gut microbiota may be the reason for the development of GD by modulating the immune system. Metabolites act as mediators or modulators between gut microbiota and thyroid. The purpose of this review is to summarize the correlations between gut microbiota, microbial metabolites and GD. Challenges in the future study are also discussed. The combination of microbiome and metabolome may provide new insight for the study and put forward the diagnosis, treatment, prevention of GD in the future.

Keywords: graves' disease (GD), metabol(n)omics, gut microbiome, autoimmunity, metabolites

INTRODUCTION

Autoimmune thyroid disease (AITD) are common organ-specific autoimmune disorders with an increasing prevalence worldwide, which involves Hashimoto thyroiditis (HT) and GD (Moshkelgosha et al., 2021). GD is caused by the autoantibodies of the thyrotropin receptor (TSHR), which leads to thyroid hyperplasia and hyperthyroidism (Bahn, 2003; Ishaq et al., 2018; Shi et al., 2019a; Moshkelgosha et al., 2021). Hyperthyroidism, fatigue, weight loss, tachycardia, and heat intolerance are common symptoms of GD. Approximately 50% of patients may develop Graves' ophthalmopathy (GO), leading to eyelid retractions and exophthalmos (Byeon et al., 2018; Yan et al., 2020). GD is the most common cause of 60–80% of hyperthyroidism and influence about 0.5% of the general population (Cooper and Stroehla, 2003; Smith and Hegedüs, 2016; Ejtahed et al., 2020). It frequently occurs in the population between 30 and 50 years old. Resemble in other autoimmune diseases, the incidence of GD is higher in women than men, the ratio of about 5/1 (Cooper and Stroehla, 2003; Ji et al., 20188; Nyström et al., 2013; Menconi et al., 2014). The risk factors of GD include genetic predisposition, environmental factors, immune factors (Covelli and Ludgate, 2017).

Hyperthyroidism is a common disease that is difficult to cure completely. Although modern medicine has brought great changes to the prevention, diagnosis, and treatment of autoimmune diseases, the etiology and pathogenesis of these diseases have not been fully illuminated. Abnormal thyroid-related indices often occur repeatedly during clinical treatment (Yang et al., 2019). Furthermore, although current treatment methods for GD can achieve a good effect, clinicians still have some concerns about the choice of treatment for safety reasons (Heyma et al., 1986; Yang et al., 2019). At present, a large number of studies have proved the relationship between intestinal microorganisms and autoimmune diseases, including Type 1 diabetes (Gianchecchi and Fierabracci, 2017; Mullaney et al., 2018), inflammatory bowel disease (Ni et al., 2017; Cao, 2018), systemic lupus

OPEN ACCESS

Edited by:
Wolfram Weckwerth,
University of Vienna, Austria

Reviewed by:
Thomas Kaiser,
University of Minnesota Twin Cities, United States

****Correspondence:***
Peiyuan Yin
yinperry@126.com
yinpeiyuan@dm.edu.cn

Specialty section:
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

Received: *22 December 2021*
Accepted: *31 January 2022*
Published: *16 February 2022*

Citation:
Liu H, Liu H, Liu C, Shang M, Wei T and Yin P (2022) Gut Microbiome and the Role of Metabolites in the Study of Graves' Disease.
Front. Mol. Biosci. 9:841223.
doi: 10.3389/fmolb.2022.841223

erythematosus (Corrêa et al., 2017), rheumatoid arthritis (Sato et al., 2017; Teng et al., 2017; Jubair et al., 2018; Picchianti-Diamanti et al., 2018) and autoimmune thyroid disease (Zhou et al., 2014). Metabolites are also considered as important mediators or modulators between gut microbiota and the thyroid. Therefore, metabolomics investigations may provide a new inside view of GD's study.

In this review, we explore the inside relationships between gut microbiota, microbiota-related metabolites and GD, and propose new ideas for prevention, diagnosis, and treatment of GD.

Brife Knowledge of Gut Microbiota

The human body is a superorganism due to the residence of trillions of prokaryotes symbiosis. Approximately 66% of the total bacteria are mainly live in the gut. Gut microbiota includes more than one thousand known species of bacteria with at least three million genes (Hehemann et al., 2010; Relman, 2012; Docimo et al., 2020). Apart from absorbing nutrients from the human body that they depend on for survival, intestinal flora also provides beneficial or harmful metabolites to the human body through their metabolic process (Turnbaugh and Gordon, 2009; Relman, 2012). These microflorae participate in the body's energy metabolism through various mechanisms, affecting the conversion of food to energy in the host, and play an essential role in the healthy state of the host (Lozupone et al., 2012; Sommer et al., 2017). When the human body is healthy, microorganisms and various organs and tissues depend on each other and act on either to form a microecological balance and jointly maintain the body's health. If the microecological balance is disturbed, it may lead to disease (Sekirov et al., 2010). Therefore, the intestinal flora is considered an "organ" with multiple regulatory functions, which greatly impacts people's health. Understanding the symbiotic relationship between microorganisms and the human body is of great significance for people to understand their health and the occurrence and development of disease (Turnbaugh and Gordon, 2009; Relman, 2012; Schmidt et al., 2018).

The technological breakthroughs in the microbiome boost the research of gut microbiota. The method of bacterial culture is a restriction of traditional bacterial research. The intestinal flora is cultured with various mediums, and the number of bacterial colonies is measured by dilution and colony count (Lagier et al., 2018). This method is sensitive but is constrained. More than 85% of the bacteria in the human intestine are anaerobic bacteria, which is difficult to cultivate in the culture medium (Lagier et al., 2015). Recently, the newly established strategy of culturomics enables the culture of microbiota that cannot be cultured before. These new methods initiate the rebirth of culture in microbiology (Kaeberlein et al., 2002; Nichols et al., 2010; Lagier et al., 2018). The development of new techniques has made it possible to study unknown gut flora.16Sr RNA high-throughput sequencing and metagenomics are commonly used methods for detecting gut microbiota. 16Sr RNA sequencing mainly studies the species composition, the evolutionary relationships among species and the diversity of communities (Laudadio et al., 2018). On the basis of 16Sr RNA sequencing analysis, metagenomic sequencing can also carry out in-depth research on gene and function, and its detection depth can reach the level of species (Wang et al., 2015; Laudadio et al., 2018; Shakya et al., 2019).

With the increasing understanding of the metabolic function of intestinal flora, the narrow sense that host metabolism is regulated by its genes is gradually expanded to co-metabolic regulation of host-symbiotic intestinal bacteria. These metabolites are often from tryptophan metabolic pathways, tyrosine and phenylalanine metabolic pathways, glucose and fatty acid metabolic pathways, classified into indoles, phenols, amino acids, peptides, etc. (Zheng et al., 2011; Van Treuren and Dodd, 2020; Fan and Pedersen, 2021). Microbiome dysbiosis is associated with various diseases, asthma, allergies, inflammatory bowel disease (Arrieta et al., 2015; Bunyavanich et al., 2016; Nishino et al., 2018), autism spectrum disorder (ASD) (Needham et al., 2021), diabetes (Giongo et al., 2011), irritable bowel syndrome (IBS) (Mars et al., 2020), obesity (Schwiertz et al., 2010a), cardiovascular disease (Jie et al., 2017), chronic kidney disease (Sircana et al., 2019). Under different disease states, the species abundance of intestinal flora and its related metabolites have various characteristics. Some studies have found that in patients with IBS, the key findings include an increase in Firmicutes to Bacteroidetes ratio (Krogius-Kurikka et al., 2009; Rajilić-Stojanović et al., 2011; Jeffery et al., 2012l; Mars et al., 2020), a decrease in Bifidobacteria and Lactobacilli (Malinen et al., 2005; Kerckhoffs et al., 2009), and an increase in Ruminococcus and Streptococci species (Kassinen et al., 2007; Rajilić-Stojanović et al., 2011; Saulnier et al., 2011; Hong and Rhee, 2014). A more coincident finding has been decreased alpha diversity. ASD showed lower levels of phylum Firmicutes and a higher abundance of Bacteroidetes (Mangiola et al., 2016; Fattorusso et al., 2019; Sharon et al., 2019). Kang and others observed significant ASD-related behavioral changes in mice with fecal microbiota transplantation (FMT) from ASD (Sharon et al., 2019) and they have developed microbiome transfer therapy (MTT) and observed a reduction in ASD-related symptoms (Kang et al., 2017).

The intestines are also the largest immune organ, gathering more than 70% of the immune cells as a vital digestive organ. Gut microbiota is also related to the host's immune system (Vatanen et al., 2016). Gut microbiota and metabolites can induce the production of helper T cells (Th) and regulator T cells (Tregs), which contribute to the maturation of host adaptive and innate immunity (Rooks and Garrett, 2016; Shi et al., 2017; Kayama et al., 2020). It can be inferred that autoimmune diseases are closely related to intestinal flora (Levy et al., 2017). There are several studies on the gut microbiota and metabolome among GD patients, and many results strongly support a role for the gut microbiota in GD and GO (Moshkelgosha et al., 2021).

GD and Gut Microbiota

Some previous studies demonstrated the connections between the gut microbiome and AITD (Köhling et al., 2017). Many studies showed that GD is related to *yersinia* enterocolitica, e.g., mice fed only with *yersinia* enterocolitica did not develop GD (Weiss et al., 1983; Wang et al., 2010). There were also significant differences in the microbiota profile between HT patients and healthy controls (Zhao et al., 2018). Zhou et al. characterized the gut microbiota in hyperthyroid patients (Zhou et al., 2014). There is limited research on the relationships between Graves' disease and the gut microbiome. However, thyroid hormone levels correlate with the gut microbiome and the diversity of gut bacteria in patients with GD (Ejtahed et al., 2020). Bacteroidetes and Firmicutes are dominant species in the human gut. The ratio of Firmicutes to Bacteroidetes is commonly considered a representative of health status (Chen et al., 2016; Indiani et al., 2018). In the disease state, these two phyla tend to show significant changes. For example, Jiang et al. showed that GD patients had reduced alpha diversity compared with healthy individuals. At the phylum level, GD patients had a significant higher proportion of Bacteroidetes and a significantly lower proportion of Firmicutes than the controls (Jiang et al., 2021). Ishaq et al. also demonstrated this phenomenon in their study (Ishaq et al., 2018). They found that the diversity of gut bacteria in GD patients was less diverse in terms of richness than in healthy people. The proportion of Firmicutes in GD was lower than that in the control group, while the proportion of Bacteroidetes was higher than in the control group (Ishaq et al., 2018). Interestingly, this finding is consistent with what was observed in obese patients. Previous studies have found that obese people tend to have more Firmicutes, while lean people tend to have more Bacteroidetes (Schwiertz et al., 2010b; Riva et al., 2017). Further research work is required about the effects of thyroid hormones on gut microbiota. Besides Firmicutes and Bacteroidetes, there were also significant changes in the ratios and abundances of other phyla. Yan et al. showed that the number of Lactobacillales, Bacilli, Megamonas, Prevotalla and Veillonella strains were increased among GD patients (Yan et al., 2020). However, the number of Rikenellaceae, Ruminococuus and Alistipes strains was decreased among GD patients. In addition, the diversity of gut flora was decreased in patients with GD (Yan et al., 2020). There were also significant changes in gut microbiota in GO patients. Shi et al. found that the bacterial diversity (Simpson and Shannon) was reduced in patients with GO compared to the controls. At the phylum levels, the proportion of Bacteroidetes increased and Firmicutes decreased significantly in GO than that in controls. There were obvious differences in bacterial profiles between the two groups (Shi et al., 2019a). Then, Shi et al. further explored the differences in the compositing of gut microbes between GO and GD patients (Shi et al., 2021). At the phylum levels, the proportion of Chloroflexi was decreased significantly in GO patients. At the genus levels, Bilophila and Subdoligranulum were increased (Shi et al., 2021). It is reported that there are three gut bacteria genera (*Bacteroides*, Prevotella, Alistipes) that could separate GD patients from healthy individuals with 85% accuracy (Su et al., 2020).

Thyrotropin receptor antibody (TRAb) is a characteristic indicator of GD, with sensitivity and specificity greater than 95% for GD diagnosis (Massart et al., 2001; Cooper, 2003). Shi et al. believed that TRAb was significantly correlated with different levels of gut microbiota. At the family level, the proportion of Succinivibrionaceae was positively correlated to TRAb. At the genus level, Subdoligranulum was positively related to TRAb. At the species level, Parabacteroides distasonis showed an opposite correlation with TRAb. Their studies also suggested that GD patients with positive TRAb showed an increased risk of developing GO (Shi et al., 2019a). Prevotella and *Bacteroides* are positively correlated with TRAb in GO patients (Shi et al., 2019b).

Metabolomics in the Study of GD

The dynamic balance of Th17 and Treg is closely related to the occurrence and development of various autoimmune diseases (Fasching et al., 2017). Treg cells are a subset of regulatory T cells that regulate the body's autoimmune response. Tregs are characterized by the transcription factor Foxp3 (major regulators of Treg) and mainly exert immune suppressive effects. Maintaining immune homeostasis by secreting inhibitory factors (TGF-B, IL-10, IL-35) mediate immune suppressive effects by regulating TCR signaling promotes secretion and differentiation of anti-inflammatory cytokines (Göschl et al., 2019). The decrease of Treg cells increases the incidence and severity of AITD. And the number of Treg cells is significantly reduced in patients with GD (Saitoh and Nagayama, 2006; Nakano et al., 2007). The Th17 cells are also a subset of T helper cells by secreting interleukin 17 (IL-17, IL-22) induces inflammation and spread. IL-17 is involved in many inflammatory and autoimmune diseases, including systemic and organ-specific autoimmune diseases (Takeuchi et al., 2020; Yasuda et al., 2019). Th17 and IL-17 were increased in GD and participated in the development of GD. In patients with AITD, the proportion of Th17 cells in peripheral blood mononuclear cells (PBMCs) increased and higher mRNA level of their specific transcription factor RORγt in PBMCs (Li et al., 2016; Li et al., 2019). However, the level of Tregs and expression of Foxp3 mRNA were greatly decreased in AITD (Li et al., 2016; Li et al., 2019). Figueroa Vega et al. found that IL-17 was elevated in the thyroid tissues of GD RORγt mRNA patients, and both IL-17 and IL-22 levels were higher than healthy controls (Figueroa-Vega et al., 2010). Di. Peng observed that the concentration of IL-17 and IL-22 in plasma of GD patients was significantly higher than that of healthy controls, which was consistent with the increase of Th17 cells and positively correlated with TSAb (Peng et al., 2013). However, some studies have shown the opposite results (Yuan et al., 2017). The metabolites of the gut microbiome have been associated with the generation of proinflammatory cytokines and the production of Th17 cells. Commensal bacteria and their metabolites can also promote Treg generation and suppress the immune system (Haase et al., 2018). SCFAs are produced by the fermentation of non-digestible carbohydrates such as dietary fiber by gut bacteria, including butyrate (C (Shi et al., 2019a)), propionate (C (Ishaq et al., 2018)) and acetate (C (Bahn, 2003)), are essential metabolites in maintaining homeostasis (Luu and Visekruna, 2019). SCFAs have been proved to alter chemotaxis and phagocytosis, changes in cell function and proliferation, induction of reactive oxygen species (ROS), anti-tumor and anti-inflammatory (Tan et al., 2014). SCFAs contribute to the maintenance of intestinal barrier integrity and its regeneration effect on the intestinal epithelium (Memba et al., 2017). SCFAs are valuable sources of nutrients for enterocytes, together with thyroid hormones (chiefly triiodothyronine), stimulating

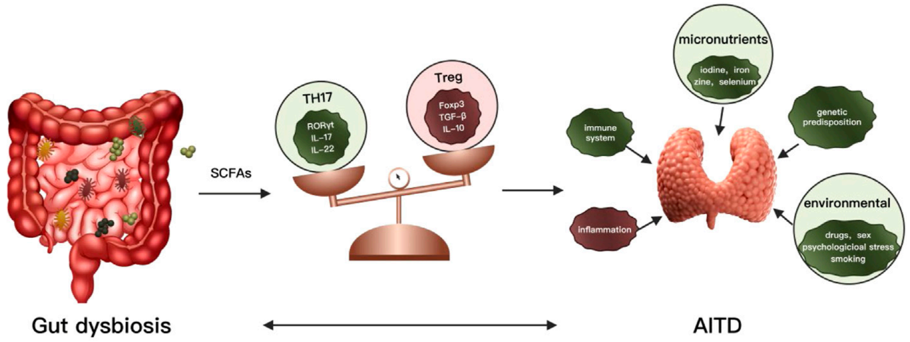


FIGURE 1 | Association between gut microbiota, metabolites, and thyroid autoimmune diseases.

enterocyte differentiation (Cayres et al., 2021; Meng et al., 1999). It also increases intercellular integrity and reduces the risk of a "leaky gut" by improving the adhesion of intestinal cells and reducing the PH in the intestinal tract, thus avoiding the invasion of pathological organisms (Memba et al., 2017; Bargiel et al., 2021). It is suggested that GD's development is often linked to a compromised intestinal barrier (Knezevic et al., 2020). Recent studies emphasized the immunomodulatory potential of SCFAs in various autoimmune diseases and inflammatory disorders such as multiple sclerosis (MS), colitis, rheumatoid arthritis and AITD. The relation between SCFAs and thyroid function seems to be confirmed by several studies in the scientific literature describing changes in the gut microbiota, including concentrations of SCFAs in impaired thyroid status (Virili et al., 2018; Liu et al., 2020). Currently, two essential functions for SCFAs have been identified, inhibition of histone deacetylases (HDACs) and activation of G-protein coupled receptors (GPCRs), particularly GPR43, GPR41 and GPR109 A (Tan et al., 2014) (Sivaprakasam et al., 2016). Butyrate has been shown to have a positive effect on rheumatoid arthritis (Hui et al., 2019), inflammatory bowel disease (IBD) (Zhou et al., 2018) and autoimmune hepatitis (AIH) (Hu et al., 2018) by rebalancing between Treg and Th17 and increasing the number of Treg cells and decreasing Th17 cells in the system (**Figure 1**). Propionate is found to affect multiple sclerosis (MS) (Duscha et al., 2020) and GD (Su et al., 2020). However, little is known about the role of the SCFAs in Graves' disease.

Struja et al. predicted the relapse of hyperthyroidism based on the assessment of metabonomics differences. Pyruvate and triglycerides are considered as predictors with AUCs of 0.73 and 0.67 (Struja et al., 2018). Al-Majdoub and others reported changes in the carnitine metabolism of GD patients prior to treatment compared to posttreatment (Al-Majdoub et al., 2017). The level of short-chain acylcarnitine decreased, medium-chain acylcarnitine increased, and long-chain acylcarnitine remained unchanged. The authors speculated that these phenomena reflect a starvation process that induced by hyperthyroidism (Al-Majdoub et al., 2017). Lipid profile from plasma and urine samples of GD patients was significantly different compared to controls. Some of Glycerophosphoethanolamine (PE), Glycerophosphoinositol (PI), Triacylglycerol (TG) and Glycerophosphoglycerol (PG) have changed significantly (Byeon et al., 2018). Polyamine metabolic profiles are also altered in AITD. GD and HT patients showed the same change relative to the control group for most of the polyamine metabolites. L-arginine (L-ARG), L-omithine (L-ORN), lysine (LYS) agmatine (AGM) are significantly and N-acetylputrescine (NPUT), spermine (SPM), 1,3-diaminopropane (DAP) are lower than the control group. However, GD and HT have different characteristics of change. GD patients had significantly lower cadverine (CAD) and higher N-acetylspermidine (NSPD), spermidine (SPD) and r-Aminobutyric (GABA) acid than the control group. But N-acetylspermine (NSPM) was decreased in HT. The anti-inflammatory effect of SPM was better than that of SDP. SPM/SPD can be more effective for estimating the anti-inflammatory effect. A decrease in SPM/SPD in patients with AITD indicated reduce in protective polyamines. SPM/SPD was negatively correlated with inflammatory chemokine IP-10 and TPOAb (Rider et al., 2007; Song et al., 2019). Ji et al. performed a non-targeted metabolomics analysis on the blood and orbital tissues of GD, GO and healthy controls. They identified ten differential metabolites in the disease group (gluconic acid, glucose, pelargonic acid, threose, fumaric acid, glycerol, mannose, pentade canoic acid, pyruvate, and 2- (4-hydroxyphenyl)ethanol) (Ji et al., 20188). The metabolite panel achieved an accuracy of 0.931 and the sensitivity and specificity are 0.787 and 0.875, respectively (Ji et al., 20188). Among the metabolite panel, almost all metabolites showed a positive correlation with the levels of TRAb (Ji et al., 20188). Propionate was significantly reduced in GD patients, which was negatively correlated with FT3, FT4, TRAb level, and positively correlated with TSH level (Su et al., 2020). At present, there are not many studies on GD metabolomics, and the specific association and mechanism still need to be further studied.

Gut dysbiosis can lead to changes in metabolites such as SCFAs. As a consequence, the balance of Th17 and Tregs would be damaged, leading to an autoimmune response and causing autoimmune thyroid diseases. AITD: autoimmune thyroid diseases; IL: interleukin; Th: T helper cell; Tregs: regulatory T cells.

Microbiome and Metabolome in GD Study

In the last 20 years, it has been established that the gut microbiome plays an essential role in maintaining host health and the occurrence and progression of the disease. Metabolites are the primary way that gut microbes interact with hosts. The small molecules generated or modified from microorganisms can be detected in urine, serum, feces, cerebrospinal fluid, and other tissues (Holmes et al., 2011; Del Rio et al., 2017). The homeostasis of a healthy intestinal environment is regulated by the balance of microbiota, metabolites, and immune systems. In the state of disease, the intestinal balance is usually destroyed. Studies showed that gut dysbiosis leads to Treg/Th17 imbalance through the propionic acid regulation pathway, which, together with other pathogenic factors, promotes GD occurrence (Su et al., 2020). Gut dysbiosis was mainly manifested by a significant decrease in SCFAs-producing bacteria and SCFAs. *Bacteroides fragilis* YCH46 strain in GD patients was obviously reduced compared to healthy controls. It can produce propionic acid, increase the number of Treg cells and reduce the number of Th17 cells. Therefore, *B. fragilis* YCH46 was a natural activator of Treg cells and inhibitor of Th17 cells (Rios-Covian et al., 2015). YCH46 strain of *B. fragilis* provides a new direction for the treatment of GD. It can improve immune dysfunction in GD patients and be used as an immunomodulator or as an auxiliary treatment for GD patients to reduce recurrence rate (Su et al., 2020). A recent study found significant differences in metabolic pathways between GD groups and healthy controls. Formaldehyde assimilation and allantoin degradation, mevalonate and isoprene biosynthesis significantly increased in the GD patients. In contrast, the microbial metabolic abilities of fatty acid biosynthesis, pyruvate fermentation to hexanol, anaerobic energy metabolism, creatinine degradation and gluconeogenesis decreased significantly in relative abundance in the patients. The change of gut microbiota is Butyricimonas *faecalis*, Faecalibacterium prausnitzii, Akkermansia muciniphila and Bifidobacterium adolescentis decreased in the GD, whereas Veillonella parvula, Eggerthella lenta, *Fusobacterium* mortiferum, *Streptococcus* parasanguinis, and *Streptococcus* salivarius were enriched. And use propionic acid, acetic acid, cholate and chenodeoxycholate as potential biomarkers (Zhu et al., 2021). Jiang et al. found that Blautia, Eubacterium and Anaerostipes were decreased in GD. Eubacterium and Anaerostipes produce butyric acid and maintain the integrity of the intestinal epithelium as well as induce the generate of Treg cells to strengthen the tightness of the intestinal mucosal barrier (Duncan et al., 2004; Venkataraman et al., 2016; Jiang et al., 2021). The primary metabolite of Blautia is butyric acid and has been shown to have anti-inflammatory effects (Jenq et al., 2015). The decrease of these three butyric acid-producing bacteria leads to the reduction of butyric acid and inhibits the differentiation of Treg cells, resulting in immune system dysfunction and eventually the development of AITD (Jiang et al., 2021).

DISCUSSION

Autoimmune diseases are still challenging for the clinic. Changes in the composition and abundance of the gut microbiota, as well as related metabolites, are closely linked to the occurrence of GD. These findings provide some potential biomarkers for early diagnosis of GD, and some new probiotics related to GD can be used for adjunctive treatment and prevention of recurrence. However, related studies on gut microbiota metabolome in patients with GD are relatively lacking, and further studies are needed. It is believed that probiotics have positive effects on thyroid diseases, which has been confirmed *in vitro* cell studies and animal studies. However, these effects on human beings still require intensive investigations. Accurate qualitative-quantitative characterization of probiotics according to different pathological stages are also needed. Current metabolomics studies provide the correlations between gut micrbiota and the disease, however, the molecular mechanism between gut microbiota and GD remain unclear. One of the key point is how the metabolites synthesized by the gut microbiota. This is essential for the following development of related medicines.

The ultimate goal for the multi-omics study is to develop new diagnostic standards (microbial/metabolite biomarkers) and treatment strategies (probiotics/targeted microbial therapy or functional metabolites) for GD, with an individual treatment plan for each patient to achieve a complete cure and prevent a recurrence.

AUTHOR CONTRIBUTIONS

HaL write the manuscript, PY revised the paper. HuL, CL, MS, and TW review the manuscript.

REFERENCES

Al-Majdoub, M., Lantz, M., and Spégel, P. (2017). Treatment of Swedish Patients with Graves' Hyperthyroidism Is Associated with Changes in Acylcarnitine Levels. *Thyroid* 27, 1109–1117. doi:10.1089/thy.2017.0218

Arrieta, M.-C., Stiemsma, L. T., Dimitriu, P. A., Thorson, L., Russell, S., Yurist-Doutsch, S., et al. (2015). Early Infancy Microbial and Metabolic Alterations Affect Risk of Childhood Asthma. *Sci. Transl. Med.* 7, 307ra152. doi:10.1126/scitranslmed.aab2271

Bahn, R. S. (2003). Pathophysiology of Graves' Ophthalmopathy: The Cycle of Disease. *J. Clin. Endocrinol. Metab.* 88, 1000–1946. doi:10.1210/jc.2002-030010

Bargiel, P., Szczuko, M., Stachowska, L., Prowans, P., Czapla, N., Markowska, M., et al. (2021). Microbiome Metabolites and Thyroid Dysfunction. *Jcm* 10, 3609. doi:10.3390/jcm10163609

Bunyavanich, S., Shen, N., Grishin, A., Wood, R., Burks, W., Dawson, P., et al. (2016). Early-life Gut Microbiome Composition and Milk Allergy Resolution. *J. Allergy Clin. Immunol.* 138, 1122–1130. doi:10.1016/j.jaci.2016.03.041

Byeon, S. K., Park, S. H., Lee, J. C., Hwang, S., Ku, C. R., Shin, D. Y., et al. (2018). Lipidomic Differentiation of Graves' Ophthalmopathy in Plasma and Urine from Graves' Disease Patients. *Anal. Bioanal. Chem.* 410, 7121–7133. doi:10.1007/s00216-018-1313-2

Cao, S. S. (2018). Cellular Stress Responses and Gut Microbiota in Inflammatory Bowel Disease. *Gastroenterol. Res. Pract.* 2018, 1–13. doi:10.1155/2018/7192646

Cayres, L. C. d. F., de Salis, L. V. V., Rodrigues, G. S. P., Lengert, A. v. H., Biondi, A. P. C., Sargentini, L. D. B., et al. (2021). Detection of Alterations in the Gut Microbiota and Intestinal Permeability in Patients with Hashimoto Thyroiditis. *Front. Immunol.* 12, 579140. doi:10.3389/fimmu.2021.579140

Chen, S., Cheng, H., Wyckoff, K. N., and He, Q. (2016). Linkages of Firmicutes and Bacteroidetes Populations to Methanogenic Process Performance. *J. Ind. Microbiol. Biotechnol.* 43, 771–781. doi:10.1007/s10295-016-1760-8

Cooper, D. S. (2003). Hyperthyroidism. *The Lancet* 362, 459–468. doi:10.1016/s0140-6736(03)14073-1

Cooper, G. S., and Stroehla, B. C. (2003). The Epidemiology of Autoimmune Diseases. *Autoimmun. Rev.* 2 (3), 119–125. doi:10.1016/s1568-9972(03)00006-5

Corrêa, J. D., Calderaro, D. C., Ferreira, G. A., Mendonça, S. M. S., Fernandes, G. R., Xiao, E., et al. (2017). Subgingival Microbiota Dysbiosis in Systemic Lupus Erythematosus: Association with Periodontal Status. *Microbiome* 5, 34. doi:10.1186/s40168-017-0252-z

Covelli, D., and Ludgate, M. (2017). The Thyroid, the Eyes and the Gut: a Possible Connection. *J. Endocrinol. Invest.* 40, 567–576. doi:10.1007/s40618-016-0594-6

Del Rio, D., Zimetti, F., Caffarra, P., Tassotti, M., Bernini, F., Brighenti, F., et al. (2017). The Gut Microbial Metabolite Trimethylamine-N-Oxide Is Present in Human Cerebrospinal Fluid. *Nutrients* 9, 1053. doi:10.3390/nu9101053

Docimo, G., Cangiano, A., Romano, R. M., Pignatelli, M. F., Offi, C., Paglionico, V. A., et al. (2020). The Human Microbiota in Endocrinology: Implications for Pathophysiology, Treatment, and Prognosis in Thyroid Diseases. *Front. Endocrinol.* 11, 586529. doi:10.3389/fendo.2020.586529

Duncan, S. H., Louis, P., and Flint, H. J. (2004). Lactate-utilizing Bacteria, Isolated from Human Feces, that Produce Butyrate as a Major Fermentation Product. *Appl. Environ. Microbiol.* 70 (10), 5810–5817. doi:10.1128/AEM.70.10.5810-5817.2004

Duscha, A., Gisevius, B., Hirschberg, S., Yissachar, N., Stangl, G. I., Eilers, E., et al. (2020). Propionic Acid Shapes the Multiple Sclerosis Disease Course by an Immunomodulatory Mechanism. *Cell* 180, 1067–1080. e16. doi:10.1016/j.cell.2020.02.035

Ejtahed, H.-S., Angoorani, P., Soroush, A.-R., Siadat, S.-D., Shirzad, N., Hasani-Ranjbar, S., et al. (2020). Our Little Friends with Big Roles: Alterations of the Gut Microbiota in Thyroid Disorders. *Emiddt* 20, 344–350. doi:10.2174/1871530319666190930110605

Fan, Y., and Pedersen, O. (2021). Gut Microbiota in Human Metabolic Health and Disease. *Nat. Rev. Microbiol.* 19, 55–71. doi:10.1038/s41579-020-0433-9

Fasching, P., Stradner, M., Graninger, W., Dejaco, C., and Fessler, J. (2017). Therapeutic Potential of Targeting the Th17/Treg Axis in Autoimmune Disorders. *Molecules* 22, 134. doi:10.3390/molecules22010134

Fattorusso, A., Di Genova, L., Dell'Isola, G., Mencaroni, E., and Esposito, S. (2019). Autism Spectrum Disorders and the Gut Microbiota. *Nutrients* 11, 521. doi:10.3390/nu11030521

Figueroa-Vega, N., Alfonso-Pe´rez, M., Benedicto, I., Sa´nchez-Madrid, F., Gonza´lez-Amaro, R., and Marazuela, M. (2010). Increased Circulating Pro-inflammatory Cytokines and Th17 Lymphocytes in Hashimoto's Thyroiditis. *J. Clin. Endocrinol. Metab.* 95, 953–962. doi:10.1210/jc.2009-1719

Gianchecchi, E., and Fierabracci, A. (2017). On the Pathogenesis of Insulin-dependent Diabetes Mellitus: the Role of Microbiota. *Immunol. Res.* 65 (1), 242–256. doi:10.1007/s12026-016-8832-8

Giongo, A., Gano, K. A., Crabb, D. B., Mukherjee, N., Novelo, L. L., Casella, G., et al. (2011). Toward Defining the Autoimmune Microbiome for Type 1 Diabetes. *ISME J.* 5, 82–91. doi:10.1038/ismej.2010.92

Göschl, L., Scheinecker, C., and Bonelli, M. (2019). Treg Cells in Autoimmunity: from Identification to Treg-Based Therapies. *Semin. Immunopathol* 41, 301–314. doi:10.1007/s00281-019-00741-8

Haase, S., Haghikia, A., Wilck, N., Müller, D. N., and Linker, R. A. (2018). Impacts of Microbiome Metabolites on Immune Regulation and Autoimmunity. *Immunology* 154, 230–238. doi:10.1111/imm.12933

Hehemann, J.-H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., and Michel, G. (2010). Transfer of Carbohydrate-Active Enzymes from marine Bacteria to Japanese Gut Microbiota. *Nature* 464, 908–912. doi:10.1038/nature08937

Heyma, P., Harrison, L. C., and Robins-Browne, R. (1986). Thyrotrophin (TSH) Binding Sites on Yersinia Enterocolitica Recognized by Immunoglobulins from Humans with Graves' Disease. *Clin. Exp. Immunol.* 64 (2), 249–254.

Holmes, E., Li, J. V., Athanasiou, T., Ashrafian, H., and Nicholson, J. K. (2011). Understanding the Role of Gut Microbiome-Host Metabolic Signal Disruption in Health and Disease. *Trends Microbiol.* 19, 349–359. doi:10.1016/j.tim.2011.05.006

Hong, S. N., and Rhee, P-L. (2014). Unraveling the Ties between Irritable Bowel Syndrome and Intestinal Microbiota. *Wjg* 20, 2470–2481. doi:10.3748/wjg.v20.i10.2470

Hu, E.-D., Chen, D.-Z., Wu, J.-L., Lu, F.-B., Chen, L., Zheng, M.-H., et al. (2018). High Fiber Dietary and Sodium Butyrate Attenuate Experimental Autoimmune Hepatitis through Regulation of Immune Regulatory Cells and Intestinal Barrier. *Cell Immunol.* 328, 24–32. doi:10.1016/j.cellimm.2018.03.003

Hui, W., Yu, D., Cao, Z., and Zhao, X. (2019). Butyrate Inhibit Collagen-Induced Arthritis via Treg/IL-10/Th17 axis. *Int. Immunopharmacology* 68, 226–233. doi:10.1016/j.intimp.2019.01.018

Indiani, C. M. d. S. P., Rizzardi, K. F., Castelo, P. M., Ferraz, L. F. C., Darrieux, M., and Parisotto, T. M. (2018). Childhood Obesity and Firmicutes/Bacteroidetes Ratio in the Gut Microbiota: A Systematic Review. *Child. Obes.* 14 (8), 501–509. doi:10.1089/chi.2018.0040

Ishaq, H. M., Mohammad, I. S., Shahzad, M., Ma, C., Raza, M. A., Wu, X., et al. (2018). Molecular Alteration Analysis of Human Gut Microbial Composition in Graves' Disease Patients. *Int. J. Biol. Sci.* 14, 1558–1570. doi:10.7150/ijbs.24151

Jeffery, I. B., O'Toole, P. W., Öhman, L., Claesson, M. J., Deane, J., Quigley, E. M. M., et al. (2012l). An Irritable Bowel Syndrome Subtype Defined by Species-specific Alterations in Faecal Microbiota. *Gut* 61 (7), 997–1006. doi:10.1136/gutjnl-2011-301501

Jenq, R. R., Taur, Y., Devlin, S. M., Ponce, D. M., Goldberg, J. D., Ahr, K. F., et al. (2015). Intestinal Blautia Is Associated with Reduced Death from Graft-Versus-Host Disease. *Biol. Blood Marrow Transplant.* 21, 1373–1383. doi:10.1016/j.bbmt.2015.04.016

Ji, D. Y., Park, S. H., Park, S. J., Kim, K. H., Ku, C. R., Shin, D. Y., et al. (2018). Comparative Assessment of Graves' Disease and Main Extrathyroidal Manifestation, Graves' Ophthalmopathy, by Non-targeted Metabolite Profiling of Blood and Orbital Tissue. *Sci. Rep.* 8, 9262. doi:10.1038/s41598-018-27600-0

Jiang, W., Yu, X., Kosik, R. O., Song, Y., Qiao, T., Tong, J., et al. (2021). Gut Microbiota May Play a Significant Role in the Pathogenesis of Graves' Disease. *Thyroid* 31, 810–820. doi:10.1089/thy.2020.0193

Jie, Z., Xia, H., Zhong, S.-L., Feng, Q., Li, S., Liang, S., et al. (2017). The Gut Microbiome in Atherosclerotic Cardiovascular Disease. *Nat. Commun.* 8, 845. doi:10.1038/s41467-017-00900-1

Jubair, W. K., Hendrickson, J. D., Severs, E. L., Schulz, H. M., Adhikari, S., Ir, D., et al. (2018). Modulation of Inflammatory Arthritis in Mice by Gut Microbiota through Mucosal Inflammation and Autoantibody Generation. *Arthritis Rheumatol.* 70, 1220–1233. doi:10.1002/art.40490

Kaeberlein, T., Lewis, K., and Epstein, S. S. (2002). Isolating "Uncultivable" Microorganisms in Pure Culture in a Simulated Natural Environment. *Science* 296, 1127–1129. doi:10.1126/science.1070633

Kang, D.-W., Adams, J. B., Gregory, A. C., Borody, T., Chittick, L., Fasano, A., et al. (2017). Microbiota Transfer Therapy Alters Gut Ecosystem and Improves Gastrointestinal and Autism Symptoms: an Open-Label Study. *Microbiome* 5, 10. doi:10.1186/s40168-016-0225-7

Kassinen, A., Krogius-Kurikka, L., Mäkivuokko, H., Rinttilä, T., Paulin, L., Corander, J., et al. (2007). The Fecal Microbiota of Irritable Bowel Syndrome Patients Differs Significantly from that of Healthy Subjects. *Gastroenterology* 133, 24–33. doi:10.1053/j.gastro.2007.04.005

Kayama, H., Okumura, R., and Takeda, K. (2020). Interaction between the Microbiota, Epithelia, and Immune Cells in the Intestine. *Annu. Rev. Immunol.* 38, 23–48. doi:10.1146/annurev-immunol-070119-115104

Kerckhoffs, A. P., Samsom, M., van der Rest, M. E., de Vogel, J., Knol, J., Ben-Amor, K., et al. (2009). Lower Bifidobacteria Counts in Both Duodenal Mucosa-Associated and Fecal Microbiota in Irritable Bowel Syndrome Patients. *Wjg* 15, 2887. doi:10.3748/wjg.15.2887

Knezevic, J., Starchl, C., Tmava Berisha, A., and Amrein, K. (2020). Thyroid-Gut-Axis: How Does the Microbiota Influence Thyroid Function? *Nutrients* 12, 1769. doi:10.3390/nu12061769

Köhling, H. L., Plummer, S. F., Marchesi, J. R., Davidge, K. S., and Ludgate, M. (2017). The Microbiota and Autoimmunity: Their Role in Thyroid Autoimmune Diseases. *Clin. Immunol.* 183, 63–74. doi:10.1016/j.clim.2017.07.001

Krogius-Kurikka, L., Lyra, A., Malinen, E., Aarnikunnas, J., Tuimala, J., Paulin, L., et al. (2009). Microbial Community Analysis Reveals High Level Phylogenetic Alterations in the Overall Gastrointestinal Microbiota of Diarrhoea-Predominant Irritable Bowel Syndrome Sufferers. *BMC Gastroenterol.* 9, 95. doi:10.1186/1471-230X-9-95

Lagier, J.-C., Dubourg, G., Million, M., Cadoret, F., Bilen, M., Fenollar, F., et al. (2018). Culturing the Human Microbiota and Culturomics. *Nat. Rev. Microbiol.* 16, 540–550. doi:10.1038/s41579-018-0041-0

Lagier, J.-C., Hugon, P., Khelaifia, S., Fournier, P.-E., La Scola, B., and Raoult, D. (2015). The Rebirth of Culture in Microbiology through the Example of Culturomics to Study Human Gut Microbiota. *Clin. Microbiol. Rev.* 28, 237–264. doi:10.1128/CMR.00014-14

Laudadio, I., Fulci, V., Palone, F., Stronati, L., Cucchiara, S., and Carissimi, C. (2018). Quantitative Assessment of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut Microbiome. *OMICS: A J. Integr. Biol.* 22, 248–254. doi:10.1089/omi.2018.0013

Levy, M., Kolodziejczyk, A. A., Thaiss, C. A., and Elinav, E. (2017). Dysbiosis and the Immune System. *Nat. Rev. Immunol.* 17, 219–232. doi:10.1038/nri.2017.7

Li, C., Yuan, J., Zhu, Y.-f., Yang, X.-j., Wang, Q., Xu, J., et al. (2016). Imbalance of Th17/Treg in Different Subtypes of Autoimmune Thyroid Diseases. *Cell Physiol Biochem* 40, 245–252. doi:10.1159/000452541

Li, Q., Wang, B., Mu, K., and Zhang, J. A. (2019). The Pathogenesis of Thyroid Autoimmune Diseases: New T Lymphocytes - Cytokines Circuits beyond the Th1−Th2 Paradigm. *J. Cell Physiol* 234, 2204–2216. doi:10.1002/jcp.27180

Liu, S., An, Y., Cao, B., Sun, R., Ke, J., and Zhao, D. (2020). The Composition of Gut Microbiota in Patients Bearing Hashimoto's Thyroiditis with Euthyroidism and Hypothyroidism. *Int. J. Endocrinol.* 2020, 1–9. doi:10.1155/2020/5036959

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012). Diversity, Stability and Resilience of the Human Gut Microbiota. *Nature* 489, 220–230. doi:10.1038/nature11550

Luu, M., and Visekruna, A. (2019). Short-chain Fatty Acids: Bacterial Messengers Modulating the Immunometabolism of T Cells. *Eur. J. Immunol.* 49, 842–848. doi:10.1002/eji.201848009

Malinen, E., Rinttila, T., Kajander, K., Matto, J., Kassinen, A., Krogius, L., et al. (2005). Analysis of the Fecal Microbiota of Irritable Bowel Syndrome Patients and Healthy Controls with Real-Time PCR. *Am. J. Gastroenterol.* 100 (2), 373–382. doi:10.1111/j.1572-0241.2005.40312.x

Mangiola, F., Ianiro, G., Franceschi, F., Fagiuoli, S., Gasbarrini, G., and Gasbarrini, A. (2016). Gut Microbiota in Autism and Mood Disorders. *Wjg* 22, 361–368. doi:10.3748/wjg.v22.i1.361

Mars, R. A. T., Yang, Y., Ward, T., Houtti, M., Priya, S., Lekatz, H. R., et al. (2020). Longitudinal Multi-Omics Reveals Subset-specific Mechanisms Underlying Irritable Bowel Syndrome. *Cell* 182, 1460–1473. e17. doi:10.1016/j.cell.2020.08.007

Massart, C., Orgiazzi, J., and Maugendre, D. (2001). Clinical Validity of a New Commercial Method for Detection of TSH-Receptor Binding Antibodies in Sera from Patients with Graves' Disease Treated with Antithyroid Drugs. *Clinica Chim. Acta* 304, 39–47. doi:10.1016/s0009-8981(00)00385-5

Memba, R., Duggan, S. N., Ni Chonchubhair, H. M., Griffin, O. M., Bashir, Y., O'Connor, D. B., et al. (2017). The Potential Role of Gut Microbiota in Pancreatic Disease: A Systematic Review. *Pancreatology* 17, 867–874. doi:10.1016/j.pan.2017.09.002

Menconi, F., Marcocci, C., and Marinò, M. (2014). Diagnosis and Classification of Graves' Disease. *Autoimmun. Rev.* 13 (4-5), 398–402. doi:10.1016/j.autrev.2014.01.013

Meng, S., Wu, J. T., Archer, S. Y., and Hodin, R. A. (1999). Short-chain Fatty Acids and Thyroid Hormone Interact in Regulating Enterocyte Gene Transcription. *Surgery* 126, 293–298. doi:10.1016/s0039-6060(99)70168-6

Moshkelgosha, S., Verhasselt, H. L., Verhasselt, H. L., Masetti, G., Covelli, D., Biscarini, F., et al. (2021). Modulating Gut Microbiota in a Mouse Model of Graves' Orbitopathy and its Impact on Induced Disease. *Microbiome* 9 (1), 45. doi:10.1186/s40168-020-00952-4

Mullaney, J. A., Stephens, J. E., Costello, M.-E., Fong, C., Geeling, B. E., Gavin, P. G., et al. (2018). Correction to: Type 1 Diabetes Susceptibility Alleles Are Associated with Distinct Alterations in the Gut Microbiota. *Microbiome* 6, 51. doi:10.1186/s40168-018-0438-z

Nakano, A., Watanabe, M., Iida, T., Kuroda, S., Matsuzuka, F., Miyauchi, A., et al. (2007). Apoptosis-induced Decrease of Intrathyroidal CD4+CD25+ Regulatory T Cells in Autoimmune Thyroid Diseases. *Thyroid* 17, 25–31. doi:10.1089/thy.2006.0231

Needham, B. D., Adame, M. D., Serena, G., Rose, D. R., Preston, G. M., Conrad, M. C., et al. (2021). Plasma and Fecal Metabolite Profiles in Autism Spectrum Disorder. *Biol. Psychiatry* 89, 451–462. doi:10.1016/j.biopsych.2020.09.025

Ni, J., Wu, G. D., Albenberg, L., and Tomov, V. T. (2017). Gut Microbiota and IBD: Causation or Correlation? *Nat. Rev. Gastroenterol. Hepatol.* 14, 573–584. doi:10.1038/nrgastro.2017.88

Nichols, D., Cahoon, N., Trakhtenberg, E. M., Pham, L., Mehta, A., Belanger, A., et al. (2010). Use of Ichip for High-Throughput *In Situ* Cultivation of "Uncultivable" Microbial Species. *Appl. Environ. Microbiol.* 76, 2445–2450. doi:10.1128/AEM.01754-09

Nishino, K., Nishida, A., Inoue, R., Kawada, Y., Ohno, M., Sakai, S., et al. (2018). Analysis of Endoscopic brush Samples Identified Mucosa-Associated Dysbiosis in Inflammatory Bowel Disease. *J. Gastroenterol.* 53, 95–106. doi:10.1007/s00535-017-1384-4

Nyström, H. F., Jansson, S., and Berg, G. (2013). Incidence Rate and Clinical Features of Hyperthyroidism in a Long-Term Iodine Sufficient Area of Sweden (Gothenburg) 2003-2005. *Clin. Endocrinol.* 78, 768–776. doi:10.1111/cen.12060

Peng, D., Xu, B., Wang, Y., Guo, H., and Jiang, Y. (2013). A High Frequency of Circulating Th22 and Th17 Cells in Patients with New Onset Graves' Disease. *PLoS One* 8, e68446. doi:10.1371/journal.pone.0068446

Picchianti-Diamanti, A., Panebianco, C., Salemi, S., Sorgi, M., Di Rosa, R., Tropea, A., et al. (2018). Analysis of Gut Microbiota in Rheumatoid Arthritis Patients: Disease-Related Dysbiosis and Modifications Induced by Etanercept. *Ijms* 19, 2938. doi:10.3390/ijms19102938

Rajilić-Stojanović, M., Biagi, E., Heilig, H. G., Kajander, K., Kekkonen, R. A., Tims, S., et al. (2011). Global and Deep Molecular Analysis of Microbiota Signatures in Fecal Samples from Patients with Irritable Bowel Syndrome. *Gastroenterology* 141 (5), 1792–1801. doi:10.1053/j.gastro.2011.07.043

Relman, D. A. (2012). The Human Microbiome: Ecosystem Resilience and Health. *Nutr. Rev.* 70 (Suppl. 1), S2–S9. doi:10.1111/j.1753-4887.2012.00489.x

Rider, J. E., Hacker, A., Mackintosh, C. A., Pegg, A. E., Woster, P. M., and Casero, R. A. (2007). Spermine and Spermidine Mediate protection against Oxidative Damage Caused by Hydrogen Peroxide. *Amino Acids* 33, 231–240. doi:10.1007/s00726-007-0513-4

Rios-Covian, D., Sánchez, B., Salazar, N., Martínez, N., Redruello, B., Gueimonde, M., et al. (2015). Different Metabolic Features of Bacteroides Fragilis Growing in the Presence of Glucose and Exopolysaccharides of Bifidobacteria. *Front. Microbiol.* 6, 825. doi:10.3389/fmicb.2015.00825

Riva, A., Borgo, F., Lassandro, C., Verduci, E., Morace, G., Borghi, E., et al. (2017). Pediatric Obesity Is Associated with an Altered Gut Microbiota and Discordant Shifts in F Irmicutes Populations. *Environ. Microbiol.* 19, 95–105. doi:10.1111/1462-2920.13463

Rooks, M. G., and Garrett, W. S. (2016). Gut Microbiota, Metabolites and Host Immunity. *Nat. Rev. Immunol.* 16, 341–352. doi:10.1038/nri.2016.42

Saitoh, O., and Nagayama, Y. (2006). Regulation of Graves' Hyperthyroidism with Naturally Occurring CD4+CD25+ Regulatory T Cells in a Mouse Model. *Endocrinology* 147, 2417–2422. doi:10.1210/en.2005-1024

Sato, K., Takahashi, N., Kato, T., Matsuda, Y., Yokoji, M., Yamada, M., et al. (2017). Aggravation of Collagen-Induced Arthritis by Orally Administered Porphyromonas Gingivalis through Modulation of the Gut Microbiota and Gut Immune System. *Sci. Rep.* 7, 6955. doi:10.1038/s41598-017-07196-7

Saulnier, D. M., Riehle, K., Mistretta, T. A., Diaz, M. A., Mandal, D., Raza, S., et al. (2011). Gastrointestinal Microbiome Signatures of Pediatric Patients with Irritable Bowel Syndrome. *Gastroenterology* 141, 1782–1791. doi:10.1053/j.gastro.2011.06.072

Schmidt, T. S. B., Raes, J., and Bork, P. (2018). The Human Gut Microbiome: From Association to Modulation. *Cell* 172, 1198–1215. doi:10.1016/j.cell.2018.02.044

Schwiertz, A., Taras, D., Schäfer, K., Beijer, S., Bos, N. A., Donus, C., et al. (2010). Microbiota and SCFA in Lean and Overweight Healthy Subjects. *Obesity (Silver Spring)* 18 (1), 190–195. doi:10.1038/oby.2009.167

Schwiertz, A., Taras, D., Schäfer, K., Beijer, S., Bos, N. A., Donus, C., et al. (2010). Microbiota and SCFA in Lean and Overweight Healthy Subjects. *Obes. Silver Spring Md.* 18, 190–195. doi:10.1038/oby.2009.167

Sekirov, I., Russell, S. L., Antunes, L. C., and Finlay, B. B. (2010). Gut Microbiota in Health and Disease. *Physiol. Rev.* 90 (3), 859–904. doi:10.1152/physrev.00045.2009

Shakya, M., Lo, C.-C., and Chain, P. S. G. (2019). Advances and Challenges in Metatranscriptomic Analysis. *Front. Genet.* 10, 904. doi:10.3389/fgene.2019.00904

Sharon, G., Cruz, N. J., Kang, D.-W., Gandal, M. J., Wang, B., Kim, Y.-M., et al. (2019). Human Gut Microbiota from Autism Spectrum Disorder Promote Behavioral Symptoms in Mice. *Cell* 177, 1600–1618. e17. doi:10.1016/j.cell.2019.05.004

Shi, N., Li, N., Duan, X., and Niu, H. (2017). Interaction between the Gut Microbiome and Mucosal Immune System. *Mil. Med Res* 4, 14. doi:10.1186/s40779-017-0122-9

Shi, T.-T., Hua, L., Wang, H., and Xin, Z. (2019). The Potential Link between Gut Microbiota and Serum TRAb in Chinese Patients with Severe and Active Graves' Orbitopathy. *Int. J. Endocrinol.* 2019, 1–12. doi:10.1155/2019/9736968

Shi, T.-T., Xin, Z., Hua, L., Wang, H., Zhao, R.-X., Yang, Y.-L., et al. (2021). Comparative Assessment of Gut Microbial Composition and Function in Patients with Graves' Disease and Graves' Orbitopathy. *J. Endocrinol. Invest.* 44, 297–310. doi:10.1007/s40618-020-01298-2

Shi, T.-T., Xin, Z., Hua, L., Zhao, R.-X., Yang, Y.-L., Wang, H., et al. (2019). Alterations in the Intestinal Microbiota of Patients with Severe and Active Graves' Orbitopathy: a Cross-Sectional Study. *J. Endocrinol. Invest.* 42, 967–978. doi:10.1007/s40618-019-1010-9

Sircana, A., De Michieli, F., Parente, R., Framarin, L., Leone, N., Berrutti, M., et al. (2019). Gut Microbiota, Hypertension and Chronic Kidney Disease: Recent Advances. *Pharmacol. Res.* 144, 390–408. doi:10.1016/j.phrs.2018.01.013

Sivaprakasam, S., Prasad, P. D., and Singh, N. (2016). Benefits of Short-Chain Fatty Acids and Their Receptors in Inflammation and Carcinogenesis. *Pharmacol. Ther.* 164, 144–151. doi:10.1016/j.pharmthera.2016.04.007

Smith, T. J., and Hegedüs, L. (2016). Graves' Disease. *N. Engl. J. Med.* 375, 1552–1565. doi:10.1056/NEJMra1510030

Sommer, F., Anderson, J. M., Bharti, R., Raes, J., and Rosenstiel, P. (2017). The Resilience of the Intestinal Microbiota Influences Health and Disease. *Nat. Rev. Microbiol.* 15, 630–638. doi:10.1038/nrmicro.2017.58

Song, J., Shan, Z., Mao, J., and Teng, W. (2019). Serum Polyamine Metabolic Profile in Autoimmune Thyroid Disease Patients. *Clin. Endocrinol.* 90, 727–736. doi:10.1111/cen.13946

Struja, T., Eckart, A., Kutz, A., Huber, A., Neyer, P., Kraenzlin, M., et al. (2018). Metabolomics for Prediction of Relapse in Graves' Disease: Observational Pilot Study. *Front. Endocrinol.* 9, 623. doi:10.3389/fendo.2018.00623

Su, X., Yin, X., Liu, Y., Yan, X., Zhang, S., Wang, X., et al. (2020). Gut Dysbiosis Contributes to the Imbalance of Treg and Th17 Cells in Graves' Disease Patients by Propionic Acid. *J. Clin. Endocrinol. Metab.* 105, dgaa511. doi:10.1210/clinem/dgaa511

Takeuchi, Y., Hirota, K., and Sakaguchi, S. (2020). Impaired T Cell Receptor Signaling and Development of T Cell-Mediated Autoimmune Arthritis. *Immunol. Rev.* 294, 164–176. doi:10.1111/imr.12841

Tan, J., McKenzie, C., Potamitis, M., Thorburn, A. N., Mackay, C. R., and Macia, L. (2014). The Role of Short-Chain Fatty Acids in Health and Disease. *Adv. Immunol.* 121, 91–119. doi:10.1016/B978-0-12-800100-4.00003-9

Teng, F., Felix, K. M., Bradley, C. P., Naskar, D., Ma, H., Raslan, W. A., et al. (2017). The Impact of Age and Gut Microbiota on Th17 and Tfh Cells in K/BxN Autoimmune Arthritis. *Arthritis Res. Ther.* 19, 188. doi:10.1186/s13075-017-1398-6

Turnbaugh, P. J., and Gordon, J. I. (2009). The Core Gut Microbiome, Energy Balance and Obesity. *J. Physiol.* 587, 4153–4158. doi:10.1113/jphysiol.2009.174136

Van Treuren, W., and Dodd, D. (2020). Microbial Contribution to the Human Metabolome: Implications for Health and Disease. *Annu. Rev. Pathol. Mech. Dis.* 15, 345–369. doi:10.1146/annurev-pathol-020117-043559

Vatanen, T., Kostic, A. D., d'Hennezel, E., Siljander, H., Franzosa, E. A., Yassour, M., et al. (2016). Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell* 165, 842–853. doi:10.1016/j.cell.2016.04.007

Venkataraman, A., Sieber, J. R., Schmidt, A. W., Waldron, C., Theis, K. R., and Schmidt, T. M. (2016). Variable Responses of Human Microbiomes to Dietary Supplementation with Resistant Starch. *Microbiome* 4 (1), 33. doi:10.1186/s40168-016-0178-x

Virili, C., Fallahi, P., Antonelli, A., Benvenga, S., and Centanni, M. (2018). Gut Microbiota and Hashimoto's Thyroiditis. *Rev. Endocr. Metab. Disord.* 19, 293–300. doi:10.1007/s11154-018-9467-y

Wang, W.-L., Xu, S.-Y., Ren, Z.-G., Tao, L., Jiang, J.-W., and Zheng, S.-S. (2015). Application of Metagenomics in the Human Gut Microbiome. *Wjg* 21, 803–814. doi:10.3748/wjg.v21.i3.803

Wang, Z., Zhang, Q., Lu, J., Jiang, F., Zhang, H., Gao, L., et al. (2010). Identification of Outer Membrane Porin F Protein ofYersinia enterocoliticaRecognized by Antithyrotopin Receptor Antibodies in Graves' Disease and Determination of its Epitope Using Mass Spectrometry and Bioinformatics Tools. *J. Clin. Endocrinol. Metab.* 95, 4012–4020. doi:10.1210/jc.2009-2184

Weiss, M., Ingbar, S. H., Winblad, S., and Kasper, D. L. (1983). Demonstration of a Saturable Binding Site for Thyrotropin in Yersinia Enterocolitica. *Science* 219, 1331–1333. doi:10.1126/science.6298936

Yan, H.-x., An, W.-c., Chen, F., An, B., Pan, Y., Jin, J., et al. (2020). Intestinal Microbiota Changes in Graves' Disease: a Prospective Clinical Study. *Biosci. Rep.* 40, BSR20191242. doi:10.1042/BSR20191242

Yang, M., Sun, B., Li, J., Yang, B., Xu, J., Zhou, X., et al. (2019). Alteration of the Intestinal flora May Participate in the Development of Graves' Disease: a Study Conducted Among the Han Population in Southwest China. *Endocr. Connect.* 8, 822–828. doi:10.1530/EC-19-0001

Yasuda, K., Kitagawa, Y., Kawakami, R., Isaka, Y., Watanabe, H., Kondoh, G., et al. (2019). Satb1 Regulates the Effector Program of Encephalitogenic Tissue Th17 Cells in Chronic Inflammation. *Nat. Commun.* 10, 549. doi:10.1038/s41467-019-08404-w

Yuan, Q., Zhao, Y., Zhu, X., and Liu, X. (2017). Low Regulatory T Cell and High IL-17 mRNA Expression in a Mouse Graves' Disease Model. *J. Endocrinol. Invest.* 40, 397–407. doi:10.1007/s40618-016-0575-9

Zhao, F., Feng, J., Li, J., Zhao, L., Liu, Y., Chen, H., et al. (2018). Alterations of the Gut Microbiota in Hashimoto's Thyroiditis Patients. *Thyroid* 28, 175–186. doi:10.1089/thy.2017.0395

Zheng, X., Xie, G., Zhao, A., Zhao, L., Yao, C., Chiu, N. H. L., et al. (2011). The Footprints of Gut Microbial-Mammalian Co-metabolism. *J. Proteome Res.* 10, 5512–5522. doi:10.1021/pr2007945

Zhou, L., Li, X., Ahmed, A., Wu, D., Liu, L., Qiu, J., et al. (2014). Gut Microbe Analysis between Hyperthyroid and Healthy Individuals. *Curr. Microbiol.* 69, 675–680. doi:10.1007/s00284-014-0640-6

Zhou, L., Zhang, M., Wang, Y., Dorfman, R. G., Liu, H., Yu, T., et al. (2018). Faecalibacterium Prausnitzii Produces Butyrate to Maintain Th17/Treg Balance and to Ameliorate Colorectal Colitis by Inhibiting Histone Deacetylase 1. *Inflamm. Bowel Dis.* 24, 1926–1940. doi:10.1093/ibd/izy182

Zhu, Q., Hou, Q., Huang, S., Ou, Q., Huo, D., Vázquez-Baeza, Y., et al. (2021). Compositional and Genetic Alterations in Graves' Disease Gut Microbiome Reveal Specific Diagnostic Biomarkers. *ISME J.* 15, 3399–3411. doi:10.1038/s41396-021-01016-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 01 March 2022
doi: 10.3389/fmolb.2022.839051

Assessment of Greenhouse Tomato Anthesis Rate Through Metabolomics Using LASSO Regularized Linear Regression Model

Ratklao Siriwach[1†], Jun Matsuzaki[1†], Takeshi Saito[2], Hiroshi Nishimura[3], Masahide Isozaki[3], Yosuke Isoyama[3], Muneo Sato[1], Masanori Arita[1,4], Shotaro Akaho[5], Tadahisa Higashide[2], Kentaro Yano[6] and Masami Yokota Hirai[1]*

[1]RIKEN Center for Sustainable Resource Science, Yokohama, Japan, [2]Institute of Vegetable and Floriculture Science, NARO, Tsukuba, Japan, [3]Mie Prefecture Agricultural Research Institute, Matsusaka, Japan, [4]National Institute of Genetics, Mishima, Japan, [5]National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan, [6]Bioinformatics Laboratory, Department of Life Sciences, School of Agriculture, Meiji University, Kawasaki, Japan

OPEN ACCESS

Edited by:
Wolfram Weckwerth,
University of Vienna, Austria

Reviewed by:
José Juan Ordaz-Ortiz,
Instituto Politécnico Nacional de México (CINVESTAV), Mexico
Zhongda Zeng,
Dalian University, China

***Correspondence:**
Masami Yokota Hirai
masami.hirai@riken.jp

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

Received: 19 December 2021
Accepted: 03 February 2022
Published: 01 March 2022

Citation:
Siriwach R, Matsuzaki J, Saito T, Nishimura H, Isozaki M, Isoyama Y, Sato M, Arita M, Akaho S, Higashide T, Yano K and Hirai MY (2022) Assessment of Greenhouse Tomato Anthesis Rate Through Metabolomics Using LASSO Regularized Linear Regression Model. Front. Mol. Biosci. 9:839051. doi: 10.3389/fmolb.2022.839051

While the high year-round production of tomatoes has been facilitated by solar greenhouse cultivation, these yields readily fluctuate in response to changing environmental conditions. Mathematic modeling has been applied to forecast phenotypes of tomatoes using environmental measurements (e.g., temperature) as indirect parameters. In this study, metabolome data, as direct parameters reflecting plant internal status, were used to construct a predictive model of the anthesis rate of greenhouse tomatoes. Metabolome data were obtained from tomato leaves and used as variables for linear regression with the least absolute shrinkage and selection operator (LASSO) for prediction. The constructed model accurately predicted the anthesis rate, with an R^2 value of 0.85. Twenty-nine of the 161 metabolites were selected as candidate markers. The selected metabolites were further validated for their association with anthesis rates using the different metabolome datasets. To assess the importance of the selected metabolites in cultivation, the relationships between the metabolites and cultivation conditions were analyzed *via* correspondence analysis. Trigonelline, whose content did not exhibit a diurnal rhythm, displayed major contributions to the cultivation, and is thus a potential metabolic marker for predicting the anthesis rate. This study demonstrates that machine learning can be applied to metabolome data to identify metabolites indicative of agricultural traits.

Keywords: metabolome, metabolites, tomato, anthesis rate, machine learning, LASSO, trigonelline

1 INTRODUCTION

Tomatoes (*Solanum lycopersicum* L.) are produced worldwide, with the highest rates of production among non-grain crops after potatoes (FAOSTAT, 2018). The high year-round production of tomato fruits has been facilitated by greenhouse cultivation in many countries. Greenhouse cultivation provides the optimal environmental conditions, such as temperature, humidity, and light conditions, needed to grow plants (Peet and Welles, 2005). However, in addition to the automatic control of environmental conditions, prompt treatment by tomato growers is necessary to mitigate the effects of extreme weather conditions. For example, extreme heat causes pre-harvest physiological disorders, resulting in fruit cracking and blossom drop in tomato plants. For such extreme heat, temporary equipment and/or manual control is required to lower the temperature in the greenhouse (Liebisch et al., 2009; Saure, 2014). Therefore, for greenhouse cultivation, there is a need to continuously and adequately manage the environmental conditions inside greenhouses. Moreover, the morphological or physiological status of tomato plants can be used to infer subsequent plant growth and outcome (crop harvest). This means that more favorable growth conditions could be investigated and elucidated to enhance plant growth and maximize tomato fruit production. At present, tomato growers empirically control the growth conditions in greenhouses according to extreme weather conditions and plant vigor.

Recently, omics data have been utilized in phenotype prediction and the identification of genes that control traits of interest. Among the omics data, gene expression data have been employed, as gene expression profiles can be easily collected by microarray experiments or sequencing technologies (Yamamoto et al., 2016; Gao et al., 2018; Liabeuf et al., 2018). Yano et al. (2006) introduced an accurate prediction method for phenotypes with comprehensive gene expression profiles using a model on a statistical index and correspondence analysis (CA). In addition to transcriptome analysis, comprehensive metabolite profiles (patterns of metabolite contents across a wide range of experimental conditions) have also become practical with high-throughput mass spectrometry-based technologies. Since metabolites are directly related to phenotypes rather than events of gene expression, phenotype prediction using metabolome data is a promising strategy with which to considerably improve predictability.

There are both direct and indirect approaches to the omics analysis of a target trait. Omics data (e.g., gene expression and/or metabolic profiles) obtained from a given organ represent the genetic and physiological status of the same organ. Therefore, omics data are directly available to identify genes and/or metabolites controlling a given trait in an organ. For example, omics data from the fruit of tomato plants rather than other organs (e.g., leaves) are suitable for the detection of genes and metabolites that play a key role in fruit development. However, the direct approach is unfavorable because for the collection of omics data, fruits need to be removed from the plant. To maximize the quantity of fruit production in the greenhouse, it is better to use vegetative organs, such as, rather of the fruit, for the collection of omics data. If omics data from vegetative organs is able to accurately represent the status of tomato fruit, the indirect approach could also prove to be effective and efficient for the identification of genes and metabolites for a trait, as well as for phenotype prediction.

The metabolic profiling of vegetative organs has been reported to be highly correlated with the quantity of tomato fruit produced. For example, the association between vegetative and reproductive growth of greenhouse tomatoes has been studied for a long time (Khan and Sagar, 1969; Tanaka and Fujita, 1974). The allocation of assimilated carbon between vegetative organs (leaves) and reproductive organs (flowers and fruits) is controlled by genetic and environmental factors, such as light intensity and temperature (Dinar and Rudich, 1985; Heuvelink and Buiskool, 1995). Previous studies have also suggested that the metabolic profiles of vegetative organs, rather than reproductive organs, are attractive and suitable for the construction of a prediction model for fruit yield.

When the metabolic profiles in a vegetative organ are effective in accurately predicting fruit yield, the profiles of a metabolite(s) must be strongly associated with yield. The metabolite(s) allows us to predict not only the yield, but also the traits that are highly correlated with the yield. For example, the effective number of flowers that eventually develop mature fruits is correlated with the yield. This suggests that the effective number of flowers newly generated within a period (e.g., a week) in the greenhouse, referred to as the “anthesis rate” in this study, is an effective index for the prediction of fruit production. In addition, this index has practical and diagnostic advantages for maximizing fruit production. When the predicted anthesis rate is too low for commercial fruit production, the environmental condition can be reconsidered to increase the rate. The improvement enhances the subsequent plant growth and increases the effective number of flowers, then maximizes tomato fruit production.

In this study, we present a statistical model with comprehensive metabolic profiles aimed at maximizing tomato fruit production in greenhouses, wherein the metabolic profiles in leaves were employed to predict the anthesis rate. Because metabolome data is a high-dimensional multivariate data, variable selection is a crucial step to characterize the underlying patterns of these variables and narrow them down to find significant variables. Sparse modeling including the least absolute shrinkage and selection operator (LASSO) model that we applied in this study is widely used in various areas of data-driven science (Rasmussen and Bro, 2012; Rish and Grabarnik, 2014). LASSO model has the ability to perform variable selection by reducing the number of variables. In the LASSO model, significantly contributing variables are weighted with large coefficients, while non-contributing variables are weighted with zero or near-zero coefficients. Consequently, we also identified metabolites that strongly contributed to the prediction of the anthesis rate. To date, the control of the environmental conditions in greenhouses has mainly relied on the experience and knowledge of experts in tomato fruit production. However, the use of machine learning and multivariate analysis with comprehensive metabolic profiles in vegetative organs allows us to not only predict fruit production, but also to adjust the environmental conditions for the enhancement of tomato growth without a need for abundant practical experience. This novel strategy will provide innovative knowledge and skills in greenhouse cultivation for all tomato growers, as well as facilitate the economically efficient production of other crops under greenhouse conditions.

2 MATERIALS AND METHODS

2.1 Plant Materials and Growth Conditions

Tomato plants were grown in greenhouses located in Tsukuba (36°2′4.88″ N, 140°6′2.9″ E) and Matsusaka (34°37′51.7″ N, 136°29′39.5″ E), Japan.

2.1.1 Tsukuba Greenhouse (TK01)

In Tsukuba, in the experiment designated TK01, the seeds of the tomato cultivar Ringyoku (National Agricultural Research Organization, Tsukuba, Japan) and rootstock cultivar Maxifort (*S. lycopersicum* × *S. habrochaites*; De Ruiter Seeds, Bergschenhoek, Netherlands) were sown on 16 May 2016. CF Momotaro York (CFMY) seeds (Takii Seed, Kyoto, Japan) were sown on 23 May 2016. On day 14 after sowing (DAS), Ringyoku scions were grafted onto Maxifort rootstocks. On DAS 28 (13 June 2016), all seedlings were transplanted into rockwool blocks (Delta4, Grodan, Roermond, Netherlands) and placed on rockwool slabs (Grotop expert, Grodan) in a greenhouse with a plant density of 3.3 plants/m^2. Culture liquid with an electrical conductivity (EC) of 3.4 mS/cm (15.8 me/L nitrate, 4.5 me/L P, 9.8 me/L K, 9.3 me/L Ca, 4.6 me/L Mg, 0.07 me/L Fe, 0.103 me/L B, 0.017 me/L Mn, 0.076 me/L Zn, 0.00120 me/L Cu, and 0.00083 me/L Mo) was administered *via* a drip. After 14 days of transplanting, culture liquid with an EC of 2.6 mS/cm was administered. To control the cultivation environment, a ubiquitous environment control system (Fujitsu, Kawasaki, Japan) was used. The greenhouse was ventilated during the day and heated overnight so that the daily mean temperature was maintained at 25°C. A heat pump (Green Package; Nepon, Tokyo, Japan) was operated from 20:00 to 04:00, with a target range of 16–20°C. The daytime relative humidity was controlled at 75% until 30 days after transplanting, and maintained at 70% thereafter. Nineteen days after transplanting, CO$_2$ was added from 05:00 to 07:00 to reach a concentration of 800 ppm. Then, and until 105 days after transplanting (26 September 2016), CO$_2$ was added to a concentration of 400 ppm all day.

2.1.2 Matsusaka Greenhouse (IA04)

In Matsusaka, two sets of experiments (IA04 and IA06) were conducted. In the experiment designated IA04, the seeds of the tomato cultivars CFMY, C5-159 (Sakata Seed Co., Japan), C5-160 (Sakata Seed Co.), and C6-164 (Sakata Seed Co.) were sown on 27 July 2016. The seedlings grafted onto Maxifort rootstocks were transplanted on 1 September 2016. The plant density was set at 2.4 plants/m^2 and then rearranged to be 3.6 plants/m^2 in late January 2017. A rockwool culture system with drip fertigation was used in the greenhouse. The culture liquid was supplied with an EC of 3.0 mS/cm (16 me/L N, 4 me/L P, 8.0 me/L K, 8 me/L Ca, and 4 me/L Mg). The interior air temperature was controlled within the range of 13–27°C. The ideal humidity was 80%, and the CO$_2$ concentration was 800 ppm normally without ventilation and 400 ppm with ventilation during cloudy weather.

2.1.3 Matsusaka Greenhouse (IA06)

In another experiment, designated IA06, the seeds of the tomato cultivars CFMY, Ringyoku, and Managua (RIJK ZWAAN, Netherlands) were sown on 4 October 2016. The seedlings grafted onto Maxifort rootstocks were transplanted on 31 October 2016. The plant density was 2.4 plants/m^2 in the first 3 months and then rearranged to 3.6 plants/m^2. A rockwool culture system with drip fertigation was used in the greenhouse. The culture liquid was supplied with an EC of 3.0 mS/cm (16 me/L N, 4 me/L P, 8.0 me/L K, 8 me/L Ca, and 4 me/L Mg). The environmental conditions were controlled as in experiment IA04.

2.2 Measurement of Anthesis Rates

To measure the anthesis rates, we periodically counted the number of flowers that had not fallen off of each plant. The cumulative numbers of flowers (“cumulative anthesis”) were plotted (see **Section 3** for details). From the cumulative anthesis plot, the anthesis rates were calculated from the gradients of a straight line between two neighboring time-points on the horizontal axis.

2.3 Metabolome Analysis

2.3.1 Sampling of Tomato Leaves

In Tsukuba (TK01), the most basal leaflet of a fully developed and sunlit leaf was sampled for two replications every 2 h continuously for 24 h at one-week intervals for 4 weeks. A total of 192 leaf samples were collected from 16 August 2016 to 6 September 2016 (Ringyoku; *n* = 96, CFMY; *n* = 96). In Matsusaka, the fully developed upper leaves were sampled during 10:00–14:00 on 13 October 2016, and 19 January 2017, for IA04 for three replications, except for C5-160 for two replications (CFMY; *n* = 6, C5-159; *n* = 6, C5-160; *n* = 4, C6-164; *n* = 6) and on 19 January 2017 (6 replications) and 9 March 2017 (8 replicates) for IA06 (Ringyoku; *n* = 14, CFMY; *n* = 14, Managua; *n* = 14). The leaves were collected and flash-frozen in liquid nitrogen.

2.3.2 Widely Targeted Metabolomic Analysis

The frozen leaf samples were freeze-dried and powdered. A small amount of samples (0.5–8.9 mg dry weight) was weighed and 1 ml/10 mg (TK01) or 4 mg (IA04 and IA06) dry weight of extraction solvent [80% (v/v) methanol and 0.1% (v/v) formic acid, with 8.4 nmol/L lidocaine and 210 nmol/L 10-camphorsulfonic acid as internal standards] was added. This mixture was shaken using a Shake Master Neo for 2 min at 1,000 rpm to extract the metabolites. After centrifugation for 1 min at 9,100 × g, the supernatant was diluted with the extraction solvent to obtain 0.4 mg/ml extracts. Next, 25 µL of the extract was dried, dissolved in 250 µL of ultra-pure water, and filtered using Millipore MultiScreenHTS384 well (Merck KGaA, Darmstadt, Germany). A 1-µL aliquot of this filtrate (0.04 mg/ml) was subjected to widely targeted metabolomics using liquid chromatography coupled with a tandem quadrupole mass spectrometer (LC-QqQ-MS) (UPLC coupled with Xevo TQ-S, Waters, Milford, MA, United States) (Sawada et al., 2009; Sawada et al., 2019). The analytical conditions are described in detail in **Supplementary Tables S1–S3**. The metabolome data were deposited in the DROP Met in PRIMe (the Platform for RIKEN Metabolomics) (DM0041, http://prime.psc.riken.jp/archives/data/DropMet/059/).

2.3.3 Measurement of Relative Metabolite Contents

For the Tsukuba data (TK01), the peak areas of 501 target metabolites (including two internal standards) were processed as follows. Values below the detection limit were set to zero. The peak area of each metabolite in a leaf sample was divided by the mean peak area in the extraction solvent control from the same leaf sample to obtain the signal-to-noise ratio. In total, 161 metabolites were detected with signal-to-noise ratios above two in more than half of the leaf samples (**Supplementary Table S3**). The peak area of each metabolite was divided by that of the internal standard (lidocaine or 10-camphorsulfonic acid) to obtain the relative metabolite content.

The peak areas from the Matsusaka data (IA04 and IA06) were processed in the same manner as those from the Tsukuba data (TK01). After calculating the signal-to-noise ratio, the peak area of each metabolite was divided by that of the internal standard (lidocaine or 10-camphorsulfonic acid) to obtain the relative metabolite content.

2.4 Least Absolute Shrinkage and Selection Operator Regularized Linear Regression Model Analysis

LASSO regularization was used to extract essential metabolites to predict an anthesis rate. We constructed a prediction model of the anthesis rate using LASSO regularized linear regression analysis, called the LASSO model, to identify the "predictor metabolites" for the anthesis rate.

2.4.1 Least Absolute Shrinkage and Selection Operator Model to Predict the Anthesis Rate in TK01

A LASSO model using metabolome data from TK01, named "M-model", was constructed. Before training the model, the relative metabolite contents of each metabolite in all leaf samples were normalized to have a mean of zero and a standard deviation of one (that is, standardization). The LASSO model was implemented using sklearn.linear_model.Lasso in the Scikit-learn package (McKinney, 2010; Pedregosa et al., 2011).

The M-model was constructed by training the metabolic profiles of 161 metabolites from 192 leaf samples. The linear regression is expressed as:

$$y_i = w_0 + w_1 X_{i1} + \ldots + w_m X_{im}, \quad i \in [1, n], \tag{1}$$

where y_i is the anthesis rate of the plant with the ith leaf samples ($1 \le i \le n$, $n = 192$), X_{ij} is the relative metabolite content of the jth metabolite in the ith sample ($1 \le j \le m$, $m = 161$), w_j is the model coefficient of the jth metabolite ($1 \le j \le m$), and w_0 is an intercept term. Here, y_i and X_{ij} are elements of a vector $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ and an $n \times m$ matrix X, respectively. The linear regression was trained with L1 regularization to perform both feature selection and regularization. The objective function to minimize is:

$$\min_{w} \frac{1}{2n} \left\|Xw - y\right\|_2^2 + \alpha \left\|w\right\|_1, \tag{2}$$

where $||Xw - y||_2^2 = \sum_{i=1}^{n} (X_i w - y_i)^2$ is the sum of the squared errors, $||w||_1 = \sum_{j=1}^{m} |w_j|$ is the L1-norm of the coefficient vector, and $\alpha \ge 0$ is the penalty constant (Tibshirani, 1996). Thus, in the M-model, significantly contributing metabolites, called the selected metabolites, were weighted with large coefficients (either positive or negative), while non-contributing metabolites were weighted with zero coefficients. R^2 value of the M-model was calculated. The prediction accuracy was assessed by 10-fold cross-validation. R^2 value and the mean squared error (MSE) were used as accuracy metrics.

In addition, the second and third LASSO model training with environmental data (E-model) and combined metabolome and environmental data (C-model), respectively, were constructed in the same manner as the M-model. In the E-model, the X matrix contained only environmental factor data (solar irradiance, ambient temperature, relative humidity, and CO_2 concentration). The X matrix in the C-model consisted of the metabolic profiles of 161 metabolites and environmental factor data.

2.4.2 Least Absolute Shrinkage and Selection Operator Model for the Assessment of the Prediction Accuracy of the Predictor Metabolites

We also used the LASSO model to assess the ability and strength of the predictor metabolites in the M-model by expanding the metabolome data from different experimental designs. The predictor metabolites selected from the M-model were used to reconstruct the LASSO model with additional leaf samples from IA04 and IA06. The model was reconstructed in the same manner as the M-model by training the metabolic profiles of the predictor metabolites of 256 leaf samples from three greenhouses (TK01, IA04, and IA06).

2.5 Classification of Leaf Samples by Principal Component Analysis

The differences in leaf samples were evaluated by the PCA of their metabolic profiles. The relative metabolite content of each metabolite in all leaf samples was standardized. The PCA tool in the Scikit-learn package was used. The first two principal components of each leaf sample were used to project the leaf samples into a two-dimensional space. PCA was performed with two datasets, TK01 and a combined data of TK01, IA04, and IA06. For the PCA of TK01, the metabolic profiles of 161 metabolites from 192 leaf samples were used. For the PCA of data combined from TK01, IA04 and IA06, the metabolic profile of the predictor metabolites of 256 leaf samples from the three greenhouses (TK01, IA04, and IA06) were used.

2.6 Hierarchical Clustering Analysis of the Predictor Metabolites

To evaluate the similarities among the predictor metabolites, the metabolic profiles of 256 leaf samples from the three greenhouses (TK01, IA04, and IA06) were used for HCL. The Pearson correlation coefficient (r) of the relative metabolite contents for each pair of metabolites was calculated (**Supplementary Figure S3** and **Supplementary Table S4**). Then, the distances between metabolites, namely, the "correlation distance" (1–r), were employed for agglomerative clustering. Linkage methods were applied to compute the distances between sub-clusters; then, a dendrogram was generated to mine metabolites showing similar profiles. The optimum linkage method was determined based on the cophenetic correlation coefficient. The best linkage method, which yielded the maximum cophenetic correlation coefficient, was used to create a hierarchical dendrogram (Jones et al., 2001). HCL was implemented using the Python library Scipy.

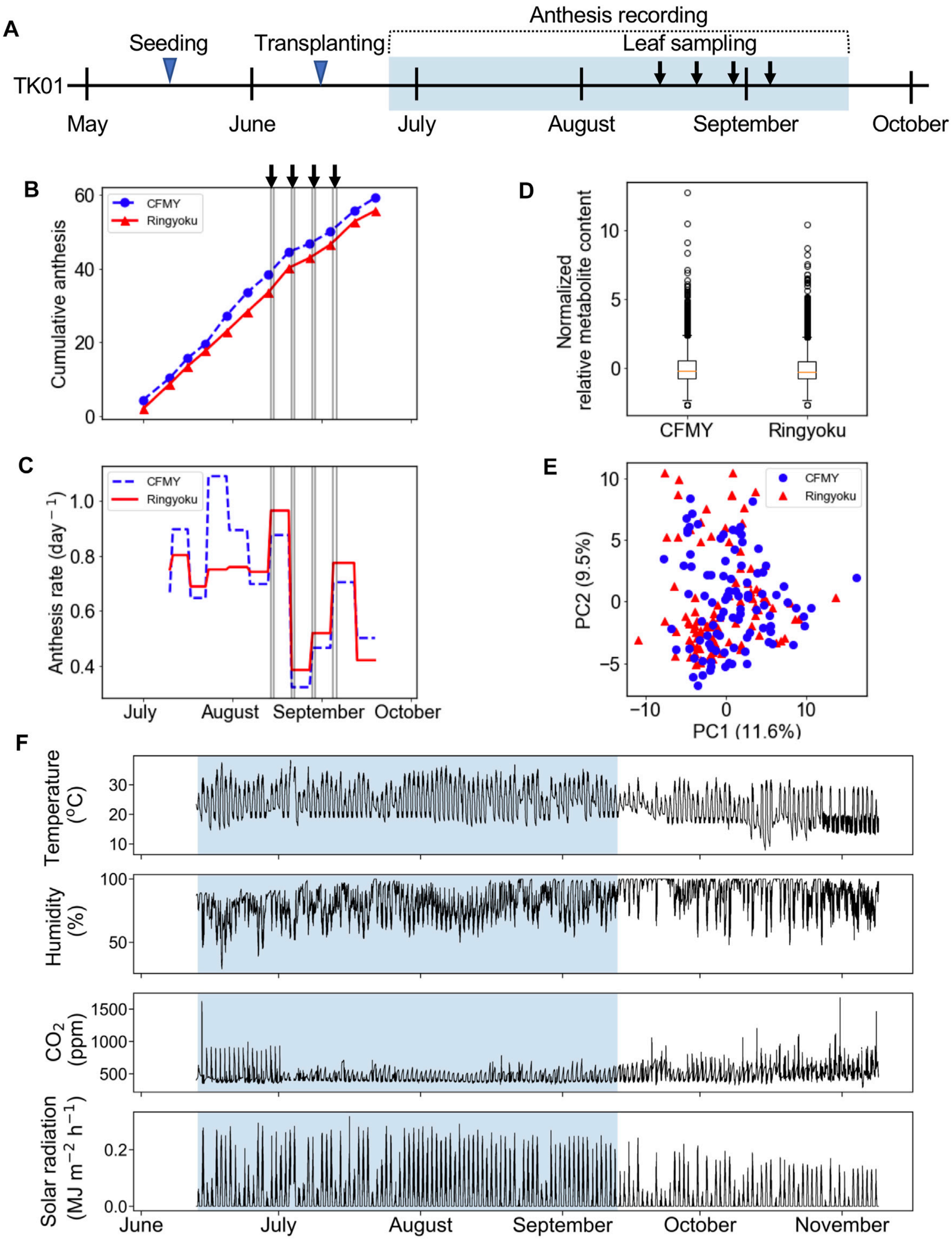


FIGURE 1 | Experimental design of TK01. **(A)** Experimental timeline for leaf sampling and observation of anthesis of cultivars CFMY and Ringyoku. The blue rectangle indicates the period of the measurement of environmental factors (e.g., temperature). **(B)** Cumulative anthesis. The arrows and gray vertical lines indicate the dates of leaf sampling for metabolome analysis. **(C)** Distributions of anthesis rates were statistically the same between cultivars (Mann-Whitney U test, $p > 0.05$, CFMY; $n = 21$, Ringyoku; $n = 21$). **(D)** Box plot of the standardized relative metabolite contents of 161 metabolites in 192 leaf samples (CFMY; $n = 96$, Ringyoku; $n = 96$). **(E)** PCA score plot of the first two components (PC1 and PC2) of leaf samples (CFMY; $n = 96$, Ringyoku; $n = 96$). The metabolic profiles of the 161 metabolites were used for PCA. The numbers in parentheses in the axes are contribution ratios. **(F)** Environmental conditions measured in the experimental timeline. The environmental data in the blue background color used for LASSO analysis (E-model and C-model). The period in the blue background color is consistent with the period for leaf sampling.

2.7 Network Analysis of the Predictor Metabolites With Correspondence Analysis

CA is a multivariate technique and is conceptually similar to PCA. In previous studies, CA has been used to clarify the associations between genes and experimental conditions in microarray analyses (Yano et al., 2006; de Tayrac et al., 2009). We employed CA for network analysis to discover the associations between the predictor metabolites and the associations between the predictor metabolites and the leaf sample characteristics, that is, experimental designs, cultivars, and sampling times.

CA was executed against metabolic profiles. The metabolome data were arranged in a matrix where the columns and rows correspond to the predictor metabolites selected by the M-model and 256 leaf samples from the three experimental designs, respectively. The relative metabolite contents of each metabolite in all leaf samples were standardized, and the minimum value was subtracted to prevent negative values. CA was performed using the FactoMineR library in R (Lê et al., 2008). Coordinates with m-1 dimensions were assigned to each metabolite and leaf sample, where m is the number of predictor metabolites. The coordinate values of all dimensions were retrieved (**Supplementary Table S5**).

2.7.1 Network Analysis Between the Predictor Metabolites and the Leaf Sample Characteristics

The Euclidean distances for each pair of a metabolite and leaf sample were calculated using coordinates in all dimensions from CA. Theoretically, a smaller Euclidean distance indicates a higher association. Based on the histograms of the Euclidean distance (**Supplementary Figure S4A**), the 15th percentile of all distances was set as a threshold value to define a significant association. Pairs of a metabolite and leaf sample with distances less than the threshold were selected (**Supplementary Table S6**). The mean of the distances between each metabolite and each leaf sample characteristics were integrated to construct metabolic networks. Networks were constructed using py2cytoscape and NetworkX libraries in Python, and Cytoscape software (version 3.6.1) (Shannon et al., 2003; Hagberg et al., 2008; Ono et al., 2015). The associations between the metabolites were also evaluated in the same manner.

2.7.2 Network Analysis Among the Predictor Metabolites

CA was used to determine the association among the predictor metabolites. The same process was performed to obtain pairwise Euclidean distances between the metabolites (**Supplementary Tables S7, S8**). The distances that passed the threshold were integrated to construct the metabolite networks.

Statistical Analysis for the Anthesis Rates

In TK01, the significance of the anthesis rates between the cultivars was analyzed using the Mann-Whitney U test. The significance of the anthesis rates among the experimental designs (TK01, IA04, and IA06) was analyzed using the Kruskal–Wallis test with Conover's multiple comparison test. Scipy in Python was used for the statistical analyses.

3 RESULTS

3.1 Data Collection for Anthesis Rate, Leaf Metabolome, and Environmental Factors

In the experiment designated TK01, two tomato cultivars, Ringyoku and CFMY, were grown in Tsukuba, Japan. After transplanting the tomatoes into a greenhouse, the cumulative number of anthesis occurrences was recorded in parallel with leaflet sampling (**Figure 1A**). The cumulative number of anthesis occurrences was used to calculate anthesis rates (**Figures 1B,C**, respectively). The anthesis rates of the Ringyoku and CFMY cultivars were similar and gradually decreased over the growing period. No significant differences were observed between cultivars. During the growing period, fully developed basal and sunlit leaves were collected from plants. Leaf sampling every 2 h for 24 h was conducted four times at one-week intervals. The sampled leaves were subjected to a widely targeted metabolome analysis using a liquid chromatography-mass spectrometer. From a total of 499 targeted metabolites, 161 metabolites above the signal-to-noise ratio threshold were selected (**Supplementary Table S3**). The relative metabolite contents of each metabolite in all leaf samples were standardized prior to further analysis. The boxplot (**Figure 1D**) and PCA score plot (**Figure 1E**) indicated that Ringyoku and CFMY had similar metabolic profiles. Thus, we pooled the metabolic profile data obtained from the two cultivars (192 leaf samples × 161 metabolites) for further analysis. In addition, environmental data (solar irradiance, ambient temperature, relative humidity, and CO_2 concentration) were also obtained (**Figure 1F**).

3.2 Least Absolute Shrinkage and Selection Operator Model for Anthesis Rate Prediction in TK01

We constructed three models (M-model, E-model, and C-model) to predict the anthesis rates in TK01. The model was trained and optimized to obtain predictor metabolites.

For the construction of the M-model, the metabolic profiles of 161 metabolites in 192 leaf samples were employed. During model training, we optimized the model by assigning a range of the penalty constant (α) and then measuring the prediction accuracy by cross-validation. The penalty constant (α) of the M-model was fine-tuned to optimize the best prediction model with the selected metabolites. The iteration training was performed by varied α from 5×10^{-5} to 0.5 (**Supplementary Figure S1A**). At each given α, different sets of metabolites with optimized LASSO coefficients (w) were selected (**Supplementary Figure S1A**). In each loop of a given α, the R^2 value of the M-model was calculated, and the prediction accuracy of the M-model was assessed by 10-fold cross-validation. The R^2 value and the mean squared error (MSE) of the 10-fold cross-validation were also calculated (**Supplementary Figure S1B**). The R^2 values of the training and cross-validation were used to determine an optimum M-model that contained the selected metabolites as the predictor metabolites for the anthesis rate (**Figure 2A**).

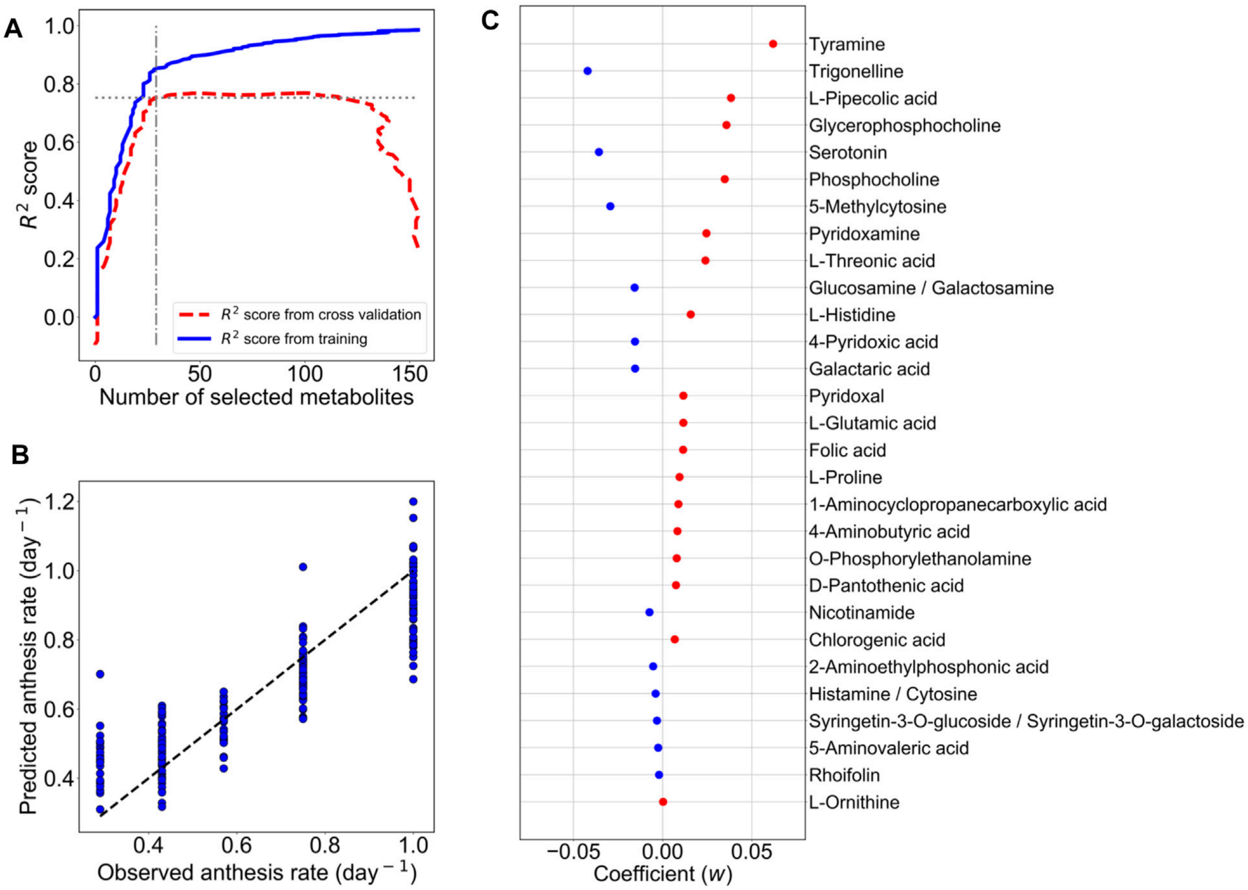


FIGURE 2 | LASSO model with ten-fold cross-validation for the prediction of the anthesis rate in TK01. For LASSO regression analysis, the metabolic profiles of 161 metabolites in 192 leaf samples were employed. **(A)** The numbers of metabolites used for predictor variables versus R^2 value. The elbow point suggests the optimum set of metabolites for the prediction model. **(B)** Comparison of anthesis rates between observed and predicted values. Predicted values were obtained from the M-model with 29 selected metabolites. The dotted line represents the agreement between the observed and predicted values. **(C)** Coefficients (w) of 29 metabolites selected by the M-model. Red dots are positive coefficients, while blue dots are negative coefficients.

From model optimization, increasing the number of metabolites in the model increases the predictive accuracy (R^2 values) in both training and cross-validation. Until cross-validation R^2 stopped improving while model R^2 continued to increase, this indicates overfitting in a high number of metabolites. Thus, we selected α, where the cross-validation R^2 started to plateau and was closest to training R^2 as our optimal model. In **Figure 2A**, the optimum number of metabolites was determined to be 29 at the elbow point on the graph that yielded the closest R^2 values between the training and cross-validation. Using the contributions of these 29 metabolites (**Figure 2B**) as predictor metabolites, we constructed a prediction model for TK01 (M-model). The M-model provided good prediction performance for the anthesis rates (**Figure 2C**). The R^2 value of the M model, R^2 s value, and MSE of the 10-fold cross-validation are summarized in **Table 1**.

To examine the possibility of including environmental factors in the prediction model, we also attempted to construct a LASSO model, the E-model, using four environmental parameters (interior air temperature, interior relative humidity, interior CO_2 concentration, and cumulative solar irradiance) recorded at 5-min intervals (**Figure 1F**). The prediction performance of the environmental parameters was poor (**Table 1** and **Supplementary Figure S2A**). Finally, the C-model model was constructed using a combination of metabolites and environmental factors. The combination slightly improved the prediction accuracy of the anthesis rate (**Table 1** and **Supplementary Figure S2B**).

3.3 Assessment of the Accuracy of Anthesis Rate Prediction Using the Predictor Metabolites

To assess the prediction accuracy of the anthesis rates by the contents of the 29 selected metabolites as the predictor metabolites from the M-model, datasets from two greenhouses (IA04 and IA06) were used.

3.3.1 Differences in Metabolic Profiles Among Experimental Designs

In IA04 and IA06, the experimental designs were conducted at a different greenhouse location (Matsusaka) from TK01 (Tsukuba).

TABLE 1 | The prediction accuracies of the three models in TK01.

Variable used for LASSO model construction	R^2 value (LASSO model)	Cross-validation	
		R^2 value	MSE
The M-model with metabolic profiles of 29 metabolites	0.85	0.75	0.013
The E-model (only environmental factors)	0.11	0.10	0.055
The C-model with metabolic profiles of 36 metabolites and environmental factors	0.89	0.83	0.010

In addition, these three experiments were performed in different growth seasons. Moreover, in addition to Ringyoku and CFMY, four additional cultivars were also used in IA04 and IA06 (**section 2.1**). During the recording of the cumulative numbers of anthesis occurrences, the leaflets were sampled for metabolome analysis at one time point around noon on 2 days (**Figure 3A**). Therefore, metabolic profiles must be varied by differences in the experimental designs. The relative metabolite contents of the 29 predictor metabolites on TK01, IA04, and IA06 is shown in a boxplot in **Figure 3B**. The distribution of the relative metabolite contents in TK01 was relatively compact, while the IA04 and IA06 data exhibited relatively larger variances. This was caused by the mixed effects of different cultivars, greenhouse conditions, and seasons. In addition, PCA for the relative metabolite contents of the 29 predictor metabolites and all leaf samples (*n* = 256) from the three greenhouses were performed to investigate the differences among the experimental designs. The TK01 leaf samples were noticeably separable from the IA04 and IA06 leaf samples, while the IA04 and IA06 leaf samples were clustered together (**Figure 3C**). In addition to the metabolic profiles, the anthesis rates differed among the three experimental designs (**Figure 3D**). The anthesis rate in IA04 was slightly higher than that in TK01, while IA06 showed the highest anthesis rate among the three experimental designs. The differences in the metabolic profiles and anthesis rate of TK01 and the two experimental designs (IA04 and IA06) made it difficult to obtain a good prediction by imputing data from IA04 and IA06 into the M-model.

3.3.2 Least Absolute Shrinkage and Selection Operator Model to Assess the Prediction Accuracy of the Predictor Metabolites

We evaluated the predictive ability of 29 predictor metabolites selected from the M-model. If the predictor metabolites are biologically associated with the anthesis rate, broaden number of leaf samples from different experimental designs should provide a good prediction model. To clarify whether a more universal model could be established, the relative metabolite content of the 29 predictor metabolites and the anthesis rates obtained in TK01, IA04, and IA06 were combined and subjected to the LASSO model. A total of 13 out of the 29 metabolites that yielded the minimum MSE were selected (R^2 = 0.75) (**Figure 3E**). The 10-fold cross-validation results demonstrated the acceptable fitting and prediction accuracy of the model (MSE = 0.26). The model showed good prediction performance across the three datasets (cross-validated R^2 = 0.69) (**Figure 3F**). This result indicates that the predictor metabolites selected by the LASSO model as contributing variables in a specific dataset (TK01) could be effective for the prediction of the anthesis rate in general.

Among the two sets of metabolites selected from the M-model and this combined data model, the LASSO coefficients of the selected metabolites showed that tyramine, trigonelline, glycerophosphocholine, and L-threonic acid had a high association with the anthesis rate in both models.

3.4 Candidate Metabolites Associated With the Anthesis Rate

Metabolites showing significant associations with anthesis rates are attractive candidates for markers of reproductive traits, including anthesis rates, fruit development, and production. We detected candidate metabolites related to anthesis rates by LASSO analysis (**Section 3.3**). To understand the biological characteristics of the 29 predictor metabolites and identify candidate metabolites for future use as prediction markers, we investigated the association between the 29 selected metabolites and anthesis rates using hierarchical clustering analysis (HCL) and correspondence analysis (CA).

First, HCL was used to visualize the metabolic profiles of TK01, IA04, and IA06. Pearson correlation coefficients (*r*) between each pair of the 29 predictor metabolites were obtained to evaluate the similarity in the profiles (**Supplementary Figure S3**). Strong correlations were observed, particularly in the top selected metabolites, such as tyramine, trigoneline, glycerophosphocholine, and serotonin, of the M-model (**Figure 4A**). This result suggests that each of these metabolites plays a similar and important role in anthesis rate estimation. It indicates that it is possible to choose only a small number of metabolites as key predictors of anthesis rates.

Next, CA was conducted for network analysis to elucidate the associations among the 29 predictor metabolites. In the metabolic network (**Figure 4B**), all of the connected metabolites were amines, except for chlorogenic acid, rhoifolin, and L-threonic acid. Thus, the nitrogen-containing metabolites showed similar accumulation patterns across the leaf samples (**Figure 4B**). Among all metabolite-to-metabolite edges, trigonelline has the most edges linked to other metabolites, indicating that trigonelline is a major coexisting metabolite with others.

CA was also conducted for network analysis to elucidate the associations between the 29 predictor metabolites and leaf sample characteristics, that is, experimental designs, cultivars, and sampling times. Among the leaf sample characteristics, the experimental design was the only factor displaying a clear separation in PCA (**Figure 3C**), whereas the cultivars and sampling times did not show distinct separation

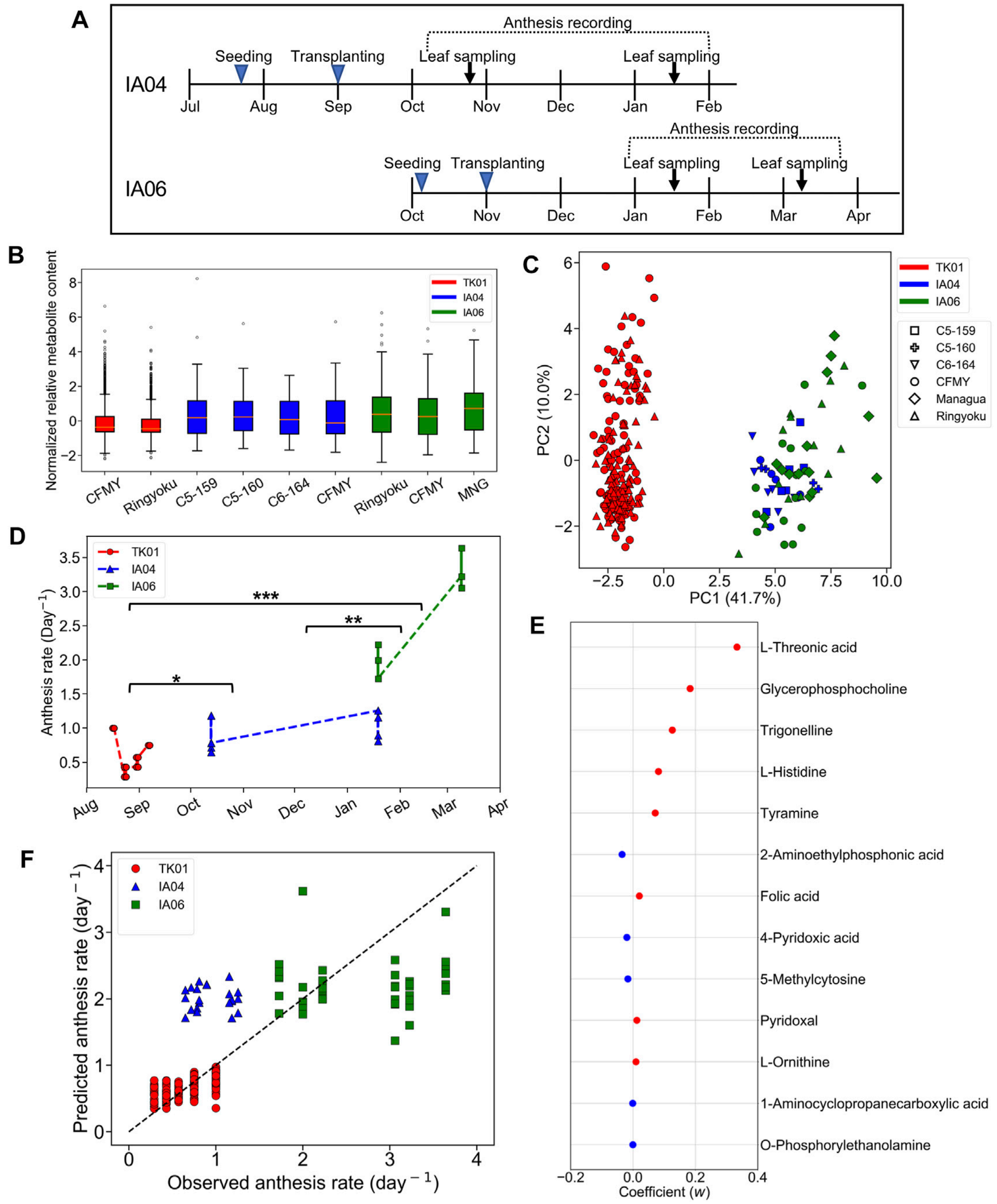


FIGURE 3 | Predictability assessment of the 29 predictor metabolites with expanding metabolome datasets. **(A)** Experimental timeline for leaf sampling and anthesis measurements in IA04 (cultivars: CFMY, C5-159, and C5-16) and IA06 (cultivars: CFMY, Ringyoku, and MNG). **(B)** Box plot of standardized relative metabolite contents of the 29 predictor metabolites in each cultivar in three experimental designs (TK01, IA04, and IA06). The numbers of leaf samples (*n*): CFMY (*n* = 96) and Ringyoku (*n* = 96) in TK01, CFMY (*n* = 6), C5-159 (*n* = 6), C6-164 (*n* = 6), and C5-160 (*n* = 4) in IA04, and Ringyoku (*n* = 14), CFMY (*n* = 14), and Managua (*n* = 14) in IA06. **(C)** PCA score plot of leaf samples (*n* = 256) by using metabolic profiles of the 29 predictor metabolites from three experimental designs (TK01, IA04 and IA06). The contribution ratio is shown in parentheses for the first and second principal component (axis). The colors indicate the experimental designs, and the markers represent the cultivars. **(D)** Anthesis rates used for the LASSO model (TK01, *n* = 16; IA04, *n* = 8; IA06, *n* = 6). Asterisks indicate significant differences according to the Kruskal-Wallis test with Conover's multiple comparison test (*, $p < 0.05$; **, $p < 0.01$; and ***, $p < 0.001$). **(E)** Model coefficients (*w*) of 13 metabolites selected in the LASSO model construction with metabolome datasets from three experimental designs (TK01, IA04 and IA06). The red dots are positive coefficients, while the blue dots are negative coefficients. **(F)** Comparison of anthesis rates between observed and predicted values obtained from the model constructed by the three datasets. The dotted line represents the agreement between the observed and predicted values.

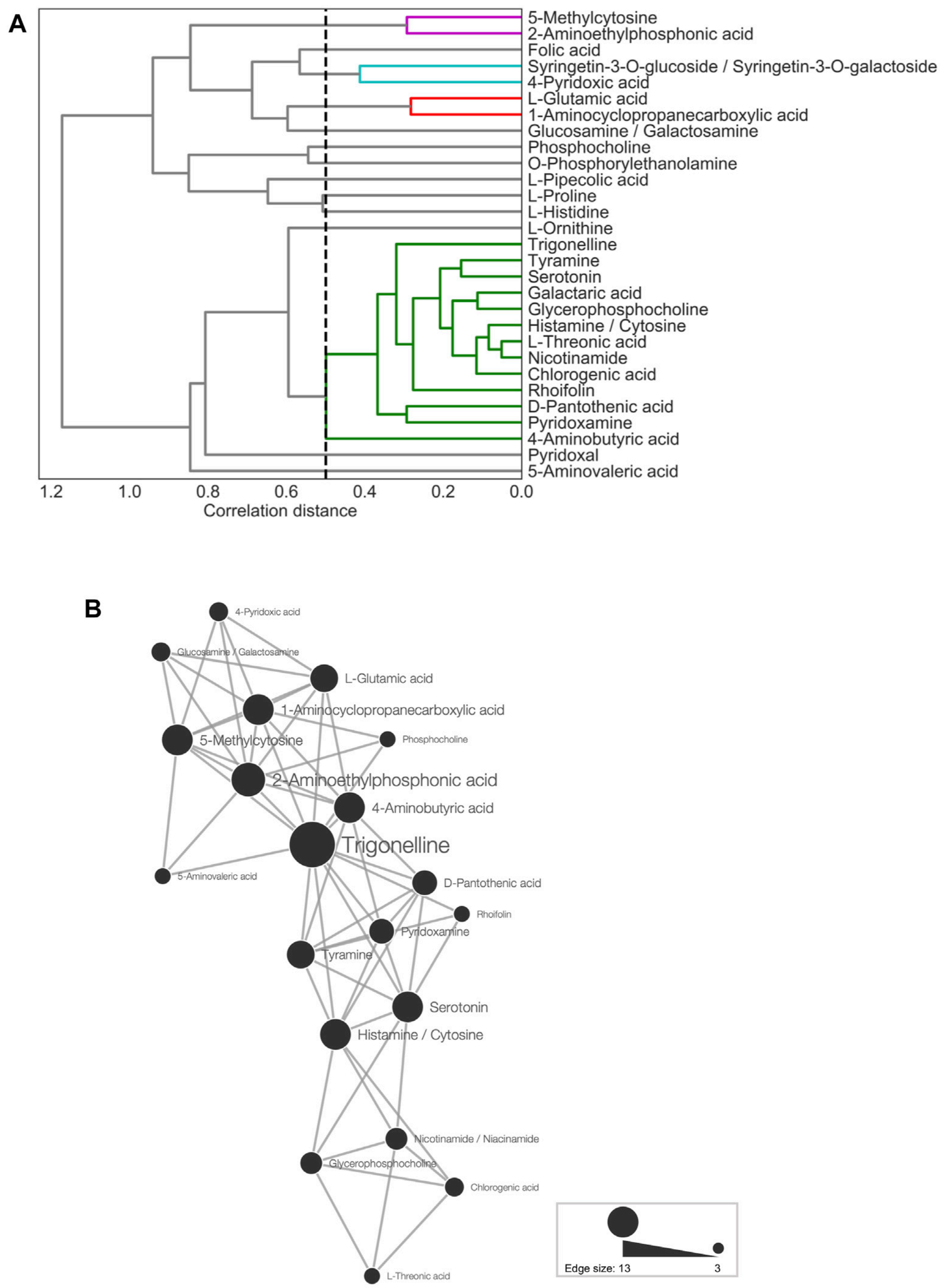


FIGURE 4 | Metabolite association of 29 predictor metabolites. **(A)** Dendrogram representing agglomerative clustering of the correlation distances of the 29 selected metabolites in average linkage. The cluster threshold was 0.5, as indicated by the black dotted line. Lines of the same color represent the same clusters. **(B)** Network for metabolites (threshold: ≤15th percentile of Euclidean distances). The node size represents the number of edges linked to other metabolites.

(**Supplementary Figures S4B,C**). Thus, in CA, we first examined the network between the predictor metabolites and the experimental designs (TK01, IA04, and IA06). In the network (**Figure 5A**), IA04 and IA06 shared seven similarly dominant metabolites. Four out of the seven metabolites, glycerophosphocholine, serotonin, trigonelline, and tyramine, were in the top five of the 29 predictor metabolites (**Figure 2C**). TK01 had nine highly associated metabolites. Among them, one metabolite, trigonelline, was linked to all three experimental designs in the network (**Figure 5A**). Next, we examined the association between metabolites and cultivars. In the metabolite to cultivar network (**Figure 5B**), a network

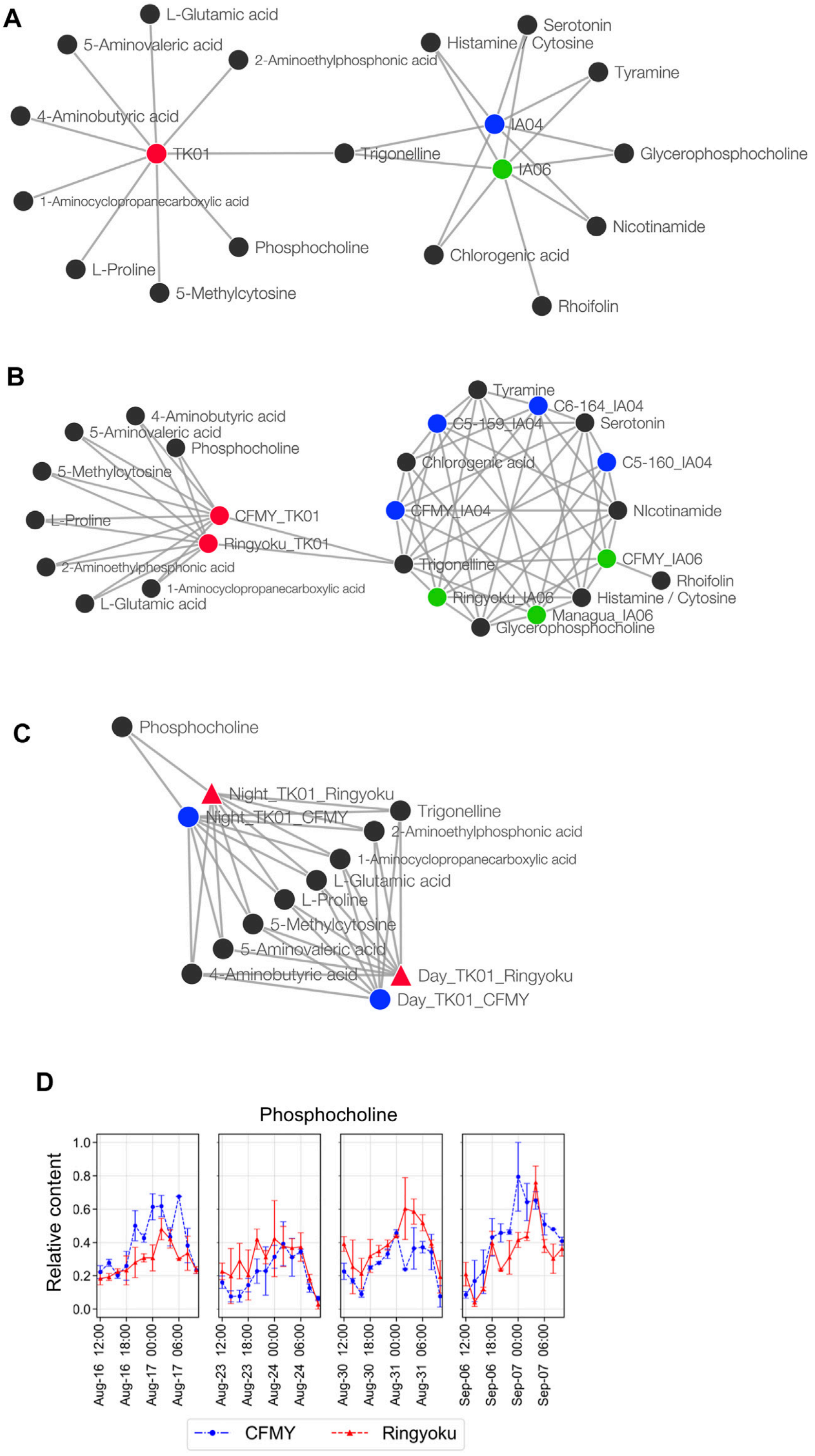


FIGURE 5 | Metabolite association with experimental designs, cultivars, and sampling times. **(A)** Network of metabolites and growth conditions. **(B)** Network of metabolites and cultivars. The cultivars were divided into subcategories of experimental design; for example, the CFMY samples were divided into three and labeled CFMY_TK01, CFMY_IA04, and CFMY_IA06. **(C)** Network of metabolites and sampling times. **(D)** Diurnal changes of the relative content of phosphocholine (scaled between 0 and 1).

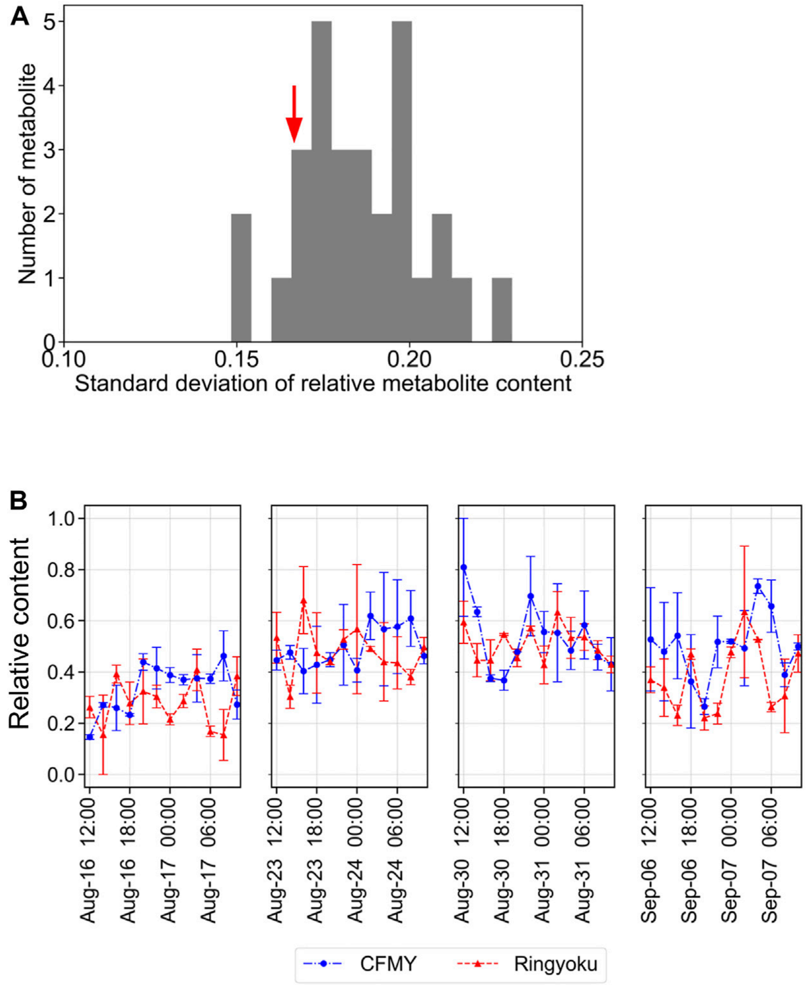


FIGURE 6 | Diurnal fluctuations of metabolite content in tomato leaves. **(A)** Distribution of the standard deviations of 29 metabolites. The red arrow indicates the standard deviation of trigonelline at 0.167. **(B)** Diurnal fluctuations of the relative content of trigonelline (scaled between 0 and 1).

pattern similar to the experimental design was observed. The cultivars IA04 and IA06 shared highly associated metabolites but did not share with the cultivars in TK01, except trigonelline, which was associated with all cultivars (**Figure 5B**).

3.5 Candidates of Stable Metabolites for the Prediction of the Anthesis Rate

Taking into account the leaf sampling time, metabolite content generally changes according to the circadian rhythm. For future use as key indicators of anthesis rate, metabolites whose contents do not change depending on the leaf sampling time are preferred. Because the leaf samples from TK01 were collected every 2 h for a day in time-series format, we constructed a Euclidean distance network of TK01 samples to identify the metabolite associated with leaf sampling time, namely day (06:00–18:00) or night (20:00–04:00) (**Figure 5C**). Among the nine metabolites strongly associated with TK01, phosphocholine was highly associated only at night. This result is consistent with the accumulation pattern of phosphocholine, which showed a diurnal bell-shaped pattern peaking at night (**Figure 5D**). Eight other metabolites, including trigonelline, shared associations during both day and night, indicating high metabolite production, which may produce stable production throughout the day.

To further evaluate the diurnal fluctuations of the 29 LASSO-selected metabolites in TK01, the relative contents of each metabolite were scaled between 0 and 1. The distribution of the standard deviations (SD) of the 29 metabolites is shown in **Figure 6A**. The standard deviations of the metabolite contents ranged from 0.148 to 0.230. Among these, the standard deviation of trigonelline was relatively small (SD = 0.167). In addition, the trigonelline content was relatively stable over the course of a day (**Figure 6B**) compared to that of the other metabolites, such as phosphocholine, glycerophosphocholine, L-glutamic acid, and 4-aminobutyric acid, which exhibited strong diurnal fluctuations (**Supplementary Figure S5**).

Taken together, our results suggest that trigonelline is an attractive metabolite for use as a marker of the anthesis rate of tomatoes. Trigonelline was one of the top five LASSO-selected metabolites for the prediction of the anthesis rate (**Figures 2C**, **3E**), showed no diurnal changes, and exhibited stable content among the different cultivation conditions and varieties (**Figures 6A,B**). Other metabolites among the top five, such as tyramine, were also available not only for the prediction of the anthesis rate, but also as markers under specific cultivation conditions.

4 DISCUSSION

Machine learning approaches have the potential to provide prediction models for agricultural traits and effectively identify metabolites, genes, or environmental factors associated with these traits (Menéndez et al., 2011; Acharjee, 2013; Das et al., 2018; Du et al., 2019; Sawada et al., 2019). Our study employed LASSO regularized linear regression model analysis to construct a prediction model of the anthesis rate using leaf metabolome data as predictor variables and identify the 29 predictor metabolites as candidate biomarkers. Importantly, we identified trigonelline as a key metabolite for anthesis rate prediction using the LASSO models and CA. Moreover, because the trigonelline content in the leaf was relatively stable over the course of a day, it was identified as an attractive biomarker of anthesis rate.

4.1 Possible Uses of Least Absolute Shrinkage and Selection Operator-Selected Metabolites as Biomarkers

The prediction of reproduction and fruit development in tomato is a powerful tool for the diagnosis of plants and the optimal management of the environmental conditions to maximize plant yields. Since anthesis is directly linked to tomato fruit production, it is a good index with which to evaluate tomato cultivation. The identification of metabolites involved in anthesis can be employed as metabolite markers for the prediction of anthesis and yield.

In the construction of the models using LASSO, unimportant metabolites were penalized by L1 regularization, leaving more prominent metabolites after variable selection. A reduction in the number of metabolites is desirable, because a smaller number of metabolites can be more easily measured for future use as biomarkers. As a result, 29 metabolites, including both primary and specialized (secondary) metabolites, were selected from among 161 metabolites. Most of the 29 selected metabolites were nitrogen-containing compounds, such as amino acids and their derivatives, alkaloids, amines, and phospholipids. The LASSO-selected metabolites could indicate the nitrogen status associated with the anthesis rate in tomatoes.

Among the 29 metabolites, trigonelline (*N*-methylnicotinate), a quaternary ammonium, exhibited a metabolic profile similar to that of the majority of the selected metabolites. (**Figure 4B**). In addition, trigonelline demonstrated the greatest association with all three growth conditions and all cultivars, while other metabolites were associated with only leaf samples from particular experiments (**Figures 5A–C**). Moreover, compared to other metabolites, trigonelline showed a relatively stable accumulation over the course of a day (**Figures 5D**, **6B** and **Supplementary Figure S5**). Among 29 metabolites associated with anthesis rate, trigonelline was shown to be a key metabolite related to anthesis rate. These results support that trigonelline is a suitable biomarker without diurnal fluctuations.

Trigonelline is known to increase in tomato leaves in response to increased nitrogen content in nutrient solutions (Tyihák et al., 1988), and can thus serve as a possible indicator of nitrogen status within the plant body. Therefore, we investigated the correlation between trigonelline content in leaves and nitrogen fertilizer absorption in IA04 and IA06 (**Supplementary Table S9**). The results showed a positive correlation (r = 0.56, p < 0.05) in IA06 and a weak correlation (r = 0.30, p < 0.05) in IA04, supporting this hypothesis. Trigonelline is synthesized from nicotinic acid, which is a metabolite of the nicotinamide adenine dinucleotide (NAD) synthesis/degradation (Ashihara, 2006). The functions of trigonelline in plants have been reported in terms of various aspects, such as cell cycle regulation, nodulation, and reduction of oxidative stress (Minorsky, 2002). A recent study reported on the function of trigonelline as a detoxified metabolite of excess nicotinic acid in the NAD cycle (Li et al., 2017). The demethylation of trigonelline regenerated nicotinic acid for utilization in NAD synthesis as a reservoir metabolite. Demethylating activity has also been observed in the leaves of some plants, as well as in coffee plant seeds, during germination (Ashihara, 2006). In *Arabidopsis thaliana*, NAD is known to play an important role in growth phase transition (Hashida et al., 2016). In a previous study, the perturbation of NAD redox homeostasis due to the overexpression of genes involved in NAD synthesis resulted in the ectopic generation of reactive oxygen species, leading to early flower stalk wilting and shortened plant longevity (Hashida et al., 2016). In addition, NAD accumulation was reported in pollen before germination, indicating that NAD metabolism plays a crucial role in pollen maturation (Hashida et al., 2013). Our hypothesis is that trigonelline may be involved in flower development via NAD homeostasis, however, further experiments are required to confirm this hypothesis.

4.2 Improving Predictability by Using Environmental Data

Although we attempted to use environmental factors to predict reproductive traits, the prediction performances of the generated models were poor (**Table 1** and **Supplementary Figure S2A**). These results support our understanding that short-term environmental data are insufficient for yield prediction. Accumulated historic datasets of environmental factors may be required to achieve more accurate predictions (Adams, 2002; Qaddoum et al., 2013; Saito et al., 2020). On the other hand, the combination of metabolome and environmental factor data resulted in improved prediction performance (**Table 1**). Considering a plant as an autotrophic production system, it is reasonable that a combination of environmental factors (system inputs) and metabolic status (a system internal condition) can produce more accurate production estimates (system outputs) than either one individually. Thus, monitoring both types of factors in a greenhouse system management is likely to yield the best prediction performance.

4.3 Machine Learning Algorithms for Metabolome Data

Among the machine learning approaches, LASSO linear regression analysis was chosen for the following reasons. First, linear regression is often used to estimate biological rates (Schneider et al., 2010). Thus, linear regression seems to be an appropriate choice for our experiments. Second, our dataset contained more variables than samples, which could lead to severe overfitting in a more complex model (Trunk, 1979). A simpler model, such as a linear regression model combined with LASSO regularization, is preferred; therefore, the LASSO linear regression method is employed in this study. In fact, we have previously tested several other regression algorithms, including ridge regression, random forest regressor, k-nearest neighbor regression, and support vector regression (Pedregosa et al., 2011; VanderPlas, 2016), all of which performed worse than or the same as the LASSO model with our dataset (data not shown). A detailed comparison of these algorithms will be described elsewhere. Based on this knowledge, LASSO was chosen for this study.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: http://prime.psc.riken.jp/archives/data/DropMet/059/.

AUTHOR CONTRIBUTIONS

RS, JM, KY, and MYH conceived and designed the study. RS and JM conducted the study, analyzed, interpreted the data, and wrote the paper. JM, TS, HN, MI, YI, MS, and TH contributed to the acquisition of data. KY and MYH supervised the research. MA, SA, TH, KY, and MYH edited and reviewed the article. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Cabinet Office, Government of Japan, Cross-ministerial Strategic Innovation Promotion Program (SIP), “Technologies for creating next-generation agriculture, forestry and fisheries” (funding agency: Bio-oriented Technology Research Advancement Institution, NARO), and MEXT KAKENHI Grant Numbers 19H04870 to KY and 18H04808 and 20H04852 to MYH. This work was also supported in part by the Research Funding for the Computational Software Supporting Program of Meiji University. Computations were partially performed on the NIG supercomputer at the ROIS National Institute of Genetics.

ACKNOWLEDGMENTS

We thank Y. Yamada for data repository on the DROP Met.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.839051/full#supplementary-material

REFERENCES

Acharjee, A. (2013). Comparison of Regularized Regression Methods for ~Omics Data. *Metabolomics* 03 (3), 126. doi:10.4172/2153-0769.1000126

Adams, S. R. (2002). Predicting the Weekly Fluctuations in Glasshouse Tomato Yields. *Acta Hortic.* 593, 19–23. doi:10.17660/ActaHortic.2002.593.1

Ashihara, H. (2006). Metabolism of Alkaloids in Coffee Plants. *Braz. J. Plant Physiol.* 18 (1), 1–8. doi:10.1590/s1677-04202006000100001

Das, B., Nair, B., Reddy, V. K., and Venkatesh, P. (2018). Evaluation of Multiple Linear, Neural Network and Penalised Regression Models for Prediction of rice Yield Based on Weather Parameters for West Coast of India. *Int. J. Biometeorol.* 62 (10), 1809–1822. doi:10.1007/s00484-018-1583-6

de Tayrac, M., Lê, S., Aubry, M., Mosser, J., and Husson, F. (2009). Simultaneous Analysis of Distinct Omics Data Sets with Integration of Biological Knowledge: Multiple Factor Analysis Approach. *BMC Genomics* 10, 32. doi:10.1186/1471-2164-10-32

Dinar, M., and Rudich, J. (1985). Effect of Heat Stress on Assimilate Metabolism in Tomato Flower Buds. *Ann. Bot.* 56 (2), 249–257. doi:10.1093/oxfordjournals.aob.a087009

Du, Q., Campbell, M., Yu, H., Liu, K., Walia, H., Zhang, Q., et al. (2019). Network-based Feature Selection Reveals Substructures of Gene Modules Responding to Salt Stress in rice. *Plant Direct* 3 (8), e00154. doi:10.1002/pld3.154

FAOSTAT (2018). *Food, Agriculture Organization of the United, Nations*. Rome, Italy: FAOSTAT Database.

Gao, N., Teng, J., Ye, S., Yuan, X., Huang, S., Zhang, H., et al. (2018). Genomic Prediction of Complex Phenotypes Using Genic Similarity Based Relatedness Matrix. *Front. Genet.* 9 (364), 364. doi:10.3389/fgene.2018.00364

Hagberg, A. A., Schult, D. A., and Swart, P. (2008). “Exploring Network Structure, Dynamics, and Function using NetworkX,” in *Proceedings of the 7th Python in Science Conference*, August 19–24, 2008. Editors G. Varoquaux, T. Vaught, and J. Millman (Pasadena, CA USA), 11–15. Available at: http://conference.scipy.org/proceedings/SciPy2008/paper_2/.

Hashida, S.-n., Itami, T., Takahara, K., Hirabayashi, T., Uchimiya, H., and Kawai-Yamada, M. (2016). Increased Rate of NAD Metabolism Shortens Plant Longevity by Accelerating Developmental Senescence inArabidopsis. *Plant Cel Physiol* 57 (11), 2427–2439. doi:10.1093/pcp/pcw155

Hashida, S.-n., Takahashi, H., Takahara, K., Kawai-Yamada, M., Kitazaki, K., Shoji, K., et al. (2013). NAD+ Accumulation during Pollen Maturation in Arabidopsis Regulating Onset of Germination. *Mol. Plant* 6 (1), 216–225. doi:10.1093/mp/sss071

Heuvelink, E., and Buiskool, R. P. M. (1995). Influence of Sink-Source Interaction on Dry Matter Production in Tomato. *Ann. Bot.* 75 (4), 381–389. doi:10.1006/anbo.1995.1036

Jones, E., Oliphant, T., and Peterson, P. (2001). *SciPy: Open Source Scientific Tools for Python*.

Khan, A., and Sagar, G. R. (1969). Alteration of the Pattern of Distribution of Photosynthetic Products in the Tomato by Manipulation of the Plant. *Ann. Bot.* 33 (4), 753–762. doi:10.1093/oxfordjournals.aob.a084322

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: AnRPackage for Multivariate Analysis. *J. Stat. Soft.* 25 (1), 18. doi:10.18637/jss.v025.i01

Li, W., Zhang, F., Wu, R., Jia, L., Li, G., Guo, Y., et al. (2017). A Novel N-Methyltransferase in Arabidopsis Appears to Feed a Conserved Pathway for Nicotinate Detoxification Among Land Plants and Is Associated with Lignin Biosynthesis. *Plant Physiol.* 174 (3), 1492–1504. doi:10.1104/pp.17.00259

Liabeuf, D., Sim, S.-C., and Francis, D. M. (2018). Comparison of Marker-Based Genomic Estimated Breeding Values and Phenotypic Evaluation for Selection of Bacterial Spot Resistance in Tomato. *Phytopathology* 108 (3), 392–401. doi:10.1094/PHYTO-12-16-0431-R

Liebisch, F., Max, J. F. J., Heine, G., and Horst, W. J. (2009). Blossom-end Rot and Fruit Cracking of Tomato Grown in Net-covered Greenhouses in Central Thailand Can Partly Be Corrected by Calcium and boron Sprays. *Z. Pflanzenernähr. Bodenk.* 172 (1), 140–150. doi:10.1002/jpln.200800180

McKinney, W. (2010). "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, June 28–July 3, 2010. Editor S.e.v.d.W.a.J. Millman (Austin, TX) 445, 51–56. doi:10.25080/Majora-92bf1922-00a

Menéndez, P., Eilers, P., Tikunov, Y., Bovy, A., and van Eeuwijk, F. (2011). Penalized Regression Techniques for Modeling Relationships between Metabolites and Tomato Taste Attributes. *Euphytica* 183 (3), 379–387. doi:10.1007/s10681-011-0374-5

Minorsky, P. V. (2002). Trigonelline: A Diverse Regulator in Plants. *Plant Physiol.* 128 (1), 7–8. doi:10.1104/pp.900014

Ono, K., Muetze, T., Kolishovski, G., Shannon, P., and Demchak, B. (2015). CyREST: Turbocharging Cytoscape Access for External Tools via a RESTful API. *F1000Res* 4, 478. doi:10.12688/f1000research.6767.1

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *J. Mach Learn. Res.* 12, 2825–2830. doi:10.1145/2786984.2786995

Peet, M. M., and Welles, G. (2005). "Greenhouse Tomato Production," in *Tomatoes*. Editor E. Heuvelink (Wallingford, UK: CABI Publishing), 257–304. doi:10.1079/9780851993966.0257

Qaddoum, K., Hines, E. L., and Iliescu, D. D. (2013). Yield Prediction for Tomato Greenhouse Using EFuNN. *ISRN Artif. Intelligence* 2013, 1–9. doi:10.1155/2013/430986

Rasmussen, M. A., and Bro, R. (2012). A Tutorial on the Lasso Approach to Sparse Modeling. *Chemometrics Intell. Lab. Syst.* 119, 21–31. doi:10.1016/j.chemolab.2012.10.003

Rish, I., and Grabarnik, G. (2014). *Sparse Modeling: Theory, Algorithms, and Applications*. Boca Raton: CRC Press.

Saito, T., Kawasaki, Y., Ahn, D.-H., Ohyama, A., and Higashide, T. (2020). Prediction and Improvement of Yield and Dry Matter Production Based on Modeling and Non-destructive Measurement in Year-Round Greenhouse Tomatoes. *Hortic. J.* 89 (4), 425–431. doi:10.2503/hortj.UTD-170

Saure, M. C. (2014). Why Calcium Deficiency Is Not the Cause of Blossom-End Rot in Tomato and Pepper Fruit - a Reappraisal. *Scientia Horticulturae* 174, 151–154. doi:10.1016/j.scienta.2014.05.020

Sawada, Y., Akiyama, K., Sakata, A., Kuwahara, A., Otsuki, H., Sakurai, T., et al. (2009). Widely Targeted Metabolomics Based on Large-Scale MS/MS Data for Elucidating Metabolite Accumulation Patterns in Plants. *Plant Cel Physiol* 50 (1), 37–47. doi:10.1093/pcp/pcn183

Sawada, Y., Sato, M., Okamoto, M., Masuda, J., Yamaki, S., Tamari, M., et al. (2019). Metabolome-based Discrimination of chrysanthemum Cultivars for the Efficient Generation of Flower Color Variations in Mutation Breeding. *Metabolomics* 15 (9), 118. doi:10.1007/s11306-019-1573-7

Schneider, A., Hommel, G., and Blettner, M. (2010). Linear Regression Analysis: Part 14 Of A Series On Evaluation Of Scientific Publications. *Dtsch Arztebl Int.* 107 (44), 776–782. doi:10.3238/arztebl.2010.0776

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Tanaka, A., and Fujita, K. (1974). Nutrio-physiological Studies on the Tomato Plant IV. Source-Sink Relationship and Structure of the Source-Sink Unit. *Soil Sci. Plant Nutr.* 20 (3), 305–315. doi:10.1080/00380768.1974.10433252

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* 58 (1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Trunk, G. V. (1979). A Problem of Dimensionality: a Simple Example. *IEEE Trans. Pattern Anal. Mach. Intell.* 1 (3), 306–307. doi:10.1109/TPAMI.1979.4766926

Tyihák, E., Sarhan, A. R. T., Cong, N. T., Barna, B., and Király, Z. (1988). The Level of Trigonelline and Other Quaternary Ammonium Compounds in Tomato Leaves in Ratio to the Changing Nitrogen Supply. *Plant Soil* 109 (2), 285–287. doi:10.1007/bf02202097

VanderPlas, J. (2016). *Python Data Science Handbook : Essential Tools for Working with Data*. O'Reilly Media.

Yamamoto, E., Matsunaga, H., Onogi, A., Kajiya-Kanegae, H., Minamikawa, M., Suzuki, A., et al. (2016). A Simulation-Based Breeding Design that Uses Whole-Genome Prediction in Tomato. *Sci. Rep.* 6, 19454. doi:10.1038/srep19454

Yano, K., Imai, K., Shimizu, A., and Hanashita, T. (2006). A New Method for Gene Discovery in Large-Scale Microarray Data. *Nucleic Acids Res.* 34 (5), 1532–1539. doi:10.1093/nar/gkl058

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Molecular Biosciences

REVIEW
published: 08 March 2022
doi: 10.3389/fmolb.2022.841373

Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation

Adam Amara[1]*, *Clément Frainay*[2], *Fabien Jourdan*[2,3], *Thomas Naake*[4], *Steffen Neumann*[5,6], *Elva María Novoa-del-Toro*[2], *Reza M Salek*[7]*, *Liesa Salzer*[8], *Sarah Scharfenberg*[5] and *Michael Witting*[9,10]*

[1]Section of Nutrition and Metabolism, International Agency for Research on Cancer (IARC-WHO), Lyon, France, [2]Toxalim (Research Centre in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, Toulouse, France, [3]MetaboHUB-Metatoul, National Infrastructure of Metabolomics and Fluxomics, Toulouse, France, [4]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany, [5]Bioinformatics and Scientific Data, Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany, [6]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany, [7]Bruker BioSpin GmbH, Ettlingen, Germany, [8]Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, Neuherberg, Germany, [9]Metabolomics and Proteomics Core, Helmholtz Zentrum München, Neuherberg, Germany, [10]Chair of Analytical Food Chemistry, TUM School of Life Sciences, Freising, Germany

OPEN ACCESS

Edited by:
Wolfram Weckwerth,
University of Vienna, Austria

Reviewed by:
Thomas O. Metz,
Pacific Northwest National Laboratory (DOE), United States
Julijana Ivanisevic,
University of Lausanne, Switzerland

***Correspondence:**
Adam Amara
AmaraA@iarc.fr
Reza M Salek
Reza.Salek@bruker.com
Michael Witting
Michael.witting@helmholtz-muenchen.de

Specialty section:
This article was submitted to Metabolomics,
a section of the journal
Frontiers in Molecular Biosciences

Received: 22 December 2021
Accepted: 18 February 2022
Published: 08 March 2022

Citation:
Amara A, Frainay C, Jourdan F, Naake T, Neumann S, Novoa-del-Toro EM, Salek RM, Salzer L, Scharfenberg S and Witting M (2022) Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation.
Front. Mol. Biosci. 9:841373.
doi: 10.3389/fmolb.2022.841373

Both targeted and untargeted mass spectrometry-based metabolomics approaches are used to understand the metabolic processes taking place in various organisms, from prokaryotes, plants, fungi to animals and humans. Untargeted approaches allow to detect as many metabolites as possible at once, identify unexpected metabolic changes, and characterize novel metabolites in biological samples. However, the identification of metabolites and the biological interpretation of such large and complex datasets remain challenging. One approach to address these challenges is considering that metabolites are connected through informative relationships. Such relationships can be formalized as networks, where the nodes correspond to the metabolites or features (when there is no or only partial identification), and edges connect nodes if the corresponding metabolites are related. Several networks can be built from a single dataset (or a list of metabolites), where each network represents different relationships, such as statistical (correlated metabolites), biochemical (known or putative substrates and products of reactions), or chemical (structural similarities, ontological relations). Once these networks are built, they can subsequently be mined using algorithms from network (or graph) theory to gain insights into metabolism. For instance, we can connect metabolites based on prior knowledge on enzymatic reactions, then provide suggestions for potential metabolite identifications, or detect clusters of co-regulated metabolites. In this review, we first aim at settling a nomenclature and formalism to avoid confusion when referring to different networks used in the field of metabolomics. Then, we present the state of the art of network-based methods for mass spectrometry-based metabolomics data analysis, as well as future developments expected in this area. We cover the use of networks applications using biochemical reactions, mass spectrometry features, chemical structural similarities, and correlations between metabolites. We also describe the application of knowledge networks such as metabolic reaction networks. Finally, we discuss the possibility of combining different networks to analyze and interpret them simultaneously.

Keywords: metabolic network, untargeted metabolomics, graph-based analysis, knowledge network, experimental network, metabolism, systems biology

INTRODUCTION

Metabolomics research is based on various opportunities to uncover the metabolites contained in biological samples. To characterize and quantify metabolites in biological samples, different types of metabolite separation techniques - such as Liquid Chromatography (LC), Gas Chromatography (GC), Capillary Electrophoresis (CE), or Ion Mobility (IM)–are coupled to a Mass-Spectrometry (MS) system. High-performance mass spectrometry systems generate increasingly complex datasets. Two major approaches are used in metabolomics: targeted methods look for a pre-selected list (or class) of metabolites, and untargeted metabolomics covers as many metabolites as possible (Schrimpe-Rutledge et al., 2016). However, in untargeted metabolomics research, processing, analyzing, and interpreting the complex datasets that are generated are major challenges. Nuclear Magnetic Resonance (NMR) techniques are also used in metabolomics (Emwas et al., 2019), but most of the network and graph methods covered here are rather focused on MS-based metabolomics. As multiple network constructions approaches presented here are relying on the specificity of data generated by MS (e.g., fragmentation or adducts).

The analysis of untargeted metabolomics datasets is frequently limited by the ability to annotate and identify metabolites at a large scale (hundreds or thousands of metabolites). Data interpretation is often reductionist and limited to a few specific metabolic processes or metabolites, found to be statistically significantly associated with a phenotype of interest. This implies that a potentially large part of the detected metabolites will be ignored if they appear not statistically significant to the question at hand. Importantly, the recent use of network and graph-based methods to analyze metabolomics data opened the possibility of metabolomics data systematic analysis (Kell and Goodacre, 2014; Perez De Souza et al., 2020).

There are two major types of networks used with metabolomics data: knowledge and experimental (**Figure 1**). Knowledge networks are generated from biochemical or biological knowledge and allow interpreting metabolomics data in the context of prior biological knowledge, such as metabolic pathways and enzymatic reactions. For instance, a metabolic network is a knowledge network, where metabolites and their known biochemical conversions are represented as nodes and edges, respectively. On the other hand, experimental networks are generated from the metabolomics data itself, based on relationships between possible or identified metabolites in the data (e.g., spectral similarity, or correlation). Notably, both types of networks (i.e., knowledge and experimental) can be used with advanced statistical methods, graph analysis, and data analysis approaches to study the interconnected data.

The words "network" and "graph" are often used interchangeably, and preferred terms depend on fields and traditions. We will refer to the curated lists of biochemical reactions and their participants (e.g., substrate, products, enzymes, and genes) as "metabolic networks" (following

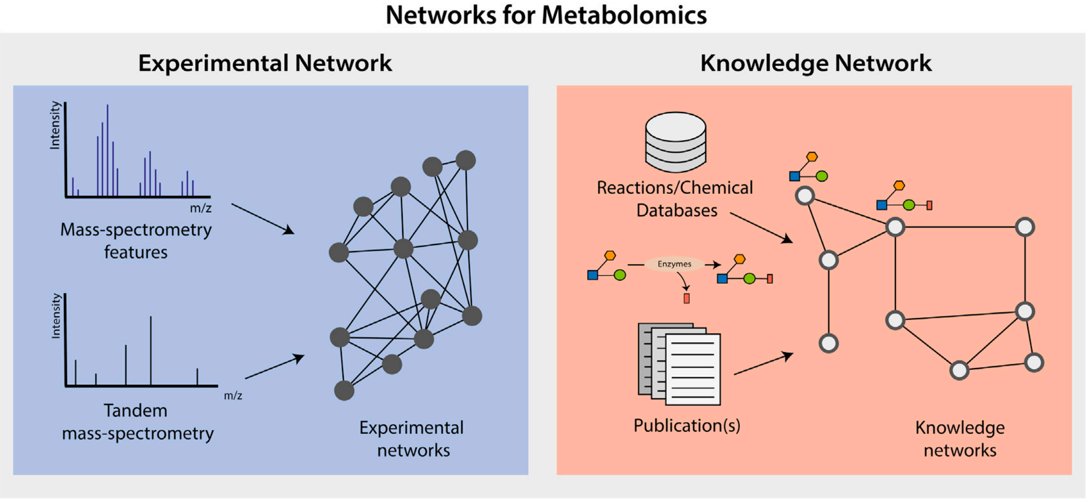


FIGURE 1 | Graphical Abstract. In this review we will be presenting two major types of networks and graphs used to analyze and interpret metabolomics data, knowledge networks and experimental networks.

current usage). We will refer as "metabolic graphs" the different entity-relationship structures that can be derived from such biochemical reaction lists to perform topological analysis (such as compound graphs and reaction graph), to avoid the ambiguity with their source material.

Metabolism consists of enzymatic and non-enzymatic reactions converting metabolites to produce energy (catabolism), build up biomass (anabolism), or respond to external stimuli. Metabolism is often seen as functional modules conserved across organisms. Examples of such functional modules are the central carbon metabolism, which is highly conserved, and the secondary (AKA specialized) metabolism, which differs vastly among organisms. Furthermore, co-metabolism in communities (such as microbiomes) increases metabolic capacities and leads to a very high diversity of metabolites. In this context, the unilateral interpretation of metabolomics data may hide complex systemic changes spanning across several pathways. This is especially the case with metabolic chart representations that are designed to focus on knowledge-based biochemical pathways and ignore the interconnections between pathways. Additionally, the lack of consensus on the partitioning of metabolic pathways or modules from one database to another can lead to major discrepancies in the analysis (Stobbe et al., 2012; Altman et al., 2013). Instead, it is possible to represent the metabolism as a network of metabolites connected by specific or promiscuous enzymatic and non-enzymatic reactions. Importantly, in such a network, we can also represent interconnections between metabolites which may look unrelated but that are connected via different pathways. Genome-Scale Metabolic Networks (GSMNs) are designed to represent this information based on genomics knowledge, providing a systemic view of the metabolism. Nevertheless, GSMNs are based on metabolism knowledge coming from genome annotation, which prevents the integration of many metabolites since there are gaps in knowledge (e.g., secondary metabolism) (Frainay et al., 2018). These gaps require us to expand those metabolic networks using experimental data from metabolomics experiments.

Untargeted MS data, either based on direct infusion or coupled to different types of separation techniques (e.g., LC, GC, CE, or IM), is characterized by features for which we measured the mass-to-charge ratio (m/z value with a mass accuracy of just a few ppm, depending on the instrumentation), the abundance (either a peak intensity or a peak area), an additional separation index (retention or migration time, mobility, or collisional cross-section value), and the associated fragmentation pattern, if collected. Based on these data, metabolites can be annotated or identified with different confidence levels, according to the Metabolomics Standard Initiative (MSI) (Fiehn et al., 2007; Sumner et al., 2007; Schymanski et al., 2014). The highest level of confidence (i.e., level 1) is achieved by a matching in at least two independent and orthogonal data (e.g., mass spectrum and retention time/index) between the metabolite feature and its authentic reference standard, both of which must be analyzed under identical conditions. This identification level is often only possible for metabolites for which reference standards are available in the respective laboratory. Indeed, recent work has shown that only a small part of the metabolites found in metabolic networks of different organisms is covered by at least one reference spectrum (Frainay et al., 2018). Lower-confidence annotations (i.e., levels 2 and 3) can be achieved by matching the metabolite feature with spectral libraries or using in-silico tools, such as MetFrag (Ruttkies et al., 2019) or CSI: FingerID (Dührkop et al., 2015), among others (Misra and van der Hooft, 2016; Spicer et al., 2017; Misra, 2021). Assessing the structural similarity relationship via spectral similarity has proven to be a powerful tool to guide annotation of unknown metabolites (Wang et al., 2016), since chances of having structurally homologous metabolites detected in parallel are high. However, metabolites are generally not detected as isolated entities, but as part of larger sets of metabolites of the same chemical classes.

Here, we will describe the current state of the art in terms of networks and graphs usage for metabolomics, detailing their characteristics and applications. We will first focus on experimental networks (such as those based on mass differences, adducts and features, structure similarities, and correlation), which are generated from metabolomics data. Notably, experimental networks have been used to annotate and identify metabolites (Loos and Singer, 2017; Schmid et al., 2021), as well as to better understand biochemical relationships between metabolites (Schollée et al., 2017; Naake and Fernie, 2019). We will also describe knowledge networks (such as ontology-based networks (Dührkop et al., 2020) and GSMNs), which are increasingly used to interpret metabolomics data (Kell and Goodacre, 2014; Frainay and Jourdan, 2017) and to annotate metabolites (Silva et al., 2014; Schmid et al., 2021). While each network (experimental or knowledge-based) covers a specific aspect of the studied biology, there are benefits in integrating them. For instance, experimental networks can help in filling the gaps in current knowledge-based networks by mapping the nodes in the knowledge-based network (i.e., metabolites) with the corresponding nodes in the experimental networks (i.e., features) and identifying missing metabolites. Importantly, knowledge-based networks provide a biological context to help interpret and analyze experimental networks. To emphasize this, we finish this review by presenting combined networks analysis approaches, such as multi-layer networks applied to the field of metabolomics.

EXPERIMENTAL NETWORKS

Experimental networks are directly derived from the acquired untargeted metabolomics data. Depending on the type of network, either MS^1, MS^2, or MS^n data is used. Each network tackles a different aspect of the compounds "metabolic relatedness", with specific assumptions and shortcomings, which we will describe in the following sections. We will discuss how mass differences, adducts and features, structure similarities and correlation data can be used to build different experimental networks.

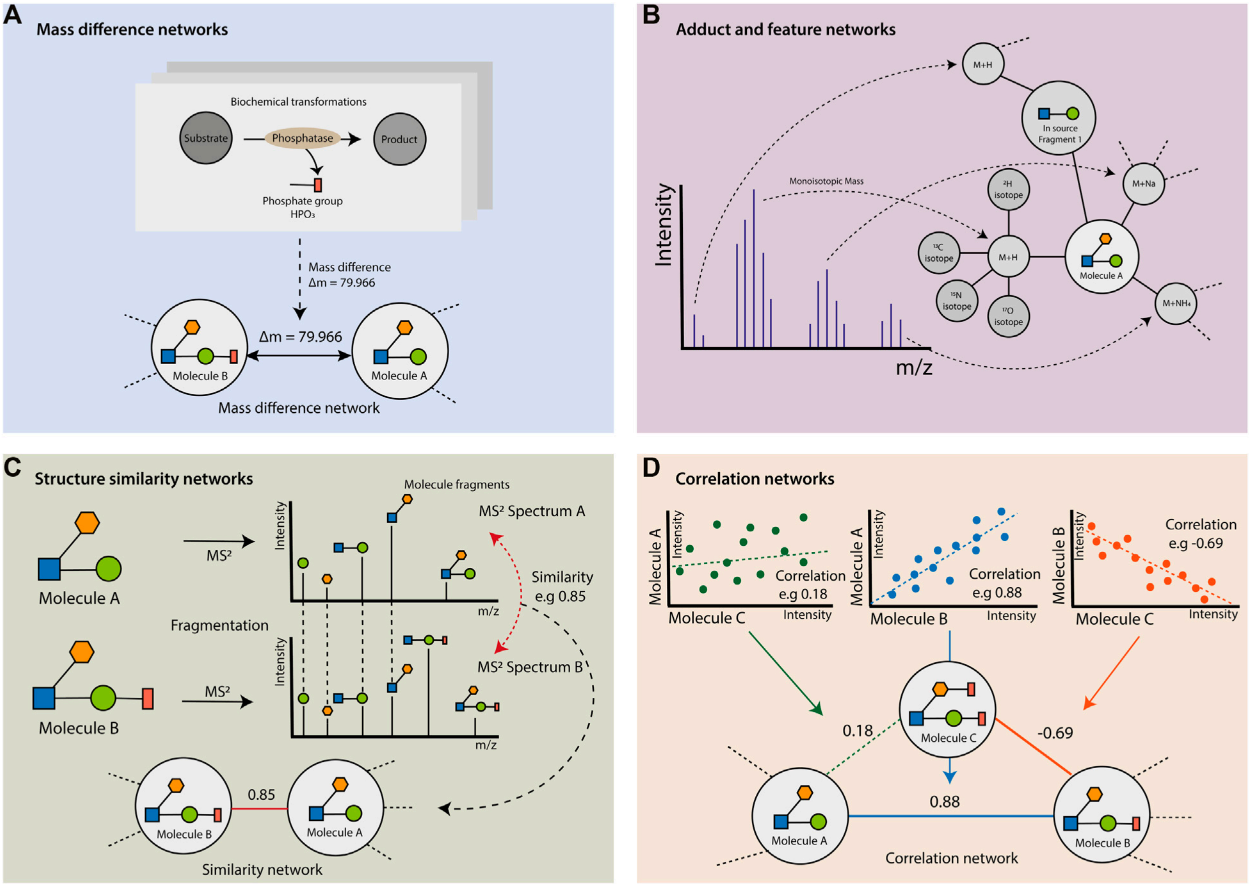


FIGURE 2 | Metabolomics-based experimental networks. **(A)** Mass difference networks: the biochemical transformations entail gains and/or losses of atoms that lead to changes in the metabolites' molecular formula and, therefore, changes in the exact mass of molecules connected via a reaction. Here, the biochemical transformation by a phosphatase causes the loss of a phosphate group (HPO3), leading to a mass difference of 79.966 between the substrate metabolite (Molecule **(B)** and the product metabolite (Molecule A). **(B)** Adduct and feature networks: metabolites have multiple possible adducts and features associated with them. Each detected adduct, isotopologue, and ion-source fragments can be represented as nodes. Adducts (e.g., M + H) are connected to corresponding or potential metabolites. Similarly, the isotopologues of an adduct are linked to the associated adduct nodes (e.g., 13C isotopologue of M + H). Finally, ion-source fragments (here in-source fragment 1) with their associated adducts and isotopologues can be linked to the corresponding node metabolite. **(C)** Structure similarity networks: the structural similarity between metabolites detected by MS methods can be observed and calculated based on their MS/MS spectra. The fragmentation patterns will be similar for two metabolites with a shared core structure (represented as circles, squares, and polygons), but a difference due to a chemical reaction (i.e., the residue represented by the red rectangle). The calculated similarity (i.e., 0.85) between two MS2 spectra is the weight of the edge linking the corresponding metabolite pair. **(D)** Correlation networks: the correlation between the abundances of two metabolites can be calculated and used as a weight for the edge (i.e., 0.88 or −0.69) between two metabolites' node (i.e., between molecules A and B, or between molecules B and C). The correlation levels considered as non-significant (i.e., 0.18) can be ignored and excluded from the correlation network (i.e., the edge between molecules A and C).

It is important to highlight that experimental networks complement each other to decipher the metabolic relationships between compounds. As two faces of the same coin, spectral similarity networks can suggest substrate-product links from expected global chemical similarity (**Figure 2C**); while mass difference networks represent the substrate-product links from characteristic differences due to local chemical structure changes (**Figure 2A**). Extra evidence of the existence of such substrate-product links can come from correlation networks, which reveal possible causal relationships between the changes in the metabolites' abundances (**Figure 2D**). Finally, the adduct and feature networks can increase the confidence in metabolites' annotations in the networks, based on characteristic patterns, associated to individual compounds in mass spectrometry (**Figure 2B**).

Mass Difference Networks

The biochemical transformations are characterized by the gain or loss of atoms, which lead to changes in the metabolites' molecular formula and, therefore, variations in the exact mass of pairs of molecules connected by a reaction. These changes can be measured in MS-based metabolomics as differences between pairs of m/z values (**Figure 2A**) to generate a mass difference network (**Table 1**).

TABLE 1 | Description of the key characteristics of mass difference networks.

Mass difference networks' main characteristics	
Nodes	Features, low level annotations (m/z + RT)
Edges	Putative substrate-product relationships from biochemical transformations' characteristic patterns
Main Hypothesis	Many biochemical reactions involve functional group transfer, yielding a characteristic mass shift. Feature pairs with mass differences matching those patterns might be involved in a reaction transferring the corresponding group
Limitations	The mass differences between a pair of features may correspond to an existing reaction, but it can happen that these two features do not correspond to a real biotransformation between the corresponding metabolites, leading to spurious edges

The mass difference approach can be used with known biotransformations and their corresponding mass differences to find potential biochemical reactions explaining the difference between m/z values (Breitling et al., 2006; Tziotis et al., 2011). Therefore, in a mass difference network, the features with their corresponding m/z values are represented as nodes, and the mass differences between pairs of m/z values that match a pre-defined transformation as edges (**Figure 2A**). Potential transformations can be derived from metabolic reaction databases, such as KEGG, MetaNetX, MetExplore, etc. (Jeffryes et al., 2015; Hadadi et al., 2016; Kanehisa et al., 2017; Cottret et al., 2018; Ebastien Moretti et al., 2021). If seed formulae (e.g., from identified metabolites) are available, information on known biochemical transformations can also be used to calculate molecular formulae, by propagating the difference formulae within the network. By comparing the frequency of certain mass differences between different conditions, conclusions on potential biochemical responses can be drawn (Moritz et al., 2016). However, this approach requires a priori hypothesis on data to generate an appropriate transformation list. Notably, features connected by a mass difference that is not included in the transformation list will not be connected in the mass difference network. Moreover, if metabolites from a reaction series are not detected by the instrument, there will be gaps (missing nodes) in the reconstructed network. For certain instances, this can be overcome, e.g., by combining several mass differences into one corresponding to multiple biotransformations. For example, gaps in the network for series of alkyl chains (C_nH_{2n+1}) can be filled by adding C_2H_4 to the transformation list to cover for two times CH_2 or by adding C_4H_8 to cover two times C_2H_4.

Another approach frequently used is to include all mass differences between all pairs of features, to generate mass difference networks. The result is a fully connected graph where all features are connected to each other, and their edges represent their mass differences. It is challenging to find meaningful network motifs in such a graph, since even non-biochemically related features would still be connected by an edge, with the sole purpose of holding the mass difference attribute. One solution to reduce irrelevant links is to filter out edges connecting features with low intensity/concentration correlation. It is also possible to filter edges following a specific Retention Time (RT) trend. For example, there is a predictable RT and mass difference between products and substrates of a specific reaction, which can be propagated from a known metabolite in the network to neighboring metabolites. This approach can result in the discovery of new biochemical transformations unbiased, as it does not use biotransformation-based mass differences (Morreel et al., 2014). However, the interpretation of the results might become complicated, as it represents a combination of several losses and gains of atoms. As an example, in the transamination reaction, transamination of pyruvate ($C_3H_4O_3$ to alanine ($C_3H_7NO_2$ is accompanied by the gain of one nitrogen and three hydrogens (NH_3 = 17.03) and the loss of one oxygen (O = 15.99), yielding to a net mass difference of 1.0316, from which no meaningful formula can be calculated.

There are different tools for the generation of mass difference networks. The tool mzGroupAnalyzer can generate a mass difference network based on an input list of transformations atom differences, it allows visualization of the metabolites elements composition with a van Krevelen diagram (based on H/C and O/C ratios) to identify patterns of structural similarity between compounds (Doerfler et al., 2014). MetaNetter is a Cytoscape plugin that performs ab initio prediction of mass difference networks from high-resolution data, such as Orbitrap or Fourier transform ion cyclotron resonance mass spectrometer (FT-ICR-MS) (Jourdan et al., 2008; Burgess et al., 2017). MetNet is an R package that represents one of the most prominent tools to generate mass difference networks based on pre-defined transformations lists; in combination with other types of information (such as RT shifts or correlations) (Naake and Fernie, 2019). The inclusion of such additional information reduces the connection degree between features, as it constrains the creation of edges between nodes with a threshold of correlations and/or with specific RT shifts.

Adducts and Features Networks

Mass differences do not only occur due to biological transformations between metabolites, but might also appear due to different physicochemical effects when introducing the metabolites to the MS. These "non-biological" mass differences can be represented in adducts and feature networks (**Table 2**). The relationships between features are used for grouping and deconvoluting the detected m/z signals, as in the R package CAMERA (Kuhl et al., 2012). Analysis of mass differences is greatly enhanced using chromatographic separation, as the RT windows help to separate metabolites features. Isotopes, adducts, as well as in-source fragments of the same metabolite show (theoretically perfect) co-elution. A particular example of co-elution is the annotation of [M +

TABLE 2 | Description of the key characteristics of adducts and features networks.

Adducts and features networks' main characteristics	
Nodes	Features, intermediate level annotations (m/z + RT + adduct information)
Edges	Putative relationships between features, such as adducts of a metabolite, in-source fragments of a metabolite, or isotopologues of an adduct
Main Hypothesis	Same as mass difference networks, but with more detailed description using RT separation and characteristic patterns associated to adducts ions formation, ion-source fragmentations, and isotope patterns
Limitations	The mass differences between two features can correspond to a chemical relationship between two features (e.g., isotopologues or adducts), but these features correspond to the same metabolite

TABLE 3 | Description of the key characteristics of structure similarity and MS/MS networks.

Structure similarity networks' main characteristics	
Nodes	Features, high level annotations (m/z + RT + MS^2)
Edges	Calculated similarity between two MS/MS spectra of two fragmented adducts
Main Hypothesis	Reactions tend to involve substrate-products pairs with high structural similarity. Thus, detected compounds with high structural similarity might be substrate/product of the same reaction
Limitations	Many compounds with structural similarity are not involved in the same reactions or pathways, which yields to false positives Depending on the methods and instruments, not every feature corresponding to a metabolite will be fragmented

$H]^+$ and $[M\text{-}H_2O + H]^+$, while $[M\text{-}H_2O + H]^+$ normally co-elutes with $[M + H]^+$, metabolites that differ in H_2O in their formulas have different chemical structures and therefore different RTs.

In-source fragmentation (ISF) is a common phenomenon that occurs in Electrospray ionization (ESI). ISF is the dissociation of a molecule that occurs within the ionization source of the mass spectrometer. During ESI, molecules gain additional internal energy that is released, resulting in the fragmentation of the molecule. This fragmentation generates additional precursor ions that can lead to false positive annotations of molecular features (Gathungu et al., 2018). There are several tools that can help with the identification of ISF of the same metabolite, e.g., CliqueMS, an R package that groups co-eluting features, based on similarity networks (Senan et al., 2019). Another recently developed R package that recognizes in-source fragments is ISFrag. ISFrag checks for co-elution, presence of the in-source fragment in the precursor MS^2 spectra, and spectral similarity (Guo et al., 2021).

Ion identity networking is used to generate a network based on the relationships between ion species linked to the same compound as well as structurally similar compounds, which enhances compound annotation (Nothias et al., 2020). The detected ion-source fragments and their associated adducts and isotopologues can be represented in the network as nodes with edges linking them to their associated metabolite nodes (**Figure 2B**).

Certain mass differences might be found in a consecutive manner, e.g., CH_2 or C_2H_4 for a homologous series, through an increase in an acyl chain length. Longer acyl chains lead to a higher RT in Reverse-Phase (RP) chromatography. Loos and Singer developed functionalities for the identification of homologous series by detecting series of mass differences following a given RT trend (Loos and Singer, 2017).

Structure Similarity Networks

Typically, molecules connected via biochemical reactions are chemically similar since they often share common substructures. This resemblance can be expressed by chemical similarity measures, such as the Tanimoto similarity (Bender and Glen, 2004; Bajusz et al., 2015). It is important to note that similarity measures can only be calculated between identified compounds, since they require chemical structures as input (**Table 3**).

In untargeted metabolomics, the MS^2 fragmentation data is mostly generated using Data-Dependent Acquisition (DDA), which results in the fragmentation of the most abundant features. Fragmentation data can be used to infer (to a certain degree) structural similarity. Consequently, chemically similar compounds are likely to show at least partially similar fragmentation patterns. Note that the spectral differences can be both varying fragment masses and neutral loss differences. Molecules that have a shared core structure (e.g., an aglycon) can have differences due to the chemical reaction (e.g., additions of glycosyl groups) linked by the similarity of their MS^2 spectra. Additionally, metabolites within the same compound class also show similar fragmentation patterns, even if they are not connected via biochemical reactions. An example is the fragmentation of glycerolipids, such as di- and tri-acylglycerols, which show characteristic neutral losses of fatty acid chains (Murphy et al., 2007).

Spectral similarity networks connect MS^2 spectra of features or metabolites that show spectral similarity values above a certain threshold (**Figure 2C**). Therefore, finding metabolites within the same compound class or a similar one connected by biochemical reactions.

Different algorithms have been developed to use spectral similarity (based on different metrics, such as cosine or modified cosine similarities) to construct molecular similarity

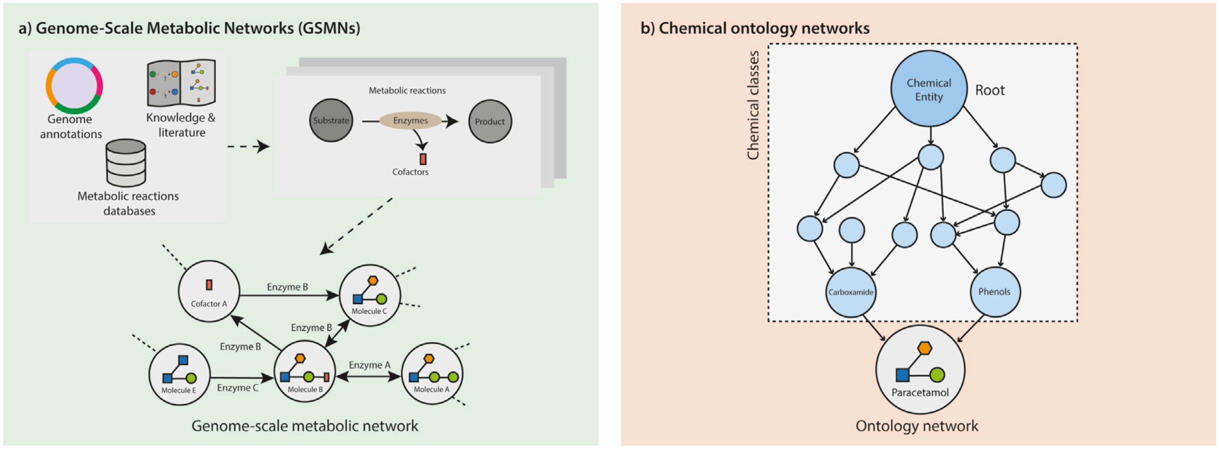


FIGURE 3 | Representation of knowledge as networks. **(A)** Genome-scale metabolic networks: reconstructed from different sources of knowledge, such as from the enzymes identified in the annotated genome of the organism under study, the metabolic reactions databases, and/or biochemical knowledge and literature. The known metabolic reactions in an organism are the basis to generate a genome-scale metabolic network, where the metabolites are represented as nodes that are linked by (directed or undirected) edges, which represent the reactions converting the metabolites. **(B)** Chemical ontology networks: structure of relationships represented as a semantic network, where the nodes represent chemicals or chemical classes as "concepts", bearing all their properties and definition, and that are connected by class membership.

networks, as a proxy for structural similarity (Demuth et al., 2004; Aguilar-Mogas et al., 2017). The first application of molecular similarity networks was proposed by Watrous et al. Their similarity measure was based on a modified cosine score, which considers the mass difference between precursor masses. The mass differences between the precursor masses are applied to the fragments in the MS^2 spectra, leading to a match of fragment peaks, either directly within a specific mass error or matching the mass plus the differences of the precursor masses. However, such spectral similarity networking only works on MS^2 spectra and merges all spectra from the same precursor m/z, ignoring the fact that different isomers might elute at different RTs (Watrous et al., 2012).

In DDA, the intensities of the fragments are often not representative of a feature abundance in different samples since the measurement of an MS^n spectrum is, in most cases, not triggered at the apex of a chromatographic peak. Feature-based molecular networking uses the abundance of the MS^1 feature (peak area or intensity), its RT, and the corresponding MS^2 spectra as input and therefore allows the differentiation of isomeric structures based on chromatography (Nothias et al., 2020). In the resulting networks, abundances can be used as an added criterion for data analysis, revealing potential biological links. However, in such networks, different adducts from a single compound might end up in separated sub-networks, based on highly similar fragmentation of adducts. Ion identity networking has been introduced to combine these sub-networks to group those adducts by combining molecular networking and MS^1 adduct detection algorithms, such as feature grouping and shape correlation (Schmid et al., 2021). This approach can also incorporate features into the network that have been identified as adducts but lack MS^2 information.

The most prominent tool-set used for molecular networking has been developed by the Global Natural Product Social Molecular Networking (GNPS; http://gnps.ucsd.edu) community (Wang et al., 2016). GNPS is an open-access platform that allows storing and analyzing MS^2 data, including molecular network generation using a modified cosine score and spectral library matching, followed by possible online visualizations.

Another example for generating molecular networks based on spectral similarity is MetGem, which utilizes the t-distributed stochastic neighbor embedding (t-SNE) algorithm to visualize the cosine scores calculated in the GNPS molecular networks. The t-SNE eases the interpretation of the molecular network by clustering together compounds that show high cosine scores, which eases the interpretation of the molecular network (Olivon et al., 2018).

There are different metrics to calculate spectral similarity. Indeed, cosine and modified cosine score might not often be the optimal choice for the construction of similarity networks. For example, compounds that show the same fragmentation pattern (i.e., the same neutral loss) but differ in the observed m/z show low cosine scores. It has been shown that Spec2Vec, a recently developed Python package that calculates spectral similarities based on fragmental relationships between large datasets, shows better overall performance than cosine-based scores, which were originally developed for matching fragmentation-rich electron ionization (EI) spectra (Huber et al., 2021).

Another approach to estimate spectral similarities is the use of hypothetical neutral loss spectra. An algorithm called core structure-based search (CSS) has been developed to calculate the spectral similarity between the mass difference between pairs of fragments ions. The CSS algorithm showed good performance in finding structurally relevant similarities (Xing et al., 2020). MS^2 data and its analysis are crucial for accessing the chemical structure of unknown metabolites. It has been shown that the combination of different bioinformatic tools further enhances

TABLE 4 | Description of the key characteristics of correlation and association networks.

Correlation and association networks' main characteristics	
Nodes	Features (intensity), low-level annotations (m/z)
Edges	Correlation between the abundances of two metabolites
Main Hypothesis	Metabolic processes imply metabolites' abundance that depends on other metabolites' abundance. Thus, metabolites with correlated abundance might be metabolically related
Limitations	The correlation or association between two metabolites (or features) does not systematically represent a metabolic relationship (e.g., substrate-product) or interdependence (e.g., co-regulation). Thus, correlation does not necessarily imply causal, biological, or chemical relationships

annotation success, which is of great importance, especially in untargeted metabolomics (Schmid et al., 2021).

Correlation and Association Networks

Metabolites that are connected in metabolic pathways often show co-dependency, which can be seen by their orchestrated concentration (i.e., abundance) changes. So, the metabolites' concentrations are correlated between metabolites that are associated or co-regulated within metabolic pathways (Rosato et al., 2018). Correlations of untargeted LC-MS metabolomics data are calculated by pairwise comparison of the peak intensity of all features, which results in a correlation adjacency matrix. In a correlation network, two metabolites are linked if their correlation value reaches a given (user-defined) threshold, which is considered as a significant correlation level (**Table 4**).

Most commonly, Pearson correlation is used to calculate correlations. However, due to tight metabolic control and the presence of long reaction sequences, standard Pearson correlation typically yields to highly connected and dense networks, which are hard to analyze and interpret. Gaussian graphical modeling uses partial instead of full correlation, and corrects for indirect correlation (i.e., when two metabolites are correlated just because they are both correlated with a third one). Therefore, using Gaussian graph modeling, only direct correlations can be found, which in turn allows us to construct meaningful networks containing potential direct reaction partners (Krumsiek et al., 2011). Benedetti et al. further compared the networks obtained using Pearson correlation, exact partial correlation, and partial correlation determined by GeneNet (Benedetti et al., 2020). They observed a dense network with an increased number of edges at increasing sample size for the Pearson correlation, whereas the partial correlation network (established with GeneNet) remained more stable. Furthermore, the statistical cut-off filter used to define the correlation threshold was more stable at varying the sample size using GeneNet than Pearson or partial correlations.

Another approach to statistically create metabolic networks is the weighted correlation network analysis, also known as weighted gene co-expression network analysis (WGCNA). In contrast to canonical correlation network analysis, the edges (which represent the correlation coefficients between features) are weighted by an exponent, such that the distribution of the weighted coefficients follows a power-law distribution, i.e., WGCNA assumes a priori a scale-free topology of the underlying network (Zhang and Horvath, 2005; Langfelder and Horvath, 2008). Nevertheless, to the best of our knowledge, it has not been proved yet if the statistical associations of the metabolites (or the subset acquired by GC- and LC-MS-based technologies) underlie such a scale-free topology.

WGCNA was originally applied to transcriptomics data, but it has also been recently employed for network generation using metabolomics data from human and human microbiome (Osterhoff et al., 2014; Pedersen and Sofia, 2018; Vernocchi et al., 2020; Murga-Garrido et al., 2021; Petersen et al., 2021), animal (Wu et al., 2021), and plants (DiLeo et al., 2011). (Samal and Martin, 2011).

KNOWLEDGE REPRESENTATION AS NETWORKS

Genome-Scale Metabolic Networks and Graphs

Genome-Scale Metabolic Networks (GSMNs) are based on the current knowledge of the metabolism of a given organism (e.g., human metabolic network Human 1 with 13,417 reactions and 4,164 metabolites) (Robinson et al., 2020). They are usually drafted from genome annotations and reaction databases, before manual curation by domain experts, using available literature and simulation results (**Table 5**). They encompass the gene–reaction–metabolite information with the matrix associating metabolites to reactions, and the association of reactions to their corresponding genes and enzymes (Thiele and Palsson, 2010) (**Figure 3A**). GSMNs are frequently used to simulate metabolic fluxes via constrained-based metabolic modeling (Becker et al., 2007; O'Brien et al., 2015). Nonetheless, we will focus here on the use of GSMNs as graphs, which we will refer to as Genome-Scale Metabolic Graphs (GSMGs). Different graphs (directed or undirected) can be derived from GSMNs (Lacroix et al., 2008). For instance, reaction graphs represent the reactions as nodes, and two reactions are connected by an edge if the product of the first reaction is the substrate of the second one. On the other hand, the nodes of a compound graph represent metabolites that are connected by edges if they are substrates and products of the same biochemical transformation. Graph-based analysis methods can be applied to GSMGs to study both the metabolism and metabolomics data (Lacroix et al., 2008; Cottret and Jourdan, 2010; Frainay and Jourdan, 2017). For instance, path searches in GSMGs have been used to infer metabolic pathways connecting metabolites of interest. While supplanted by flux methods for

TABLE 5 | Description of the key characteristics of genome-scale metabolic networks and graphs.

Main characteristics of genome-scale metabolic networks and graphs	
Nodes	A "pool" of compounds (not restricted to small molecules)
Edges	Substrate-product relationships from known reactions
Main Hypothesis	Using genome annotation, reaction databases, and manual/semi-auto curation to generate a model of an organism's metabolism
Limitations	There are gaps in the knowledge or predicted metabolic reactions in organisms, which creates an incomplete network of the metabolism

such goals, path searches are still used for metabolomics data clustering and visualization (Liggi and Griffin, 2017; del Mar Amador et al., 2018). While GSMG analysis has been mainly focused on path search, graph theory encompasses a vast range of applications. Centrality analysis, for example, aiming at identifying key nodes in a graph, is quite popular for regulation and protein interaction network analysis, and has been applied a few times to metabolic networks as well (Faust et al., 2010; Bánky et al., 2013; Frainay et al., 2019). Beyond metabolomics data analysis, graph-based metrics have been used more to characterize and compare whole metabolic networks (Ma and Zeng, 2003; Mazurie et al., 2010).

It is important to note that GSMNs do not cover all the metabolic products identified by metabolomics analysis, suggesting the absence of metabolic reactions and metabolites in the networks, as previously shown with the human GSMN (Frainay et al., 2018). This is a well-known problem as the GSMNs are biased by a reconstruction based on available genome annotations and knowledge of enzymatic reactions (Thiele et al., 2014; Pan and Reed, 2018). In consequence, gaps in the metabolic pathways are not always filled, as such gaps may also be due to enzymatic promiscuity and underground metabolisms (Notebaart et al., 2014; Pan and Reed, 2018).

The format in which GSMNs are stored can impact the graph structure and therefore the analysis of the graph, which is inconvenient. GSMNs are mainly shared in SBML format, which is an exchange format for computational models (not restricted to biochemical reactions) in biology (Hucka et al., 2003). SBML is mainly oriented towards quantitative models, which is why it has become the main support for GSMNs, given the popularity of GSMNs application for flux analysis. Building a network from a file in SBML format implies that the nodes correspond to a particular "species". It should be noted that the species nodes can represent other biological entities than metabolites (such as proteins, generic degradation products or even the whole "biomass"). Furthermore, due to the GSMNs being tailored for flux modeling, the species actually represent pools of available biological entities at a given time and location. Consequently, SBML tends to represent the same metabolite as multiple species ("pool") in different compartments, with a specific quantity that will be used for flux simulations. While SBML standard allows linking the species describing the same metabolites since version 2, in practice, those links are rarely defined. This leads to "duplicated" compartment-specific metabolites in many GSMNs, which differ from experimental networks in general, as compartment location is rarely available for metabolomics data. An alternative to SBML to represent metabolism knowledge is the BioPAX standard (Demir et al., 2010), oriented towards a semantic description of biological processes for indexing, sharing, and integration purposes, rather than quantitative modeling (Strömbäck and Lambrix, 2005). A network built from a BioPAX standard will have nodes that correspond to resources that describe biological entities, which are described using ontology vocabulary and linked to multiple information. However, BioPAX standard is mainly used at the individual pathway-level rather than the genome-scale level. Both exchange formats (SBML and BioPAX) represent knowledge about metabolism through lists of biochemical reactions, referencing metabolites as substrates or products (Strömbäck and Lambrix, 2005). A direct network translation would lead to a "bipartite metabolic graph", where both reactions and compounds are explicitly represented as nodes. Compounds are thus never directly connected by an edge, but always through a reaction node, which differs from the structure of experimental networks, where related compounds are directly linked by edges.

Chemical Ontology Networks

Chemical ontologies aim at providing a structured and formalized representation of chemical concepts. By describing an explicit structure of relationships among compounds, it can easily be represented as a semantic network that can be processed (**Table 6**).

One of the main differences with the other presented networks is that, in chemical ontologies, the links do not represent (or suggest) biochemical/metabolic relationships that involve the transformation of one node into another. Rather, they represent subsumption relations between chemical compounds and broader chemical classes. For example, the ChEBI ontology links the node "paracetamol" to "carboxamide" and "phenols", and each class back to higher classes, such as organic aromatic compounds (see **Figure 3B**). These graphs are directed acyclic graphs since they are organized hierarchically, are directed, and do not contain cycles. Importantly, in an ontology, molecules can belong to multiple parent classes. The compounds typically found in experimental networks lie as terminal nodes, and the rest of the nodes represent chemical classes. It is also mostly the case for GSMNs, but it is not rare to find nodes corresponding to classes (e.g., "a fatty acid") (Poupin et al., 2020). Chemical ontologies can also integrate other kinds of relationships directly linking molecules, such as tautomers or conjugates (which can create

TABLE 6 | Description of the key characteristics of knowledge networks and graphs.

Main characteristics of knowledge networks and graphs	
Nodes	Chemical compounds and chemical classes
Edges	Class membership (subsumption) and other optional semantic relations
Main Hypothesis	Knowledge can be organized and represented as a network/graph by manual curation from domain experts and semi-automatic class assignments
Limitations	Ontologies representing the same things might still differ and can be mapped to different concepts

cycles in the networks). The ChEBI ontology also links chemical compounds and classes to other concepts: their chemical/biological "roles" (e.g., emulsifier or neurotransmitter) (Degtyarenko et al., 2007). It is important to note that the class hierarchy of chemical ontologies is built manually by domain expert consortia, and the annotation of chemical instances to classes is either done manually or automatically if a class definition can be expressed as a set of formal rules.

Graphs built from ontologies allow detecting related compounds through their belonging to a shared class. Moreover, beyond finding "sibling" compounds, a graph distance between terminal nodes through their most precise common class can be computed to quantify relatedness between any pair of compounds. Such distances based on the ontology's graph structure are a common form of semantic similarity, which found many applications in functional ontologies, such as the Gene Ontology (GO) (Ashburner et al., 2000).

Some specific tools allow fetching the chemical classification of a compound, which can then be used to generate the chemical ontology network of each compound. For example, ClassyFire allows to automatically assign chemical classification based on the compound's structure (e.g., SMILES), using the ChemOnt ontology (Feldman et al., 2005; Djoumbou Feunang et al., 2016). Another tool, CANOPUS, can predict the chemical class based on MS^2 data using ClassyFire and the ChemOnt ontology (Dührkop et al., 2020).

COMBINING NETWORKS ANALYSIS AND MULTI-LAYER NETWORKS

Each of the previously presented networks (both knowledge-based and experimental) represents a different aspect of metabolism. The combination of two or more of such networks brings more comprehensive and informative analysis than a single network, by bringing different angles to the data and combining specific advantages of each network.

For instance, to improve annotations of metabolite features, spectral similarity networks can be combined with different information, such as chemical ontologies or mass difference networks. ChemRICH, for example, is a chemical similarity enrichment analysis that uses Tanimoto chemical similarity and ontologies to associate the metabolic structures from the similarity network with possible metabolic classes in the ontologies network (Barupal and Oliver, 2017). The main benefit of ChemRICH, as compared to classical pathway mapping, is a higher coverage because missing compounds in chemical ontologies can be mapped. Another tool developed to improve metabolite annotation is MolNetEnhancer, which combines molecular networks with chemical ontologies generated by ClassyFire and results from diverse in-silico annotation tools (Ernst et al., 2019). MolNetEnhancer shows great improvement in annotations, even without a prior library match in GNPS. FT-BLAST is a tool that uses fragmentation trees and their comparison to compounds in databases to annotate unknown compounds.

A fragmentation tree illustrates the fragmentation pattern of a compound by representing the molecular formulae of the fragments as nodes, and the neutral losses as edges (Rasche et al., 2012). Note that the in-silico annotation tool CSI:FingerID is also based on fragmentation trees (Dührkop et al., 2015). Moreover, iMet deals with the issue of metabolite annotations that were not present in any database. It uses the spectral similarity and the mass difference of the unknown compounds, and the metabolites present in the databases, in order to find putative neighbor metabolites that show high similarity and that are connected by chemical transformations (Aguilar-Mogas et al., 2017). This way, mass difference networks can be greatly enhanced by combining them with other approaches, such as correlation or spectral similarity networks (Aguilar-Mogas et al., 2017).

To further improve annotation, correlations between the concentration (i.e., abundance) of metabolites that are spectrally similar can be included to analyze metabolomics data. Indeed, it is very likely that, besides having a high spectral similarity, the concentration of metabolites that are connected via biochemical reactions also have a high correlation. Gaquerel et al. utilize in-source fragmentation patterns and correlation networks to improve MetFrag annotation results (Gaquerel et al., 2013). The combination of correlation networks with other metabolic networks can bring new insights into the metabolomics data. For example, Quell et al. demonstrated the potential of combining correlation networks (using Gaussian graphical modeling) with GSMNs and metabolite-gene association networks (derived from genome-wide association studies) to identify unknown metabolites from cohort studies (Quell et al., 2017). However, correlations and associations, in general, emerge due to different mechanisms, so the interpretation is not always straightforward (Steuer, 2006). For example, many associations between metabolite levels (e.g., strong correlations) do not happen between metabolites that are neighbors in the GSMN or that are directly involved in the same metabolic pathways. Analyzing and interpreting the association

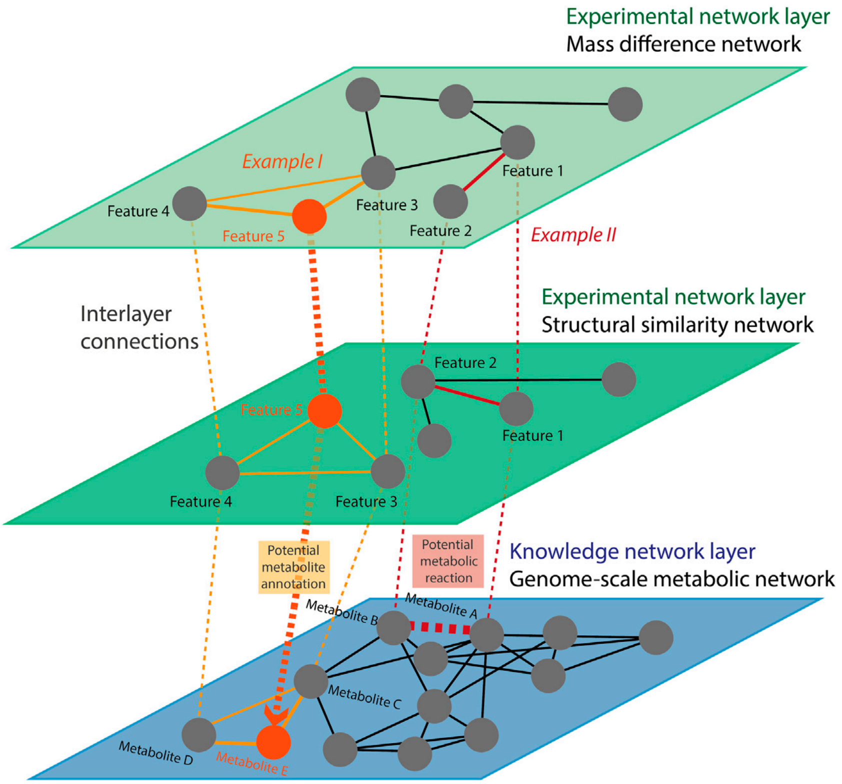


FIGURE 4 | Multi-layer networks principle. Every network (either knowledge-based or experimental) is an independent layer. Common nodes (i.e., identified metabolites) are connected to themselves across the different layers by inter-layer edges. The set of nodes is common in the experimental layers, but we omitted some nodes for the sake of simplicity. The edges of the individual layers and between them can be used, for example, to identify potential metabolite annotations (Example I) and metabolic reactions (Example II). Multi-layer networks allow preserving the topology and organization of each individual network. In Example I, features 3 and 4 were identified as metabolites C and D, respectively. In both experimental layers, these two features are connected with each other and with feature 5. Similarly, in the knowledge-based layer, metabolites C and D are connected with each other and with metabolite E. Therefore, it is likely that feature 5 corresponds to metabolite E. In the same way, features 1 and 2, identified as metabolites A and B, respectively, are connected to each other in the experimental layers but not in the knowledge-based one. In Example II, the metabolite A and B are separated by a mass difference corresponding to known biotransformation (e.g., a phosphatase as in **Figure 2A**) in the layer 1 and are connected by a high structural similarity in layer 2. This represents a potential novel metabolic reaction occurring between metabolites A and B in layer 3.

and correlation networks alongside complementary networks, such as GSMNs, help reduce spurious associations by using the biological knowledge incorporated in GSMNs (Benedetti et al., 2020).

GSMNs can help annotate untargeted metabolomics datasets, as the metabolites and their relationships via metabolic reactions can be analyzed to enhance metabolites' annotations based on the biochemical context (Silva et al., 2014). First, metabolites (and potentially their structures) present in a GSMN represent a knowledge base of the metabolome/lipidome of a given organism. It must be noted that, in the past, GSMNs often lacked detailed structural curation and chemical identifiers, and metabolite names are often rather arbitrary. However, different improvements were suggested and are slowly adopted by the GSMN community (Witting, 2020).

Here, untargeted metabolomics data could be used to help to improve the GSMNs by identifying missing metabolites and filling missing metabolic pathways. For example, metabolites predicted from the WormJam GSMN have been compared against detected metabolites in the nematode Caenorhabditis elegans (C. elegans) in different studies (Salzer and Witting, 2021). Interestingly, the overlap of detected and predicted metabolites was rather modest (less than 40%). Plenty of metabolites beyond the consensus model were found, and structural similarity (based on chemical similarity using Tanimoto distances) has been suggested as an option to identify structurally related molecules (Witting et al., 2018).

Combining experimental network methods with biochemical knowledge-based networks can open new avenues. For example, the recently published tool LINEX allows to analyze lipidomics data by combining lipid metabolic reactions networks analysis with correlation networks (Köhler et al., 2021). With this method, Kohler et al. interpreted the lipidomics correlation networks in the context of biochemical reactions and found new insights on lipid metabolism in three previously published datasets (Köhler et al., 2021). In the same context, MetDNA uses MS/MS similarity networks and metabolic reaction networks. When two metabolites are connected by a reaction in the metabolic reaction network (i.e., when they are neighbor nodes), it is likely that they also show high similarity in the MS/MS similarity network, which can be used to weight their annotation confidence (Shen et al., 2019). By providing a controlled vocabulary, chemical ontologies, such as ChEBI or ChemOnt (Feldman et al., 2005; Degtyarenko et al., 2007), also contribute to the ease of interoperability between networks and data, notably by being frequently referenced in GSMNs and used in many chemical libraries. The controlled vocabulary combined with the distances between the nodes in the ontology offer a useful opportunity for handling partial identification of metabolites in metabolomics data (e.g., in case of lipids (PC(32:1)), since they allow to map such data onto metabolic pathways, using ontology from one specific compound (as identified in the data) to a more generic class (as annotated in the network) (Poupin et al., 2020).

Another approach to combine networks and analyze them could be to construct multi-layer networks. Multi-layer networks are particularly interesting as they allow viewing the metabolism from different but complementary perspectives (one per layer) while keeping the individual features (such as the topology) of each layer (**Figure 4**). Multi-layer networks are a useful approach to bring together multiple networks and interlink information across network types, for example between experimental and knowledge networks. As shown in **Figure 4**, the links between identified metabolites (i.e., nodes with interlayer edges between the experimental layers and the knowledge-based layer, represented as dotted lines) can be used to identify unknown features (**Figure 4**, Example I) or to identify a potential novel metabolic reaction (**Figure 4**, Example II). Multi-layer networks methods are already applied to multi-omics data (Hammoud and Kramer, 2020; Malek et al., 2020), but would benefit metabolomics data analysis by integrating metabolomics experimental and knowledge-based networks.

CONCLUSIONS AND FUTURE DIRECTIONS

Fundamentally, metabolites are the small molecules that are the components of the metabolism. Metabolites are consumed or produced via metabolic reactions mostly driven by biomolecules, such as proteins and genes. In order to study the metabolism and to have a global overview, we can represent the reactions as a network. Current knowledge of metabolism and chemical compounds can be used to generate genome-scale metabolic networks (**Figure 3A**) and ontology-based networks (**Figure 3B**), respectively.

In addition, we can generate other types of networks using experimental data. Indeed, metabolomics data capture different aspects and properties of the chemical compounds that constitute the metabolism. In this review, we described the most common networks that can be built based on the interactions and relationships between the measured compounds. We divided the experimental networks into four types: mass difference networks (**Figure 2A**), adduct and feature networks (**Figure 2B**), structure and MS/MS similarity networks (**Figure 2C**), and correlation networks (**Figure 2D**). The capabilities of those networks to represent the relationships between components are used to annotate and identify metabolites in untargeted MS-based metabolomics data.

In the end, each of the networks described here is useful for specific aspects of metabolomics data analysis and/or interpretation, but they also have limitations. Hence, integrating different networks into multi-layer networks holds great promise to combine all the information and derive new biological insights (**Figure 4**). Particularly, the combination of knowledge-based networks with experimental networks would help to use prior metabolic or chemical knowledge to improve the metabolites' identification and interpretation in biologically relevant contexts.

In the future, with improved metabolite coverage, annotation, and identification, the combination of networks will enable new data analytical approaches. We therefore think that the development of approaches and algorithms for the analysis of metabolomics multi-layer networks will be at the center stage and will gain more and more attention. The multi-layer networks' approach goes beyond mere metabolomics data and will allow integrating multiple omic data (as independent layers), including metabolomics. This will finally enable the analysis of metabolism with a systems biology approach.

AUTHOR CONTRIBUTIONS

The following authors contributed particularly significantly to some specific sections: CF—knowledge-based networks, EN—overall manuscript, LS—MS^2 networks, mass difference, and correlation networks, TN—WGCNA in correlation networks and mass difference networks. The first author AA contributed to the manuscript structure, organization, figures, and across the whole manuscript. The last author MW contributed to the overall paper with significant inputs on writing the intro and experimental networks chapters.

FUNDING

This research was funded by the Agence Nationale de la Recherche (ANR, French National Research Agency)—MetaboHUB, the national metabolomics and fluxomics infrastructure (Grant ANR-INBS-0010), Project number ANR-19-CE45-0021 (MetClassNet) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project number 431572533 (MetClassNet).

REFERENCES

Aguilar-Mogas, A., Sales-Pardo, M., Navarro, M., Guimerà, R., and Yanes, O. (2017). IMet: A Network-Based Computational Tool to Assist in the Annotation of Metabolites from Tandem Mass Spectra. *Anal. Chem.* 89 (6), 3474–3482. doi:10.1021/acs.analchem.6b04512

Altman, T., Travers, M., Kothari, A., Caspi, R., and Karp, P. D. (2013). A Systematic Comparison of the MetaCyc and KEGG Pathway Databases. *BMC Bioinformatics* 14 (March), 112. doi:10.1186/1471-2105-14-112

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* 25 (1), 25–29. doi:10.1038/75556

Bajusz, D., Rácz, A., and Héberger, K. (2015). Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform* 7, 1–13. doi:10.1186/S13321-015-0069-3

Bánky, D., Iván, G., and Grolmusz, V. (2013). Equal Opportunity for Low-Degree Network Nodes: A PageRank-Based Method for Protein Target Identification in Metabolic Graphs. *PLoS ONE* 8 (1), e54204. doi:10.1371/journal.pone.0054204

Barupal, D. K., and Fiehn, O. (2017). Chemical Similarity Enrichment Analysis (ChemRICH) as Alternative to Biochemical Pathway Mapping for Metabolomic Datasets. *Sci. Rep.* 7 (1), 1–11. doi:10.1038/s41598-017-15231-w

Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø., and Herrgard, M. J. (2007). Quantitative Prediction of Cellular Metabolism with Constraint-Based Models: The COBRA Toolbox. *Nat. Protoc.* 2 (3), 727–738. doi:10.1038/nprot.2007.99

Bender, A., and Glen, R. C. (2004). Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* 2 (22), 3204–3218. doi:10.1039/b409813g

Benedetti, E., Pučić-Baković, M., Keser, T., Gerstner, N., Büyüközkan, M., Štambuk, T., et al. (2020). A Strategy to Incorporate Prior Knowledge into Correlation Network Cutoff Selection. *Nat. Commun.* 11 (1), 1–12. doi:10.1038/s41467-020-18675-3

Breitling, R., Ritchie, S., Goodenowe, D., Stewart, M. L., and Barrett, M. P. (2006). Ab Initio Prediction of Metabolic Networks Using Fourier Transform Mass Spectrometry Data. *Metabolomics* 2 (3), 155–164. doi:10.1007/s11306-006-0029-z

Burgess, K. E. V., Borutzki, Y., Rankin, N., Daly, R., and Jourdan, F. (2017). MetaNetter 2: A Cytoscape Plugin for Ab Initio Network Analysis and Metabolite Feature Classification. *J. Chromatogr. B* 1071 (December), 68–74. doi:10.1016/j.jchromb.2017.08.015

Cottret, L., Frainay, C., Chazalviel, M., Cabanettes, F., Gloaguen, Y., Camenen, E., et al. (2018). MetExplore: Collaborative Edition and Exploration of Metabolic Networks. *Nucleic Acids Res.* 46 (W1), W495–W502. doi:10.1093/NAR/GKY301

Cottret, L., and Jourdan, F. (2010). Graph Methods for the Investigation of Metabolic Networks in Parasitology. *Parasitology* 137 (9), 1393–1407. doi:10.1017/S0031182010000363

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., et al. (2007). ChEBI: A Database and Ontology for Chemical Entities of Biological Interest. *Nucleic Acids Res.* 36 (Database), D344–D350. doi:10.1093/nar/gkm791

del Mar Amador, M., Colsch, B., Lamari, F., Jardel, C., Ichou, F., Rastetter, A., et al. (2018). Targeted versus Untargeted Omics - the CAFSA story. *J. Inherit. Metab. Dis.* 41, 447–456. doi:10.1007/S10545-017-0134-3

Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., et al. (2010). The BioPAX Community Standard for Pathway Data Sharing. *Nat. Biotechnol.* 28 (9), 935–942. doi:10.1038/nbt.1666

Demuth, W., Karlovits, M., and Varmuza, K. (2004). Spectral Similarity versus Structural Similarity: Mass Spectrometry. *Analytica Chim. Acta* 516 (1–2), 75–85. doi:10.1016/J.ACA.2004.04.014

DiLeo, M. V., Strahan, G. D., den Bakker, M., and Hoekenga, O. A. (2011). Weighted Correlation Network Analysis (WGCNA) Applied to the Tomato Fruit Metabolome. *PLoS ONE* 6 (10), e26683. doi:10.1371/journal.pone.0026683

Djoumbou Feunang, Y., Eisner, R., Knox, C., Chepelev, L., Hastings, J., Owen, G., et al. (2016). ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy. *J. Cheminform* 8, 1–20. doi:10.1186/S13321-016-0174-Y

Doerfler, H., Sun, X., Wang, L., Engelmeier, D., Lyon, D., and Weckwerth, W. (2014). MzGroupAnalyzer-Predicting Pathways and Novel Chemical Structures from Untargeted High-Throughput Metabolomics Data. *PLOS ONE* 9 (5), e96188. doi:10.1371/JOURNAL.PONE.0096188

Dührkop, K., Nothias, L.-F., Fleischauer, M., Reher, R., Ludwig, M., Hoffmann, M. A., et al. (2020). Systematic Classification of Unknown Metabolites Using High-Resolution Fragmentation Mass Spectra. *Nat. Biotechnol.* 39, 462–471. doi:10.1038/s41587-020-0740-8

Dührkop, K., Shen, H., Meusel, M., Rousu, J., and Böcker, S. (2015). Searching Molecular Structure Databases with Tandem Mass Spectra Using CSI:FingerID. *Proc. Natl. Acad. Sci. USA* 112 (41), 12580–12585. doi:10.1073/PNAS.1509788112

Emwas, A.-H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G. A. N., et al. (2019). NMR Spectroscopy for Metabolomics Research. *Metabolites* 9 (7), 123. doi:10.3390/METABO9070123

Ernst, M., Kang, K. B., Caraballo-Rodríguez, A. M., Nothias, L.-F., Wandy, J., Chen, C., et al. (2019). MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. *Metabolites* 9 (7), 144. doi:10.3390/metabo9070144

Faust, K., Dupont, P., Callut, J., and van Helden, J. (2010). Pathway Discovery in Metabolic Networks by Subgraph Extraction. *Bioinformatics* 26 (9), 1211–1218. doi:10.1093/BIOINFORMATICS/BTQ105

Feldman, H. J., Dumontier, M., Ling, S., Haider, N., Hogue, C. W. V., and Hogue, V. (2005). CO: A Chemical Ontology for Identification of Functional Groups and Semantic Comparison of Small Molecules. *FEBS Lett.* 579 (21), 4685–4691. doi:10.1016/J.FEBSLET.2005.07.039

Fiehn, O., Robertson, D., Griffin, J., van der Werf, M., Nikolau, B., Morrison, N., et al. (2007). The Metabolomics Standards Initiative (MSI). *Metabolomics* 3 (3), 175–178. doi:10.1007/s11306-007-0070-6

Frainay, C., Aros, S., Chazalviel, M., Garcia, T., Vinson, F., Weiss, N., et al. (2019). MetaboRank: Network-Based Recommendation System to Interpret and Enrich Metabolomics Results. *Bioinformatics* 35 (2), 274. Alfonso Valencia. doi:10.1093/bioinformatics/bty577

Frainay, C., and Jourdan, F. (2017). Computational Methods to Identify Metabolic Sub-networks Based on Metabolomic Profiles. *Brief Bioinform* 18 (1), 43–56. doi:10.1093/bib/bbv115

Frainay, C., Schymanski, E., Neumann, S., Merlet, B., Salek, R., Jourdan, F., et al. (2018). Mind the Gap: Mapping Mass Spectral Databases in Genome-Scale Metabolic Networks Reveals Poorly Covered Areas. *Metabolites* 8 (3), 51. doi:10.3390/metabo8030051

Gaquerel, E., Kuhl, C., and Neumann, S. (2013). Computational Annotation of Plant Metabolomics Profiles via a Novel Network-Assisted Approach. *Metabolomics* 9 (4), 904–918. doi:10.1007/S11306-013-0504-2

Gathungu, R. M., Larrea, P., Sniatynski, M. J., Marur, V. R., Bowden, J. A., Koelmel, J. P., et al. (2018). Optimization of Electrospray Ionization Source Parameters for Lipidomics to Reduce Misannotation of In-Source Fragments as Precursor Ions. *Anal. Chem.* 90 (22), 13523–13532. doi:10.1021/ACS.ANALCHEM.8B03436/SUPPL_FILE/AC8B03436_SI_002.XLS

Guo, J., Shen, S., Xing, S., Yu, H., and Huan, T. (2021). ISFrag: De Novo Recognition of In-Source Fragments for Liquid Chromatography-Mass Spectrometry Data. *Anal. Chem.* 93 (29), 10243–10250. doi:10.1021/ACS.ANALCHEM.1C01644

Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A., and Hatzimanikatis, V. (2016). ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* 5 (10), 1155–1166. doi:10.1021/ACSSYNBIO.6B00054

Hammoud, Z., and Kramer, F. (2020). Multilayer Networks: Aspects, Implementations, and Application in Biomedicine. *Big Data Anal.* 5 (1), 1–18. doi:10.1186/S41044-020-00046-0

Huber, F., Ridder, L., Verhoeven, S., Spaaks, J. H., Diblen, F., Rogers, S., et al. (2021). Spec2Vec: Improved Mass Spectral Similarity Scoring through Learning of Structural Relationships. *Plos Comput. Biol.* 17 (2), e1008724. doi:10.1371/journal.pcbi.1008724

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., et al. (2003). The Systems Biology Markup Language (SBML): A Medium

Forrepresentation and Exchange of Biochemical Network Models. *Bioinformatics* 19 (4), 524–531. doi:10.1093/BIOINFORMATICS/BTG015

Jeffryes, J. G., Colastani, R. L., Elbadawi-Sidhu, M., Kind, T., Niehaus, T. D., Broadbelt, L. J., et al. (2015). MINEs: Open Access Databases of Computationally Predicted Enzyme Promiscuity Products for Untargeted Metabolomics. *J. Cheminform* 7, 1–8. doi:10.1186/S13321-015-0087-1

Jourdan, F., Breitling, R., Barrett, M. P., and Gilbert, D. (2008). MetaNetter: Inference and Visualization of High-Resolution Metabolomic Networks. *Bioinformatics* 24 (1), 143–145. doi:10.1093/bioinformatics/btm536

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* 45 (D1), D353–D361. doi:10.1093/NAR/GKW1092

Kell, D. B., and Goodacre, R. (2014). Metabolomics and Systems Pharmacology: Why and How to Model the Human Metabolic Network for Drug Discovery. *Drug Discov. Today* 19 (2), 171–182. doi:10.1016/j.drudis.2013.07.014

Köhler, N., Rose, T. D., Falk, L., and Pauling, J. K. (2021). Investigating Global Lipidome Alterations with the Lipid Network Explorer. *Metabolites* 11 (8), 488. doi:10.3390/METABO11080488

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F. J. (2011). Gaussian Graphical Modeling Reconstructs Pathway Reactions from High-Throughput Metabolomics Data. *BMC Syst. Biol.* 5, 1–16. doi:10.1186/1752-0509-5-21

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., and Neumann, S. (2012). CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* 84 (1), 283–289. doi:10.1021/AC202450G/SUPPL_FILE/AC202450G_SI_001.PDF

Lacroix, V., Cottret, L., Thébault, P., and Sagot, M.-F. (2008). An Introduction to Metabolic Networks and Their Structural Analysis. *Ieee/acm Trans. Comput. Biol. Bioinf.* 5 (4), 594–617. doi:10.1109/TCBB.2008.79

Langfelder, P., and Horvath, S. (2008). WGCNA: An R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* 9, 1–13. doi:10.1186/1471-2105-9-559

Liggi, S., and Griffin, J. L. (2017). Metabolomics Applied to Diabetes−lessons from Human Population Studies. *Int. J. Biochem. Cel Biol.* 93 (December), 136–147. doi:10.1016/J.BIOCEL.2017.10.011

Loos, M., and Singer, H. (2017). Nontargeted Homologue Series Extraction from Hyphenated High Resolution Mass Spectrometry Data. *J. Cheminform* 9 (1), 12. doi:10.1186/s13321-017-0197-z

Ma, H.-W., and Zeng, A.-P. (2003). The Connectivity Structure, Giant Strong Component and Centrality of Metabolic Networks. *Bioinformatics* 19 (11), 1423–1430. doi:10.1093/BIOINFORMATICS/BTG177

Malek, M., Zorzan, S., and Ghoniem, M. (2020). A Methodology for Multilayer Networks Analysis in the Context of Open and Private Data: Biological Application. *Appl. Netw. Sci.* 5 (1), 1–28. doi:10.1007/S41109-020-00277-Z/TABLES/21

Mazurie, A., Bonchev, D., Schwikowski, B., and Buck, G. A. (2010). Evolution of Metabolic Network Organization. *BMC Syst. Biol.* 4 (1), 59–10. doi:10.1186/1752-0509-4-59/FIGURES/2

Misra, B. B. (2021). New Software Tools, Databases, and Resources in Metabolomics: Updates from 2020. *Metabolomics* 17 (5), 49. doi:10.1007/s11306-021-01796-1

Misra, B. B., and van der Hooft, J. J. J. (2016). Updates in Metabolomics Tools and Resources: 2014-2015. *Electrophoresis* 37 (1), 86–110. doi:10.1002/elps.201500417

Moretti, S., Tran, V. D. T., Mehl, F., Ibberson, M., Pagni, M., and Pagni, M. (2021). MetaNetX/MNXref: Unified Namespace for Metabolites and Biochemical Reactions in the Context of Metabolic Models. *Nucleic Acids Res.* 49, D570–D574. doi:10.1093/nar/gkaa992

Moritz, F., Kaling, M., Schnitzler, J.-P., and Schmitt-Kopplin, P. (2017). Characterization of poplar Metabotypes via Mass Difference Enrichment Analysis. *Plant Cel Environ.* 40, 1057–1073. doi:10.1111/pce.12878

Morreel, K., Saeys, Y., Dima, O., Lu, F., Van de Peer, Y., Vanholme, R., et al. (2014). Systematic Structural Characterization of Metabolites in Arabidopsis via Candidate Substrate-Product Pair Networks. *The Plant Cell* 26 (3), 929–945. doi:10.1105/TPC.113.122242

Murga-Garrido, S. M., Hong, Q., Cross, T.-W. L., Hutchison, E. R., Han, J., Thomas, S. P., et al. (2021). Gut Microbiome Variation Modulates the Effects of Dietary Fiber on Host Metabolism. *Microbiome* 9 (1), 117. doi:10.1186/S40168-021-01061-6

Murphy, R. C., James, P. F., McAnoy, A. M., Krank, J., Duchoslav, E., and Barkley, R. M. (2007). Detection of the Abundance of Diacylglycerol and Triacylglycerol Molecular Species in Cells Using Neutral Loss Mass Spectrometry. *Anal. Biochem.* 366 (1), 59–70. doi:10.1016/J.AB.2007.03.012

Naake, T., and Fernie, A. R. (2019). MetNet: Metabolite Network Prediction from High-Resolution Mass Spectrometry Data in R Aiding Metabolite Annotation. *Anal. Chem.* 91 (3), 1768–1772. doi:10.1021/acs.analchem.8b04096

Notebaart, R. A., Szappanos, B., Kintses, B., Pal, F., Gyorkei, A., Bogos, B., et al. (2014). Network-Level Architecture and the Evolutionary Potential of Underground Metabolism. *Proc. Natl. Acad. Sci.* 111 (32), 11762–11767. doi:10.1073/pnas.1406102111

Nothias, L.-F., Petras, D., Schmid, R., Dührkop, K., Rainer, J., Sarvepalli, A., et al. (2020). Feature-Based Molecular Networking in the GNPS Analysis Environment. *Nat. Methods* 17 (9), 905–908. doi:10.1038/s41592-020-0933-6

O'Brien, E. J., Monk, J. M., and Palsson, B. O. (2015). Using Genome-Scale Models to Predict Biological Capabilities. *Cell* 161 (5), 971–987. doi:10.1016/j.cell.2015.05.019

Olivon, F., Elie, N., Grelier, G., Roussi, F., Litaudon, M., and Touboul, D. (2018). MetGem Software for the Generation of Molecular Networks Based on the T-SNE Algorithm. *Anal. Chem.* 90 (23), 13900–13908. doi:10.1021/acs.analchem.8b03099

Osterhoff, M., Frahnow, T., Seltmann, A., Mosig, A., Neunübel, K., Sales, S., et al. (2014). Identification of Gene-Networks Associated with Specific Lipid Metabolites by Weighted Gene Co-expression Network Analysis (WGCNA). *Exp. Clin. Endocrinol. Diabetes* 122 (03), P098. doi:10.1055/S-0034-1372115

Pan, S., and Reed, J. L. (2018). Advances in Gap-Filling Genome-Scale Metabolic Models and Model-Driven Experiments Lead to Novel Metabolic Discoveries. *Curr. Opin. Biotechnol.* 51, 103–108. Elsevier Ltd. doi:10.1016/j.copbio.2017.12.012

Pedersen, H. K., Forslund, S. K., Gudmundsdottir, V., Petersen, A. Ø., Hildebrand, F., Hyötyläinen, T., et al. (2018). A Computational Framework to Integrate High-Throughput '-omics' Datasets for the Identification of Potential Mechanistic linksA Computational Framework to Integrate High-Throughput '-Omics' Datasets for the Identification of Potential Mechanistic Links. *Nat. Protoc.* 13 (12), 2781–2800. doi:10.1038/s41596-018-0064-z

Perez De Souza, L., Alseekh, S., Brotman, Y., and FernieFernie, A. R. (2020). Network-Based Strategies in Metabolomics Data Analysis and Interpretation: From Molecular Networking to Biological Interpretation. *Expert Rev. Proteomics* 17 (4), 243–255. doi:10.1080/14789450.2020.1766975

Petersen, C., DaiDai, D. L. Y., Boutin, R. C. T., Sbihi, H., Sears, M. R., Moraes, T. J., et al. (2021). A Rich Meconium Metabolome in Human Infants Is Associated with Early-Life Gut Microbiota Composition and Reduced Allergic Sensitization. *Cel Rep. Med.* 2, 100260. doi:10.1016/j.xcrm.2021.100260

Poupin, N., Vinson, F., Moreau, A., Batut, A., Chazalviel, M., Colsch, B., et al. (2020). Improving Lipid Mapping in Genome Scale Metabolic Networks Using Ontologies. *Metabolomics* 16 (4), 44–11. doi:10.1007/S11306-020-01663-5/FIGURES/6

Quell, J. D., Römisch-Margl, W., Colombo, M., Krumsiek, J., Evans, A. M., Mohney, R., et al. (2017). Automated Pathway and Reaction Prediction Facilitates In Silico Identification of Unknown Metabolites in Human Cohort Studies. *J. Chromatogr. B* 1071 (December), 58–67. doi:10.1016/j.jchromb.2017.04.002

Rasche, F., Scheubert, K., Hufsky, F., Zichner, T., Kai, M., Svatoš, A., et al. (2012). Identifying the Unknowns by Aligning Fragmentation Trees. *Anal. Chem.* 84 (7), 3417–3426. doi:10.1021/AC300304U

Robinson, J. L., Kocabaş, P., Wang, H., Cholley, P.-E., Cook, D., Nilsson, A., et al. (2020). An Atlas of Human Metabolism. *Sci. Signal.* 13 (624), eaaz1482. doi:10.1126/SCISIGNAL.AAZ1482

Rosato, A., Tenori, L., Cascante, M., Saccenti, E., Martins dos Santos, P., and Saccenti, E. (2018). From Correlation to Causation: Analysis of Metabolomics Data Using Systems Biology Approaches. *Metabolomics* 14 (4), 1–20. doi:10.1007/S11306-018-1335-Y

Ruttkies, C., Neumann, S., and Posch, S. (2019). Improving MetFrag with Statistical Learning of Fragment Annotations. *BMC Bioinformatics* 20 (1), 1–14. doi:10.1186/S12859-019-2954-7

Salzer, L., and Witting, M. (2021). Quo Vadis Caenorhabditis elegans Metabolomics-A Review of Current Methods and Applications to Explore Metabolism in the Nematode. *Metabolites* 202111 (5), 284. doi:10.3390/METABO11050284

Samal, A., and Martin, O. C. (2011). Randomizing Genome-Scale Metabolic Networks. *PLOS ONE* 6 (7), e22295. doi:10.1371/JOURNAL.PONE.0022295

Schmid, R., Petras, D., Nothias, L.-F., Wang, M., Aron, A. T., Jagels, A., et al. (2021). Ion Identity Molecular Networking for Mass Spectrometry-Based Metabolomics in the GNPS Environment. *Nat. Commun.* 12, 1–12. doi:10.1038/s41467-021-23953-9

Schollée, J. E., Schymanski, E. L., Stravs, M. A., Gulde, R., Thomaidis, N. S., and Hollender, J. (2017). Similarity of High-Resolution Tandem Mass Spectrometry Spectra of Structurally Related Micropollutants and Transformation Products. *J. Am. Soc. Mass. Spectrom.* 28 (12), 2692–2704. doi:10.1007/S13361-017-1797-6

Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D., and McLean, J. A. (2016). Untargeted Metabolomics Strategies-Challenges and Emerging Directions. *J. Am. Soc. Mass. Spectrom.* 27 (12), 1897–1905. doi:10.1007/S13361-016-1469-Y

Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., et al. (2014). Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* 48 (4), 2097–2098. doi:10.1021/ES5002105

Senan, O., Aguilar-Mogas, A., Navarro, M., Capellades, J., Noon, L., Burks, D., et al. (2019). CliqueMS: a Computational Tool for Annotating In-Source Metabolite Ions from LC-MS Untargeted Metabolomics Data Based on a Coelution Similarity Network. *Bioinformatics* 35 (20), 4089–4097. doi:10.1093/bioinformatics/btz207

Shen, X., Wang, R., Xiong, X., Yin, Y., Cai, Y., Ma, Z., et al. (2019). Metabolic Reaction Network-Based Recursive Metabolite Annotation for Untargeted Metabolomics. *Nat. Commun.* 10 (1), 1–14. doi:10.1038/s41467-019-09550-x

Silva, R. R., Jourdan, F., Salvanha, D. M., Letisse, F., Jamin, E. L., Guidetti-Gonzalez, S., et al. (2014). ProbMetab: an R Package for Bayesian Probabilistic Annotation of LC-MS-based Metabolomics. *Bioinformatics* 30 (9), 1336–1337. doi:10.1093/BIOINFORMATICS/BTU019

Spicer, R., Salek, R. M., Moreno, P., Cañueto, D., and Steinbeck, C. (2017). Navigating Freely-Available Software Tools for Metabolomics Analysis. *Metabolomics* 13 (9), 106. doi:10.1007/s11306-017-1242-7

Steuer, R. (2006). Review: On the Analysis and Interpretation of Correlations in Metabolomic Data. *Brief. Bioinform.* 7 (2), 151–158. doi:10.1093/bib/bbl009

Stobbe, M. D., HoutenHouten, S. M., Kampen, A. H. C., Wanders, R. J. A., and Moerland, P. D. (2012). Improving the Description of Metabolic Networks: The TCA Cycle as Example. *FASEB j.* 26 (9), 3625–3636. doi:10.1096/FJ.11-203091

Strömbäck, L., and Lambrix, P. (2005). Representations of Molecular Pathways: An Evaluation of SBML, PSI MI and BioPAX. *Bioinformatics* 21 (24), 4401–4407. doi:10.1093/BIOINFORMATICS/BTI718

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, T. W.-M., et al. (2007). Proposed Minimum Reporting Standards for Chemical Analysis. *Metabolomics* 3 (3), 211–221. doi:10.1007/s11306-007-0082-2

Thiele, I., and Palsson, B. Ø. (2010). A Protocol for Generating a High-Quality Genome-Scale Metabolic Reconstruction. *Nat. Protoc.* 5 (1), 93–121. doi:10.1038/nprot.2009.203

Thiele, I., Vlassis, N., and Fleming, R. M. T. (2014). FastGapFill: Efficient Gap Filling in Metabolic Networks. *Bioinformatics (Oxford, England)* 30 (17), 2529–2531. doi:10.1093/bioinformatics/btu321

Tziotis, D., Hertkorn, N., and Schmitt-Kopplin, P. (2011). Kendrick-analogous Network Visualisation of Ion Cyclotron Resonance Fourier Transform Mass Spectra: Improved Options for the Assignment of Elemental Compositions and the Classification of Organic Molecular Complexity. *Eur. J. Mass. Spectrom. (Chichester)* 17 (4), 415–421. doi:10.1255/EJMS.1135

Vernocchi, P., Gili, T., Conte, F., Del Chierico, F., Conta, G., Miccheli, A., et al. (2020). Network Analysis of Gut Microbiome and Metabolome to Discover Microbiota-Linked Biomarkers in Patients Affected by Non-small Cell Lung Cancer. *Ijms* 21 (22), 8730. doi:10.3390/ijms21228730

Wang, M., CarverCarver, J. J., Phelan, L. M., Garg, N., Peng, Y., Nguyen, D. D., et al. (2016). Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* 34, 828–837. doi:10.1038/nbt.3597

Watrous, J., Roach, P., Alexandrov, T., Heath, B. S., Yang, J. Y., Kersten, R. D., et al. (2012). Mass Spectral Molecular Networking of Living Microbial Colonies. *Proc. Natl. Acad. Sci.* 109 (26), E1743–E1752. doi:10.1073/PNAS.1203689109

Witting, M., Hastings, J., Rodriguez, N., Joshi, C. J., HattwellHattwell, J. P. N., Ebert, P. R., et al. (2018). Modeling Meets Metabolomics-The WormJam Consensus Model as Basis for Metabolic Studies in the Model Organism Caenorhabditis elegans. *Front. Mol. Biosci.* 5 (NOV), 96. doi:10.3389/FMOLB.2018.00096

Witting, M. (2020). Suggestions for Standardized Identifiers for Fatty Acyl Compounds in Genome Scale Metabolic Models and Their Application to the WormJam Caenorhabditis Elegans Model. *Metabolites* 10, 130. doi:10.3390/METABO10040130

Wu, J., Ye, Y., Quan, J., Ding, R., Wang, X., Zhuang, Z., et al. (2021). Using Nontargeted LC-MS Metabolomics to Identify the Association of Biomarkers in Pig Feces with Feed Efficiency. *Porc Health Manag.* 7, 1–10. doi:10.1186/S40813-021-00219-W

Xing, S., Hu, Y., Yin, Z., Liu, M., Tang, X., Fang, M., et al. (2020). Retrieving and Utilizing Hypothetical Neutral Losses from Tandem Mass Spectra for Spectral Similarity Analysis and Unknown Metabolite Annotation. *Anal. Chem.* 92 (21), 14476–14483. doi:10.1021/ACS.ANALCHEM.0C02521

Zhang, B., and Horvath, S. (2005). A General Framework for Weighted Gene Co-expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* 4 (1), Article17. doi:10.2202/1544-6115.1128

Author Disclaimer: Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

Conflict of Interest: Author RS is employed by Bruker BioSpin GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

BRIEF RESEARCH REPORT
published: 05 April 2022
doi: 10.3389/fmolb.2022.839809

Serum NMR Profiling Reveals Differential Alterations in the Lipoproteome Induced by Pfizer-BioNTech Vaccine in COVID-19 Recovered Subjects and Naïve Subjects

Veronica Ghini[1,2,3], *Laura Maggi*[4], *Alessio Mazzoni*[4], *Michele Spinicci*[4,5], *Lorenzo Zammarchi*[4,5], *Alessandro Bartoloni*[4,5], *Francesco Annunziato*[4,6] and *Paola Turano*[1,2,3]*

[1]*Department of Chemistry, University of Florence, Florence, Italy,* [2]*Magnetic Resonance Center (CERM), University of Florence, Florence, Italy,* [3]*Consorzio Interuniversitario Risonanze Magnetiche di Metallo Proteine (CIRMMP), Florence, Italy,* [4]*Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy,* [5]*Infectious and Tropical Disease Unit, Careggi University Hospital, Florence, Italy,* [6]*Flow Cytometry Diagnostic Center and Immunotherapy, Careggi University Hospital, Florence, Italy*

OPEN ACCESS

Edited by:
Wolfram Weckwerth,
University of Vienna, Austria

Reviewed by:
Chi-Fon Chang,
Genomics Research Center,
Academia Sinica, Taiwan
Sofia Moco,
VU Amsterdam, Netherlands

Correspondence:
Paola Turano
turano@cerm.unifi.it

Specialty section:
This article was submitted to Metabolomics,
a section of the journal
Frontiers in Molecular Biosciences

Received: *20 December 2021*
Accepted: *07 March 2022*
Published: *05 April 2022*

Citation:
Ghini V, Maggi L, Mazzoni A, Spinicci M, Zammarchi L, Bartoloni A, Annunziato F and Turano P (2022) Serum NMR Profiling Reveals Differential Alterations in the Lipoproteome Induced by Pfizer-BioNTech Vaccine in COVID-19 Recovered Subjects and Naïve Subjects.
Front. Mol. Biosci. 9:839809.
doi: 10.3389/fmolb.2022.839809

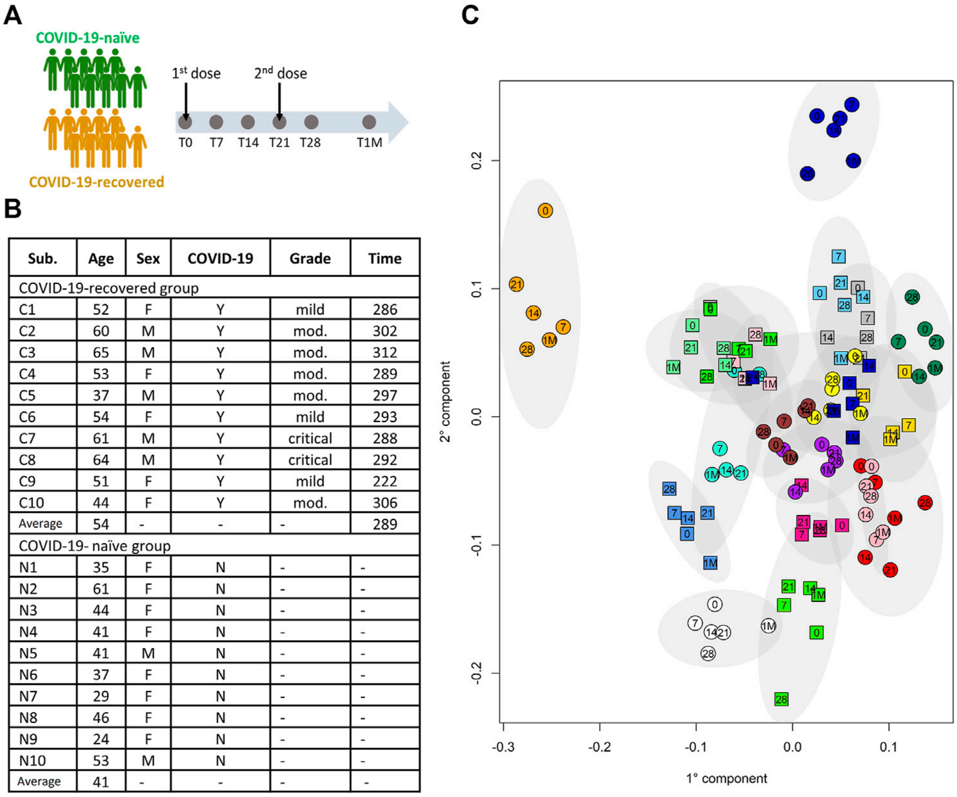
^{1}H NMR spectra of sera have been used to define the changes induced by vaccination with Pfizer-BioNTech vaccine (2 shots, 21 days apart) in 10 COVID-19-recovered subjects and 10 COVID-19-naïve subjects at different time points, starting from before vaccination, then weekly until 7 days after second injection, and finally 1 month after the second dose. The data show that vaccination does not induce any significant variation in the metabolome, whereas it causes changes at the level of lipoproteins. The effects are different in the COVID-19-recovered subjects with respect to the naïve subjects, suggesting that a previous infection reduces the vaccine modulation of the lipoproteome composition.

Keywords: SARS-CoV-2, vaccine, NMR, metabolomics, lipoproteins

INTRODUCTION

While health systems worldwide race to vaccinate people against SARS-CoV-2, several studies have appeared where the measured levels of antibodies in the blood before vaccination and then after each of the two vaccine doses (Ebinger et al., 2021, 2; Mazzoni et al., 2021, 19). These studies have highlighted different response in COVID-19-recovered or naïve subjects in terms of antibody levels, which is the most relevant information for the design and implementations of efficient mass vaccination campaigns in the context of COVID-19 emergency. One of the main outcomes of such studies in mRNA vaccines indicates that subjects who previously had COVID-19 get a strong immune response from a single dose (Levi et al., 2021; Mazzoni et al., 2021).

^{1}H nuclear magnetic resonance (NMR) spectroscopy analysis of biofluids produces profiles that show characteristic responses to changes in physiological status and has been used in a few studies in the past to monitor changes in urinary metabolite levels in mice administered different types of influenza vaccines (Sasaki et al., 2019) or to identify serum markers predictive of adverse reactions against smallpox (McClenathan et al., 2017) as well as metabolic signatures of responses induced by a series of commonly used human vaccines, as reviewed in (Diray-Arce et al., 2020). On the other



Sub.	Age	Sex	COVID-19	Grade	Time
COVID-19-recovered group					
C1	52	F	Y	mild	286
C2	60	M	Y	mod.	302
C3	65	M	Y	mod.	312
C4	53	F	Y	mod.	289
C5	37	M	Y	mod.	297
C6	54	F	Y	mild	293
C7	61	M	Y	critical	288
C8	64	M	Y	critical	292
C9	51	F	Y	mild	222
C10	44	F	Y	mod.	306
Average	54	-	-	-	289
COVID-19- naïve group					
N1	35	F	N	-	-
N2	61	F	N	-	-
N3	44	F	N	-	-
N4	41	F	N	-	-
N5	41	M	N	-	-
N6	37	F	N	-	-
N7	29	F	N	-	-
N8	46	F	N	-	-
N9	24	F	N	-	-
N10	53	M	N	-	-
Average	41	-	-	-	-

FIGURE 1 | (A) Schematic representation of the study design. **(B)** Table summarizing the main demographic characteristics of the subjects included in the study; the COVID-19-recovered and - naïve subjects are indicated with C and N, respectively; for the COVID-19-recovered group the column "Grade" refers to the grade of the disease severity, i.e. mild, moderate (mod.) or critical; the column "Time" refers to the time (in days) from COVID-19 diagnosis to the first dose of vaccine. **(C)** Individual metabolic phenotype as it results from a PCA-CA score plot (binned NOESY spectra). Each color represents a different subject; squares: COVID-19-naïve; circles: COVID-19-recovered. Numbers indicate the collection time: T0 = 0, T7 = 7, T14 = 14, T21 = 21, T28 = 28, T1M = 1M.

hand, ^{1}H NMR has been also successfully used to monitor changes in metabolites and lipoproteins induced by SARS-CoV-2 infection (Bruzzone et al., 2020; Kimhofer et al., 2020; Ballout et al., 2021; Baranovicova et al., 2021; Bizkarguenaga et al., 2021; Julkunen et al., 2021; Lodge et al., 2021; Masuda et al., 2021; Meoni et al., 2021).

Here, we monitored the time-dependent response to the mRNA Pfizer-BioNTech vaccine in a cohort of 20 healthcare workers, 10 of them had a previous history of COVID-19 and 10 were COVID-19 naïve. All of them received two doses, 21 days apart. The NMR spectra of serum samples collected at six different time points were analyzed to monitor time-dependent intra-individual changes induced by vaccination and to explore possible differences between individual previously infected with COVID-19 and individuals without prior infection. While no significant differences between the two groups exist before vaccination, the first dose is sufficient to induce changes in the lipoproteins levels (but not in metabolites), whose size and nature depends upon absence or presence of previous infection. Differences between the two groups of individuals are maintained along the monitored timeline. The second dose is essentially inconsequential in the group of COVID-19-recovered subjects.

MATERIAL AND METHODS

Study Design

The study was conducted at the beginning of the Italian vaccination campaign against COVID-19 using the Pfizer-BioNTech mRNA vaccine (January-February 2021). Twenty Caucasian healthcare workers of the Careggi University Hospital of Florence were recruited, 10 of them had a previous history of COVID-19 (hereafter called "COVID-19-recovered"), and 10 were COVID-19 naïve ("COVID-19-naïve") (**Figure 1A**). The main features of the cohort are provided in **Figure 1B**. The COVID-19-recovered subjects have been infected in the period March-April 2020, with the Wuhan strain; they recovered from the disease on average 255 days before vaccination (range 208–280 days). The inclusion/exclusion criteria were those used for Pfizer-BioNTech vaccine administration for healthcare workers.

The study was conducted in accordance with the Declaration of Helsinki. The study was approved by the Careggi University Hospital Ethical Committee (n. 19466_spe). Written informed consent was obtained from recruited subjects.

For all subjects blood serum samples were collected at six different time points: before the first dose (T0); 7 and 14 days after the first dose (T7 and T14, respectively); 21 days after the first dose, just before the second dose (T21); 28 days after the first dose and 7 days after the second dose (T28); 1 month after the second dose (T1M) (**Figure 1A**). Blood samples were collected (4 h after breakfast) in a BD vacutainer clot-activator tube for serum collection and processed within 1 hour from sample collection. After processing, all the serum samples were immediately stored at −30°C until NMR analysis (February-March 2021).

NMR Sample Preparation and Data Acquisition

NMR samples were prepared according to standard procedures (Takis et al., 2019; Vignoli et al., 2019). Frozen serum samples were thawed at room temperature. A total of 350 µl of sodium phosphate buffer (70 mM Na_2HPO_4; 20% (v/v) ^2H_2O; 6.1 mM NaN_3, 4.6 mM sodium trimethylsilyl [2,2,3,3−$^2H_4$] propionate (TMSP), pH 7.4) was added to 350 µl of each serum sample; the mixture was homogenized by vortexing for 30 s. A total of 600 µl of each mixture was transferred into a 5.00 mm NMR tube (Bruker BioSpin) for the analysis. ^{1}H-NMR spectra were acquired using a Bruker 600 MHz spectrometer (Bruker BioSpin) operating at 600.13 MHz proton Larmor frequency and equipped with a 5 mm PATXI ^1H–^{13}C–^{15}N and ^{2}H-decoupling probe including a *z* axis gradient coil, an automatic tuning-matching (ATM) and an automatic and refrigerated sample changer (SampleJet, Bruker BioSpin). A BTO 2000 thermocouple served for temperature stabilization at the level of approximately 0.1 K at the sample. Before measurement, samples were kept for 5 min inside the NMR probe head, for temperature equilibration at 310 K.

For each serum sample, three one-dimensional (1D) ^{1}H NMR spectra were acquired with water peak suppression and different pulse sequences that allowed the selective observation of different molecular components: 1) a standard NOESY 1Dpresat (noesygppr1d.comp; Bruker BioSpin) pulse sequence (using 32 scans, 98,304 data points, a spectral width of 18,028 Hz, an acquisition time of 2.7 s, a relaxation delay of 4 s and a mixing time of 0.01 s); 2) a standard CPMG (cpmgpr1d.comp; Bruker BioSpin) pulse sequence (using 32 scans, 73,728 data points, a spectral width of 12,019 Hz and a relaxation delay of 4 s); 3) a standard diffusion-edited (ledbgppr2s1d.comp; Bruker BioSpin) pulse sequence (using 32 scans, 98,304 data points, a spectral width of 18,028 Hz and a relaxation delay of 4 s). All spectra were recorded at the Magnetic Resonance Center of the University of Florence (CERM).

Free induction decays were multiplied by an exponential function equivalent to a 0.3 Hz line-broadening factor before applying Fourier transform. Transformed spectra were automatically corrected for phase and baseline distortions and calibrated (glucose doublet at δ 5.24 ppm) using TopSpin 3.5 (Bruker BioSpin).

Assignment and Quantification

The metabolites, whose peaks in the NMR spectra were well defined and resolved, were assigned and their concentrations determined; the assignment procedure was performed using an ^{1}H NMR spectra library of pure organic compounds (BBIOREFCODE, Bruker BioSpin). The concentrations of 22 metabolites (**Supplementary Table S1**) were analysed using *In Vitro* Diagnostics research (IVDr) B.I.-Quant PS tool (Bruker, BioSpin). One hundred fourteen components associated to lipoprotein main parameters, i.e. triglycerides (TG), bound and free cholesterol (Chol and Free Chol), phospholipids (PL), apolipoproteins A1, A2 and B100 (ApoA1, ApoA2 and ApoB100) in each of the main lipoprotein classes, i.e. very low-density lipoproteins (VLDL), high-density lipoproteins (HDL), intermediate-density lipoproteins (IDL), and low-density lipoproteins (LDL) and in their respective subfractions were also analysed (**Supplementary Table S2**) through the IVDr Lipoprotein Subclass Analysis B.I.-LISA tool (Bruker, BioSpin) (Jiménez et al., 2018).

Statistical Analysis

All data analyses were performed using the “R” software. Multivariate analyses were applied on NOESY binned spectra. To this aim, each spectrum in the region 10.00–0.2 ppm was divided into 0.02 ppm chemical shift bins, and the corresponding spectral areas were integrated using the AMIX software. The area of each bin was normalized to the total spectral area, calculated with exclusion of the water region (4.50–5.00 ppm). Principal component analysis (PCA) was used as unsupervised exploratory analysis to obtain an overview of the data to detect the presence of clusters (function *prcomp*); canonical analysis (CA) was used in combination with PCA to increase the supervised separation among individuals (in house developed script) and to define their individual metabolomic fingerprint (Assfalg et al., 2008; Bernini et al., 2009). The global accuracy for classification was assessed by means of a Monte Carlo cross-validation scheme.

For univariate analyses, the non-parametric Wilcoxon-Mann-Whitney test was used to infer differences between the metabolite/lipoprotein levels in the comparison between COVID-19-recovered group and COVID-19-naïve group. Instead, for pairwise comparison within each group, the paired Wilcoxon signed-rank test was used to analyzed the differences between the samples of a given individual at each time point with respect to T0 (Neuhäuser, 2011).

RESULTS

It is known that the NMR detectable part of the blood metabolome/lipoproteome contains a strong signature that defines the individual metabolic phenotype that, in the absence of pathophysiological perturbations, remains stable over a time span of the order of years (Holmes et al., 2008; Yousri et al., 2014; Ghini et al., 2015). The distribution of the metabolic phenotype of the 20 subjects under study is shown in **Figure 1C**. Notably, we don’t observe any clustering in the metabolic space of the samples from COVID-19-naïve subjects with respect to those of COVID-19-recovered subjects; this result is not unexpected given the fact that COVID-19-recovered subjects are sampled after more than 7 months from infection and do not report any long-COVID symptoms.

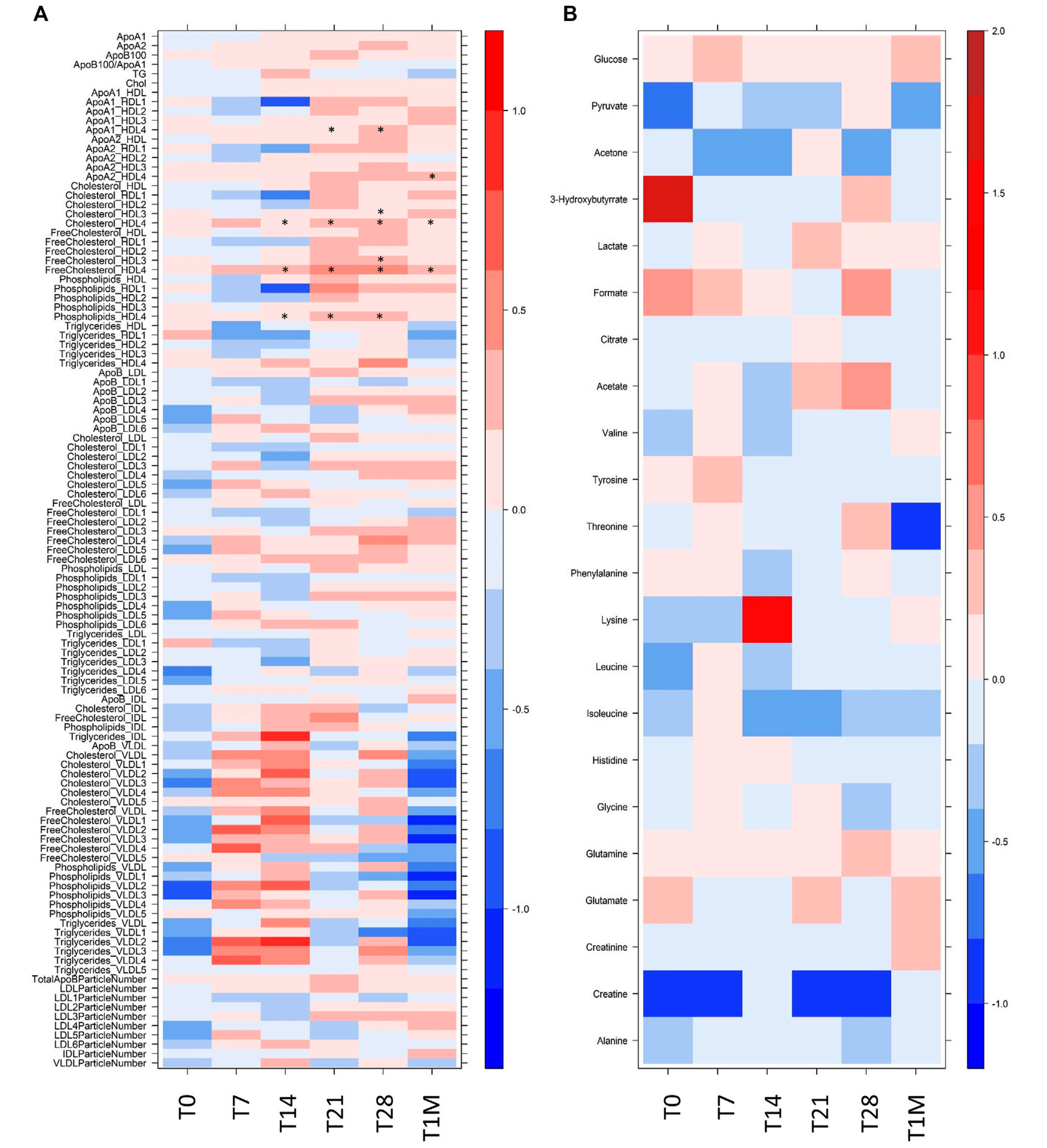


FIGURE 2 | Level plot of Log$_2$(FC) of **(A)** lipoprotein related parameters and **(B)** metabolites; red/blue values indicate higher/lower concentration at T0, T7, T14, T21, T28 and T1M samples of COVID-19-recovered group with respect to COVID-19-naïve group. The brightness of each color corresponds to the magnitude of the FC. Asterisks indicate statistical significanceThe level plot has been created using the function *levelplot* implemented in the R package "Lattice".

As shown in **Figure 2**, in our cohort the differences that exists at T0 between the two groups are not significant, although the two groups are not identical, as it is normal to expect for the comparison of any 10 randomly selected individuals against any other 10. The intra-individual differences (**Figure 2**) remain smaller than the inter-individual ones upon

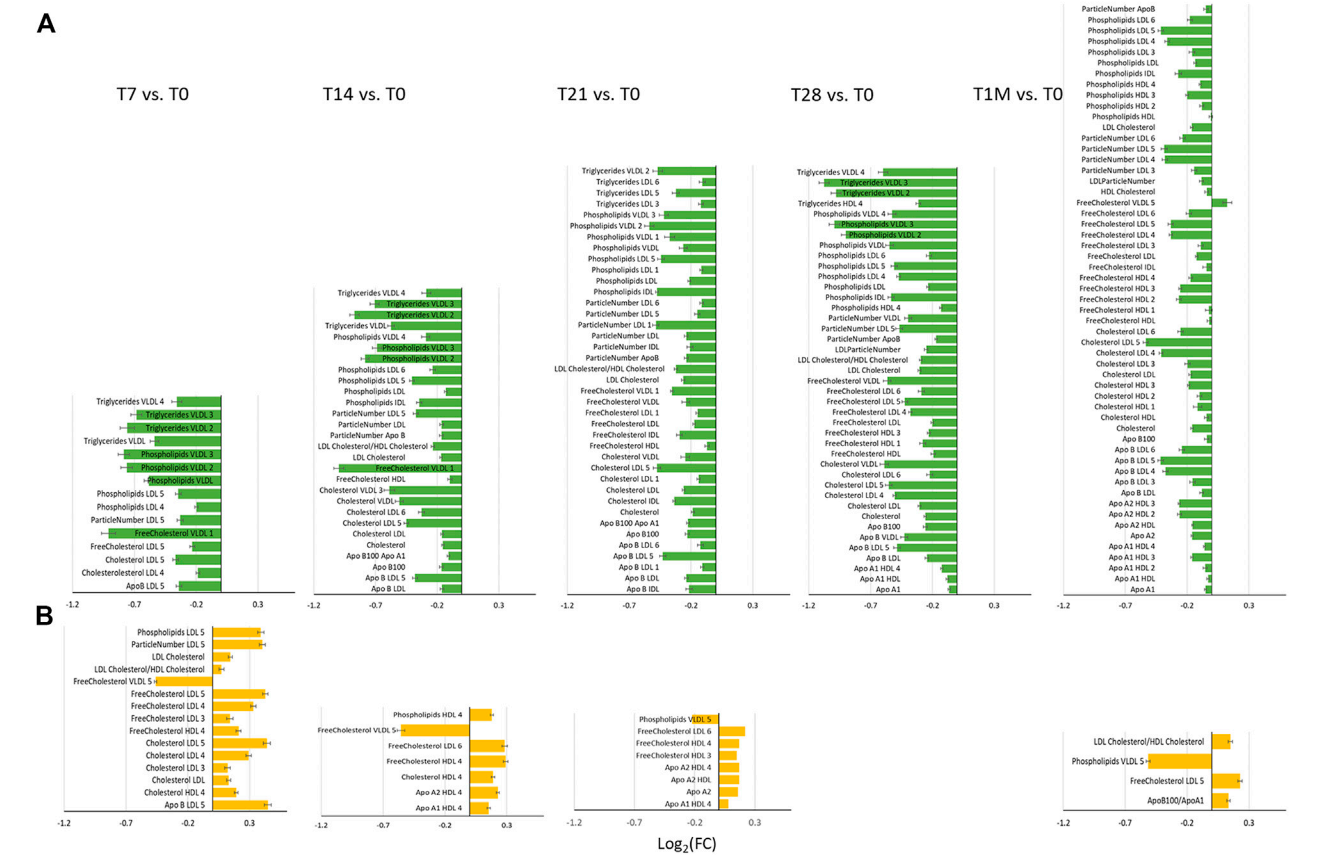


FIGURE 3 | Bar plots of Log_2 (FC) of lipoprotein related parameters significantly different for the comparison at T7, T14, T21, T28 and T1M with respect to T0, in **(A)** COVID-19-naïve (green plots) and **(B)** COVID-19-recovered (orange plots) groups. Features with Log_2(FC) positive/negative values have higher/lower concentration in T7, T14, T21, T28 and T1M samples with respect to T0.

vaccination, which therefore does not represent a major modification of the metabolic phenotype. The inter-individual discrimination considering the six samples collected for each subject is >85%. Nevertheless, in response to vaccination we could observe some common changes that are consistently occurring in all subjects within each group at a given time. As shown in **Figure 2**, the differences between the two groups are essentially restricted to a small number of lipoprotein parameters. They mainly involve HDL4 subfractions (with some *p*-value < 0.05) and appear from T14. Although not statistically significant, a clear trend is observed also for all the VLDL subfractions along the time line T0-T1M; the $\log_2$(FC) is maximum at T7 and T14 and then decreases, until at T1M it tends towards the re-establishment of the pattern observed at T0.

To better analyze the origin of the time-dependent changes, we performed a paired analysis, so to highlight the common intra-individual variations in each group. To this purpose the concentration of all measurable species for a given individual at each time point was compared to that of the same individual at T0. **Figure 3** reports the $\log_2$(FC) of the lipoprotein parameters that were observed to change significantly in the COVID-19-naïve and COVID-19-recovered groups, separately. The pattern of changes is clearly different between the two cohorts. In the former case (**Figure 3A**), we observe an overall decrease in concentration of lipoproteins with average absolute values decreasing from T7 to T21, and then increasing again after the second dose (T28) and again decreasing at T1M. Contemporarily, when the time distance from dose administration increases, we observe an increase in the number of dysregulated features. With the help of **Figure 4**, we can identify the following trends. In terms of main parameters, the most affected along the time series are the ApoB100 and total cholesterol. In terms of main fractions, we observed a continuous dysregulation of the LDL parameters, with the only exception of that associated to triglycerides; these changes persist up to T1M. The earliest (T7) changes are associated to the LDL5 subfraction. For VLDL, the affected main parameters are phospholipids and triglycerides; the largest changes are observed at T7 and T28 (i.e. at the first time point evaluated after the first and second dose, respectively), where the absolute values of their Log_2(FC) is >0.7; these changes do not persist after T28. The HDL subfractions, with the exception of those associated to triglycerides, change significantly only at T1M, but the extent of the changes is quite small. A completely different trend is observed when looking at the lipoproteins in the COVID-19-recovered subjects (**Figure 3B**), where the changes are much

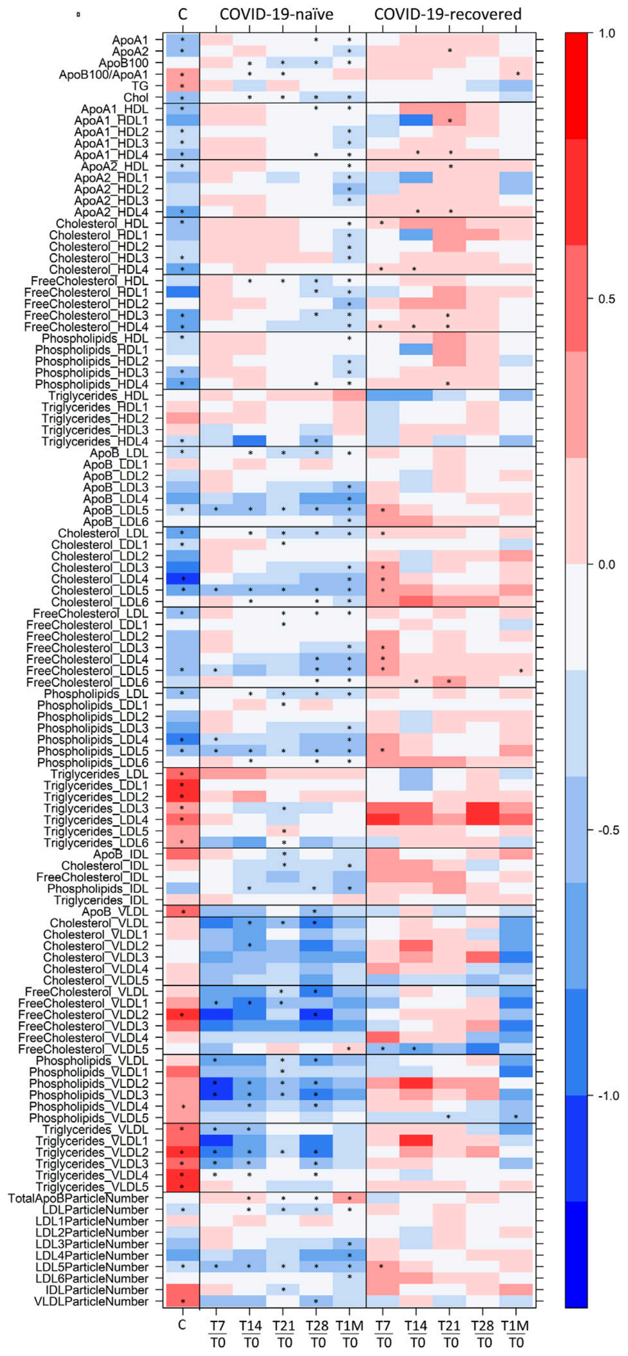


FIGURE 4 | Level plot of Log_2(FC) of the lipoproteome: for COVID-19-naïve and COVID-19-recovered groups (second and third columns, respectively), red/blue parameters indicate higher/lower concentration at T7, T14, T21, T28, and T1M serum samples with respect to T0 samples. For COVID-19 positive subjects (first column), red/blue parameters indicate higher/lower concentration in serum samples of 30 COVID-19 patients with respect to 30 sex- and age-matched control subjects (Meoni et al., 2021). The brightness of each color corresponds to the magnitude of the FC. Asterisks indicate statistical significance. The level plot was created using the function *levelplot* implemented in the R package "Lattice".

smaller in size, of the opposite sign (with the only exception of the decrease in Free Cholesterol- and Phospholipids-VLDL5), and essentially negligible after the second dose. Also the number of affected features is very small and substantially limited to HDL4 and LDL5 parameters (**Figure 3**, **Figure 4**). In neither case, COVID-19-naïve and recovered groups, the measured levels of lipoproteins exceeded the range of values typical of a population of healthy adults (Jiménez et al., 2018). Interestingly, no consistent changes could be observed for any of the metabolites at any of the sampled time points, in neither group.

DISCUSSION

^{1}H NMR provides a unique tool to measure the levels of lipoprotein main parameters, main fractions and subfractions (Jiménez et al., 2018), in addition to metabolites. Here, NMR allowed us to monitor the effects of the Pfizer-BioNTech vaccine in people who never had a contact with the virus and in those with prior COVID-19 infection. In the former group, changes are relatively large in size and mainly involve a downregulation of LDL -cholesterol, -free cholesterol, –phospholipids and–apolipoprotein B100 along with a downregulation of VLDL-phospholipids and–triglycerides; LDL5 emerges as the main dysregulated subfraction. In the latter group instead, the overall changes are small and limited to few lipoprotein components (HDL4 and LDL5 features).

Although this is a small-size pilot study, those described above are clear-cut differences that is extremely unlikely to happen due to chance. The interpretation of the observed changes is far from straightforward. An obvious comparison is with the immunological response. Indeed, the same subjects have been analyzed by some of us in terms of their immune response (Mazzoni et al., 2021). The anti–SARS-CoV-2 serum antibody levels in COVID-19–recovered subjects reach a plateau after the first dose (T7-T14), without any additional improvement after the second one. Instead, in the COVID-19-naïve subjects these levels are not reached even after the second dose (T28).

There is not a common pattern in the timeline trend of immune response and lipoprotein alterations, the only common trait being a reduced response to the second dose in the COVID-19-recovered subjects. What we observe by NMR is most probably an interplay of multiple effects, with a different modulation in the two groups of vaccinated subjects. The fact that previous infection limits the extent of the observed effects suggests that whatever process remodulates the lipoproteins following vaccination in COVID-19-recovered subjects, it has to be related to the "new" encounter with the spike protein. It is worth noting that lipid stripping from cell membrane is a phenomenon associated to the specific action of the spike protein and might be differently operative in recovered and naïve individuals. It is also known that LDL and cholesterol are key mediators of inflammation (Chróinín et al., 2014), which could also have a different extent in recovered and naïve subjects following vaccination. Notably during acute COVID-19 infection, where both lipid bilayer degradation induced by the spike protein and severe inflammation occur, cholesterol and LDL5 are also significantly altered with respect to healthy values, **Figure 4**, first column (Bruzzone et al., 2020; Kimhofer et al., 2020; Ballout et al., 2021; Bizkarguenaga et al., 2021; Lodge et al., 2021; Masuda et al., 2021; Meoni et al., 2021).

Although aware of the intrinsic limitations of the study, we believe the results could stimulate future research addressing a number of relevant aspects. This type of results, if confirmed in larger and diverse (by age, sex, ethnicity, morbidities) populations, might help defining abnormal response to vaccination with the Pfizer-BioNTech formulation and adverse events. A comparison between the effects induced by the different vaccines (Pfizer vs. Moderna; mRNA vs. DNA vaccines, etc.) might shed light on the existence of correlations between fluctuations in the lipoprotein profiles and immune status and to dissect them from the response to the specific formulation.

DATA AVAILABILITY STATEMENT

The dataset presented in this study can be found in an online repository. The name of the repository and accession number can be found below: This study is available at the NIH Common Fund's National Metabolomics Data Repository (NMDR) website, the Metabolomics Workbench, https://www.metabolomicsworkbench.org where it has been assigned Study ID ST002086. The data can be accessed directly *via* its Project DOI: http://dx.doi.org/10.21228/M8FM6D.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Careggi University Hospital Ethical Committee (n. 19466_spe). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

FA devised the main conceptual ideas. PT design the methodological framework. FA, LM, AM, MS, LZ, and AB enrolled the subjects, provided samples and associated data. VG performed the NMR and statistical analyses. LM, AM, VG, FA, and PT contributed to the interpretation of the results. PT and VG drafted the manuscript. All authors, reviewed and edited the manuscripts.

FUNDING

This study was supported by funds to the Department of Experimental and Clinical Medicine, University of Florence (Project Excellence Departments 2018–2022); The University of Florence, project RICTD2122; The Italian Ministry of Health (COVID-2020-12371849); and the Tuscany Region (TagSARSCoV 2). VG is the recipient of a post-doctoral grant provided by the Tuscany Region, project COMETA, Bando COVID-19.

ACKNOWLEDGMENTS

PT and VG acknowledge the support and the use of resources of Instruct-ERIC, a Landmark ESFRI project, and specifically the CERM/CIRMMP Italy Centre. We thank Dr. Leonardo Tenori and Dr. Marco Fondi (University of Florence) and Dr. Edoardo Saccenti (Wageningen University and Research) for their advice on statistical analyses.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.839809/full#supplementary-material

REFERENCES

Assfalg, M., Bertini, I., Colangiuli, D., Luchinat, C., Schäfer, H., Schütz, B., et al. (2008). Evidence of Different Metabolic Phenotypes in Humans. *Proc. Natl. Acad. Sci. U.S.A.* 105, 1420–1424. doi:10.1073/pnas.0705685105

Ballout, R. A., Kong, H., Sampson, M., Otvos, J. D., Cox, A. L., Agbor-Enoh, S., et al. (2021). The NIH Lipo-COVID Study: A Pilot NMR Investigation of Lipoprotein Subfractions and Other Metabolites in Patients with Severe COVID-19. *Biomedicines* 9, 1090. doi:10.3390/biomedicines9091090

Baranovicova, E., Bobcakova, A., Vysehradsky, R., Dankova, Z., Halasova, E., Nosal, V., et al. (2021). The Ability to Normalise Energy Metabolism in Advanced COVID-19 Disease Seems to Be One of the Key Factors Determining the Disease Progression-A Metabolomic NMR Study on Blood Plasma. *Appl. Sci.* 11, 4231. doi:10.3390/app11094231

Bernini, P., Bertini, I., Luchinat, C., Nepi, S., Saccenti, E., Schäfer, H., et al. (2009). Individual Human Phenotypes in Metabolic Space and Time. *J. Proteome Res.* 8, 4264–4271. doi:10.1021/pr900344m

Bizkarguenaga, M., Bruzzone, C., Gil-Redondo, R., SanJuan, I., Martin-Ruiz, I., Barriales, D., et al. (2021). Uneven Metabolic and Lipidomic Profiles in Recovered COVID-19 Patients as Investigated by Plasma NMR Metabolomics. *NMR Biomed.* 35, e4637. doi:10.1002/nbm.4637

Bruzzone, C., Bizkarguenaga, M., Gil-Redondo, R., Diercks, T., Arana, E., García de Vicuña, A., et al. (2020). SARS-CoV-2 Infection Dysregulates the Metabolomic and Lipidomic Profiles of Serum. *iScience* 23, 101645. doi:10.1016/j.isci.2020.101645

Chroinin, D. N., Marnane, M., Akijian, L., Merwick, A., Fallon, E., Horgan, G., et al. (2014). Serum Lipids Associated with Inflammation-Related PET-FDG Uptake in Symptomatic Carotid Plaque. *Neurology* 82, 1693–1699. doi:10.1212/WNL.0000000000000408

Diray-Arce, J., Conti, M. G., Petrova, B., Kanarek, N., Angelidou, A., and Levy, O. (2020). Integrative Metabolomics to Identify Molecular Signatures of Responses to Vaccines and Infections. *Metabolites* 10, 492. doi:10.3390/metabo10120492

Ebinger, J. E., Fert-Bober, J., Printsev, I., Wu, M., Sun, N., Prostko, J. C., et al. (2021). Antibody Responses to the BNT162b2 mRNA Vaccine in Individuals Previously Infected with SARS-CoV-2. *Nat. Med.* 27, 981–984. doi:10.1038/s41591-021-01325-6

Ghini, V., Saccenti, E., Tenori, L., Assfalg, M., and Luchinat, C. (2015). Allostasis and Resilience of the Human Individual Metabolic Phenotype. *J. Proteome Res.* 14, 2951–2962. doi:10.1021/acs.jproteome.5b00275

Holmes, E., Loo, R. L., Stamler, J., Bictash, M., Yap, I. K. S., Chan, Q., et al. (2008). Human Metabolic Phenotype Diversity and its Association with Diet and Blood Pressure. *Nature* 453, 396–400. doi:10.1038/nature06882

Jiménez, B., Holmes, E., Heude, C., Tolson, R. F., Harvey, N., Lodge, S. L., et al. (2018). Quantitative Lipoprotein Subclass and Low Molecular Weight Metabolite Analysis in Human Serum and Plasma by 1H NMR Spectroscopy in a Multilaboratory Trial. *Anal. Chem.* 90, 11962–11971. doi:10.1021/acs.analchem.8b02412

Julkunen, H., Cichońska, A., Slagboom, P. E., and Würtz, P.Nightingale Health UK Biobank Initiative (2021). Metabolic Biomarker Profiling for Identification of Susceptibility to Severe Pneumonia and COVID-19 in the General Population. *eLife* 10, e63033. doi:10.7554/eLife.63033

Kimhofer, T., Lodge, S., Whiley, L., Gray, N., Loo, R. L., Lawler, N. G., et al. (2020). Integrative Modeling of Quantitative Plasma Lipoprotein, Metabolic, and Amino Acid Data Reveals a Multiorgan Pathological Signature of SARS-CoV-2 Infection. *J. Proteome Res.* 19, 4442–4454. doi:10.1021/acs.jproteome.0c00519

Levi, R., Azzolini, E., Pozzi, C., Ubaldi, L., Lagioia, M., Mantovani, A., et al. (2021). One Dose of SARS-CoV-2 Vaccine Exponentially Increases Antibodies in Individuals Who Have Recovered from Symptomatic COVID-19. *J. Clin. Invest.* 131, e149154. doi:10.1172/JCI149154

Lodge, S., Nitschke, P., Kimhofer, T., Coudert, J. D., Begum, S., Bong, S.-H., et al. (2021). NMR Spectroscopic Windows on the Systemic Effects of SARS-CoV-2 Infection on Plasma Lipoproteins and Metabolites in Relation to Circulating Cytokines. *J. Proteome Res.* 20, 1382–1396. doi:10.1021/acs.jproteome.0c00876

Masuda, R., Lodge, S., Nitschke, P., Spraul, M., Schaefer, H., Bong, S.-H., et al. (2021). Integrative Modeling of Plasma Metabolic and Lipoprotein Biomarkers of SARS-CoV-2 Infection in Spanish and Australian COVID-19 Patient Cohorts. *J. Proteome Res.* 20, 4139–4152. doi:10.1021/acs.jproteome.1c00458

Mazzoni, A., Di Lauria, N., Maggi, L., Salvati, L., Vanni, A., Capone, M., et al. (2021). First-dose mRNA Vaccination Is Sufficient to Reactivate Immunological Memory to SARS-CoV-2 in Subjects Who Have Recovered from COVID-19. *J. Clin. Invest.* 131, 149150. doi:10.1172/JCI149150

McClenathan, B. M., Stewart, D. A., Spooner, C. E., Pathmasiri, W. W., Burgess, J. P., McRitchie, S. L., et al. (2017). Metabolites as Biomarkers of Adverse Reactions Following Vaccination: A Pilot Study Using Nuclear Magnetic Resonance Metabolomics. *Vaccine* 35, 1238–1245. doi:10.1016/j.vaccine.2017.01.056

Meoni, G., Ghini, V., Maggi, L., Vignoli, A., Mazzoni, A., Salvati, L., et al. (2021). Metabolomic/lipidomic Profiling of COVID-19 and Individual Response to Tocilizumab. *Plos Pathog.* 17, e1009243. doi:10.1371/journal.ppat.1009243

Neuhäuser, M. (2011). "Wilcoxon–Mann–Whitney Test," in *International Encyclopedia of Statistical Science* (Berlin, Heidelberg: Springer), 1656–1658. Available at: https://link.springer.com/referenceworkentry/10.1007/978-3-642-04898-2_615 (Accessed February 21, 2018).

Sasaki, E., Kusunoki, H., Momose, H., Furuhata, K., Hosoda, K., Wakamatsu, K., et al. (2019). Changes of Urine Metabolite Profiles Are Induced by Inactivated Influenza Vaccine Inoculations in Mice. *Sci. Rep.* 9, 16249. doi:10.1038/s41598-019-52686-5

Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., et al. (2016). Metabolomics Workbench: An International Repository for Metabolomics Data and Metadata, Metabolite Standards, Protocols, Tutorials and Training, and Analysis Tools. *Nucleic Acids Res.* 44, D463–D470. doi:10.1093/nar/gkv1042

Takis, P. G., Ghini, V., Tenori, L., Turano, P., and Luchinat, C. (2019). Uniqueness of the NMR Approach to Metabolomics. *Trac Trends Anal. Chem.* 120, 115300. doi:10.1016/j.trac.2018.10.036

Vignoli, A., Ghini, V., Meoni, G., Licari, C., Takis, P. G., Tenori, L., et al. (2019). High-Throughput Metabolomics by 1D NMR. *Angew. Chem. Int. Ed.* 58, 968–994. doi:10.1002/anie.201804736

Yousri, N. A., Kastenmüller, G., Gieger, C., Shin, S.-Y., Erte, I., Menni, C., et al. (2014). Long Term Conservation of Human Metabolic Phenotypes and Link to Heritability. *Metabolomics* 10, 1005–1017. doi:10.1007/s11306-014-0629-y

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

REVIEW
published: 27 April 2022
doi: 10.3389/fmolb.2022.882487

Studying Metabolism by NMR-Based Metabolomics

*Sofia Moco**

Division of Molecular and Computational Toxicology, Department of Chemistry and Pharmaceutical Sciences, Amsterdam Institute for Molecular and Life Sciences, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

During the past few decades, the direct analysis of metabolic intermediates in biological samples has greatly improved the understanding of metabolic processes. The most used technologies for these advances have been mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy. NMR is traditionally used to elucidate molecular structures and has now been extended to the analysis of complex mixtures, as biological samples: NMR-based metabolomics. There are however other areas of small molecule biochemistry for which NMR is equally powerful. These include the quantification of metabolites (qNMR); the use of stable isotope tracers to determine the metabolic fate of drugs or nutrients, unravelling of new metabolic pathways, and flux through pathways; and metabolite-protein interactions for understanding metabolic regulation and pharmacological effects. Computational tools and resources for automating analysis of spectra and extracting meaningful biochemical information has developed in tandem and contributes to a more detailed understanding of systems biochemistry. In this review, we highlight the contribution of NMR in small molecule biochemistry, specifically in metabolic studies by reviewing the state-of-the-art methodologies of NMR spectroscopy and future directions.

Keywords: metabolomics, NMR, metabolism, qNMR, stable isotopes, metabolite-protein interactions

OPEN ACCESS

Edited by:
Wolfram Weckwerth,
University of Vienna, Austria

Reviewed by:
Stefano Cacciatore,
International Centre for Genetic Engineering and Biotechnology (ICGEB), South Africa
Young Hae Choi,
Leiden University, Netherlands

***Correspondence:**
Sofia Moco
s.moco@vu.nl

In memory of Prof. Dr. Jacques Vervoort

Specialty section:
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

Received: *23 February 2022*
Accepted: *24 March 2022*
Published: *27 April 2022*

Citation:
Moco S (2022) Studying Metabolism by NMR-Based Metabolomics. Front. Mol. Biosci. 9:882487. doi: 10.3389/fmolb.2022.882487

INTRODUCTION–NMR, A TOOLSET OF STRATEGIES IN STUDYING METABOLISM

Nuclear Magnetic Resonance (NMR) is a spectroscopic technique that takes advantage of the energetic transition of nuclear spins in the presence of a strong magnetic field. Since the first NMR spectrum published in 1940s, the use of NMR as an analytical chemistry discipline has matured into numerous areas (Claridge, 2006). NMR has proven to be an essential tool in life sciences including in the identification and structure elucidation of organic molecules and specifically metabolites; in studying the dynamics of macromolecules such as proteins and nucleic acids; and more recently in the field of metabolomics (Cohen et al., 1995; Vignoli et al., 2019). Because NMR measurements of molecules are so sensitive to the chemical environment, it offers selective chemical information about molecules in their physiological setting.

The use of NMR in metabolic studies has a long history. ^{31}P NMR was firstly used to monitor phosphorous-containing metabolites, such as nucleotide and sugar phosphates, including redox species, in cells and tissues (Hoult et al., 1974; Shulman et al., 1979; Gadian and Radda, 1981). Researchers in the late 1970s optimistically stated 'it is now possible to obtain on metabolites *in vivo* the kinds of detailed information about structure, motion, reaction rates, and binding sites that have been obtained by NMR studies of purified biomolecules in solution'. (Shulman et al., 1979). Many of these topics are still the subject of research using NMR methodologies today.

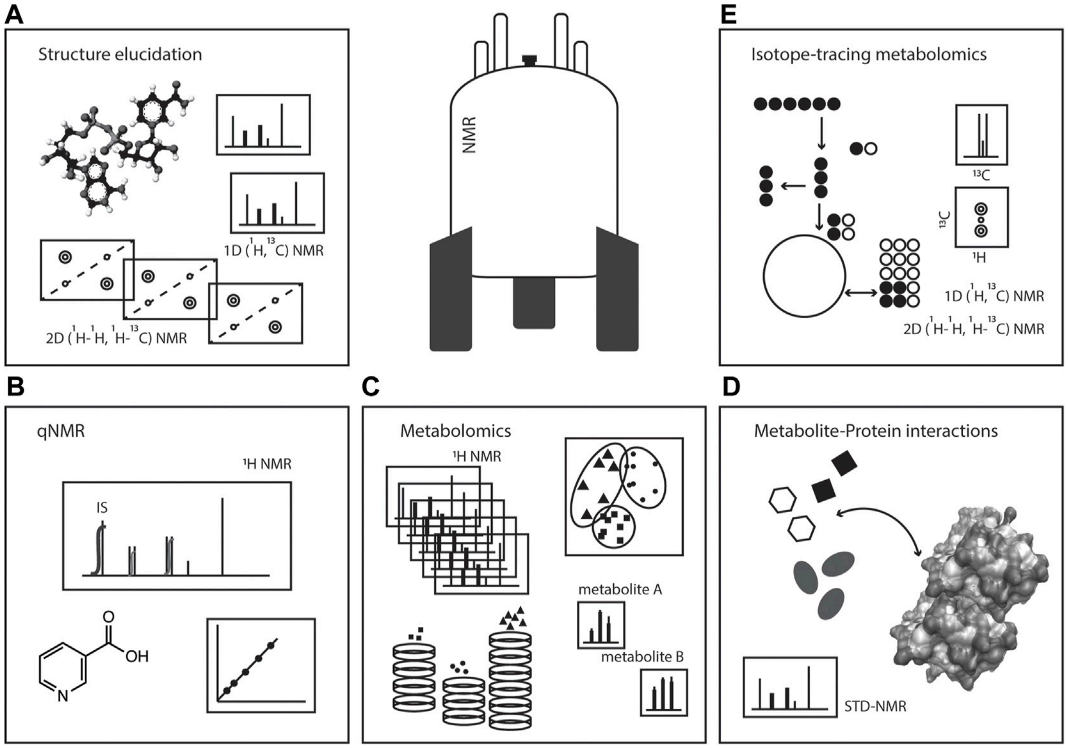


FIGURE 1 | NMR spectroscopy: a toolset in metabolism studies. Pictorial representation of the various ways NMR spectroscopy can be used in metabolic studies, such as **(A)** structure elucidation, **(B)** quantitative NMR (qNMR), **(C)** metabolomics, **(D)** metabolite-protein interactions, and **(E)** isotope-tracing metabolomics or stable isotope resolved metabolomics (SIRM).

Radioactive tracers were the gold standard in studying metabolic fate of molecules in biological systems with widespread application in fields such as medicine, nutrition, toxicology, environmental sciences, and pharmacology. Radioactive isotopes have been progressively replaced by safe stable isotope tracers with the development of labelled supplies and improved detection strategies (Matwiyoff and Ott, 1973; Jang et al., 2018). Stable isotope resolved metabolomics (SIRM) can determine activities of many metabolic reactions across a wide variety of metabolic pathways and has been used to determine absolute metabolic fluxes (Buescher et al., 2015; Lane et al., 2019). Mass spectrometry (MS) and NMR are the techniques of choice in analysing labelling experiments (Lane et al., 2019). NMR is able to provide positional labelling information, a recognisable advantage in discerning metabolite information (Fan et al., 2012).

The establishment of a new era of biological mixture analyses - metabolomics - has emerged because of the development of advanced technologies. Confidence in measuring metabolites has become widespread. These technological advancements, in addition to the pressing societal need in understanding metabolic diseases, boosted a refreshed interest in metabolic studies over the past decade. After all, metabolism pervades every aspect of biology (Deberardinis and Thompson, 2012). While mass spectrometry (MS) has been adopted in many laboratories for metabolic and metabolomics studies because of its wide coverage and high sensitivity, NMR remains used by a smaller community of scientists. NMR gathers several advantages (Emwas et al., 2019). NMR measurements are highly robust: inter-laboratory measurements are reproducible (Ward et al., 2010) and the stability of instrumental response can be months to years if samples are appropriately stored (Pinto et al., 2014). In regular NMR experiments, samples are in tubes and no chromatographic methods are used: hence, the sample is not in contact with the instrument eliminating the need for cleaning the instrument. NMR spectrometers can easily be shared among users with diverse applications, without risk of contamination or carry-over. NMR is quantitative, so both relative and absolute metabolite concentrations can be obtained. Most isomers lead to distinct spectra, making NMR an indispensable tool in structure elucidation.

In this review, we will focus on several NMR strategies of interest in studying metabolism: i) metabolomics analyses; ii) metabolite identification and structure elucidation; iii) quantification (qNMR) of metabolites; iv) the use of stable isotopes in metabolism studies; and v) metabolite-protein interactions, **Figure 1**. The versatility of NMR makes this spectroscopy a powerful toolset in tackling metabolism questions in a variety of biological systems, aiding in unravelling fundamental aspects of biochemistry including metabolite identification, quantification and turnover, metabolic activities, organelle compartmentalisation, and

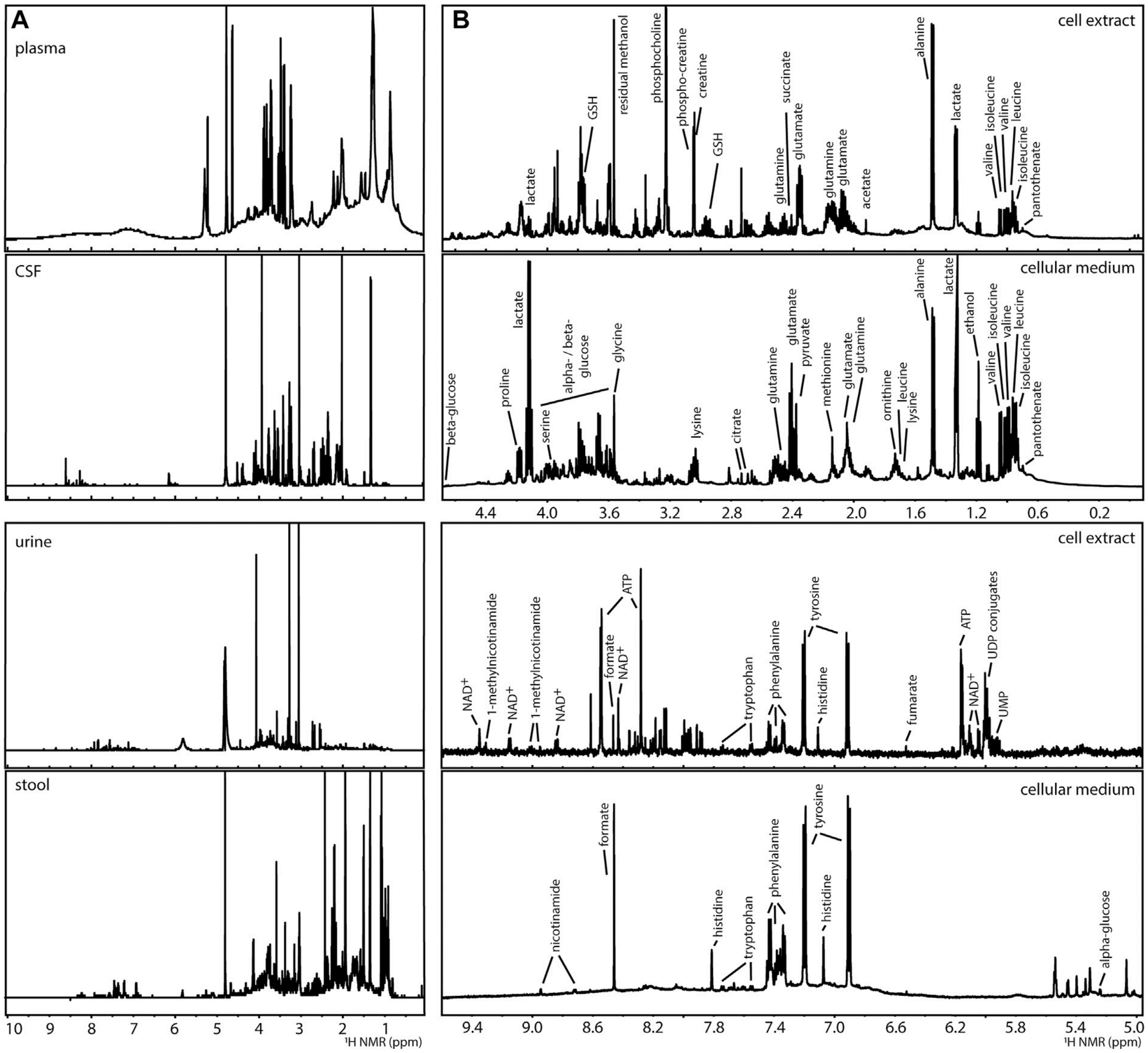


FIGURE 2 | Examples of ^{1}H NMR spectra of metabolomics analyses of **(A)** human biofluids (prepared in phosphate buffer saline in D_2O pH 7.4): plasma, cerebral spinal fluid (CSF), urine, and extract of faecal water (stool) and, **(B)** a human liver cell model (HepG2) after 24 h of culture: intracellular content (cell extract prepared by methanolic extraction) and extracellular content (cellular medium), indicating some of the detected metabolites.

metabolite interaction with macromolecules for enzymology or regulatory events.

NMR METABOLOMICS AND METABOLIC PROFILING

The analysis of complex mixtures (as in metabolomics) by NMR has been used in the characterisation of foods, natural extracts, and biological samples (Moco et al., 2007; Larive et al., 2015; Hatzakis, 2019). A variety of biological samples, such as extracts of microorganisms from the gut microbiome (Klünemann et al., 2021), mammalian cell systems (Kostidis et al., 2017), mammalian (Beckonert et al., 2007) and plant (Kim et al., 2011) tissues, and clinical tissues and biofluids such as plasma, urine, cerebral spinal fluid or faecal water (Beckonert et al., 2007; Martin et al., 2012; Da Silva et al., 2013) have been described, **Figure 2A**. ^{1}H NMR spectra, such as NOESY-1D (1D Nuclear Overhauser Effect Spectroscopy), are commonly utilised generating catalogues of profiles of a large number of metabolites. About 60 metabolites can be identified in an untargeted ^{1}H NMR spectrum using a 600 MHz NMR spectrometer in samples (such as human urine) with little effort in sample preparation (Takis et al., 2017; Vignoli et al., 2019). The ^{1}H NMR analysis of blood matrices such as serum, in addition to small molecules (metabolites) also allows for the detection of lipoprotein classes (Soininen et al., 2009). For example, the analysis of a human cell system detects amino acids, organic acids, sugars, and other metabolites mainly belonging to central carbon metabolism and connected pathways, **Figure 2B** (Kostidis et al., 2017). While in most metabolomics applications, biological samples are placed in solution, analysis of intact tissues can be done by high resolution (HR) magic angle spinning (MAS)-NMR (Chan et al., 2009). However, all obtained ^{1}H NMR spectra in metabolomics suffer from considerable signal overlap since sample preparation is minimal, each metabolite often leads to several signals in the spectrum, and many metabolites can be detected.

The use ^{1}H NMR spectra for metabolomics requires consistent solvent suppression and a flat baseline. Since many biological matrices are water-based, suppressing the solvent signal allows for a better detection of lower abundant compounds and increased sensitivity. Solvent suppression also decreases radiation damping in cryoprobes (Barding et al., 2012). Although there are different pulse sequences that suppress solvent signals, such as WET, WATERGATE or PURGE, NOESY-1D with presaturation and Carr-Purcell-Meiboom-Gill (CPMG) are probably the most widely used in metabolomics (Giraudeau et al., 2015). A flat baseline is essential for subsequent statistical analysis and metabolite quantification (Barding et al., 2012; Emwas et al., 2015; Giraudeau et al., 2015).

While 2D NMR experiments such as ^{1}H,^{13}C-Heteronuclear Single Quantum Coherence (HSQC) (Bingol et al., 2014), ^{1}H-^{1}H-Total Correlation Spectroscopy (TOCSY) (Jiang et al., 2020), 2D-^{1}H-*J*-resolved (JRes) or 2D-^{1}H-Diffusion-ordered NMR spectroscopy (DOSY) or Concentration-ordered NMR spectroscopy (CORDY) (Huang et al., 2015) offer a more deconvoluted picture of a mixture, these experiments take considerably more time and are computationally more intensive to process. Consequently, 2D NMR experiments are infrequently used for fingerprinting purposes in large studies. Even though ^{1}H NMR is the mostly widely nucleus used in metabolomics, other nuclei such as ^{13}C (Clendinen et al., 2015), ^{15}N (Bhinderwala et al., 2018) and ^{31}P (Bhinderwala et al., 2020) have been applied in direct NMR analyses. These nuclei are usually studied through ^{1}H magnetisation in 2D NMR experiments. The ubiquity of ^{1}H in most metabolites and its high NMR sensitivity make ^{1}H NMR the ideal nucleus in NMR-based metabolomics.

An important advantage of NMR-based metabolomic studies is the reproducibility among laboratories (Ward et al., 2010). Given the robustness of the NMR measurement, standardisation of procedures has become progressively easier, especially in clinical applications such as the analysis of human urine, blood serum and plasma. Urine samples are obtained (of course) non-invasively which has led to the development of research and clinical diagnostics. Standardisation of procedures is essential for clinical applications (Emwas et al., 2015). The speed and robustness of sample biomarker profiling with NMR spectroscopy has been extended to thousands of samples. For example, human blood plasma samples of approximately 121,000 participants from UK Biobank have been analysed, leading to an extended clinical chemistry panel consisting of 249 biomarkers and ratios, based on metabolite signals of lipoproteins, lipids, amino acids and a few glycolysis intermediates (Ritchie et al., 2021).

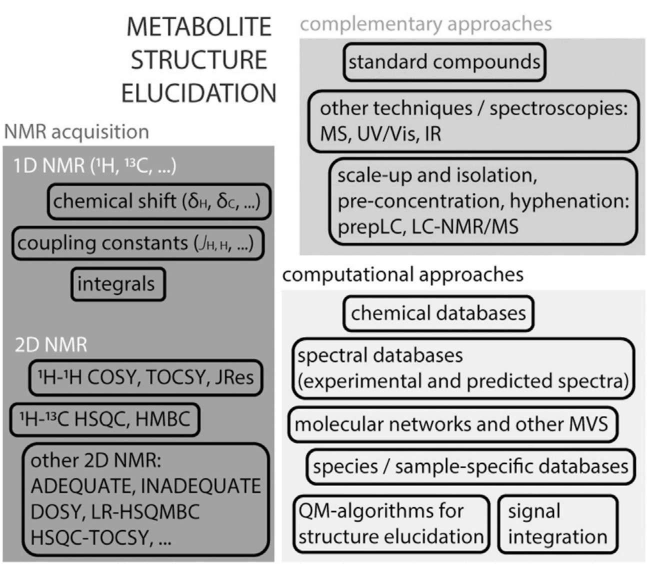

METABOLITE STRUCTURE ELUCIDATION
complementary approaches
standard compounds
other techniques / spectroscopies: MS, UV/Vis, IR
scale-up and isolation, pre-concentration, hyphenation: prepLC, LC-NMR/MS
NMR acquisition
1D NMR (¹H, ¹³C, ...)
chemical shift (δH, δC, ...)
coupling constants (JH,H, ...)
integrals
2D NMR
¹H-¹H COSY, TOCSY, JRes
¹H-¹³C HSQC, HMBC
other 2D NMR: ADEQUATE, INADEQUATE DOSY, LR-HSQMBC HSQC-TOCSY, ...
computational approaches
chemical databases
spectral databases (experimental and predicted spectra)
molecular networks and other MVS
species / sample-specific databases
QM-algorithms for structure elucidation
signal integration

FIGURE 3 | Structure elucidation strategies using NMR. Acquisition of a ^{1}H NMR spectrum on an isolated compound provides essential information about the molecule's structure such as the chemical shift (electronegativity of neighbouring protons and possible functional groups), coupling constants (multiplicity of signals reflects the influence of neighbouring protons), signal integral (assessment of equivalent protons). Atom connectivity can be assessed by homonuclear and heteronuclear 2D NMR spectra, usually ^{1}H–^{1}H or ^{1}H-^{13}C. In certain cases, other 2D or 3D NMR experiments are useful to obtain more detailed information. Identification and spectral deconvolution in NMR benefits from available computational approaches, including databases, multivariate statistical approaches (MVS) and quantum-mechanic-based algorithms. And the availability of complementary information, such as the use of authentic standards or mass spectrometry is generally helpful. In the case of complex mixtures, as in metabolomics, scale-up and metabolite isolation are often unavoidable, in particular in the case of unknown metabolites.

Since metabolomics relies on the comparative analysis of a system challenged by a perturbation relative to its control, it can be applied to a many biochemical questions related to metabolism: such as drug-induced metabolic perturbations (Vinaixa et al., 2011), aetiology of metabolic diseases, like cancer (Vignoli et al., 2021), or cellular development and differentiation (Moussaieff et al., 2015).

METABOLITE STRUCTURE VERIFICATION AND ELUCIDATION

NMR is perhaps mostly known for its ability to elucidate chemical structures of small molecules, **Figure 3**. A ^{1}H NMR spectrum of a given molecule provides information about functional groups (position of chemical shifts), spatial or connecting protons (multiplicity of signals and coupling patterns), and number of equivalent protons (signal integrals). The interpretation of these signals can in many cases lead to an unambiguous identification of the molecule. NMR is efficient in distinguishing many isomers, by their unique spectra. The exception are enantiomers, that require chiral agents to be derivatised into diastereomers for analyses by regular NMR spectroscopy. While ^{1}H NMR provides crucial and sometimes sufficient information to resolve a structure, the complexity of certain molecules requires additional strategies. The interpretation of certain 1D ^{1}H NMR spectra, in particular in the presence of complex multiplicities and second order effects, can profit from quantum-chemistry algorithms (Elyashberg et al., 2016; Pauli et al., 2021). For example, the web-based Cosmic Truth (CT) software uses experimental spectra to calculate coupling constants in complex multiplets, and thereby provide higher certainty on metabolite identification in ^{1}H NMR spectra (Achanta et al., 2021; Pauli et al., 2021).

The next level of obtaining structural information came with the establishment of 2D homonuclear and heteronuclear NMR pulse sequences, which obtains enhanced atom connectivity and spatial information within spin systems. While identification of most purified metabolites can be done using a combination of standard 1D NMR ^{1}H and ^{13}C NMR and 2D NMR (as ^{1}H, ^{1}H-COSY (COrrelation SpectroscopY); ^{1}H, ^{1}H-TOCSY; ^{1}H, ^{13}C-HSQC; ^{1}H, ^{13}C-HMBC (Heteronuclear Multiple Bond Correlation)) experiments, identifying certain molecules requires additional information because of their complexity (Elyashberg, 2015). Methods such as ADEQUATE (Adequate Sensitivity Double-Quantum Spectroscopy), INADEQUATE (Incredible Natural Abundance DoublE QUAntum Transfer Experiment), HSQC-TOCSY and LR-HSQMBC (long-range heteronuclear single quantum multiple bond correlation) aid in putting in evidence certain properties towards resolving more complex spin systems in, for example, natural compounds (Elyashberg, 2015). Pure shift NMR spectroscopy that includes methods such as PSYCHE (Pure Shift Yielded by Chirp Excitation Suppressing) collapse multiplet signals into singlets in ^{1}H NMR spectra to improve spectral resolution (Foroozandeh et al., 2014). For complete structure elucidation, nested super-pulse sequences that encompass a series of existing 2D NMR methods (e.g., HMQC-HSQC-COSY-NOESY) can be used. One example is NOAH: NMR by Ordered Acquisition using ^{1}H-detection (Kupče and Claridge, 2017). A sample containing 50 mM cyclosporine in benzene-d_6 was acquired with the NOAH-5 super-sequence (that combines ^{1}H-^{15}N HMQC, multiplicity edited ^{1}H-^{13}C HSQC, ^{1}H-^{13}C HMBC, COSY and NOESY pulse sequences), producing five 2D spectra in one experiment in 44 min (Kupče and Claridge, 2017).

Metabolite identification by NMR benefits from additional chemical information which can be done by integration of complementary pieces of information. Mass spectrometry (MS) can assist in providing the molecular mass of a molecule, and thereby a putative molecular formula, as well as some structural information by MS/MS fragmentation. The combination of chemical information provided by NMR and MS in combination is highly efficient in metabolite identification (Moco et al., 2007). The access to online resources with experimental and/or predicted NMR spectral databases, such as HMDB (Wishart et al., 2021), BMRB (Ulrich et al., 2007) and NMRShiftDB (Steinbeck et al., 2003) are important tools to deduce possible molecules.

Metabolite identification in metabolomics can be challenging given the presence of many overlapping signals. Libraries of metabolites found in common matrices such as urine, plasma, and serum are useful resources (Wishart et al., 2021). Structure verification in metabolomic studies is done by comparing profiles to standards in spectral databases, as well as acquisition of 2D NMR and MS directly on mixtures. The integration of NMR and MS analyses can provide confirmatory and complementary information on the underlying metabolites, avoiding metabolite isolation (Moco et al., 2008). Multivariate statistical tools as Statistical Total Correlation Spectroscopy (STOCSY) (Cloarec et al., 2005), Subset Optimization by Reference Matching (STORM) (Posma et al., 2012) or Resolution EnhanceD SubseT Optimization by Reference Matching (RED-STORM) (Posma et al., 2017), have been used in biofluid spectra, highlighting spectral regions of differential metabolites. The multiple resonances of a metabolite can be correlated across metabolite datasets.

Concentrating the sample or compound isolation prior to NMR analysis in inevitable when dealing with unknown molecules or unknown matrices. Hyphenated techniques, such as liquid chromatography (LC)-NMR-MS or LC-solid phase extraction (SPE)-diode array detection (DAD)-MS/NMR (Moco and Vervoort, 2012; Garcia-Perez et al., 2020) have been developed. For example, the identification of the uremic toxins *N*^{1}-methyl-2-pyridone-5-carboxamide and *N*^{1}-methyl-4-pyridone-5-carboxamide in C57BL/6 mice's urine was possible through a combination of sequential 1D NMR, STOCSY, 2D NMR, SPE, 2D NMR and spiking of standards (Garcia-Perez et al., 2020). Complex matrices such as a plant extract or a biological sample can be separated by LC, detected by MS with the NMR spectra acquired in a subsequent integrated step (Wolfender et al., 2019). Experimental data in combination with computational tools, including chemometrics, are used in concerted ways to fully describe complex mixtures, often with few if any sample separation steps (Wolfender et al., 2019).

METABOLITE QUANTIFICATION

NMR is inherently quantitative. However, for many applications, qualitative analyses suffice, and the quantitative aspect is overlooked. Quantitative NMR (qNMR) is progressively gaining attention, with applications to drugs, vaccines, natural products, and mixtures such as biological samples and plant extracts (Holzgrabe, 2010; Simmler et al., 2014; Giraudeau, 2017; Li and Hu, 2017; Giancaspro et al., 2021). The basic principle of qNMR relies on the intensity of the NMR signal of an analyte being proportional to the number of nuclei. One of the important determinants in quantitative analysis is the optimisation of longitudinal relaxation time, T1, of protons in ^{1}H qNMR. To obtain truly quantitative spectra, long delays are often required to allow full proton relaxation, as the delay is set to be at least 5 times T1 to have >99.3% of protons to return to original position (Holzgrabe, 2010). For example, the T1 of maleic acid in D_2O phosphate buffer saline pH 7.4, a commonly used internal standard in qNMR, is ~6.5 s. If this is the longest T1 of the resonances found in the sample to quantify, the inter-scan relaxation delay should be set to >30 s. The challenging part of implementing a qNMR routine is maintaining the exactness of procedures, from consistently using the same acquisition and processing parameters to taking into account the physico-chemical properties of the sample (pH, ionic strength, solubility, chemical interactions and interferences, storage) and calibration of scales and pipets (Bharti and Roy, 2012). The quantification of pure compounds or simple mixtures is done by purity analysis and often reference materials are used. This is quite commonplace in pharmaceutical formulations. When implemented, qNMR can lead to superb results in accuracy (<1% error) and robustness (Holzgrabe, 2010; Mahajan and Singh, 2013).

qNMR in 1D-^{1}H NMR of complex mixtures acquired for metabolomic studies is usually less accurate, yielding a trueness of 10–20% (Giraudeau, 2017) since many of the analytical and instrumental parameters are cannot be optimised. qNMR on 2D homonuclear and heteronuclear NMR spectra have also been reported (Giraudeau, 2017; Li and Hu, 2017), which has the advantage of additional metabolite deconvolution compared to 1D NMR. The increased use of fast 1D ^{1}H NMR metabolomics analyses has generated interest in developing quantification strategies in these spectra. Internal reference signals as ERETIC and PULCON avoid the use of external references that crowd spectra (Wishart, 2008; Holzgrabe, 2010).

qNMR in metabolomics is often performed by chemometric analyses (Wishart, 2008; Madrid-Gambin et al., 2020). Metabolomic analytical procedures require high consistency to obtain (semi-) quantitative values. Standardized pre-laboratory procedures (sample collection, storage), sample preparation, spectral acquisition, pre-processing of spectra (referencing, phasing, baseline correction, etc) and statistical analyses or machine learning are all necessary to obtain quantitative results (Wishart, 2008). Signal identification and integration across a series of spectra, including either binning or dynamic integration with alignment and normalisation before quantification are common steps. Given that certain metabolite signals are likely to overlap, it is important to define the least overlapped signals as the metabolite quantifier. While other techniques such as LC-MS, require the use of labelled internal standards and laborious method development and validation, qNMR is easier to implement. Metabolites quantification within the μM to mM range is feasible in NMR metabolomics spectra. For example, the extracellular metabolites in media of mammalian cells was quantified by NMR with a ~15% error (Kostidis et al., 2017). Lineshape fitting models have been used in deconvoluting metabolites in ^{1}H NMR spectra of ultrafiltrated human serum samples, integrating 42 metabolites and explaining >92% of the spectrum (Mihaleva et al., 2014). Computational strategies for qNMR are likely to be further developed for metabolite quantification especially for clinical research studies and acceptance for use in clinical diagnostics.

STABLE ISOTOPE RESOLVED METABOLOMICS

Metabolite analysis of cells, organelles as mitochondria, and organs was initiated by Shulman and co-workers in 1970s. By using ^{31}P and ^{13}C NMR they were able to study aspects of cellular metabolism as oxidative phosphorylation and kinetics of glycolysis in *E. coli* and rat liver cells (Radda and Seeley, 1979; Shulman et al., 1979). Metabolic networks are often highly homeostatic and branched making it difficult to understand metabolic regulation by assessing metabolite concentrations in steady state conditions. A more informative approach is to assess metabolite turnover, defined as the quantity of the metabolite moving through its pool per unit time (McCabe and Previs, 2004; Fan and Lane, 2008). Metabolic turnover is studied with (usually isotope) tracers. Labelled compounds allow the determination of rates of a metabolic flux. Stable isotopes have largely replaced radioactive tracers and use MS or NMR as technologies.

NMR is particularly useful in metabolic studies because it can provide quantitative information of many metabolites at the same time as well as to distinguishing positional labelling. Stable isotope tracer analysis (or stable isotope resolved metabolomics, SIRM) by NMR commonly uses ^{13}C, but also ^{15}N (Lapidot and Gopher, 1997) or ^{2}H (Mahar et al., 2020) labelled tracers. Different tracers may be used according to the pathway of interest. For example [$^{13}C_{1,2}$]-glucose can be used to distinguish between oxidative and non-oxidative branches of the pentose phosphate pathways because of the distribution of ^{13}C in different downstream metabolites; [U-^{13}C]-Glutamine is used to study glutaminolysis, as well as TCA cycle, amino acid metabolism and pyrimidine biosynthesis; and [U-^{13}C]-palmitic acid is often used to study ß-oxidation (Fan and Lane, 2008; Jang et al., 2018; Saborano et al., 2019), **Figure 4**. Since NMR allows to measure isotopomers and metabolites in biological samples, compartmentalisation and exchange dynamics of metabolic pools can be studied. Spatial and temporal events are fundamental to understand metabolism of a given system (Fan and Lane, 2008).

The majority of isotopomer analysis in NMR makes use of ^{13}C tracers for direct measurements of labelled carbons. Specifically, direct 1D ^{1}H NMR experiments allow detecting ^{13}C satellite signals, enable isotopomer distribution analysis, and are generally used for quantification of label incorporation (Fan and Lane, 2008; Vinaixa et al., 2017). Since labelled and non-labelled signals are detected in crude cell extracts, the ^{1}H NMR spectrum is often crowded because of the number of metabolites present. Therefore 2D homonuclear and heteronuclear NMR experiments are used to detect characteristic labelling patterns (COSY, TOCSY, HSQC, HMBC and HCCH-TOCSY and HSQC-TOCSY) (Fan et al., 2005; Fan and Lane, 2008; Lane and Fan, 2017). ^{1}H–^{13}H HSQC are regularly used for ^{13}C metabolic flux analysis, even if there is a large range of coupling constants (typically 120–210 Hz) in the metabolites detected (Reed et al., 2019). To make up for lengthy acquisition times of 2D NMR experiments, ultrafast 2D NMR has been applied to specific isotopic enrichments in complex biological mixtures, considerably reducing acquisition times (Giraudeau et al., 2011). ^{13}C-filtered 1D spectra (and 2D spectra to reduce signal overlap) appear to be accurate in calculating label incorporation in sparsely labelled metabolic samples (Reed et al., 2019). Combining MS and NMR-based SIRM can be a strategy to obtain isotopomer distributions in a model-free way and a wider coverage of the involved intermediates (Chong et al., 2017).

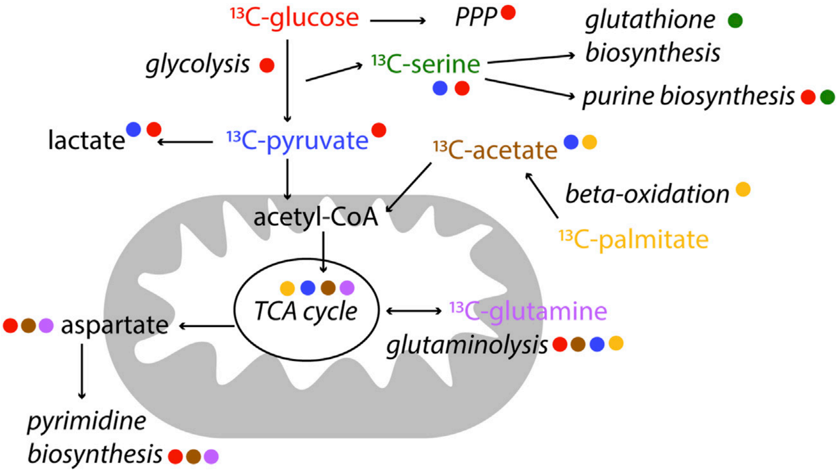


FIGURE 4 | Examples of label incorporation schemes in central metabolism by NMR. Certain tracers are better suited to study specific pathways. In this scheme, the colour of the tracer is indicated next to the pathway name it is used for. ^{13}C-tracers are represented, however alternative tracers with other labelled (^{2}H, ^{15}N) nuclei might be used, as well as other available labelled precursors.

Stable isotope administration can be combined with magnetic resonance spectroscopy (MRS), allowing for *in vivo* metabolic phenotyping in pre-clinical and clinical settings (Leftin et al., 2013). The use of dynamic nuclear polarization (DNP) has offered a major advantage in achieving metabolic studies *in vivo*, as it leads to an outstanding increase of sensitivity, >10,000 (Ardenkjær-Larsen et al., 2003), by using hyperpolarised substrates. For example, the use of hyperpolarized [^{13}C$_1$]pyruvate allowed to study skeletal muscle stimulation *in vivo* (Leftin et al., 2013).

Since central metabolism (including glycolysis, gluconeogenesis, TCA cycle and pentose phosphate pathway) is the heart of metabolism and bioenergetics, most SIRM is being performed on these pathways. For example, patients with early-stage non–small-cell lung cancer were infused with U-^{13}C-glucose before tissue resection. Through SIRM analysis, it was determined that the cancerous tissues in these patients had enhanced pyruvate carboxylase (PC) activity over glutaminase 1 (GLS1), compared to non-cancerous tissues. Both PC and GLS1 are important enzymes in anaplerotic reactions, replenishing TCA cycle intermediates, needed in highly proliferating cells, as cancer cells (Sellers et al., 2015). An example of using NMR to trace other metabolic pathways besides central metabolism, is the comparison of two cancer cells line models (bladder, UMUC3, and prostate, PC3) to assess lipid metabolism turnover (Lin et al., 2021). The use of isotope tracers has also been discussed for studying nucleotide metabolism (Lane and Fan, 2015).

To perform SIRM, detection of metabolic intermediates is required, and of course, knowledge of the metabolic map (Kanehisa et al., 2014; Wishart et al., 2021). Reconstructions of metabolic models are valuable, even if certain aspects remain challenging such as compartmentalization and tissue specificity (Magnúsdóttir et al., 2016). Informatic tools become increasingly important for metabolic flux analysis calculations (Arita, 2003; Rahim et al., 2022).

Isotopes are also used to unravel metabolic fate of certain drugs or unknown molecules. In this case the aim is to study ADMET (absorption, distribution, metabolism, excretion, and toxicity) of a therapeutic agent. Many pharmaceutical drugs contain ^{19}F, so ^{19}F NMR can be used to monitor parent compounds and resulting metabolic products (Rietjens and Vervoort, 1989; Lindon et al., 2004; Keun et al., 2008; Reid and Murphy, 2008). ^{13}C- or ^{15}N-labelled drugs can also be analysed for metabolic fate using 1D- and 2D-NMR in cells and animals (Fan and Lane, 2008; Mutlib, 2008; Fan et al., 2012).

METABOLITE-PROTEIN INTERACTIONS

Interactions between metabolites and proteins are a pre-requisite in enzymatic and allosteric events, defining metabolism and its regulation. While many methodologies to study interactions between macromolecules (e.g., protein-protein interactions) have been developed, methods to systematically assess protein-metabolite interactions are still scarce and often limited to hydrophobic metabolites (Li et al., 2010; Nikolaev et al., 2016; Piazza et al., 2018). NMR has a long history of studying protein dynamics *in vitro*, including changes in protein conformation upon ligand binding (Cohen et al., 1995) specifically by monitoring amino acid residues in the protein backbone. However, a set of ligand-observed NMR experiments can be used to specifically monitor the binding event via the ligand (in opposition to monitoring the protein). Saturation transfer difference (STD) (Mayer and Meyer, 1999; Viegas et al., 2011), **Figure 5**, water–ligand observation with gradient spectroscopy (WaterLOGSY) (Dalvit et al., 2001), time constant of spin-lattice relaxation in rotating frame (T1rho) and CPMG (Gossert and Jahnke, 2016) are some examples of NMR methods to monitor ligand binding to purified (non-isotopically labelled) proteins (Gossert and Jahnke, 2016). Ligand-observed NMR has been

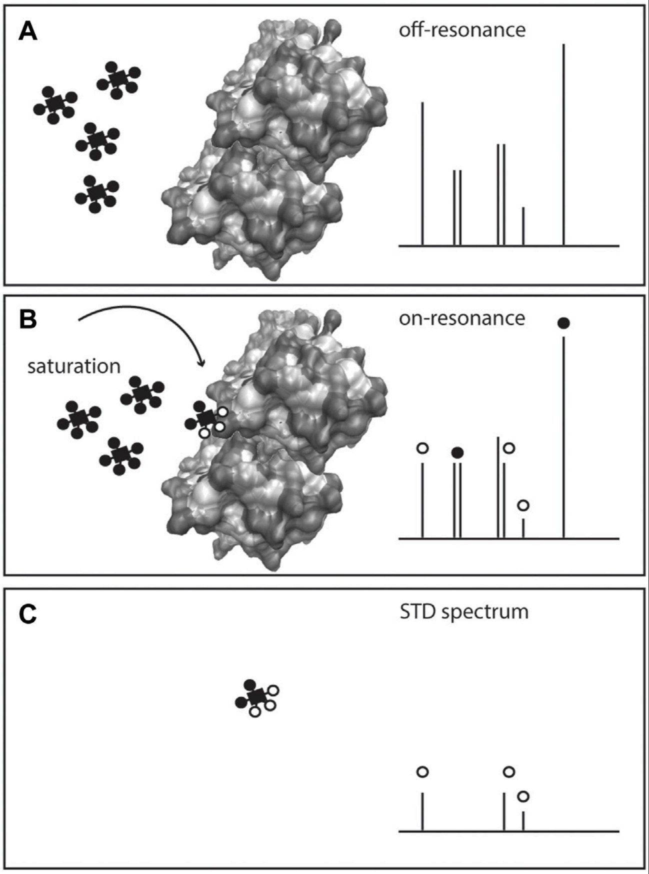


FIGURE 5 | Scheme of Saturation Transfer Difference (STD)-NMR for studying protein-metabolite interactions. The protein is exposed to an excess of ligand(s) or metabolite(s) and a ^{1}H NMR spectrum is recorded (off-resonance spectrum), **(A)** Given the low amount of protein in solution, only the metabolite(s) signals are visible. On a second acquisition, selected saturation is applied to the protein that is transferred to the bound metabolite through the nuclear Overhauser effect, inducing the bound ligand resonances to broaden or disappear (on-resonance spectrum), **(B)** The difference spectrum (off-resonance on-resonance), STD spectrum, **(C)** exhibits the resonances of the metabolite bound to the protein, and confirms the presence of the protein-metabolite interaction.

primarily used in high throughput fragment screening conducted for drug discovery (Pellecchia et al., 2008; Gossert and Jahnke, 2016).

Ligand-observed NMR methods depend on certain conditions. The ligand is added in excess (10–20 fold to the protein amount) to a large protein (>30 kDa), and the interactions are typically weak, with dissociation constants (K_D) 1 μM-10 mM. The ligand is in a fast exchange with the protein, and upon binding, the signal experiences a strong relaxation as evidenced by proton signal broadening or disappearance from the spectrum, **Figure 5**. By analysis of bound and unbound states, binding can be obtained via NOEs (STD and water-mediated NOEs, WaterLOSGY) on ligand signals (Meyer and Peters, 2003; Gossert and Jahnke, 2016). Calculation of K_D and even epitope mapping of the ligand interaction can be obtained. Competition between ligands in ligand mixtures can also be assessed (Viegas et al., 2011; Monaco et al., 2017).

Ligand-observed NMR has been applied to the systematic identification of endogenous metabolites-protein interactions, an example of which are the central carbon metabolism proteins of *E. coli* (Nikolaev et al., 2016; Diether et al., 2019). Solutions of up to 55 metabolites were exposed to 29 purified metabolic enzymes. This approach identified 76 novel interactions between endogenous metabolites and central metabolism enzymes (Diether et al., 2019).

While this type of approach is quite fast to set-up from the NMR acquisition side, it remains dependent on the availability of purified proteins (or at least enriched protein cell suspensions) with a defined number of metabolites. There are however efforts underway to test small molecule-macromolecule interactions in cellular environments (Siegal and Selenko, 2019). For example, whole-cell STD measurements have been used to determine the binding mode of ligands to an intracellular protein in live bacteria (Bouvier et al., 2019) and cancer cells (Primikyri et al., 2018).

COMPUTATIONAL TOOLS AND RESOURCES

Data analysis is inherent to data acquisition. Many NMR applications have been traditionally processed manually, as for example in the elucidation or confirmation of small molecule structures. However, all NMR applications benefit from computational tools and resources, for faster and more accurate extraction of biochemical information. Some of these tools and resources were mentioned in previous sections of this review. Overviews of the many publicly available resources and open source metabolomics tools have been comprehensively listed, so readers should consult them (Eghbalnia et al., 2017; Stanstrup et al., 2019; Shea and Misra, 2020).

Parsing of data into signal lists or matrices in NMR-based metabolomics is a requirement for performing multivariate statistical or machine learning analyses. This pre-processing step can be done by binning spectral data or by peak picking and integration using commercial software or public algorithms, such as, for example, AlpsNMR (Madrid-Gambin et al., 2020). ^{1}H NMR spectra are sensitive to solvent, pH, ionic strength, and temperature, and thus slight shifts in proton resonances are likely to occur. This can be corrected with alignment algorithms before signal integration. There are many algorithms able to handle NMR data for various purposes, including automated putative metabolite identification according to spectral databases as HMDB (Wishart et al., 2021). Multivariate analyses or machine learning methods for NMR metabolomics spectra can be done prior to or after metabolite assignment. Typically, these analyses will assist in identifying differential significant metabolites (or metabolite features) in datasets. Details on various possibilities of handling NMR-based metabolomics data can be consulted elsewhere (Blaise et al., 2021; Debik et al., 2022). Beyond statistical treatment, web-based tools like MetaboAnalyst (Chong et al., 2018) allow to visualise metabolomics data in an user-friendly way, and are able to perform additional tasks, as for example pathway enrichment analysis (Wieder et al., 2021).

Overviews of metabolic pathways can be consulted in databases like KEGG (Kanehisa et al., 2017), HMDB (Wishart et al., 2021), WikiPathways (Kutmon et al., 2016) and Recon3D (Brunk et al., 2018) and are valuable to map metabolites from NMR spectra to specific pathways. Labelling patterns obtained from SIRM are useful for metabolic network reconstruction, however it remains challenging to fully exploit it (Lane et al., 2019).

Reports on integration of metabolome data with other omics has been attempted since genes, proteins, and metabolites collectively contribute to metabolism and its regulation. Even though multiple strategies are necessary (Jendoubi, 2021), no single or universal method is applicable for all experiments given the limitation of detections, time scales of the different omics, and specific research questions of each study.

SUMMARY AND FUTURE DIRECTIONS

NMR spectroscopy can be used to study various aspects of metabolism. This review discussed metabolomics analyses qNMR, structure elucidation, SIRM, metabolite-protein interactions and computational approaches studied by NMR.

One of the disadvantages of NMR is its relative lower signal-to-noise, compared to other analytical techniques. The development of microprobes and cryoprobes (Anklin, 2016), as well as the use of hyperpolarised substrates by DNP (Lerche et al., 2015; Plainchont et al., 2018) are some of the strategies being used to enhance sensitivity in NMR measurements.

The development of novel methods capable of deconvoluting complex signals, such as pure shift experiments (Zangger, 2015) and fast experiments based on non-uniform sampling (Mobli and Hoch, 2014), are developments that will assist in obtaining more and faster structural information.

Efforts to standardize procedures for NMR-based metabolomics, and in particular for clinical applications, are on-going (Ritchie et al., 2021). Development of guidelines are likely to allow the use of NMR measurements as an enhanced clinical chemistry panel with applications in screening large biobanks. In this case, quantification of metabolites directly from biofluids will be essential. Hence development of computational tools for spectral deconvolution, integration and quantification will be needed.

With the worldwide increase of metabolic diseases, studying metabolic turnover of pathways through SIRM is likely to be more frequently used in research and clinical settings. Thus, diversification of tracers and NMR strategies, as well as computational tools, are likely to be further developed in this area.

The shift towards human systems and particularly in-cell environments are inevitable since these systems better mimic physiological conditions. It will be important to develop strategies to monitor metabolite-protein interactions in these environments. Studying interaction will contribute to further knowledge of catalytic and allosteric events essential in metabolic regulation.

A more detailed overview of metabolism and its dynamics at the organelle-level - in its cellular compartments - and at the organism level - in specific organs - will be essential to dissect. The interplay between these metabolite pools is indispensable to understand metabolic regulation in health and disease at a systems biochemistry level. While NMR will never make up for its lack of sensitivity, it will enable the study of the many aspects of the spectroscopy of life.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

SM acknowledges the funding of the EU project RISK-HUNT3R (RISK assessment of chemicals integrating HUman centric Next generation Testing strategies promoting the 3Rs) number 964537.

ACKNOWLEDGMENTS

SM acknowledges the critical reading of this manuscript by Dr Jim Kaput. SM is grateful for the NMR facility of the Faculty of Sciences of the VU Amsterdam, managed by Dr Elwin (A.C.) Janssen.

REFERENCES

Achanta, P. S., Jaki, B. U., McAlpine, J. B., Friesen, J. B., Niemitz, M., Chen, S.-N., et al. (2021). Quantum Mechanical NMR Full Spin Analysis in Pharmaceutical Identity Testing and Quality Control. *J. Pharm. Biomed. Anal.* 192, 113601. doi:10.1016/j.jpba.2020.113601

Anklin, C. (2016). "Chapter 3. Small-Volume NMR: Microprobes and Cryoprobes," in *Modern NMR Approaches to the Structure Elucidation of Natural Products, Instrumentation and Software*. Editors G. E. Martin, A. J. Williams, and D. Rovnyak (Cambridge, United Kingdom: Royal Society of Chemistry), 38–57. doi:10.1039/9781849735186-00038

Ardenkjær-Larsen, J. H., Fridlund, B., Gram, A., Hansson, G., Hansson, L., Lerche, M. H., et al. (2003). Increase in Signal-To-Noise Ratio of > 10,000 Times in Liquid-State NMR. *Proc. Natl. Acad. Sci. U.S.A.* 100, 10158–10163. doi:10.1073/pnas.1733835100

Arita, M. (2003). In Silico atomic Tracing by Substrate-Product Relationships in *Escherichia coli* Intermediary Metabolism. *Genome Res.* 13, 2455–2466. doi:10.1101/gr.1212003

Barding, G. A., Salditos, R., and Larive, C. K. (2012). Quantitative NMR for Bioanalysis and Metabolomics. *Anal. Bioanal. Chem.* 404, 1165–1179. doi:10.1007/s00216-012-6188-z

Beckonert, O., Keun, H. C., Ebbels, T. M. D., Bundy, J., Holmes, E., Lindon, J. C., et al. (2007). Metabolic Profiling, Metabolomic and Metabonomic Procedures for NMR Spectroscopy of Urine, Plasma, Serum and Tissue Extracts. *Nat. Protoc.* 2, 2692–2703. doi:10.1038/nprot.2007.376

Bharti, S. K., and Roy, R. (2012). Quantitative 1H NMR Spectroscopy. *Trac Trends Anal. Chem.* 35, 5–26. doi:10.1016/j.trac.2012.02.007

Bhinderwala, F., Evans, P., Jones, K., Laws, B. R., Smith, T. G., Morton, M., et al. (2020). Phosphorus NMR and its Application to Metabolomics. *Anal. Chem.* 92, 9536–9545. doi:10.1021/acs.analchem.0c00591

Bhinderwala, F., Lonergan, S., Woods, J., Zhou, C., Fey, P. D., and Powers, R. (2018). Expanding the Coverage of the Metabolome with Nitrogen-Based NMR. *Anal. Chem.* 90, 4521–4528. doi:10.1021/acs.analchem.7b04922

Bingol, K., Bruschweiler-Li, L., Li, D.-W., and Brüschweiler, R. (2014). Customized Metabolomics Database for the Analysis of NMR ^{1}H-^{1}H TOCSY and ^{13}C-^{1}H HSQC-TOCSY Spectra of Complex Mixtures. *Anal. Chem.* 86, 5494–5501. doi:10.1021/ac500979g

Blaise, B. J., Correia, G. D. S., Haggart, G. A., Surowiec, I., Sands, C., Lewis, M. R., et al. (2021). Statistical Analysis in Metabolic Phenotyping. *Nat. Protoc.* 16, 4299–4326. doi:10.1038/s41596-021-00579-1

Bouvier, G., Simenel, C., Jang, J., Kalia, N. P., Choi, I., Nilges, M., et al. (2019). Target Engagement and Binding Mode of an Antituberculosis Drug to its Bacterial Target Deciphered in Whole Living Cells by NMR. *Biochemistry* 58, 526–533. doi:10.1021/acs.biochem.8b00975

Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Mih, N., et al. (2018). Recon3D Enables a Three-Dimensional View of Gene Variation in Human Metabolism. *Nat. Biotechnol.* 36, 272–281. doi:10.1038/nbt.4072

Buescher, J. M., Antoniewicz, M. R., Boros, L. G., Burgess, S. C., Brunengraber, H., Clish, C. B., et al. (2015). A Roadmap for Interpreting 13C Metabolite Labeling Patterns from Cells. *Curr. Opin. Biotechnol.* 34, 189–201. doi:10.1016/j.copbio.2015.02.003

Chan, E. C., Koh, P. K., Mal, M., Cheah, P. Y., Eu, K. W., Backshall, A., et al. (2009). Metabolic Profiling of Human Colorectal Cancer Using High-Resolution Magic Angle Spinning Nuclear Magnetic Resonance (HR-MAS NMR) Spectroscopy and Gas Chromatography Mass Spectrometry (GC/MS). *J. Proteome Res.* 8, 352–361. doi:10.1021/pr8006232

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., et al. (2018). MetaboAnalyst 4.0: Towards More Transparent and Integrative Metabolomics Analysis. *Nucleic Acids Res.* 46, W486–W494. doi:10.1093/nar/gky310

Chong, M., Jayaraman, A., Marin, S., Selivanov, V., de Atauri Carulla, P. R., Tennant, D. A., et al. (2017). Combined Analysis of NMR and MS Spectra (CANMS). *Angew. Chem. Int. Ed.* 56, 4140–4144. doi:10.1002/anie.201611634

Claridge, T. D. W. (2006). *High-Resolution NMR Techniques in Organic Chemistry*. Editors J. E. Baldwin and R. M. Williams. 2nd ed. (Elsevier)

Clendinen, C. S., Pasquel, C., Ajredini, R., and Edison, A. S. (2015). 13C NMR Metabolomics: Inadequate Network Analysis. *Anal. Chem.* 87, 5698–5706. doi:10.1021/acs.analchem.5b00867

Cloarec, O., Dumas, M.-E., Craig, A., Barton, R. H., Trygg, J., Hudson, J., et al. (2005). Statistical Total Correlation Spectroscopy: An Exploratory Approach for Latent Biomarker Identification from Metabolic 1H NMR Data Sets. *Anal. Chem.* 77, 1282–1289. doi:10.1021/ac048630x

Cohen, J. S., Jaroszewski, J. W., Kaplan, O., Ruiz-Cabello, J., and Collier, S. W. (1995). A History of Biological Applications of NMR Spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* 28, 53–85. doi:10.1016/0079-6565(95)01020-3

Da Silva, L., Godejohann, M., Martin, F.-P. J., Collino, S., Bürkle, A., Moreno-Villanueva, M., et al. (2013). High-Resolution Quantitative Metabolome Analysis of Urine by Automated Flow Injection NMR. *Anal. Chem.* 85, 5801–5809. doi:10.1021/ac4004776

Dalvit, C., Fogliatto, G., Stewart, A., Veronesi, M., and Stockman, B. (2001). WaterLOGSY as a Method for Primary NMR Screening: Practical Aspects and Range of Applicability. *J. Biomol. NMR* 21, 349–359. doi:10.1023/A:1013302231549

Deberardinis, R. J., and Thompson, C. B. (2012). Cellular Metabolism and Disease: What Do Metabolic Outliers Teach Us? *Cell* 148, 1132–1144. doi:10.1016/j.cell.2012.02.032

Debik, J., Sangermani, M., Wang, F., Madssen, T. S., and Giskeødegård, G. F. (2022). Multivariate Analysis of NMR-based Metabolomic Data. *NMR Biomed.* 35, 1–21. doi:10.1002/nbm.4638

Diether, M., Nikolaev, Y., Allain, F. H., and Sauer, U. (2019). Systematic Mapping of Protein-Metabolite Interactions in Central Metabolism of *Escherichia coli*. *Mol. Syst. Biol.* 15, 1–16. doi:10.15252/msb.20199008

Eghbalnia, H. R., Romero, P. R., Westler, W. M., Baskaran, K., Ulrich, E. L., and Markley, J. L. (2017). Increasing Rigor in NMR-Based Metabolomics through Validated and Open Source Tools. *Curr. Opin. Biotechnol.* 43, 56–61. doi:10.1016/j.copbio.2016.08.005

Elyashberg, M. E., Williams, A. J., and Blinov, K. A. (2016). "Application of Computer- Assisted Structure Elucidation (CASE) Methods and NMR Prediction to Natural Products," in *Modern NMR Approaches to the Structure Elucidation of Natural Products, Instrumentation and Software*. Royal Society of Chemistry: Cambridge, United Kingdom

Elyashberg, M. (2015). Identification and Structure Elucidation by NMR Spectroscopy. *Trac Trends Anal. Chem.* 69, 88–97. doi:10.1016/j.trac.2015.02.014

Emwas, A.-H., Luchinat, C., Turano, P., Tenori, L., Roy, R., Salek, R. M., et al. (2015). Standardizing the Experimental Conditions for Using Urine in NMR-Based Metabolomic Studies with a Particular Focus on Diagnostic Studies: a Review. *Metabolomics* 11, 872–894. doi:10.1007/s11306-014-0746-7

Emwas, A.-H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G. A. N., et al. (2019). Nmr Spectroscopy for Metabolomics Research. *Metabolites* 9, 123. doi:10.3390/metabo9070123

Fan, T. W.-M., and Lane, A. N. (2008). Structure-Based Profiling of Metabolites and Isotopomers by NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* 52, 69–117. doi:10.1016/j.pnmrs.2007.03.002

Fan, T. W.-M., Lorkiewicz, P. K., Sellers, K., Moseley, H. N. B., Higashi, R. M., and Lane, A. N. (2012). Stable Isotope-Resolved Metabolomics and Applications for Drug Development. *Pharmacol. Ther.* 133, 366–391. doi:10.1016/j.pharmthera.2011.12.007

Fan, T. W. M., Bandura, L. L., Higashi, R. M., and Lane, A. N. (2005). Metabolomics-Edited Transcriptomics Analysis of Se Anticancer Action in Human Lung Cancer Cells. *Metabolomics* 1, 325–339. doi:10.1007/s11306-005-0012-0

Foroozandeh, M., Adams, R. W., Meharry, N. J., Jeannerat, D., Nilsson, M., and Morris, G. A. (2014). Ultrahigh-Resolution NMR Spectroscopy. *Angew. Chem. Int. Ed.* 53, 6990–6992. doi:10.1002/anie.201404111

Gadian, D. G., and Radda, G. K. (1981). NMR Studies of Tissue Metabolism. *Annu. Rev. Biochem.* 50, 69–83. doi:10.1146/annurev.bi.50.070181.000441

Garcia-Perez, I., Posma, J. M., Serrano-Contreras, J. I., Boulangé, C. L., Chan, Q., Frost, G., et al. (2020). Identifying Unknown Metabolites Using NMR-Based Metabolic Profiling Techniques. *Nat. Protoc.* 15, 2538–2567. doi:10.1038/s41596-020-0343-3

Giancaspro, G., Adams, K. M., Bhavaraju, S., Corbett, C., Diehl, B., Freudenberger, J. C., et al. (2021). The qNMR Summit 5.0: Proceedings and Status of qNMR Technology. *Anal. Chem.* 93, 12162–12169. doi:10.1021/acs.analchem.1c02056

Giraudeau, P. (2017). Challenges and Perspectives in Quantitative NMR. *Magn. Reson. Chem.* 55, 61–69. doi:10.1002/mrc.4475

Giraudeau, P., Massou, S., Robin, Y., Cahoreau, E., Portais, J.-C., and Akoka, S. (2011). Ultrafast Quantitative 2D NMR: An Efficient Tool for the Measurement of Specific Isotopic Enrichments in Complex Biological Mixtures. *Anal. Chem.* 83, 3112–3119. doi:10.1021/ac200007p

Giraudeau, P., Silvestre, V., and Akoka, S. (2015). Optimizing Water Suppression for Quantitative NMR-Based Metabolomics: A Tutorial Review. *Metabolomics* 11, 1041–1055. doi:10.1007/s11306-015-0794-7

Gossert, A. D., and Jahnke, W. (2016). NMR in Drug Discovery: A Practical Guide to Identification and Validation of Ligands Interacting with Biological Macromolecules. *Prog. Nucl. Magn. Reson. Spectrosc.* 97, 82–125. doi:10.1016/j.pnmrs.2016.09.001

Hatzakis, E. (2019). Nuclear Magnetic Resonance (NMR) Spectroscopy in Food Science: A Comprehensive Review. *Compr. Rev. Food Sci. Food Saf.* 18, 189–220. doi:10.1111/1541-4337.12408

Holzgrabe, U. (2010). Quantitative NMR Spectroscopy in Pharmaceutical Applications. *Prog. Nucl. Magn. Reson. Spectrosc.* 57, 229–240. doi:10.1016/j.pnmrs.2010.05.001

Hoult, D. I., Busby, S. J. W., Gadian, D. G., Radda, G. K., Richards, R. E., and Seeley, P. J. (1974). Observation of Tissue Metabolites Using 31P Nuclear Magnetic Resonance. *Nature* 252, 285–287. doi:10.1038/252285a0

Huang, Y., Zhang, Z., Chen, H., Feng, J., Cai, S., and Chen, Z. (2015). A High-Resolution 2D J-Resolved NMR Detection Technique for Metabolite Analyses of Biological Samples. *Sci. Rep.* 5, 8390. doi:10.1038/srep08390

Jang, C., Chen, L., and Rabinowitz, J. D. (2018). Metabolomics and Isotope Tracing. *Cell* 173, 822–837. doi:10.1016/j.cell.2018.03.055

Jendoubi, T. (2021). Approaches to Integrating Metabolomics and Multi-Omics Data: A Primer. *Metabolites* 11, 184. doi:10.3390/metabo11030184

Jiang, L., Howlett, K., Patterson, K., and Wang, B. (2020). Introduction of a New Method for Two-Dimensional NMR Quantitative Analysis in Metabolomics Studies. *Anal. Biochem.* 597, 113692. doi:10.1016/j.ab.2020.113692

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* 45, D353–D361. doi:10.1093/nar/gkw1092

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, Information, Knowledge and Principle: Back to Metabolism in KEGG. *Nucl. Acids Res.* 42, D199–D205. doi:10.1093/nar/gkt1076

Keun, H. C., Athersuch, T. J., Beckonert, O., Wang, Y., Saric, J., Shockcor, J. P., et al. (2008). Heteronuclear 19F−1H Statistical Total Correlation Spectroscopy as a Tool in Drug Metabolism: Study of Flucloxacillin Biotransformation. *Anal. Chem.* 80, 1073–1079. doi:10.1021/ac702040d

Kim, H. K., Choi, Y. H., and Verpoorte, R. (2011). NMR-Based Plant Metabolomics: where Do We Stand, where Do We Go? *Trends Biotechnol.* 29, 267–275. doi:10.1016/j.tibtech.2011.02.001

Klünemann, M., Andrejev, S., Blasche, S., Mateus, A., Phapale, P., Devendran, S., et al. (2021). Bioaccumulation of Therapeutic Drugs by Human Gut Bacteria. *Nature* 597, 533–538. doi:10.1038/s41586-021-03891-8

Kostidis, S., Addie, R. D., Morreau, H., Mayboroda, O. A., and Giera, M. (2017). Quantitative NMR Analysis of Intra- and Extracellular Metabolism of Mammalian Cells: A Tutorial. *Analytica Chim. Acta* 980, 1–24. doi:10.1016/j.aca.2017.05.011

Kupče, Ē., and Claridge, T. D. W. (2017). NOAH: NMR Supersequences for Small Molecule Analysis and Structure Elucidation. *Angew. Chem. Int. Ed.* 56, 11779–11783. doi:10.1002/anie.201705506

Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E. L., Bohler, A., et al. (2016). WikiPathways: Capturing the Full Diversity of Pathway Knowledge. *Nucleic Acids Res.* 44, D488–D494. doi:10.1093/nar/gkv1024

Lane, A. N., and Fan, T. W.-M. (2017). NMR-Based Stable Isotope Resolved Metabolomics in Systems Biochemistry. *Arch. Biochem. Biophys.* 628, 123–131. doi:10.1016/j.abb.2017.02.009

Lane, A. N., and Fan, T. W.-M. (2015). Regulation of Mammalian Nucleotide Metabolism and Biosynthesis. *Nucleic Acids Res.* 43, 2466–2485. doi:10.1093/nar/gkv047

Lane, A. N., Higashi, R. M., and Fan, T. W.-M. (2019). NMR and MS-based Stable Isotope-Resolved Metabolomics and Applications in Cancer Metabolism. *Trac Trends Anal. Chem.* 120, 115322. doi:10.1016/j.trac.2018.11.020

Lapidot, A., and Gopher, A. (1997). Quantitation of Metabolic Compartmentation in Hyperammonemic Brain by Natural Abundance 13C-NMR Detection of 13C-15N Coupling Patterns and Isotopic Shifts. *Eur. J. Biochem.* 243, 597–604. doi:10.1111/j.1432-1033.1997.00597.x

Larive, C. K., Barding, G. A., and Dinges, M. M. (2015). NMR Spectroscopy for Metabolomics and Metabolic Profiling. *Anal. Chem.* 87, 133–146. doi:10.1021/ac504075g

Leftin, A., Degani, H., and Frydman, L. (2013). *In Vivo* magnetic Resonance of Hyperpolarized [13C1]pyruvate: Metabolic Dynamics in Stimulated Muscle. *Am. J. Physiology-Endocrinology Metab.* 305, E1165–E1171. doi:10.1152/ajpendo.00296.2013

Lerche, M. H., Jensen, P. R., Karlsson, M., and Meier, S. (2015). NMR Insights into the Inner Workings of Living Cells. *Anal. Chem.* 87, 119–132. doi:10.1021/ac501467x

Li, X., Gianoulis, T. A., Yip, K. Y., Gerstein, M., and Snyder, M. (2010). Extensive *In Vivo* Metabolite-Protein Interactions Revealed by Large-Scale Systematic Analyses. *Cell* 143, 639–650. doi:10.1016/j.cell.2010.09.048

Li, X., and Hu, K. (2017). Quantitative NMR Studies of Multiple Compound Mixtures. *Annu. Rep. NMR Spectrosc* 90, 85–143. doi:10.1016/bs.arnmr.2016.08.001

Lin, P., Dai, L., Crooks, D. R., Neckers, L. M., Higashi, R. M., Fan, T. W.-M. T., et al. (2021). Nmr Methods for Determining Lipid Turnover via Stable Isotope Resolved Metabolomics. *Metabolites* 11, 202. doi:10.3390/metabo11040202

Lindon, J., Holmes, E., and Nicholson, J. (2004). Toxicological Applications of Magnetic Resonance. *Prog. Nucl. Magn. Reson. Spectrosc.* 45, 109–143. doi:10.1016/j.pnmrs.2004.05.001

Madrid-Gambin, F., Oller-Moreno, S., Fernandez, L., Bartova, S., Giner, M. P., Joyce, C., et al. (2020). AlpsNMR: An R Package for Signal Processing of Fully Untargeted NMR-Based Metabolomics. *Bioinformatics* 36, 2943–2945. doi:10.1093/bioinformatics/btaa022

Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., et al. (2016). Generation of Genome-Scale Metabolic Reconstructions for 773 Members of the Human Gut Microbiota. *Nat. Biotechnol.* 35, 81–89. doi:10.1038/nbt.3703

Mahajan, S., and Singh, I. P. (2013). Determining and Reporting Purity of Organic Molecules: Why qNMR. *Magn. Reson. Chem.* 51, 76–81. doi:10.1002/mrc.3906

Mahar, R., Donabedian, P. L., and Merritt, M. E. (2020). HDO Production from [2H7]glucose Quantitatively Identifies Warburg Metabolism. *Sci. Rep.* 10, 1–10. doi:10.1038/s41598-020-65839-8

Martin, F.-P. J., Montoliu, I., Nagy, K., Moco, S., Collino, S., Guy, P., et al. (2012). Specific Dietary Preferences Are Linked to Differing Gut Microbial Metabolic Activity in Response to Dark Chocolate Intake. *J. Proteome Res.* 11, 6252–6263. doi:10.1021/pr300915z

Matwiyoff, N. A., and Ott, D. G. (1973). Stable Isotope Tracers in the Life Sciences and Medicine. *Science* 181, 1125–1133. doi:10.1126/science.181.4105.1125

Mayer, M., and Meyer, B. (1999). Characterization of Ligand Binding by Saturation Transfer Difference NMR Spectroscopy. *Angew. Chem. Int. Ed.* 38, 1784–1788. doi:10.1002/(sici)1521-3773(19990614)38:12<1784:aid-anie1784>3.0.co;2-q

McCabe, B. J., and Previs, S. F. (2004). Using Isotope Tracers to Study Metabolism: Application in Mouse Models. *Metab. Eng.* 6, 25–35. doi:10.1016/j.ymben.2003.09.003

Meyer, B., and Peters, T. (2003). NMR Spectroscopy Techniques for Screening and Identifying Ligand Binding to Protein Receptors. *Angew. Chem. Int. Ed.* 42, 864–890. doi:10.1002/anie.200390233

Mihaleva, V. V., Korhonen, S.-P., Van Duynhoven, J., Niemitz, M., Vervoort, J., and Jacobs, D. M. (2014). Automated Quantum Mechanical Total Line Shape Fitting Model for Quantitative NMR-Based Profiling of Human Serum Metabolites. *Anal. Bioanal. Chem.* 406, 3091–3102. doi:10.1007/s00216-014-7752-5

Mobli, M., and Hoch, J. C. (2014). Nonuniform Sampling and Non-Fourier Signal Processing Methods in Multidimensional NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* 83, 21–41. doi:10.1016/j.pnmrs.2014.09.002

Moco, S., Forshed, J., De Vos, R. C. H., Bino, R. J., and Vervoort, J. (2008). Intra- and Inter-Metabolite Correlation Spectroscopy of Tomato Metabolomics Data Obtained by Liquid Chromatography-Mass Spectrometry and Nuclear Magnetic Resonance. *Metabolomics* 4, 202–215. doi:10.1007/s11306-008-0112-8

Moco, S., and Vervoort, J. (2012). *Chemical Identification Strategies Using Liquid Chromatography-Photodiode Array-Solid-Phase Extraction-Nuclear Magnetic Resonance/Mass Spectrometry*. Editors N. W. Hardy and R. D. Hall (Totowa, NJ: Humana Press). doi:10.1007/978-1-61779-594-7

Moco, S., Vervoort, J., Moco, S., Bino, R. J., De Vos, R. C. H., and Bino, R. (2007). Metabolomics Technologies and Metabolite Identification. *Trac Trends Anal. Chem.* 26, 855–866. doi:10.1016/j.trac.2007.08.003

Monaco, S., Tailford, L. E., Juge, N., and Angulo, J. (2017). Differential Epitope Mapping by STD NMR Spectroscopy to Reveal the Nature of Protein-Ligand Contacts. *Angew. Chem. Int. Ed.* 56, 15289–15293. doi:10.1002/anie.201707682

Moussaieff, A., Rouleau, M., Kitsberg, D., Cohen, M., Levy, G., Barasch, D., et al. (2015). Glycolysis-Mediated Changes in Acetyl-CoA and Histone Acetylation Control the Early Differentiation of Embryonic Stem Cells. *Cel Metab.* 21, 392–402. doi:10.1016/j.cmet.2015.02.002

Mutlib, A. E. (2008). Application of Stable Isotope-Labeled Compounds in Metabolism and in Metabolism-Mediated Toxicity Studies. *Chem. Res. Toxicol.* 21, 1672–1689. doi:10.1021/tx800139z

Nikolaev, Y. V., Kochanowski, K., Link, H., Sauer, U., and Allain, F. H.-T. (2016). Systematic Identification of Protein-Metabolite Interactions in Complex Metabolite Mixtures by Ligand-Detected Nuclear Magnetic Resonance Spectroscopy. *Biochemistry* 55, 2590–2600. doi:10.1021/acs.biochem.5b01291

O'Shea, K., and Misra, B. B. (2020). Software Tools, Databases and Resources in Metabolomics: Updates from 2018 to 2019. *Metabolomics* 16, 36. doi:10.1007/s11306-020-01657-3

Pauli, G., Ray, G. J., Bzhelyansky, A., Jaki, B., Corbett, C., Szabo, C., et al. (2021). Essential Terminology Connects NMR and qNMR Spectroscopy to its Theoretical Foundation. *ChemRxiv* 49, 1-49. doi:10.33774/chemrxiv-2021-l3dhr

Pellecchia, M., Bertini, I., Cowburn, D., Dalvit, C., Giralt, E., Jahnke, W., et al. (2008). Perspectives on NMR in Drug Discovery: A Technique Comes of Age. *Nat. Rev. Drug Discov.* 7, 738–745. doi:10.1038/nrd2606

Piazza, I., Kochanowski, K., Cappelletti, V., Fuhrer, T., Noor, E., Sauer, U., et al. (2018). A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication. *Cell* 172, 358–372. doi:10.1016/j.cell.2017.12.006

Pinto, J., Domingues, M. R. M., Galhano, E., Pita, C., Almeida, M. d. C., Carreira, I. M., et al. (2014). Human Plasma Stability during Handling and Storage: Impact on NMR Metabolomics. *Analyst* 139, 1168–1177. doi:10.1039/c3an02188b

Plainchont, B., Berruyer, P., Dumez, J.-N., Jannin, S., and Giraudeau, P. (2018). Dynamic Nuclear Polarization Opens New Perspectives for NMR Spectroscopy in Analytical Chemistry. *Anal. Chem.* 90, 3639–3650. doi:10.1021/acs.analchem.7b05236

Posma, J. M., Garcia-Perez, I., De Iorio, M., Lindon, J. C., Elliott, P., Holmes, E., et al. (2012). Subset Optimization by Reference Matching (STORM): An Optimized Statistical Approach for Recovery of Metabolic Biomarker Structural Information from 1H NMR Spectra of Biofluids. *Anal. Chem.* 84, 10694–10701. doi:10.1021/ac302360v

Posma, J. M., Garcia-Perez, I., Heaton, J. C., Burdisso, P., Mathers, J. C., Draper, J., et al. (2017). Integrated Analytical and Statistical Two-Dimensional Spectroscopy Strategy for Metabolite Identification: Application to Dietary Biomarkers. *Anal. Chem.* 89, 3300–3309. doi:10.1021/acs.analchem.6b03324

Primikyri, A., Sayyad, N., Quilici, G., Vrettos, E. I., Lim, K., Chi, S. W., et al. (2018). Probing the Interaction of a Quercetin Bioconjugate with Bcl-2 in Living Human Cancer Cells with In-cell NMR Spectroscopy. *FEBS Lett.* 592, 3367–3379. doi:10.1002/1873-3468.13250

Radda, G. K., and Seeley, P. J. (1979). Recent Studies on Cellular Metabolism by Nuclear Magnetic Resonance. *Annu. Rev. Physiol.* 41, 749–769. doi:10.1146/annurev.ph.41.030179.003533

Rahim, M., Ragavan, M., Deja, S., Merritt, M. E., Burgess, S. C., and Young, J. D. (2022). INCA 2.0: A Tool for Integrated, Dynamic Modeling of NMR- and MS-based Isotopomer Measurements and Rigorous Metabolic Flux Analysis. *Metab. Eng.* 69, 275–285. doi:10.1016/j.ymben.2021.12.009

Reed, M. A. C., Roberts, J., Gierth, P., Kupče, Ē., and Günther, U. L. (2019). Quantitative Isotopomer Rates in Real-Time Metabolism of Cells Determined by NMR Methods. *ChemBioChem* 20, 2207–2211. doi:10.1002/cbic.201900084

Reid, D., and Murphy, P. (2008). Fluorine Magnetic Resonance *In Vivo*: A Powerful Tool in the Study of Drug Distribution and Metabolism. *Drug Discov. Today* 13, 473–480. doi:10.1016/j.drudis.2007.12.011

Rietjens, I. M. C. M., and Vervoort, J. (1989). Microsomal Metabolism of Fluoroanilines. *Xenobiotica* 19, 1297–1305. doi:10.3109/00498258909043181

Ritchie, S. C., Surendran, P., Karthikeyan, S., Lambert, S. A., Bolton, T., Pennells, L., et al. (2021). Quality Control and Removal of Technical Variation of NMR Metabolic Biomarker Data in ~120,000 UK Biobank Participants. *medRxiv* 9, 1–25. Available at: https://www.medrxiv.org/content/10.1101/2021.09.24.21264079v1.abstract. doi:10.1101/2021.09.24.21264079

Saborano, R., Eraslan, Z., Roberts, J., Khanim, F. L., Lalor, P. F., Reed, M. A. C., et al. (2019). A Framework for Tracer-Based Metabolism in Mammalian Cells by NMR. *Sci. Rep.* 9, 1–13. doi:10.1038/s41598-018-37525-3

Sellers, K., Fox, M. P., Bousamra, M., Slone, S. P., Higashi, R. M., Miller, D. M., et al. (2015). Pyruvate Carboxylase Is Critical for Non-small-cell Lung Cancer Proliferation. *J. Clin. Invest.* 125, 687–698. doi:10.1172/JCI72873DS1

Shulman, R. G., Brown, T. R., Ugurbil, K., Ogawa, S., Cohen, S. M., and den Hollander, J. A. (1979). Cellular Applications of ^{31}P and ^{13}C Nuclear Magnetic Resonance. *Science* 205, 160–166. doi:10.1126/science.36664

Siegal, G., and Selenko, P. (2019). Cells, Drugs and NMR. *J. Magn. Reson.* 306, 202–212. doi:10.1016/j.jmr.2019.07.018

Simmler, C., Napolitano, J. G., McAlpine, J. B., Chen, S.-N., and Pauli, G. F. (2014). Universal Quantitative NMR Analysis of Complex Natural Samples. *Curr. Opin. Biotechnol.* 25, 51–59. doi:10.1016/j.copbio.2013.08.004

Soininen, P., Kangas, A. J., Würtz, P., Tukiainen, T., Tynkkynen, T., Laatikainen, R., et al. (2009). High-throughput Serum NMR Metabonomics for Cost-Effective Holistic Studies on Systemic Metabolism. *Analyst* 134, 1781. doi:10.1039/b910205a

Stanstrup, J., Broeckling, C., Helmus, R., Hoffmann, N., Mathé, E., Naake, T., et al. (2019). The metaRbolomics Toolbox in Bioconductor and Beyond. *Metabolites* 9, 200. doi:10.3390/metabo9100200

Steinbeck, C., Krause, S., and Kuhn, S. (2003). NMRShiftDBConstructing a Free Chemical Information System with Open-Source Components. *J. Chem. Inf. Comput. Sci.* 43, 1733–1739. doi:10.1021/ci0341363

Takis, P. G., Schäfer, H., Spraul, M., and Luchinat, C. (2017). Deconvoluting Interrelationships between Concentrations and Chemical Shifts in Urine Provides a Powerful Analysis Tool. *Nat. Commun.* 8, 1662. doi:10.1038/s41467-017-01587-0

Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., et al. (2007). BioMagResBank. *Nucleic Acids Res.* 36, D402–D408. doi:10.1093/nar/gkm957

Viegas, A., Manso, J., Nobrega, F. L., and Cabrita, E. J. (2011). Saturation-Transfer Difference (STD) NMR: A Simple and Fast Method for Ligand Screening and Characterization of Protein Binding. *J. Chem. Educ.* 88, 990–994. doi:10.1021/ed101169t

Vignoli, A., Ghini, V., Meoni, G., Licari, C., Takis, P. G., Tenori, L., et al. (2019). High-Throughput Metabolomics by 1D NMR. *Angew. Chem. Int. Ed.* 58, 968–994. doi:10.1002/anie.201804736

Vignoli, A., Risi, E., McCartney, A., Migliaccio, I., Moretti, E., Malorni, L., et al. (2021). Precision Oncology via Nmr-Based Metabolomics: A Review on Breast Cancer. *Int. J. Mol. Sci.* 22, 4687–4726. doi:10.3390/ijms22094687

Vinaixa, M., Rodríguez, M. A., Aivio, S., Capellades, J., Gómez, J., Canyellas, N., et al. (2017). Positional Enrichment by Proton Analysis (PEPA): A One-Dimensional 1 H-NMR Approach for 13 C Stable Isotope Tracer Studies in Metabolomics. *Angew. Chem. Int. Ed.* 56, 3531–3535. doi:10.1002/anie.201611347

Vinaixa, M., Rodriguez, M. A., Samino, S., Díaz, M., Beltran, A., Mallol, R., et al. (2011). Metabolomics Reveals Reduction of Metabolic Oxidation in Women with Polycystic Ovary Syndrome after Pioglitazone-Flutamide-Metformin Polytherapy. *PLoS One* 6, e29052. doi:10.1371/journal.pone.0029052

Ward, J. L., Baker, J. M., Miller, S. J., Deborde, C., Maucourt, M., Biais, B., et al. (2010). An Inter-laboratory Comparison Demonstrates that [1H]-NMR Metabolite Fingerprinting Is a Robust Technique for Collaborative Plant Metabolomic Data Collection. *Metabolomics* 6, 263–273. doi:10.1007/s11306-010-0200-4

Wieder, C., Frainay, C., Poupin, N., Rodríguez-Mier, P., Vinson, F., Cooke, J., et al. (2021). Pathway Analysis in Metabolomics: Recommendations for the Use of Over-representation Analysis. *Plos Comput. Biol.* 17, e1009105. doi:10.1371/journal.pcbi.1009105

Wishart, D. S., Guo, A., Oler, E., Wang, F., Anjum, A., Peters, H., et al. (2021). HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Res.* 50, D622–D631. doi:10.1093/nar/gkab1062

Wishart, D. S. (2008). Quantitative Metabolomics Using NMR. *Trac Trends Anal. Chem.* 27, 228–237. doi:10.1016/j.trac.2007.12.001

Wolfender, J.-L., Nuzillard, J.-M., Van Der Hooft, J. J. J., Renault, J.-H., and Bertrand, S. (2019). Accelerating Metabolite Identification in Natural Product Research: Toward an Ideal Combination of Liquid Chromatography-High-Resolution Tandem Mass Spectrometry and NMR Profiling, In Silico Databases, and Chemometrics. *Anal. Chem.* 91, 704–742. doi:10.1021/acs.analchem.8b05112

Zangger, K. (2015). Pure Shift NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* 86-87, 1–20. doi:10.1016/j.pnmrs.2015.02.002

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 01 June 2022
doi: 10.3389/fmolb.2022.896226

The Effects of Carbon Source and Growth Temperature on the Fatty Acid Profiles of *Thermobifida fusca*

Dirk C. Winkelman and Basil J. Nikolau*

Department of Biochemistry, Biophysics and Molecular Biology and the Center of Metabolic Biology, Iowa State University, Ames, IA, United States

The aerobic, thermophilic *Actinobacterium*, *Thermobifida fusca* has been proposed as an organism to be used for the efficient conversion of plant biomass to fatty acid-derived precursors of biofuels or biorenewable chemicals. Despite the potential of *T. fusca* to catabolize plant biomass, there is remarkably little data available concerning the natural ability of this organism to produce fatty acids. Therefore, we determined the fatty acids that *T. fusca* produces when it is grown on different carbon sources (i.e., glucose, cellobiose, cellulose and avicel) and at two different growth temperatures, namely at the optimal growth temperature of 50°C and at a suboptimal temperature of 37°C. These analyses establish that *T. fusca* produces a combination of linear and branched chain fatty acids (BCFAs), including *iso*-, *anteiso*-, and 10-methyl BCFAs that range between 14- and 18-carbons in length. Although different carbon sources and growth temperatures both quantitatively and qualitatively affect the fatty acid profiles produced by *T. fusca*, growth temperature is the greater modifier of these traits. Additionally, genome scanning enabled the identification of many of the fatty acid biosynthetic genes encoded by *T. fusca*.

Keywords: *Thermobifida fusca*, *Actinomycete*, fatty acid biosynthesis pathway, principal component analysis, gas chromatography- mass spectrometry, branched chain fatty acids, fatty acid synthase

Edited by:
Wolfram Weckwerth,
University of Vienna, Austria

Reviewed by:
Bo Pang,
Apertor Pharmaceuticals, Inc.,
United States
Francisco Solano,
University of Murcia, Spain
José Juan Ordaz-Ortiz,
Instituto Politécnico Nacional de
México (CINVESTAV), Mexico
Fidele Tugizimana,
Omnia, South Africa

****Correspondence:***
Basil J. Nikolau
dimmas@iastate.edu

Specialty section:
This article was submitted to
Metabolomics,
a section of the journal
Frontiers in Molecular Biosciences

Received: *14 March 2022*
Accepted: *16 May 2022*
Published: *01 June 2022*

Citation:
Winkelman DC and Nikolau BJ (2022)
The Effects of Carbon Source and
Growth Temperature on the Fatty Acid
Profiles of Thermobifida fusca.
Front. Mol. Biosci. 9:896226.
doi: 10.3389/fmolb.2022.896226

INTRODUCTION

Plants possess the photosynthetic ability to chemically reduce atmospheric carbon dioxide and generate lignocellulosic biomass, providing the world with a feedstock that could be utilized for production of bio-based chemicals or biofuels (Zoghlami and Paës, 2019). Because plant lignocellulosic biomass can be derived from agricultural waste, it can serve as a feedstock without compromising global food security (Brethauer and Studer, 2014; Li et al., 2021). Fatty acids are a class of energy-dense biomolecules that are similar to petroleum-derived fuels and chemicals, making them potential replacements of fossil-carbon products currently in the marketplace if they can be produced from biorenewable feedstocks (Nikolau et al., 2008; Janssen and Steinbüchel, 2014; Shanks and Keeling, 2017). Unfortunately, this process is hindered by the composition of plant lignocellulosic biomass (i.e., a mixture of cellulose, hemicelluloses, and lignin), which is difficult to catabolize and naturally recalcitrant to microbial and enzymatic degradation (Zoghlami and Paës, 2019). Current methods for breaking down lignocellulosic biomass are costly as they require chemical pretreatments, which inhibit subsequent enzymatic catabolism and add to economic infeasibility (Isikgor and Becer, 2015). Several lignocellulosic degrading microbes are under consideration to serve in consolidated bioprocessing (CBP) strategies in an effort to lower costs (Xiong et al., 2018). A CBP approach would take advantage of a microbe's natural cellulolytic capabilities and allow simultaneous fermentation of derived sugar monomers to synthesize the desired bioproducts, such as fatty acids.

Thermobifida fusca is a thermophilic, cellulolytic *Actinobacterium* that is capable of breaking down lignocellulose. It naturally resides in warmer organic materials, including manure piles, compost heaps, and rotting hay (Mccarthy and Cross, 1984; Zhang et al., 1998). Its ability to hydrolyze plant biomass at higher temperatures (optimum growth at 50°C) and grow over a broad pH range makes it a prime candidate for larger scale CBP applications. The readily available *T. fusca* genome sequence reveals that it has the capacity to express many enzymes useful for hydrolyzing biomass, including numerous cellulases, xylanases, and carbohydrate transporters for sugar uptake (Lykidis et al., 2007). Many of these thermally stable enzymes have been heterologously expressed in alternative microbial hosts and analyzed for their applicability to biomass conversion (Ghangas and Wilson, 1987; Ali et al., 2015; Klinger et al., 2015; Saini et al., 2015; Zhao et al., 2015; Setter-Lamed et al., 2017; Yan and Fong, 2018; Ali et al., 2020).

Although *T. fusca* shows a high propensity to degrade plant biomass, little is known about the fatty acid products that it naturally produces or the fatty acid biosynthetic machinery that the microbe possesses. In this manuscript, we identify many of the fatty acid biosynthetic genes encoded by the *T. fusca* genome. Furthermore, we have determined the fatty acid profiles of *T. fusca* when it is grown on four different carbon sources (i.e., glucose, cellobiose, cellulose, and avicel) at both the optimal growth temperature (50°C) or at a suboptimal temperature (37°C). *T. fusca* has the ability to produce linear saturated and unsaturated fatty acids, but primarily produces a suite of branched-chain fatty acids (BCFAs), particularly *iso*-, *anteiso*-, and 10-methyl BCFAs that are primarily between 14- and 18-carbons in length. In addition, *T. fusca* fatty acid profiles can be affected by environmental changes associated with carbon source and growth temperature, with the latter being the more significant factor driving the fatty acid composition.

MATERIALS AND METHODS

Thermobifida fusca Media and Growth Conditions

T. fusca (strain BAA-629) was obtained from the American Type Culture Collection (ATCC) (Manassas, Virginia). The frozen stock was revived as directed by ATCC using their standard TYG 741 media, and cultures were incubated at 50°C. Experimental cultures were grown in 100 ml of Hagerdahl media (ATCC medium 2382) supplemented with 0.5% (w/v) of a carbon source: glucose, cellobiose, cellulose, or avicel. Each culture was initiated with 3 ml of inoculum and cultures were grown at either 37°C or 50°C for up to 2 days.

Harvesting *Thermobifida fusca* Cells

Cells were collected from each culture by centrifugation at 5000 × g for 5 min, and the wet weight of the cell pellet was determined. Cultures grown at the suboptimal temperature (37°C) did not consume all of the insoluble solid carbon source (i.e., cellulose or avicel). These cell pellets were washed with sterile water and the final, washed cell pellet was weighed. Cell pellets were flash frozen in liquid nitrogen, lyophilized for 48 h, and the dry weight was recorded prior to fatty acid extraction.

Fatty Acid Analysis

Fatty acids were extracted from three aliquots of cells taken from a *T. fusca* culture. Cells were pelleted and lyophilized. Lyophilized cell pellet aliquots (10 mg each) were transferred to a glass tube and spiked with 10 µg of nonadecanoic acid as an internal standard. The pellets were then suspended by vortexing for 15 min in a solution of 5% (v/v) sulfuric acid in methanol and the suspension was incubated at 80°C for 1 h. After cooling to room temperature, 1 ml of hexane: chloroform (4:1 v/v) solution and 1 ml of 0.9% (w/v) NaCl were added to each tube, and the mixture was vortexed for 5 min. The organic and aqueous phases were separated by centrifugation, and the organic phase containing the resulting fatty acid methyl esters was transferred to a separate test tube. The aqueous phase was extracted an additional time with hexane: chloroform (4:1 v/v) solution, and the organic phases were collected and pooled. The extracts containing fatty acid methyl esters were concentrated by evaporation under a stream of nitrogen gas. GC-MS analysis was performed with an Agilent 6890 GC equipped with a DB-1 MS capillary column (Agilent 122–0112). Chromatography was performed using helium gas at a flow-rate of 1.2 ml/min, using an inlet temperature set at 280°C. Individual fatty acid methyl esters were identified by GC-MS fragmentation patterns in tandem with NIST AMDIS software (Stein, 1999), as well as by comparing their retention times to known fatty acid methyl ester standards obtained from Supelco Inc. (Bellefonte, PA) and Metraya LLC (State College, PA). Peak areas of individual fatty acid methyl esters were integrated with AMDIS software, and these were converted to fatty acid concentrations relative to the peak area of the known amount of nonadecanoic acid internal standard that was added to each sample.

Principal Component Analysis

Principal component analysis was conducted using Metaboanalyst software (Pang et al., 2021) with data from all replicates of *T. fusca* cultures grown in each of the eight growing conditions. Quantitative fatty acid abundance data (µmoles/g dry cell weight) were uploaded from Microsoft Excel to Metaboanalyst statistical software and autoscaling was done as enabled by Metaboanalyst to create a PCA plot and a PCA biplot.

Computational Identification of Enzymatic Components of the *Thermobifida fusca* Fatty Acid Biosynthesis Machinery

The sequenced *T. fusca* genome was queried with the BLASTP algorithm (Altschul et al., 1990) using query sequences of experimentally confirmed acetyl-CoA carboxylase (ACCase) and Type II fatty acid synthase (FAS) components, either originating from *Actinomycetes* or from *Escherichia coli*

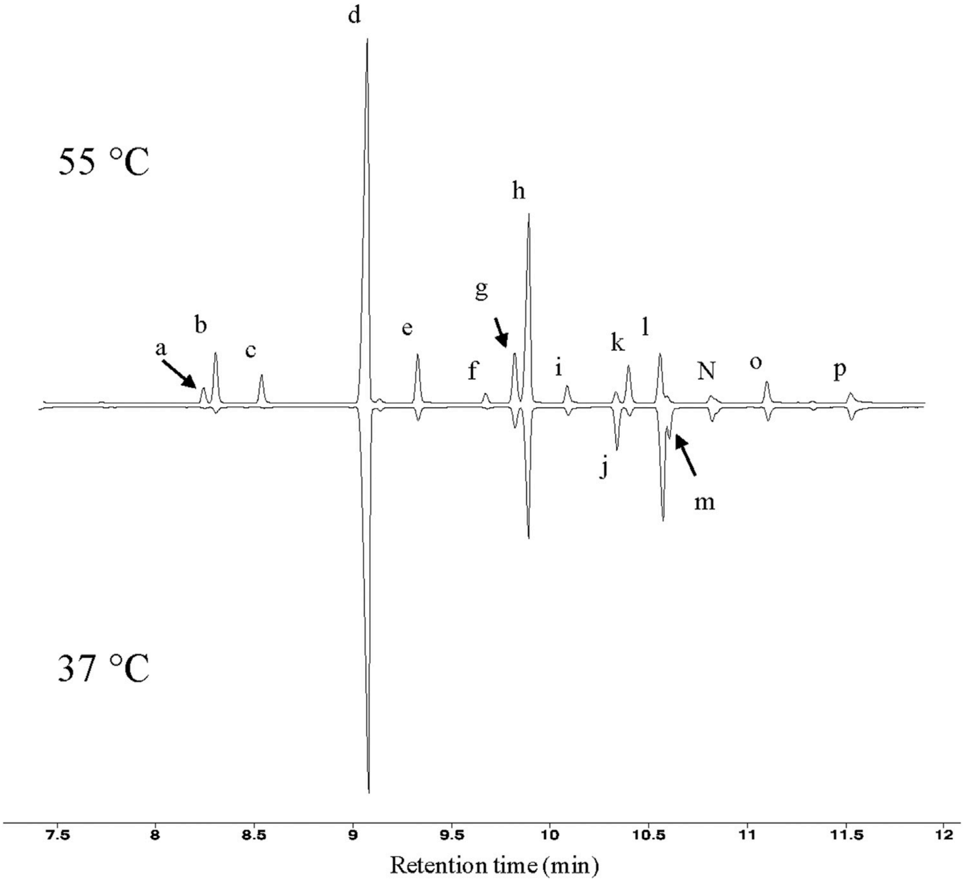


FIGURE 1 | GC profiles of *T. fusca* fatty acids. Typical GC profiles of fatty acid methyl esters isolated from *T. fusca* cultures grown on cellobiose at the indicated growth temperatures. Fatty acids were identified by mass-spectrometry and by comparing retention time with commercial standards. a = *iso*-15:0; b = *anteiso*-15:0; c = *n*-15:0; d = *iso*-16:0; e = *n*-16:0; f = 10-methyl-16:0; g = *iso*-17:0; h = *anteiso*-17:0; i = *n*-17:0; j = unknown; k = 10-methyl-17:0; l = *iso*-18:0; m = unknown; n = *n*-18:0; o = 10-methyl-18:0; *p* = *n*-19:0.

(Cronan and Thomas, 2009; Shivaiah et al., 2021). Additionally, genes were identified using the KEGG genome browser (Kanehisa et al., 2002) based on sequence homology, key processes, and operon organization.

RESULTS

Fatty Acids Produced by *Thermobifida fusca* at Optimal Growing Conditions

Fatty acid analyses show that when *T. fusca* was grown at 50°C it primarily synthesizes fatty acids that are between 14 and 18 carbons in length, although trace amounts of 13-carbon fatty acids were also detected. The majority of the fatty acids produced were BCFAs, particularly *iso*- or *anteiso*-BCFAs (i.e., the methyl branch is present at the ω-1 or ω-2 position of the acyl chain, respectively) (**Figure 1**). Small amounts of mid-chain BCFAs were also detected, these being 10-methyl BCFAs ranging between 16 and 18 carbons in length. Linear fatty acids were also present, accounting for approximately 10% of total fatty acids produced. Additionally, we determined that *T. fusca* can produce minor amounts (<5%) of mono-unsaturated C16 and C18 fatty acids with a single double bond at the 9th position.

Fatty Acid Profiles Are Affected by Carbon Source and Growth Temperature

Overall fatty acid yield was 2- to 3- fold higher when *T. fusca* was grown at the optimal growth temperature (50°C) as compared to growth at the suboptimal temperature (37°C) (**Figure 2**), and this phenomenon was observed independent of the carbon sources that were evaluated. Principal component analysis of the fatty acid compositional data visualized the factors contributing to the different fatty acid profiles of *T. fusca* through generation of a PCA plot (**Figure 3**). Principal component 1 accounts for ~52% of the sample variation, indicating that the samples are primarily separated by growth temperature, whereas carbon source contributed to ~26% of the sample variation (represented by principal component 2). Indeed, the data points in the PCA plot cluster distinctly by growth temperature, and to a lesser

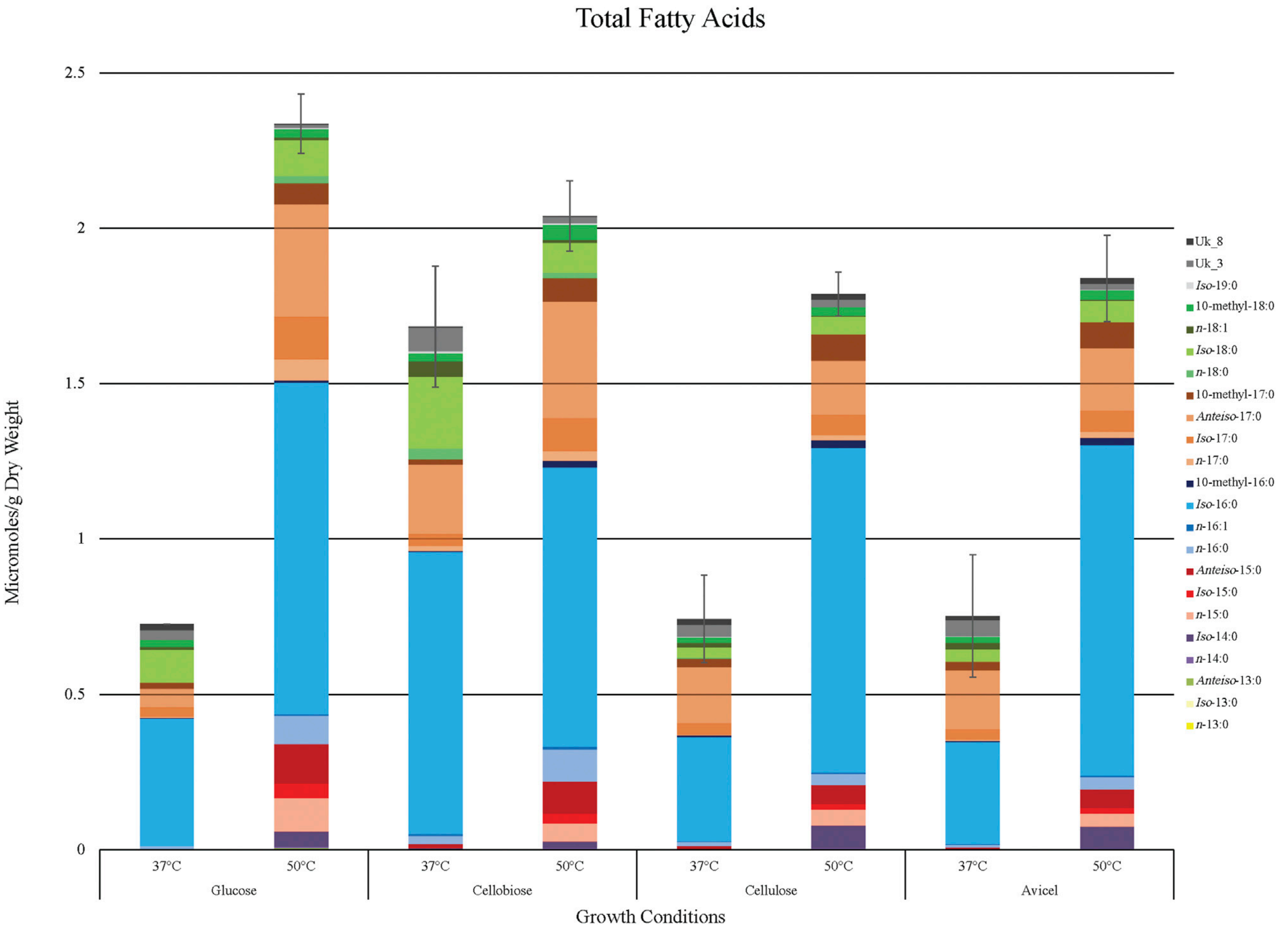


FIGURE 2 | Fatty acid yield generated by *T. fusca* in various growing conditions. Total accumulation of all identified fatty acid products (µmoles/g dry weight) when *T. fusca* was grown on glucose, cellobiose, cellulose, or Avicel as carbon source and cultured at either 37°C or 50°C, respectively. Fatty acid species are stacked in order of increasing chain length. Error bars represent standard error from three replicates.

extent by carbon source, with soluble and insoluble carbon sources typically clustering together within each growth temperature cluster. The conclusion that growth temperature is the primary driver of fatty acid composition is indicated by the PCA biplot (**Supplementary Figure S1**), as most of the vectors representing the different fatty acids are pointed horizontally, indicating that they contributed more to principal component 1. Additionally, some of the longer-chain fatty acids (18-carbons and longer) and the 10-methyl BCFAs are more associated with changes in the carbon source, as their vectors point more vertically.

The fatty acid profile of *T. fusca* shifted when it was cultured at its suboptimal temperature; the relative abundance of BCFAs was increased when the bacterium was cultured at 37°C (**Supplementary Figure S2**). Growth temperature also affected the acyl-chain lengths of the fatty acids produced by *T. fusca*. While C13, C14, and C15 fatty acid species account for approximately 10–15% of the fatty acids present at 50°C, they are almost completely absent from the cultures grown at 37°C (**Supplementary Figure S3**). Although unsaturated fatty acids were only detected at trace amounts at the optimum growth temperature, they make up a more significant proportion of the total fatty acid content when *T. fusca* was cultured at 37°C (~5%).

While growth temperature was the primary factor determining the types of fatty acids produced, carbon source also contributed to a shift in fatty acid profiles. Changes were primarily observed when comparing cultures grown on soluble versus insoluble carbon sources. Specifically, the fatty acid profiles of *T. fusca* grown on cellulose were very similar to those obtained when *T. fusca* was grown on Avicel, while fatty acid profiles of cultures grown on cellobiose resembled those grown on glucose. The main shift between the cultures grown on the soluble versus insoluble carbon sources can be attributed to the types of BCFAs present. While *iso*-BCFAs are the main species present in all growth conditions, when *T. fusca* was grown on soluble carbon sources we observed an even higher proportion of *iso*-BCFAs at 37°C. This relationship is reversed when using insoluble carbon sources; namely there is a higher proportion of *iso*-BCFAs at 50°C. In both cases,

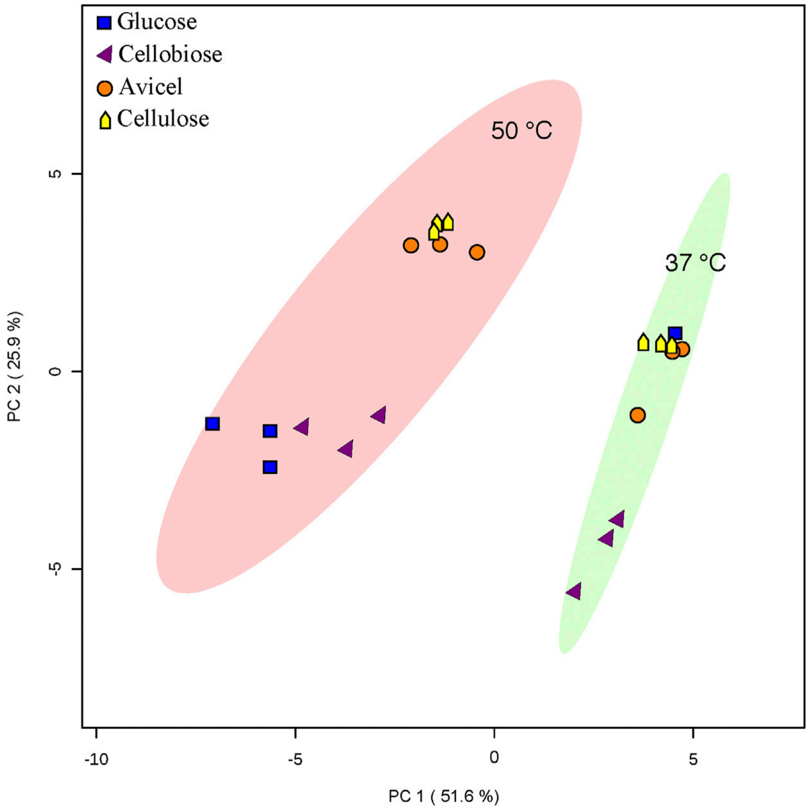


FIGURE 3 | PCA analysis. PCA analysis was conducted with Metaboanalyst software. An ellipse indicating 95% confidence regions for each heterotic group (37°C or 50°C growth temperature) is provided. Only one replicate of *T. fusca* supplemented with glucose at 37°C is depicted.

the change in abundance of *iso*-BCFAs is complemented with a corresponding change in the abundance of *anteiso*-BCFAs.

Identification of *Thermobifida fusca* Fatty Acid Biosynthesis Machinery

Although many of the enzymes involved in fatty acid biosynthesis have not been specifically characterized from *T. fusca*, they are identifiable by their sequence homology to enzymes from other bacteria, and by the operon organization in the *T. fusca* genome (**Table 1**). Querying the sequence of the *T. fusca* genome indicates that like many other bacteria, *T. fusca* utilizes a Type II fatty acid synthase (FAS) system to assemble fatty acids (Cronan and Thomas, 2009; Gago et al., 2011; Gago et al., 2018). In most organisms, FAS utilizes acetyl-CoA and malonyl-CoA as substrates, and the latter substrate is generated by the carboxylation of the former, a reaction catalyzed by acetyl-CoA carboxylase (Waldrop et al., 2012). Multiple iterations of the FAS cycle using these two substrates generates linear, saturated fatty acids, but these are minor components in *T. fusca*. The *iso*-and *anteiso*-BCFAs that account for a large portion of the fatty acids of *T. fusca* are produced by this FAS system by using branched-chain acyl-CoA substrates, rather than acetyl-CoA, which can be generated by the deamination of branched chain amino acids (i.e., valine, leucine, or isoleucine) (Kaneda, 1991; Beck et al., 2004; Zhu et al., 2005). Thus, the fatty acid biosynthetic machinery of *T. fusca* can be considered as consisting of at least four modules (**Figure 4**): 1) the module that generates the acyl-CoA starting substrate for FAS; 2) the module that generates the malonyl-CoA elongating substrate for FAS; 3) the FAS system itself; and 4) a fatty acid modifying module which generates the unsaturated and the internally BCFAs.

The acyl-CoA substrates required by this organism's FAS system are products of primary metabolism from sugars, generating acetyl-CoA, as well as products of branched chain amino acid metabolism, generating isobutyryl-CoA and 2-methylbutrylyl-CoA. Sequence homology identified multiple candidates for both the branched chain aminotransferase (Tfu_0616, Tfu_2112) and branched chain α-keto acid dehydrogenase (Tfu_0180, Tfu_0181, Tfu_0182) enzymes required to generate isobutyryl-CoA and 2-methylbutyryl-CoA from valine, leucine, or isoleucine. Alternatively, these branched chain acyl-CoAs may be generated from the α-keto acids that are intermediates of branched chain amino acid biosynthesis. Indeed, the branched chain aminotransferase encoded by Tfu_0616 is located in the genome adjoining genes that encode enzymes involved in branched chain amino acid biosynthesis, including genes with high sequence homology to acetolactate synthase (Tfu_0611, Tfu_0612), keto-acid isomeroreductase

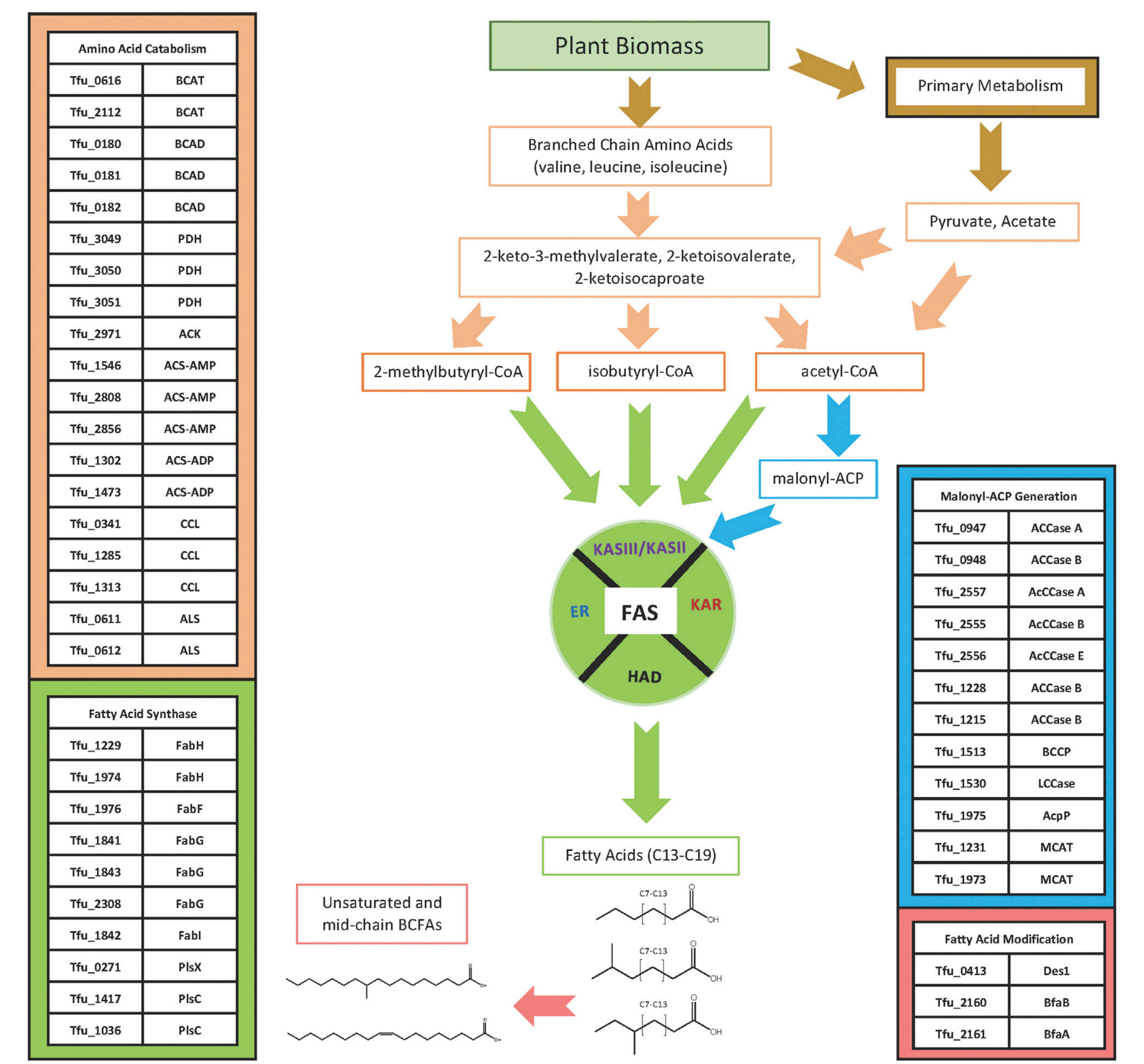
TABLE 1 | Identification of *T. fusca* fatty acid biosynthesis machinery. Genes were identified using the BLASTP algorithm using query sequences of experimentally confirmed enzymes from *E. coli* or from *Actinomycetes*.

Enzyme	Description	Gene name
ACCase A	ACCase BC and BCCP subunits	Tfu_0947
ACCase B	ACCase CT subunit	Tfu_0948
AcCCase A	AcCCase BC and BCCP subunit	Tfu_2557
AcCCase B	AcCCase CT subunit	Tfu_2555
AcCCase E	AcCCase E subunit	Tfu_2556
ACCase B	Additional ACCase CT subunit	Tfu_1228, Tfu_1215
BCCP	Biotin Carboxyl-Carrier Protein	Tfu_1513
LCCase	LCCase	Tfu_1530
AcpP	Acyl-carrier protein (ACP)	Tfu_1975
MCAT	Malonyl-CoA:ACP transacylase	Tfu_1231, Tfu_1973
FabH	3-ketoacyl-ACP synthase III	Tfu_1229, Tfu_1974
FabF	3-ketoacyl-ACP synthase III isozyme	Tfu_1976
FabG	3-ketoacyl-ACP reductase	Tfu_1841, Tfu_1843, Tfu_2308
FabA	3-hydroxyacyl-ACP dehydratase	Unknown
FabI	Enoyl-ACP reductase	Tfu_1842
PlsX	Glycerol-3-phosphate acyltransferase	Tfu_0271
PlsC	Glycerol-3-phosphate acyltransferase	Tfu_1417, Tfu_1036
BCAT	Branched chain amino acid aminotransferase	Tfu_0616, Tfu_2112
BCAD	Branched Chain alpha keto acid dehydrogenase	Tfu_0180, Tfu_0181, Tfu_0182
Des1	Delta-9 acyl-CoA desaturase	Tfu_0413
BfaB	Δ9 unsaturated fatty acid methyl transferase	Tfu_2160
BfaA	10-methylene BCFA reductase	Tfu_2161
PDH	Pyruvate Dehydrogenase Complex	Tfu_3049, Tfu_3050, Tfu_3051
ACK	Acetate Kinase	Tfu_2971
ACS-AMP	AMP-Forming Acetyl-CoA Synthetase	Tfu_1546, Tfu_2808, Tfu_2856
ACS-ADP	ADP-Forming Acetyl-CoA Synthetase	Tfu_1302, Tfu_1473
CCL	Citryl-CoA lyase	Tfu_0341, Tfu_1285, Tfu_1313
ALS	Acetolactate synthase	Tfu_0611, Tfu_0612

(Tfu_0613), 3-isopropylmalate dehydrogenase (Tfu_0615), 2-isopropylmalate synthase (Tfu_0617), and 3-isopropylmalate dehydratase (Tfu_0626, Tfu_0627). (Franco and Blanchard, 2017). Acetyl-CoA can be produced through several biological processes (Krivoruchko et al., 2015), including the oxidative decarboxylation of pyruvate catalyzed by the pyruvate dehydrogenase complex (PDH) (Tfu_3049, Tfu_3050, Tfu_3051). Alternatively, acetyl-CoA can be generated through the activation of acetate catalyzed by: 1) an acetate kinase (Tfu_2971); 2) an AMP-forming acetyl-CoA synthetase (Tfu_1546, Tfu_2808, Tfu_2856); or 3) an ADP-forming acetyl-CoA synthetase (Tfu_1302, Tfu_1473). *T. fusca* also possesses three genes that encode for proteins that resemble citryl-CoA lyase (Tfu_0341, Tfu_1285, Tfu_1313), a component of the reductive TCA cycle capable of converting citryl-CoA to acetyl-CoA and oxaloacetate (Aoshima et al., 2004; Hügler et al., 2005; Hügler et al., 2007; Katiyar et al., 2018).

The biotin-containing enzyme, acetyl-CoA carboxylase (ACCase) converts acetyl-CoA to malonyl-CoA, a reaction that is classically considered the first and rate-limiting reaction of fatty acid biosynthesis. As with all biotin enzymes, sequences of these proteins can be recognized by sequence homology among three different functional domains: the biotin carboxylase (BC), biotin carboxyl carrier protein (BCCP), and the carboxyl transferase (CT) domains (Cronan and Waldrop, 2002; Tong, 2013). The tertiary and quaternary organization of these domains varies considerably, depending on the phylogeny of the organism. *E. coli*, for examples, has ACCase components that are organized as individual proteins that come together to form the enzyme complex. In contrast, several ACCases from *Actinomycetes* consist of two subunits: the A subunit that encompasses both the BC and BCCP functional domains, and the B subunit that encompasses the CT domain (Gago et al., 2011; Gago et al., 2018). Additionally, some *Actinomycete* ACCases have a third non-catalytic subunit, E, that is needed for proper assembly of the holoenzyme complex (Shivaiah et al., 2021). Moreover, the A and B subunit quaternary organization of biotin enzymes is also common to propionyl-CoA carboxylase (Tong, 2013) and methylcrotonyl-CoA carboxylase (Song et al., 1994; McKean et al., 2000; Wurtele and Nikolau, 2000), which complicates the sequence-based identification of ACCase in the *T. fusca* genome.

Previous studies have experimentally characterized the *T. fusca* operon (Tfu_2555, Tfu_2556, Tfu_2557) that encodes the B, E, and A subunits of an acyl-CoA carboxylase (AcCCase) (Shivaiah et al., 2021). This enzyme is promiscuous and can carboxylate acetyl-CoA, propionyl-CoA, and butyryl-CoA. Other *Actinobacteria* (e.g., *Streptomyces coelicolor*) also express such a promiscuous



Amino Acid Catabolism	
Tfu_0616	BCAT
Tfu_2112	BCAT
Tfu_0180	BCAD
Tfu_0181	BCAD
Tfu_0182	BCAD
Tfu_3049	PDH
Tfu_3050	PDH
Tfu_3051	PDH
Tfu_2971	ACK
Tfu_1546	ACS-AMP
Tfu_2808	ACS-AMP
Tfu_2856	ACS-AMP
Tfu_1302	ACS-ADP
Tfu_1473	ACS-ADP
Tfu_0341	CCL
Tfu_1285	CCL
Tfu_1313	CCL
Tfu_0611	ALS
Tfu_0612	ALS

Fatty Acid Synthase	
Tfu_1229	FabH
Tfu_1974	FabH
Tfu_1976	FabF
Tfu_1841	FabG
Tfu_1843	FabG
Tfu_2308	FabG
Tfu_1842	FabI
Tfu_0271	PlsX
Tfu_1417	PlsC
Tfu_1036	PlsC

Malonyl-ACP Generation	
Tfu_0947	ACCase A
Tfu_0948	ACCase B
Tfu_2557	AcCCase A
Tfu_2555	AcCCase B
Tfu_2556	AcCCase E
Tfu_1228	ACCase B
Tfu_1215	ACCase B
Tfu_1513	BCCP
Tfu_1530	LCCase
Tfu_1975	AcpP
Tfu_1231	MCAT
Tfu_1973	MCAT

Fatty Acid Modification	
Tfu_0413	Des1
Tfu_2160	BfaB
Tfu_2161	BfaA

FIGURE 4 | Fatty acid biosynthesis pathway. Acyl-CoA starting substrates for FAS are generated from glycolytic catabolism of glucose or from α-keto acids that can be produced *via* the catabolism of branched-chain amino acids or as the penultimate intermediates in the biosynthesis of branched chain amino acids. The malonyl-ACP substrate used for the elongation reaction of FAS is synthesized from acetyl-CoA by ACCases. The acyl-CoA and malonyl-ACP substrates are used by the FAS system to elongate *iso-*, *ante-iso*, and linear *n*-fatty acids, which can be modified to produce unsaturated and 10-methyl branched chain fatty acids.

carboxylase, in addition to a highly specific propionyl-CoA carboxylase (Gago et al., 2018). The more promiscuous AcCCase enzymes can thereby generate not only malonyl-CoA, but also methylmalonyl-CoA and ethylmalonyl-CoA, and may therefore have multiple metabolic functions. For example, the catabolism of valine, isoleucine and odd-numbered fatty acids generates propionyl-CoA, which is further metabolized *via* the TCA cycle after the sequential conversion to methylmalonyl-CoA and succinyl-CoA (Wongkittichote et al., 2017). Alternatively, methylmalonyl-CoA and ethylmalonyl-CoA can be used as substrates by polyketide synthases, generating polyketides with methyl- or ethyl-branches in the final structure (Khosla and Keasling, 2003; Risdian et al., 2019).

Our BLAST-based search of the *T. fusca* genome identified additional genes that encode homologs of biotin-containing carboxylase proteins, namely Tfu_0947, Tfu_0948, Tfu_1215, Tfu_1228, Tfu_1513, and Tfu_1530. The adjacent Tfu_0947 and Tfu_0948 genes suggest that they are on a single operon, with Tfu_0947 encoding a subunit with the BC and BCCP domains, and Tfu_0948 encoding a subunit carrying the CT domain. This subunit/domain organization suggests that this operon may encode either a propionyl-CoA carboxylase (Gago et al., 2018) or methylcrotonyl-CoA carboxylase (Tomassetti et al., 2018), although ACCases with such quaternary subunit organizations also occur in many *Actinomycetes*, including *Streptomyces coelicolor* and *Mycobacterium tuberculosis* (Gago et al., 2011; Tong, 2013).

The Tfu_1228 and Tfu_1215 genes encode proteins that resemble CT subunits of biotin-carboxylases, but neither gene lies within an operon that houses other functional subunits necessary to form the holoenzyme complex. It is a common feature among *Actinomycetes* to mix and match different CT-subunits with a common BC/BCCP subunit, and thus generate different enzymatic capability with a single BC/BCCP subunit (Gago et al., 2011; Gago et al., 2018). The Tfu_1228 and Tfu_1215 genes may instill such a mechanism, and thus these genes could also provide a means for generating malonyl-CoA for FAS. A similar mechanism may occur in *T. fusca*, as Tfu_1228 and Tfu_1215 could be part of additional ACCase complexes that use an A subunit from another operon (such as Tfu_2557 or Tfu_0947). An additional gene with sequence homology to known ACCase components is Tfu_1513, which encodes a protein with high sequence homology to a BCCP that is not located near a BC or CT domain. Such a genome organization is similar to that found in *E. coli*, where the BC, BCCP, and CT components are separated into four individual proteins encoded by 3 separate operons (Cronan and Waldrop, 2002; Gago et al., 2018).

The Tfu_1530 gene encodes a large protein of 1849 residues, and it appears to encompass all three catalytic domains required for the carboxylation reaction (i.e., BC, BCCP and CT domains). This homomeric domain organization is common to such biotin carboxylases as ACCases from eukaryotes (i.e., plants, fungi, and animals) (Nikolau et al., 2003; Sasaki and Nagano, 2004), pyruvate carboxylases (Jitrapakdee and Wallace, 1999) and a long chain acyl-CoA carboxylase from *Mycobacterium* species (Tran et al., 2015; Lyonnet et al., 2017). The latter enzyme is involved in the biosynthesis of mycolic acid, a fatty acid specifically associated with the *Mycobacterium* genus (Marrakchi et al., 2014). Thus, the specific enzymatic function encoded by the Tfu_1530 gene is not recognizable by sequence homology but may include the ability to generate malonyl-CoA for FAS.

Type II FAS systems use cyclic iterations of four reactions that are each catalyzed by distinct enzymes (Cronan and Rock, 2008). Each cycle of the process adds two carbon atoms to the growing acyl-chain, with the donor of these two carbon subunits being the malonyl moiety of malonyl-CoA. The malonyl-moiety is first loaded onto the acyl-carrier protein (ACP) subunit of the FAS system, a reaction catalyzed by malonyl-CoA: ACP transacylase (MCAT). Subsequently, each FAS cycle begins with a Claisen condensation reaction between a preexisting “starting” acyl-CoA or acyl-ACP and malonyl-ACP to generate a 3-ketoacyl-ACP intermediate, which is 2-carbons longer than the initial acyl moiety. The first of these Claisen condensation reactions is between an acyl-CoA and malonyl-ACP, catalyzed by a 3-ketoacyl-ACP synthase III (encoded by the *FabH* gene). The chemical nature of the acyl-CoA substrate used in this condensation reaction determines the nature of the ω-end of the resulting fatty acid product; namely, utilizing acetyl-CoA, isobutyryl-CoA or methylbutyryl-CoA as the substrate leads to the generation of linear, *iso*-BCFA or *anteiso*-BCFA, respectively. The subsequent three reactions of each FAS cycle involve sequential reduction, dehydration and further reduction, catalyzed by 3-ketoacyl-ACP reductase (*FabG*), 3-hydroxyacyl-ACP dehydratase (*FabA*), and enoyl-ACP reductase (*FabI*), respectively. The product of each FAS cycle results in the generation of an acyl-ACP product that is two carbons longer than the pre-loaded acyl chain, and it serves as the substrate for the Claisen condensation reaction of the next round of the FAS cycle; these subsequent Claisen condensation reactions with a malonyl-ACP substrate are catalyzed by a 3-ketoacyl-ACP synthase II isozyme (*FabF*). These catalytic processes generate saturated fatty acids, and typically the process is terminated by transfer of the acyl moiety from acyl-ACP to glycerol-3-phosphate (glycerol-3-P), a reaction catalyzed by glycerol-3-P acyltransferases (GPATs) (Yao and Rock, 2013). The substrate specificity of GPATs determine the chain-lengths of the fatty acids produced by FAS, and in most bacteria they are typically of between 14 and 18 carbon atoms.

The *T. fusca* genome contains multiple operons that could encode for FAS genes, although the similarity to enzymatic functions associated with polyketide biosynthesis, catalyzed by Type II polyketide synthase or a nonribosomal peptide biosynthetic system (Corre and Challis, 2009), may confound these identifications. Specifically, Tfu_1973 (MCAT) is part of a large operon that includes Tfu_1974 (*FabH*), Tfu_1975 (*ACP*), and Tfu_1976 (*FabF*). Another copy of MCAT is encoded by Tfu_1231, which is also positioned near an additional copy of *FabH* (Tfu_1229). We found three *FabG* genes in the *T. fusca* genome that appear to encode for the 3-ketoacyl-ACP reductase. Two of these *FabG* genes, Tfu_1841 and Tfu_1843, are located in the same operon that also contains Tfu_1842 (*FabI*). A third copy of *FabG* (Tfu_2308), is found separately and does not appear to be part of a larger operon.

The fatty acid modification module includes genes that are required to generate unsaturated and 10-methyl BCFAs. Recent studies have indicated that *Actinomycetes* express a three-reaction pathway that metabolically links these two fatty acids. Specifically, a Δ-9 acyl-CoA desaturase (Tfu_0413) (Lykidis et al., 2007) can generate monounsaturated fatty acids, which are substrates for BfaB (Tfu_2160) and BfaA (Tfu_2161) enzymes, which transform the monounsaturated acyl-CoA to 10-methylene BCFA and to 10-methyl BCFA, respectively (Blitzblau et al., 2021).

DISCUSSION

While plants have presented the world with a large renewable feedstock of lignocellulosic biomass, the recalcitrant nature of this material has provided a challenge to make its utilization economically feasible (Isikgor and Becer, 2015; Zoghlami and Paës, 2019; Li et al., 2021). Fortunately, biological evolution has produced organisms capable of breaking down plant biomass (Brethauer and Studer, 2014; Saini et al., 2015; Xiong et al., 2018). One such organism that has received attention for its ability to catabolize biomass is *T. fusca* (Lykidis et al., 2007; del Pulgar and Saadeddin, 2014; Deng et al., 2016; Vanee et al., 2017). However, the breakdown of lignocellulosic biomass must be coupled to the conversion of the derived carbon to molecular structures that have utility as replacements of fossil-carbon based chemicals, fuels and materials. Biologically produced fatty acids have chemo-physical properties that make them highly desirable as substitutes of fossil-carbon products (Nikolau et al., 2008; Janssen and Steinbüchel, 2014; Shanks and Keeling, 2017). We have therefore evaluated the types of energy-dense fatty acid molecules that *T. fusca* is capable of producing.

Specifically, we have profiled the fatty acids that *T. fusca* produces in eight different growth conditions, and in parallel we have queried the sequenced genome of this organism to identify many of the genes that may be involved in the conversion of lignocellulosic-derived carbon to fatty acids. In these analyses fatty acids were chemically converted to methyl esters, which facilitated their subsequent GC-based identification and quantification. Because this conversion was based on transmethylation chemistry, which converts existing fatty acyl esters to methyl esters, we infer that the fatty acids that we profiled are acyl moieties of more complex lipids. Such lipids could include membrane associated polar lipids (e.g., phospholipids) or non-polar storage lipids (e.g., triacylglycerols). Although some *Actinomycetes* are capable of producing triacylglycerols, the absence of a gene for diacylglycerol acyltransferases in the *T. fusca* genome (Lykidis et al., 2007) that is necessary for triacylglycerol biosynthesis would preclude the occurrence of these lipids. Moreover, in *Actinomycetes,* triacylglycerols usually accumulate during the stationary phase of growth (Olukoshi and Packter, 1994; Alvarez and Steinbüchel, 2002; Comba et al., 2013; Santucci et al., 2019). Thus, we surmise that the fatty acids identified in this study are primarily membrane-associated phospholipids, which accumulate to facilitate the exponential growth of bacteria.

The primary fatty acids of *T. fusca* are three types of BCFAs: *iso*-branched, *anteiso*-branched, and mid-chain branched. We evaluated the metabolic responsiveness of *T. fusca* in relation to different growth temperature and different carbon sources and found that the fatty acid profiles shifted particularly in response to growth temperature. This temperature induced change in the fatty acid profile is consistent with the need to maintain membrane fluidity at the cooler temperature and results in increased proportion of BCFAs (Mansilla et al., 2004; Mendoza, 2014), although this shift in fatty acid composition is also dependent on carbon source. For example, when grown on glucose or cellobiose, there is a proportional increase in *iso*-BCFAs at 37°C. In many organisms, maintaining membrane fluidity at lower temperatures is achieved by increasing the degree of fatty acid unsaturation, which also occurs in *T. fusca*. Because the proportion of unsaturated fatty acids was always <5% of the total recovered fatty acids, it is difficult to envision that the unsaturated fatty acids are solely responsible for maintaining membrane fluidity at the cooler temperature. Thus, as with other microbial species (Klein et al., 1999; Saunders et al., 2016; Hassan et al., 2020), we suggest that the alteration in BCFAs is the main mechanism by which *T. fusca* maintains membrane fluidity at colder temperatures.

BFCAs have chemo-physical attributes that make them of particular interest for engineering applications. Specifically, the methyl-branch in the alkyl chain has the effect of lowering the melting point of the fatty acid, as compared to the linear-chain fatty acids of the same number of carbon atoms (Yao and Hammond, 2006). Therefore, BCFAs can maintain a fluid state at lower temperatures, which is desirable for biodiesel or lubrication application purposes in colder climates (Bart et al., 2013). Although unsaturated fatty acids can be used for these purposes, they are more susceptible to oxidation in these applications that expose them to oxygen at higher temperatures and pressures, as occurs in combustion engines. Thus, *T. fusca* can be viewed as a source of saturated BCFAs, with enhanced application potential as biofuels and biolubricants.

Despite this potential for producing desirable bioproducts, the yield of fatty acids from *T. fusca* cultures is relatively low as compared to such oleaginous organisms, such as plant oil seed crops (e.g., sunflower, canola, safflower), yeasts (e.g., *Yarrowia lipolytica* and *Rhodosporidium toruloides*) and microalgae (e.g., *Botryococcus braunii*). These later organisms hyperaccumulate fatty acid containing neutral lipids that can account for 20–60% of the dry biomass (Banerjee et al., 2002; Sharafi et al., 2015; Patel et al., 2019; Liu H. et al., 2021; Liu Y. et al., 2021; Petraru et al., 2021; Zemour et al., 2021). In contrast, *T. fusca* accumulates fatty acids at approximately 1 mg/g dry biomass. Increasing fatty acid titers from non-model organisms such as *T. fusca* is becoming increasingly viable, either by mutagenesis and selection strategies (Guo et al., 2019; Südfeld et al., 2021), or direct targeted genetic engineering strategies (Zhang et al., 2019; Racharaks et al., 2021). The latter strategy needs prior knowledge of the fatty acid biosynthesis machinery that would be targeted for genetic engineering, whereas the former strategy can be informative of regulatory mechanisms, once mutant alleles can be genetically mapped relative to the fatty acid biosynthesis machinery encoded by the target genome. We therefore examined the *T. fusca* genome to identify the components of the fatty acid biosynthesis machinery, which would facilitate these strategies for improving fatty acid yields.

T. fusca appears to possess much of the fatty acid biosynthesis machinery that is common to other bacteria, particularly *Actinomycetes* (Gago et al., 2011; Lyonnet et al., 2017; Gago et al., 2018). At the core is the Type II FAS system, which appears to be encoded primarily within two operons, encompassing the genes Tfu_1973 to Tfu_1975 and Tfu_1841 to Tfu_1843. Additional homologs of some of the FAS enzymatic components are also encoded in the genome, but these may be associated with a polyketide synthase system or the nonribosomal peptide biosynthetic system responsible for the biosynthesis of the siderophore, fuscachelins (Dimise et al., 2008; Corre and Challis, 2009).

The FAS system needs to be provided with four different substrates: acetyl-CoA, isobutyryl-CoA, 2-methylbutyryl-CoA and malonyl-CoA. The former substrates are used to initiate FAS reactions, and thereby generate linear fatty acids and two types of BCFAs; whereas malonyl-CoA is the substrate that is used to elongate the fatty acid in the FAS catalyzed reaction. We identified five potential enzymatic systems that can generate acetyl-CoA: the pyruvate dehydrogenase complex, AMP-forming acetyl-CoA synthetase, ADP-forming acetyl-CoA synthetase, acetate kinase, and citryl-CoA lyase. These systems give *T. fusca* the ability to produce acetyl-CoA from multiple carbon sources, including pyruvate and acetate. Citryl-CoA lyase represents the final step in the reductive TCA cycle that also requires a 2-oxoglutarate: ferredoxin oxidoreductase (Tfu_2674, Tfu_2675). The presence of these enzymes potentially gives *T. fusca* the ability to synthesize acetyl-CoA from carbon dioxide (Aoshima et al., 2004; Hügler et al., 2005), a process that has previously been explored in *Mycobacterium tuberculosis* (Boshoff and Barry, 2005; Watanabe et al., 2011; Katiyar et al., 2018). The reductive TCA cycle can be considered the reversal of the oxidative TCA cycle that is prevalent in aerobic organisms and oxidizes carbon from acetyl-CoA to generate carbon dioxide. The reductive TCA cycle is considered to be a primordial pathway for carbon dioxide fixation facilitating autotrophic metabolism in the earliest life-forms on earth, possibly prior to the advent of the currently prevalent photosynthetic pentose phosphate cycle (Buchanan and Arnon, 1990; Romano and Conway, 1996; Smith and Morowitz, 2004). Thus, *T. fusca* has the metabolic flexibility for generating this important intermediate in metabolism, and moreover these mechanisms provide opportunities for improving acetyl-CoA generation *via* genetic engineering strategies, as has been done in a number of microbial systems (Lian et al., 2014; Liu et al., 2017; Soma et al., 2017; Ku et al., 2020).

Several biotin-containing enzyme systems were identifiable *via* sequence homology, which could generate the malonyl-CoA substrate that is used by FAS to elongate the fatty acid. These enzymes carboxylate acyl-CoA substrates, including acetyl-CoA, and thereby generate malonyl-CoA (Nikolau et al., 2003), which is considered to be the rate-limiting reaction for fatty acid biosynthesis (Davis et al., 2000; Chaturvedi et al., 2021). Prior experimental characterization has identified one of these biotin-containing enzymes as being capable of carboxylating acetyl-CoA, propionyl-CoA, or butyryl-CoA (Shivaiah et al., 2021), but whether *T. fusca* encodes other proteins that can also carboxylate acetyl-CoA will require similar characterizations.

Finally, the *T. fusca* genome features enzymes capable of modifying fatty acids following their assembly by the FAS system. These include an acyl-CoA desaturase to produce unsaturated fatty acids, which are substrates for a methylase and reductase needed to generate 10-methyl BCFAs (Blitzblau et al., 2021). These three enzymes, in addition to the flexibility of the FAS system to use an assortment of α-carboxyl acyl-CoAs, indicate that *T. fusca* possesses the metabolic machinery to convert plant biomass to a wide array of fatty acids that have applications as biorenewable products. In particular, *T. fusc*a appears to be an organism that can generate unique combinations of BCFAs, including those with a mid-chain branch, which have potential applications as feedstocks for novel bioproducts, such as bio-lubricants (Blitzblau et al., 2021).

Collectively the fatty acid profiling data integrated with the genomics data identified the genetic potential of *T. fusca*. Additional data that evaluates the expression of this genetic potential would precisely deduce the catalytic and regulatory circuit(s) that define the metabolic system that generates the diversity of fatty acids produced by this organism. Such genetic expression data will require extensive steady state transcriptomics, proteomics and metabolomics data, integrated with flux analyses to further expand the comprehension of the metabolic system that converts sugar-feedstocks to fatty acids, and thereby make *T. fusca* an even more attractive candidate to produce energy-dense biomolecules in a consolidated bioprocessing system.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

DW and BJN both contributed to design of this study as well as manuscript preparation and editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.896226/full#supplementary-material

Supplementary Figure S1 | PCA biplot of T. fusca fatty acid data. Overlay of PCA score plot and loading plot. Cultures numbered 1-3 were grown at 37°C, while cultures numbered 4-6 were grown at 50°C. CB, cellobiose, GL, glucose, CL, cellulose, AV, avicel. Only a single determination of *T. fusca* supplemented with glucose at 37°C is depicted.

Supplementary Figure S2 | Proportion of fatty acid classes identified in *T. fusca* cultures. Mole percent of total fatty acids produced by *T. fusca* in different growth conditions, organized by lipid species. "*n*-FA" indicates linear fatty acid chain.

Supplementary Figure S3 | Proportion of fatty acid chain lengths identified in *T. fusca* cultures. Mole percent of total fatty acids produced by *T. fusca* in different growth conditions, organized by chain length.

REFERENCES

Ali, I., Asghar, R., Ahmed, S., Sajjad, M., Tariq, M., and Akhtar, M. W. (2015). mRNA Secondary Structure Engineering of Thermobifida Fusca Endoglucanase (Cel6A) for Enhanced Expression in *Escherichia coli. World J. Microbiol. Biotechnol.* 31 (3), 499–506. doi:10.1007/s11274-015-1806-5

Ali, I., Rehman, H. M., Mirza, M. U., Akhtar, M. W., Asghar, R., Tariq, M., et al. (2020). Enhanced Thermostability and Enzymatic Activity of cel6A Variants from Thermobifida Fusca by Empirical Domain Engineering. *Biology* 9 (8), 214. doi:10.3390/biology9080214

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* 215 (3), 403–410. doi:10.1016/s0022-2836(05)80360-2

Alvarez, H., and Steinbüchel, A. (2002). Triacylglycerols in Prokaryotic Microorganisms. *Appl. Microbiol. Biotechnol.* 60 (4), 367–376. doi:10.1007/s00253-002-1135-0

Aoshima, M., Ishii, M., and Igarashi, Y. (2004). A Novel Enzyme, Citryl-CoA Lyase, Catalysing the Second Step of the Citrate Cleavage Reaction in Hydrogenobacter Thermophilus TK-6. *Mol. Microbiol.* 52 (3), 763–770. doi:10.1111/j.1365-2958.2004.04010.x

Banerjee, A., Sharma, R., Chisti, Y., and Banerjee, U. C. (2002). Botryococcus Braunii: A Renewable Source of Hydrocarbons and Other Chemicals. *Crit. Rev. Biotechnol.* 22 (3), 245–279. doi:10.1080/07388550290789513

Bart, J. C. J., Gucciardi, E., and Cavallaro, S. (2013). "6 - Chemical Transformations of Renewable Lubricant Feedstocks," in *Biolubricants*. Editors J.C.J. Bart, E. Gucciardi, and S. Cavallaro (Sawston, UK: Woodhead Publishing), 249–350. doi:10.1533/9780857096326.249

Beck, H. C., Hansen, A. M., and Lauritsen, F. R. (2004). Catabolism of Leucine to Branched-Chain Fatty Acids in Staphylococcus Xylosus. *J. Appl. Microbiol.* 96 (5), 1185–1193. doi:10.1111/j.1365-2672.2004.02253.x

Blitzblau, H. G., Consiglio, A. L., Teixeira, P., Crabtree, D. V., Chen, S., Konzock, O., et al. (2021). Production of 10-methyl Branched Fatty Acids in Yeast. *Biotechnol. Biofuels* 14 (1), 12. doi:10.1186/s13068-020-01863-0

Boshoff, H. I. M., and Barry, C. E. (2005). Tuberculosis - Metabolism and Respiration in the Absence of Growth. *Nat. Rev. Microbiol.* 3 (1), 70–80. doi:10.1038/nrmicro1065

Brethauer, S., and Studer, M. H. (2014). Consolidated Bioprocessing of Lignocellulose by a Microbial Consortium. *Energy Environ. Sci.* 7 (4), 1446–1453. doi:10.1039/C3EE41753K

Buchanan, B. B., and Arnon, D. I. (1990). A Reverse KREBS Cycle in Photosynthesis: Consensus at Last. *Photosynth Res.* 24, 47–53. doi:10.1007/bf00032643

Chaturvedi, S., Gupta, A. K., Bhattacharya, A., Dutta, T., Nain, L., and Khare, S. K. (2021). Overexpression and Repression of Key Rate-limiting Enzymes (Acetyl CoA Carboxylase and HMG Reductase) to Enhance Fatty Acid Production from Rhodotorula Mucilaginosa. *J. Basic Microbiol.* 61 (1), 4–14. doi:10.1002/jobm.202000407

Comba, S., Menendez-Bravo, S., Arabolaza, A., and Gramajo, H. (2013). Identification and Physiological Characterization of Phosphatidic Acid Phosphatase Enzymes Involved in Triacylglycerol Biosynthesis in Streptomyces Coelicolor. *Microb. Cell Fact.* 12 (1), 9. doi:10.1186/1475-2859-12-9

Corre, C., and Challis, G. L. (2009). New Natural Product Biosynthetic Chemistry Discovered by Genome Mining. *Nat. Prod. Rep.* 26 (8), 977–986. doi:10.1039/B713024B

Cronan, J. E., Jr., and Waldrop, G. L. (2002). Multi-subunit Acetyl-CoA Carboxylases. *Prog. Lipid Res.* 41 (5), 407–435. doi:10.1016/s0163-7827(02)00007-3

Cronan, J. E., Jr., and Rock, C. O. (2008). Biosynthesis of Membrane Lipids. *EcoSal Plus* 3 (1). doi:10.1128/ecosalplus.3.6.4

Cronan, J. E., and Thomas, J. (2009). Chapter 17 Bacterial Fatty Acid Synthesis and its Relationships with Polyketide Synthetic Pathways. *Methods Enzym.* 459, 395–433. doi:10.1016/S0076-6879(09)04617-5

Davis, M. S., Solbiati, J., and Cronan, J. E., Jr. (2000). Overproduction of Acetyl-CoA Carboxylase Activity Increases the Rate of Fatty Acid Biosynthesis in *Escherichia coli*. *J. Biol. Chem.* 275 (37), 28593–28598. doi:10.1074/jbc.M004756200

del Pulgar, E. M. G., and Saadeddin, A. (2014). The Cellulolytic System ofThermobifida Fusca. *Crit. Rev. Microbiol.* 40 (3), 236–247. doi:10.3109/1040841x.2013.776512

Deng, Y., Mao, Y., and Zhang, X. (2016). Metabolic Engineering of a Laboratory-evolvedThermobifida fuscamuC Strain for Malic Acid Production on Cellulose and Minimal Treated Lignocellulosic Biomass. *Biotechnol. Prog.* 32 (1), 14–20. doi:10.1002/btpr.2180

Dimise, E. J., Widboom, P. F., and Bruner, S. D. (2008). Structure Elucidation and Biosynthesis of Fuscachelins, Peptide Siderophores from the Moderate Thermophile Thermobifida Fusca. *Proc. Natl. Acad. Sci. U.S.A.* 105 (40), 15311–15316. doi:10.1073/pnas.0805451105

Franco, T. M. A., and Blanchard, J. S. (2017). Bacterial Branched-Chain Amino Acid Biosynthesis: Structures, Mechanisms, and Drugability. *Biochemistry* 56 (44), 5849–5865. doi:10.1021/acs.biochem.7b00849

Gago, G., Arabolaza, A., Diacovich, L., and Gramajo, H. (2018). "Components and Key Regulatory Steps of Lipid Biosynthesis in Actinomycetes," in *Biogenesis of Fatty Acids, Lipids and Membranes*. Editor O. Geiger (Cham: Springer International Publishing), 1–25. doi:10.1007/978-3-319-43676-0_65-1

Gago, G., Diacovich, L., Arabolaza, A., Tsai, S.-C., and Gramajo, H. (2011). Fatty Acid Biosynthesis in Actinomycetes. *FEMS Microbiol. Rev.* 35 (3), 475–497. doi:10.1111/j.1574-6976.2010.00259.x

Ghangas, G. S., and Wilson, D. B. (1987). Expression of a Thermomonospora Fusca Cellulase Gene in Streptomyces Lividans and Bacillus Subtilis. *Appl. Environ. Microbiol.* 53 (7), 1470–1475. doi:10.1128/aem.53.7.1470-1475.1987

Guo, M., Cheng, S., Chen, G., and Chen, J. (2019). Improvement of Lipid Production in Oleaginous Yeast Rhodosporidium Toruloides by Ultraviolet Mutagenesis. *Eng. Life Sci.* 19 (8), 548–556. doi:10.1002/elsc.201800203

Hassan, N., Anesio, A. M., Rafiq, M., Holtvoeth, J., Bull, I., Haleem, A., et al. (2020). Temperature Driven Membrane Lipid Adaptation in Glacial Psychrophilic Bacteria. *Front. Microbiol.* 11, 824. doi:10.3389/fmicb.2020.00824

Hügler, M., Huber, H., Molyneaux, S. J., Vetriani, C., and Sievert, S. M. (2007). Autotrophic CO_2fixation via the Reductive Tricarboxylic Acid Cycle in Different Lineages within the Phylum Aquificae: Evidence for Two Ways of Citrate Cleavage. *Environ. Microbiol.* 9 (1), 81–92. doi:10.1111/j.1462-2920.2006.01118.x

Hügler, M., Wirsen, C. O., Fuchs, G., Taylor, C. D., and Sievert, S. M. (2005). Evidence for Autotrophic CO_2 Fixation via the Reductive Tricarboxylic Acid Cycle by Members of the ε Subdivision of Proteobacteria. *J. Bacteriol.* 187 (9), 3020–3027. doi:10.1128/JB.187.9.3020-3027.2005

Isikgor, F. H., and Becer, C. R. (2015). Lignocellulosic Biomass: A Sustainable Platform for the Production of Bio-Based Chemicals and Polymers. *Polym. Chem.* 6 (25), 4497–4559. doi:10.1039/C5PY00263J

Janssen, H., and Steinbüchel, A. (2014). Fatty Acid Synthesis in *Escherichia coli* and its Applications towards the Production of Fatty Acid Based Biofuels. *Biotechnol. Biofuels* 7 (1), 7. doi:10.1186/1754-6834-7-7

Jitrapakdee, S., and Wallace, J. C. (1999). Structure, Function and Regulation of Pyruvate Carboxylase. *Biochem. J.* 340 (Pt 1), 1–16. doi:10.1042/bj3400001

Kaneda, T. (1991). Iso- and Anteiso-Fatty Acids in Bacteria: Biosynthesis, Function, and Taxonomic Significance. *Microbiol. Rev.* 55 (2), 288–302. doi:10.1128/mr.55.2.288-302.1991

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002). The KEGG Databases at GenomeNet. *Nucleic Acids Res.* 30 (1), 42–46. doi:10.1093/nar/30.1.42

Katiyar, A., Singh, H., and Azad, K. K. (2018). Identification of Missing Carbon Fixation Enzymes as Potential Drug Targets in Mycobacterium Tuberculosis. *J. Integr. Bioinforma.* 15 (3), 20170041. doi:10.1515/jib-2017-0041

Khosla, C., and Keasling, J. D. (2003). Metabolic Engineering for Drug Discovery and Development. *Nat. Rev. Drug Discov.* 2 (12), 1019–1025. doi:10.1038/nrd1256

Klein, W., Weber, M. H. W., and Marahiel, M. A. (1999). Cold Shock Response of *Bacillus Subtilis* : Isoleucine-dependent Switch in the Fatty Acid Branching Pattern for Membrane Adaptation to Low Temperatures. *J. Bacteriol.* 181 (17), 5341–5349. doi:10.1128/JB.181.17.5341-5349.1999

Klinger, J., Fischer, R., and Commandeur, U. (2015). Comparison of Thermobifida Fusca Cellulases Expressed in *Escherichia coli* and Nicotiana Tabacum Indicates Advantages of the Plant System for the Expression of Bacterial Cellulases. *Front. Plant Sci.* 6, 1047. doi:10.3389/fpls.2015.01047

Krivoruchko, A., Zhang, Y., Siewers, V., Chen, Y., and Nielsen, J. (2015). Microbial Acetyl-CoA Metabolism and Metabolic Engineering. *Metab. Eng.* 28, 28–42. doi:10.1016/j.ymben.2014.11.009

Ku, J. T., Chen, A. Y., and Lan, E. I. (2020). Metabolic Engineering Design Strategies for Increasing Acetyl-CoA Flux. *Metabolites* 10 (4), 166. doi:10.3390/metabo10040166

Li, C., Lin, X., Ling, X., Li, S., and Fang, H. (2021). Consolidated Bioprocessing of Lignocellulose for Production of Glucaric Acid by an Artificial Microbial Consortium. *Biotechnol. Biofuels* 14 (1), 110. doi:10.1186/s13068-021-01961-7

Lian, J., Si, T., Nair, N. U., and Zhao, H. (2014). Design and Construction of Acetyl-CoA Overproducing *Saccharomyces cerevisiae* Strains. *Metab. Eng.* 24, 139–149. doi:10.1016/j.ymben.2014.05.010

Liu, H., Song, Y., Fan, X., Wang, C., Lu, X., and Tian, Y. (2021). Yarrowia Lipolytica as an Oleaginous Platform for the Production of Value-Added Fatty Acid-Based Bioproducts. *Front. Microbiol.* 11, 608662. doi:10.3389/fmicb.2020.608662

Liu, W., Zhang, B., and Jiang, R. (2017). Improving Acetyl-CoA Biosynthesis in *Saccharomyces cerevisiae* via the Overexpression of Pantothenate Kinase and PDH Bypass. *Biotechnol. Biofuels* 10 (1), 41. doi:10.1186/s13068-017-0726-z

Liu, Y., Koh, C. M. J., Yap, S. A., Cai, L., and Ji, L. (2021). Understanding and Exploiting the Fatty Acid Desaturation System in Rhodotorula Toruloides. *Biotechnol. Biofuels* 14 (1), 73. doi:10.1186/s13068-021-01924-y

Lykidis, A., Mavromatis, K., Ivanova, N., Anderson, I., Land, M., DiBartolo, G., et al. (2007). Genome Sequence and Analysis of the Soil Cellulolytic Actinomycete *Thermobifida Fusca* YX. *J. Bacteriol.* 189 (6), 2477–2486. doi:10.1128/JB.01899-06

Lyonnet, B. B., Diacovich, L., Gago, G., Spina, L., Bardou, F., Lemassu, A., et al. (2017). Functional Reconstitution of the *Mycobacterium tuberculosis* Long-chain acyl-CoA Carboxylase from Multiple acyl-CoA Subunits. *Febs J.* 284 (7), 1110–1125. doi:10.1111/febs.14046

Mansilla, M. C., Cybulski, L. E., Albanesi, D., and de Mendoza, D. (2004). Control of Membrane Lipid Fluidity by Molecular Thermosensors. *J. Bacteriol.* 186 (20), 6681–6688. doi:10.1128/jb.186.20.6681-6688.2004

Marrakchi, H., Lanéelle, M.-A., and Daffé, M. (2014). Mycolic Acids: Structures, Biosynthesis, and beyond. *Chem. Biol.* 21 (1), 67–85. doi:10.1016/j.chembiol.2013.11.011

Mccarthy, A. J., and Cross, T. (1984). A Taxonomic Study of Thermomonospora and Other Monosporic Actinomycetes. *Microbiology* 130 (1), 5–25. doi:10.1099/00221287-130-1-5

McKean, A. L., Ke, J., Song, J., Che, P., Achenbach, S., Nikolau, B. J., et al. (2000). Molecular Characterization of the Non-biotin-containing Subunit of 3-Methylcrotonyl-CoA Carboxylase. *J. Biol. Chem.* 275 (8), 5582–5590. doi:10.1074/jbc.275.8.5582

Mendoza, D. d. (2014). Temperature Sensing by Membranes. *Annu. Rev. Microbiol.* 68 (1), 101–116. doi:10.1146/annurev-micro-091313-103612

Nikolau, B. J., Ohlrogge, J. B., and Wurtele, E. S. (2003). Plant Biotin-Containing Carboxylases. *Archives Biochem. Biophysics* 414 (2), 211–222. doi:10.1016/S0003-9861(03)00156-5

Nikolau, B. J., Perera, M. A. D. N., Brachova, L., and Shanks, B. (2008). Platform Biochemicals for a Biorenewable Chemical Industry. *Plant J.* 54 (4), 536–545. doi:10.1111/j.1365-313X.2008.03484.x

Olukoshi, E. R., and Packter, N. M. (1994). Importance of Stored Triacylglycerols in Streptomyces: Possible Carbon Source for Antibiotics. *Microbiology* 140 (Pt 4), 931–943. doi:10.1099/00221287-140-4-931

Pang, Z., Chong, J., Zhou, G., de Lima Morais, D. A., Chang, L., Barrette, M., et al. (2021). MetaboAnalyst 5.0: Narrowing the Gap between Raw Spectra and Functional Insights. *Nucleic Acids Res.* 49 (W1), W388–W396. doi:10.1093/nar/gkab382

Patel, A., Antonopoulou, I., Enman, J., Rova, U., Christakopoulos, P., and Matsakas, L. (2019). Lipids Detection and Quantification in Oleaginous Microorganisms: An Overview of the Current State of the Art. *BMC Chem. Eng.* 1 (1), 13. doi:10.1186/s42480-019-0013-9

Petraru, A., Ursachi, F., and Amariei, S. (2021). Nutritional Characteristics Assessment of Sunflower Seeds, Oil and Cake. Perspective of Using Sunflower Oilcakes as a Functional Ingredient. *Plants* 10 (11), 2487. doi:10.3390/plants10112487

Racharaks, R., Arnold, W., and Peccia, J. (2021). Development of CRISPR-Cas9 Knock-In Tools for Free Fatty Acid Production Using the Fast-Growing Cyanobacterial Strain Synechococcus Elongatus UTEX 2973. *J. Microbiol. Methods* 189, 106315. doi:10.1016/j.mimet.2021.106315

Risdian, C., Mozef, T., and Wink, J. (2019). Biosynthesis of Polyketides in Streptomyces. *Microorganisms* 7 (5), 124. doi:10.3390/microorganisms7050124

Romano, A. H., and Conway, T. (1996). Evolution of Carbohydrate Metabolic Pathways. *Res. Microbiol.* 147 (6), 448–455. doi:10.1016/0923-2508(96)83998-2

Saini, A., Aggarwal, N. K., Sharma, A., and Yadav, A. (2015). Actinomycetes: A Source of Lignocellulolytic Enzymes. *Enzyme Res.* 2015, 1–15. doi:10.1155/2015/279381

Santucci, P., Johansen, M. D., Point, V., Poncin, I., Viljoen, A., Cavalier, J.-F., et al. (2019). Nitrogen Deprivation Induces Triacylglycerol Accumulation, Drug Tolerance and Hypervirulence in Mycobacteria. *Sci. Rep.* 9 (1), 8667. doi:10.1038/s41598-019-45164-5

Sasaki, Y., and Nagano, Y. (2004). Plant Acetyl-CoA Carboxylase: Structure, Biosynthesis, Regulation, and Gene Manipulation for Plant Breeding. *Biosci. Biotechnol. Biochem.* 68 (6), 1175–1184. doi:10.1271/bbb.68.1175

Saunders, L. P., Sen, S., Wilkinson, B. J., and Gatto, C. (2016). Insights into the Mechanism of Homeoviscous Adaptation to Low Temperature in Branched-Chain Fatty Acid-Containing Bacteria through Modeling FabH Kinetics from the Foodborne Pathogen Listeria Monocytogenes. *Front. Microbiol.* 7, 1386. doi:10.3389/fmicb.2016.01386

Setter-Lamed, E., Moraïs, S., Stern, J., Lamed, R., and Bayer, E. A. (2017). Modular Organization of the Thermobifida Fusca Exoglucanase Cel6B Impacts Cellulose Hydrolysis and Designer Cellulosome Efficiency. *Biotechnol. J.* 12 (10), 1700205. doi:10.1002/biot.201700205

Shanks, B. H., and Keeling, P. L. (2017). Bioprivileged Molecules: Creating Value from Biomass. *Green Chem.* 19 (14), 3177–3185. doi:10.1039/C7GC00296C

Sharafi, Y., Majidi, M. M., Goli, S. A. H., and Rashidi, F. (2015). Oil Content and Fatty Acids Composition inBrassicaSpecies. *Int. J. Food Prop.* 18 (10), 2145–2154. doi:10.1080/10942912.2014.968284

Shivaiah, K.-K., Upton, B., and Nikolau, B. J. (2021). Kinetic, Structural, and Mutational Analysis of Acyl-CoA Carboxylase from Thermobifida Fusca YX. *Front. Mol. Biosci.* 7, 615614. doi:10.3389/fmolb.2020.615614

Smith, E., and Morowitz, H. J. (2004). Universality in Intermediary Metabolism. *Proc. Natl. Acad. Sci. U.S.A.* 101 (36), 13168–13173. doi:10.1073/pnas.0404922101

Soma, Y., Yamaji, T., Matsuda, F., and Hanai, T. (2017). Synthetic Metabolic Bypass for a Metabolic Toggle Switch Enhances Acetyl-CoA Supply for Isopropanol Production by *Escherichia coli*. *J. Biosci. Bioeng.* 123 (5), 625–633. doi:10.1016/j.jbiosc.2016.12.009

Song, J., Wurtele, E. S., and Nikolau, B. J. (1994). Molecular Cloning and Characterization of the cDNA Coding for the Biotin-Containing Subunit of 3-Methylcrotonoyl-CoA Carboxylase: Identification of the Biotin Carboxylase and Biotin-Carrier Domains. *Proc. Natl. Acad. Sci. U.S.A.* 91 (13), 5779–5783. doi:10.1073/pnas.91.13.5779

Stein, S. E. (1999). An Integrated Method for Spectrum Extraction and Compound Identification from Gas Chromatography/mass Spectrometry Data. *J. Am. Soc. Mass Spectrom.* 10 (8), 770–781. doi:10.1016/S1044-0305(99)00047-1

Südfeld, C., Hubáček, M., Figueiredo, D., Naduthodi, M. I. S., van der Oost, J., Wijffels, R. H., et al. (2021). High-throughput Insertional Mutagenesis Reveals Novel Targets for Enhancing Lipid Accumulation in Nannochloropsis Oceanica. *Metab. Eng.* 66, 239–258. doi:10.1016/j.ymben.2021.04.012

Tomassetti, M., Garavaglia, B. S., Vranych, C. V., Gottig, N., Ottado, J., Gramajo, H., et al. (2018). 3-methylcrotonyl Coenzyme A (CoA) Carboxylase Complex Is Involved in the Xanthomonas Citri Subsp. Citri Lifestyle during Citrus Infection. *PLoS One* 13 (6), e0198414. doi:10.1371/journal.pone.0198414

Tong, L. (2013). Structure and Function of Biotin-dependent Carboxylases. *Cell. Mol. Life Sci.* 70 (5), 863–891. doi:10.1007/s00018-012-1096-0

Tran, T. H., Hsiao, Y.-S., Jo, J., Chou, C.-Y., Dietrich, L. E. P., Walz, T., et al. (2015). Structure and Function of a Single-Chain, Multi-Domain Long-Chain Acyl-CoA Carboxylase. *Nature* 518 (7537), 120–124. doi:10.1038/nature13912

Vanee, N., Brooks, J., and Fong, S. (2017). Metabolic Profile of the Cellulolytic Industrial Actinomycete Thermobifida Fusca. *Metabolites* 7 (4), 57. doi:10.3390/metabo7040057

Waldrop, G. L., Holden, H. M., and Maurice, M. S. (2012). The Enzymes of Biotin Dependent CO_2 metabolism: What Structures Reveal about Their Reaction Mechanisms. *Protein Sci.* 21 (11), 1597–1619. doi:10.1002/pro.2156

Watanabe, S., Zimmermann, M., Goodwin, M. B., Sauer, U., Barry, C. E., 3rd, and Boshoff, H. I. (2011). Fumarate Reductase Activity Maintains an Energized Membrane in Anaerobic *Mycobacterium tuberculosis*. *PLoS Pathog.* 7 (10), e1002287. doi:10.1371/journal.ppat.1002287

Wongkittichote, P., Ah Mew, N., and Chapman, K. A. (2017). Propionyl-CoA Carboxylase - A Review. *Mol. Genet. Metabolism* 122 (4), 145–152. doi:10.1016/j.ymgme.2017.10.002

Wurtele, E. S., and Nikolau, B. J. (2000). Characterization of 3-Methylcrotonyl-CoA Carboxylase from Plants. *Methods Enzymol.* 324, 280–292. doi:10.1016/s0076-6879(00)24238-9

Xiong, W., Reyes, L. H., Michener, W. E., Maness, P. C., and Chou, K. J. (2018). Engineering Cellulolytic Bacterium Clostridium Thermocellum to Co-ferment Cellulose- and Hemicellulose-derived Sugars Simultaneously. *Biotechnol. Bioeng.* 115 (7), 1755–1763. doi:10.1002/bit.26590

Yan, Q., and Fong, S. S. (2018). Cloning and Characterization of a Chitinase from Thermobifida Fusca Reveals Tfu_0580 as a Thermostable and Acidic Endochitinase. *Biotechnol. Rep.* 19, e00274. doi:10.1016/j.btre.2018.e00274

Yao, J., and Rock, C. O. (2013). Phosphatidic Acid Synthesis in Bacteria. *Biochimica Biophysica Acta (BBA) - Mol. Cell Biol. Lipids* 1831 (3), 495–502. doi:10.1016/j.bbalip.2012.08.018

Yao, L., and Hammond, E. G. (2006). Isolation and Melting Properties of Branched-Chain Esters from Lanolin. *J. Am. Oil Chem. Soc.* 83 (6), 547–552. doi:10.1007/s11746-006-1238-3

Zemour, K., Adda, A., Labdelli, A., Dellal, A., Cerny, M., and Merah, O. (2021). Effects of Genotype and Climatic Conditions on the Oil Content and its Fatty Acids Composition of Carthamus tinctorius L. Seeds. *Agronomy* 11 (10), 2048. doi:10.3390/agronomy11102048

Zhang, Y., Wang, J., Wang, Z., Zhang, Y., Shi, S., Nielsen, J., et al. (2019). A gRNA-tRNA Array for CRISPR-Cas9 Based Rapid Multiplexed Genome Editing in *Saccharomyces cerevisiae. Nat. Commun.* 10 (1), 1053. doi:10.1038/s41467-019-09005-3

Zhang, Z., Wang, Y., and Ruan, J. (1998). Reclassification of Thermomonospora and Microtetraspora. *Int. J. Syst. Bacteriol.* 48, 411–422. doi:10.1099/00207713-48-2-411

Zhao, L., Geng, J., Guo, Y., Liao, X., Liu, X., Wu, R., et al. (2015). Expression of the Thermobifida Fusca Xylanase Xyn11A in Pichia pastoris and its Characterization. *BMC Biotechnol.* 15 (1), 18. doi:10.1186/s12896-015-0135-y

Zhu, K., Ding, X., Julotok, M., and Wilkinson, B. J. (2005). Exogenous Isoleucine and Fatty Acid Shortening Ensure the High Content of Anteiso-C15:0 Fatty Acid Required for Low-Temperature Growth of Listeria Monocytogenes. *Appl. Environ. Microbiol.* 71 (12), 8002–8007. doi:10.1128/AEM.71.12.8002-8007.2005

Zoghlami, A., and Paës, G. (2019). Lignocellulosic Biomass: Understanding Recalcitrance and Predicting Hydrolysis. *Front. Chem.* 7, 874. doi:10.3389/fchem.2019.00874

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 18 July 2022
doi: 10.3389/fmolb.2022.857505

Systematic Investigation of LC Miniaturization to Increase Sensitivity in Wide-Target LC-MS-Based Trace Bioanalysis of Small Molecules

Veronika Fitz[1,2], *Yasin El Abiead*[1], *Daniel Berger*[1] *and Gunda Koellensperger*[1,3,4]*

[1]Department of Analytical Chemistry, Faculty of Chemistry, University of Vienna, Vienna, Austria, [2]Vienna Doctoral School in Chemistry (DoSChem), University of Vienna, Vienna, Austria, [3]Vienna Metabolomics Center (VIME), University of Vienna, Vienna, Austria, [4]Chemistry Meets Biology, University of Vienna, Vienna, Austria

OPEN ACCESS

Edited by:
Wolfram Weckwerth,
University of Vienna, Austria

Reviewed by:
Zdenek Spacil,
Masaryk University, Czechia
Ann Van Schepdael,
KU Leuven, Belgium
Roccaldo Sardella,
University of Perugia, Italy
Malcolm Donald McLeod,
Australian National University,
Australia

***Correspondence:**
Gunda Koellensperger
gunda.koellensperger@univie.ac.at

Specialty section:
This article was submitted to
Metabolomics,
a section of the journal
Frontiers in Molecular Biosciences

Received: 18 January 2022
Accepted: 06 June 2022
Published: 18 July 2022

Citation:
Fitz V, El Abiead Y, Berger D and
Koellensperger G (2022) Systematic
Investigation of LC Miniaturization to
Increase Sensitivity in Wide-Target LC-
MS-Based Trace Bioanalysis of
Small Molecules.
Front. Mol. Biosci. 9:857505.
doi: 10.3389/fmolb.2022.857505

Covering a wide spectrum of molecules is essential for global metabolome assessment. While metabolomics assays are most frequently carried out in microbore LC-MS analysis, reducing the size of the analytical platform has proven its ability to boost sensitivity for specific *-omics* applications. In this study, we elaborate the impact of LC miniaturization on exploratory small-molecule LC-MS analysis, focusing on chromatographic properties with critical impact on peak picking and statistical analysis. We have assessed a panel of small molecules comprising endogenous metabolites and environmental contaminants covering three flow regimes—analytical, micro-, and nano-flow. Miniaturization to the micro-flow regime yields moderately increased sensitivity as compared to the nano setup, where median sensitivity gains around 80-fold are observed in protein-precipitated blood plasma extract. This gain resulting in higher coverage at low µg/L concentrations is compound dependent. At the same time, the nano-LC-high-resolution mass spectrometry (HRMS) approach reduces the investigated chemical space as a consequence of the trap-and-elute nano-LC platform. Finally, while all three setups show excellent retention time stabilities, rapid gradients jeopardize the peak area repeatability of the nano-LC setup. Micro-LC offers the best compromise between improving signal intensity and metabolome coverage, despite the fact that only incremental gains can be achieved. Hence, we recommend using micro-LC for wide-target small-molecule trace bioanalysis and global metabolomics of abundant samples.

Keywords: miniaturization, chromatography, LC-MS, metabolomics, exposomics, coverage, sensitivity

1 INTRODUCTION

The physicochemical diversity and wide concentration ranges of metabolites in biological samples to date prevent comprehensive coverage of the metabolome by a single (or even a few) analytical methods. Methods based on liquid chromatography coupled to mass spectrometry (LC-MS) offer the best sensitivity, highest versatility regarding physicochemical coverage and dynamic ranges between

Abbreviations: EIC, extracted ion chromatogram; ESI, electrospray ionization; HRMS, high-resolution mass spectrometry; i.d., inner diameter; LC, liquid chromatography; LOD, limit of detection; MS, mass spectrometry; o.d., outer diameter; rsd, relative standard deviation.

two and four orders of magnitude. Specifically, chromatographic separation supports the identification of isomers, reduces ion suppression and improves detection of low-abundant compounds (Lu et al., 2017; Alseekh et al., 2021). LC-MS-based metabolomics experiments are most frequently carried out in microbore scale (i.e., 1.5–3.2 mm inner column diameter and flow rates of 100–500 µl/min) (Vasconcelos Soares Maciel et al., 2020). Microbore systems are robust and convenient to use, accommodate short and steep gradients and provide high-performance chromatography with peak widths around 3 s (full width at half maximum). The workflows established for high resolution mass spectrometry (HRMS)-based analysis depend on highly repeatable features with regard to retention time and signal intensity. Microbore LC systems allow robust plug-and-play operation while providing reproducible retention times, peak shapes and signal intensities, supporting automated data processing in the course of non-targeted experiments. Likewise, narrow peak shape, technical reproducibility of signal intensities, minimal system carryover and linear detector response build the basis for (relative and absolute) quantification.

In contrast to small-molecule *-omics*, proteomics and peptide LC-MS-analyses are commonly carried out on the nano-scale (i.e., 10–150 µm column i.d. and flow rates of 0.1–1 µl/min). Gradients in proteomics and peptide analysis are typically much longer (in the order of an hour) and eluent composition covers a narrower span of organic eluent content. Nano-LC coupled to nano-ESI-MS offers unrivalled mass sensitivity essential for the analysis of low-volume samples. The assets but also challenges of nano-LC are related to the low flow rates employed. On the one hand, sensitivity can be vastly increased by reduced on-column sample dilution and compatibility with nano-ESI, offering itself unique benefits for ionization (Juraschek et al., 1999; Schmidt et al., 2003; Kourtchev et al., 2020). On the other hand, the low flow rates and small column dimensions make the whole system more susceptible to void volumes, clogging of column and capillaries/emitter, mass overload and associated system carryover, etc. (Noga et al., 2007). In principle, these stressors also affect microbore LC, but are more pronounced at the very low flow rates of nano-LC, and complicate successful handling in practice. Hence, nano-LC is not as widely established in small-molecule *-omics* as it is in proteomics, but it has been successfully applied for (xeno-) metabolomics analysis especially for cases where low available sample volumes demanded the smallest possible analysis platform (Lanckmans et al., 2006; Nakatani et al., 2020; Geller et al., 2022). In fact, LC miniaturization for small-molecule analysis is iteratively discussed in literature as a means of increasing sensitivity when dealing with low sample amounts (Chetwynd and David, 2018; Nakatani et al., 2020; Sanders and Edwards, 2020).

While analytical flow and nano-flow LC-MS platforms are routinely used in metabolomics and proteomics analyses (Shi et al., 2004; Wilson et al., 2015; Yi et al., 2017), the interest for micro-flow platforms is increasing in both research communities as a means to enhance sensitivity and save analysis cost (coming from analytical flow) (Greco et al., n.d.; Gray et al., 2016; Cebo et al., 2020; King et al., 2020) or to enhance robustness and reproducibility of the analysis (coming from nano-flow) (Bian et al., 2020). LC miniaturization maximizes signal intensities from a given amount of injected sample and potentially extends the analysis scope toward molecules with low abundance or detector response. Signal intensity is key for compound identification as signal intensity thresholds determine the triggering, acquisition and quality of MS/MS spectra.

A handful of studies have compared the performance of specifically optimized miniaturized LC-MS platforms with their established microbore LC-MS workflows for metabolomics or other multi-residue small-molecule analyses (Chetwynd et al., 2014; Nakatani et al., 2020; Zardini Buzatto et al., 2020; Geller et al., 2022). With the present study, we address the following question: Assuming that sample volume is not a limiting factor, would LC miniaturization allow to broaden the analyte scope by extending coverage toward low abundant analytes? Injecting the same sample volume on a smaller analytical platform equals a large volume injection, which is successfully applied for, e.g., proteomics or environmental analysis, but without increasing the actual amount of injected sample–a considerable asset for metabolomics experiments since they usually deal with dense sample matrices and long sequence runs. LC miniaturization holds the potential for maximizing sensitivity without the cost of polluting ion source and mass spectrometer with additional sample. Here, we pinpoint the benefit and challenges of LC miniaturization for non-targeted multicomponent small molecule analysis in practice by transferring a typical metabolomics method from analytical to micro- and nano-flow regime while holding injection volume, mobile and stationary phases, gradient and detection parameters constant.

2 EXPERIMENTAL

We compared a standard analytical scale (250 µl/min) reversed-phase metabolomics method with two grades of miniaturization, micro- (57 µl/min) and nano-flow (0.3 µl/min) (Vasconcelos Soares Maciel et al., 2020), by injecting a series of standards and matrix samples on each platform. We analyzed spiked exogenous compounds at different concentrations and endogenous human plasma metabolites at natural abundance levels. The analytes were selected to cover a wide range of physicochemical properties and show different grades of reversed phase retention (see **Figure 6**) to monitor chromatographic enrichment and retention-related impacts on signal intensity.

2.1 Standards and Solvents

Acetonitrile (ACN) and water were of LC-MS grade and ordered at Sigma-Aldrich (Vienna, Austria) and Fisher Scientific (Vienna, Austria). Formic acid ≥99% and methanol (MeOH) were also of LC-MS purity and ordered at VWR International (Vienna, Austria).

The mycotoxins aflatoxin B1 and G2, ochratoxin A, sterigmatocystin, T2-toxin and zearalenone were obtained from RomerLabs (Tulln, Austria). Aflatoxin M1, aflatoxicol, alternariol, and ochratoxin alpha were obtained from Toronto Research Chemicals (Ontario, Canada). A total of 10 mycotoxins were analyzed. Pharmaceutical and agrochemical standards were kindly provided by Eurofins Umwelt Österreich GmbH & Co KG (39 and 9 compounds, respectively). Standards were obtained in dissolved form or weighed and dissolved in appropriate solvent to obtain single stock solutions. All standards were of HPLC-grade or LC-MS-grade purity. Molecules and sum formulas are listed in **Supplementary Table S1**.

In the following, we compare peak width and peak shape, repeatability of detector response, retention time stability, signal intensity and peak concentration for model molecules that are detected in all three setups in spiked plasma extract at a concentration of 1 μg/L, except ceftiofur and coumaphos, which are assessed at 10 μg/L. Linear range, sensitivity, matrix effect and limit of detection are assessed and compared for molecules with a linear relation of concentration and detector response in all three setups. Coverage and signal intensity ratios are additionally assessed using a panel of molecules detected in plasma extract at naturally occurring abundances (48 endogenous metabolites, 3 xenobiotics). Details are listed in **Supplementary Tables S2–S5**.

2.2 Spiking Solutions

Single stocks were stored at −20°C until they were volumetrically combined to give two multicomponent mixtures: one containing pharmaceuticals/agrochemicals and one containing mycotoxins. For both mixtures, single stocks were combined volumetrically and evaporated to dryness in a vacuum centrifuge at room temperature. The pharmaceutical/agrochemical residue was reconstituted in MeOH to give a multicomponent standard with a concentration of 50,000 μg/L. Further 1:10 dilution steps with MeOH yielded multicomponent standards with concentrations of 5,000, 500, 50, 5 and 0.5 μg/L. The mycotoxin residue was thoroughly reconstituted in 5% (v/v) ACN to give a multicomponent standard with a concentration of 500 μg/L. Further 1:10 dilution steps with 5% (v/v) ACN yielded multicomponent standards with concentrations of 50 and 5 μg/L.

2.3 Plasma Extraction

Pooled human blood plasma (two donors) was purchased in frozen form (dry ice) at Innovative Research, Inc. (46430 Peary Court, Novi, Michigan, United States) and stored at −20°C until sample preparation. After thawing at room temperature, 2 ml of plasma was transferred to a 15 ml Falcon tube and mixed with 6 ml acidified ACN (ACN +0.1% (v/v) formic acid). The mixture was vortexed for 3 min and kept at −20°C for 1 h to allow protein precipitation, then it was vortexed again and mildly centrifuged for 5 min at 2,000 rcf. The supernatant was collected and transferred to Eppendorf tubes for high-speed centrifugation (14,000 rcf, 10 min). Centrifugation steps were performed at room temperature on a HERMLE Z446K centrifuge. The supernatants were carefully aspirated and mingled in 5 ml Eppendorf tubes. Seven aliquots of 400 μl each were prepared in brown 1.5 ml HPLC glass vials with screw caps and septum.

2.4 Samples

Experiments were based on spiked plasma extract (matrix samples) and spiked neat solvent. For neat solvent samples, 2 ml of LC-MS-grade water were taken through the same sample preparation procedure as described for plasma extracts (**Section 2.3**). Next, 400 μl aliquots of plasma extract or solvent were spiked with the previously prepared multicomponent mixtures (pharmaceuticals/agrochemicals and mycotoxins, respectively), giving six concentration levels (5,000, 1,000, 100, 10, 1, and 0.1 μg/L) plus one zero sample for each of the two matrices. Mycotoxins were not spiked to samples of the highest concentration and were diluted 1:10 in all other concentration levels as compared with the spiked pharmaceuticals/agrochemicals, giving sample concentrations of 100, 10, 1, 0.1, and 0.01 μg/L plus zero sample. Spiked samples were evaporated to dryness, thoroughly reconstituted in 5% (v/v) ACN and transferred to 1.5 ml brown-glass HPLC-vials with 200 μL glass inserts and screw caps with slit septum for analysis. Portions of sample that were not needed at the moment were kept in their original HPLC-vials and stored at −20°C in dissolved form.

2.5 Instrumental Setups

Three analytical setups were compared in this analysis. All were based on C18 HSS T3 column chemistry with acidified H_2O/ACN as eluent system, and on detection *via* HRMS with a Q Exactive HF quadrupole-Orbitrap mass spectrometer (Thermo Scientific). The microplatform was stringently scaled to maintain the same linear flow velocity as in the analytical flow equivalent. The three platforms employed the same sub-2μm stationary material and were operated with volumetric flow-rates close to their van Deemter optima.

Owing to the great structural diversity of metabolites, it is impossible to assess the entire metabolome with one analytical platform (Patti, 2011; Lu et al., 2017). Nevertheless, the aim is to capture as many molecules as possible and separations in global metabolomics are not tailored to specific molecules but most frequently use generic LC gradients spanning very low to very high organic eluent content and medium run times. Fully wettable stationary phases tolerate 100% (v/v) aqueous eluent composition and offer retention for the more polar analytes that would be flushed away with minimal organic solvent in the eluent compared with conventional C18 phases. Next to analyte enrichment, several other parameters along the analytical process influence the intensity of the resulting detector signal: Matrix density, dilution, solvent and volume of the injected sample, extra-column volumes and flow-rate, amount and mass loadability of the stationary phase, ionization efficiency depending on analyte chemistry, eluent, coeluting matrix, droplet size related to emitter geometry, spray voltage; and finally, ion transfer, width of *m/z* scan window, ion suppression effects in the c-trap and detection efficiency in the mass analyzer. All of these parameters should be adapted to the type of sample and analytes of interest and should finally suit the analytical platform to achieve the highest possible sensitivity. Optimization of the whole analytical procedure is indeed quite specific for each application. With this study, we want to elaborate the sensitivity potential enabled by LC miniaturization,

TABLE 1 | Key features of the three analytical setups.

	Analytical setup	Micro-setup	Nano-setup
LC instrument	Vanquish Duo UHPLC	Vanquish Duo UHPLC	UltiMate 3000 RSLCnano
Column i.d.	2.1 mm	1.0 mm	0.075 mm (separation) 0.3 mm (trap)
Flow rate	250 µL/min	57 µL/min	Separation: 0.3 µL/min Loading: 30 µL/min
Inject. volume	3 µL	3 µL	3 µL
ESI source	Ion Max with HESI-II-probe	Ion Max with HESI-II-probe	Nanospray Flex
Emitter i.d.	100 µm	50 µm	30 µm
Spray voltage	3.5 kV	3.5 kV	1.9 kV
Other source parameters	Flow rate default for temperatures and gas flows	Flow rate default for temperatures and gas flows	Manually adjusted emitter position

i.e., reduced column inner diameter and reduced flow-rate, for wide-target small-molecule analysis. It serves comparability to keep as many of the parameters constant as possible (injection volume, mass spectrometer, detection parameters) and adapt only parameters that are directly related to the flow rate (LC instrument, flow rate, ion source parameters). For the adapted parameters, we followed vendor recommendations as far as possible to ensure we operated each instrumental platform under the respective optimal conditions while offering suitable conditions for a wide variety of analytes.

An overview of the key method features is given in **Table 1**. Further details can be found in the text below.

2.5.1 Analytical Flow Setup

The standard LC-setup was built upon a Vanquish Duo UHPLC system (Thermo Scientific) consisting of a solvent rack, two binary pumps, a split sampler with two injection valves, and a column compartment. The capillary setup was optimized for analytical flow regimes and consisted of 100 µm i.d. Viper-capillaries (Thermo Fisher Scientific) pre- and post-column. An Acquity UPLC HSS T3 column (2.1 mm i.d. × 150 mm, 100 Å, 1.8 µm, Waters) equipped with a VanGuard Pre-Column (2.1 mm i.d. × 5 mm, Waters) was eluted in gradient-mode with a flow rate of 250 µl/min at 35°C. Mobile phase A was H_2O + 0.1% (v/v) formic acid, mobile phase B was ACN +0.1% (v/v) formic acid. The following gradient was applied: 0–1 min 1% B, 1–5 min ramp to 50% B, 5–12 min ramp to 99% B, 12–15 min hold at 99% B, at 15 min switch to 1% B, followed by 15–22 min re-equilibration at 1% B. The injection volume was 3 µl and the injector needle was washed with 80% ACN for 15 s after each injection. The column was connected to an Ion Max Source with a Heated Electrospray Ionization (HESI-II) Probe and a 100 µm i.d. stainless steel emitter (Thermo Fisher Scientific) via a 100 µm i.d. Viper-capillary, a zero dead-volume grounding union, and a piece of 100 µm i.d. PEEK-capillary.

2.5.2 Micro-Flow Setup

This LC-setup was built upon the same Vanquish Duo UHPLC system (Thermo Scientific) as the analytical flow setup. An Acquity UPLC HSS T3 column (1 mm i.d. × 150 mm, 100 Å, 1.8 µm, Waters) equipped with a VanGuard Pre-Column (2.1 mm i.d. × 5 mm, Waters) was eluted in gradient-mode. The flow rate was volumetrically scaled to maintain the same linear flow velocity as in the analytical setup and was held constant at 57 µl/min. Column temperature, eluents, and gradient were the same as described for the analytical setup. The same 100 µm i.d. Viper capillaries were used pre-column as described above, which led to slight gradient delay in combination with the lower flow rate. It was necessary to prolong the re-equilibration step to 9 min, resulting in a total runtime of 24 min: 0–1 min 1% B, 1–5 min ramp to 50% B, 5–12 min ramp to 99% B, 12–15 min hold at 99% B, at 15 min switch to 1% B, followed by 15–24 min re-equilibration at 1% B. Injection volume, needle wash and column temperature were the same as in the analytical flow setup. To avoid post-column peak broadening, the post-column flow path was adapted to the lower volumetric flow rate: The column was connected to an Ion Max Source with a Heated Electrospray Ionization (HESI-II) Probe *via* a 50 µm i.d. × 350 mm nanoViper-capillary, a zero dead-volume grounding union, and 50 µm i.d. × 150 mm nanoViper-capillary. The ion source was equipped with a 50 µm i.d. stainless steel emitter. Flow-path adaptations were made according to (Greco et al., n.d.).

2.5.3 Nano-Flow Setup

For the nano-flow setup, a trap-and-elute configuration was chosen to increase loading capacity and loading flow rate. An UltiMate 3000 RSLCnano system (Thermo Scientific) consisting of SRD-3400 solvent rack, NCS-3500RS pump module containing the column compartment and an 850 bar 10-port switching valve, and a WPS-3000TPL RS temperature-controlled autosampler equipped with a 350-bar 8-port-valve and an 850-bar injection valve. Fluidic setup and capillary dimensions followed vendor recommendations to minimize pre-column extra-column volumes. Micro-flow (30 µL/min) was delivered by a ternary micro pump for preconcentrating the sample on a trapping column. The loading pump delivered the flow through the autosampler injection valve via the 10-port switching valve in the column compartment onto the trapping column. Nano-flow (0.3 µL/min) was delivered by a nano/capillary pump and directed onto the nano-column. Flow rate was regulated with an integrated ProFlow flowmeter. The nano/capillary pump delivered the flow to the nano-column *via* the 10-port switching valve in the column compartment. A 3 µL sample plug was drawn in microliter-pickup mode through a 2.4 µL injection needle and into a 20 µL sample loop. LC-MS-grade water with 0.1% (v/v) formic acid served as pickup-fluid. The sample was injected into the loading-flow path and accumulated on the trapping column for 1 min. The trapping column effluent was directed to waste during the loading procedure. Subsequent analysis was carried out in back-flush mode, i.e., the trapping column was switched in line with the nano-column by rotating the column compartment 10-port valve, now carrying the entire nano/capillary-pump-gradient through the trapping column in reversed direction and through the nano-column to the MS. Pre-concentration setups with commercial equipment typically employ the same stationary phase chemistry for trap column and analytical column (Wilson et al., 2015), but with larger particle size and hence less retentivity of the trap column. A nanoEase *M/Z* HSS T3 trap column (0.3 × 50 mm, 100Å, 5 µm, Waters) was used for pre-concentration and a nanoEase *M/Z* HSS T3 nano-column (0.075 × 150 mm, 100Å, 1.8 µm, Waters) for analyte separation. The column compartment was kept at 35°C.

Eluent composition was the same for trapping and separation: Eluent A was H_2O + 0.1% (v/v) formic acid and eluent B was ACN +0.1% (v/v) formic acid. Trapping was pursued for 1 minute with 0% B (isocratic). While the aliphatic groups of ordinary C18 material collapse in 100% aqueous environment, the HSS T3 chemistry is fully wettable. We chose this material to allow a loading step without organic modifier to retain polar compounds as far as possible. The trapping column was switched in line with the nano-column 1 min after injection and with another 1-min delay, a gradient from 1 to 99% B in 11 min was delivered by the nano/capillary-pump for separation on the nano-column. The gradient was followed by a 6-min flush with 99% B and 23 min re-equilibration at 1% B. The nano-column was connected to a nano-ESI source with a piece of 20 µm i.d./ 280 µm o.d. fused silica tubing, a PTFE sleeve and a zero dead volume PEEK union. Considering the complexity of samples obtained by non-selective liquid-liquid-extraction and centrifugation, a stainless-steel emitter with an i.d. of 30 µm was chosen to ensure longer durability and avoid clogging as compared to silica emitters with lower inner diameter. Emitter position was adjusted manually. Spray voltage was 1.9 kV in positive ionization mode.

2.5.4 Mass Spectrometry

HRMS was performed with a Q Exactive HF quadrupole-Orbitrap mass spectrometer (Thermo Scientific). The following parameters were used for all three setups: MS1 spectra (profile mode), scan range 80–1200 *m/z*, positive polarity, resolution 120,000, AGC target 3e6, maximum injection time 200 ms, and S-lens RF-level 50. For ionization, two different electrospray sources were used: A Nanospray Flex ion source equipped with a 30 µm i.d. steel emitter for the nano-scale setup, and an Ion Max source equipped with a HESI-II-probe and steel emitter for micro- and analytical setup. Both sources were purchased from Thermo Fisher Scientific. Emitter i.d. was 100 µm for the analytical setup, while the micro-setup required a reduced emitter i.d. of 50 µm. We optimized the ESI parameters and carried out the experiments under optimum condition for the respective flow regime. Spray voltage was 1.9 kV for the nano-setup and 3.5 kV for micro- and analytical setup, respectively. Flow rate sensitive parameters for HESI-ionization were adapted according to vendor recommendations: for micro-setup (57 µL/min), capillary temperature was 250°C, sheath gas 30.70, auxiliary gas 10.00, spare gas 1.00, and probe heater temperature 157°C. For analytical setup (250 µL/min), capillary temperature was 253.13°C, sheath gas 46.25, auxiliary gas 10.63, spare gas 2.13, and probe heater temperature 406.25°C.

2.6 Data Evaluation

After data acquisition, vendor-specific profile mode files were centroided with the msConvert GUI application (version 3.0.19014-f9d5b8a3b) from the ProteoWizard Toolkit applying the peakPicking-filter (vendor msLevel = 1-1) and mzML as output format (Chambers et al., 2012). Centroided data were subjected to targeted data evaluation in Skyline (Adams et al., 2020). A mass extraction window of 10 ppm was used to generate extracted ion chromatograms of the target compounds. The chosen procedure outweighed calibration-related differences in mass accuracy between the datasets and avoided loss of peak area in the *m/zm/z* dimension for all three setups (Vereyken et al., 2019). Evaluation of chromatographic parameters focused on $[M + H]^+$ adducts of the monoisotopic peaks. For selected compounds, extracted ion chromatograms were generated based on $[M]^+$, $[M + NH_4]^+$ or $[M + H\text{-}H_2O]^+$ adducts. For spiked exogenous compounds, area values of monoisotopic EICs were used to characterize signal intensity, signal reproducibility, matrix effect, slope and linear range of calibration curves, and limit of detection. Furthermore, we assessed chromatographic peak width and symmetry, retention time stability and peak concentration. Signals with less than three consecutive data points per peak were dismissed. Since some compounds showed background noise, signals below 2× the averaged signal of a matrix-specific zero sample (solvent or plasma extract with a spiked concentration of 0 µg/L, $n = 4$) was put in place as additional filter. Formulas for the calculation of the parameters can be found in the results section.

The same method characteristics were assessed for a panel of metabolites detected in non-spiked plasma extract, except for matrix effect, calibration curve and LOD. Peak concentration was expressed relative to the analytical flow setup. The chosen metabolites represent major metabolite groups found in the human serum metabolome database (Psychogios et al., 2011): Small organic acids, nucleobases, steroid hormones, sugar phosphates, amino acids, and lysophospholipids. The molecules were identified by retention time comparison with authentic standards or literature as noted in **Supplementary Table S1**.

3 RESULTS

This study elaborates the impact of LC miniaturization in small-molecule LC-MS analysis. We compared two miniaturized platforms with different grades of miniaturization, micro- and nano-flow, with the standard analytical flow platform, for depicting the metabolome of an abundant sample. The platforms were characterized by performance parameters critical for current practices of non-target bioanalysis-like peak picking and statistical analysis, emphasizing sensitivity and metabolome coverage. We compared analytical figures of merit for model molecules that were detected in all three setups in spiked plasma extract at a concentration of 1 µg/L, except ceftiofur and coumaphos, which were assessed at 10 µg/L, or for molecules that showed a linear relation of signal intensity and concentration in all three setups (linear range, sensitivity, matrix effect, limit of detection), respectively. Coverage and signal intensity ratios were additionally assessed using a panel of endogenous metabolites detected in plasma extract at naturally occurring abundances. We analyzed a panel of molecules with wide chemical diversity with logP values between −4.4 and 7.7 and molecular weights spanning approximately 85–550 Da. The results are listed in **Supplementary Tables S2–S5**.

Non-targeted experiments should cover both ionization polarities since some metabolite classes ionize more effectively as anions (constituents of the central energy metabolism like small organic acids, sugars and their di-/triphosphates, etc.) and others as cations (amino acids, nucleobases, nucleosides and nucleotides, steroids, several lipid classes, etc.). Acquiring positive and negative mode data in an automated fashion or even in one chromatographic run is desirable. In particular, the Orbitrap mass analyzer supports fast polarity switching, which allows to record positive and negative ionization mode data near-simultaneously and offers a substantial increase in analysis throughput. The used nanoESI source, however, did not allow fast polarity switching. The emitter position needs to be adjusted for each ionization mode manually, precluding automated acquisition of positive and negative mode data in one run. Additionally, the nanoESI spray is less stable in negative mode especially for eluent compositions with a high aqueous content, as observed by us and others (Nguyen-Khuong et al., 2018). For our experiments we therefore used positive ionization mode. The described chromatographic phenomena apply to both polarities. When interpreting our results it should be kept in mind that ionization efficiency and matrix effect can differ in negative ionization mode.

3.1 Chromatographic Quality, Repeatability, and Peak Shape

3.1.1 Peak Width and Symmetry

Peak width (here: width at half-maximum) and peak symmetry are related to chromatographic resolution and enrichment success. Narrow peaks ensure maximal separation efficiency and signal intensity, resulting in cleaner MS/MS spectra and better detection limits. Peak width homogeneity affects the quality of fragment spectra as the average peak width is set for data-dependent MS/MS acquisition (average peak width for dynamic exclusion and apex trigger). It also influences the quality of peak picking by commonly applied software like XCMS, where a window of peak widths needs to be defined to help differentiate chromatographic peaks from background signals. Miniaturization to modular LC-systems holds the risk of peak broadening because the ratio of column-volume and flow rate to extra-column volumes tends to be less favorable compared to columns with a greater inner diameter even with optimized flow-path connections. Median peak width was comparable between the setups (**Figure 1**). Phospholipids eluted as broader peaks on all platforms with median peak widths around 5 s under micro- and analytical flow, and around 7 s under nano-flow conditions. Sn-1 and sn-2 positional isomers were fully separated in the former two setups but were not baseline separated in the nano-platform. Thiophosphates (acephate, coumaphos, methamidophos), sulfonamides (sulfachlorpyridazine, sulfadiazine, sulfadimethoxine, sulfamethazine, sulfamethoxazole, sulfamethoxypyridazine, sulfaquinoxaline, sulfathiazole) and the sulfonate florfenicol, as well as several nitrates-containing compounds (dimetridazole, dinoterb, furazolidone, ronidazole) eluted as very broad peaks in the nano-setup, while the same compounds exhibited excellent peak shapes in the micro- and analytical flow regime. Since the same was observed for direct injection mode without enrichment column (data not shown), we assume that the pronounced compound-class specific peak shape distortion in the nano-setup is linked to surface interactions. Unwanted surface interactions are enhanced in nano-LC systems, leading to unexpected chromatographic effects even for compounds with otherwise good retention, which complicates the choice of optimal peak width for data-dependent MS/MS acquisition and non-targeted peak picking. Additionally, most molecules displayed compound-specific tailing in the nano-setup at all tested concentrations. Analytical and micro-platform, on the other hand, showed almost perfect peak symmetry. Tailing peaks reduce chromatographic resolution and hold the risk of masking low abundant analytes through ion suppression. Overlapping peaks lead to chimeric spectra during fragmentation, undermining the accuracy of compound identification.

3.1.2 Signal Stability

Among the molecules that were detected in all setups, the median repeatability of area values was around 3.7% relative standard

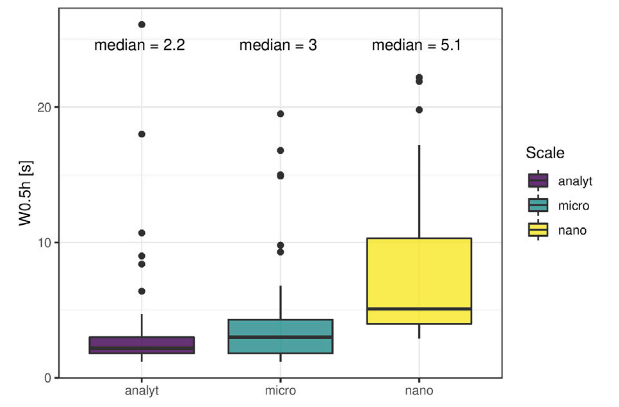


FIGURE 1 | Peak width. Full width at half maximum was assessed for 53 molecules comprising endogenous (ceftiofur and coumaphos 10 µg/L, caffeine, paraxanthine and theobromine at naturally occurring abundance; all others 1 µg/L) molecules in plasma extract. Lysophospholipids were broader compared to the rest in all setups. Further, thiophosphate, sulfonamide/sulfonate and nitrate-containing compounds had distorted peak shapes in the nano setup.

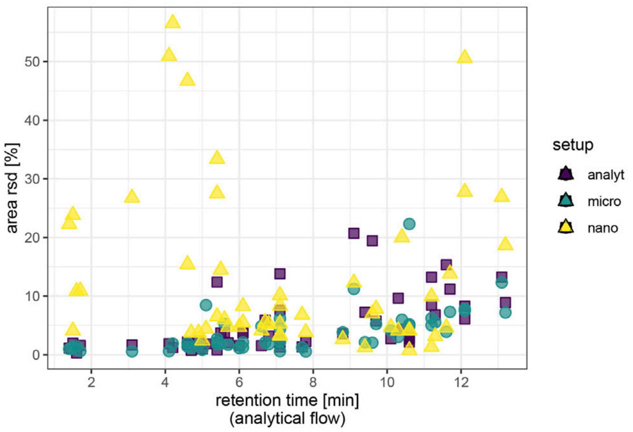


FIGURE 2 | Repeatability of signal intensity based on repeated injections (N = 4) of spiked plasma extract. Endogenous metabolites and caffeine/ caffeine metabolites were assessed at natural abundance; ceftiofur and coumaphos at 10 µg/L, all other exogenous compounds at 1 µg/L. Area values are background corrected.

deviation in the analytical flow setup for the spiked exogenous test molecules and excellent 2.3% for the endogenous molecules investigated. Repeatability improved upon miniaturization for well-retained exogenous compounds with previously low signal intensity. We observed a decrease of area repeatability with retention time in the analytical flow regime and to a lesser extent in the micro-flow regime (**Figure 2**). The nano-flow platform showed elevated but satisfactory area repeatability for most of the exogenous compounds. Sulfonamides and florfenicol (sulfonate), nitrates, and thiophosphonates-displayed reduced area repeatability. Notably, the nano-setup displayed optimum repeatability only for a specific retention segment, while the more hydrophilic metabolites (amino acids) were less reproducible due to suboptimal chromatographic enrichment in the trap-and-elute configuration and area rsd had a tendency to increase for the very lipophilic metabolites (lysophospholipids) due to spray destabilization at high proportions of organic solvent in the eluent. Additionally, repeatability did not improve analogously with signal intensity in the nano-flow setup. It is difficult to maintain a stable electrospray throughout the wide gradient with only one spray voltage setting and the delicate stability of the nanospray is even more affected by rapidly changing eluent conditions.

3.1.3 Retention Time Stability

We assessed retention time stability based on repeated injections of a quality control sample throughout the injection sequence (c = 1 µg/L, *n* = 6) spanning 8.5 h (analytical), 9 h (micro) and 15 h (nano). Retention times were adequately stable for all three investigated setups, as retention times deviated from the mean less than 5 s during 23 injections (**Figure 3**). Retention time stability of the nano-setup was comparable to micro- and analytical flow regime, even though absolute retention times were around twice as high due to pronounced gradient delay and resulting duration of the method. However, compounds that eluted as broad peaks (sulfonamides/sulfonate, nitrates, thiophosphates) also displayed reduced retention time repeatability.

3.2 Sensitivity

3.2.1 Signal Intensity/Sensitivity

Yielding the highest signal intensity out of a given amount of sample is the principal goal of LC miniaturization. The assumption for our study is that abundant sample material is available and sample volume is not a limiting factor for sensitivity (e.g., plasma analysis of adult humans). Hence, we injected the same sample volume on each platform. Signal intensity was assessed for all molecules, sensitivity (expressed as slope of calibration curves in linear range) was additionally reported for the spiked exogenous compounds. On average, signal intensity and sensitivity were improved through both miniaturized setups. Area ratios increased with retention time in the nano-setup, underlining that chromatographic enrichment was an important factor to maximize sensitivity. For the micro-setup, this relation was not as straightforward. The actual extent of intensity increase depended on the specific molecule in both setups (**Figure 4**). Using the micro-flow setup we observed a median increase of signal intensity of around 2-fold for spiked plasma extract, with individual signal intensity ratios ranging between 0.7 and 20 (excluding LPC 20:1, which increased to 100-fold due to very low signal intensity in the analytical flow setup). Downscaling to nano-flow multiplied signal intensities compared to the analytical flow regime: a median 45-fold for the investigated endogenous metabolites and around 75-fold for exogenous molecules. Signal intensity ratios of the more polar compounds fell below 30-fold increase, while individual rather lipophilic molecules exceeded 1,000-fold increase.

3.2.2 Peak Concentration

Volumetric flow rate of the analytical setup (250 µL/min) had been scaled to the smaller column dimensions of the micro-setup (1 mm i.d. vs. 2.1 mm i.d.) to maintain approximately the same linear flow velocity (**Eqs 1**, **2**). Under the assumption that peak width in the miniaturized setups is as narrow as under analytical flow regime, the analyte band is more concentrated and signal intensities theoretically increase by a factor of 4.4 after injecting the same amount of sample. Likewise, the theoretical increase in signal intensity using the nano-setup is 880-fold. However, transferring the whole analytical platform to a lower flow regime opens a multitude of factors that can influence actually obtainable signal intensities. First, chromatographic effects like peak broadening and tailing upon miniaturization can hamper signal intensity and signal-to-noise ratios; second, the different flow rates and peak concentrations can impact ionization efficiency, ion transmission and collection, and Orbitrap analysis. We compared peak concentrations (**Eq. 3**) and found that (mild) peak broadening affected the concentration of the analyte band in the miniaturized setups. Broader peaks elute in a higher volume of eluent and the concentration reaching the detector is therefore reduced. The theoretical gain in peak concentration as described above is therefore not reached in practice. While both, peak concentration gain and signal

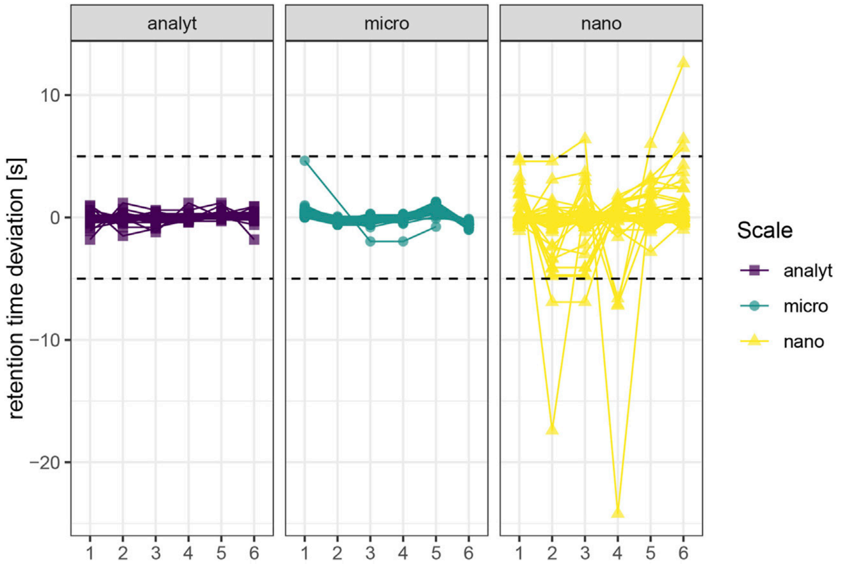


FIGURE 3 | Retention time stability. A quality control sample (c = 1 µg/L) was injceted six times (*n* = 6) spanning 8.5 h (analytical), 9 h (micro) and 15 h (nano). Replicates 4–6 were injected right after another. The dashed line marks 5 s deviation from the mean retention time. Molecules with distorted peak shapes also display reduced retention time stability due to imprecise automatic detection of the peak apex.

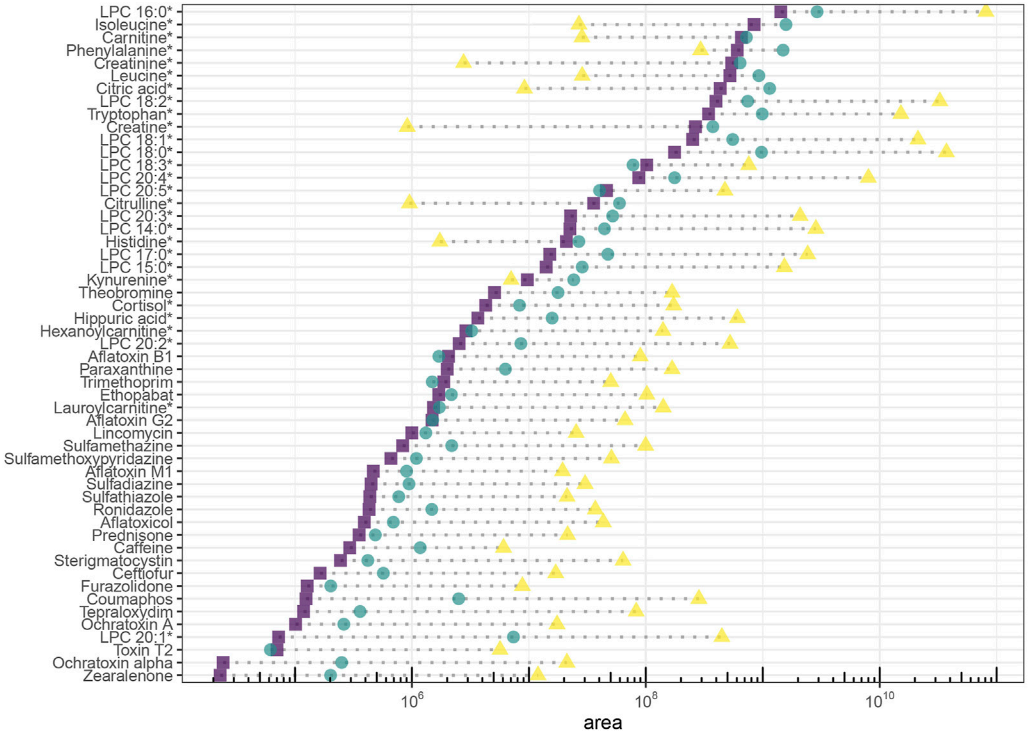


FIGURE 4 | Signal intensity of endogenous (natural abundance) and exogenous molecules (ceftiofur and coumaphos 10 µg/L, caffeine, paraxanthine and theobromine naturally occuring abundance, all others 1 µg/L) in plasma extract. Areas are background corrected. Endogenous molecules are marked with an asterisc.

intensity, increase were lower than in theory, the actual profit in signal intensity was again lower than for peak concentrations. This finding points to factors beyond chromatographic enrichment that influence detector response. The most polar analytes including amino acids were quantitatively lost during the loading step in the nano-setup. Detector response of the other molecules and in the micro-setup was affected (i.e., mostly reduced) by extra-column effects, for example, signal suppression (during ionization or in the C-trap) due to up-concentrated analyte and matrix, up-concentration of analytes altering adduct formation, and non-linear ESI-response (Yu et al., 2020).

$$Z_a = Y_a{*}60{*}\,\frac{4}{\pi * d_a^2}, \tag{1}$$

$$Y_m = \frac{Z_a}{60} * \frac{\pi * d_m^2}{4}, \tag{2}$$

$$c_{Peak} = \frac{IV * c}{Y * fwhm}, \tag{3}$$

where Y_a. Y_m = volumetric flow rate of analytical and micro-setup [mL/min]
Z_a, Z_m = linear flow velocity of analytical and micro-setup [cm/h]
d_a, d_m = column inner diameter of analytical and micro-setup [cm]
c_{Peak} = peak concentration, average conc. across whole peak volume [µg/L]
IV= injection volume (3 µL)
c = concentration of injected sample (1 µg/L)
Y = flow rate of respective setup [µL/min]
$Fwhm$ = peak width at half maximum [min]

3.2.3 Linear Range

Non-targeted exploratory *-omics* investigation typically involves fold-change analysis of signal intensities between studied sample groups and ideally, the linear range of the analytical platform should cover the range of analyte concentrations in the different sample groups to properly compare them (Alseekh et al., 2021).

Linear range was estimated for a panel of exogenous molecules according to The Fitness for Purpose of Analytical Methods: A Laboratory Guide to Method Validation and Related Topics (Eurachem, 2nd ed. 2014) and spanned between 2 and 4 orders of magnitude (Eurachem, 2014). On average, it was shorter and reached lower concentrations in the nano-setup compared to micro- and analytical flow platform (**Supplementary Figure S1**). Linear range length of individual molecules was affected by matrix effects in all setups. A wide linear dynamic range is advantageous in assays where the individual analyte is expected to appear in a wide concentration span, as in metabolomics, and the shorter linear range in the nano-setup arguably complicates quantitative comparison of samples or sample types that contain vastly different quantities of analyte. However, quantification of low abundant analytes profits from higher sensitivity. It depends on expected analyte concentration and goal of the analysis which of these aspects is given more importance.

3.2.4 Limit of Detection

The limit of detection (LOD) estimates the concentration at which an assay can accurately predict if a compound is present in the sample or not. As such, it views signal intensity in conjunction with a level of certainty, which is represented by signal repeatability at concentrations on the verge of being undetectable. To account for possible compound-specific carryover and background and consistently select appropriately low concentrations for each platform, LOD was calculated based on relative area standard deviation of four repeated injections of the lowest standard/sample concentration in the linear range (**Eq. 4**). In account of the lower linear range concentrations of the nano setup (**Supplementary Figure S1**), the average concentration of standards used for LOD calculation were lower for the nano-setup than for the other two. LOD values averaged around 0.09 and 0.32 µg/L for standard and plasma extract in the analytical flow setup, respectively. LODs in the micro-setup reflected the gains in signal intensity (0.06 and 0.15 µg/L for standard and plasma extract, respectively). For the nano-setup, the gains in signal intensity did not equally translate to higher reproducibility of chromatographic peak area, resulting in LOD values much higher than expected judging from the high signal intensities (median values of 0.08 µg/L and 0.21 µg/L for standard and plasma extract, respectively). Drawing on the retention time, specific distributions of signal intensity and repeatability in the three setups, the actually obtainable profit regarding LOD was very much compound dependent. For some compounds finding already ideal chromatographic conditions and sufficient signal intensity in the analytical flow setup, LODs nominally even decreased upon miniaturization (**Supplementary Figure S2**). There are several ways to calculate the LOD and all give slightly different results, favoring systems with high signal stability over those with high signal intensity, or vice versa. At any rate, detection limits need to be understood as an estimate only. Regarding the nano-setup, we can draw from this comparison that LODs did not improve equally to signal intensity due to practical instrumental issues like carryover and impaired signal stability.

$$Limit\ of\ detection = \frac{3 * sd}{slope}, \tag{4}$$

where sd = standard deviation of the chromatographic peak area (background corrected) of repeated injections of standard/spiked plasma extract with a concentration equal to the lowest calibration point in the linear range (n = 4)
$slope$ = slope of calibration curve in the linear range

3.2.5 Matrix Effect

Matrix effect was calculated as the ratio of calibration curve slopes between plasma extract and standard for a panel of exogenous molecules. A crude human plasma extract was chosen as model matrix to challenge the systems with matrix complexity often encountered in *-omics* experiments of biological samples. On average, all setups showed signal suppression and the steepness of calibration curves was reduced through the matrix (**Figure 5**) without any obvious relation to retention time. The effect was especially notable in the nano-setup with an average sensitivity loss of almost 50% compared to matrix-free samples. The trap-and-elute configuration removed hydrophilic matrix components like salts that hamper ionization, but other matrix components were retained on the trap column and eluted in the relevant retention

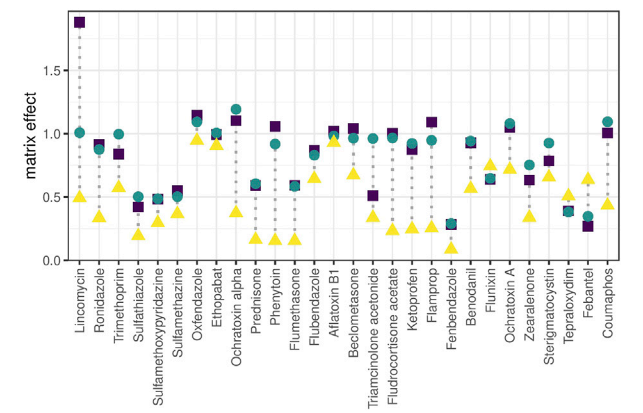


FIGURE 5 | Matrix effect. Matrix effect is calculated for exogenous compounds as the calibration curve slope (linear range) obtained for spiked plasma extract relative to the slope obtained for pure solvent. Ratio <1: matrix-related ion suppression, ratio >1: matrix-related signal enhancement, ratio = 1: signal intensity was not influenced by matrix effects. Molecules are ordered from low to high retention time (left to right).

time window together with the targeted analytes in up-concentrated form. The nano-system showed even higher ion suppression compared to micro- and analytical setup. This is attributed to reduced chromatographic resolution due to the observed chromatogram compression. In fact, this experiment showed that the nano-system is hardly compatible with the fast and steep gradients applied in wide-target small-molecule analysis.

4 DISCUSSION

For the present study we selected three platforms representing practical solutions in different *-omics* disciplines and demonstrated how different grades of LC miniaturization fundamentally affect chromatographic parameters related to successful non-targeted LC-ESI-MS-based *-omics* analysis. From a chromatographic perspective, system optimization for quantitative assays includes adjustments of separation, minimizing peak widths and maximizing signal intensity and chromatographic compound resolution (determined by retention, selectivity and efficiency) in the shortest possible time. In the application of non-target HRMS, this optimization strategy needs to be reassessed.

Miniaturization to micro-flow regime on average yielded moderately increased sensitivity as expected. The flow rate for the micro-flow setup ranged in a similar magnitude as the one used for the analytical flow setup and used the same ion source. The theoretical signal intensity increase is 4.4-fold based on reduced radial dilution on column. In practice, we saw that the actual profit is largely compound-dependent, and does for most of the molecules not entirely reflect enrichment success (expressed as peak concentration). This is related to the detection process. ESI-MS does not necessarily respond with twofold intensity when analyzing a sample with twice the concentration (Patti, 2011). Rather, increasing (peak) concentration to n-fold leads to a lower than n-fold increase in signal intensity, an effect coined as fold-change compression (Yu et al., 2020). Micro-LC falls far behind nano-LC regarding sensitivity increase, but the gain comes at almost no cost: Micro-LC can be installed on the same instruments as microbore LC and thus offers equal robustness, method adaptability and ease of use. An indispensable feature for the employed workflow was facile stopping and re-starting of the eluent flow, which enabled just-in-time (offline) mass calibration and optimum mass accuracy conditions, thus exploiting the full identification selectivity of HRMS. Notably, while offering only incremental improvement of signal intensity, micro-LC-ESI-MS equals the analytical flow platform regarding chromatographic selectivity, positive-negative-switching ability, peak shape, handling of eluent compositions, and steep gradients. Micro-LC suits the chemical diversity of small molecules as much as established analytical flow platforms with slightly increased signal intensities for most of the molecules (**Figure 6A**) and around ¼ of the eluent consumption.

Nano-LC coupled to MS is valued for enhancing ionization (Wilm, 2011) and reducing matrix effects, which potentially allows increasing signal intensities beyond chromatographic enrichment. Exploiting the benefits of “true nano-ESI” would be a unique argument in favor of using nano-LC for *-omics* analysis and could make up for the practical complications of using a nano-platform even when sample size is not limited. In practice and comparable to the micro-flow setup, we found that signal intensities were far below the theoretical 880-fold increase derived from the large-volume injection, and we did not see more efficient ionization or reduced matrix effect. The nano-platform was configured in accordance with practical proteomics solutions to maximize robustness, whereas the benefits of “true nano-ESI” only emerge at lower flow-rates (<50 nL/min) and with narrower spray tips (outer diameter of few µm) (Schmidt et al., 2003). The hardware setup we employed offers advantageous practical handling for metabolomics analysis–the (relatively) higher flow-rates can be precisely controlled and the larger inner diameter prevents the emitter from clogging, thus enabling serial analysis of protein-precipitated samples. Signal intensity and sensitivity were multiplied for many molecules upon miniaturization to the nano-flow platform, including several compounds below LOD under micro- and analytical flow. However, the platform’s unrivaled mass sensitivity based on chromatographic enrichment is tightly connected to a specific logP and *m/z* segment and was limited for smaller (<200 Da) and more hydrophilic (logP < −0.5) metabolites of the investigated panel. Some of the more polar compounds including most amino acids were completely lost to the nano-LC investigation (**Figure 6B**). Additionally, the nano-platform did not allow automated positive-negative-switching as the emitter needs to be positioned manually for negative ionization mode. This is necessary to maximize sensitivity and avoid corona discharge or breakdown of the spray. The chemical range covered by one run is thus reduced and the manual adjustments compromise automatability and quantitative reproducibility. Moreover, electrospray obtained with the nano-ESI source is not as stable as with the heated ESI source when adopting a wide range of eluent compositions (1–99% (v/v) organic). Overhead times for spray stabilization forbid just-in-time offline mass calibration and create

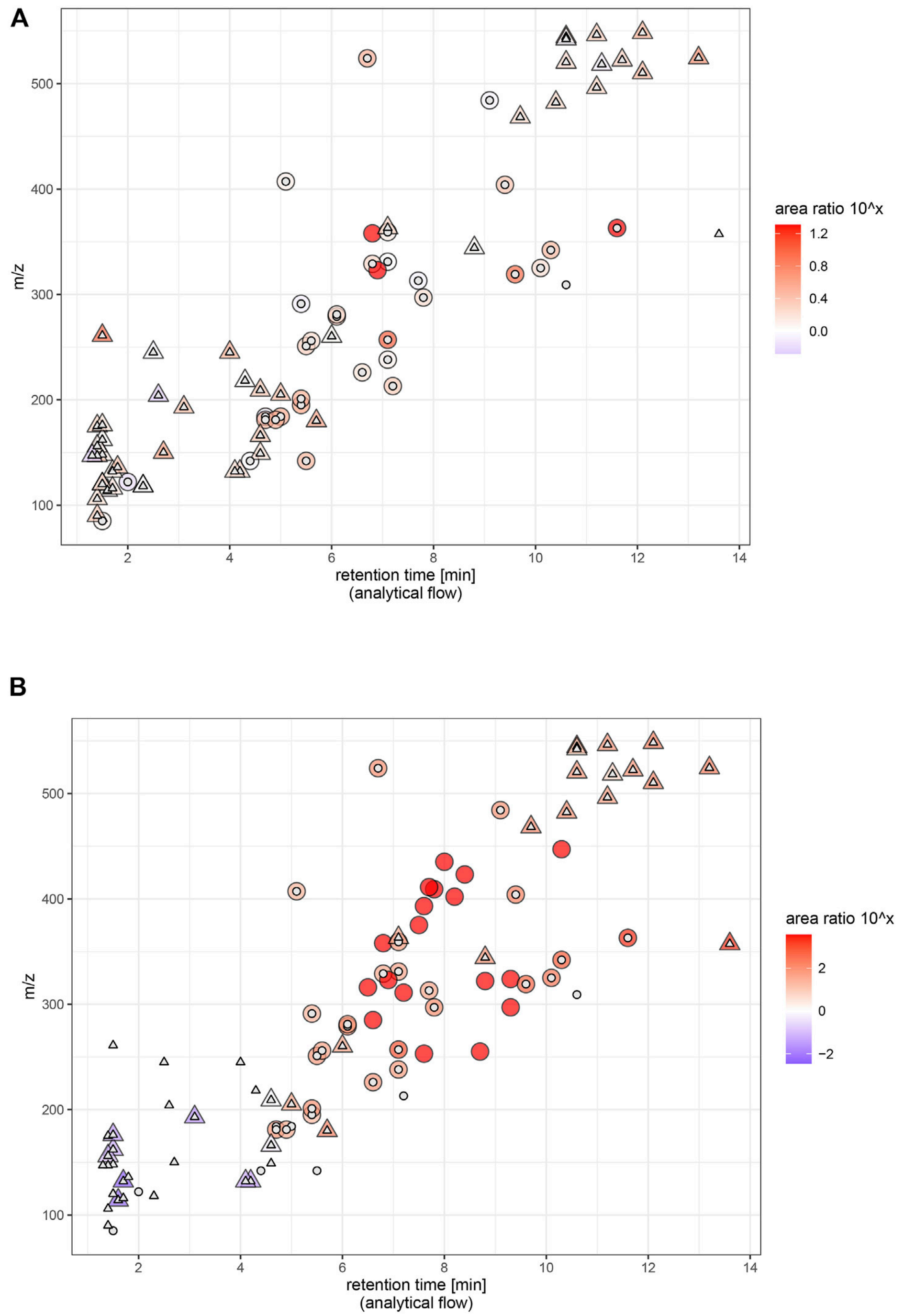


FIGURE 6 | Signal intensity (relative) and physicochemical coverage. Endogenous and exogenous molecules (ceftiofur and coumaphos 10 µg/L, caffeine, paraxanthine and theobromine naturally occuring abundance, all others 1 µg/L) in plasma extract. Panel **(A)** micro, panel **(B)** nano. Ratio = area in miniaturized setup/area in analytical flow setup. Triangles represent endogenous metabolites, circles represent exogenous analytes. Large icon = molecule has been found in miniaturized setup, small icon: molecule has been found in analytical flow setup. Areas are background corrected. LPC 20:1 was excluded to facilitate visual comparison (ratio ~6,000 in nano and ~200 in micro).

reluctance toward spontaneous method adaptations once the system is successfully running. Varying chromatographic peak shapes complicate parameter optimization for data-dependent fragmentation and peak-picking in non-targeted assays.

LC miniaturization is most promising for analyte panels with similar chemical properties, which make it possible to tailor chromatography and maximize chromatographic enrichment. As such, miniaturized chromatography has been successfully applied for highly sensitive peptidomics and lipidomics and has facilitated chemical residue analysis in environmental research (Wilson et al., 2015; Yi et al., 2017; Zardini Buzatto et al., 2020). However, is it also a viable approach for global metabolomics considering the broad diversity of metabolites? The covered physicochemical spectrum was demonstrably reduced under high degrees of miniaturization and we conclude that specificity of enrichment and the need to adapt chromatographic parameters more stringently to the compounds of interest, the problematic implementation of steep gradients and gradient extremes and the lack of positive-negative-switching capability contradict wide-spectrum small-molecule analysis with trap-and-elute nano-LC. Only by focusing specific compound classes with similar physicochemical properties or equalizing retention and ionization properties through chemical derivatization (Luo and Li, 2017), miniaturization to nano-flow regime will exert its true potential. Conversely, micro-LC offers the best compromise between improving signal intensity and metabolome coverage, despite the fact that only incremental gains can be achieved. Hence, we recommend using micro-LC for global metabolomics experiments.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Conceptualization: VF, GK, and YA. Data curation, formal analysis, methodology, visualization, and writing–original draft preparation, writing–review and editing: VF. Funding acquisition, resources, supervision, and writing–review and editing: GK. Investigation/practical work: VF and DB. Software: VF and YA; and all authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS

The authors thank all members of the Koellensperger Lab (University of Vienna) for continuous support and Prof. Benedikt Warth for supplying them with mycotoxin standards. They would further like to thank Mauro De Pra and Oleksandr Boychenko from Thermo Fisher Scientific for the scientific collaboration and Thermo Fisher Scientific for making an Ultimate 3000 RSLC nano-instrument available to their working group.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.857505/full#supplementary-material

Supplementary Figure S1 | Linear range in spiked plasma extract **(A)** and pure standard **(B)**. Instrument response was plotted against concentration and linearity was assessed by visual inspection of the resulting plot and linear regression line, supported by statistics and appropriate R^2 values (The Fitness for Purpose of Analytical Methods: A Laboratory Guide to Method Validation and Related Topics 2014). Tested concentration range: 0.01–100 µg/L for mycotoxins, 0.1–5,000 µg/L for all other molecules. Molecules are ordered by retention time (top left: lower rt, bottom right: higher rt). Sulfonamides/sulfonate, nitrates, and thiophosphates are marked with (*).

Supplementary Figure S2 | Limit of detection obtained for test compounds spiked to plasma extract based on area standard deviation of N = 4 repeated injections at the lowest concentration in the linear range. Calculation is described in the main text. Sulfonamides/sulfonate, nitrates, and thiophosphates are marked with (*).

REFERENCES

Adams, K. J., Pratt, B., Bose, N., Dubois, L. G., St. John-Williams, L., Perrott, K. M., et al. (2020). Skyline for Small Molecules: A Unifying Software Package for Quantitative Metabolomics. *J. Proteome Res.* 19 (4), 1447–1458. doi:10.1021/acs.jproteome.9b00640

Alseekh, S., Aharoni, A., Brotman, Y., Contrepois, K., D'Auria, J., Ewald, J., et al. (2021). Mass Spectrometry-Based Metabolomics: A Guide for Annotation, Quantification and Best Reporting Practices. *Nat. Methods* 18 (7), 747–756. doi:10.1038/s41592-021-01197-1

Bian, Y., Zheng, R., Bayer, F. P., Wong, C., Chang, Y.-C., Meng, C., et al. (2020). Robust, Reproducible and Quantitative Analysis of Thousands of Proteomes by Micro-flow LC–MS/MS. *Nat. Commun.* 11, 157. doi:10.1038/s41467-019-13973-x

Cebo, M., Fu, X., Gawaz, M., Chatterjee, M., and Lämmerhofer, M. (2020). Micro-UHPLC-MS/MS Method for Analysis of Oxylipins in Plasma and Platelets. *J. Pharm. Biomed. Anal.* 189, 113426. doi:10.1016/j.jpba.2020.113426

Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., et al. (2012). A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nat. Biotechnol.* 30 (10), 918–920. doi:10.1038/nbt.2377

Chetwynd, A. J., and David, A. (2018). A Review of Nanoscale LC-ESI for Metabolomics and its Potential to Enhance the Metabolome Coverage. *Talanta* 182, 380–390. doi:10.1016/j.talanta.2018.01.084

Chetwynd, A. J., David, A., Hill, E. M., and Abdul-Sada, A. (2014). Evaluation of Analytical Performance and Reliability of Direct NanoLC-NanoESI-High Resolution Mass Spectrometry for Profiling the (Xeno)Metabolome. *J. Mass Spectrom.* 49 (10), 1063–1069. doi:10.1002/jms.3426

Eurachem (2014). *The Fitness for Purpose of Analytical Methods: A Laboratory Guide to Method Validation and Related Topics.* 2nd ed. 2014. Available at: https://www.eurachem.org/images/stories/Guides/pdf/MV_guide_2nd_ed_EN.pdf (Accessed February 5, 2022).

Geller, S., Lieberman, H., Belanger, A. J., Yew, N. S., Kloss, A., and Ivanov, A. R. (2022). Comparison of Microflow and Analytical Flow Liquid Chromatography Coupled to Mass Spectrometry Global Metabolomics Methods Using a Urea Cycle Disorder Mouse Model. *J. Proteome Res.* 21 (1), 151–163. doi:10.1021/acs.jproteome.1c00628

Gray, N., Adesina-Georgiadis, K., Chekmeneva, E., Plumb, R. S., Wilson, I. D., and Nicholson, J. K. (2016). Development of a Rapid Microbore Metabolic Profiling Ultraperformance Liquid Chromatography-Mass Spectrometry Approach for High-Throughput Phenotyping Studies. *Anal. Chem.* 88 (11), 5742–5751. doi:10.1021/acs.analchem.6b00038

Greco, G., Boychenko, A., and Swart, R. (2016). *Robust LC-MS Analysis of Pesticides with 1.0mm i.d. Column Using the Vanquish Horizon UHPLC System*, 7. Waltham, MA: Thermo Fisher Scientific.

Juraschek, R., Dülcks, T., and Karas, M. (1999). Nanoelectrospray—More Than Just a Minimized-Flow Electrospray Ionization Source. *J. Am. Soc. Mass Spectrom.* 10 (4), 300–308. doi:10.1016/S1044-0305(98)00157-3

King, A. M., Trengove, R. D., Mullin, L. G., Rainville, P. D., Isaac, G., Plumb, R. S., et al. (2020). Rapid Profiling Method for the Analysis of Lipids in Human Plasma Using Ion Mobility Enabled-Reversed Phase-Ultra High Performance Liquid Chromatography/Mass Spectrometry. *J. Chromatogr. A* 1611, 460597. doi:10.1016/j.chroma.2019.460597

Kourtchev, I., Szeto, P., O'Connor, I., Popoola, O. A. M., Maenhaut, W., Wenger, J., et al. (2020). Comparison of Heated Electrospray Ionization and Nanoelectrospray Ionization Sources Coupled to Ultra-High-Resolution Mass Spectrometry for Analysis of Highly Complex Atmospheric Aerosol Samples. *Anal. Chem.* 92 (12), 8396–8403. doi:10.1021/acs.analchem.0c00971

Lanckmans, K., Van Eeckhaut, A., Sarre, S., Smolders, I., and Michotte, Y. (2006). Capillary and Nano-Liquid Chromatography-Tandem Mass Spectrometry for the Quantification of Small Molecules in Microdialysis Samples: Comparison with Microbore Dimensions. *J. Chromatogr. A* 1131 (1–2), 166–175. doi:10.1016/j.chroma.2006.07.090

Lu, W., Su, X., Klein, M. S., Lewis, I. A., Fiehn, O., and Rabinowitz, J. D. (2017). Metabolite Measurement: Pitfalls to Avoid and Practices to Follow. *Annu. Rev. Biochem.* 86, 277–304. doi:10.1146/annurev-biochem-061516-044952

Luo, X., and Li, L. (2017). Metabolomics of Small Numbers of Cells: Metabolomic Profiling of 100, 1000, and 10000 Human Breast Cancer Cells. *Anal. Chem.* 89 (21), 11664–11671. doi:10.1021/acs.analchem.7b03100

Nakatani, K., Izumi, Y., Hata, K., and Bamba, T. (2020). An Analytical System for Single-Cell Metabolomics of Typical Mammalian Cells Based on Highly Sensitive Nano-Liquid Chromatography Tandem Mass Spectrometry. *Mass Spectrom.* 9 (1), A0080. doi:10.5702/massspectrometry.A0080

Nguyen-Khuong, T., Pralow, A., Reichl, U., and Rapp, E. (2018). Improvement of Electrospray Stability in Negative Ion Mode for Nano-PGC-LC-MS Glycoanalysis via Post-Column Make-up Flow. *Glycoconj. J.* 35 (6), 499–509. doi:10.1007/s10719-018-9848-1

Noga, M., Sucharski, F., Suder, P., and Silberring, J. (2007). A Practical Guide to Nano-LC Troubleshooting. *J. Sep. Sci.* 30 (14), 2179–2189. doi:10.1002/jssc.200700225

Patti, G. J. (2011). Separation Strategies for Untargeted Metabolomics. *J. Sep. Sci.* 34 (24), 3460–3469. doi:10.1002/jssc.201100532

Psychogios, N., Hau, D. D., Peng, J., Guo, A. C., Mandal, R., Bouatra, S., et al. (2011). The Human Serum Metabolome. *PLoS ONE* 6 (2), e16957. doi:10.1371/journal.pone.0016957

Sanders, K. L., and Edwards, J. L. (2020). Nano-Liquid Chromatography-Mass Spectrometry and Recent Applications in Omics Investigations. *Anal. Methods* 12 (36), 4404–4417. doi:10.1039/D0AY01194K

Schmidt, A., Karas, M., and Dülcks, T. (2003). Effect of Different Solution Flow Rates on Analyte Ion Signals in Nano-ESI MS, or: When Does ESI Turn into Nano-ESI? *J. Am. Soc. Mass Spectrom.* 14 (5), 492–500. doi:10.1016/S1044-0305(03)00128-4

Shi, Y., Xiang, R., Horváth, C., and Wilkins, J. A. (2004). The Role of Liquid Chromatography in Proteomics. *J. Chromatogr. A* 1053 (1), 27–36. doi:10.1016/s0021-9673(04)01204-x,

Vasconcelos Soares Maciel, E., de Toffoli, A. L., Sobieski, E., Domingues Nazário, C. E., and Lanças, F. M. (2020). Miniaturized Liquid Chromatography Focusing on Analytical Columns and Mass Spectrometry: A Review. *Anal. Chim. Acta* 1103, 11–31. doi:10.1016/j.aca.2019.12.064

Vereyken, L., Dillen, L., Vreeken, R. J., and Cuyckens, F. (2019). High-Resolution Mass Spectrometry Quantification: Impact of Differences in Data Processing of Centroid and Continuum Data. *J. Am. Soc. Mass Spectrom.* 30 (2), 203–212. doi:10.1007/s13361-018-2101-0

Wilm, M. (2011). Principles of Electrospray Ionization. *Mol. Cell. Proteomics* 10 (7), M111.009407. doi:10.1074/mcp.M111.009407

Wilson, S. R., Vehus, T., Berg, H. S., and Lundanes, E. (2015). Nano-LC in Proteomics: Recent Advances and Approaches. *Bioanalysis* 7 (14), 1799–1815. doi:10.4155/bio.15.92

Yi, L., Piehowski, P. D., Shi, T., Smith, R. D., and Qian, W.-J. (2017). Advances in Microscale Separations towards Nanoproteomics Applications. *J. Chromatogr. A* 1523, 40–48. doi:10.1016/j.chroma.2017.07.055

Yu, H., Xing, S., Nierves, L., Lange, P. F., and Huan, T. (2020). Fold-Change Compression: An Unexplored but Correctable Quantitative Bias Caused by Nonlinear Electrospray Ionization Responses in Untargeted Metabolomics. *Anal. Chem.* 92 (10), 7011–7019. doi:10.1021/acs.analchem.0c00246

Zardini Buzatto, A., Kwon, B. K., and Li, L. (2020). Development of a NanoLC-MS Workflow for High-Sensitivity Global Lipidomic Analysis. *Anal. Chim. Acta* 1139, 88–99. doi:10.1016/j.aca.2020.09.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor WW declared a shared affiliation with the author(s) at the time of review.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 12 August 2022
DOI 10.3389/fmolb.2022.859787

Investigating the role of GLUL as a survival factor in cellular adaptation to glutamine depletion *via* targeted stable isotope resolved metabolomics

Şafak Bayram[1†], Yasmin Sophiya Razzaque[1†], Sabrina Geisberger[1†], Matthias Pietzke[1,2], Susanne Fürst[1,3], Carolina Vechiatto[1], Martin Forbes[1], Guido Mastrobuoni[1] and Stefan Kempa[1*]

[1]Proteomics and Metabolomics Platform, Max-Delbrück-Center for Molecular Medicine (MDC), Berlin Institute for Medical Systems Biology (BIMSB), Berlin, Germany, [2]Mass Spectrometry Facility, Max Planck Institute for Molecular Genetics, Berlin, Germany, [3]Theoretical Chemistry Quantum Chemistry, Institute for Chemistry, Technische Universität Berlin, Berlin, Germany

Cellular glutamine synthesis is thought to be an important resistance factor in protecting cells from nutrient deprivation and may also contribute to drug resistance. The application of "targeted stable isotope resolved metabolomics" allowed to directly measure the activity of glutamine synthetase in the cell. With the help of this method, the fate of glutamine derived nitrogen within the biochemical network of the cells was traced. The application of stable isotope labelled substrates and analyses of isotope enrichment in metabolic intermediates allows the determination of metabolic activity and flux in biological systems. In our study we used stable isotope labelled substrates of glutamine synthetase to demonstrate its role in the starvation response of cancer cells. We applied ^{13}C labelled glutamate and ^{15}N labelled ammonium and determined the enrichment of both isotopes in glutamine and nucleotide species. Our results show that the metabolic compensatory pathways to overcome glutamine depletion depend on the ability to synthesise glutamine *via* glutamine synthetase. We demonstrate that the application of dual-isotope tracing can be used to address specific reactions within the biochemical network directly. Our study highlights the potential of concurrent isotope tracing methods in medical research.

KEYWORDS

targeted stable isotope resolved metabolomics, GLUL, nucleotide biosynthesis, glutamine addiction, cancer metabolism, glutamine synthetase

Check for updates

OPEN ACCESS

EDITED BY
Wolfram Weckwerth,
University of Vienna, Austria

REVIEWED BY
Mariafrancesca Scalise,
University of Calabria, Italy
Denise Toscani,
University of Parma, Italy

*CORRESPONDENCE
Stefan Kempa,
stefan.kempa@mdc-berlin.de

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Metabolomics,
a section of the journal
Frontiers in Molecular Biosciences

RECEIVED 21 January 2022
ACCEPTED 15 July 2022
PUBLISHED 12 August 2022

CITATION
Bayram Ş, Razzaque YS, Geisberger S, Pietzke M, Fürst S, Vechiatto C, Forbes M, Mastrobuoni G and Kempa S (2022), Investigating the role of GLUL as a survival factor in cellular adaptation to glutamine depletion via targeted stable isotope resolved metabolomics.
Front. Mol. Biosci. 9:859787.
doi: 10.3389/fmolb.2022.859787

1 Introduction

Reprogramming of cellular metabolism was the earliest molecular phenotypes described in cancer cells; Otto Warburg described the preference of cancer cells to ferment pyruvate into lactic acid even in the presence of oxygen and Hanahan and Weinberg finally included metabolic reprogramming into their list of hallmarks of cancer (Warburg, 1956; Vander Heiden et al., 2009; Hanahan and Weinberg, 2011). Nowadays more and more detailed studies show a complex deregulation of cancer cell metabolism that is connected to growth and proliferation, immune cell evasion and also drug resistance mechanisms.

Despite a profound activation of glucose metabolism, cancer cells metabolise amino acids, such as glutamine (Lieu et al., 2020). Glutamine is a non-essential amino acid, and the most abundant in human blood plasma. Besides providing a source of energy, glutamine is required for several processes, including macromolecule biosynthesis, amino acid uptake, inhibition of autophagy and it triggers target of rapamycin (mTOR) kinase activation (Nicklin et al., 2009). Glutamine derived carbon fuels the tricarboxylic acid (TCA) cycle. The amino group of glutamine contributes to the synthesis of non-essential amino acids *via* the transaminase network and the amido-group serves as an obligate nitrogen donor for *de novo* nucleotide synthesis and hexosamine synthesis, specifically reactions that use the amido-nitrogen make glutamine a conditional essential metabolite (Altman et al., 2016; Bott et al., 2019).

Four different glutamine transport systems are characterized so far. These are known as SNAT3 (System N, SLC38A3) which is important in glutamine uptake in periportal cells in liver and in renal proximal tuble cells. SNAT1 (SLC38A1) is important in glutamine uptake by neuronal cells and ASCT2 (SLC1A5) is essential for glutamine uptake by rapidly growing epithelial cells and tumour cells in culture; the brush border membrane transporter B0 AT1 (SLC6A19) facilitates the uptake of glutamine across the kidney and intestinal brush border (McGivan and Bungard, 2007).

In the absence of sufficient extracellular glutamine, intracellular *de novo* synthesis can provide this essential metabolite. Glutamine synthetase (GS), also referred to as glutamate-ammonia ligase (GLUL) ligates glutamate with ammonia in an ATP-dependent condensation reaction (Nicklin et al., 2009). Several studies have revealed that the depletion of glutamine causes cell death (Eagle, 1955; Yuneva et al., 2007). This phenomenon, termed glutamine addiction, has been observed in a variety of cancer types in *in vitro* and *in vivo* studies (Wise and Thompson, 2011). Additionally, the reprogrammed metabolism of glutamine was shown to be crucial in tumorigenesis and tumour development (Yoo et al., 2020). Nevertheless, the molecular mechanisms underlying glutamine addiction are still not fully resolved.

Recent studies have demonstrated that glutamine deprived cells can be rescued by asparagine supplementation, for unclear reasons (Zhang et al., 2014). In the absence of glutamine, cells were rescued to a greater extent by asparagine supplementation relative to α-ketoglutarate, aspartate or glutamate (Zhu et al., 2017). In summary, asparagine has been demonstrated to regulate cell growth and rescue glutamine deficiency *via* several potential mechanisms (Zhang et al., 2014; Krall and Christofk, 2015; Zhang et al., 2017; Zhu et al., 2017; Pavlova et al., 2018). Understanding these mechanisms is fundamental to the development and efficacy of metabolic therapies targeting asparagine and glutamine metabolism. Interestingly rat stem cells transformed by the oncogenic Kaposi's sarcoma-associated herpesvirus (KSHV) demonstrated the capacity to utilise the amido group from both glutamine and asparagine for purine and pyrimidine biosynthesis (Zhu et al., 2017). In our study we have shown that the ability of colon cancer cells to compensate glutamine withdrawal by asparagine supplementation did solely depend on intracellular *de novo* glutamine synthesis by glutamine synthetase (GLUL).

Dejure and Royla examined the growth behaviour of a panel of cell lines under glutamine supplemented and glutamine depleted conditions (Dejure et al., 2017). All tested colon cancer cells (HCT116, GEO, HT29, SW480, RKO) stopped proliferation in glutamine depleted conditions, while HEK293 cells were able to proliferate. Our investigations revealed that the ability of HEK293 cells to proliferate in glutamine deprived conditions was abolished when dialyzed serum was used in the growth medium. Therefore, the previously observed "glutamine independency" of HEK293 cells may be attributed to remaining small molecules enabling glutamine synthesis. To identify which amino acids enable cell growth in glutamine depleted conditions, cell growth assays were performed with supplementation of either glutamine, asparagine, glutamate, aspartate and alanine with or without ammonium. GLUL's substrates, glutamate and ammonium, were associated with the greatest proliferation rate in the absence of glutamine in HEK293 and HCT116 cells. RKO cells were unable to proliferate in the absence of glutamine. The application of a competitive inhibitor of GLUL, methionine sulfoximine (MSO), prevented proliferation in the absence of glutamine also in HEK293 and HCT116 cells. Taken together, these findings pointed towards a key role for GLUL in adaptation to glutamine depletion.

To demonstrate GLUL activity and to determine the metabolic fate of GLUL's substrates, a dual-tracer and targeted Stable Isotope Resolved Metabolomics (SIRM) method was established. We developed a "targeted SIRM" dual isotope tracing technique in which substrates specific to a biological reaction are differentially labelled and monitored *via* high resolution mass spectrometry. In this case, the simultaneous application of ^{13}C-glutamate and ^{15}N-ammonium allowed us to detect the relative contribution of extracellular glutamate and ammonium to intracellular glutamine synthesis, as well as monitor the downstream contribution of glutamine's carbon and nitrogen to *de-novo* nucleotide biosynthesis. We also performed a time resolved dual-isotope tracing analysis and found the kinetics of glutamine synthesis in HEK293 and HCT116 cells are distinct, RKO cells did not show *de novo* glutamine synthesis, although the protein could be detected in proteomics analyses and western blot experiments.

Furthermore, we present growth conditions that preserve the viability of glutamine-dependent cancer cells under glutamine depletion. Our data show that all different amino acid supplementations that enable cell survival and proliferation with or without ammonium depend finally on the intracellular activity of GLUL. With the new established method of dual tracing and targeted pulsed stable isotope resolved metabolomics we could analyze the dynamics of intracellular glutamine synthesis.

2 Materials and methods

2.1 Cell culture

The standard cell culture medium (glutamine-supplemented medium) comprised Dulbecco's Modified Eagle Medium (DMEM, Thermo Fisher) without glucose (Glc), glutamine (Gln), phenol red or sodium pyruvate, supplemented with 10% dialyzed fetal bovine serum (dFBS), 2.5 g/L Glc, and 2 mM Gln. HEK293, HCT116 and RKO cells were grown in 10 cm plates at 37°C, 5% CO_2, 21% $O_{2,}$ and 85% relative humidity, and were passaged every 3,4 days to avoid contact inhibition and supply new media. When a confluency of at least 70% was reached, cells were washed once with 1x PBS and detached from the plate *via* trypsinization with TrypLE (GIBCO). Pre-warmed medium was added to cease trypsinization. The volume of medium added was calculated according to the desired splitting ratio and the cells were resuspended before the appropriate fraction of the cell suspension was transferred to a new plate. For cell growth assays and subsequent experiments, cells were harvested at a confluency of 80%–90% before being transferred to new plates at a seeding density of $2{\times}10^6$ cells which prevents contact inhibition.

2.2 Cell growth analysis

For the cell growth assays, pre-cultivated cells were seeded on 10 cm plates. The following day, the viable cell count was measured for the 0 hour time point and the cell culture medium was changed to that containing the appropriate condition (Gln: 2 mM; Alanine (Ala), Asparagine (Asn), Aspartate (Asp), Glutamate (Glu): all 1 mM, NH_4^+: 0.8 mM). Cells were passaged once they reached a confluency of at least 60%, upon which the cell count was determined. Media was replaced every 3,4 days to avoid limiting nutrients. Viability and cell number were monitored using the TC20 automated cell counter (Biorad).

2.3 Methionine sulfoximine inhibitor proliferation assay

Pre-cultivated cells were seeded on 6-well plates at a seeding density of $3{\times}10^5$ cells and $12{\times}10^5$ cells for HCT116 and HEK293 cells, respectively. The following day, the viable cell count was measured for the 0 hour time point and the cell culture medium was changed to that containing the appropriate culture condition (see Section 2.1) treated with either 500 µM Methionine Sulfoximine (MSO, Sigma Aldrich) or, as a negative control, sterile water (H_2O). The viable cell count was determined at 24, 48, 72, and 96 h post-treatment. Media was replaced daily to replenish substrates and remove secreted reaction products.

2.4 Western blotting

Cells grown in standard media conditions (not starved for glutamine) were washed with PBS and harvested in 1 ml ice-cold RIPA buffer. Cell lysates of HCT116, RKO and HEK293 were denatured in loading buffer for 5 min at 95°C. 40 µg of proteins were loaded and separated on a 10% SDS gel and run for 1 h at 70 V and 1 h at 120 V. The gel was transferred to a nitrocellulose membrane (0.2 µm, Biorad) at 25 V, 1 A for 30 min (Biorad TransBlot Turbo V1.02). The membrane was blocked for 1.5 h in 5% milk in TBS-T at room temperature and cut below the 70 kDa band. The membranes were incubated with primary antibodies against Vinculin (1:2,000 dilution, Sigma, V9131) and against GLUL (1:1,000 dilution, Thermo Fisher, PA1-46165) in 5% milk in TBS-T over-night at 4°C. After washing the membranes in TBST, the membranes were incubated in the HRP-conjugated secondary antibodies (NEB, 7074S; NEB, 7076S) for 1 h at room temperature. After washing the membranes in TBST and TBS, the membrane was developed using an ECL Western Blotting detection reagent (Amersham, RPN2109) according to the manufacturer's protocol. The Vilber FX gel system was used to record the luminescence (Vilber Lourmat, France).

2.5 Targeted stable isotope resolved metabolomics and pSIRM

Cells were pre-cultivated in Glu + NH_4^+-supplemented medium for at least 3 days prior to stable SIRM analysis. HEK293 and HCT116 cells were able to proliferate in Glu + NH_4^+-supplemented dFBS medium and were therefore pre-cultivated for over one month for cell growth assays before being seeded at a density of 2E+6 cells on 10 cm plates. RKO cells were unable to proliferate in Glu + NH_4^+-supplemented medium and were therefore pre-cultivated in Gln-supplemented dFBS medium and changed to medium containing Glu + NH_4^+ performed 3 days prior to the SIRM experiment.

For SIRM experiments, cells were then labelled for 24 h with ^{13}C labelled glutamate and ^{15}N labelled ammonium and treated in parallel with either 500 µM Methionine Sulfoximine (MSO, Sigma Aldrich) or, as a negative control, sterile water (H_2O). For pSIRM experiments, cells were pre-treated for 6 h with either 1 mM MSO or, as a negative control, sterile water, in fresh Glu + NH_4^+-supplemented dFBS medium. Afterwards, cells were labelled for 15 min, 30 min, 1 h, and 3 h with ^{13}C labelled glutamate and ^{15}N labelled ammonium, with or without 1 mM MSO. In both experiments, SIRM and pSIRM, 1 mM ^{13}C labelled glutamate was used. However, for SIRM experiments 96 µM ^{15}N labelled ammonium were used, whereas for pSIRM 0.8 mM ^{15}N labelled ammonium were used. Cells were harvested and extracted in a methanol-chloroform-water solution as described elsewhere [DOI: 10.1016/B978-0-12-801329-8.00009-X].

Intracellular amino acids were measured as TBDMS derivatives by high-resolution GC-MS. Dried cellular extracts were mixed with 25 µl MTBSTFA (Sigma) and 25 µl ACN and incubated at constant shaking for 1 h at 80°C. Derivatization was automatized on a TriPlus RSH auto-sampler (Thermo Fisher) and each sample was injected immediately after the derivatization. Samples were injected into a Q Exactive GC Orbitrap system (Thermo Fisher) with a splitof 1:5 (1 µl injection volume) in a temperature-controlled injector (TriPlus RSH auto-sampler, Thermo Fisher) with a baffled glass liner. The initial temperature was 80°C for 15 s, followed by an increase of 7°C/s up to 260°C, which is held for 3 min at the end of the temperature program. Gas chromatographic separation was carried out on a Trace 1,300 GC (Thermo Fisher) equipped with a TG-5SILMS column (30 m length, 250 µm inner diameter, 0.25 µm film thickness (Thermo Fisher). Helium was used as the carrier gas (1.2 ml/min flow rate). Gas chromatography was performed with an initial temperature of 68°C for 2 min, followed by an increase of 5°C/min up to 120°C, followed by an increase of 7°C/min up to 200°C, followed by an increase of 12°C/min up to 320°C which is held for 6 min. The spectra were recorded in a mass range of m/z = 60––600 with resolution at 200 m/z set at 120,000.

The elemental composition of different fragments for glutamine were calculated based on the exact mass and compared with known literature-values. To extract the intensities for the different isotopic masses we constructed a compound library including the mass shifts induced by ^{13}C and ^{15}N. Mass shifts were calculated *via* a custom R-Script based on the known masses for the fragments and the number of potentially incorporated carbon and nitrogen atoms. Each incorporated ^{13}C or ^{15}N increased the target mass by 1.0033548 or 0.99693689, respectively.

For the SIRM experiment, samples were then processed and peaks were integrated with Tracefinder 5.0 (Thermo Fisher), by importing this target list as a Tracefinder Compound database and extracting the extracted ion chromatograms (EIC) within a 5 ppm window. For the pSIRM experiment, samples were processed and peaks integrated in Xcalibur Quanbrowser (Thermo Fisher), extracting EIC with a mass tolerance of 2.5 ppm. For both, peak integration quality was visually checked and finally all peak areas were exported.

2.6 Measurement of free nucleotides

Free nucleotides were measured by direct infusion MS on a Q Exactive HF (Thermo Fisher) coupled to a Triversa Nanomate (Advion) nanoESI ion source. The Triversa Nanomate was operated in negative mode, with 1.5 kV spray voltage and 0.5 psi head gas pressure. The spectra were recorded for a duration of 3 min in a mass range of m/z = 140–850 m/z mass units with resolution at 200 m/z set at 240,000. A target list with 48 compounds was prepared in a similar way as described above. The M-H fragment was further calculated by subtracting 1.00728 from the exact mass of the uncharged molecule. For the extraction of the peak intensities the raw files were first converted to.dta2d files using TOPPAS FileConverter tool (Kohlbacher et al., 2007a; Kohlbacher et al., 2007b; Sturm et al., 2008).

The.dta2d files were then processed with a custom R script. Briefly, zero intensities and TIC intensities were removed from the datafiles as well the first and last five scans as these scans tend to be instable. All masses that fit into a 5 ppm window for each mass in the target list was associated to that specific compound. To extract only the apex the most intense mass per compound and scan was kept. Finally, the median and the standard deviation for all the scans was calculated to obtain a single readout per compound and sample.

Natural abundance correction for both types of experiment was performed using the Accucor package (URL: https://doi.org/10.1021/acs.analchem.7b00396).

3 Results

3.1 Cell growth assay in fetal bovine serum vs. dialyzed fetal bovine serum

Dejure and Royla, tested the effect of glutamine starvation on GEO, HCT116, HEK293, HT29, RKO and SW480 cells and found that all cell lines were not able to proliferate except HEK293 (Dejure et al., 2017).

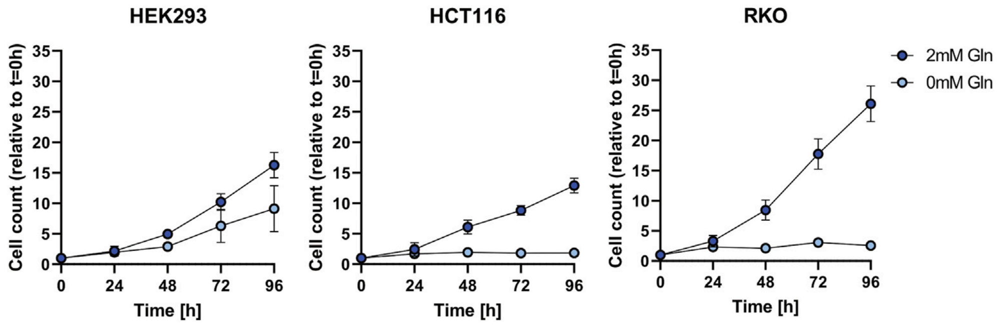


FIGURE 1
Cell growth assay for HEK293, HCT116, and RKO cells in non-dialyzed FBS with 2 mM or 0 mM glutamine (Gln) in the cell culture media. Cell count was determined every 24 h over the course of 96 h and is shown relative to *t* = 0 as mean ± SD.

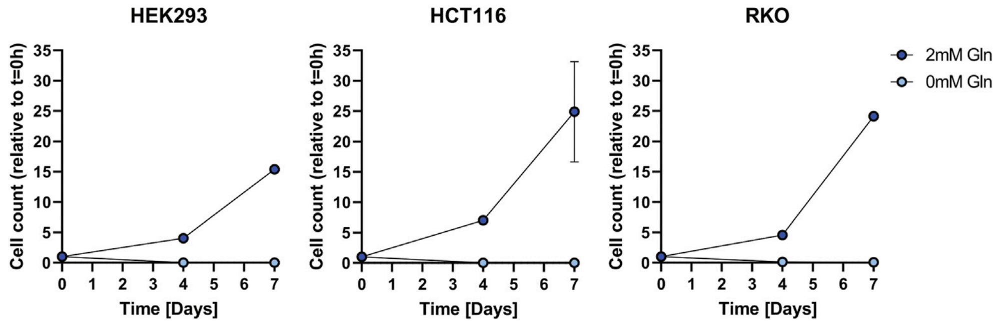


FIGURE 2
Cell growth assay for HEK293, HCT116, and RKO cells in dialyzed FBS with 2 mM or 0 mM glutamine (Gln) in the cell culture media. Cell count was determined over the course of 7 days and is shown relative to *t* = 0 as mean ± SD.

We investigated as to whether two colon cancer cell lines HCT116 and RKO can can adapt to glutamine depletion and removed glutamine from the medium for several days (Figure 1). Also in this experiment HEK293 cells exhibited glutamine independence, but RKO and HCT116 cells were not able to proliferate. In order to exclude that the glutamine independence of HEK293 cells is not caused by small molecules provided by the fetal bovine serum (FBS) we used dialyzed FBS and repeated the proliferation experiment (Figure 2). Interestingly, in this case also HEK293 cells were not able to proliferate when glutamine was deprived. Thus, glutamine was also essential for HEK293 cells when using dialyzed FBS.

3.2 Cell growth assay in supplemented dialyzed fetal bovine serum

In order to identify which metabolic pathways are efficiently utilised in glutamine-depleted condition, we monitored cell survival and growth upon supplementation with substrates of the glutamine-centric metabolic network. To achieve this, we supplemented: Alanine (Ala), Ala + ammonium (NH_4^+), Asparagine (Asn), Aspartate (Asp), Asp + NH_4^+, Glutamate (Glu), Glu + NH_4^+ in dialyzed FBS medium (Figure 3).

Viable cell count was determined at every passage over the course of 31 days. Cell count data were log2-transformed and graphically represented. Cell doubling time was calculated based on the division of culture duration by delta in log2-transformed cell counts. All three tested cell lines exhibit the highest proliferation rate and lowest doubling time in Gln-supplemented medium (Figure 3). In HEK293 and HCT116 cells the proliferation rate in Glu + NH_4^+-supplemented medium is close to that in Gln-supplemented medium. For HCT116 cells proliferation in Asp + NH_4^+-supplemented media is remarkably high. In HEK293 cells also the addition of glutamate leads to intermediate cell proliferation rates. Contrary, RKO cells can not compensate glutamine withdrawal under any condition. RKO cell viability decreased and cell death occurred, preventing the possibility of obtaining viable cell count data after 5 days onwards. Therefore, the

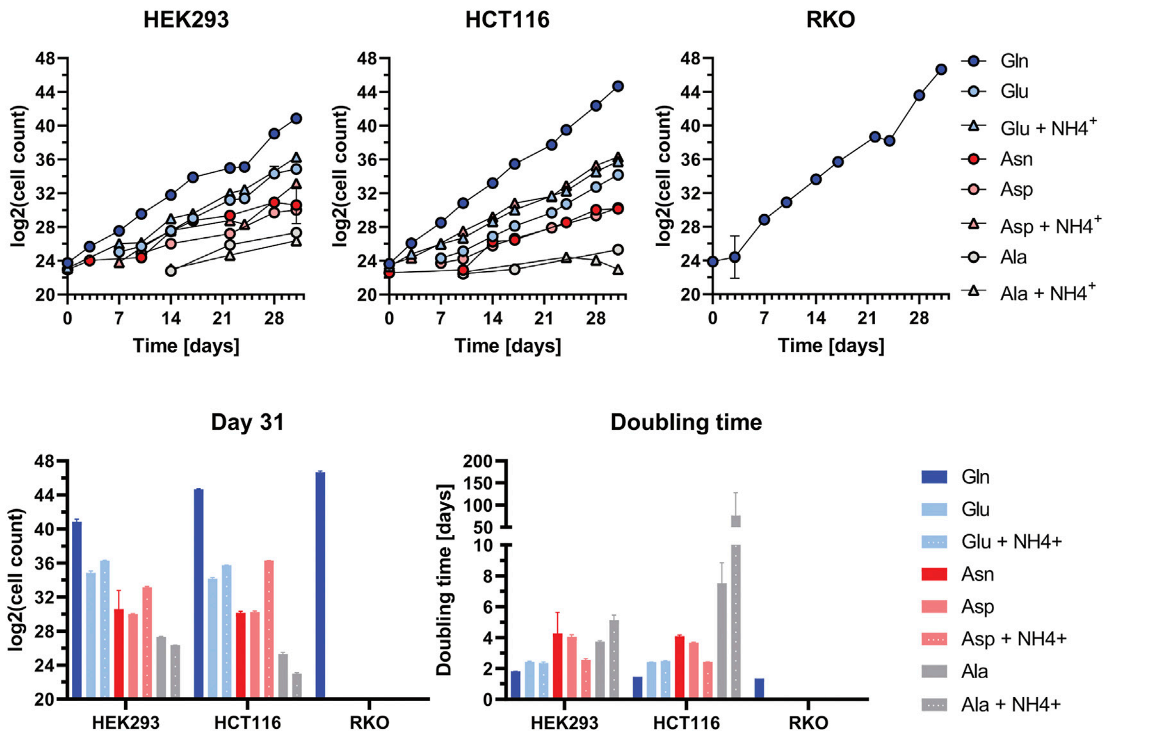


FIGURE 3
Cell Growth Assay in supplemented dialyzed FBS in HEK293, HCT116, and RKO cells. Cell growth upon the application of various amino acid substrates: Alanine (Ala): 1 mM; Ala: 1 mM + NH_4^+: 0.8 mM; Asparagine (Asn): 1 mM; Aspartate (Asp): 1 mM; Asp: 1 mM + NH_4^+: 0.8 mM; Glutamate (Glu): 1 mM; Glu: 1 mM + NH_4^+: 0.8 mM; Glutamine (Gln): 2 mM. Viable cell count was determined at every passage over the course of 31 days. Cell count data (each *n* = 2) were Log2-transformed and are shown as mean ± SD. The doubling time was calculated based on the duration in culture and the number of duplications underwent during this time (i.e., duration/Δlog2(cell count)).

proliferation rate for Ala, Ala + NH_4^+, Asn, Asp- Asp + NH_4^+, Glu, Glu + NH_4^+-supplemented media is 0.

3.3 Methionine sulfoximine inhibitor proliferation assay

The cell growth assays show that in glutamine-depleted dialyzed FBS conditions, HEK293 and HCT116 cells proliferate best when supplemented with the substrates of GLUL: Glu and NH_4^+(Figure 3). Based on this result, an inhibitor assay was performed to assess the effect of blocking *de novo* glutamine synthesis. Therefore, cells were treated with MSO, a competitive inhibitor of GLUL. As RKO cells are unable to proliferate in glutamine-depleted conditions, they were not subjected to this assay.

A pilot experiment (data not shown) was performed in HEK293 and HCT116 cells and an inhibitor concentration of 500 µM was found to be effective. The inhibitor-containing medium was refreshed and viable cell count was determined every 24 h over a 96 h time period. A parallel assay was performed using water, the solvent control, instead of MSO. Each measurement was taken from three biological replicates. The mean and standard deviation of the viable cell counts for each time point were graphically represented (Figure 4). Untreated HEK293 and HCT116 cells exhibit similar proliferation rates to those observed in the previous growth assay for all tested different conditions. However, MSO-treated HEK293 and HCT116 cells only show proliferation in Gln-supplemented medium. Showing that the chosen inhibitor concentration is not harmful to cells if glutamine is provided.

3.4 Targeted stable isotope resolved metabolomics and methionine sulfoximine-treatment

We designed a dual tracer stable isotope resolved metabolomics (SIRM) study using ^{13}C and ^{15}N labelled substrates (^{13}C glutamate, ^{15}N ammonium) to determine whether the cells utilise extracellular substrates for *de novo* glutamine synthesis and subsequent nucleotide biosynthesis. Based on the results of the cell growth assays, Glu + NH_4^+-supplemented medium was chosen to trace nitrogen and carbon

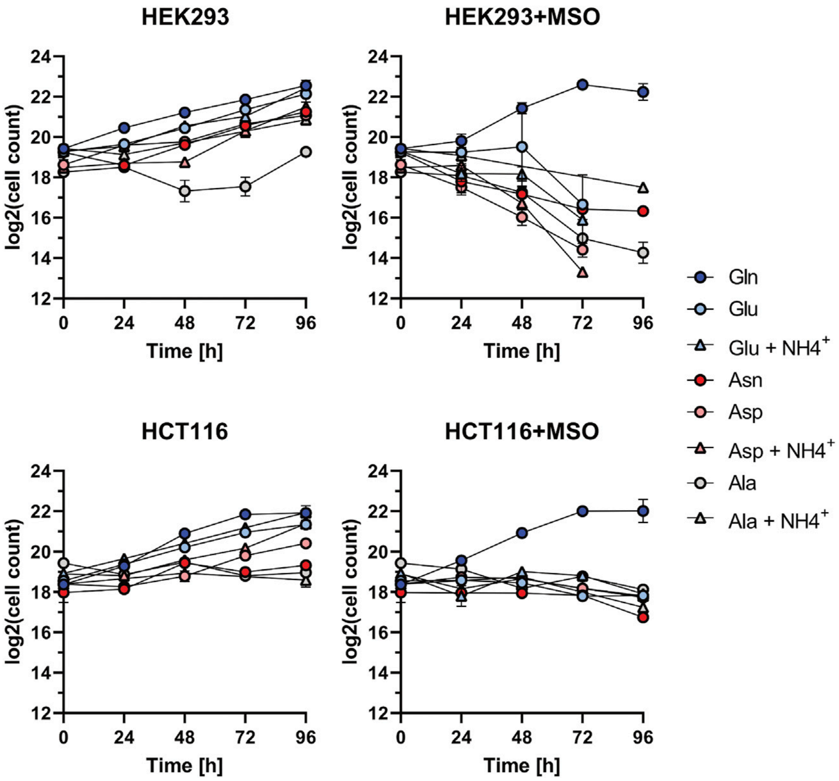


FIGURE 4
Proliferation inhibition assay for HEK293, HCT116, and RKO cells. Investigation of cell growth upon the application of GLUL inhibitor MSO (500 µM) or H_2O with Alanine (Ala): 1 mM; Ala: 1 mM + NH_4^+: 0.8 mM; Asparagine (Asn): 1 mM; Aspartate (Asp): 1 mM; Asp: 1 mM + NH_4^+: 0.8mM; Glutamate (Glu): 1mM; Glu: 1 mM + NH_4^+: 0.8 mM; Glutamine (Gln): 2 mM. Cell count (each *n* = 2) was determined every 24 h over the course of 96 h and is shown relative to *t* = 0 as mean ± SD.

incorporation into glutamine. The experiment was performed in three biological replicates.

Glutamine can be synthesised by GLUL-mediated ligation of glutamate and ammonium, with ammonium providing the amido-group. The ligation of $^{13}C_5$-glutamate and ^{15}N-ammonium was monitored by GC-MS detection of $^{13}C_5$-, $^{14}N_1$-, and $^{13}C_5$- $^{15}N_1$-glutamine isotopologues. Automated peak extraction from GC-MS spectra was performed *via* Tracefinder (Thermo Fisher) and mean ^{13}C and ^{15}N enrichment calculations were performed using custom R scripts. In HCT116 and HEK293 cells, ^{13}C- and ^{15}N-incorporation into glutamine was detected in several characteristic glutamine-3TBDMS fragments, as indicated by the corresponding ^{13}C- and ^{15}N-induced mass shift of the peaks. Glutamine-3TBDMS fragments comprising a 5C skeleton and 2N atoms underwent a mass shift of approximately 6 Da (m+6) while fragments comprising a 4C skeleton and 2N atoms underwent a mass shift of 5 Da (m+5), corresponding to $^{13}C_5$-$^{15}N_1$ and $^{13}C_4$- $^{15}N_1$ isotopologues, respectively. The cleanest signal was obtained for the fragment at 431 m/z and therefore this fragment was used for further analysis.

In the pilot experiment HCT116 and HEK293 cells were labelled for 24 h with 13C glutamate and 15N ammonium. In HEK293 cells 51% enrichment of ^{13}C and 2% enrichment of ^{15}N in the glutamine-3TBDMS fragment were monitored. HCT116 cells have almost twice as much ^{13}C enrichment at 87% and ^{15}N enrichment at 22%. In both HEK293 and HCT116 cells MSO treatment abolished ^{13}C and ^{15}N enrichment to 0%. RKO cells do not demonstrate ^{13}C or ^{15}N enrichment in untreated and MSO-treated conditions (Figure 5).

The contribution of newly synthesised glutamine isotopologues to nucleotide biosynthesis was monitored by direct-infusion MS detection of nucleotide isotopologues. Peak extraction from direct infusion-MS spectra was performed manually with XCalibur Qualbrowser (Thermo Fisher). In the *de novo* purine biosynthesis pathway, glutamine donates two nitrogen atoms to IMP and AMP, and three nitrogen atoms to GMP. HEK293 and HCT116 demonstrate ^{15}N enrichment in AMP and GMP while RKO cells do not. In HEK293 and HCT116 cells, one ^{15}N atom (N1) is incorporated into each AMP and GMP. HEK293 cells exhibit 25% enrichment of $^{15}N_1$-AMP and 20% enrichment of $^{15}N_1$-GMP, while HCT116 cells

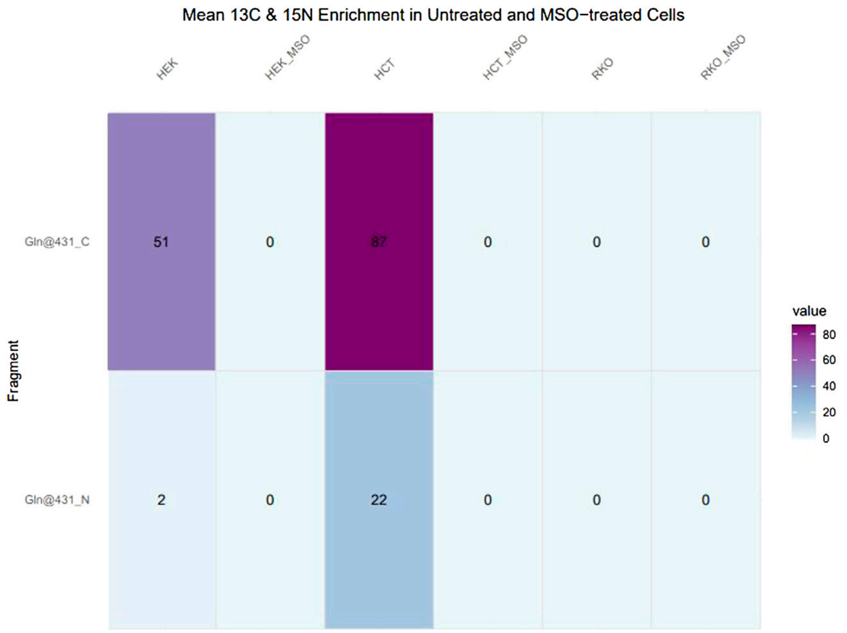


FIGURE 5
^{13}C and ^{15}N enrichment in glutamine in untreated and MSO-treated HEK293, HCT116 and RKO Cells. Cells were cultivated with ^{13}C5-glutamate and ^{15}N-ammonium for 24 h. After obtaining mass spectra, peak areas were extracted and natural isotope abundance correction and isotope enrichment calculations were performed. Data represent mean ^{13}C and ^{15}N enrichment (%).

exhibit 35% enrichment of $^{15}N_1$-AMP and 15% enrichment of $^{15}N_1$GMP (Figure 6).

In a second step we analyzed the dynamics of glutamine synthesis in HCT116 and HEK293 cells in a time course manner. The cell lines were incubated for 15 min, 30 min 1 h and 3 h with ^{13}C labeled glutamate and ^{15}N labeled ammonium. Label incorporation in glutamine was analyzed as described above (Figure 7). Interestingly label incorporation in glutamine can be found already after 15 min in HCT116 cells and after 30 min in HEK293 cells. Both cell lines show a fast glutamine synthesis but the kinetics are different. In HCT116 cells the incorporation of ^{15}N labeled ammonium into glutamine exceeds the formation of carbon and nitrogen labeled glutamine, this argues for faster ammonium import into HCT116 cells compared to HEK293 cells.

Interestingly, we performed a western blot analysis to analyze GLUL protein expression in the three cell lines and found that GLUL protein is present in all cell line even under normal conditions (Supplementary Material). We compared the western blot result with proteomics data (not shown) and could also find specific peptides for GLUL in all cell lines. Thus, the reason for the lacking GLUL activity in RKO cells cannot be explained by missing GLUL protein levels but must be caused by other reasons, e.g., mutations in the GLUL gene or impaired transport of substrates for GLUL reaction.

4 Discussion

So far stable isotope tracing studies with multiple isotopic tracers were performed without specific applications to demonstrate how this technology can add more information, compared to single isotope tracing methods. Here we show for the first time that this technology can be used to address specific reactions in the metabolic network and to address clinically relevant questions. In our study we analyzed the activity of glutamine synthetase (GLUL) by applying both substrates glutamate and ammonium labelled with stable isotopes.

Glutamine synthetase (GLUL), is of major interest, because this enzyme may be a resistance factor in metabolic cancer treatments; like the asparaginase treatment for acute lymphoblastic leukemia (ALL) and solid cancer (Rotoli et al., 2005). Glutamine is an important nutrient supporting cell growth and proliferation, oncogenic mutations often render cancer cells glutamine-dependent (Altman et al., 2016). In glutamine-depleted conditions, α-ketoglutarate, aspartate and glutamate supplementation have been demonstrated to rescue cell growth of glutamine-dependent cancer cells to a certain extend (Zhu et al., 2017).

In our study we analyzed the nature of glutamine addiction of three selected cell lines. HEK293 cells were partially glutamine auxotroph and HCT116 and RKO cells glutamine addicted (Dejure et al., 2017). We found that the

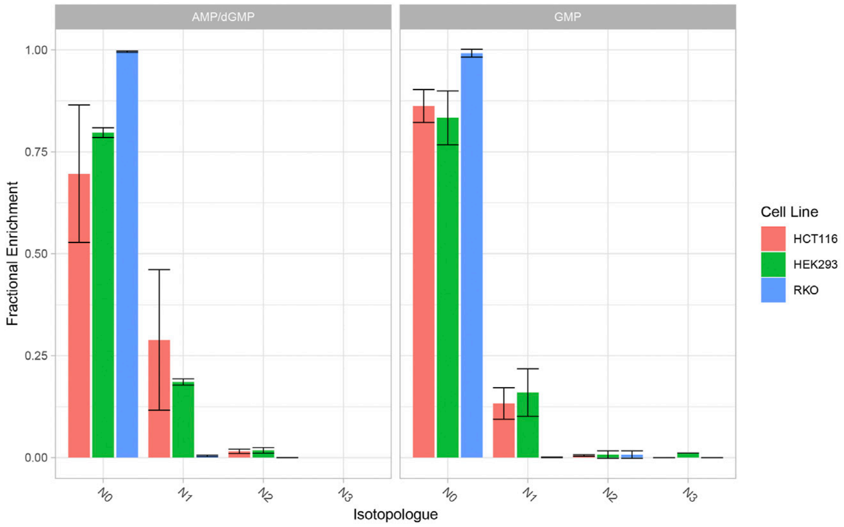


FIGURE 6
^{15}N Enrichment in AMP/GMP (purine nucleotides) in HEK293, HCT116 and RKO cells. Cells were cultivated with $^{13}C_5$-glutamate and ^{15}N-ammonium for 24 h. Nucleotide isotopologues were measured *via* direct-infusion MS and relative quantities are graphically represented. Data represent mean ± SD of three biological replicates.

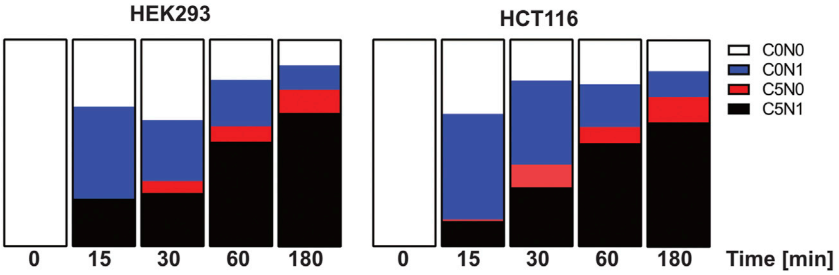


FIGURE 7
Isotope incorporation into glutamine after pulse labelling with $^{13}C_5$-glutamate and ^{15}N-ammonium in HCT116 and HEK293 cells. HCT116 and HEK293 cells were incubated with $^{13}C_5$-glutamate and ^{15}N-ammonium for 15 min, 30 min, 1 h, and 3 h (*n* = 2 each). Shown are the relative pool sizes of non-labelled (C0N0), ^{15}N labelled (C0N1), $^{13}C_5$ labelled (C5N0) ^{15}N-$^{13}C_5$ labelled (C5N1) glutamine.

ability of HEK293 cells to proliferate under glutamine deprived conditions did depend on the usage of non-dialyzed FBS, if we used dialyzed serum also HEK293 cells did depend on external glutamine supply. This demonstrates that glutamine addiction is found in all tested cell lines in our study.

In order to investigate the metabolic pathways that can contribute to glutamine autotrophy we applied a selected set of amino acids in a defined knock out medium. Although these conditions are artificial or synthetic compared to the natural environment of a cancer cell, this experiment can be used to understand the metabolic wires around glutamine (Figure 8); we supplemented: Alanine (Ala), Ala + ammonium (NH4+), Asparagine (Asn), Aspartate (Asp), Asp + NH4+, Glutamate (Glu), Glu + NH4+ in dialyzed FBS medium. Because of the absence of GLUL activity in RKO cells none of the supplements could contribute to cell growth and survival. This result clearly shows that all pathways that allow glutamine independency funnel into *de-novo* glutamine synthesis *via* GLUL. Similarly, the application of the specific GLUL inhibitor MSO abolished the capacity of HEK293 and HCT116 cells to proliferate without glutamine using the supplemented nutrients. Our results demonstrate once more that GLUL is the major player in glutamine independence.

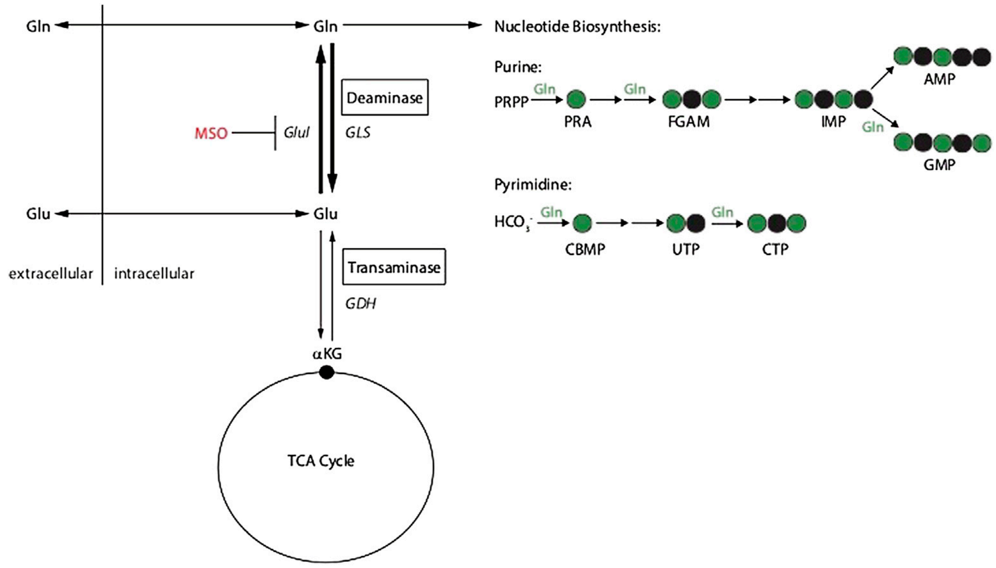


FIGURE 8
Schematic of glutamine metabolism and MSO inhibition of glutamine synthetase (GLUL).

Using high resolution mass spectrometry, we were able to monitor GLUL activity *via* a dual-tracer targeted SIRM approach. This method allowed us to measure a specified reaction by the application of multiple isotopic tracers. We observed in the pilot experiment a cell-line specific incorporation of extracellular ammonium into glutamine: HCT116 cells displayed a higher incorporation of extracellular ammonium for glutamine synthesis than HEK293 cells. However, we do not know if the available ammonium is spent within 24 h. In subsequent experiments the ammonium concentration was increased. In order to retrieve dynamic information about the uptake rates of individual substrates, a time course experiment was performed. The time and stable isotope resolved metabolomics experiments using multiple tracers delivered valuable information about the metabolic activity facilitated in the cell lines. The data show that HCT116 cells have a faster uptake of glutamate and that internal ammonium is used within the first 15 min of the pulse experiment. The data indicate that HCT116 cells possess higher glutamine synthesis rates. Both cell lines show high levels of stable isotope enrichment in glutamine after 30 min labeling time.

By using direct-infusion MS, we detected ^{13}C and ^{15}N enrichment in the *de novo* purine biosynthesis pathway in HEK293 and HCT116 cells, when ^{13}C glutamate and ^{15}N ammonium were supplied. To demonstrate the essentiality of GLUL activity to *de novo* glutamine synthesis and downstream nucleotide synthesis, inhibition of GLUL *via* MSO treatment ablated ^{13}C and ^{15}N incorporation.

Interestingly, RKO cells do not demonstrate ^{13}C and ^{15}N incorporation even in the absence of MSO treatment, indicating that either GLUL is inactive or the import of these substrates is compromised. The results from the dual-tracer targeted SIRM study reflect the observations from the cell growth and inhibitor assays (Figure 6). Taken together; the cell growth assays, inhibitor studies and SIRM analyses reveal that, in glutamine depleted conditions cell growth is dependent upon *de novo* glutamine synthesis.

The usage of the dual isotope tracing strategy to measure targeted and enzymatic activity in the cellular network in a time resolved manner is an advantage. This was not done so far, and our study is the first showing the power of this technology also for a clinically relevant question. We could show at multiple layers that, despite glutamine synthetases is expressed at the protein level, GLUL activity is absent in RKO cells, thus we show that expression levels alone cannot explain all metabolic activities.

The complex growth experiment highlights the role of an active glutamine synthesis to rescue glutamine withdrawal by using other amino acids, or a combination of amino acids and ammonium. We could also show that in all the reactions that we tested GLUL is the key enzyme and consequently; blocking glutamine synthetase with the inhibitor MSO abolishess growth and proliferation in the tested cell lines. Therefore, we propose that this method can be applied in clinical studies assessing different kinds of tumour cells and measuring glutamine synthetase activity *in vivo*.

Overall, we show that concurrent stable isotope labelling serves as a powerful tool for probing not only metabolic pathways, but also independent enzymatic reactions. Leveraging this tool enabled us to validate our observations from *in vitro* cell-based assays and demonstrate an essential reaction underlying the capacity of cells to adapt to glutamine-depletion. However, to detect mass shifts induced by small molecules such as ^{13}C and ^{15}N atoms, a very high resolution is needed (Su et al., 2017). In the absence of such a high resolution, mathematical models can be used to calculate the relative contribution of these molecules. We utilised the R package IsoCorrectoR to calculate the relative contribution of ^{13}C and ^{15}N in glutamine. For the purine nucleotides, we were able to resolve ^{13}C and ^{15}N incorporation from the raw data.

Our dual-tracer targeted SIRM study highlights the potential for high resolution mass spectrometry to monitor specific biological reactions at the atomic level. In future it can be envisioned to study more enzymatic reactions using concurrent isotope tracing techniques. We propose that all metabolic reactions that require two or more substrates that can be addressed with diverse isotopic labeling can be analyzed using this method, e.g., reactions within the *de novo* nucleotide biosynthesis or hexose amine biosynthesis. This will be of mayor advantage if enzymatic activity essays are not established.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

SB, YR, SG, and CV performed experiments. SB, YR, SG, MF, MP, SF, GM, and SK analyzed the data. SB, YR, MF, MP, and SK wrote manuscript. SK supervised the study.

Funding

SB was supported by the Sander Foundation and SF by DKTK (German consortium for translational cancer research). MP was funded by the Helmholtz association (AMPRO), YR and MF by Deutsche Krebshilfe (Enable). SG was supported by the Bundesministerium für Bildung und Forschung funding MSTARS (Multimodal Clinical Mass Spectrometry to Target Treatment Resistance).

Acknowledgments

We thank Jenny Grobe for excellent technical assistance.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.859787/full#supplementary-material

References

Altman, B. J., Stine, Z. E., and Dang, C. V. (2016). From Krebs to clinic: Glutamine metabolism to cancer therapy. *Nat. Rev. Cancer* 16 (10), 619–634. doi:10.1038/nrc.2016.71

Bott, A. J., Maimouni, S., and Zong, W. X. (2019). The pleiotropic effects of glutamine metabolism in cancer. *Cancers (Basel)*, 11, 770. doi:10.3390/cancers11060770

Dejure, F. R., Royla, N., Herold, S., Kalb, J., Walz, S., Ade, C. P., et al. (2017). The MYC mRNA 3′-UTR couples RNA polymerase II function to glutamine and ribonucleotide levels. *EMBO J.* 36, 1854–1868. doi:10.15252/embj.201796662

Eagle, H. (1955). Nutrition needs of mammalian cells in tissue culture. *Science* 122 (3168), 501–514. doi:10.1002/9780470114735.hawley00624

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell.* 144 (5), 646–674. doi:10.1016/j.cell.2011.02.013

Kohlbacher, O., Reinert, K., Gröpl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., et al. (2007). TOPP—The OpenMS proteomics pipeline. *Bioinformatics* 23 (2), e191–e197. doi:10.1093/bioinformatics/btl299

Kohlbacher, O., Reinert, K., Gröpl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., et al. (2007). Topp - the OpenMS proteomics pipeline. *Bioinformatics* 23, e191–e197. doi:10.1093/bioinformatics/btl299

Krall, A. S., and Christofk, H. R. (2015). Rethinking glutamine addiction. *Nat. Cell. Biol.* 17, 1515–1517. doi:10.1038/ncb3278

Lieu, E. L., Nguyen, T., Rhyne, S., and Kim, J. (2020). Amino acids in cancer. *Exp. Mol. Med.* 52, 15–30. doi:10.1038/s12276-020-0375-3

McGivan, J. D., and Bungard, C. I. (2007). The transport of glutamine into mammalian cells. *Front. Biosci.* 12, 874–882. doi:10.2741/2109

Nicklin, P., Bergman, P., Zhang, B., Triantafellow, E., Wang, H., Nyfeler, B., et al. (2009). Bidirectional transport of amino acids regulates mTOR and autophagy. *Cell.* 136 (3), 521–534. doi:10.1016/j.cell.2008.11.044

Pavlova, N. N., Hui, S., Ghergurovich, J. M., Fan, J., Intlekofer, A. M., White, R. M., et al. (2018). As extracellular glutamine levels decline, asparagine becomes an essential amino acid. *Cell. Metab.* 27, 428–438. doi:10.1016/j.cmet.2017.12.006

Rotoli, B. M., Uggeri, J., Dall'Asta, V., Visigalli, R., Barilli, A., Gatti, R., et al. (2005). Inhibition of glutamine synthetase triggers apoptosis in asparaginase-resistant cells. *Cell. Physiol. biochem.* 15 (6), 281–292. doi:10.1159/000087238

Sturm, M., Bertsch, A., Gröpl, C., Hildebrandt, A., Hussong, R., Lange, E., et al. (2008). OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinforma.* 9, 163. doi:10.1186/1471-2105-9-163

Su, X., Lu, W., and Rabinowitz, J. D. (2017). Metabolite spectral accuracy on orbitraps. *Anal. Chem.* 89 (11), 5940–5948. doi:10.1021/acs.analchem.7b00396

Vander Heiden, M., Cantley, L., and Thompson, C. (2009). Understanding the Warburg effect: The metabolic requirements of cell proliferation. *Sci.* 324 (5930), 1029–1033. doi:10.1126/science.1160809

Warburg, O. (1956). On the origin of cancer cells on the origin of cance. *Source Sci. New Ser.* 123 (123), 309–314. doi:10.1126/science.123.3191.309

Wise, D. R., and Thompson, C. B. (2011). Glutamine addiction: A new therapeutic target in cancer. *Trends Biochem. Sci.* 35 (8), 427–433. doi:10.1016/j.tibs.2010.05.003

Yoo, H. C., Park, S. J., Nam, M., Kang, J., Kim, K., Yeo, J. H., et al. (2020). A variant of SLC1A5 is a mitochondrial glutamine transporter for metabolic reprogramming in cancer cells. *Cell. Metab.* 31, 267–283. doi:10.1016/j.cmet.2019.11.020

Yuneva, M., Zamboni, N., Oefner, P., Sachidanandam, R., and Lazebnik, Y. (2007). Deficiency in glutamine but not glucose induces MYC-dependent apoptosis in human cells. *J. Cell. Biol.* 178 (1), 93–105. doi:10.1083/jcb.200703099

Zhang, J., Fan, J., Venneti, S., Cross, J. R., Takagi, T., Bhinder, B., et al. (2014). Asparagine plays a critical role in regulating cellular adaptation to glutamine depletion. *Mol. Cell.* 56, 205–218. doi:10.1016/j.molcel.2014.08.018

Zhang, J., Pavlova, N. N., and Thompson, C. B. (2017). Cancer cell metabolism: The essential role of the nonessential amino acid, glutamine. *EMBO J.* 36, 1302–1315. doi:10.15252/embj.201696151

Zhu, Y., Li, T., Da Silva, S. R., Lee, J., Lu, C., Eoh, H., et al. (2017). A critical role of glutamine and asparagine γ-Nitrogen in nucleotide biosynthesis in cancer cells hijacked by an oncogenic virus. *MBio* 122, 501–514. doi:10.1128/mBio.01179-17

frontiers | Frontiers in Molecular Biosciences

TYPE Review
PUBLISHED 08 September 2022
DOI 10.3389/fmolb.2022.972190

Check for updates

OPEN ACCESS

EDITED BY
Wolfram Weckwerth,
University of Vienna, Austria

REVIEWED BY
Oleg Mayboroda,
Leiden University Medical Center,
Netherlands
Fumio Matsuda,
Osaka University, Japan

*CORRESPONDENCE
R. Verpoorte,
VERPOORT@Chem.Leidenuniv.NL

SPECIALTY SECTION
This article was submitted to
Metabolomics,
a section of the journal
Frontiers in Molecular Biosciences

RECEIVED 17 June 2022
ACCEPTED 11 August 2022
PUBLISHED 08 September 2022

CITATION
Verpoorte R, Kim HK and Choi YH
(2022), Trivialities in metabolomics:
Artifacts in extraction and analysis.
Front. Mol. Biosci. 9:972190.
doi: 10.3389/fmolb.2022.972190

Trivialities in metabolomics: Artifacts in extraction and analysis

R. Verpoorte[1]*, H. K. Kim[1] and Y. H. Choi[1,2]

[1]Natural Products Laboratory, Institute of Biology Leiden, Leiden University, Leiden, The Netherlands, [2]College of Pharmacy, Kyung Hee Univeristy, Seoul, South Korea

The aim of this review is to show the risks of artifact formation in metabolomics analyses. Metabolomics has developed in a major tool in system biology approaches to unravel the metabolic networks that are the basis of life. Presently TLC, LC-MS, GC-MS, MS-MS and nuclear magnetic resonance are applied to analyze the metabolome of all kind of biomaterials. These analytical methods require robust preanalytical protocols to extract the small molecules from the biomatrix. The quality of the metabolomics analyses depends on protocols for collecting and processing of the biomaterial, including the methods for drying, grinding and extraction. Also the final preparation of the samples for instrumental analysis is crucial for highly reproducible analyses. The risks of artifact formation in these steps are reviewed from the point of view of the commonly used solvents. Examples of various artifacts formed through chemical reactions between solvents or contaminations with functional groups in the analytes are discussed. These reactions involve, for example, the formation of esters, *trans*-esterifications, hemiacetal and acetal formation, N-oxidations, and the formation of carbinolamines. It concerns chemical reactions with hydroxyl-, aldehyde-, keto-, carboxyl-, ester-, and amine functional groups. In the analytical steps, artifacts in LC may come from the stationary phase or reactions of the eluent with analytes. Differences between the solvent of the injected sample and the LC-mobile phase may cause distortions of the retention of analytes. In all analytical methods, poorly soluble compounds will be in all samples at saturation level, thus hiding a potential marker function. Finally a full identification of compounds remains a major hurdle in metabolomics, it requires a full set of spectral data, including methods for confirming the absolute stereochemistry. The putative identifications found in supplemental data of many studies, unfortunately, often become "truly" identified compounds in papers citing these results. Proper validation of the protocols for preanalytical and analytical procedures is essential for reproducible analyses in metabolomics.

KEYWORDS

artifacts, metabolomics, extraction, solvents, decomposition

1 Introduction

In the past 2 decades metabolomics has rapidly developed as an important tool in studying various biological and medical questions. The aim of metabolomics is the qualitative and quantitative analysis of all small molecules present in biological samples. By comparing the metabolome of organisms under different conditions, information can be obtained about the regulation of the metabolic network and the potential biological role of specific compounds. The analysis of the metabolome is complicated because a wide spectrum of compounds with totally different physico-chemical properties are present in a wide range of concentrations. That makes the measurement of all small molecules in an organism in one operation the major challenge for the 21st century's (aL)chemistry.

The analysis of the small molecules is done by means of chromatographic separations (TLC, LC or GC) coupled to UV-, mass- (MS), or nuclear magnetic resonance (NMR) spectrometry. Alternatively the analysis is done directly by MS/MS or NMR. Each of these methods has its own advantages and disadvantages. These aspects will be dealt with in other chapters. In this chapter we want to focus on some basic problems that one should keep in mind when developing a metabolomics analysis.

The goal of this chapter is to highlight the problems of artifact formation in the preanalytical and analytical procedures. Artifacts are described as the compounds that are not present in an intact metabolome but are formed in the process of harvesting, drying, grinding, extracting, and preparing samples for analysis, and during the separation and detection phase of the analysis.

Artifacts can be formed by a reaction of an analyte with the solvents itself or with contaminants in solvents. In the books on Chromatography of Alkaloids (Baerheim Svendsen and Verpoorte 1983; Verpoorte and Baerheim Svendsen. 1984) we have reviewed some of these problems in connection with alkaloids. In other papers we discussed some basic aspects of metabolomic analyses (Verpoorte et al., 2008) and artifact formations with solvents (Maltese et al., 2009). The present chapter summarizes these earlier papers as well as some more recent examples. This information should be useful to have at hand in a book on metabolomics. This is not rocket science, but it concerns trivial things that people forget when running automated instrumental analyses. For newcomers not trained in natural products chemistry or analytical chemistry this may be new information. At least when reviewing papers in this field we are often surprised by the ignorance about basic methods to prepare the sample (extract) for the final high-tech analysis.

2 Solvents for extraction and chromatography

To discuss artifact formation of all known compounds would be a huge task, not to speak about all unknown compounds. Instead, we will focus on the most common solvents and their known contaminations (Table 1) that may play a role in generating artifacts. For a comprehensive survey of contaminations in solvents and reagents is referred to Middleditch (1989) and Venditti (2020). In the former publication, the molecular weights and mass spectra of common solvents, additives, and contaminations like plasticizers, paper whiteners, rubber constituents, and antioxidants, are described. In addition, there are review papers on specific group of metabolites. Capon (2020) reviewed the artifact formation for marine natural products. The possible mechanisms of the artifact formation are discussed in depth. They vary from simple esterification, solvolysis and oxidation to highly complex chemical rearrangements. Artifact formation of various terpenoids was reviewed by Hanson (2017). Dehydration, rearrangements, and oxidation, among others, cause formation of artifacts from all kinds of terpenes. Xu and colleagues (2020) reported methods to predict potential artifacts by reactions with methanol or oxidation. Venditti (2020) particularly discussed monoterpenoid artifacts.

3 Stability of analytes

Papers describing stability of compounds under different conditions often claim that a compound is not stable, but there are very few papers that really have known standards as controls. In our experience terpenoid indole alkaloids are not very stable. People working with isoquinoline alkaloids claimed that these alkaloids were very labile. However, working on both types of alkaloids, it was obvious that most isoquinoline alkaloids were more stable than the indole alkaloids. Apparently, there are more feelings, than there is understanding. It is difficult to predict solubility and stability of pure compounds. In general, the experience is that light and heat are an important factors. Keep the compounds, when dissolved, always in the dark, at the lowest possible temperature. This also means that extractions using Soxhlet equipment are extremely detrimental for the analysis of the true metabolome of an organism. Alcohols are in general the best solvents to store compounds and extracts. In general, the stability of compounds in the halogenated solvents is low, compounds like the anhydronium indole alkaloids serpentine and alstonine do not survive dissolving in chloroform (Verpoorte and Sandberg 1971; Baerheim Svendsen and Verpoorte 1983). Also reserpine and related indole alkaloids are rapidly oxidized in chloroform (Wright and Tang, 1972). The pH does play a role, though for each compound the optimum can be different. Moreover in mixtures there can be differences, e.g. by the presence of natural antioxidants in extracts. In NMR-based metabolomics the use of fully deuterated solvents, like methanol and D_2O, may cause the replacement of certain protons with deuterium, e.g. in aldehydes and ketones via a keto-enol equilibrium. For example, naringenin has been shown to have two phloroglucinol protons to be completely

TABLE 1 The example of impurities and reactions of the solvents most commonly used in phytochemistry.

Solvent class	Solvents	Contaminations	artifacts
Alcohol	Methanol, ethanol, propanediol, glycol, glycerol	Aldehydes	Esters, acetals, hemiacetals, carbinolamines
Ethers	Diethyl ether, tetrahydrofuran	Peroxides, aldehydes, alcohols	N-oxides, carbinolamines, esters, acetals, hemiacetals
Esters	Ethyl acetate	Acetaldehyde	*Trans*-esterifications, Esters, acetals, hemiacetals, carbinolamines
Acetonitrile	Acetonitrile		Acetamide, BHT, dichlorobenzene, glutatonitrile, succinonitrile
Chloroform	Chloroform	Phosgene, CH_2BrCl, CH_2Cl_2	Quaternary amines
Dichloromethane	Dichloromethane	CH_2BrCl, CNCl	Quaternary amines cyanides
Aromatic hydrocarbons	Toluene		Various Hydrocarbons

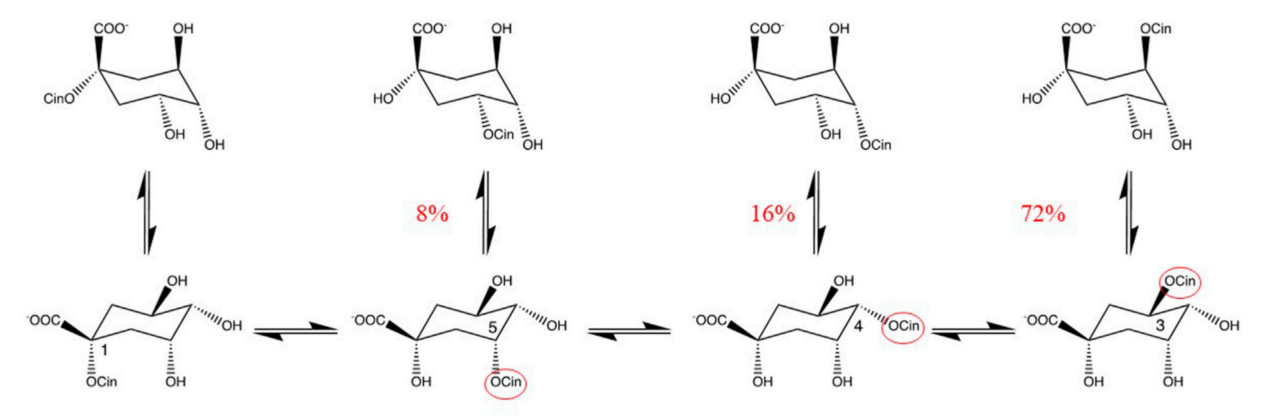


FIGURE 1
Isomerization of chlorogenic acids. Intramolecular migration of the cinnamoyl-group (cin) in pure chlorogenic acid (100% pure 3-cinnamoylquinic acid) when 3 min in phosphate buffer pH 7, at 90°C (Hanson, 1965). Percentages given are quantities relative the pure chlorogenic acid. cin = cinnamoyl.

exchanged in the presence of a solvent with a deuterated hydroxyl group (Verpoorte et al., 2008). Finally, one should also keep in mind that the stationary phases used in chromatography also play a role. For example, Pauli and co-workers (Tang et al., 2021) showed that silica affects the oxidation of prenyl groups in various natural products.

4 Reactions of solvent or solvent's impurities with analytes

Many preanalytical protocols have been reported for extraction and sample preparation. Here we will confine us to the solvent itself as a chemical that may react with analytes and illustrate this with some examples.

Alcohols Alcohols are often used in extraction. Methanol is commonly used to extract biological samples for metabolomic analysis. However, methanol is toxic and thus for applications in food, medicines or cosmetics, ethanol, 1,2-dihydroxypropane, glycol and glycerol are preferred. In liquid chromatography methanol is often used as component of the mobile phase. The reactive site of alcohols is a hydroxyl group. The reaction of carboxyl group(s) of analytes with alcohols may yield esters. Even, inter- and intra-molecular *trans*-esterifications may occur. The fast isomerization of chlorogenic acid is a good example of an intramolecular *trans*-esterification of the cinnamoyl group. Within 3 min at 90°C in water pH 7, about 28% of the pure chlorogenic acid was found to be isomerized (Hanson, 1965; Clifford et al., 1989) (Figure 1). This example shows that one should be very careful in drawing conclusions from any changes in the levels of these compounds, which are ubiquitous in plants.

Methanol and ethanol have been reported to react with fatty acids to generate esters during extractions (Lough et al., 1962; Johnson et al., 1976; Xu et al., 2020). Brondz and colleagues (2007) reported esterification of the carboxylic acid group in β-carboline alkaloids. The effect of methanol was studied in more

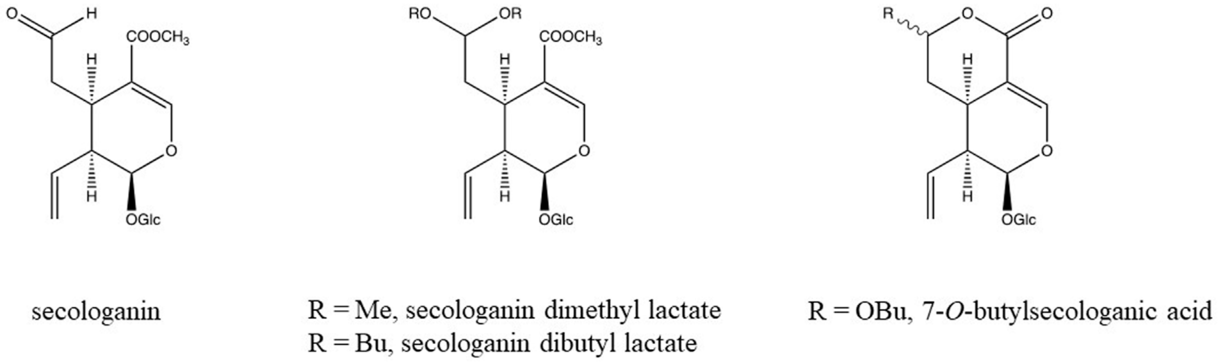

COOCH3
H
O
OGlc
RO
OR
R
secologanin
R = Me, secologanin dimethyl lactate
R = Bu, secologanin dibutyl lactate
R = OBu, 7-O-butylsecologanic acid

FIGURE 2
Chemical structure of secologanin and artifacts formed during isolation in alcoholic solvents (Verpoorte unpublished results, Tomassini et al., 1995).

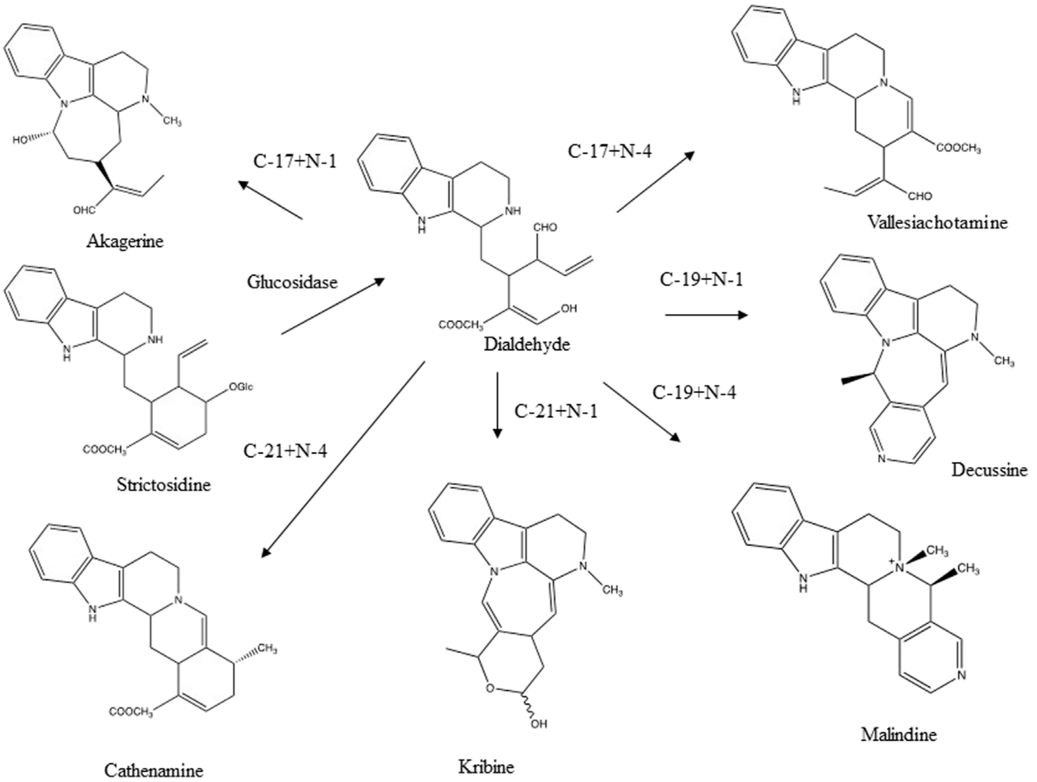


FIGURE 3
Various skeletons of terpenoid indole alkaloids formed through the intramolecular reaction of the aldehyde group and an amine function after glucolysis of strictosidine (Verpoorte, 2000).

detail by Xu and coworkers (2020). They developed a chemometric tool to predict potential artifacts by reactions with methanol or oxidation. Some benzylisoquinoline alkaloids and caffeic acid derivatives were used as examples. Sauerschnig and colleagues (2018) reported many other examples of artifact formation with methanol. By using

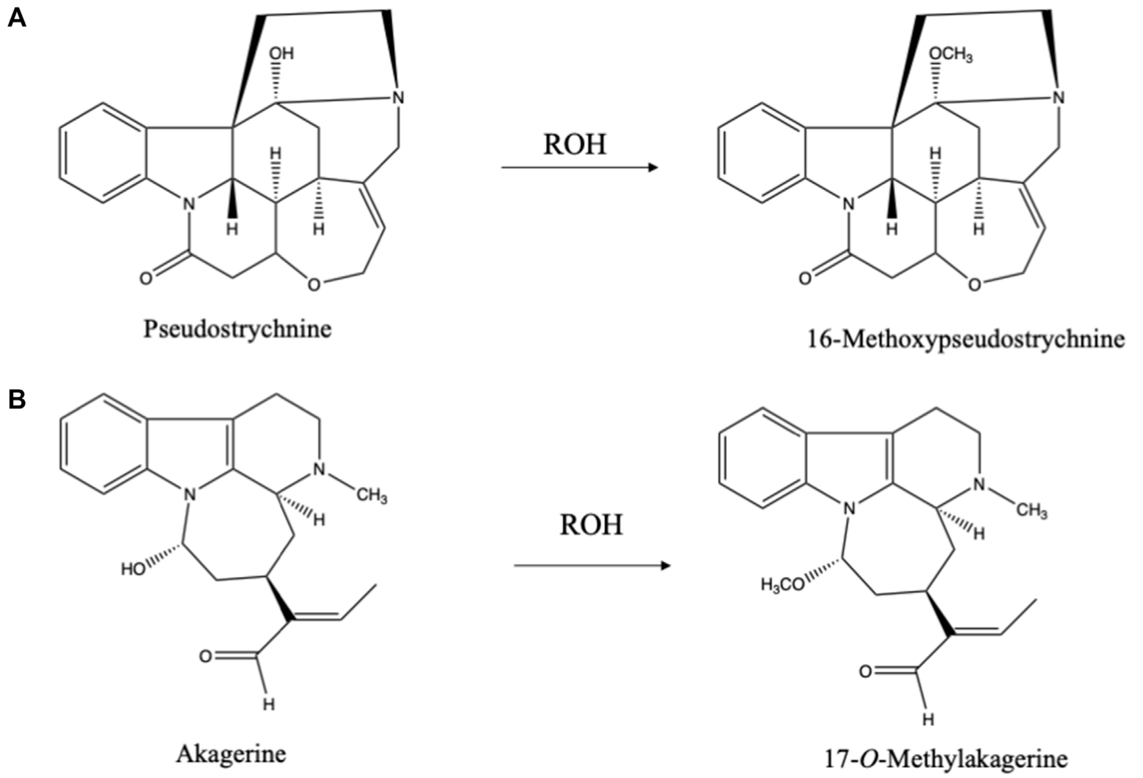


FIGURE 4
(A) Artifacts formed during isolation of pseudostrychnine in alcoholic solvent (Bisset et al., 1965). (B) Artifact formed during isolation of akagerine in alcoholic solvent (Rolfsen et al., 1978).

deuterated methanol, they showed by LC-MS that 8% of the more than 1,100 detected metabolites were artifacts containing a deuterated OMe group.

Another reaction concerns aldehyde- and keto-groups. They may react with alcohols to yield hemiacetals and acetals. In chromatography a single pure compound may show several peaks due to these reactions. Secologanin is an example of such a compound (Figure 2). Both intra- and intermolecular reactions may be involved. Acetone may even form adducts with secologanin (Verpoorte unpublished results, Tomassini et al., 1995). Aldehyde- and keto-groups maybe involved in all kinds of internal rearrangements, like in the biosynthesis of terpenoid indole alkaloids, in which strictosidine is the precursor for a large number of pathways leading to different skeletons (Verpoorte, 2000) (Figure 3). The first reaction is the loss of a glucose. This leads to the opening of the acetal containing ring, in which the molecule unfolds to give two reactive aldehyde functions and two reactive amino groups. This opens the biosynthetic pathways leading to different structures. The reactions of an aldehyde- or keto-group with an amine or an hydroxyl group are important reactions to keep in mind as they are a major source for artifacts.

The alkaloid gentianine is an example of a non-natural alkaloid that is formed during extraction with an ammonia containing extraction solvent. The ammonia may react with the aldehyde group in the iridoid sweroside, yielding gentianine (Phillipson et al., 1974; Popov et al., 1988). Bunel and coworkers (2014) showed that 2-hydroxy-4-methoxybenzaldehyde reacts with ammonia to give an alkaloid like compound. Wenkert and co-workers (1965) reported the artifact formation of an abietane-type of diterpene when ammonia was used in the extraction.

The hydroxyl group in carbinolamines easily reacts with alcohols (just like hemiacetals) yielding an *O*-Methyl derivative in case of methanol as extraction solvent. 16-Methoxypseudostrychnine (Bisset et al., 1965) (Figure 4A) and 17-*O*-methylakagerine (Rolfsen et al., 1978) (Figure 4B) are examples of such artifacts. In this connection the use of ethanol as extraction solvent has advantages. First of all, it is a greener solvent than methanol and is less toxic. Moreover, in case of any reaction with an alcohol during the extraction one will find an ethoxy group instead of a methoxy group. As ethoxy groups are rare in nature, this is an excellent method for for identifying potential artifacts.

A keto-group containing solvent like acetone, may form adducts with ammonia or amines that give alkaloid-positive color reactions, particularly when running preparative column

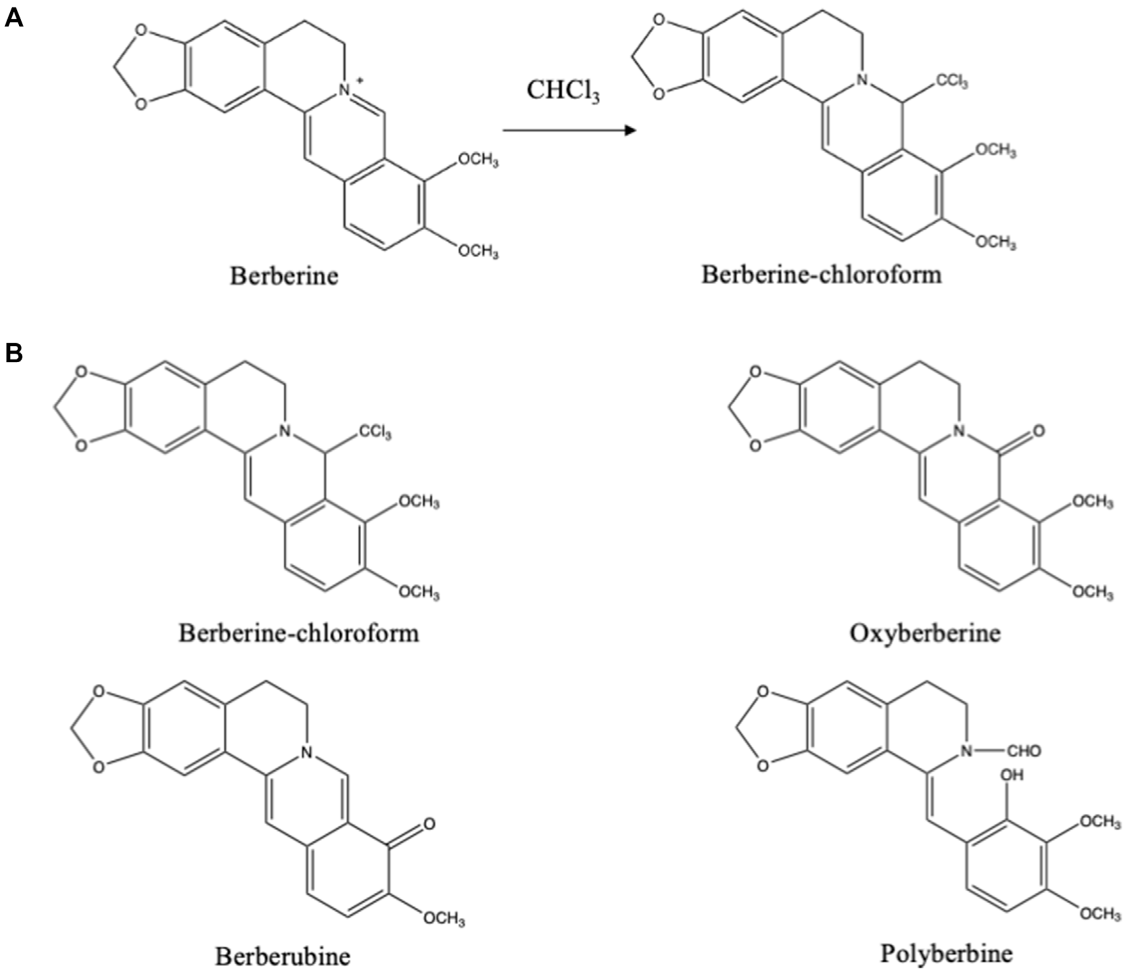


FIGURE 5
(A) Artifact formed from berberine during its isolation using chloroform (Miana, 1973). **(B)** Artifacts derived from berberine formed during column chromatography using chloroform:methanol (99:1) as eluting system (Shamma and Rahimizadeh, 1986).

chromatography on silica (Householder and Camp 1965). Alcohols may contain aldehydes and carboxylic acids as contaminations. One should keep in mind that commercial chloroform always contains 1–2% of ethanol. That means that the above-mentioned reactions also occur in chloroform solution (see below). In general, the experience is that dissolved in alcohols compounds are reasonably stable.

Ethers Though ethers are rather inert in terms of reactivity if compared to alcohols, their major problem is the formation of peroxides. In the use of ethers, great caution is required when evaporating ethers because of high risks of explosions. These peroxides mediate artifact formation, as complex natural products can be oxidized by these peroxides. The most common one is the well-known *N*-oxidation of amines. The *N*-oxides formed may further react and cause ring openings, exemplified by the case of strychnine where the *N*-oxide rearranges into the hydroxy derivative (pseudostrychnine) (Figure 4A) (Bisset et al., 1965). Via this carbinolamine a ring can be opened. N-oxidation is a common step in the catabolism of alkaloids and nitrogen containing medicines. When choosing for diethyl ether as solvent, one should always check for the presence of peroxides.

Esters The intramolecular *trans*-esterifications in chlorogenic acid and their analogues were mentioned above. Using esters (e.g. ethyl acetate) as solvent incurs the risk of *trans*-esterifications. The combination of ammonia and ethyl acetate may lead to crystallization of acetamide.

Halogenated solvents For the extractions of medium polar compounds and for liquid-liquid purifications, halogenated solvents such as chloroform and dichloromethane are often used. For toxicity reasons dichloromethane is recommended to be used instead of the more toxic chloroform. Because of

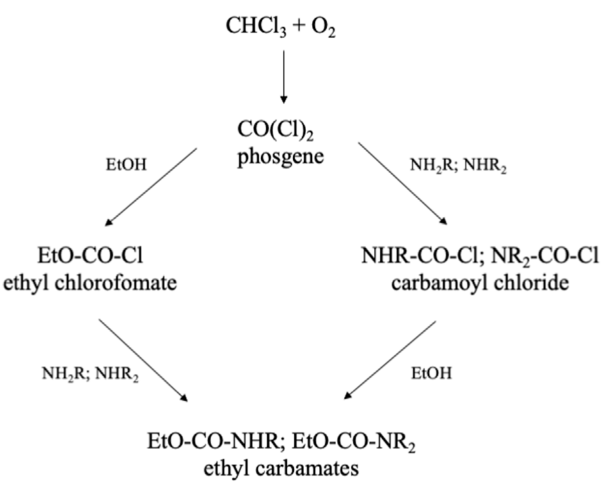


FIGURE 6
Formation of ethyl chloroformate in chloroform and possible resulting adducts (Siek et al., 1977; Cone et al., 1982; Maudens et al., 2007).

physico-chemical properties chloroform has some advantages in better dissolving medium polar compounds, and in particular alkaloids. In terms of artifacts both have serious disadvantages (Baerheim Svendsen and Verpoorte 1983). Besselièvre and coworkers (1972) asked the question "is dichloromethane a solvent or a reagent". They found that the indole alkaloid tubotaiwine is rapidly converted to the quaternary dichloromethotubotaiwine when dissolved in dichloromethane. Strychnine, brucine (Phillipson and Bisset 1972; Verpoorte et al., 2008) and atropine (Vincze and Geven, 1978) have been reported to also give such quaternization of an amine function with dichloromethane. These quaternary dichlorometho alkaloids have lost the lipophilic properties of the tertiary alkaloids. Also, with chloroform these artifacts were formed, though at lower levels as they are formed with dichloromethane and dichlorobromomethane present in chloroform as contaminations (Phillipson and Bisset 1972). Hansen (1977) reported on the artifact formation through *N*-alkylation of amines. In case of strychnine and brucine, in addition to dichlorometho artifacts and N-alkylation, also *N*-oxides and the pseudo-strychnine and -brucine were formed in the chlorinated solvents

In dichloromethane cyanogen chloride (CNCl) might be present in variable quantities as contamination (Franklin, et al., 1978). Primary and secondary amines may form nitriles with this impurity. In chloroform, however, no CNCl could be detected. Chloroform itself reacts with protoberberine type of alkaloids. Especially, during column chromatography using chloroform-methanol as mobile phase, a trichloro compound was formed from berberine (Figure 5A) and related alkaloids (Miana, 1973). Through oxidation other artifacts were formed from the these alkaloids (Figure 5B) (Shamma and Rahimizadeh, 1986).

Another problem with chloroform is its oxidation in the light, yielding phosgene, a well-known chemical warfare gas. This highly reactive gas reacts with all kinds of compounds. To neutralize phosgene, chloroform always contains 0.5–2% of ethanol. Ethanol reacts with phosgene (Figure 6), and thus keeps the level of phosgene low, but still there will be some artifacts formed with analytes. In the analysis of normeperidine the ethyl chloroformate derivative of the target compound was detected (Siek et al., 1977). Cone and colleagues (1982) reported artifact formation in the analysis of metabolites of codeine, when chloroform was used to extract the alkaloids from biological fluids. A similar study was published for the extraction of anthracyclines (Maudens et al., 2007). The presence of ethanol in chloroform may also be the cause of artifacts as described above for alcohols. By distillation chloroform can be purified, but always some alcohol should be added after distillation.

5 Chromatography related problems

5a Identification of compounds

Pimms et al. (1995) made an estimation of the number of organisms on earth. The estimation was between 10 and 100 million organisms, among which 250.000 plant species. A search made at the end of last century in the NAPRALERT database (NAtural PRoducts ALERT, focused on natural products and their bioactivities) showed that of total plant biodiversity, around 15% of the species had been studied to some extent for secondary metabolites and only about 5% for one or a few biological activities (Verpoorte 1998, 2000; Verpoorte et al., 2006). In the Dictionary of Natural Products (2022) the present number of compounds is 328,000. If we assume that every species can make one unique compound, there must be millions of yet unknown compounds present in nature. Based on the number of genes in a plant we estimate that a plant may contain 20,000 to 50,000 different compounds with a very broad range of polarities and with a huge dynamic range.

With the high resolution of the state-of-the-art hyphenated metabolomics methods, we expect many new compounds to be reported in the coming years. Putative identifications of compounds are made through searching various databases with information about retention behavior, molecular weight, and MS-fragmentation patterns. Tools like the recently developed molecular networking are helpful in identification of known compounds, as well as in predicting the chemical structures of novel compounds (Aron et al., 2020). However, for proper identification of a compound and for structure elucidation of novel compounds, the mentioned information is not sufficient for a full identification. A complete set of spectroscopic data is needed to confirm the chemical

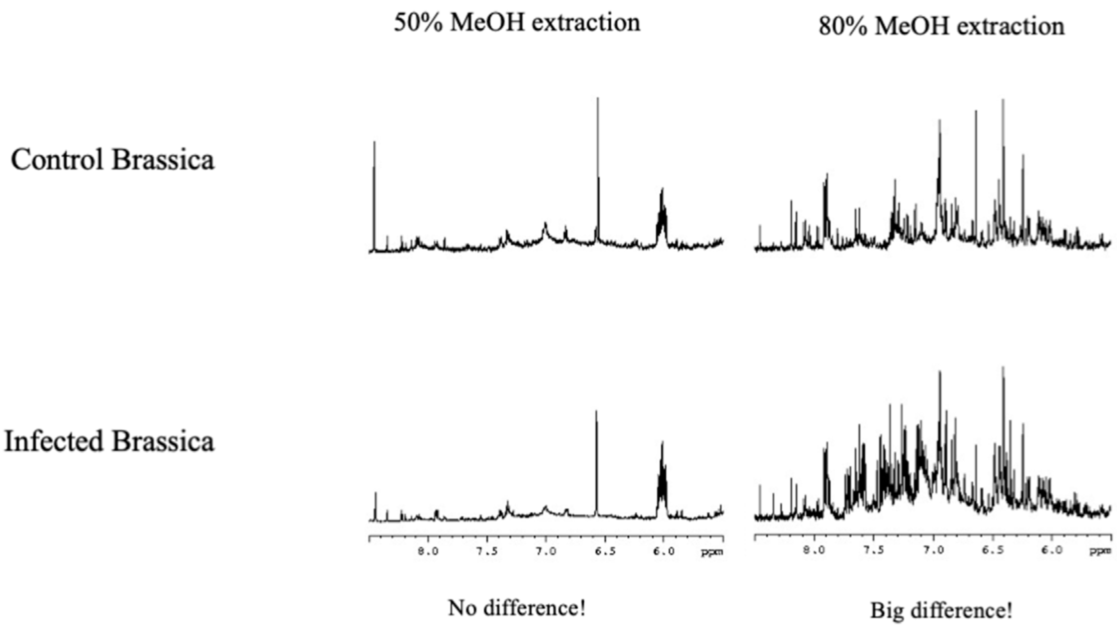


FIGURE 7
Difference between NMR metabolomics data depending on the extraction solvent. Control *Brassica nigra* and infected *Brassica nigra* leaves were extracted using two different solvents containing 50% MeOH **(A)** and 80% MeOH **(B)**. There was less differences between control and infected leaves when extracted with 50% MeOH. However when it is extracted using 80% MeOH, a big difference could be found in both extracts, showing the choice of extraction solvent is important (Verpoorte et al., 2007).

structures and the stereochemistry. Unfortunately, in literature one may find publications with supplementary data with a long list of putative "identified" compounds, based on retention time, molecular weight, MS-fragmentation and comparison with existing databases. These "identifications" are later often cited in other papers as identified compounds, without further new evidence for the identification. One may consider such identifications also as artifacts generated by the automated analytical methods. To avoid doubt about identifications, recommendations have been made for the level of confidence of an identification from metabolomics data (Blaženovic et al., 2018). Whether it is acceptable to say that a component is with 90% confidence compound X, or similar "statistical" support, is in our view doubtful. At least for any marker molecule identified through metabolomics analysis, one should have hard spectral evidence for the identity.

5b Injection samples

Extracts must be dissolved at a certain point in the preanalytical processing for further separation or analysis. There are numerous solvents to extract or to redissolve extracts or molecules obtained from a biological sample. In the different analytical procedures, different procedures are needed. In the Liquid chromatography (TLC and HPLC) the extract must be completely dissolved in a proper solvent. For TLC the choice is based on the solubility of the extract, and on the ease of applying the sample on the plate, where after application the solvent has to be evaporated, before the development of the TLC plate. In case of HPLC the solvent of the sample for injection in the LC-system must be compatible with the column and the eluent. For example, in developing a protocol for a metabolomic analysis, the final solvent used to inject a sample in LC should be as similar as possible to the mobile phase in terms of polarity and pH, including the presence of compounds that may affect retention behavior (e.g. ion-pairing). Ion-pairing LC is used in the analysis of alkaloids, e.g. using long chain sulfonic acids as additive in the mobile phase. But also, ions like acetate, formate, trifluoroacetate, chloride, bromide, and iodide may act as ion-pairing agents, and in liquid-liquid extractions they may cause loss of analytes (Hermans and Verpoorte 1986). In case of large differences between injection solvent and mobile phase the retention of some analytes may be affected. Even the formation of multiple peaks for a single molecule may occur in case of a large difference. In the analysis of complex chromatograms this may easily be overlooked, resulting in errors in quantitation and identification. In case of GC the solvent for dissolving an extract should enable the derivatization required to volatilize the various components present in the metabolome. In case of NMR the NMR-solvent can be used to extract the biomaterial, the reduction in workload is a major advantage of NMR. In case of mass spectrometry solid extracts or biomaterials can be used when the equipment offers this option. The analytical method applied determines the necessary preanalytical procedures. Obviously there will be differences between the various analytical procedures for risks of artifacts formation.

Metabolomics is used to identify markers for certain conditions, by comparing metabolomes of different materials, and identifying what signals correspond with what condition. Using a polar solvent to extract biomaterials means that certain compounds are poorly dissolved. Consequently, the signals of these compounds represent the peak height of a saturated solution of those compounds and will be similar for all samples. By only focusing on differences between metabolomes one might miss markers that are poorly soluble. From the fact that certain signals not seem to change, no conclusions can be drawn. Changing the extraction solvent may have a great impact on the visible metabolome and new markers might become visible (see Figure 7) (Verpoorte et al., 2007).

Another example of this problem is in the extraction of a given weight of sample with different amounts of solvent, e.g., 2 ml or 10 ml solvent. The total amount of a poorly dissolved compound differs a factor 5 between these extracts. When these extracts are taken to dryness and then redissolved in a well-defined amount of another solvent in which the compound is very well soluble, you will find a large difference for the amount detected in the two samples, though they could be similar quantities in the extracted materials. In liquid chromatography the choice of mobile phase is crucial for the separation. The pH is an important factor to keep in mind, as spectral data may be quite different for a compound when measured at different pH. An example is magnoflorine, a quaternary alkaloid. For many years there where two alkaloids mentioned in the literature, *N,N*-dimethyllindcarpine and magnoflorine, but finally it turned out that is was one and the same compound, with quite different spectral data if measured at high or low pH (Stermitz et al., 1980). Schripsema and colleagues (1986) used the pH effect as a tool for structure elucidation, as with trifluoroacetic acid one could deconvolute NMR spectra with a lot of overlapping signals, because the pH strongly affected the shift of protons close to nitrogen atoms. In identification of compounds by spectral data of LC-MS or GC-MS one should keep in mind that the pH of the injected sample may greatly affect retention and MS-fragmentation.

Finally, one other experience we want to share is about salicylic acid. An interesting compound as it is a signal compound that affects the metabolome of plants. We noted in literature that reported recoveries of salicylic acid varied from 30–60%. Un unacceptable difference, that will invalidate any conclusions when measuring this compound. We studied this in some detail and found that the problem is that salicylic acid is volatile. When taking an extract to full dryness, it disappears completely (Verberne et al., 2002). By adding a small amount of sodium hydroxide the evaporation was avoided, and high reproducible recovery was obtained.

6 Conclusion

Metabolomes of biological materials are complex, because of the large number of compounds with a wide range of polarities and concentrations. In preparing samples for metabolomic analysis extraction with organic solvents is a common step. These solvents may interact with various analytes through chemical reactions. Also contaminations in the solvents may be involved in the formation of artifacts. Particularly hydroxyl-, aldehyde-, keto-, carboxyl-, ester-, and amine functional groups are involved in the artifact formation. Oxidation, esterification, hydrolysis, glycolysis are common reactions that may occur in the preanalytical steps of the sample preparation. Considering the problems with some of the classic organic solvents, in terms of artifacts formation, their toxicity and their ecological damage, future research should be focused on developing novel green solvents for analytical chemistry, like the use of ionic liquids or natural deep eutectic solvents (Dai et al., 2013). In the LC analytical steps it is differences in injection solvent and mobile phase that are sources of artifacts, like distorted peaks or even double peak formation. Saturated solutions of poorly soluble compounds may hide markers. Finally the proper identification of compounds is a major hurdle, as it requires the full set of spectral data (UV, IR, MS, NMR), and methods for proving the full stereochemistry. Identification on the basis of UV, MS and retention is not sufficient. The development of a metabolomics analysis protocol should include a proper validation. For reproducible results the quality of all used chemicals and solvents should be controlled. For future reference, registration of the metadata from all steps of the protocol from collection to final chemometric analysis is essential.

Author contributions

RV wrote the paper HK contributed with examples, created figures YC contributed with examples.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer OM declared a shared affiliation with the authors to the handling editor at the time of review.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aron, A. T., Gentry, E. C., McPhail, K. L., Nothias, L. F., Nothias-Esposito, M., Bouslimani, A., et al. (2020). Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat. Protoc.* 15, 1954–1991. doi:10.1038/s41596-020-0317-5

Baerheim Svendsen, A., and Verpoorte, R. (1983). *Chromatography of alkaloids, part I, thin-layer chromatography*, Vol. 23A. Amsterdam: Chromatography LibraryElsevier, 534.

Besselièvre, R., Langlois, N., and Potier, P. (1972). Chlorure de methylene, solvant ou reactif? *Bull. Soc. Chim. Fr.* 4, 1477–1478.

Bisset, N. G., Casinovi, C. G., Galeffi, C., and Marini Bettolo, G. B. (1965). On some 16-alkoxy-strychnines. *Ric. Sci. 2. Ser. Pt. 2. Rend. B* 35, 273–274.

Blaženovic, I., Kind, T., Ji, J., and Fiehn, O. (2018). Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites* 8, 31. doi:10.3390/metabo8020031

Brondz, I., Ekeberg, D., Hoiland, K., Bell, D. S., and Annino, A. R. (2007). The real nature of the indole alkaloids in *Corinarius infractus*: Evaluation of artifact formation through solvent extraction method development. *J. Chromatogr. A* 1148, 1–7. doi:10.1016/j.chroma.2007.02.074

Bunel, V., Hamel, M., Duez, P., and Stevigny, C. (2014). Artifactual generation of an alkaloid in the course of *Mondia whitei* (Hook.f.) skeels roots extraction: A clue to endogenous-formed bioactive compounds? *Phytochem. Lett.* 10, 101–106. doi:10.1016/j.phytol.2014.08.012

Capon, R. J. (2020). Extracting value: Mechanistic insights into the formation of natural product artifacts – case studies in marine natural products. *Nat. Prod. Rep.* 37, 55–79. doi:10.1039/c9np00013e

Clifford, M. N., Kellard, B., and Birch, G. G. (1989). Characterisation of chlorogenic acids by simultaneous isomerisation and transesterification with tetramethylammonium hydroxide. *Food Chem. x.* 33, 115–123. doi:10.1016/0308-8146(89)90114-3

Cone, E. J., Buchwald, W. F., and Darwin, W. D. (1982). Analytical controls in drug metabolic studies. II. Artifact formation during chloroform extraction of drugs and metabolites with amine substituents. *Drug Metab. Dispos.* 10, 561–567.

Dai, Y., van Spronsen, J., Witkamp, G. J., Verpoorte, R., and Choi, Y. H. (2013). Ionic liquids and deep eutectic solvents in natural products research: Mixtures of solids as extraction solvents. *J. Nat. Prod* 76, 2162–2173. doi:10.1021/np400051w

DNP (2022). *Dictionary of natural products*. http://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml;jsessionid=DB01289ACAA79C222859E1CD8A98A894 (Accessed January 23, 2022).

Franklin, R. A., Heatherington, K., Morrison, B. J., Sherron, P., and Ward, T. J. (1978). Communication. Detection of cyanogen chloride as an impurity in dichloromethane and its significance in drug and metabolite analysis. *Analyst* 103, 660–662. doi:10.1039/an9780300660

Hansen, S. H. (1977). Analytical aspects of the N-alkylating properties of chloroform, dichloromethane and 1, 2-dichloroethane. *Arch. Pharm. Chem.* 5, 194–200.

Hanson, J. R. (2017). Pseudo-natural products, some artefacts formed during the isolation of terpenoids. *J. Chem. Res.* 41, 497–503. doi:10.3184/174751917x15021050367558

Hanson, K. R. (1965). Chlorogenic acid biosynthesis. Chemical synthesis and properties of the mono-O-cinnamoylquinic acids. *Biochemistry* 4, 2719–2731. doi:10.1021/bi00888a023

Hemingway, S. R., Phillipson, J. D., and Verpoorte, R. (1981). *Meconopsis cambrica* alkaloids. *J. Nat. Prod. (Gorakhpur).* 44, 67–74. doi:10.1021/np50013a012

Hermans-Lokkerbol, A., and Verpoorte, R. (1986). Droplet counter-current chromatography of alkaloids. The influence of ph-gradients and ion-pair formation on the retention of alkaloids. *Planta Med.* 52, 299–302. doi:10.1055/s-2007-969158

Housholder, D. E., and Camp, B. J. (1965). Formation of alkaloid artifacts in plant extracts by the use of ammonium hydroxide and acetone. *J. Pharm. Sci.* 54, 1676–1677. doi:10.1002/jps.2600541128

Johnson, A. R., Fogerty, A. C., Hood, R. L., Kozuharov, S., and Ford, G. L. (1976). Gas-liquid chromatography of ethyl ester artifacts formed during the preparation of fatty acid methyl esters. *J. Lipid Res.* 17, 431–432. doi:10.1016/s0022-2275(20)34930-0

Lough, A. K., Felinski, L., and Garto, G. A. (1962). The production of methyl esters of fatty acids as artifacts during the extraction or storage of tissue lipids in the presence of methanol. *J. Lipid Res.* 3, 478–480. doi:10.1016/s0022-2275(20)40396-7

Maltese, F., Erkelens, C., van der Kooy, F., Choi, Y. H., and Verpoorte, R. (2009). Identification of natural epimeric flavanone glycosides by NMR spectroscopy. *Food Chem. x.* 116, 575–579. doi:10.1016/j.foodchem.2009.03.023

Maltese, F., van der Kooy, F., and Verpoorte, R. (2009). Solvent derived artifacts in natural products chemistry. *Nat. Prod. Commun.* 4, 1934578X0900400–454. doi:10.1177/1934578x0900400326

Maudens, K. E., Will, S. M. R., and Lambert, W. E. (2007). Traces of phosgene in chloroform: Consequences for extraction of anthracyclines. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* 848, 384–390. doi:10.1016/j.jchromb.2006.10.073

May, R. M. (1988). How many species are there on Earth? *Science* 241, 1441–1449. doi:10.1126/science.241.4872.1441

Miana, G. A. (1973). Tertiary dihydroprotoberberine alkaloids of Berberis lycium. *Phytochemistry* 12, 1822–1823. doi:10.1016/0031-9422(73)80415-7

Middleditch, B. S. (1989). *HPLC, TLC and PC journal of chromatography library*, 44. Amsterdam: Elsevier.Analytical artifacts. GC, MS

NAPRALERT (2022). https://pharmacognosy.pharmacy.uic.edu/napralert/ assessed 17-07-2022.

Paton, A. J. (2013). From working list to online flora of all known plants: Looking forward with hindsight 1. *Ann. Mo. Bot. Gard.* 99, 206–213. doi:10.3417/2011115

Phillipson, J. D., and Bisset, N. G. (1972). Quaternisation and oxidation of strychnine and brucine during plant extraction. *Phytochemistry* 11, 2547–2553. doi:10.1016/s0031-9422(00)88534-9

Phillipson, J. D., Hemingway, S. R., Bisset, N. G., Houghton, P. J., and Shellard, E. J. (1974). Angustine and related alkaloids from species of *mitragyna, nauclea, uncaria*, and *strychnos*. *Phytochemistry* 13, 973–978. doi:10.1016/s0031-9422(00)91432-8

Pimm, S. L., Gareth, G. J., Gittleman, J. L., and Brooks, T. M. (1995). The future of biodiversity. *Science* 269, 347–350. doi:10.1126/science.269.5222.347

Popov, S. S., Marekov, N. L., and Do, T. N. (1988). *In vitro* transformations of gentiopicroside and swertiamarin. *J. Nat. Prod.* 4, 765–768. doi:10.1021/np50058a018

Rolfsen, W., Bohlin, L., Yeboah, S. K., Geevaratne, M., and Verpoorte, R. (1978). New indole alkaloids of *Strychnos dale* and *Strychnos elaeocarpa*. *Planta Med.* 34, 264–273. doi:10.1055/s-0028-1097449

Sauerschnig, S., Doppler, M., Bueschl, C., and Schuhmacher, R. (2018). Methanol generates numerous artifacts during sample extraction and storage of extracts in metabolomics research. *Metabolites* 8, 1. doi:10.3390/metabo8010001

Schripsema, J., Verpoorte, R., and Baerheim Svendsen, A. (1986). Trifluoroacetic acid, a ^{1}H NMR shift reagent for alkaloids. *Tetrahedron Lett.* 27, 2523–2526. doi:10.1016/s0040-4039(00)84574-8

Shamma, M., and Rahimizadeh, M. (1986). The identity of chileninone with berberrubine. The problem of true natural products vs. artifacts of isolation. *J. Nat. Prod. (Gorakhpur).* 49, 398–405. doi:10.1021/np50045a003

Siek, T. J., Eichmeier, L. S., Caplis, M. E., and Esposito, F. E. (1977). The reaction of normeperidine with an impurity in chloroform. *J. Anal. Toxicol.* 1, 211–214. doi:10.1093/jat/1.5.211

Stermitz, F. R., Castedo, L., and Dominguez, D. (1980). Magnoflorine and N, N-dimethyllindcarpine. *J. Nat. Prod. (Gorakhpur).* 43, 140–142. doi:10.1021/np50007a013

Tang, Y., Friesen, J. B., Nikolić, D. S., Lankin, D. C., McAlpine, J. B., Chen, S. N., et al. (2021). Silica gel-mediated oxidation of prenyl motifs generates natural product-like artifacts. *Planta Med.* 87, 998–1007. doi:10.1055/a-1472-6164

Tomassini, L., Cometa, M. F., Serafini, M., and Nicoletti, M. (1995). Isolation of secoiridoid artifacts from *Lonicera japonica*. *J. Nat. Prod. (Gorakhpur).* 58, 1756–1758. doi:10.1021/np50125a020

Venditti, A. (2020). What is and what should never be: Artifacts, improbable phytochemicals, contaminants and natural products. *Nat. Prod. Res.* 34, 1014–1031. doi:10.1080/14786419.2018.1543674

Verberne, M. C., Brouwer, N., Delbianco, F., Linthorst, H. J. M., Bol, J. F., and Verpoorte, R. (2002). Method for the extraction of the volatile compound salicylic acid from tobacco leaf material. *Phytochem. Anal.* 13, 45–50. doi:10.1002/pca.615

Verpoorte, R., and Baerheim Svendsen, A. (1984). *Chromatography of alkaloids, Part II, GLC and HPLC*. Amsterdam: Chromatography Library Elsevier.

Verpoorte, R., Choi, Y. H., and Kim, H. K. (2007). NMR-based metabolomics at work in phytochemistry. *Phytochem. Rev.* 6, 3–14. doi:10.1007/s11101-006-9031-3

Verpoorte, R., Choi, Y. H., Mustafa, N. R., and Kim, H. K. (2008). Metabolomics: Back to basics. *Phytochem. Rev.* 7, 525–537. doi:10.1007/s11101-008-9091-7

Verpoorte, R. (1998). Exploration of nature's chemodiversity: The role of secondary metabolites as leads in drug development. *Drug Discov. Today* 3, 232–238. doi:10.1016/s1359-6446(97)01167-7

Verpoorte, R., Kim, H. K., and Choi, Y. H. (2006). in *Plants as source of medicines: New perspectives*. Editors R. J. Bogers, L. E. Craker, and D. Lange (Dordrecht: Springer), 261–274.

Verpoorte, R. (2000). Pharmacognosy in the new millennium: Leadfinding and biotechnology. *J. Pharm. Pharmacol.* 52, 253–262. doi:10.1211/0022357001773931

Verpoorte, R. (2000). "Plant secondary metabolism," in *Metabolic engineering of plant secondary metabolism*. Editors R. Verpoorte and A. W. Alfermann (Dordrecht: Kluwer Academic Publishers), 1–30.

Verpoorte, R., and Sandberg, F. (1971). Alkaloids of *Strychnos camptoneura*. *Acta Pharm. Suec.* 8, 119–122.

Vincze, A., and Gefen, L. (1978). Solvent caused quaternization as a possible source of error in the mass spectral quantitation of tertiary amines. I. Methylene chloride quaternization. *Isr. J. Chem.* 17, 236–238. doi:10.1002/ijch.197800041

Wenkert, E., Fuchs, A., and McChesney, J. D. (1965). Chemical artifacts from the family Labiatae. *J. Org. Chem.* 30, 2931–2934. doi:10.1021/jo01020a012

Xu, T., Chen, W., Zhou, J., Dai, J., Li, Y., and Zhao, Y. (2020). Virtual screening for reactive natural products and their probable artifacts of solvolysis and oxidation. *Biomolecules* 10, 1486. doi:10.3390/biom10111486

Yang, J. Y., Sanchez, L. M., Rath, C. M., Liu, X., Boudreau, P. D., Bruns, N., et al. (2013). Molecular Networking as a dereplication strategy. *J. Nat. Prod.* 76, 1686–1699. doi:10.1021/np400413s

frontiers | Frontiers in Molecular Biosciences

TYPE Review
PUBLISHED 17 October 2022
DOI 10.3389/fmolb.2022.926623

Interpretable machine learning methods for predictions in systems biology from omics data

David Sidak[1], Jana Schwarzerová[1,2], Wolfram Weckwerth[1,3] and Steffen Waldherr[1]*

[1]Department of Functional and Evolutionary Ecology, Faculty of Life Sciences, Molecular Systems Biology (MOSYS), University of Vienna, Vienna, Austria, [2]Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Brno, Czech Republic, [3]Vienna Metabolomics Center (VIME), Faculty of Life Sciences, University of Vienna, Vienna, Austria

Check for updates

OPEN ACCESS

EDITED BY
Julien Boccard,
Université de Genève, Switzerland

REVIEWED BY
Nathan Lewis,
University of California, United States
Miguel De Figueiredo,
Université de Genève, Switzerland

*CORRESPONDENCE
Steffen Waldherr,
steffen.waldherr@univie.ac.at

SPECIALTY SECTION
This article was submitted to Metabolomics,
a section of the journal
Frontiers in Molecular Biosciences

RECEIVED 22 April 2022
ACCEPTED 15 August 2022
PUBLISHED 17 October 2022

CITATION
Sidak D, Schwarzerová J, Weckwerth W and Waldherr S (2022), Interpretable machine learning methods for predictions in systems biology from omics data.
Front. Mol. Biosci. 9:926623.
doi: 10.3389/fmolb.2022.926623

Machine learning has become a powerful tool for systems biologists, from diagnosing cancer to optimizing kinetic models and predicting the state, growth dynamics, or type of a cell. Potential predictions from complex biological data sets obtained by "omics" experiments seem endless, but are often not the main objective of biological research. Often we want to understand the molecular mechanisms of a disease to develop new therapies, or we need to justify a crucial decision that is derived from a prediction. In order to gain such knowledge from data, machine learning models need to be extended. A recent trend to achieve this is to design "interpretable" models. However, the notions around interpretability are sometimes ambiguous, and a universal recipe for building well-interpretable models is missing. With this work, we want to familiarize systems biologists with the concept of model interpretability in machine learning. We consider data sets, data preparation, machine learning methods, and software tools relevant to omics research in systems biology. Finally, we try to answer the question: "What is interpretability?" We introduce views from the interpretable machine learning community and propose a scheme for categorizing studies on omics data. We then apply these tools to review and categorize recent studies where predictive machine learning models have been constructed from non-sequential omics data.

KEYWORDS

multi-omics, interpretable machine learning, deep learning, explainable artificial intelligence, metabolomics, proteomics, transcriptomics

1 Introduction

Machine learning (ML) is advancing rapidly, with new methods introduced almost daily. As the field progresses, also its methods become better accessible to researchers from other disciplines due to the development and release of new software tools. Many fundamental ML methods can be applied to almost any data set. Nonetheless, the real-world goals of researchers that apply these methods to their own data sets may diverge from the objectives of the ML model itself (Lipton, 2016). While a researcher may want to understand the molecular mechanisms of a disease or may want to know why a ML model classifies a patient as having a disease, the ML model may aim to minimize the number of wrong predictions. Understanding predictions is especially important in a clinical context, where medical professionals need to justify healthcare decisions (Barredo Arrieta et al., 2020). Bringing real-world and ML objectives into harmony asks for methods that make ML models more interpretable (Lipton, 2016). The research field behind this goal is *interpretable machine learning* (Murdoch et al., 2019), which falls under the umbrella of *explainable artificial intelligence (XAI)* (Barredo Arrieta et al., 2020). Advances in this domain are becoming even more important as ML models are increasing in complexity. Further, using data-driven approaches like machine learning to not just predict from data but also to learn about the biological mechanisms that generate the data in the first place is an attractive concept. Mechanistic approaches like kinetic models take long to develop and require a detailed prior understanding of a system, while machine learning models can make better predictions and sometimes answer the same biological questions with less effort (Costello and Martin, 2018).

Consequently, interpretable ML has received more and more attention in biology in recent years. Various studies that apply machine learning to biological data sets have been published, many claiming to implement "interpretable" (Wang et al., 2020; Oh et al., 2021; Sha et al., 2021), "explainable" (Manica et al., 2019), "gray-box" (Nguyen et al., 2021), "white-box" (Yang et al., 2019) or "visible" (Ma et al., 2018) machine learning frameworks. All these terms refer to the urge to gain valuable biological knowledge from data with the help of machine learning, which falls under the keyword "interpretability" (Lipton, 2016; Murdoch et al., 2019). Now, the question arises, what is interpretability?, or, more specifically, what makes a machine learning model interpretable? The answer to this fundamental question is under debate in the machine learning community for some time now. Many answers have been proposed (Lipton, 2016; Murdoch et al., 2019; Barredo Arrieta et al., 2020), but a clear consensus is still missing. Generally, "interpretability [itself] is a broad, poorly defined concept (Murdoch et al., 2019)," which is probably the main reason why definitions in a machine learning context are complicated to fix. Clearly, there are different perspectives to view interpretability in machine learning: e.g., it can mean how much we can learn from data by using a ML model (Murdoch et al., 2019), how well we understand the ML model itself (i.e., comprehend how it makes a prediction), or how much extra information the model can provide that supports predictions (Lipton, 2016). *Interpretation methods*, the techniques by which we gain biological insight from data with machine learning besides predictions, may divide into "model-based" and "post hoc" methods (Murdoch et al., 2019). While model-based methods rely on adapting the model before training it, post-hoc methods operate on already trained models (Murdoch et al., 2019).

In machine learning, there are three main ways to train models, namely reinforcement learning, unsupervised learning, and supervised learning. Throughout this review, we want to focus on *supervised learning* because of its prevalence in general (LeCun et al., 2015) and in the context of predictive systems biology. Supervised learning presents models with a set of training *samples* (e.g., omics profiles from multiple patients) for which the outcome of a prediction (e.g., health conditions) is already known (Presnell and Alper, 2019). Based on this training data set, supervised learning tries to produce a model that accurately predicts the target for samples without a known solution (Angermueller et al., 2016). Supervised machine learning techniques have been applied to high-throughput omics data to predict a broad range of clinical, phenotypical, and physiological observations.

While diagnosing various diseases (Leitner et al., 2017; Trainor et al., 2017; Hu et al., 2018; Pai et al., 2019; Stamate et al., 2019; Nguyen et al., 2021; Sha et al., 2021; van Dooijeweert et al., 2021) or predicting clinical outcomes (Bahado-Singh et al., 2019; Pai et al., 2019; Zhang et al., 2021) seem common, possible applications reach up to inference of the fluxome (Alghamdi et al., 2021) or growth rate (Culley et al., 2020) of a cell from transcript levels. Besides using machine learning for predictions, many studies attempt to gain additional biological knowledge by implementing post-hoc or model-based interpretation methods (Alakwaa et al., 2018; Date and Kikuchi, 2018; Hu et al., 2018; Bahado-Singh et al., 2019; Wang et al., 2020; Nguyen et al., 2021; Wang et al., 2021). Further, interpretability can improve by incorporating prior biological knowledge into a research project (Nguyen et al., 2021; Wang et al., 2021).

This review was written from an interdisciplinary perspective and is intended for an audience with systems biological background but not necessarily experience in machine learning, who are interested in machine learning approaches for generating biological insight. We aim to familiarize readers with the term interpretability and equip them with a fundamental machine learning background necessary for understanding the concept. To achieve this, we take an example-based approach by highlighting studies that successfully extract biological insight from non-sequential omics data sets with the help of interpretation methods.

Furthermore, we present a scheme for categorizing research papers based on two criteria, 1) the use of interpretation methods and 2) at which point prior knowledge enters a research project. With this categorization system, we hope to contribute to the establishment of terms associated with interpretability and allow ML projects to be compared in their interpretability. In this work, we have assigned a total of 26 publications to 9 categories that our scheme outlines.

We start with a characterization of the utilized data sets, what studies predict from them, and how to prepare them for machine learning. Then we present supervised learning methods that systems biologists applied to omics data and showcase available software tools for data manipulation, visualization, and up to fully automatic ML solutions for omics data analysis. We try to answer the question: "What is interpretability?" by introducing fundamental concepts, describing our categorization scheme, and highlighting exemplary works in systems biology. With this work, we want to raise awareness for interpretable machine learning and its potential for gaining insight from omics data.

2 Data sets

Due to the *data-driven* nature of machine learning, data is essential for a successful ML project (Mendez et al., 2019). Ultimately, any machine learning model tries to learn discriminative features, relationships, patterns, or structures found within a data set. In a data set for supervised learning, a sample consists of variables that describe its properties (the *features*, e.g., molecule abundances) and has one or more outcome variables associated with it that provide the corresponding prediction target (the *labels*) (Shalev-Shwartz and Ben-David, 2013; Angermueller et al., 2016; Deisenroth et al., 2020). Labels can be any variables we wish to predict, ranging from categorical variables describing cancer (sub)types (Alakwaa et al., 2018; Sharma et al., 2019; Zhang et al., 2021) to continuous specifications of cell growth (Kim et al., 2016; Culley et al., 2020). Based on whether labels are categorical or quantitative variables, one differentiates between the two supervised prediction tasks, *classification* or *regression* (Bishop, 2006, p. 3). A feature can be any variable we expect to be predictive of a target variable, such as metabolite abundances (Trainor et al., 2017; Stamate et al., 2019; Sha et al., 2021), "traditional risk factors" (Liu et al., 2017), metabolic fluxes (Culley et al., 2020), and even kinetic parameters when the goal is to predict the feasibility of kinetic models (Andreozzi et al., 2016). Samples with known labels provide the "ground truth" enabling the ML model to learn how predictions for unlabeled samples should optimally look like (Martorell-Marugán et al., 2019).

Usually, the data set that holds all collected and labeled samples is divided into at least a *training set* and an independent *test set* (Trainor et al., 2017; Alakwaa et al., 2018; Sharma et al., 2019; Culley et al., 2020; van Dooijeweert et al., 2021). A *learning algorithm* uses the training set to improve/construct a ML model (Bousquet and Elisseeff, 2002), e.g., by estimating parameters or functional forms. Since the model is fit to the training data, the model's error on this data can be drastically smaller on unseen data like the test set (Maceachern and Forkert, 2021), which means that the model struggles on new samples drawn from the same underlying distribution, i.e., the model has a poor "generalization" ability (Shalev-Shwartz and Ben-David, 2013, sect. 1.1). This phenomenon is known as *overfitting*. Guiding high-level modeling decisions (i.e., *hyperparameters* like the number of layers in a neural network) with the test set can similarly overfit the model to this data (Bishop, 2006, p. 32). It is, therefore, required to use a third separate *validation set* (Angermueller et al., 2016) or, if samples are rare, use other techniques like cross-validation that avoid using the test set for such optimization purposes (Bishop, 2006, p. 32f). After tuning the design and training, a model's realistic *performance*, i.e., "predictive accuracy" (Murdoch et al., 2019) is measured on the out-of-sample test set (Angermueller et al., 2016).

With omics data sets becoming more readily available, they are also more frequently exposed to machine learning algorithms. Alone in this review, the categorized studies covered eight distinct data types characterizing a biological system—not counting network-type data. Omics data sets lend themselves to interpretable machine learning solutions because of their sheer complexity, making them hard to interpret by visual inspection or simple statistical methods. Table 1 provides an overview of the reviewed studies that demonstrates a wide diversity of prediction targets. We compile some of the targets into the categories "Diagnosis," "Clinical Outcome," and "Physiology." Physiology includes phenotypic predictions, genetic properties, cellular state and dynamics, etc. Predictions that did not fit any of these categories were regional origin of an organism (Date and Kikuchi, 2018), type of a cell (Wang et al., 2020; Wang et al., 2021), "feasibility" of kinetic models (Andreozzi et al., 2016), and body region where a tumor emerged (Zhang et al., 2021). The most common category was Diagnosis with 16 examples. Among the diagnosed diseases, cancer is most prevalent. One reason is the commendable availability of large omics data sets enabled by The Cancer Genome Atlas (TCGA) program. Unarguably, precision medicine, especially cancer research and diagnostics has benefited a lot from machine learning in recent years (Grapov et al., 2018; Chiu et al., 2020). Another trend that seems to arise is the application of machine learning to problems that have been traditionally solved with mechanistic models, like the estimation of metabolic fluxes (Alghamdi et al., 2021) and metabolite changes over time (Costello and Martin, 2018). Phenotypic discrimination is also very apparent. This includes predicting cell growth (Kim et al., 2016; Culley et al., 2020), patient biological sex (Zhang et al., 2021), and organism body size (Asakura et al., 2018). Zhang et al. (2021) demonstrated that even multiple predictions, ranging from cancer type classification and stratification over patient age and sex to patient survival, are possible from the same integrated data source. Building large "multi-task" (Zhang et al., 2021) machine learning frameworks that can predict multiple biological system properties for one sample seem promising as data collections grow and become more well-curated, as exemplified by Kim et al. (2016).

TABLE 1 Overview of the categorized studies.

Omics data type	Prediction Method(s)	Effective raw features	Effective raw samples	Prediction type	Ref
Metabolomics	**Ensemble DNN,** DNN, RF, SVM	106 NMR peaks	502 profiles	Regression (Physiology; fish body size)	Asakura et al. (2018)
Metabolomics	LogReg	24 metabolites	1571 profiles	Binary Classification (Clinical Outcome; prospective type 2 diabetes)	Liu et al. (2017)
Metabolomics	SVM	1737 metabolites	58 profiles	Binary Classification (Diagnosis; Diamond Blackfan Anaemia)	van Dooijeweert et al. (2021)
Metabolomics	**RF**, AdaBoost, SVM, NBC	109 metabolites	12–18[a] profiles	Binary Classification (Physiology; pathway presence in tomato pericarp)	Toubiana et al. (2019)
Metabolomics	DNN, XGBoost (DT), RF	347 metabolites	357 profiles	Binary Classification (Diagnosis; alzheimer-type dementia)	Stamate et al. (2019)
Metabolomics	**LGP**, LogReg	70 metabolites	389 profiles	Binary Classification (Diagnosis; knee osteoarthritis)	Hu et al. (2018)
Metabolomics	**LGP**, SVM, RF	242 metabolites	114–115 profiles	Binary Classification (Diagnosis; alzheimer's disease, amnestic mild cognitive impairment)	Sha et al. (2021)
Metabolomics	**DNN**, PLS-DA, RF, SVM	≤106 NMR peaks	1022 profiles	Binary Classification (Other; regional origin of fish)	Date and Kikuchi (2018)
Metabolomics	PLS-DA, Sparse PLS-DA, RF, SVM, kNN, NBC, ANN	≤1032[b] metabolites	38 profiles	Multi-class Classification (Diagnosis; cardio vascular disease)	Trainor et al. (2017)
Metabolomics	PLS-DA, Sparse PLS-DA, RF, SVM, kNN, NBC, ANN	≤431[b] metabolites	not assigned[a]	Binary Classification (Diagnosis; adenocarcinoma lung cancer)	Trainor et al. (2017)
Metabolomics	PLS-DA, Sparse PLS-DA, RF, SVM, kNN, NBC, ANN	not assigned[a]	not assigned[a]	Multi-class Classification (Physiology; genotype)	Trainor et al. (2017)
Metabolomics	**DNN**, RF, SVM, DT, LDA, NSC, GBM	162 metabolites	271 profiles	Binary Classification (Diagnosis; breast cancer stratification)	Alakwaa et al. (2018)
Metabolomics	SVM, PLS-DA	16 and 131 metabolites	21 and 32 profiles	Binary Classification (Diagnosis; gestational diabetes mellitus)	Leitner et al. (2017)
Proteomics	LDA, SVM, kNN, RF	123 peptides	183 profiles	Multi-class Classification (Physiology; genotypes)	Hoehenwarter et al. (2011)
Transcriptomics	**CNN**, RF, DT, AdaBoost	60483 genes	6216 profiles	Multi-class Classification (Diagnosis; different cancer types)	Sharma et al. (2019)
Transcriptomics	SimNet	≤17814[b] genes	348 profiles	Binary Classification (Diagnosis; breast cancer stratification)	Pai et al. (2019)
Transcriptomics	SimNet	not assigned[a]	194 profiles	Binary Classification (Diagnosis; asthma)	Pai et al. (2019)
Transcriptomics	SVR, RF, DNN, BEMKL, BRF, MMANN	≥68[c] genes	1229 profiles	Regression (Physiology; eukaryotic growth rate)	Culley et al. (2020)
single-cell Transcriptomics	GNN	862 genes	162 single-cell profiles	Multi-class Classification (Other; cell type)[d]	Alghamdi et al. (2021)
single-cell transcriptomics	**CapsNet**, SVM, RF, LDA, kNN, ANN	3346 genes	17933[a] single-cell profiles	Multi-class Classification (Other; cell type)[d]	Wang et al. (2020)
single-cell transcriptomics	CapsNet	9437 genes	4993 profiles	Multi-class Classification (Other; cell type)[d]	Wang et al. (2021)
Epigenomics	**VAE in combination with different ML methods,** RBF SVM, RF, ANN, DNN	438831 DNA methylation sites	3905 profiles	Multi-class Classification (Diagnosis; brain cancer subtypes)	Zhang et al. (2021)

(Continued on following page)

TABLE 1 (*Continued*) Overview of the categorized studies.

Omics data type	Prediction Method(s)	Effective raw features	Effective raw samples	Prediction type	Ref
Multi-omics (DNA copy number, Transcriptomics, Proteomics)	**modified NSC**, SVM, NSC	≤16266[a] proteins, ≤17282[a] genes	103 profiles per omics-type	Multi-class Classification (Diagnosis; breast cancer stratification)	Koh et al. (2019)
Multi-omics (Transcriptomics, Proteomics, microRNA Transcriptomics, DNA methylation, DNA copy number)	SimNet	not assigned[a]	150, 252, 77 and 155 profiles per omics-type in four independent data sets	Binary Classification (Clinical Outcome; cancer patient survival)	Pai et al. (2019)
Multi-omics (mRNA Transcriptomics, microRNA Transcriptomics, Epigenomics)	**VAE in combination with different ML methods**, RBF SVM(R), RF(R), ANN(R), DNN(R), CoxPH	58043 genes, 438831 DNA methylation sites, 1881 miRNAs	9736–11538 profiles per omics-type	Multi-class Classification (Diagnosis; different cancer types), Regression (Physiology; patient age), Binary Classification (Physiology; patient biological sex), Multi-class Classification (Physiology; tumour stage, Other; body region of tumor emergence), Regression (Clinical Outcome; patient survival function)	Zhang et al. (2021)
Multi-omics (Proteomics, Metabolomics)	SVM, GLM, NSC, RF, LDA, DNN	≤141[b] metabolites, ≤ 27[a] proteins	26 profiles per omics-type	Binary Classification (Clinical Outcome; perinatal outcome in asymptomatic women with short cervix)	Bahado-Singh et al. (2019)
Multi-omics (Transcriptomics, SNP-omics (genetic variants))	**DNN with Lasso**, DSPN, AdaBoost, DT, SVM, ANN, RF, kNN, GP, NBC, RBM, RBF SVM, SVM with Lasso, LogReg with Lasso	2598 genes, 127304 SNPs	1378 profiles per omics-type	Binary Classification (Diagnosis; schizophrenia)	Nguyen et al. (2021)
Multi-omics (Transcriptomics, SNP-omics (genetic variants))	**DNN with Lasso**, DSPN, AdaBoost, DT, SVM, ANN, RF, kNN, GP, NBC, RBM, RBF SVM, SVM with Lasso, LogReg with Lasso	118 genes, 332 SNPs	248 profiles per omics-type	Binary Classification (Diagnosis; lung cancer stage)	Nguyen et al. (2021)
Multi-omics (Transcriptomics, Proteomics, Metabolomics, Fluxomics)	RNN, LassoReg, Ensemble LassoReg	4096 genes, 1001 proteins, 356 metabolites, ≤ 120[b] fluxes	≤3579[b] transcriptomics profiles, ≤71[b] proteomics profiles, ≤696[b] metabolomics profiles, ≤43[b] fluxomics profiles	Regression (Physiology; expression level of mRNAs, proteins and metabolites, prokaryotic growth rate)	Kim et al. (2016)
Multi-omics (Fluxomics, Metabolomics)	DT	≤106[a] metabolites, ≤175[a] fluxes	not assigned[a]	Binary Classification (Other; feasibility of kinetic parameter sets)	Andreozzi et al. (2016)
Multi-omics (time-series Proteomics and Metabolomics)	Models found by TPOT	≤86[e] metabolites, ≤76[e] proteins	21 profiles per omics-type	Regression (Physiology; metabolite time derivatives)	Costello and Martin (2018)

[a]True number not clearly obvious from the descriptions found in the main body of the work.

[b]Number might be lower because some (additional) raw features or samples might have been filtered out.

[c]Value varies between different prediction methods.

[d]This prediction task was repeated on other data sets from the same omics type(s) that are not listed here.

[e]Estimated from provided supplementary material.

Table notes: Counts for effective raw features/samples are explained in detail in Section 2.1. Additionally, non-omics features are not listed. The listed prediction methods are generic types, meaning that they may describe any derived method. Please consult the referenced publications for details on the utilized method. Supervised methods that were not used for predictions but e.g., in preprocessing, the post-hoc phase, or for additional analysis are not listed. Bold methods indicate which methods were presented as the authors' methods of choice or which were primarily used for predictions. Abbreviations: DNN, Deep Neural Network; RF, Random Forest; SVM, Support Vector Machine; DT, Decision Tree; LDA, Linear Discriminant Analysis; NSC, Nearest Shrunken Centroid; GBM, Gradient Boosting Machine (Boosted Tree Model, Generalized Boosted Model, Gradient Boosted Tree); TPOT, Python package for automatic model selection (see Supplementary Table S1); PLS-DA, Partial Least Squares Discriminant Analysis; RBF, Radial Basis Function Kernel; ANN, feed-forward Artificial Neural Network; LogReg, Logistic Regression; XGBoost, Extreme Gradient Boosting; Lasso, Lasso (L1) Regularization; LassoReg, Lasso Regression; SVR, Support Vector Regression; BEMKL, Bayesian Efficient Multiple Kernel Learning; BRF, Bagged RF; MMANN, Multi-Modal ANN; VAE, Variational Autoencoder; RNN, Recurrent Neural Network; Ensemble *X*, combination of multiple base models of type *X*; GNN, Graph Neural Network; NBC, Naïve Bayes Classifier; CapsNet, Capsule Network; GLM, Generalized Linear Model; LGP, Linear Genetic Program; AdaBoost, Adaptive Boosting; GP, Gaussian Process; RBM, Restricted Boltzmann Machine; SimNet, Similarity Network; *X*(R), Regression variant of method *X*; CoxPH, Cox Proportional Hazard Model; miRNA, micro Ribonucleic Acid; SNP, Single-Nucleotide Polymorphism; kNN, k-Nearest Neighbors; CNN, Convolutional Neural Network; DSPN, Deep Structured Phenotype Network.

2.1 Data set dimension and size

The number of features (i.e., data set dimension) and samples (i.e., data set size) can be an important factor for a ML model's performance. Alakwaa et al. (2018) found that their neural network model under-performed when data set size was low but out-performed other ML methods when the training set was sufficiently large. Further, Mendez et al. (2019) compared the performance of several ML models on multiple metabolomics data sets and suggested that, at least in their study, classification error was impacted less by a change in the ML method than by a change in the number of training samples. We have, therefore, also included this information in Table 1. However, one should be explicit when listing data set dimensions and sizes. In a ML project, the original data set is often heavily processed: original features are scaled, new features are created, some original samples or features are omitted, etc. In this work, we summarize the part of the workflow that starts after raw data tables have been constructed and manipulates data before it reaches the ML model for prediction as *data preprocessing*. A raw data table in this context summarizes one omics type and contains one value per omics entity for every observed entity (e.g., one abundance value per metabolite for every patient). Data preprocessing is outlined in more detail in Section 2.2. Preprocessing often changes the dimension and size of a data set, sometimes creating completely new features and samples. As an example, Toubiana et al. (2019) derived a set of 444 graph-based features for 339 pathways from a few repeated profiles of 106 metabolites by characterizing pathways in metabolite correlation networks. Sample conversions that change the entity a sample belongs to, e.g., from a "biological replicate" to a pathway (Toubiana et al., 2019), seem relatively rare. However, since feature conversions are frequently encountered (Andreozzi et al., 2016; Koh et al., 2019; Pai et al., 2019; Sharma et al., 2019; Toubiana et al., 2019; Culley et al., 2020; Zhang et al., 2021) we need to clarify what the numbers found in Table 1 mean.

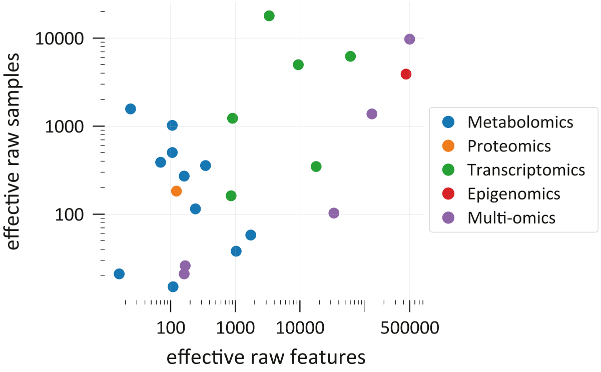


FIGURE 1
Comparison of effective raw data set dimensions and sizes in the categorized studies. Each point represents a data set that was used for optimizing and testing at least one predictive model. In *multi-omics*, a data set includes measurements from multiple omics sources. Each data set is plotted at the position that corresponds to its effective raw dimension and size. Please refer to the main text for explanations on the meaning of effective raw feature and sample counts (Section 2.1). Note that the graph shows only a selection of all ML-ready data sets from all studies. Supplementary Figure S1 provides references to the shown data points.

Typically, specifications of dimension and size characterize only either the raw data set or the ML-ready data set used in optimizing and testing a ML model. In our opinion, a reasonable alternative approach to express data set dimensions and sizes is one that quantifies the amount of raw data that ultimately contributes to the ML-ready data set. We call the corresponding values *effective raw feature/sample counts*. These metrics describe the number of raw features (i.e., variables of genes, SNPs, DNA methylation sites, proteins, metabolites, fluxes, etc.) and raw samples (e.g., omics feature profiles) from the raw data sets that contribute information to a single data set available for ML. Hence raw features or samples that are not integrated into the ML-ready data set because they were filtered out during preprocessing are not counted towards these values. However, even if raw features partially become target variables (Kim et al., 2016) they can still be considered effective. Since effective raw features and samples are part of the raw data set, it is important to not confuse their counts with specifications that refer to final features and samples of the ML-ready data set, which might be quite different. We argue that effective raw feature and sample counts allow comparison of ML-ready data sets even under extreme data set transformations and reductions. Although these numbers seem relevant they are unfortunately often difficult to reconstruct from a reader's perspective without analysing the original data and code. Further, when the same raw data set yields multiple distinct ML-ready data sets, effective counts can vary a lot between models, as noticeable in the study by Culley et al. (2020).

Figure 1 shows effective counts for ML-ready data sets in the 26 categorized publications. Generally, we find that studies that use solely metabolomics data (Leitner et al., 2017; Liu et al., 2017; Trainor et al., 2017; Alakwaa et al., 2018; Asakura et al., 2018; Date and Kikuchi, 2018; Hu et al., 2018; Stamate et al., 2019; Toubiana et al., 2019; Sha et al., 2021; van Dooijeweert et al., 2021) use a lower number of effective raw features for predictions than studies employing only transcriptomics (Sharma et al., 2019; Culley et al., 2020; Wang et al., 2020; Alghamdi et al., 2021; Wang et al., 2021). The two exceptions on the transcriptomics side (Culley et al., 2020; Alghamdi et al., 2021) originally had more raw features but some of them were omitted for at least one major analysis because some genes were not present in a metabolic network model. Due to technical limitations, metabolomics still struggles to reach high throughputs, such that either the number of raw features or the number of raw samples is restricted. This depends also on the experimental method. All metabolomics studies in Figure 1 with more than 200 effective raw features (Trainor et al., 2017; Stamate et al., 2019; Sha et al., 2021) use liquid chromatography coupled to mass spectrometry (LC-MS) or LC-MS together with another method, respectively. While the study with the second-lowest number of effective raw features (Liu et al., 2017) used LC-MS together with nuclear magnetic resonance (NMR) spectrometry, in this case, the authors reduced their raw feature count from originally 261 to 24 effective metabolite features for predictions. Although methods of 2-dimensional gas chromatography can detect respectable amounts of molecules (Phillips et al., 2013), studies that used solely gas chromatography (Alakwaa et al., 2018) or NMR (Asakura et al., 2018; Date and Kikuchi, 2018) did not reach more than 200 compounds. Another concern of metabolomics is that the exact identity of some of the raw features is often unclear (Weckwerth, 2011). Recently, some efforts have been made to solve this metabolite annotation problem also with machine learning approaches (Nguyen et al., 2019). The biological meaning of features is especially important when results should be interpreted. Consequently, interpretation methods that evaluate the importance of individual features might struggle to generate meaningful biological insight when applied to metabolomics data with unreliable annotations.

On the other end of the scope, transcriptomics oftentimes easily reaches over 3,000 effective raw features (Sharma et al., 2019; Wang et al., 2020; Wang et al., 2021) and studies that use measurements from multiple omics sources can have and retain close to 500,000 raw features due to the high-dimensionality of epigenomics data and strategies to condense this information (Zhang et al., 2021). However, taking into account more features for a prediction is not always favourable. Besides technical difficulties linked to data sets with many features, like storing large feature vectors and computational cost (Bommert et al., 2022), working with high-dimensional samples causes diverse issues. The machine learning literature summarizes challenges that arise in high-dimensional data sets under the "curse of dimensionality" (Bishop, 2006; Shalev-Shwartz and Ben-David, 2013; Forsyth, 2019). Especially, when relevant information in the data is "sparse," meaning that only a few features truly influence the prediction target, like it is often the case for transcriptomics data (Vikalo et al., 2007), considering additional features only "add[s] noise to the data" (Culley et al., 2020). Having high-dimensional samples, while the number of samples is much lower, is even worse. One major problem is that the same number of samples are often spread over wider distances in a higher-dimensional space (Forsyth, 2019, p. 77f) and it would, therefore, require much more samples to similarly populate this space (Bishop, 2006, p. 35). In this case, the risk of overfitting to the training data is increased (Kim and Tagkopoulos, 2018; Jiang et al., 2020). A way to mitigate the "curse" is by reducing the number of dimensions by combining original features to find a new lower-dimensional description for each original sample or by omitting some original features (Zhang et al., 2021). The corresponding methods are often called *feature extraction* and *feature selection* and summarized as *dimensionality reduction techniques* (Reel et al., 2021). These methods are frequently "unsupervised," meaning that they do not use the information stored in the labels (Cai et al., 2022) and are almost always advisable when dealing with a large number of raw features. Feature selection methods can make ML models more accurate (Chen et al., 2020) and better interpretable (Bommert et al., 2022). For more details, see the following section about data preprocessing (Section 2.2).

In addition, sometimes omics data such as metabolite amounts reference information that is changing over time. These dynamics are important to consider when modeling with data collected at multiple time points, as it may affect the reliability of ML predictions. One possible innovation for correcting algorithms that have to deal with input data representing dynamic information is by analysing *concept drift* (Agrahari and Singh, 2021). Concept drift in machine learning arises when the statistical properties of the target variable change over time, usually due to the fact that the identity of the input data that the model was trained on has significantly changed over time. Then, a model that is unaware of this change can no longer make accurate predictions. It has already been shown that metabolomics data is subject to concept drift, making prediction models not taking the dynamics into account less reliable (Schwarzerova et al., 2021).

2.2 Data preprocessing

In the machine learning community there is a popular saying: "garbage in, garbage out." It means that every successful machine learning project lives and dies with the quality of the data set it uses. Besides the experimental procedure that determines the raw data quality, data preprocessing, the step that takes raw data and turns it into a data set suitable for learning, is critical (Kotsiantis et al., 2007), especially for omics data (Kim and Tagkopoulos, 2018). Figure 2 illustrates the flow of data and information through a modeling framework, indicating the vital role of data preprocessing. Data preprocessing can involve many steps, and these often heavily depend on the raw data and application. In particular, during preprocessing

- data from different sources might be combined (**data integration**), e.g., microRNA and mRNA expression levels might be "concatenated" (Cai et al., 2022),

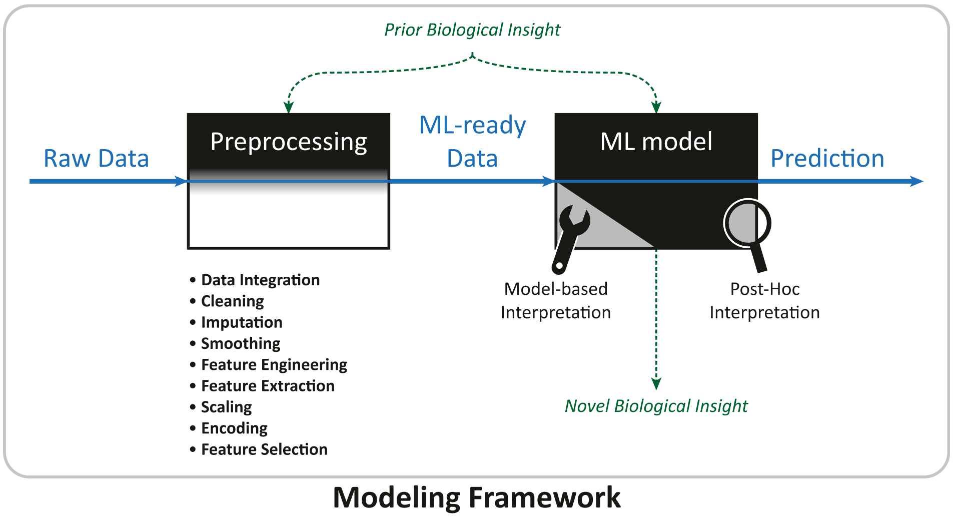


FIGURE 2
Data and information flow in a modeling framework. The modeling framework inhabits the complete work-flow of a machine learning project, from the raw data set to producing a final prediction. Data preprocessing converts the raw data set into a data set suitable for machine learning. In the machine learning phase, model-based or post-hoc interpretation methods might be applied to generate novel biological knowledge. Prior biological insight (see Table 2 for examples) might enter at different steps, sometimes improving the interpretability of the ML model.

- samples might be deleted (**cleaning**), e.g., because a patient might be an obvious outlier, the diagnosis is unclear, or a value is obviously corrupted like a negative abundance record,
- missing values need to be filled in (**imputation**), e.g., by inferring them from other measurements,
- noise might be reduced (**smoothing**), e.g., by “smoothing methods” (Simonoff, 1996),
- new features and data representations might be created with the help of dimensionality reduction techniques and/or expert knowledge (**feature extraction** [Guyon and Elisseeff, 2006] and **feature engineering** [Kuhn and Johnson, 2019]), e.g., an “autoencoder” (see Section 3.3.2 for explanation) might find a compact vector description of a large epigenomics profile (Zhang et al., 2021), or the fluxome might be inferred from transcript levels via constraint-based models (Culley et al., 2020),
- the scale of variables might be changed (**scaling**), e.g., normalizing and/or standardizing gene expression values within genes,
- the format of variables might be changed (**encoding**), e.g., “0” might indicate absence of a gene and “1” its presence (Kim et al., 2016),
- a subset of the initial variables might be selected (**feature selection**), e.g., some metabolite features can be disregarded because they are linked to pharmacotherapy of the disease of interest (Liu et al., 2017) or because they were previously reported to be irrelevant for disease prediction.

There is no universal recipe that, when applied to any data set, will yield good results (Kotsiantis et al., 2007). Hence, finding a preprocessing procedure that works well for a given problem sometimes requires testing several methods (Forsyth, 2019, p. 376). In many cases some preprocessing steps are not needed or they might need to be done in a different order. Additionally, prior biological knowledge might enter into the modeling framework at several points throughout preprocessing. A few examples are as follows: Culley et al. (2020) incorporated a genome-scale metabolic model into their modeling framework to derive simulated fluxome-level features by bounding reactions with experimental transcriptomics data. Pai et al. (2019) created features for groups of genes from transcript-level features by using known gene-pathway associations. Andreozzi et al. (2016) used prior knowledge about the kinetic properties of enzymes to help create multiple kinetic models that served as input to their machine learning model. Koh et al. (2019) used biological networks to calculate interaction-level features from the abundances of interaction partners (i.e., genes and proteins). Possibilities in finding new data representations seem very diverse. Omics profiles can be converted to images by mapping expression levels of genes or pathways onto pixels with unsupervised techniques, making them accessible for “convolutional neural networks” (Sharma et al., 2019; Oh et al., 2021), which are explained later in Section 3.3. Autoencoders can condense almost 500,000 biological features from three omics sources into a single feature vector with 128 entries informative for several subsequent predictions (Zhang et al., 2021).

Although preprocessing can reduce computational cost and significantly improve predictions (Zhang et al., 2019), it can also hurt performance when valuable information is accidentally thrown away during a preparation step (Bishop, 2006, p. 3; Guyon and Elisseeff, 2006, p. 4). This is observable in the work of Culley et al. (2020). In their performance comparison, distinct regression models that were trained on original experimental transcriptomic features consistently out-performed those trained on artificial flux features derived from the same experimental data. Culley et al. (2020) observed only performances similar to ML models trained solely on the original data when they combined information from the original and converted data. In one case, the integrated data slightly outperformed the original gene expression data. This example may demonstrate that mechanistic insight (e.g., constraint-based modeling) can enrich experimental data (Culley et al., 2020). Nonetheless, converting features from one omics layer to another should be done with care, since blindly trusting new features while disregarding the original data could lead to poorer results (Guyon and Elisseeff, 2006, p. 4). For details on how to prepare raw omics data sets for machine learning the work of Kim and Tagkopoulos (2018) is a good starting point. Further, there are great books (Guyon and Elisseeff, 2006; Kuhn and Johnson, 2019) for learning how to manipulate and select features in order to improve performance.

A common problem in omics data sets is that the number of features is much higher than the number of samples. In that case, dimensionality reduction through feature extraction, engineering, or selection is useful to reduce the impact of data sparsity on the prediction reliability.

3 Toolbox for supervised machine learning

With the growing interest in machine learning in recent years, the toolbox of available methods and platforms to apply them grows constantly. As a consequence, selecting a method that works well for a given task and data set can be daunting for non-experts in the field of data science. There is “no free lunch” (Wolpert and Macready, 1997) in supervised machine learning, meaning that there exists no “universal” model that works well in any situation (Shalev-Shwartz and Ben-David, 2013, sect. 5.1). Instead expertise about the specific biological problem is important for a successful ML project (Shalev-Shwartz and Ben-David, 2013, sect. 5.1.1). In this section, we provide an overview of some of the supervised learning methods that have been applied to omics data sets. Due to the sheer diversity of methods that have been introduced to systems biological problems (see Table 1), describing them all in detail would go beyond the scope of this work.

From a very general point of view, supervised learning is the task of learning a mapping (a “hypothesis”; Shalev-Shwartz and Ben-David, 2013, sect. 2.1) between a set of variables (the features) and one or more target variables (the labels) given a set of pairs of these two (the training data) to discriminate among target variables (Angermueller et al., 2016). The ML model normally receives features in the form of a vector (Angermueller et al., 2016). By convention this feature vector is denoted $\mathbf{x} \in \mathbb{R}^d$, where d is the dimension of the vector (Bishop, 2006; Forsyth, 2019; Deisenroth et al., 2020). For simplicity, we will now consider only the case where there is a single target variable. Depending on the type of this label one discriminates between two categories of supervised machine learning methods, namely *classification* and *regression*. In a classification problem setting, a label, y_i, describes to which class a sample, i, belongs and can take one of two in binary classification ($y_i \in \{C_0, C_1\}$) or one of many possible values in multi-class classification ($y_i \in \{C_0, C_1, \ldots, C_n\}$). If our goal is to predict if a tumor belongs to a cancer subtype, possible classes could be: “subtype-A,” “subtype-B,” or “subtype-C,” which could be encoded to the numerical values {0, 1, 2}. For regression the label is a real number, $y_i \in \mathbb{R}$ (Deisenroth et al., 2020, p. 289).

When using a training set of the form $T = \{(\mathbf{x}_0, y_0), (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ (Deisenroth et al., 2020, p. 370) the task for a supervised ML algorithm is now to select a suitable hypothesis (Shalev-Shwartz and Ben-David, 2013, chpt. 2). A hypothesis maps a feature vector, $\mathbf{x} \in X$, to a label, $y \in Y$, h: $X \to Y$ (Shalev-Shwartz and Ben-David, 2013, sect. 2.1). During learning, the algorithm picks from a set of possible hypotheses, the *hypothesis class*, $h \in \mathcal{H}$ (Shalev-Shwartz and Ben-David, 2013, sect. 2.3). To tell the learning algorithm which hypothesis works well, we have to define a criterion that measures how large the error is between the true label (y_i; known from the training data) and a prediction made by the model, $\hat{y}_i = h(\mathbf{x})$. This criterion is known as a *loss function*, $l_h(y_i, \hat{y}_i)$ (Deisenroth et al., 2020, p. 260). The optimization problem is now to minimize the mean error over all our training samples (Deisenroth et al., 2020, p. 260; Shalev-Shwartz and Ben-David, 2013, sect. 2.2 and 2.3). This is known in statistical learning theory as *empirical risk minimization (ERM) with inductive bias* (Shalev-Shwartz and Ben-David, 2013, sect. 2.3).

It is important to note that we usually want to find a model that minimizes the error on data not presented during training (Deisenroth et al., 2020, p. 261), like samples from patients we want to diagnose in order to give them the right medical treatment. However, minimizing this error would require unlimited training samples (Deisenroth et al., 2020, p. 261). The fact that we have only access to a restricted training set (Deisenroth et al., 2020, p. 262) is why one should always test a trained model on data the model was not fit to. A predictor that performs well on training data but poorly on new data has learned a bad hypothesis, one that does not generalize to new samples drawn from the same data-generating distribution (Maceachern and Forkert, 2021), as mentioned earlier in Section 2.

3.1 Classification

3.1.1 Support vector machine

Support Vector Machines (SVM) are frequently used for binary classification purposes (Leitner et al., 2017; Alakwaa et al., 2018; Date and Kikuchi, 2018; Sha et al., 2021; van Dooijeweert et al., 2021). In this basic setting, SVMs aim to find a "decision boundary" in the form of a hyperplane (Bishop, 2006, p. 326f) that segregates the two classes of data points (Forsyth, 2019, p. 21). In the case where the two classes are perfectly separable there exists an endless number of possible hyperplanes that correctly classify all training samples (Deisenroth et al., 2020, p. 374). SVMs select the hyperplane that lies half-way between the two data point clusters. More specifically, they choose the hyperplane that is farthest away (in terms of "perpendicular distance") from the nearest data point (Bishop, 2006, p. 327). SVMs can also be applied to problems where classes are not perfectly separable (Cortes et al., 1995) by permitting some data points to be incorrectly labeled (Deisenroth et al., 2020, p. 379). The error function that allows SVMs to find an optimal solution is the *hinge loss* (Forsyth, 2019, p. 23). There are several great books that introduce SVMs in detail (Bishop, 2006; Shalev-Shwartz and Ben-David, 2013; Forsyth, 2019; Deisenroth et al., 2020).

Support vector machines are probably one of the most classical machine learning methods and frequently serve as base-line models in performance comparisons for omics data sets (Asakura et al., 2018; Date and Kikuchi, 2018; Koh et al., 2019; Wang et al., 2020; Nguyen et al., 2021; Sha et al., 2021). van Dooijeweert et al. (2021) chose a SVM as their primary method to classify individuals based on their metabolomics signatures as either healthy or potentially having Diamond Blackfan Anaemia (DBA).

3.1.2 Decision trees, random forests and boosted trees

Decision trees classify samples based on a tree-like hierarchical decision process. Starting from a *root node* and proceeding towards one of many *leaf nodes*, a sample is classified by following a path within the tree that is controlled by making a decision at each step (i.e., at each "internal node"; Shalev-Shwartz and Ben-David, 2013, chpt. 18). A final decision leads to a leaf that determines the class label for the given sample (Shalev-Shwartz and Ben-David, 2013, chpt. 18). Decisions within the tree use certain properties of the sample, which can be viewed as asking a yes/no question similar to, Is the expression of gene A higher than a threshold? and then proceeding along the corresponding branch (Shalev-Shwartz and Ben-David, 2013, chpt. 18). Decision trees can be automatically constructed by repeatedly choosing questions ("splitting rules") from a pool of questions while each time evaluating the benefit of using a particular question with the help of a *gain measure* (Shalev-Shwartz and Ben-David, 2013, sect. 18.2).

The ability to verbalize and visualize a decision tree in terms of simple yes/no questions makes them a common example of a likely interpretable machine learning method (Shalev-Shwartz and Ben-David, 2013; Lipton, 2016; Murdoch et al., 2019). As long as its "depth" [i.e., the number of decisions to reach a leaf (Shalev-Shwartz and Ben-David, 2013, sect. 21.1)] stays within the limits of human comprehension a decision tree is usually a simulatable classifier (see Section 4.1 for explanation) as implied by Lipton (2016). However, decision trees have a known disadvantage, i.e., a single decision tree of arbitrary size tends to overfit data (Shalev-Shwartz and Ben-David, 2013, sect. 18.1 and 18.2). By combining multiple decision trees into a *random forest* (Breiman, 2001), letting them "vote" on labels, and choosing the one that gets the most votes, overfitting can be circumvented (Shalev-Shwartz and Ben-David, 2013, sect. 18.3). Using a "**b**ootstrap **agg**regat**ing**" (short "**bagging**") method (Breiman, 1996) is a common way to construct random forests (Forsyth, 2019, p. 41f).

Another approach that combines decision trees is *boosting* (Friedman, 2002). In short, boosting constructs a series of "base" models (e.g., decision trees) in which each model has a different voting power and they are trained such that more attention is brought to samples incorrectly labeled by earlier models (Bishop, 2006, p. 657). For a more detailed description of random forests and boosting please refer to the work of Breiman (2001) or to Bishop's (2006) book for bagging and boosting. Andreozzi et al. (2016) provide an illustrative toy example of a decision tree and demonstrate how the rules learned by the tree can be utilized to improve the "feasibility" of a population of kinetic models. Similar to support vector machines, random forests are popular for performance comparisons in systems biology (Alakwaa et al., 2018; Asakura et al., 2018; Date and Kikuchi, 2018; Wang et al., 2020; Nguyen et al., 2021; Sha et al., 2021).

3.1.3 k-nearest neighbors

k-nearest neighbors (kNN) is a method that classifies new data points based on how similar they are in their features to samples in the training data set for which the true class label is known (Forsyth, 2019, p. 7). More specifically, a new sample is given the label that is most probable when looking at its *k-nearest neighbors* (Bishop, 2006, p. 125f) in terms of an appropriate measure of distance in feature space (Forsyth, 2019, p. 8). kNN classifiers are sometimes used in performance comparisons (Trainor et al., 2017; Wang et al., 2020; Nguyen et al., 2021), however, from the 26 considered studies in this review, none presented kNN as their method of choice for predictions.

3.1.4 Nearest shrunken centroid

Nearest shrunken centroid (NSC) is a modified version of the nearest-centroid classifier and was proposed by Tibshirani et al. (2002) for inferring tumor classes from trancriptomics data. Its advantage over the original classifier (i.e., nearest-centroid) lies in that it allows for an inherent selection of features that are most distinct between sample classes (Tibshirani et al., 2002). Thus, it is suitable for data sets with a high number of features that may simultaneously contain only a few relevant signals like transcriptomics data.

Following the steps in the original publication (Tibshirani et al., 2002): First, the algorithm calculates an average sample (i.e., the *centroid*) for each class and the whole data set. Then, the similarity between the class centroids and the global centroid is evaluated by a *t*-statistic for every feature and class. This *t*-statistic is then numerically "shrunken" by subtracting a constant, Δ. In the final classifier, a feature effectively loses its ability to distinguish between classes if all of its corresponding values dropped beneath zero or became zero in this step. This way, features that are unimportant for predictions can be gradually removed as Δ increases (Tibshirani et al., 2002). Koh et al. (2019) modified the original version of NSC such that it takes into account also related features when calculating test statistics.

3.2 Regression

Regression is the task of finding a mapping from a feature vector to a real number (Jiang et al., 2020). In a regression setting, a fundamental assumption is that our labels are subject to some random measurement error; hence, there is no relationship between the labels and features in the form of a deterministic function (Deisenroth et al., 2020, p. 289). An example of a regression problem would be the prediction of an organism's body size from metabolomic measurements (Asakura et al., 2018).

3.2.1 Linear regression

In *linear regression* we assume that a straight line that is randomly displaced from the origin relates features and labels (Forsyth, 2019, p. 209). Given a training data set, suitable model parameters (a.k.a. fitting the line) are usually found by so-called "maximum likelihood estimation" using a "gradient descent" algorithm (Deisenroth et al., 2020, p. 293), which is, in this context, the same as finding the minimum of the sum of squared residuals between model predictions and the training labels (Bishop, 2006, p. 141).

3.2.2 Lasso regression

Lasso is a *regularization method* that was proposed by Tibshirani (1996) and can eliminate non-informative features by setting their contributions to zero, potentially yielding a *sparse model* (i.e., a model that effectively uses only some of the given features; Forsyth, 2019, p. 262f). Generally, regularization tries to avoid overfitting during training, e.g., by keeping parameters in reasonable ranges, embedding feature selection into the model (Jiang et al., 2020), or randomly switching neurons on and off in a neural network (Angermueller et al., 2016). In lasso regression, this is achieved by adding a regularization term to the loss function of the regression model that shrinks some parameters to zero, eliminating the contributions made by the corresponding features (Bishop, 2006, p. 144f). Kim et al. (2016) primarily used lasso regression in their modular ML approach to predict quantities in several omics layers and Nguyen et al. (2021) incorporated lasso regularization into their deep neural network for selecting predictive features. Lasso regression was also applied to omics data as a feature selection strategy for the final predictive model (Leitner et al., 2017; Liu et al., 2017; Pai et al., 2019). Leitner et al. (2017) used this approach to select for the most suitable set of metabolites for early prediction of gestational diabetes mellitus (GDM). A combination of two different data sets, blood and urine samples, showed the highest prediction accuracy with a SVM model.

3.2.3 Partial least squares regression

Partial least squares (PLS) regression was introduced by Wold (1975) and constructs a set of *latent variables* that are most predictive of multiple target variables from the original features (Abdi, 2010). PLS works well when there are less samples than features and when features are suspected to be highly correlated with each other (Abdi, 2010; Trainor et al., 2017). Consequently, metabolomics data lends itself to PLS, e.g., because of its oftentimes low number of samples with many features and correlated metabolites (Mendez et al., 2019). Additionally, PLS is well-accessible for post-hoc interpretations that measure feature importance (Fonville et al., 2010; Leitner et al., 2017; Mendez et al., 2019).

A variant of PLS that is sometimes used to classify omics profiles is *partial least squares discriminant analysis (PLS-DA)* (Trainor et al., 2017; Date and Kikuchi, 2018). In this case, the target variables are categorical and a threshold on the predictions made by a corresponding regression model determines the predicted labels (Brereton and Lloyd, 2014).

For in-depth mathematical descriptions of the regression and the classification approach, see Abdi (2010) and Brereton and Lloyd (2014). Fonville et al. (2010) discuss some interpretability aspects of PLS and related methods in metabonomics.

3.3 Neural networks

Neural networks comprise a large group of machine learning methods that all have in common that they contain entities called *neurons* (Sengupta et al., 2020). Real biological neurons and how they wire and learn together initially served as a model for these mathematical units (Macukow et al., 2016). Nonetheless, modern artificial neural networks (ANN) have only little in common with nervous systems. A neuron can be seen as a function that takes an input feature vector, $\mathbf{x}$, and returns a value, y, that represents its current *activity* (Angermueller et al., 2016). A typically non-linear *activation function* determines how the neuron responds to inputs weighted by learnable *weight* parameters (Mendez et al., 2019; Sengupta et al., 2020). Another learnable parameter, the *bias*, is added before the input-to-output conversion and determines how easily the neuron activates (Sengupta et al., 2020). Generally, one could speak of a neural network when a neuron receives input from another neuron.

In the most classical type of neural networks, called "feed-forward neural networks," neurons are organized into *layers* (Mendez et al., 2019). Each layer holds a number of neurons that solely receive input from neurons in the previous layer and pass their output only to neurons in the next layer. However, some neurons might receive no input and instead show a steady activation (Shalev-Shwartz and Ben-David, 2013, sect. 20.1). Nonetheless, normally two consecutive layers are "fully connected," meaning that every neuron in a subsequent layer receives a vector, $\mathbf{y}^{(i)}$, corresponding to all outputs from a preceding layer (Angermueller et al., 2016). In a feed-forward neural network there are three types of layers. The *input layer* feeds the feature vector of a sample for which a prediction is to be made into the network. This input signal is then propagated through one or more *hidden layers* until the last layer, the *output layer*, is reached. The outputs, $\mathbf{y}^{(\mathrm{out})}$, of the neurons in the output layer can for instance represent probabilities for cancer classes (Alakwaa et al., 2018) or even metabolite concentration change over time (Costello and Martin, 2018). In a binary classification task, the output layer often has only one neuron. At any hidden layer, an output vector, $\mathbf{y}^{(\mathrm{h})}$, can be seen as a new set of internal "features" for an input sample abstracted automatically by the hidden neurons from their input vector (LeCun et al., 2015). This ability, to sequentially find new, more discriminative, features, allows feed-forward neural networks to enrich the information relevant for predictions (Forsyth, 2019, p. 367) and filter out less relevant information (LeCun et al., 2015).

When neural networks contain more than one hidden layer they are often termed "multilayer" or *deep neural networks (DNNs)* (Shrestha and Mahmood, 2019; Zhang et al., 2019). Deep neural networks have the advantage that they avoid having to carefully construct (i.e., "hand-engineer") input features—instead the original raw features can be used directly in most cases (LeCun et al., 2015). *Backpropagation* is the key ingredient that allows DNNs to learn efficiently (Macukow et al., 2016). During backpropagation, the model's prediction error is traced back to individual model parameters, hence allowing them to be appropriately adjusted (LeCun et al., 2015).

Neural networks can be applied to a variety of problems (Shrestha and Mahmood, 2019). When we allow neural networks with a particular activation function to have an unlimited number of hidden layers they can theoretically simulate any function connecting input features and target variables (Hanin, 2019).

3.3.1 Specialized neural networks

There are a lot of different neural network architectures that were mostly designed to perform well on one specific task. Examples of specialized neural networks that have been applied to omics data sets are *convolutional neural networks* (Sharma et al., 2019; Oh et al., 2021), *recurrent neural networks* (Kim et al., 2016), *graph neural networks* (Alghamdi et al., 2021), *capsule networks* (Wang et al., 2020; Wang et al., 2021), and *autoencoders* (Zhang et al., 2021).

Convolutional neural networks (CNNs) were developed to work with data in which features have a known spatial relation, e.g., sequential data, image-like data, and stacks of image-like data (LeCun et al., 2015). They can learn to recognize complex objects such as animals in pictures by internally decomposing their input (LeCun et al., 2015). This ability is partly due to the fact that consecutive layers are not fully linked such that a neuron sees only a part of the whole picture, the "local receptive field" (Shrestha and Mahmood, 2019). Sharma et al. (2019) applied CNNs to transcriptomics data by assigning RNAs to pixels according to their similarity in the training data and then integrating RNA abundances into these pixels for every sample.

Recurrent neural networks (RNNs) perform well on time-series data, where "information of previous time steps" needs to be remembered because it is relevant for later time points (Sengupta et al., 2020). Unlike in classical feed-forward architectures (e.g., multi-layer feed-forward neural networks), in RNNs, neurons receive information extracted from earlier inputs additionally to the present input (Sengupta et al., 2020). Kim et al. (2016) used a RNN to predict transcript levels in a cell from genetic and environmental features in the hope of replicating the behaviour of cycles frequently found in transcriptional regulatory networks.

Graph neural networks (GNNs) is an umbrella term for neural networks which can work with data that can be represented as graphs (Zhou et al., 2018) and there are many subtypes of them (Wu et al., 2019). For instance, "Message Passing Neural Networks (MPNN)" (Gilmer et al., 2017) are a type of "convolutional graph neural networks" (Wu et al., 2019) in which vertices in the graph store information and share information along edges with neighboring vertices in a step-wise process until an output is generated by taking into account the final states of vertices (Gilmer et al., 2017) for local "node-level" or global "graph-level" predictions (Wu et al., 2019). Alghamdi et al. (2021) used a GNN to infer metabolic reaction rates in individual cells from transcriptomics data by viewing the metabolic network as a factor graph.

In the next sections, we will discuss autoencoders and capsule networks in more detail. We highlight autoencoders because of their ability to serve as powerful feature extractors, as demonstrated on multi-omics data (Zhang et al., 2021), and capsule networks because of their young age and distinct nature to "regular" neural networks. Shrestha and Mahmood (2019) and

Sengupta et al. (2020) review many more specialized neural network architectures, and Zhou et al. (2018) and Wu et al. (2019) discuss graph neural networks in great detail.

3.3.2 Autoencoders

An autoencoder is a special feed-forward neural network architecture that, rather than trying to predict target variables from an input, learns to output its given input (Martorell-Marugán et al., 2019). Since they only use feature information they can be classified as "unsupervised DNN[s]" (Shrestha and Mahmood, 2019). The important detail about this architecture is that it includes a hidden layer with usually only a few neurons (Sengupta et al., 2020). This characteristic layer is sometimes called the *bottleneck*. Since information is passed on from layer to layer, at the bottleneck the model is forced to find a description of the input with low dimension (Sengupta et al., 2020). In contrast to principal component analysis for dimensionality reduction, non-linear activation functions allow autoencoders to compress their inputs non-linearly (Shrestha and Mahmood, 2019), which can lead to more informative descriptions (Charte et al., 2018). The bottleneck divides autoencoders into two parts, the *encoder*, and the *decoder* (Shrestha and Mahmood, 2019). While the encoder tries to extract the most relevant information from the original input to condense it at the bottleneck, the decoder tries to reproduce the input in the output layer from it (Shrestha and Mahmood, 2019). Once an autoencoder was trained, it can generate a compact description from a sample which may then serve as input for predictive models or can be used to plot the data when the new description has only two or three dimensions (Zhang et al., 2021).

There is a wide variety of autoencoders that can serve other purposes than just dimensionality reduction. For instance, when an autoencoder is challenged to reproduce original samples from samples that were randomly perturbed the model can learn to remove similar "noise" from new samples (Gondara, 2016). Another commonly used version is a *variational autoencoder (VAE)*. Rather than learning discrete sample descriptions, VAEs learn the parameters of a normal distribution from which new descriptions can be drawn (Zhang et al., 2021). As such an VAE can act as a sample generator that could theoretically come up with omics measurements for imaginary patients when decoding a newly drawn description (Shrestha and Mahmood, 2019; Zhang et al., 2021). Furthermore, model parameters learned by an autoencoder can serve as first drafts for those of a supervised neural network, allowing effective "pre-training" of supervised models (Erhan et al., 2010) as demonstrated on omics data (Alakwaa et al., 2018).

3.3.3 Capsule networks

Capsule Networks (CapsNets) are a novel type of neural network that was introduced by the team of Geoffrey E. Hinton (Sabour et al, 2017). CapsNets have challenged the state-of-the-art CNNs in image identification. CapsNets aim to overcome some of the flaws of CNNs, like the loss of local information during a typical filter operation and difficulties with recognizing objects when they appear in new orientations (Sabour et al, 2017). CapsNets are exceptionally good at resolving objects when they are shown on top of each other (Sabour et al, 2017). According to the authors (Sabour et al, 2017), in a capsule network multiple neurons are configured into "capsules" that each detect the presence and characteristics of an associated "entity." In an transciptomics profile, an individual capsule can be set up to predict the presence of a specific protein and indicate its properties (Wang et al., 2021). A capsule returns a vector that corresponds to the activities of its neurons and indicates the probability that the entity is present with its scale and the entity's characteristics by its orientation (Sabour et al, 2017). Capsules are further organized into layers that follow a child-parent like hierarchy. As an example, in the implementation of Wang et al. (2020), the capsules in the last capsule layer each indicated the presence of a cell class that the authors aimed to predict. In a later work (Wang et al., 2021), child capsules of these parent capsules representing cell classes were encouraged to portray transcription factors or groups of interacting proteins. When processing samples, an innovative *dynamic routing* protocol ensures that each capsule signals mostly to a single parent capsule, i.e., the one whose output harmonizes well with its own, which amplifies plausible relationships between capsules and, consequently, between their entities (Sabour et al, 2017).

3.4 Software implementation

In terms of software implementation, three main programming languages, namely, Python, R and Matlab are frequently used in omics analysis. Currently, Python is coming to the fore in machine learning in general (Srinath, 2017). Despite many Python innovations, R offers numerous libraries and packages for biological analyses, including ones specifically for handling omics data (Chong and Xia, 2018; Picart-Armada et al., 2018). This is mainly due to the history of bioinformatics analysis using the Bioconductor repository (Gentleman et al., 2005). Nevertheless, we must point out that R has its original roots in statistical analysis. Thus, R also offers methods developed at the borderline between computer science and statistics (Torsten Hothorn, 2022).

The main difference in software implementations using Python or R is usually the target application. Mostly, R packages are created and tested for one data type with very specific properties, see Supplementary Table S1. As a result, the R language in omics analysis is seldomly used directly for developing neural networks, but rather for optimizing more classical learning methods such as linear regression or Bayesian

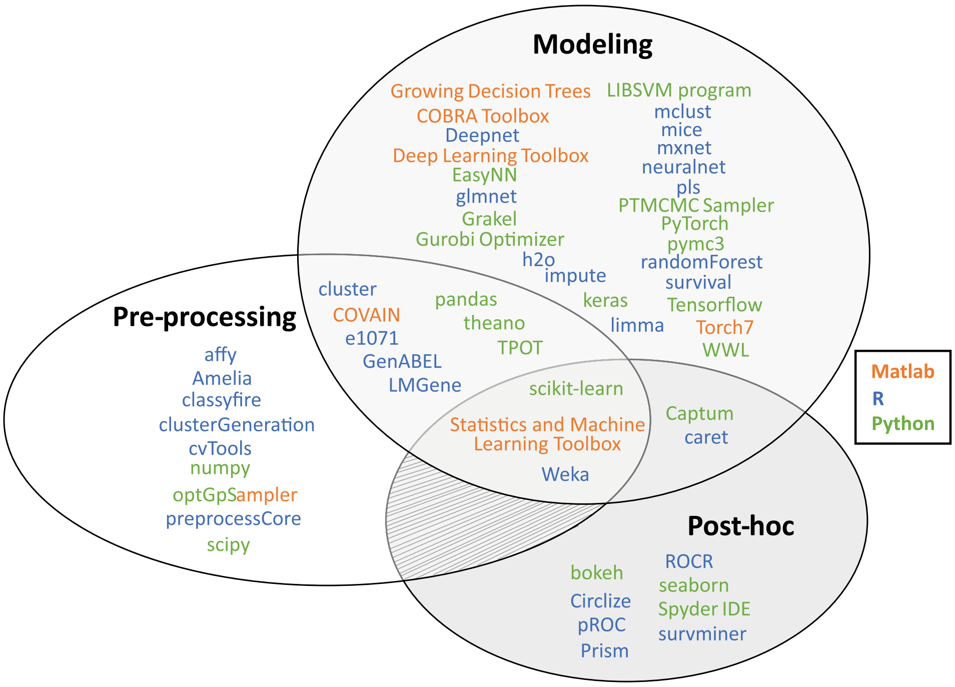


FIGURE 3
Overview of useful software packages for machine learning implementations from the most prevalent programming languages in computational biology (i.e., R, Python, and Matlab). All listed packages have been applied in an omics data analysis context (see Supplementary Table S1 for references). Most packages focus on either data pre-processing, the modeling phase (i.e., model-based interpretations and designing, training or executing a ML model in general), or the post-hoc analysis phase (i.e., post-hoc interpretations and data visualization).

methods. In addition, a large part of scientific research regarding ML algorithms is conducted in Matlab. Nowadays, Matlab also offers many new innovations related mostly to training and proper optimization of error functions in neural networks.

A combination of different languages also offers more analysis options. Appropriate interfaces exist for example to use Python in R[1] and Matlab[2]. A summary of useful software packages for (interpretable) machine learning can be found in Figure 3 and Supplementary Table S1.

4 What is interpretability?

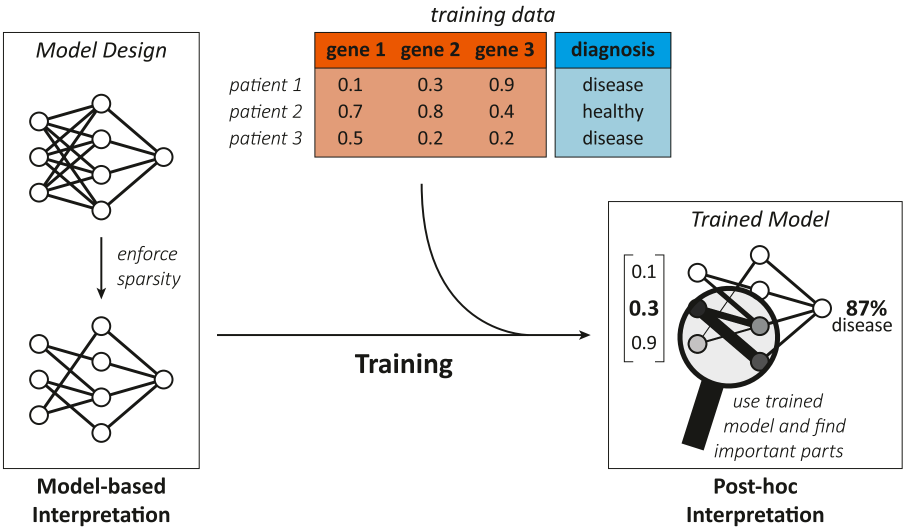
4.1 Basic concepts of interpretability

The concept of interpretability has been thoroughly discussed in recent years in the machine learning community, leading to a diversity of different perceptions, terms, and attempts at its definition (Lipton, 2016; Murdoch et al., 2019; Barredo Arrieta et al., 2020). Terms that are strongly associated with interpretable machine learning are *transparency* (Lipton, 2016; Barredo Arrieta et al., 2020), *white-box* (Loyola-Gonzalez, 2019), *explainability*, *understandability*, and *comprehensibility* (Barredo Arrieta et al., 2020). While all of these terms might capture different notions of the same overall concept (Lipton, 2016; Barredo Arrieta et al., 2020), they seem to refer to the same underlying desires, which are to trust, understand, or interpret the decision-making process or the results obtained from a machine learning model. Besides its controversial nature, there is a strong agreement that the topic of interpretability is important in machine learning (Lipton, 2016; Barredo Arrieta et al., 2020), especially for experts and scientists that deploy ML models to real-world problems (Murdoch et al., 2019).

Due to its many facets, it is necessary to fix a definition of interpretability when writing about it (Lipton, 2016). Interpretability can be defined as “the ability to explain or to provide the meaning in understandable terms to a human” (Barredo Arrieta et al., 2020) or to be able to extract “relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the

[1] https://www.rstudio.com/blog/reticulate-r-interface-to-python/

[2] http://mathworks.com/help/matlab/call-python-libraries.html



	gene 1	gene 2	gene 3	diagnosis
patient 1	0.1	0.3	0.9	disease
patient 2	0.7	0.8	0.4	healthy
patient 3	0.5	0.2	0.2	disease

FIGURE 4
Illustration showing the difference between model-based and post-hoc interpretation methods. A model-based interpretation strategy could be to design a sparse model by limiting the possible connections in a neural network (e.g., with knowledge about biological networks). Once the ML model is trained, post-hoc analysis can reveal the model parts that are most important for predictions, hinting on genes or biological interactions relevant for the disease.

model" (Murdoch et al., 2019). In this review, we would like to adapt the second definition and define it in the context of this work as the ability to generate biological insight from data with the help of machine learning methods.

4.1.1 Reliability of interpretations

To gain real insight, any information we extract from a ML model and interpret needs to be reliable. As Murdoch et al. (2019) describe, this depends on two criteria: *predictive accuracy*, i.e., performance of the model, and *descriptive accuracy*, i.e., performance of the interpretation method. They argue that interpretations would be unreliable if either the ML model fails to model the data accurately or the interpretation method is unable to correctly extract information from the model. Furthermore, we argue that interpretability relies on every step that leads towards an interpretation. This includes the whole analysis framework: we need to trust 1) that the raw data contains the desired information in an unbiased manner, 2) that the data preprocessing steps retain the relevant information from the raw data, 3) that, as suggested by Murdoch et al. (2019), the ML model correctly captures relevant information from the training data, and 4) that the interpretation method effectively conveys this information.

All of these points need to work correctly to avoid misleading interpretations. In particular, raw data quality is very important. If the raw data is flawed, both predictions and interpretations will automatically be inaccurate/misleading. Raw data quality relies on the experimental procedure, a topic we hardly touch on in this review. This further demonstrates the broad scope of interpretability.

Preprocessing depends on the properties of the available data, the problem of interest, and the ML model. Thus, individual preprocessing steps might need to be validated for every implementation. Generally, it is crucial to not accidentally lose valuable information during preprocessing, as discussed in Section 2.2.

Regarding the ML model, Murdoch et al. (2019) emphasize that "one must appropriately measure predictive accuracy." For this, samples in the test set must not be involved in model optimization and training, since they simulate how the model would predict labels of new/unknown samples. Further, one should collect test samples without bias, slight changes in the training set and model should not heavily impact predictive accuracy, and predictions should be equally accurate for all types of samples (Murdoch et al., 2019).

Murdoch et al. (2019) suggest that descriptive accuracy depends on the interpretation and ML method and that some ML methods offer either superior descriptive or predictive accuracy: while, e.g., a deep neural network may outperform a decision tree, the decision tree may be easier to interpret. In systems biology, we frequently want to achieve both, e.g., correctly diagnose a disease and understand the reasoning behind the diagnosis. Therefore, we may have to balance the two objectives (Murdoch et al., 2019).

4.1.2 Interpretation methods

There are two general classes of interpretation methods, namely *post-hoc* (Lipton, 2016; Murdoch et al., 2019; Barredo Arrieta et al., 2020) and *model-based* techniques (Murdoch et al., 2019), which Figure 4 exemplifies. Model-based interpretations rely on the implementation of ML models “readily providing insight into the relationships they have learned” (Murdoch et al., 2019), whereas post-hoc interpretations only take place after the designing and training process and try to produce relevant biological knowledge just from the finalized model (Murdoch et al., 2019; Barredo Arrieta et al., 2020).

Model-based interpretation methods

Model-based interpretability can be achieved by enforcing three different properties in a model: “sparsity,” “simulatability,” and “modularity” (Murdoch et al., 2019).

Sparsity arises when some parameters are set to zero by the ML model itself or explicitly by the designer with prior knowledge, thereby decreasing the number of variables that need to be comprehended (Murdoch et al., 2019). Further, sparsity can associate parts of the ML model with biological entities, which allows additional interpretations and is discussed in Section 4.4.1. Methods that enforce sparsity require that there is indeed only a limited number of relevant connections between the features and the prediction target as indicated by Murdoch et al. (2019). When too many or the wrong parameters are eliminated, the model might learn an inaccurate/misleading relation. Additionally, any parameter that influences an interpretation should have similar values when we retrain the model with a slightly different training set (Murdoch et al., 2019), e.g., one where a single sample was changed or omitted/added. This requirement is generally known as *stability* in learning theory (Bousquet and Elisseeff, 2002). Methods that offer sparsity are, for instance, lasso regularized models (Murdoch et al., 2019) and nearest shrunken centroid because they intrinsically eliminate contributions of unimportant features.

Simulatability refers to the degree at which a person can comprehend and could theoretically think/run through the whole procedure of computing an output for a given input (Murdoch et al., 2019; Barredo Arrieta et al., 2020) “in reasonable time” (Lipton, 2016). Human comprehension demands that the following properties are sufficiently low: the complexity of the studied problem [referred to as the complexity of “the underlying relationship” by Murdoch et al. (2019)], the samples’ dimension (Murdoch et al., 2019), the model’s overall complexity, and the number of steps from input to output (Lipton, 2016). Therefore, making simulatability a requirement would drastically shrink the space of available methods and biological problems (Murdoch et al., 2019). Examples of models that usually exhibit a high level of simulatability are linear and logistic regression models, single decision trees, k-nearest neighbor classifiers, rule-based models, single neuron neural networks (Barredo Arrieta et al., 2020), and linear genetic programs (LGPs).

Modularity is a property where the model includes elements (i.e., “modules”) that make the model partially understandable because they are interpretable on their own (Murdoch et al., 2019). In the two case studies of modular designs (Kim and Tagkopoulos, 2018; Alghamdi et al., 2021) we highlight later, modules allow restricted insight because their inputs and outputs are biologically meaningful. Consequently, the module as a whole depicts a biological mechanism. It is the biological process that connects transparent input and output [e.g., transcription and its regulation; translating the genotype and environmental context to the transcriptome (Kim et al., 2016)]. Nonetheless, the way a module mathematically models a biological process could be elusive. This type of modularity seems related to what Lipton (2016) and Barredo Arrieta et al. (2020) call *decomposability*, which they describe as that the model is fully composed of elements (i.e., features, internal variables, computations) that make instinctively sense. Hence, we might call these cases partially decomposable. Neural network based models with a modular design and “generalized additive models” offer modularity (Murdoch et al., 2019), while decision trees and linear models can be fully decomposable (Lipton, 2016).

Post-hoc interpretation methods

Post-hoc interpretation techniques act after training and aim to reveal some of the hidden “relationships” the model has internalized by viewing the training samples (Murdoch et al., 2019). We see post-hoc interpretations more generally as the action of extracting valuable information from a trained model. Thus, a post-hoc interpretation could be as simple as communicating naturally meaningful coefficients of a linear model to a human interpreter. There exist various post-hoc approaches for different ML models that try to interpret a trained model by, e.g., assessing the importance of input features or relationships between them (Murdoch et al., 2019), visualizations, providing exemplary predictions, simplifying the model, putting reasonings into words, or elucidating individual properties of the model (Barredo Arrieta et al., 2020).

For additional examples, and further clarifications on the mentioned terms regarding interpretability great resources are the works of Lipton (2016), Murdoch et al. (2019), and Barredo Arrieta et al. (2020).

TABLE 2 Categorization of research studies applying machine learning techniques to non-sequential omics data sets. Summary of interpretation methods, assigned category and the approach demonstrated in the publication that led to this classification (*top*). Summary of utilized modeling frameworks, assigned category and prior knowledge that entered the modeling framework (*bottom*).

Interpretation Method

Approach	Category	Ref
Sparse model	model-based	Koh et al. (2019); Pai et al. (2019); Nguyen et al. (2021); Wang et al. (2021)
Modular design	model-based	Kim et al. (2016); Alghamdi et al. (2021)
Well-simulatable model	model-based	Andreozzi et al. (2016); Hu et al. (2018); Sha et al. (2021)
Input-response analysis	post-hoc	Alakwaa et al. (2018); Costello and Martin (2018); Wang et al. (2020); Zhang et al. (2021)
Feature importance	post-hoc	Leitner et al. (2017); Alakwaa et al. (2018); Asakura et al. (2018); Date and Kikuchi (2018); Bahado-Singh et al. (2019); Culley et al. (2020); van Dooijeweert et al. (2021)
	no interpretation methods	Hoehenwarter et al. (2011); Liu et al. (2017); Trainor et al. (2017); Mendez et al. (2019); Sharma et al. (2019); Stamate et al. (2019); Toubiana et al. (2019)

Modeling Framework

Incorporated Prior Knowledge	Category	Ref
Biological network information		
Transcriptional regulatory network	light gray-box	Kim et al. (2016)[a]; Koh et al. (2019); Nguyen et al. (2021)[a]; Wang et al. (2021)[a]
Protein-protein interaction network	light gray-box	Kim et al. (2016); Koh et al. (2019); Wang et al. (2021)[a]
Co-expression protein network	light gray-box	Kim et al. (2016)
Metabolic network	light gray-box	Alghamdi et al. (2021)[a]
Pathways of metabolites	dark gray-box	Toubiana et al. (2019)[b]
Other biological relationships		
Chromosomal allocation of CpG sites	light gray-box	Zhang et al. (2021)[a]
Expression quantitative trait loci	light gray-box	Nguyen et al. (2021)[a]
Chemical composition	light gray-box	Alghamdi et al. (2021)
Constraint-based metabolic modeling	light gray-box	Kim et al. (2016)[b]
	dark gray-box	Andreozzi et al. (2016)[b]; Culley et al. (2020)[b]
Reaction kinetics	dark gray-box	Andreozzi et al. (2016)[b]
No prior knowledge	black-box	Hu et al. (2018); Sha et al. (2021); Alakwaa et al. (2018); Date and Kikuchi (2018); Bahado-Singh et al. (2019), Wang et al. (2020); van Dooijeweert et al. (2021); Leitner et al. (2017); Costello and Martin (2018); Asakura et al. (2018); Sharma et al. (2019); Mendez et al. (2019); Stamate et al. (2019); Liu et al. (2017); Trainor et al. (2017); Hoehenwarter et al. (2011)

[a] Knowledge was used to select connections in a neural network
[b] Knowledge was used to create new features/variables.

4.2 Interpretability categorization scheme

In this work, we have developed a scheme which allows us to categorize research studies that applied ML models to biological data sets. In this scheme, studies are classified into a total of nine combined categories according to two criteria, 1) the used interpretation method and 2) if and at which point prior biological insight was incorporated into the project. Table 2 summarizes similarities between the reviewed studies in these two characteristics and states the corresponding categorizations.

4.2.1 Use of interpretation methods

Following the definitions laid out in Section 4.1 we differentiate between:

- **No interpretation methods.** Studies that do not implement post-hoc or model-based interpretation methods.
- **Post-hoc interpretations.** Studies that gain biological insight by analyzing a trained ML model with post-hoc interpretation methods.

- **Model-based interpretations.** Studies that gain biological insight by either using a *well-interpretable* machine learning model as their primary model or modifying a machine learning model such that its sparsity, simulatability, modularity or decomposability is increased.

We consider machine learning models to be "well-interpretable" if they were explicitly declared to frequently demonstrate sparsity, simulatability, modularity, or decomposability by the interpretable machine learning community, or if they obviously display one of these properties. In particular, this includes, methods that use lasso regularization or "sparse coding" (Murdoch et al., 2019), decision trees (Lipton, 2016; Murdoch et al., 2019; Barredo Arrieta et al., 2020), linear regression models, logistic regression models, k-nearest neighbor classifiers, single neuron neural networks, rule-based models, Bayesian models (Barredo Arrieta et al., 2020), generalized additive models (Murdoch et al., 2019; Barredo Arrieta et al., 2020), neural network based models with a modular design (Murdoch et al., 2019), nearest shrunken centroid, and linear genetic programs.

4.2.2 Use of prior knowledge

Using prior knowledge to guide the design of a ML model can boost interpretability and even performance, e.g., when introducing sparsity (Murdoch et al., 2019). If neural networks are wired according to known biological relationships, elements of the ML model can be virtually coupled to biological entities. This possibility was demonstrated for cellular components (Ma et al., 2018), genes (Nguyen et al., 2021), regulatory proteins, and protein interaction clusters (Wang et al., 2021). For defining categories with respect to the integration of prior biological knowledge, we adopt a view from the field of system identification (SI). SI discriminates between the three categories black-box, gray-box and white-box for mathematical models based on the amount of theoretical and experimental knowledge that went into their construction (Sjöberg et al., 1995; Isermann and Münchhof, 2011).

Machine learning models are often tightly embedded into a much larger *modeling framework*. This modeling framework includes all data preprocessing steps as explained in Section 2.2 and can be seen as anything that supports the data flow from the initial raw data to a final prediction. Sometimes, this modeling framework can be enormous (Andreozzi et al., 2016), representing a significant portion of the added scientific value of a study. Because prior knowledge can enter not only in the ML model itself but also during preprocessing, we want to utilize this categorization criterion to capture a property of the modeling framework. With this in mind, we differentiate between:

- **Black-box.** Modeling frameworks that do not incorporate any prior biological knowledge—they are purely determined by measurement data ("data-driven").
- **Dark gray-box.** Modeling frameworks that incorporate prior biological knowledge in any step before the machine learning model that makes the final prediction.
- **Light gray-box.** Modeling frameworks that incorporate prior biological knowledge into their machine learning model. This category also includes cases where prior biological knowledge enters at both points, before the machine learning model, and within it.

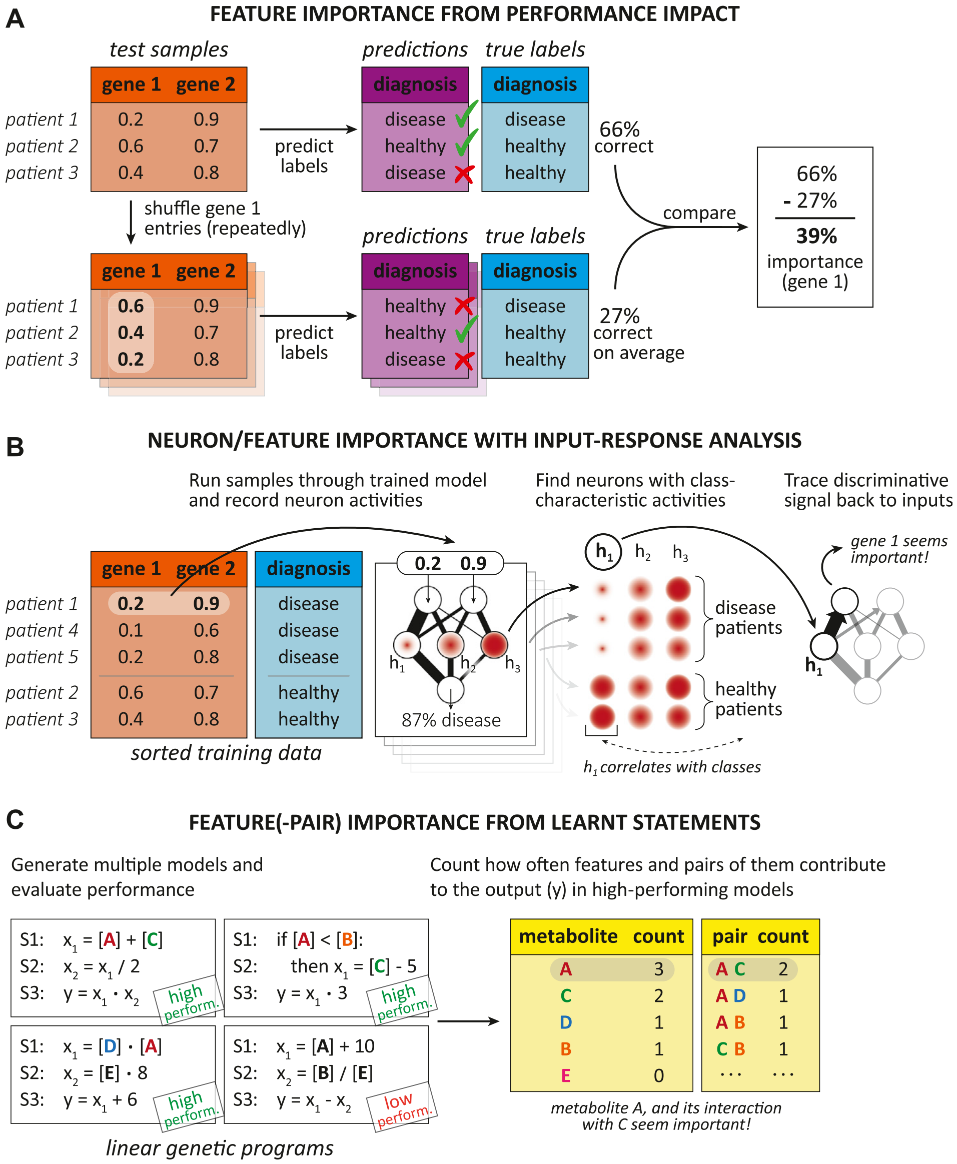
Please note that because of how SI (Sjöberg et al., 1995; Isermann and Münchhof, 2011) defines "white-box" models, a corresponding category would inherently exclude any approach that includes a ML model. This is because in SI, the term white-box describes models in which every mechanism and parameter is known from theoretical knowledge (i.e., previous experience and first principles), without relying on any measurement data (Sjöberg et al., 1995). In machine learning, a learning algorithm automatically integrates measurement data into mathematical models, which contradicts with the white-box definition from SI. Consequently, a white-box category does not appear in our scheme. Please further consider that the terms "black-box" and "white-box" frequently pop up in the machine learning literature and try to convey the level of interpretability of a ML model (Lipton, 2016; Loyola-Gonzalez, 2019; Murdoch et al., 2019; Barredo Arrieta et al., 2020). However, we avoid these notions because they seem vaguely defined and overused. We want to emphasize that they should not be confused with the well-established homonyms found in SI (Sjöberg et al., 1995; Ljung et al., 1998; Isermann and Münchhof, 2011) upon which we base our second criterion.

4.2.3 Additional considerations and examples

Although we try to outline clear categories, it is possible to encounter studies whose allocation seems uncertain. In this section, we provide additional considerations together with examples to make assignments more conclusive.

The model-based interpretations category does not exclude the use of post-hoc interpretation methods. From the fact that data is the target of interpretations (Murdoch et al., 2019) and how we defined post-hoc methods follows that post-hoc interpretations must always accompany a model-based strategy. For instance, the post-hoc method *integrated gradients* (Sundararajan et al., 2017) is applied by Nguyen et al. (2021) to a ML model that was modified to exhibit sparsity.

Whether a machine learning model is well-interpretable is difficult to judge. For instance, the notion of simulatability depends on the complexity of the model (Lipton, 2016; Murdoch et al., 2019; Barredo Arrieta et al., 2020). Decomposability demands that all features are meaningful (Lipton, 2016; Barredo Arrieta et al., 2020), which depends on



metabolite	count
A	3
C	2
D	1
B	1
E	0

pair	count
A C	2
A D	1
A B	1
C B	1
···	···

FIGURE 5
Examples of post-hoc interpretation methods from Section 4.3 in simplified form. **(A)** The impact of perturbing individual features on the overall performance of the ML model can indicate how important features are. This is the basic principle behind "mean decrease accuracy," as it was performed by Date and Kikuchi (2018). Thereby, entries of individual features are shuffled between test samples multiple times. Model performance is measured each time and compared to the performance obtained by using the unchanged test set. **(B)** Activation patterns of neurons can allow insight into the ML model, revealing important neurons that have learnt to discriminate between classes and important features that enable this discriminative ability. This input-response analysis was part of the post-hoc interpretation strategy of Alakwaa et al. (2018). They used it to verify that some neurons have learnt to distinguish between two cancer subclasses and to discover metabolites that are primarily associated with one subclass. In the generalized version illustrated here, activities of neurons in the first hidden layer are recorded while the already trained neural network processes the training samples. Comparing the activities between different classes can then identify characteristic neurons. Inputs that strongly connect to these class-characteristic neurons are likely important. **(C)** Statements found in well-performing linear genetic programs (LGPs) can reveal important input features and might further indicate important feature interactions. This was demonstrated by Hu et al. (2018) and Sha et al. (2021) and is shown here in a simplified way. As described by Hu et al. (2018) and Sha et al. (2021), LGPs are made up of a sequence of statements that convert some input features, [X], to an output variable, y, and are generated by a process similar to biological evolution. Since LGPs do not need to use all features, individual and pairwise counts of features that influence the output in well-performing programs may indicate the importance of features and their relationships (Sha et al., 2021).

the raw data and preprocessing. For modular designs, individual modules need to be interpretable on their own (Murdoch et al., 2019), which depends on their nature, context, and relationship to each other. Reducing the number of variables that need to be comprehended by building sparse models (Murdoch et al., 2019) might generally improve interpretability. However, if too many variables remain in the sparse model, interpretations may still be limited, as implied by Murdoch et al. (2019). All these factors vary between implementations of the same general ML method. To judge if a ML model is well-interpretable we have taken an Occam's Razor approach. We assume every implementation of a ML method is well-interpretable if the interpretable machine learning community (Lipton, 2016; Murdoch et al., 2019; Barredo Arrieta et al., 2020) mentions that the method usually displays sparsity, simulatability, modularity, or decomposability.

All categories that assess aspects of the machine learning method are based on *primary ML models*. Many studies develop a machine learning approach and then compare it to a set of well-established base-line models (Asakura et al., 2018; Date and Kikuchi, 2018; Wang et al., 2020; Sha et al., 2021; Zhang et al., 2021). We call the models that the authors present as their methods of choice/interest (or which they primarily use) for predictions the primary ML models. We considered interpretability aspects only of primary models and viewed the modeling framework from their perspective. Consequently, any additional ML models, e.g., base-line models, lasso regression to select features for a primary model (Leitner et al., 2017; Liu et al., 2017; Pai et al., 2019) or kNN for data imputation (Alakwaa et al., 2018; Stamate et al., 2019) did not influence our categorizations.

We considered models whose individual predictions were combined (Kim et al., 2016) as one large primary model. On the other hand, if models receive the same input but predict different target variables (Costello and Martin, 2018) these were not seen as one model.

Biological insight that is a direct consequence of a prediction was not considered to be generated by an interpretation method. For example, Toubiana et al. (2019) mapped pathways onto correlation networks, derived graph-based features and used these to predict if the pathways are part of the tomato metabolism. By repeating this procedure with unlabeled pathways testable hypotheses about their affiliation to tomato can be proposed (Toubiana et al., 2019). Here, the output is directly subject to interpretation, while the model itself is left untouched. We did not consider this case to be an interpretation method. Hypothetically, any prediction made by a ML model could be experimentally tested as long as the output has a biological meaning.

With all of this in mind, we are now ready to highlight some of the works we have categorized in Table 2 in more detail in the next sections. These sections focus on post-hoc and model-based interpretation methods. If studies have integrated prior biological knowledge in an original way, this will also be discussed.

4.3 Post-hoc interpretations

4.3.1 Discovering biomarkers by simple feature importance measures

Probably the most frequent approach to extract knowledge from a ML model is to assess *feature importance* in some way. Knowing how individual genes or metabolites influence the predicted probability of a disease can provide a first glimpse into the mechanisms of the disease. Investigating in which biological subsystems (e.g., pathways) predictive genes or metabolites participate lets us narrow down the origin of the disease within the system. A proposed set of relevant molecules could serve as biomarkers, enabling us to develop diagnostic tools that do not require untargeted omics screens. Further, reducing the number of considered variables may lead to more accurate predictions by lowering the noise brought by unnecessary information (Culley et al., 2020). In order to gain biological insight with feature importance scores, the inputs need to have a clear connection to a biological entity. For instance, when working with principal components as inputs, which could be linear combinations of measurements from over 60,000 genes, then, their relative importances most certainly provide no immediate biological insight. Nonetheless, in this specific case, importance scores could be backtraced to meaningful raw features (i.e., genes) by knowing the PCA loadings.

An advantage of methods that evaluate feature importance is that they are convenient to implement, as they come with many software packages for machine learning. Bahado-Singh et al. (2019) used functions from the packages *caret* and *h2o* in R to score patient properties, including metabolomic and proteomic measurements, according to their ability to discriminate between clinical outcomes. This allowed them to propose a single metabolite as a promising biomarker for premature delivery in pregnant women with the same physiology.

Leitner et al. (2017) ranked untargeted metabolomic features according to their importance in a PLS-DA model. The top-ranked metabolite in this analysis pointed them toward a specific metabolic pathway. Experimentally targeting this pathway by Stable Isotope Diluted Direct Infusion Electrospray Ionisation Mass Spectrometry (SID-MS) yielded new metabolomic features that improved predictions with a SVM model when combined with untargeted features. This demonstrates that novel biological insight from interpretations can also allow us to build better predictive models.

Date and Kikuchi (2018) estimated the relevance of metabolic markers in their deep neural network (DNN) in terms of "Mean Decrease Accuracy (MDA)." MDA measures the impact of perturbing an individual feature on the performance of the ML model (Figure 5A). To compute MDA for a feature, the authors compared the original performance of their ML model to multiple cases in which the entries of the feature were shuffled between data samples. A feature whose entries are mixed between samples loses some of its predictive power because the labels stay fixed, disconnecting many entries from their correct label. Multiple such iterations dampen the stochastic effects of random shuffling. Date and Kikuchi (2018) demonstrated that calculated MDA scores were similar among different ML methods and they allowed them to hypothesize about relevant metabolic markers for a sample's regional origin.

4.3.2 Biological insight from recording how the model responds to different inputs

Since supervised ML models learn how to map their inputs to different desired outputs, they can react very differently to different samples. Apart from the output itself, there are often internal responses that arise while processing a sample. For instance, neurons in neural networks activate differently, capsules in capsule networks couple to their parents differently, decision trees follow different paths to get to a leaf. Although these responses can be quite distinct, we can expect that they are mostly similar for samples with similar labels (e.g., those belonging to the same class). Monitoring these responses can be a handy tool to extract novel biological knowledge from a ML model. We call this general approach *input-response analysis*.

Alakwaa et al. (2018) addressed feature importance with the same method as Bahado-Singh et al. (2019) and additionally identified relevant metabolites and pathways by tracking how individual neurons in a neural network respond when presented with distinct inputs (Figure 5B). They trained their ML model on metabolite measurements from breast cancer patients belonging to the estrogen receptor positive or negative class, which are associated with distinct survival rates. Depending on the input, neurons found in each layer will activate differently. Alakwaa et al. (2018) noticed significant differences amongst the two cancer classes in the responses of some neurons in the first hidden layer of their trained model. By backtracing these discriminative signals over the strongest neuronal connections to the inputs, they could find relevant metabolites. The authors reported that some of these molecules were indicated to be linked to breast cancer by other studies. Finally, they looked at pathways harbouring relevant metabolites to further investigate their role in cancer metabolism. For this purpose, also data of enzymes showing distinct expression levels between the cancer classes was used. This study demonstrates how learnt connection weights together with neuron response patterns can allow a glimpse into the inner workings of a ML method often thought to be incomprehensible.

Wang et al. (2020) implemented a capsule network (see Section 3.3.3 for explanation) to predict a cell's type based on its single-cell gene expression pattern. They adopted an interpretation strategy very similar to that of Alakwaa et al. (2018). They analyzed how their capsule network responds to samples from different classes and backtracked the observed signals to the inputs of the network (i.e., transcript levels). This enabled the authors to hypothesize about a set of "core genes" typical for every cell class and allowing it to be discriminated from other cell classes. Their model is divided into two parts, a "feature extractor" and the actual capsule network. The feature extractor consists of several neural networks that each aim to find an informative vector description of the expression levels and supply it to a different primary capsule. A process called "Dynamic routing" connects the primary capsules to higher-level capsules (Sabour et al, 2017). In the implementation of Wang et al. (2020), the higher-level capsules each represent a cell class and their activation levels are used to classify a single-cell mRNA sample. During dynamic routing, so-called *coupling coefficients* are calculated that determine the contribution a primary capsule makes to the activity (Sabour et al, 2017) of a "cell type capsule" (Wang et al., 2020). These coupling coefficients depend on the input and the authors computed their mean values for every cell class. This way, they were able to find the primary capsules that received gene expression information that was most valuable for identifying each cell class. Since every primary capsule receives input from a single neural network, analysis of the weights learned by each network could identify genes characteristic of a cell class. Taken together, the work of Wang et al. (2020) further demonstrates that even very complex model architectures that have many parameters can allow the extraction of novel biological insight.

Nguyen et al. (2021) used the method *integrated gradients* (Sundararajan et al., 2017) to assess the relevance of their features (i.e., SNPs and genes). Integrated gradients presents the trained model with a series of artificial inputs that progressively contain more information from a real sample while looking on how the output changes in response (Sundararajan et al., 2017). With their calculated scores, the authors found genes and SNPs that most influenced probability for schizophrenia. Additionally, they designed their neural network such that links between the input and first hidden layer convey a biological interpretation, representing either SNP-gene or gene-gene interactions. Using a method derived from integrated gradients termed *Conductance* (Dhamdhere et al., 2018) together with their special architecture allowed Nguyen et al. (2021) to evaluate also the importance of the biologically meaningful connections in the neural network. They reported that many of their results are supported by literature, and additional data, respectively. Since the step of assigning neural network links to biological interactions involves altering the ML model, this approach falls under model-based interpretation methods and will be discussed in more detail in Section 4.4.1. Implementations of integrated gradients and conductance are available in the Python package Captum.

The work of Costello and Martin (2018) exemplifies how input-output analysis of a trained ML model can provide biological insight that is experimentally testable. The authors trained multiple models to each predict the current rate of change for another metabolite given metabolite and protein abundances at the same time instant. This design was chosen with the hope of challenging traditional kinetic models in their ability to pursue a metabolic system over time. The models were trained with samples from smoothed metabolite and protein trajectories of measured time series from two biotechnologically interesting pathways. Costello and Martin (2018) demonstrated how their models can generate novel biological knowledge. For that, synthetic data of multiple artificial "strains" was created. Each strain differed in how its protein timeseries was generated (i.e., by changing the parameters of hill function expression models). Using their ML models, they predicted potential product yields for each strain to identify proteins whose over-/underexpression influence yield. This analysis was done with partial least squares (PLS) regression. They also demonstrated that even if their ML models were trained only on two experimental data sets, they could exceed the accuracy of a carefully "handcrafted" kinetic model in predicting metabolite trajectories within a pathway. The ML framework of Costello and Martin (2018) as a whole could be argued to be interpretable because of its modular appearance. Every individual ML algorithm receives the same inputs and predicts another quantity, while both inputs (i.e., protein and metabolite levels) and output (i.e., dynamics of a single metabolite) have a clear biological interpretation. Nonetheless, we see their regression models as separate units and not parts of a modular design since their predictions are not combined and they are trained independently. Further, their ML models are very distinct and rather incomprehensible, comprising completely different methods discovered by the software tool TPOT that automatically generates efficient machine learning solutions for a given task (see Supplementary Table S1 for more information on TPOT).

4.3.3 Biological insight from ML methods that are frequently simulatable

Hu et al. (2018) utilized a supervised method that generates so-called *linear genetic programs (LGPs)* (Figure 5C). The authors used it to separate patients with osteoarthritis from healthy individuals based on their metabolome characteristics. As explained by Hu et al. (2018) and Sha et al. (2021), a linear genetic program is a sequence of "statements" that describes how features (i.e., metabolite abundances) should be combined with themselves or with other variables and under which conditions. At the end, a special variable constitutes the output of the program (i.e., chance for osteoarthritis). LGP classifiers are improved by an algorithm that essentially mimics biological evolution (Hu et al., 2018; Sha et al., 2021). After "training," Hu et al. (2018) evaluated the number of times a metabolite feature was present in one of their best performing models, and how often two metabolites appeared in the same LGP. With this information and the help of graph analysis they identified potential metabolic markers and showed that they correlate in their incidence in the top LGPs.

In their recent study (Sha et al., 2021), Hu et al. applied the same method to discover metabolites that can differentiate between patients with Alzheimer's Disease (AD), patients with amnestic mild cognitive impairment, and healthy individuals. Many of the top metabolites found to be predictive of AD were also suggested by two other ML methods (i.e., RF and SVM), however, with some discrepancies.

Alakwaa et al. (2018), Wang et al. (2020), and Hu et al. (2018) found their own method to rank features by their predictive power. When extracting such importance scores from a model, comparing the results to those obtained by established methods can be critical. The result, which features are important, should not depend on the utilized method because feature importance should be fundamentally determined by the causal relationships found in the biological process that created the data. Sha et al. (2021) reported that from 20 relevant metabolites found by their method, 10 overlapped with 20 they had identified using another post-hoc interpretation method on another ML model. Although in total 242 metabolites were considered, this could indicate that some of the top-ranked metabolites from one method might not be good biomarkers. Individual linear genetic programs demonstrate high simulatability because they can be read like a piece of computer code as suggested by Sha et al. (2021). However, as demonstrated by the results of Hu et al. (Hu et al., 2018; Sha et al., 2021), LGPs performing well on the same data can be very diverse, using different features and relationships between them. Hence, we note that one should be careful to not over-interpret a single LGP.

Andreozzi et al. (2016) expanded their previously developed "ORACLE" framework by a machine learning part. According to the authors, ORACLE integrates experimental data, including metabolomics and fluxomics, and theoretical knowledge about enzyme kinetics and creates a collection of kinetic models. The aim of their decision tree algorithm "iSCHRUNK" was then to learn from these kinetic models what makes some of them "feasible" and others not. Kinetic models generated by ORACLE were labeled as either feasible or not feasible. Models were considered feasible if they had a locally stable steady state, and matched theoretical knowledge as well as the available experimental data. The parameter values and feasibility label of each kinetic model embody their training data set. After training, the learned "splitting rules" (see Section 3.1.2 for explanation) can be interpreted as kinetic parameter ranges that partition the parameter space. Drawing from a feasible region of the space allowed the authors to discover new parameter sets corresponding to presumably feasible kinetic models.

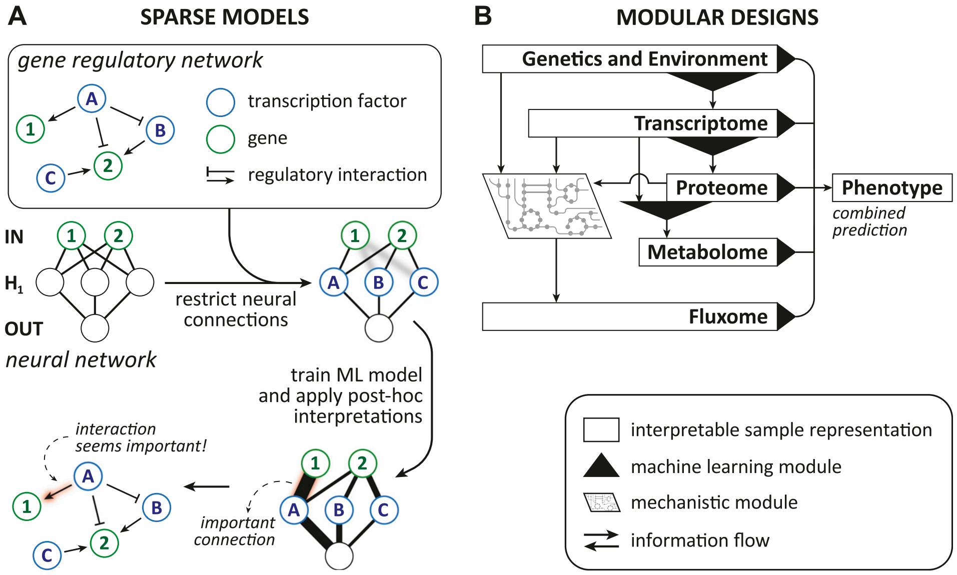


FIGURE 6
Examples of model-based interpretation methods from Section 4.4 in simplified form. **(A)** Guiding the topology of a neural network by a biological network can increase interpretability. This example captures the fundamental principle of the model-based interpretation strategy presented by Nguyen et al. (2021) and further by Wang et al. (2021). By assigning each neuron in the first hidden layer, H_1, to a biological entity (i.e., a transcription factor in this example), connections to biologically meaningful inputs (i.e., genes) can be wired according to a biological network. This limits the possible connections in the neural network, introducing sparsity. After training, post-hoc techniques could measure the importance (*red glow*) of the interpretable connections, revealing potentially relevant biological interactions. **(B)** Simplified version of the modular design described by Kim et al. (2016) for predicting the growth phenotype, metabolic dynamics, and expression levels of a cell from its genetics and environment. In general, a modular design may describe a sample or aspects of it in different biologically meaningful ways (i.e., interpretable sample representations). Modules then convert between these transparent representations and may rely on machine learning or mechanistic principles. In the portrayed example (Kim et al., 2016), ML modules connect the genetic and environmental inputs and different omics representations. A mechanistic module (i.e., a metabolic model) is embedded into the design and infers the fluxome under constraints derived from multiple representations. Finally, predictions from multiple ML modules are combined to estimate the phenotype.

Since decision trees are frequently outperformed by more complex methods like random forests (Alakwaa et al., 2018; Sharma et al., 2019), the work of Andreozzi et al. (2016) is an excellent example of how supposably sacrificing predictive accuracy by choosing a simple ML model can drastically increase descriptive accuracy. Using a decision tree allowed them to exploit it as a generator for high-quality kinetic models, which could probably not have been done so easily using more complex ML models like neural networks.

4.4 Model-based interpretations

As outlined in Section 4.1.2, model-based interpretation techniques rely on modifying the ML algorithm to increase interpretability or choosing a well-interpretable model (Murdoch et al., 2019). Note that the following examples focus on improving interpretability by design choices rather than selecting archetypally interpretable models. A general pattern that can be recognized is that most studies mentioned here couple parts of their ML model to biological entities with the help of biological network information.

4.4.1 Sparse models allow more in-depth interpretations

In their recent study, Wang et al. (2021) enhanced their capsule network that we described earlier in Section 4.3.2 by incorporating prior insight from biological networks. Their ML model was designed to take a single-cell transcriptomics profile as input and predict the type of the corresponding cell. Expression information from all genes (the inputs) was fed into every primary capsule via its own neural network. In this work (Wang et al., 2021), sparsity was enforced because only genes who are regulated by the same transcription factor (TF) or genes whose proteins interact (i.e., participate in the same interaction subnetwork) provide input to the same primary capsule. This way, primary capsules are primed to represent individual TFs or protein interaction clusters. Gene-TF and gene-cluster relationships were inferred from a transcriptional regulatory network (TRN), and a protein-protein interaction (PPI) network, respectively. After training, they applied a similar post-hoc interpretation strategy as in their previous study (Wang et al., 2020). Again, by calculating mean coupling coefficients for every cell type, the relationships between primary capsules and their parents (the "cell type capsules") could be unraveled. This time, the mean coupling coefficients could be directly interpreted as relevances of individual TFs and protein clusters for classifying a certain cell type. Their results supported their interpretation approach. They reported that important TFs and PPI clusters were predominantly associated with only a single cell type and many of these affiliations were known from literature. Wang et al.'s work demonstrates that with the help of prior knowledge more in-depth interpretations are possible [compare Wang et al. (2020) with Wang et al. (2021)]. Their previous black-box modeling framework was converted to a gray-box and by invoking sparsity they successfully implemented a model-based interpretation method.

Nguyen et al. (2021) deployed a sparse deep neural network to learn about potential biological relationships. Their model infers a diagnosis for schizophrenia from transcriptomics and genetic variants (SNPs) data. Features from both biological data types served as the inputs for the neural network. However, neurons in the first hidden layer were allowed to receive only information from inputs that are associated with the same gene (Figure 6A). This way, these neurons were tied to individual genes similar to the primary capsules in the work of Wang et al. (2021). Associations between the gene neurons and inputs were inferred from expression quantitative trait loci (the gene's expression is influenced by the input SNP), and transcriptional regulatory interactions (the gene is regulated by the input gene). Since, in their design, the inputs, the first hidden neurons, and connections between them have a biological meaning, more advanced post-hoc interpretations were possible, as described in Section 4.3.2. Additionally, the authors' neural network was lasso regularized (see Section 3.2.2 for explanation) such that inputs from genes and SNPs with low predictive power are ignored. Both limited connectivity and lasso regularization increase the sparsity of the ML model, making interpretations easier.

Koh et al. (2019) employed a network-focused strategy to classify breast cancer tumors based on their multi-omics signatures and learn about molecular subsystems that characterize tumor subclasses. Raw transcriptomics, proteomics, and gene copy number features were converted to one feature per molecular interaction. Considered interactions were either TF-gene or protein-protein interactions from a TRN, or PPI network, respectively. The new interaction-level features should reflect the probabilities of each interaction and were, thus, calculated such that they were high if both interaction partners were overexpressed, and low if both were underexpressed. Gene copy number information served as a tool to scale mRNA abundances, reducing/increasing them when the corresponding gene was over-/underrepresented. For learning from the new features, the authors modified the original nearest shrunken centroid (NSC) algorithm. As many other supervised methods, vanilla NSC cannot integrate any prior knowledge about how features might influence each other. However, their modification allowed NSC to consider also the features of interactions that are close in the biological network context when deciding whether an interaction's feature is important for discriminating between classes. This allowed the authors to favor interactions that form subsystems. The authors suggested that these subsystems are biologically more meaningful than important interactions that are dispersed over the biological network. Importantly, NSC chooses a set of relevant features for every class separately (Tibshirani et al., 2002). Consequently, the subsystems discovered by Koh et al. (2019) varied between tumor subclasses. Further, identifying annotated pathways that agree with important subsystems facilitated interpretability and enabled the authors to hypothesize about pathway over-/underexpression in breast cancer subclasses.

The work of Koh et al. (2019) demonstrates that biological expertise can help us to carefully engineer new interpretable features that allow us to view our data from a different (e.g., network) perspective. Notably, the authors reported that in comparison to unmodified NSC applied directly on proteomics and transcriptomics features, their method performed worse on experimental data and better on synthetic data. Although their interaction-level features seem to have captured most of the valuable information stored in the raw features while offering great interpretability, this example again emphasizes that one should be careful when replacing original features, as mentioned earlier in Section 2.2.

4.4.2 Modular designs are partially transparent

Kim et al. (2016) curated a large data compendium for *E. coli* "Ecomics," which harbours measurements from five different omics layers, together with data about the experiments and network-type data. With this data collection they predicted the complete state (i.e., levels of mRNA, proteins, metabolites, and metabolic fluxes) and growth dynamics of a cell based on its genetics (i.e., strain, genetic perturbations) and environmental factors (i.e., medium, stress). Their design (Figure 6B) was divided into modules that each predict quantities from only one omics layer. The metabolic fluxes were predicted with constraint-based metabolic modeling, while all other modules used machine learning (i.e., a recurrent neural network or lasso regression). Modules were partly exchanging information, providing and/or receiving predicted values to/from other modules. Information from all modules was compiled to collectively predict growth rate. Most interesting for this review is their recurrent neural network (RNN) for predicting transcript abundances. The RNN received a description of the experimental condition and was trained to match experimental transcript profiles with its predictions. The authors selected a RNN for this task to account for cycles (i.e., “feedback loops”) frequently found in transcriptional regulatory networks. The authors hoped that signals would propagate through the iterative layers of the RNN similar to signals traveling in a loop in the biological network. Further, they chose a sigmoid activation function partly because of its similarity to the Hill function. Intriguingly, when neural connections (referred to as “network topology” by the authors) in the RNN were guided by a transcriptional regulatory network, predictions were more accurate. As a whole their modeling framework is a good example of an interpretable machine learning framework due to its pronounced modularity. Every input and output of a module has a clear biological meaning. Besides providing transparency, modular designs have the advantage that modules can be trained/tuned independently as long as data for a module’s input and output is available, which allowed Kim et al. (2016) to use most of their data collection as training data. Further, they integrated prior knowledge at several points in their design, including the metabolic model for fluxome predictions. This makes their modeling framework a light gray-box.

Alghamdi et al. (2021) developed a graph neural network that can estimate metabolic fluxes in one cell from single-cell transcriptomics data. For that the metabolic network was viewed as a directed factor graph (i.e., a special bigraph). In this bigraph, “factor nodes” were individual metabolites and “variable nodes” embodied the reactions in which connected metabolites participate. Directed links indicated whether the metabolite acts as a product or a substrate in a reaction. This graph was constructed from the stoichiometry of a global metabolic network, and then reduced in size to cope with the computational cost linked to finding global flux solutions. Reductions were realized by combining reactions and omitting certain metabolites. To train their model a tailored loss function (see Section 3 for explanation) was designed. Therein reasonable solutions were defined to minimize the “flux imbalance” (i.e., influx versus outflux) of all metabolites, harbour zero or positive fluxes with an appropriate scale, and possess consistency with experimental data. Each rate of a combined reaction was estimated from the transciptomic features of its associated genes via a deep neural network (DNN), resulting in a total of 169 parallel DNNs that need to be trained in harmony. For training, Alghamdi et al. (2021) used their own algorithm. They tested their approach on various data sets, including their own, where they compared predicted flux changes due to genetic and environmental perturbations with experimentally observed metabolite concentration changes, confirming the predictive ability of their approach. We see their complete graph neural network as a modular system with good model-based interpretability. Every input (i.e., single-cell transcriptomic features) and output (i.e., metabolic fluxes) of each DNN module has a clear biological interpretation. Modules are arranged/connected according to a biological network topology, allowing network analysis. This possibility was demonstrated by the authors: by specifically up- or downregulating groups of genes (e.g., in glycolysis) certain metabolic subnetworks (e.g., the Krebs cycle) were impacted as expected. Further, they showed that targeting individual genes can reveal the genes that most influence certain fluxes.

5 Conclusions and outlook

In this review, we have categorized 26 scientific papers according to their interpretation strategies and the integration of prior knowledge and discussed some of them in detail. We have found that despite the large diversity of machine learning methods utilized in these studies, some parallels in their interpretation methods can be established. The majority of studies computed scores that assess the importance of input features (Alakwaa et al., 2018; Asakura et al., 2018; Date and Kikuchi, 2018; Hu et al., 2018; Bahado-Singh et al., 2019; Koh et al., 2019; Wang et al., 2020; Nguyen et al., 2021; Sha et al., 2021; Wang et al., 2021). These scores were then sometimes used to discover molecular subsystems (e.g., pathways) of interest (Alakwaa et al., 2018; Koh et al., 2019; Wang et al., 2020; Nguyen et al., 2021). Most model-based interpretation methods relied on either coupling parts of a machine learning model to comprehensible biological entities [e.g., genes (Nguyen et al., 2021), TFs (Wang et al., 2021), interacting proteins (Wang et al., 2021), fluxes (Alghamdi et al., 2021)] or associations between them [e.g., regulatory interactions (Wang et al., 2021), SNP-gene links (Nguyen et al., 2021)] or implementing ML methods that can be considered simulatable (Andreozzi et al., 2016; Hu et al., 2018; Sha et al., 2021). Many papers integrated prior knowledge in the form of biological networks into their modeling frameworks (Koh et al., 2019; Toubiana et al., 2019; Alghamdi et al., 2021; Nguyen et al., 2021; Wang et al., 2021), thereby turning them into gray-boxes; while some studies even incorporated whole constraint-based models (Kim et al., 2016; Culley et al., 2020). Whenever extracting knowledge from machine learning approaches, it is important to make sure that the results are in-line with available literature. One reason why this is especially critical is that many ML models use stochastic training algorithms that can produce drastically different parameterizations on the same training set. When these parameters then influence interpretation results, e.g., by calculating importance scores, we need to make sure that the results are not due to random effects. In other words, results found by interpretation methods should be consistent between different training runs and methods, to not fall into the trap of overinterpretation.

Because we find that the combinatorial space of distinct biological data sets (in source/type, dimension, and size) and what we could learn from them seems endless, interpretation methods might always need to be tailored to a specific scientific problem. Just like in data preprocessing (see Section 2.2) there is no universal recipe for good results. This is, despite some fundamental similarities, reflected in the diversity of approaches we highlighted in this review. A consequence of this diversity is that putting interpretation strategies into well-defined categories can be complicated. One reason for this is the fuzziness of the notions associated with interpretability. For instance, the definition of simulatability is very subjective. At which point is a ML model like a decision tree simple enough for a human to reconstruct its decision-making process? Apart from the ambiguity in terminology arising from different notions, we see a high relevance of interpretable machine learning in systems biology research.

Author contributions

DS and SW planned the study and conducted literature analysis. DS drafted major parts of the text. JS performed and drafted the analysis of software implementations. WW interpreted analysis results. SW contributed to and critically revised the text. All authors read and approved the final version of the manuscript.

Funding

JS has been supported by grant FEKT-K-21-6878 realised within the project Quality Internal Grants of BUT (KInG BUT), Reg. No. CZ.02.2.69/0.0/0.0/19_073/0016948, which is financed from the OP RDE.

Acknowledgments

We would like to thank Kevin Sidak, Nicolas Nemeth, and Christiane Derra for their suggestions, support, and informative discussions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.926623/full#supplementary-material

References

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *WIREs Comp. Stat.* 2, 97–106. doi:10.1002/WICS.51

Agrahari, S., and Singh, A. K. (2021). "Concept drift detection in data stream mining : A literature review," in *Journal of King Saud University - Computer and Information Sciences* (Amsterdam, Netherlands: Elsevier). doi:10.1016/J.JKSUCI.2021.11.006

Alakwaa, F. M., Chaudhary, K., and Garmire, L. X. (2018). Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. *J. Proteome Res.* 17, 337–347. doi:10.1021/ACS.JPROTEOME.7B00595

Alghamdi, N., Chang, W., Dang, P., Lu, X., Wan, C., Gampala, S., et al. (2021). A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data. *Genome Res.* 31, 1867–1884. doi:10.1101/GR.271205.120

Andreozzi, S., Miskovic, L., and Hatzimanikatis, V. (2016). iSCHRUNK - in silico approach to characterization and reduction of uncertainty in the kinetic models of genome-scale metabolic networks. *Metab. Eng.* 33, 158–168. doi:10.1016/J.YMBEN.2015.10.002

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol. Syst. Biol.* 12, 878. doi:10.15252/MSB.20156651

Asakura, T., Date, Y., and Kikuchi, J. (2018). Application of ensemble deep neural network to metabolomics studies. *Anal. Chim. Acta* 1037, 230–236. doi:10.1016/J.ACA.2018.02.045

Bahado-Singh, R. O., Sonek, J., McKenna, D., Cool, D., Aydas, B., Turkoglu, O., et al. (2019). Artificial intelligence and amniotic fluid multiomics: Prediction of perinatal outcome in asymptomatic women with short cervix. *Ultrasound Obstet. Gynecol.* 54, 110–118. doi:10.1002/UOG.20168

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, 82–115. doi:10.1016/J.INFFUS.2019.12.012

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer. doi:10.1007/978-0-387-45528-0

Bommert, A., Welchowski, T., Schmid, M., and Rahnenführer, J. (2022). Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Brief. Bioinform.* 23, bbab354–13. doi:10.1093/BIB/BBAB354

Bousquet, O., and Elisseeff, A. (2002). Stability and generalization. *J. Mach. Learn. Res.* 2, 499–526. doi:10.1162/153244302760200704

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi:10.1007/BF00058655

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/A:1010933404324

Brereton, R. G., and Lloyd, G. R. (2014). Partial least squares discriminant analysis: Taking the magic away. *J. Chemom.* 28, 213–225. doi:10.1002/CEM.2609

Cai, Z., Poulos, R. C., Liu, J., and Zhong, Q. (2022). Machine learning for multi-omics data integration in cancer. *iScience* 25, 103798. doi:10.1016/J.ISCI.2022.103798

Charte, D., Charte, F., García, S., del Jesus, M. J., and Herrera, F. (2018). A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. *Inf. Fusion* 44, 78–96. doi:10.1016/J.INFFUS.2017.12.007

Chen, Z., Pang, M., Zhao, Z., Li, S., Miao, R., Zhang, Y., et al. (2020). Feature selection may improve deep neural networks for the bioinformatics problems. *Bioinformatics* 36, 1542–1552. doi:10.1093/BIOINFORMATICS/BTZ763

Chiu, Y. C., Chen, H. I., Gorthi, A., Mostavi, M., Zheng, S., Huang, Y., et al. (2020). Deep learning of pharmacogenomics resources: Moving towards precision oncology. *Brief. Bioinform.* 21, 2066–2083. doi:10.1093/BIB/BBZ144

Chong, J., and Xia, J. (2018). MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. *Bioinformatics* 34, 4313–4314. doi:10.1093/BIOINFORMATICS/BTY528

Cortes, C., Vapnik, V., and Saitta, L. (1995). Support-vector networks. *Mach. Learn.* 320, 273–297. doi:10.1007/BF00994018

Costello, Z., and Martin, H. G. (2018). A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *NPJ Syst. Biol. Appl.* 4, 19–14. doi:10.1038/s41540-018-0054-3

Culley, C., Vijayakumar, S., Zampieri, G., and Angione, C. (2020). A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. *Proc. Natl. Acad. Sci. U. S. A.* 117, 18869–18879. doi:10.1073/pnas.2002959117

Date, Y., and Kikuchi, J. (2018). Application of a deep neural network to metabolomics studies and its performance in determining important variables. *Anal. Chem.* 90, 1805–1810. doi:10.1021/ACS.ANALCHEM.7B03795

Deisenroth, M. P., Faisal, A. A., and Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge, United Kingdom: Cambridge University Press. doi:10.1017/9781108679930

Dhamdhere, K., Sundararajan, M., and Yan, Q. (2018). *How important is a neuron?* arXiv. doi:10.48550/arXiv.1805.12233

Erhan, D., Bengio, Y., Courville, A., Ca, P. A. M., Ca, P. V., and Com, B. (2010). Why does unsupervised pre-training help deep learning? Pierre-antoine manzagol pascal vincent samy bengio. *J. Mach. Learn. Res.* 11, 625–660. doi:10.5555/1756006

Fonville, J. M., Richards, S. E., Barton, R. H., Boulange, C. L., Ebbels, T. M., Nicholson, J. K., et al. (2010). The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. *J. Chemom.* 24, 636–649. doi:10.1002/CEM.1359

Forsyth, D. (2019). *Applied machine learning*. Cham: Springer Nature Switzerland. doi:10.1007/978-3-030-18114-7

Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Statistics Data Analysis* 38, 367–378. doi:10.1016/S0167-9473(01)00065-2

Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (2005). *Bioinformatics and computational biology solutions using R and bioconductor*, 1. New York: Springer.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). "Neural message passing for quantum chemistry," in 34th international conference on machine learning, ICML 2017, Sydney, NSW, Australia. Editors D. Precup and Y. W. Teh (Bookline, MA: JMLR, Inc. and Microtome Publishing), 3, 1263–1272.

Gondara, L. (2016). Medical image denoising using convolutional denoising autoencoders. *IEEE Int. Conf. Data Min. Work. ICDMW* 0, 241–246. doi:10.1109/ICDMW.2016.0041

Grapov, D., Fahrmann, J., Wanichthanarak, K., and Khoomrung, S. (2018). Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. *OMICS A J. Integr. Biol.* 22, 630–636. doi:10.1089/omi.2018.0097

Guyon, I., and Elisseeff, A. (2006). *Feature extraction*, 207. Berlin, Heidelberg: Springer. doi:10.1007/978-3-540-35488-8

Hanin, B. (2019). Universal function approximation by deep neural nets with bounded width and ReLU activations. *Mathematics* 20197, 992992. doi:10.3390/MATH7100992

Hoehenwarter, W., Larhlimi, A., Hummel, J., Egelhofer, V., Selbig, J., Van Dongen, J. T., et al. (2011). MAPA distinguishes genotype-specific variability of highly similar regulatory protein isoforms in potato tuber. *J. Proteome Res.* 10, 2979–2991. doi:10.1021/PR101109A/ASSET/IMAGES/MEDIUM/PR-2010-01109A_0008.GIF

Hu, T., Oksanen, K., Zhang, W., Randell, E., Furey, A., Sun, G., et al. (2018). An evolutionary learning and network approach to identifying key metabolites for osteoarthritis. *PLoS Comput. Biol.* 14, e1005986. doi:10.1371/JOURNAL.PCBI.1005986

Isermann, R., and Münchhof, M. (2011). *Identification of dynamic systems: An introduction with applications*. Berlin, Heidelberg: Springer, 1–705. doi:10.1007/978-3-540-78879-9

Jiang, T., Gradus, J. L., and Rosellini, A. J. (2020). Supervised machine learning: A brief primer. *Behav. Ther.* 51, 675–687. doi:10.1016/J.BETH.2020.05.002

Kim, M., Rai, N., Zorraquino, V., and Tagkopoulos, I. (2016). Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli*. *Nat. Commun.* 7, 13090. doi:10.1038/ncomms13090

Kim, M., and Tagkopoulos, I. (2018). Data integration and predictive modeling methods for multi-omics datasets. *Mol. Omics* 14, 8–25. doi:10.1039/C7MO00051K

Koh, H. W., Fermin, D., Vogel, C., Choi, K. P., Ewing, R. M., and Choi, H. (2019). iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. *NPJ Syst. Biol. Appl.* 5, 22. doi:10.1038/S41540-019-0099-Y

Kotsiantis, S. B., Kanellopoulos, D., and Pintelas, P. E. (2007). *Data preprocessing for supervised leaning*. doi:10.5281/ZENODO.1082415

Kuhn, M., and Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. 1 edn. New York: CRC Press, 1–297. doi:10.1201/9781315108230

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi:10.1038/nature14539

Leitner, M., Fragner, L., Danner, S., Holeschofsky, N., Leitner, K., Tischler, S., et al. (2017). Combined metabolomic analysis of plasma and urine reveals AHBA, tryptophan and serotonin metabolism as potential risk factors in Gestational Diabetes Mellitus (GDM). *Front. Mol. Biosci.* 4, 84. doi:10.3389/FMOLB.2017.00084

Lipton, Z. C. (2016). The mythos of model interpretability. *Commun. ACM* 61, 36–43. doi:10.1145/3233231

Liu, J., Semiz, S., Van Der Lee, S. J., Van Der Spek, A., Verhoeven, A., Van Klinken, J. B., et al. (2017). Metabolomics based markers predict type 2 diabetes in a 14-year follow-up study. *Metabolomics* 1, 104. doi:10.1007/s11306-017-1239-2

Ljung, L. (1998). "System identification," in *Signal analysis and prediction*. Editors A. Prochazka, J. Uhlir, P. W. J. Rayner, and N. G. Kingsbury (Boston: Birkhäuser), 163–173. doi:10.1007/978-1-4612-1768-8_11

Loyola-Gonzalez, O. (2019). Black-box vs. White-Box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access* 7, 154096–154113. doi:10.1109/ACCESS.2019.2949286

Mendez, K. M., Reinke, S. N., David, Â., and Broadhurst, I. (2019). A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics* 15, 150. doi:10.1007/s11306-019-1612-4

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., et al. (2018). Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* 15, 290–298. doi:10.1038/nmeth.4627

Maceachern, S. J., and Forkert, N. D. (2021). Machine learning for precision medicine. *Genome* 64, 416–425.10.1139/GEN-2020-0131/ASSET/IMAGES/LARGE/GEN-2020-0131F1.JPEG

Macukow, B. (2016). "Neural networks-state of art, brief history, basic models and architecture," in *Computer information systems and industrial management*. Editors K. Saeed and W. Homenda (Cham: Springer), 9842, 3–14. doi:10.1007/978-3-319-45378-1_1

Manica, M., Oskooei, A., Born, J., Subramanian, V., Sáez-Rodríguez, J., and Martínez, M. R. (2019). Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Mol. Pharm.* 16, 4797–4806. doi:10.1021/ACS.MOLPHARMACEUT.9B00520

Martorell-Marugán, J., Siham Tabik, S., Benhammou, Y., Del Val, C., Zwir, I., Herrera, F., et al. (2019). in *Deep learning in omics data analysis and precision medicine. Computational biology*. Editor H. Husi (Brisbane City: Exon Publications), 37–53. doi:10.15586/COMPUTATIONALBIOLOGY.2019.CH3

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. U. S. A.* 116, 22071–22080. doi:10.1073/PNAS.1900654116

Nguyen, D. H., Nguyen, C. H., and Mamitsuka, H. (2019). Recent advances and prospects of computational methods for metabolite identification: A review with emphasis on machine learning approaches. *Brief. Bioinform.* 20, 2028–2043. doi:10.1093/BIB/BBY066

Nguyen, N. D., Jin, T., and Wang, D. (2021). Varmole: A biologically drop-connect deep neural network model for prioritizing disease risk variants and genes. *Bioinformatics* 37, 1772–1775. doi:10.1093/BIOINFORMATICS/BTAA866

Oh, J. H., Choi, W., Ko, E., Kang, M., Tannenbaum, A., and Deasy, J. O. (2021). PathCNN: Interpretable convolutional neural networks for survival prediction and pathway analysis applied to glioblastoma. *Bioinformatics* 37, i443–i450. doi:10.1093/BIOINFORMATICS/BTAB285

Pai, S., Hui, S., Isserlin, R., Shah, M. A., Kaka, H., and Bader, G. D. (2019). netDx: interpretable patient classification using integrated patient similarity networks. *Mol. Syst. Biol.* 15, e8497. doi:10.15252/MSB.20188497

Phillips, M., Cataneo, R. N., Chaturvedi, A., Kaplan, P. D., Libardoni, M., Mundada, M., et al. (2013). Detection of an extended human volatome with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry. *PloS one* 8, e75274. doi:10.1371/JOURNAL.PONE.0075274

Picart-Armada, S., Fernández-Albert, F., Vinaixa, M., Yanes, O., and Perera-Lluna, A. (2018). Fella: an R package to enrich metabolomics data. *BMC Bioinforma.* 19, 538–539. doi:10.1186/s12859-018-2487-5

Presnell, K. V., and Alper, H. S. (2019). Systems metabolic engineering meets machine learning: A new era for data-driven metabolic engineering. *Biotechnol. J.* 14, e1800416. doi:10.1002/BIOT.201800416

Reel, P. S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol. Adv.* 49, 107739. doi:10.1016/J.BIOTECHADV.2021.107739

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. *Adv. Neural Inf. Process. Syst.*, 3857–3867. doi:10.5555/3294996.3295142

Schwarzerova, J., Bajger, A., Pierdou, I., Popelinsky, L., Sedlar, K., and Weckwerth, W. (2021). "An innovative perspective on metabolomics data analysis in biomedical research using concept drift detection," in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX. Editors Y. Huang, L. A. Kurgan, F. Luo, X. Hu, Y. Chen, E. R. Dougherty, et al. (New York City, NY: Institute of Electrical and Electronics Engineers (IEEE)), 3075–3082. doi:10.1109/BIBM52615.2021.9669418

Sengupta, S., Basak, S., Saikia, P., Paul, S., Tsalavoutis, V., Atiah, F., et al. (2020). A review of deep learning with special emphasis on architectures, applications and recent trends. *Knowledge-Based Syst.* 194105596. doi:10.1016/J.KNOSYS.2020.105596

Sha, C., Cuperlovic-Culf, M., and Hu, T. (2021). Smile: Systems metabolomics using interpretable learning and evolution. *BMC Bioinforma.* 22, 284. doi:10.1186/S12859-021-04209-1

Shalev-Shwartz, S., and Ben-David, S. (2013). *Understanding machine learning: From theory to algorithms*, 9781107057135. Cambridge, United Kingdom: Cambridge University Press, 1–397. doi:10.1017/CBO9781107298019

Sharma, A., Vans, E., Shigemizu, D., Boroevich, K. A., and Tsunoda, T. (2019). DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Sci. Rep.* 9, 11399. doi:10.1038/s41598-019-47765-6

Shrestha, A., and Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE Access* 7, 53040–53065. doi:10.1109/ACCESS.2019.2912200

Simonoff, J. S. (1996). *Smoothing methods in statistics. Springer series in statistics*. New York: Springer. doi:10.1007/978-1-4612-4026-6

Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P. Y., et al. (1995). Nonlinear black-box modeling in system identification: A unified overview. *Automatica* 31, 1691–1724. doi:10.1016/0005-1098(95)00120-8

Srinath, K. (2017). Python–the fastest growing programming language. *Int. Res. J. Eng. Technol. (IRJET)* 4, 354–357.

Stamate, D., Kim, M., Proitsi, P., Westwood, S., Baird, A., Nevado-Holgado, A., et al. (2019). A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: Results from the European Medical Information Framework for Alzheimer disease biomarker discovery cohort. *Alzheim. Dement. Translat. Res. Clin. Intervent.* 5 (1), 933–938. doi:10.1016/j.trci.2019.11.001

Sundararajan, M., Taly, A., and Yan, Q. (2017). "Axiomatic attribution for deep networks," in 34th international conference on machine learning, ICML 2017, Sydney, NSW, Australia. Editors D. Precup and Y. W. Teh (Bookline, MA: JMLR, Inc. and Microtome Publishing), 7, 3319–3328.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 99, 6567–6572. doi:10.1073/PNAS.082099299

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58, 267–288. doi:10.1111/J.2517-6161.1996.TB02080.X

Torsten Hothorn (2022). CRAN Task View: Machine Learning & Statistical Learning. Version 2022-03-07. Available at: https://CRAN.R-project.org/view=MachineLearning (Accessed June 29, 2022).

Toubiana, D., Puzis, R., Wen, L., Sikron, N., Kurmanbayeva, A., Soltabayeva, A., et al. (2019). Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data. *Commun. Biol.* 2, 214. doi:10.1038/s42003-019-0440-4

Trainor, P. J., de Filippis, A. P., and Rai, S. N. (2017). Evaluation of classifier performance for multiclass phenotype discrimination in untargeted metabolomics. *Metabolites* 7, E30. doi:10.3390/METABO7020030

van Dooijeweert, B., Broeks, M. H., van Beers, E. J., Verhoeven-Duif, N. M., van Solinge, W. W., Nieuwenhuis, E. E., et al. (2021). Dried blood spot metabolomics reveals a metabolic fingerprint with diagnostic potential for Diamond Blackfan Anaemia. *Br. J. Haematol.* 193, 1185–1193. doi:10.1111/BJH.17524

Vikalo, H., Parvaresh, F., and Hassibi, B. (2007). "On recovery of sparse signals in compressed DNA microarrays," in Conference Record - Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA (IEEE), 693–697. doi:10.1109/ACSSC.2007.4487303

Wang, L., Miao, X., Nie, R., Zhang, Z., Zhang, J., and Cai, J. (2021). MultiCapsNet: A general framework for data integration and interpretable classification. *Front. Genet.* 12, 767602. doi:10.3389/fgene.2021.767602

Wang, L., Nie, R., Yu, Z., Xin, R., Zheng, C., Zhang, Z., et al. (2020). An interpretable deep-learning architecture of capsule networks for identifying cell-type gene expression programs from single-cell RNA-sequencing data. *Nat. Mach. Intell.* 2, 693–703. doi:10.1038/s42256-020-00244-4

Weckwerth, W. (2011). Unpredictability of metabolism-the key role of metabolomics science in combination with next-generation genome sequencing. *Anal. Bioanal. Chem.* 400, 1967–1978. doi:10.1007/s00216-011-4948-9

Wold, H. (1975). "Path models with latent variables: The NIPALS approach," in *Quantitative sociology*. Editors H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, and V. Capecchi (Cambridge, Massachusetts: Academic Press), 307–357. doi:10.1016/B978-0-12-103950-9.50017-4

Wolpert, D. H., and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* 1, 67–82. doi:10.1109/4235.585893

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2019). A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 4–24. doi:10.1109/TNNLS.2020.2978386

Yang, J. H., Wright, S. N., Hamblin, M., McCloskey, D., Alcantar, M. A., Schrübbers, L., et al. (2019). A white-box machine learning approach for revealing antibiotic mechanisms of action. *Cell* 177, 1649–1661. doi:10.1016/J.CELL.2019.04.016

Zhang, X., Xing, Y., Sun, K., and Guo, Y. (2021). OmiEmbed: A unified multi-task deep learning framework for multi-omics data. *Cancers* 13, 3047. doi:10.3390/CANCERS13123047

Zhang, Z., Zhao, Y., Liao, X., Shi, W., Li, K., Zou, Q., et al. (2019). Deep learning in omics: A survey and guideline. *Brief. Funct. Genomics* 18, 41–57. doi:10.1093/BFGP/ELY030

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., et al. (2018). Graph neural networks: A review of methods and applications. *AI Open* 1, 57–81. doi:10.1016/j.aiopen.2021.01.001

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 16 November 2022
DOI 10.3389/fmolb.2022.953189

Metabolic diversity in a collection of wild and cultivated *Brassica rapa* subspecies

Shuning Zheng[1], Jędrzej Szymański[1,2], Nir Shahaf[1], Sergey Malitsky[1], Sagit Meir[1], Xiaowu Wang[3], Asaph Aharoni[1] and Ilana Rogachev[1]*

[1]Department of Plant and Environmental Sciences, Weizmann Institute of Science, Rehovot, Israel, [2]Department of Molecular Genetics, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany, [3]Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China

OPEN ACCESS

EDITED BY
Wolfram Weckwerth,
University of Vienna, Austria

REVIEWED BY
Taoyang Wu,
University of East Anglia,
United Kingdom
Jian Chen,
Jiangsu Academy of Agricultural Sciences, China

*CORRESPONDENCE
Ilana Rogachev,
ilana.rogachev@weizmann.ac.il

SPECIALTY SECTION
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 25 May 2022
ACCEPTED 29 August 2022
PUBLISHED 16 November 2022

CITATION
Zheng S, Szymański J, Shahaf N, Malitsky S, Meir S, Wang X, Aharoni A and Rogachev I (2022), Metabolic diversity in a collection of wild and cultivated Brassica rapa subspecies.
Front. Mol. Biosci. 9:953189.
doi: 10.3389/fmolb.2022.953189

Brassica rapa (*B. rapa*) and its subspecies contain many bioactive metabolites that are important for plant defense and human health. This study aimed at investigating the metabolite composition and variation among a large collection of *B. rapa* genotypes, including subspecies and their accessions. Metabolite profiling of leaves of 102 *B. rapa* genotypes was performed using ultra-performance liquid chromatography coupled with a photodiode array detector and quadrupole time-of-flight mass spectrometry (UPLC-PDA-QTOF-MS/MS). In total, 346 metabolites belonging to different chemical classes were tentatively identified; 36 out of them were assigned with high confidence using authentic standards and 184 were those reported in *B. rapa* leaves for the first time. The accumulation and variation of metabolites among genotypes were characterized and compared to their phylogenetic distance. We found 47 metabolites, mostly representing anthocyanins, flavonols, and hydroxycinnamic acid derivatives that displayed a significant correlation to the phylogenetic relatedness and determined four major phylometabolic branches; 1) Chinese cabbage, 2) yellow sarson and rapid cycling, 3) the mizuna-komatsuna-turnip-caitai; and 4) a mixed cluster. These metabolites denote the selective pressure on the metabolic network during *B. rapa* breeding. We present a unique study that combines metabolite profiling data with phylogenetic analysis in a large collection of *B. rapa* subspecies. We showed how selective breeding utilizes the biochemical potential of wild *B. rapa* leading to highly diverse metabolic phenotypes. Our work provides the basis for further studies on *B. rapa* metabolism and nutritional traits improvement.

KEYWORDS

***Brassica rapa*, metabolite profiling, mass spectrometry, multivariate data analysis, phylogenetic analysis, specialized metabolism, selective breeding**

1 Introduction

Brassica rapa (*B. rapa*) is an economically important crop species of the genus *Brassica* and is widely cultivated and consumed worldwide. During the long history of selective breeding, it reached an enormous morphological diversity and a wide range of useful purposes, including leafy vegetables (e.g., Chinese cabbage, pak choi, and mizuna), inflorescence vegetables (e.g., caixin and broccoletto), floral shoot and stem vegetables (e.g., purple caitai and turnip top), enlarged root vegetables or fodders (e.g., turnip), as well as oilseed crops (e.g., yellow sarson). Due to its strong adaptability, short growth period, high yield, unique flavor, and nutritional benefits, *B. rapa* is increasingly popular worldwide (Salehi et al., 2021).

Brassica vegetables have been widely acknowledged for their beneficial effects on human health. Epidemiological studies have indicated that increased consumption of *Brassica* vegetables is strongly associated with a reduced risk of cancer, cardiovascular disease, diabetes, and immune dysfunction (Raiola et al., 2018; Salehi et al., 2021). These health-related properties have been attributed to nutrients and health-promoting phytochemicals, such as *Brassica*-specific glucosinolates, carotenoids, vitamins, and phenolic compounds (Paul et al., 2019). Glucosinolates and their breakdown products have been reported to reduce the

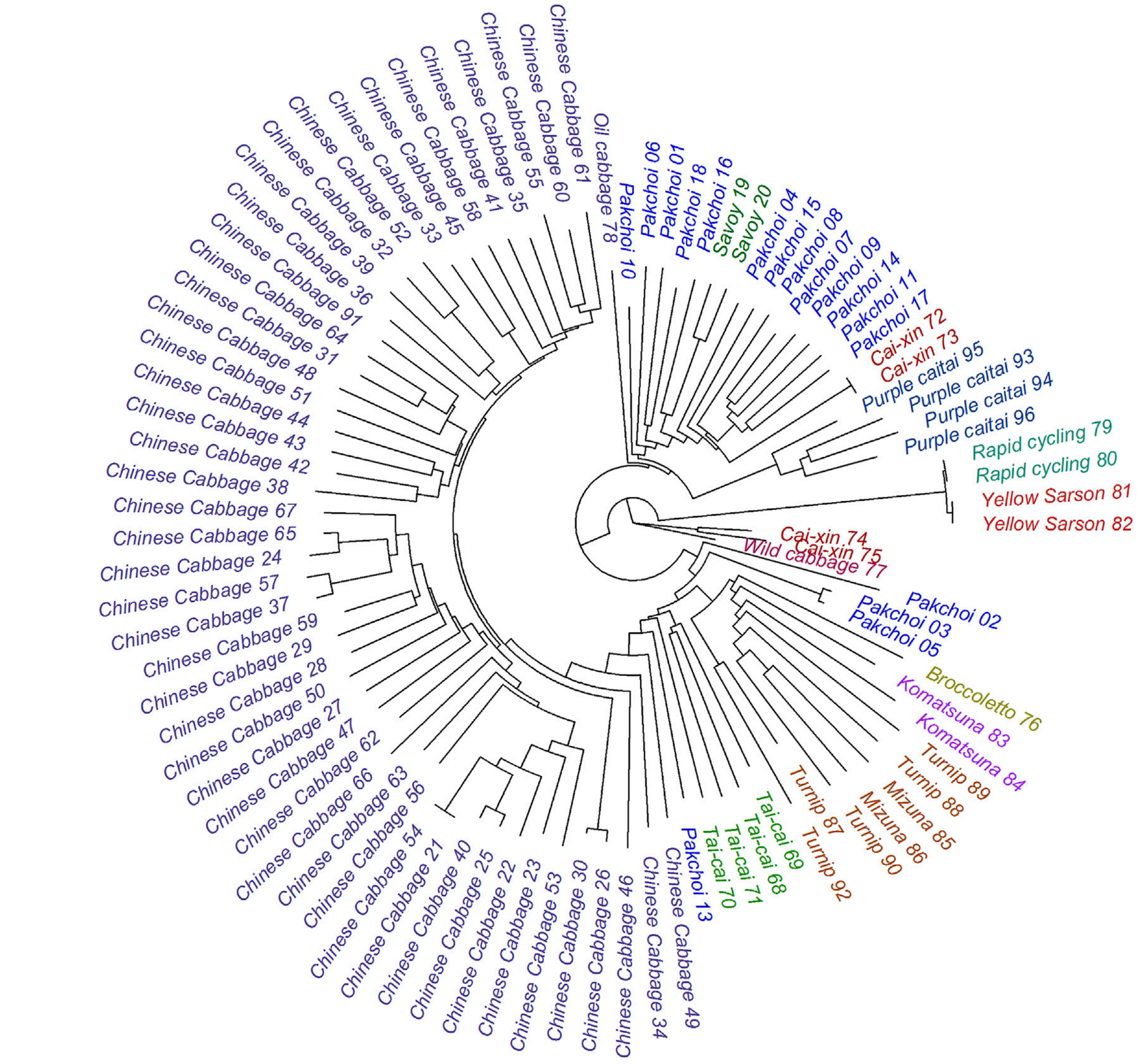


FIGURE 1
The phylogenetic distance tree of the 102 analyzed *B. rapa* accessions based on the DNA sequence variation—SNPs with MAF >0.05 as described by Cheng et al. (2016). The tree is rooted in the wild-type cabbage genotype. Classification of the sub-species is represented by the color code.

TABLE 1 Summary of the 102 *B. rapa* accessions in this study.

Accession name	Subspecies	Sample number	Samples in total
Chinese cabbage	ssp. *Pekinensis*	#21–#67, #91	48
Pak choi	ssp. *Chinensis*	#01–#18	18
Purple caitai	ssp. *Chinensis* var. *Purpurea*	#93–#102	10
Turnip	ssp. *Rapa*	#87–#90, #92	5
Caixin	ssp. *Parachinensis*	#72–#75	4
Taicai	ssp. *Chinensis* var. *Tai-tsai*	#68–#71	4
Savoy	ssp. *Narinosa*	#19–#20	2
Rapid cycling		#79–#80	2
Yellow sarson	ssp. *Tricolaris*	#81–#82	2
Komatsuna	var. *Pervidis*	#83–#84	2
Mizuna	ssp. *Nipposinica*	#85–#86	2
Broccoletto	ssp. *Broccoletto*	#76	1
Wild cabbage		#77	1
Oil cabbage		#78	1

risk of lung, colon, and other types of cancer (Mandrich and Caputo, 2020). Phenolic compounds in plants possess potential health-promoting effects, including antioxidant, anti-inflammatory, anti-microbial, anti-obesity, and anti-tumour activities (Cao et al., 2021).

The potential activity and bioavailability of dietary phytochemicals in *B. rapa* depend on the chemical structure, modifications, and content. Most previous studies on *B. rapa* have focused on specific classes of targeted compounds, such as glucosinolates, organic acids, or phenolic compounds. For example, glucosinolate profiles in different *B. rapa* varieties have been reported (Liu et al., 2020; Zou et al., 2021). Phenolic compounds have been investigated in turnip (Chihoub et al., 2019), Chinese cabbage (Managa et al., 2020), pak choi (Jeon et al., 2018; Yeo et al., 2021), and mizuna (Kyriacou et al., 2021), establishing flavonoids and hydroxycinnamic acids as main phenolic compounds. However, the morphological, flavor, and taste diversity of *B. rapa*, suggest much wider metabolic complexity, potentially including new interesting compounds and biochemistry.

In the present study, we performed comprehensive metabolic characterization of 102 representative *B. rapa* genotypes, covering 14 subspecies and their individual accessions exhibiting a wide variety of morphological traits (Figure 1). Clustering analysis revealed similarities of accessions and metabolites in metabolic composition and chemical structure, respectively. Furthermore, we carried out phylogenetic analysis and assessed the relationship between metabolic composition and genetic relatedness of various *B. rapa* accessions. This highlighted the biochemical effects of selective breeding of *B. rapa*.

2 Materials and methods

2.1 Chemicals

All solvents were of HPLC grade. Methanol and acetonitrile were purchased from Merck KGaA (Darmstadt, Germany). Formic acid was purchased from J.T. Baker (Germany). Ultrapure water was produced using a Milli-Q water purification system (Millipore, Bedford, MA, United States).

2.2 Plant material

We selected 102 representative *B. rapa* genotypes belonging to 14 main *B. rapa* subspecies groups for metabolite profiling analysis, including accessions of Chinese cabbage, pak choi, caixin, turnip, savoy, mizuna, taicai, komatsuna, purple caitai, rapid cycling, yellow sarson, broccoletto, oil cabbage, and wild cabbage (see Table 1 and Figure 1). All of selected *B. rapa* accessions were previously genotyped (Cheng et al., 2016), and SNPs with MAF (minor allele frequency) > 0.05 were retrieved, as described by Cheng et al. (2016). Leaf samples were obtained from the Institute of Vegetable and Flowers, Chinese Academy of Agricultural Sciences (IVF-CAAS, Beijing, China). All plants were cultivated in a greenhouse under the same growth conditions in the fall of 2012. Fifty days after seeding, two or three fresh leaves (about 15–20 g) of uniform size and free from decay and mechanical damage were harvested, snap-frozen in liquid nitrogen, and lyophilized. The freeze-dried samples were then ground into fine powder and stored at −80°C until further analysis. Three biological replicates were taken for each genotype.

2.3 Metabolite extraction and sample preparation

Powdered *B. rapa* leaf material (200 mg) was extracted with 80% (v/v) aqueous methanol containing 0.1% (v/v) formic acid by 20 min sonication at room temperature. The extract was centrifuged at 13,000 g for 15 min, and the supernatant was filtered through a 0.22-μm syringe PVDF filter and transferred to an HPLC vial for LC-MS analysis.

2.4 UPLC-PAD-QTOF-MS/MS analyses

Non-targeted metabolite analysis was performed on a UPLC-qTOF system (Waters Synapt) with the UPLC column connected in-line to a PDA detector and then to the MS detector (Synapt, Water Corp, Manchester, United Kingdom) equipped with electrospray ionization (ESI) source. Chromatographic separation was carried out using an UPLC BEH C18 column (100 × 2.1 mm i. d, 1.7 μm, Waters Acquity). The mobile phase consisted of two solvents: 0.1% formic acid in acetonitrile/water (5:95, v/v) (A) and 0.1% formic acid in acetonitrile (B). The linear gradient program was as follows: 100%–72% A over 22 min, 72%–60% A over 0.5 min, 60%–0% A over 0.5 min, held at 100% B for a further 1.5 min, then returned to the initial conditions (100% A) in 0.5 min and conditioning at 100% A for 1 min. The flow rate was 0.3 ml min^{-1}, and the column temperature was maintained at 35 °C. The injection volume was 4 μl. UV-vis spectra were recorded in the range of 210–500 nm.

The MS conditions were as follows: capillary voltage of 3.0 kV, cone voltage of 28 V, source temperature of 125°C, desolvation temperature of 275°C, desolvation gas flow rate of 650 L h^{-1}, and cone gas flow rate of 25 L h^{-1}. Nitrogen was used as desolvation and cone gas, and argon was utilized as the collision gas. Data were acquired in MS^E mode from *m/z* 50 to 1,500 in centroid mode at negative ion mode, comprising two interleaved full scan acquisition functions: the low energy function and the high energy function. The low energy function employed collision energy at 4 eV to acquire accurate mass data for intact precursor ions. For the high energy function, a collision energy ramp of 10–35 eV was applied for fragmentation information. The MS system was calibrated using sodium formate. Leucine enkephalin was used as a reference lock-mass compound to ensure mass accuracy. The $[M-H]^-$ ion at *m/z* 554.2615 was detected via the independent LockSpray™ channel. A mixture of 15 standard compounds, injected after each batch of 10 biological samples, was used for instrument quality control. MassLynx software version 4.1 (Waters) was used to control the instrument and calculate accurate masses.

2.5 Data processing and statistical analysis

LC-MS raw data files were converted to NetCDF format using MassLynx DataBridge (version 4.1; Waters Corp.). Peak picking, retention time correction, and alignment were then performed using R packages XCMS (Smith et al., 2006) and CAMERA (Kuhl et al., 2012). Data normalization, analysis, and visualization were performed using R 3.1.2 (Ihaka and Gentleman, 1996). The relative peak intensities were normalized to the median intensity of each chromatogram and subsequently scaled between the minimum non-zero and the maximum value of the original dataset. Hierarchical clustering analysis (HCA) in heat map was performed using Euclidean distance and average linkage on Z-transformed variables (either rows/metabolites or columns/samples). Significant differences in the accumulation of metabolites in measured accessions were identified by one-way ANOVA (Ritchie et al., 2015) with FDR ≤0.05 and followed by Tukey's HSD test (Supplementary Table S2).

2.6 Phylogenetic distance analysis

The neighbor-joining tree was constructed by the BioNJ algorithm (Gascuel, 1997) using J-C distance in PHYLIP 3.6 software (Felsenstein, 2004) and all SNPs with MAF >0.05. The tree was rooted using a midpoint method (Farris, 1972). Phylogenetic signal was computed in a "picante" R package (v1.8.2 2020; Kembel et al., 2010) using Blomberg's K statistics (Blomberg et al., 2003) on the background of a Brownian motion model of the trait evolution. The significance of the phylogenetic signal was obtained in 9,999 random permutations of the phylogenetic tree labels (Supplementary Table S2).

2.7 Clustering of metabolites

For all metabolites SMILES codes have been obtained and translated to the standard molecular fingerprints as described by Faulon et al. (2003) using the rcdk R package (v3.6.0 2021; Guha, 2007). The Tanimoto similarity has been calculated according to the method of Fligner et al. (2002) and the result has been displayed as a hierarchical clustering tree using the complete agglomerative linkage method. Metabolites sharing the same fingerprint have been grouped and treated as one single compound.

3 Result and discussion

3.1 Strategy for metabolite identification

To comprehensively characterize the metabolome of *B. rapa* leaves, a total of 102 representative *B. rapa* genotypes, belonging to 14 major *B. rapa* subspecies groups, were selected to cover their large genetic and phenotypic variations (Figure 1).

Untargeted metabolite analysis was performed using UPLC-PDA-QTOF-MS/MS and representative total ion chromatograms (TIC) of five *B. rapa* accessions are shown in Supplementary Figure S1. Out of 7,286 quantified mass features, a total of 346 metabolites were identified at different confidence levels. In our study, two metabolite identification strategies were used: 1) high-confidence metabolite identification based on authentic standards and 2) putative identification based on literature and public databases. In total, 37 metabolites were identified with the "high confidence" strategy through comparison of retention time (RT), UV/Vis spectra, accurate mass, isotopic distribution, and fragmentation pattern with those of authentic standards using the WEIZMASS library, a reference spectral library comprising spectra of 3,540 highly pure plant metabolites (Shahaf et al., 2016). Of these, 13 metabolites were identified for the first time in *B. rapa*. Respectively, 309 metabolites were putatively identified in *B. rapa* leaves by surveying the literature and public databases (KNapSack, DNP, Massbank, KEGG, and ReSpect). Metabolites previously reported in the Brassicaceae family were collected in a custom reference database that included metabolite names, molecular formulas, molecular weight, chemical structures, biological sources, and literature or database resources. Mass features following XCMS and Camera clustering were first searched against this reference database using a homemade script. The accurate mass of the molecule and adduct ions as well as their isotope distribution patterns were considered as main search parameters. Next, the structural information from UV/vis spectra and mass fragmentation patterns of the hits were used for putative identification. In this study, a large number of metabolite isomers were identified and discriminated in *B. rapa* leaves based on retention time and/or MS fragments (see Supplementary Material).

3.2 Chemical complexity of the metabolic profiles

The 346 putatively identified metabolites belong to various chemical classes, including 105 flavonols, 93 hydroxycinnamic acid derivatives, 51 monolignol and oligolignol derivatives, 33 glucosinolates, 14 anthocyanins, 10 organic acid, 8 indolics, 5 benzenoids, 3 amino acids, and 24 others. These metabolites are mostly products and intermediates of specialized metabolism pathways associated with nutritional and health-promoting effects of *B. rapa* as well as flavor and aroma (Salehi et al., 2021). To our knowledge, 184 of the detected metabolites were identified in *B. rapa* for the first time. The complete list of all identified metabolites with respective chemical, analytical, and biological descriptors is provided in Supplementary Table S1. Tanimoto similarity analysis showed that the identified metabolites are linked to twelve major clusters based on their chemical structure (Figure 2). Expectedly, most of these clusters were enriched by a specific class of metabolites. However, the non-biased grouping highlighted also four clusters containing metabolites of diverse classes (clusters I, J, K, and L) and included mostly precursors and intermediates upstream in the major pathways of specialized metabolism.

3.3 Characterization of major specialized metabolite classes in *B. rapa*

All putatively identified metabolites were categorized into ten major chemical classes based on structures and fragmentation patterns. Both the composition and relative content of metabolites varied significantly among measured accessions (Supplementary Table S2), indicating the impact of genetic diversity on the metabolic variation within the *B. rapa* species.

3.3.1 Glucosinolates

Glucosinolates are a group of nitrogen- and sulfur-containing specialized metabolites and are classified into aliphatic, aromatic, and indole glucosinolates, according to whether they originate from aliphatic amino acids, aromatic amino acids or tryptophan, respectively. Glucosinolates contain a β-D- glucopyranosyl common core moiety and a variable side chain. Several studies presented the typical MS fragmentation of glucosinolates (Fabre et al., 2007; Francisco et al., 2009). First, based on the common core structure, glucosinolate could produce characteristic fragments at *m/z* 96.96, 195.03, 241.00, 259.01, and 274.99 via the cleavage of bonds on either side of the sulfur atoms. However, not all fragments could always be observed in MS^E fragmentation. We used the most abundant fragment ions at *m/z* 96.96 (sulfate anion) and *m/z* 259.01 (sulfated glucose anion) as diagnostic ions to preliminarily check the presence of glucosinolates. In addition, glucosinolates undergo consistent and characteristic neutral losses of sulfur trioxide (SO_3, 79.96 amu), anhydroglucose (Glc, 162.05 amu), dehydroxythioglucose (SGlc-OH, 178.03 amu), thioglucose (SGlc, 196.04 amu) as well as combined loss of sulfur trioxide and anhydroglucose (Glc + SO_3, 242.01 amu), parameters that could be used for identification of variable side chains. Finally, the variable side chain could also produce unique fragments. For example, glucohesperin is an aliphatic glucosinolate with a deprotonated molecular ion at *m/z* 464.07 and formula as C14H27NO10S3. The MS fragments showed characteristic fragment ions at *m/z* 79.95, 274.99, 259.01, 241.00, and 195.03. Neutral loss fragments from deprotonated molecular were observed at *m/z* 449.04 (loss of a methyl moiety from the side chain, −15.03 amu), *m/z* 384.11 (loss SO3, −79.96 amu), *m/z* 226.06 (loss Glc + SO3, −242.01 amu) and *m/z* 400.07 (loss of a methylsulfinyl moiety from the side chain, −64.00 amu). In addition, the dimethylsulfinyl fragment ion (*m/z* 400.07) underwent further neutral loss to give the product ions at *m/z*

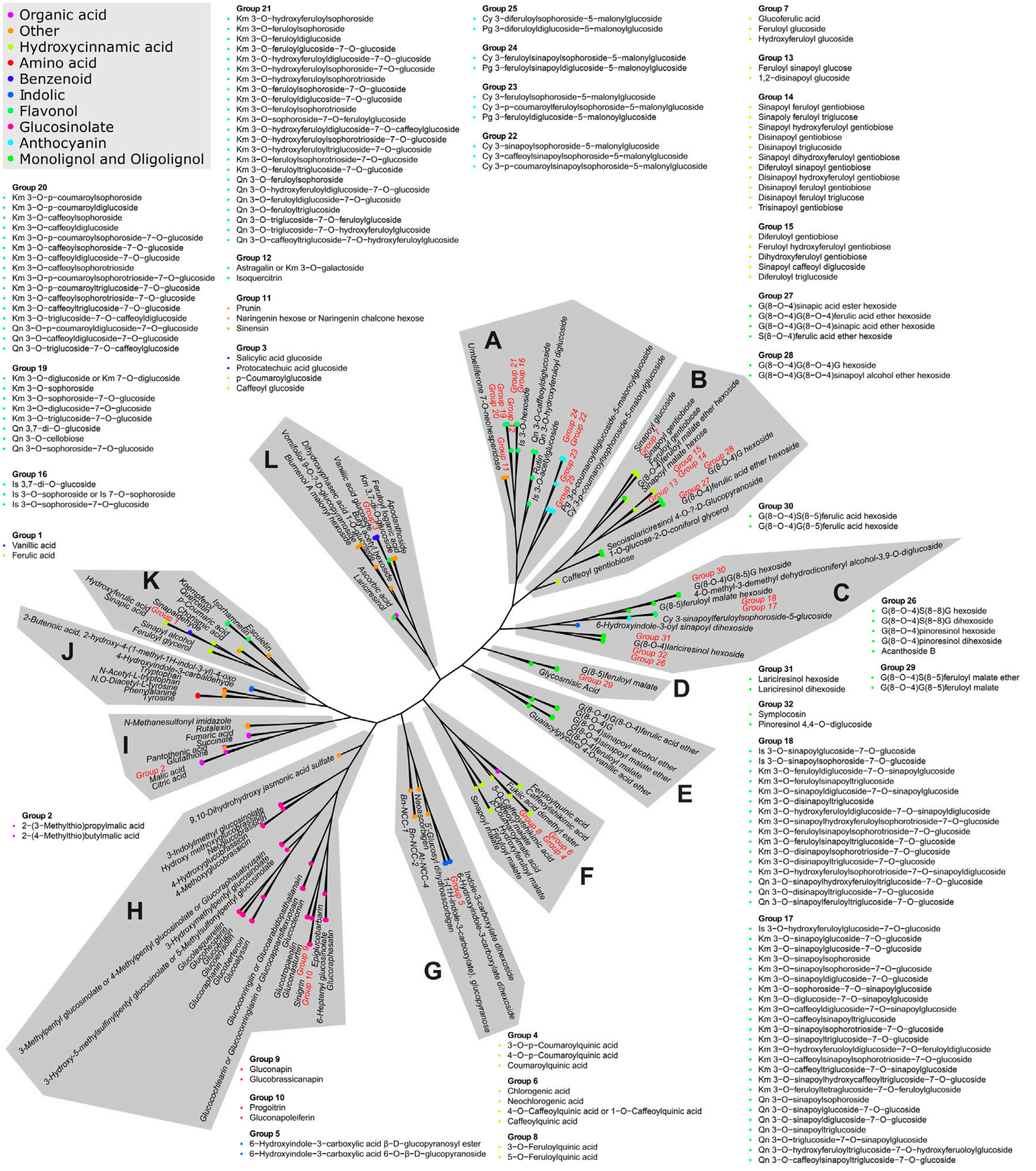

Organic acid
Other
Hydroxycinnamic acid
Amino acid
Benzenoid
Indolic
Flavonol
Glucosinolate
Anthocyanin
Monolignol and Oligolignol
Group 20
Km 3−O−p−coumaroylsophoroside
Km 3−O−p−coumaroyldiglucoside
Km 3−O−caffeoylsophoroside
Km 3−O−caffeoyldiglucoside
Km 3−O−p−coumaroylsophoroside−7−O−glucoside
Km 3−O−caffeoylsophoroside−7−O−glucoside
Km 3−O−caffeoyldiglucoside−7−O−glucoside
Km 3−O−caffeoylsophorotrioside
Km 3−O−p−coumaroylsophorotrioside−7−O−glucoside
Km 3−O−p−coumaroyltriglucoside−7−O−glucoside
Km 3−O−caffeoylsophorotrioside−7−O−glucoside
Km 3−O−caffeoyltriglucoside−7−O−glucoside
Km 3−O−triglucoside−7−O−caffeoyldiglucoside
Qn 3−O−p−coumaroyldiglucoside−7−O−glucoside
Qn 3−O−caffeoyldiglucoside−7−O−glucoside
Qn 3−O−triglucoside−7−O−caffeoylglucoside
Group 19
Km 3−O−diglucoside or Km 7−O−diglucoside
Km 3−O−sophoroside
Km 3−O−sophoroside−7−O−glucoside
Km 3−O−diglucoside−7−O−glucoside
Km 3−O−triglucoside−7−O−glucoside
Qn 3,7−di−O−glucoside
Qn 3−O−cellobiose
Qn 3−O−sophoroside−7−O−glucoside
Group 16
Is 3,7−di−O−glucoside
Is 3−O−sophoroside or Is 7−O−sophoroside
Is 3−O−sophoroside−7−O−glucoside
Group 1
Vanillic acid
Ferulic acid
Group 21
Km 3−O−hydroxyferuloylsophoroside
Km 3−O−feruloylsophoroside
Km 3−O−feruloyldiglucoside
Km 3−O−feruloylglucoside−7−O−glucoside
Km 3−O−hydroxyferuloyldiglucoside−7−O−glucoside
Km 3−O−hydroxyferuloylsophoroside−7−O−glucoside
Km 3−O−hydroxyferuloylsophorotrioside
Km 3−O−feruloylsophoroside−7−O−glucoside
Km 3−O−feruloyldiglucoside−7−O−glucoside
Km 3−O−feruloylsophorotrioside
Km 3−O−sophoroside−7−O−feruloylglucoside
Km 3−O−hydroxyferuloyldiglucoside−7−O−caffeoylglucoside
Km 3−O−hydroxyferuloylsophorotrioside−7−O−glucoside
Km 3−O−hydroxyferuloyltriglucoside−7−O−glucoside
Km 3−O−feruloylsophorotrioside−7−O−glucoside
Km 3−O−feruloyltriglucoside−7−O−glucoside
Qn 3−O−feruloylsophoroside
Qn 3−O−hydroxyferuloyldiglucoside−7−O−glucoside
Qn 3−O−feruloyldiglucoside−7−O−glucoside
Qn 3−O−feruloyltriglucoside
Qn 3−O−triglucoside−7−O−feruloylglucoside
Qn 3−O−triglucoside−7−O−hydroxyferuloylglucoside
Qn 3−O−caffeoyltriglucoside−7−O−hydroxyferuloylglucoside
Group 12
Astragalin or Km 3−O−galactoside
Isoquercitrin
Group 11
Prunin
Naringenin hexose or Naringenin chalcone hexose
Sinensin
Group 3
Salicylic acid glucoside
Protocatechuic acid glucoside
p−Coumaroylglucoside
Caffeoyl glucoside
Group 25
Cy 3−diferuloylsophoroside−5−malonylglucoside
Pg 3−diferuloyldiglucoside−5−malonoylglucoside
Group 24
Cy 3−feruloylsinapoylsophoroside−5−malonylglucoside
Pg 3−feruloylsinapoyldiglucoside−5−malonoylglucoside
Group 23
Cy 3−feruloylsophoroside−5−malonylglucoside
Cy 3−p−coumaroylferuloylsophoroside−5−malonylglucoside
Pg 3−feruloyldiglucoside−5−malonoylglucoside
Group 22
Cy 3−sinapoylsophoroside−5−malonylglucoside
Cy 3−caffeoylsinapoylsophoroside−5−malonylglucoside
Cy 3−p−coumaroylsinapoylsophoroside−5−malonylglucoside
Group 7
Glucoferulic acid
Feruloyl glucoside
Hydroxyferuloyl glucoside
Group 13
Feruloyl sinapoyl glucose
1,2−disinapoyl glucoside
Group 14
Sinapoyl feruloyl gentiobiose
Sinapoly feruloyl triglucose
Sinapoyl hydroxyferuloyl gentiobiose
Disinapoyl gentiobiose
Disinapoyl triglucoside
Sinapoyl dihydroxyferuloyl gentiobiose
Diferuloyl sinapoyl gentiobiose
Disinapoyl hydroxyferuloyl gentiobiose
Disinapoyl feruloyl gentiobiose
Disinapoyl feruloyl triglucose
Trisinapoyl gentiobiose
Group 15
Diferuloyl gentiobiose
Feruloyl hydroxyferuloyl gentiobiose
Dihydroxyferuloyl gentiobiose
Sinapoyl caffeoyl diglucoside
Diferuloyl triglucoside
Group 27
G(8−O−4)sinapic acid ester hexoside
G(8−O−4)G(8−O−4)ferulic acid ether hexoside
G(8−O−4)G(8−O−4)sinapic acid ether hexoside
S(8−O−4)ferulic acid ether hexoside
Group 28
G(8−O−4)G(8−O−4)G hexoside
G(8−O−4)G(8−O−4)sinapoyl alcohol ether hexoside
A
Umbelliferone 7-O-neohesperidose
Group 11
Group 19
Group 20
Group 12
Group 21
Group 16
Is 3-O-hexoside
Is 3-O-acetylglucoside
Rutin
Qn 3-O-caffeoyldiglucoside
Qn 3-O-hydroxyferuloyl diglucoside
Group 25
Group 23
Group 24
Group 22
Pg 3-p-coumaroyldiglucoside-5-malonoylglucoside
Cy 3-p-coumaroylsophoroside-5-malonylglucoside
B
Sinapoyl glucoside
Group 7
Sinapoyl gentiobiose
Feruloyl gentiobiose
G(8-O-4)feruloyl malate ether hexoside
Sinapoyl malate hexose
Group 15
Group 14
Group 13
Group 28
G(8-O-4)G hexoside
Group 27
G(8-O-4)ferulic acid ether hexoside
Secoisolariciresinol 4-O-?-D-Glucopyranoside
1-O-glucose-2-O-coniferol glycerol
Caffeoyl gentiobiose
Group 30
G(8-O-4)G(8-5)G hexoside
4-O-methyl-3-demethyl dehydrodiconiferyl alcohol-3,9-O-diglucoside
G(8-5)feruloyl malate hexoside
Group 18
Group 17
Cy 3-sinapoylferuloylsophoroside-5-glucoside
6-Hydroxyindole-3-oyl sinapoyl dihexoside
Group 31
G(8-O-4)lariciresinol hexoside
Group 32
Group 26
C
Group 30
G(8−O−4)S(8−5)ferulic acid hexoside
G(8−O−4)G(8−5)ferulic acid hexoside
Group 26
G(8−O−4)S(8−8)G hexoside
G(8−O−4)S(8−8)G dihexoside
G(8−O−4)pinoresinol hexoside
G(8−O−4)pinoresinol dihexoside
Acanthoside B
Group 31
Lariciresinol hexoside
Lariciresinol dihexoside
Group 29
G(8−O−4)S(8−5)feruloyl malate ether
G(8−O−4)G(8−5)feruloyl malate
Group 32
Symplocosin
Pinoresinol 4,4−O−diglucoside
G(8-5)feruloyl malate
Group 29
Glycosmisic Acid
D
G(8-O-4)G(8-O-4)ferulic acid ether
G(8-O-4)G
G(8-O-4)sinapoyl alcohol ether
G(8-O-4)sinapoyl malate ether
G(8-O-4)feruloyl malate
Guaiacylglycerol 4-O-vanillic acid ether
E
Feruloylquinic acid
Caffeoylshikimic acid
Fukiic acid dimethyl ester
Group 6
Group 4
5-O-Caffeoylshikimic acid
Group 8
Caffeoyl malate
p-Coumaroylmalic acid
Hydroxyferuloyl malate
Feruloyl malate
Sinapoyl malate
F
Group 18
Is 3−O−sinapoylglucoside−7−O−glucoside
Is 3−O−sinapoylsophoroside−7−O−glucoside
Km 3−O−feruloyldiglucoside−7−O−sinapoylglucoside
Km 3−O−feruloylsinapoyltriglucoside
Km 3−O−sinapoyldiglucoside−7−O−sinapoylglucoside
Km 3−O−disinapoyltriglucoside
Km 3−O−hydroxyferuloyldiglucoside−7−O−sinapoyldiglucoside
Km 3−O−sinapoylhydroxyferuloylsophorotrioside−7−O−glucoside
Km 3−O−feruloylsinapoylsophorotrioside−7−O−glucoside
Km 3−O−feruloylsinapoyltriglucoside−7−O−glucoside
Km 3−O−disinapoylsophorotrioside−7−O−glucoside
Km 3−O−disinapoyltriglucoside−7−O−glucoside
Km 3−O−hydroxyferuloylsophorotrioside−7−O−sinapoyldiglucoside
Qn 3−O−sinapoylhydroxyferuloyltriglucoside−7−O−glucoside
Qn 3−O−disinapoyltriglucoside−7−O−glucoside
Qn 3−O−sinapoylferuloyltriglucoside−7−O−glucoside
Group 17
Is 3−O−hydroxyferuloylglucoside−7−O−glucoside
Km 3−O−sinapoylglucoside−7−O−glucoside
Km 3−O−sinapoylglucoside−7−O−glucoside
Km 3−O−sinapoylsophoroside
Km 3−O−sinapoylsophoroside−7−O−glucoside
Km 3−O−sinapoyldiglucoside−7−O−glucoside
Km 3−O−sophoroside−7−O−sinapoylglucoside
Km 3−O−diglucoside−7−O−sinapoylglucoside
Km 3−O−caffeoyldiglucoside−7−O−sinapoylglucoside
Km 3−O−caffeoylsinapoyltriglucoside
Km 3−O−sinapoylsophorotrioside−7−O−glucoside
Km 3−O−sinapoyltriglucoside−7−O−glucoside
Km 3−O−hydroxyferuoloyldiglucoside−7−O−feruloyldiglucoside
Km 3−O−caffeoylsinapoylsophorotrioside−7−O−glucoside
Km 3−O−caffeoyltriglucoside−7−O−sinapoylglucoside
Km 3−O−sinapoylhydroxycaffeoyltriglucoside−7−O−glucoside
Km 3−O−feruloyltetraglucoside−7−O−feruloylglucoside
Qn 3−O−sinapoylsophoroside
Qn 3−O−sinapoylglucoside−7−O−glucoside
Qn 3−O−sinapoyldiglucoside−7−O−glucoside
Qn 3−O−sinapoyltriglucoside
Qn 3−O−triglucoside−7−O−sinapoylglucoside
Qn 3−O−hydroxyferuloyltriglucoside−7−O−hydroxyferuoloylglucoside
Qn 3−O−caffeoylsinapoyltriglucoside−7−O−glucoside
Indole-3-carboxylate dihexoside
6-Hydroxyindole-3-carboxylate dihexoside
Group 5
1-(1H-indole-3-carboxylate), glucopyranose
5'-Glucosyl dihydroascorbigen
At-NCC-4
Neoascorbigen
Bn-NCC-2
Bn-NCC-1
G
Group 4
3−O−p−Coumaroylquinic acid
4−O−p−Coumaroylquinic acid
Coumaroylquinic acid
Group 6
Chlorogenic acid
Neochlorogenic acid
4−O−Caffeoylquinic acid or 1−O−Caffeoylquinic acid
Caffeoylquinic acid
Group 8
3−O−Feruloylquinic acid
5−O−Feruloylquinic acid
Group 9
Gluconapin
Glucobrassicanapin
Group 10
Progoitrin
Gluconapoleiferin
Group 5
6−Hydroxyindole−3−carboxylic acid β−D−glucopyranosyl ester
6−Hydroxyindole−3−carboxylic acid 6−O−β−D−glucopyranoside
H
9,10-Dihydrohydroxy jasmonic acid sulfate
3-Indolylmethyl glucosinolate
Hydroxy methoxyglucobrassicin
Neoglucobrassicin
4-Hydroxyglucobrassicin
4-Methoxyglucobrassicin
3-Methylpentyl glucosinolate or 4-Methylpentyl glucosinolate or Glucoraphasatiyusain
3-Hydroxy-5-methylsulfinylpentyl glucosinolate or 5-Methylsulfonylpentyl glucosinolate
3-Hydroxymethylpentyl glucosinolate
Glucolesquerellin
Glucohesperin
Glucoerucin
Glucoerysolin
Glucoraphanin
Glucoberteroin
Glucoalyssin
Glucocochlearin or Glucoconringianin or Glucocapparisflexuosain
Glucoconringiin or Glucoarabidopsithalianain
Glucocleomin
Glucotropaeolin
Gluconasturtiin
Sinigrin
Group 9
Group 10
Epiglucobarbarin
6-Heptenyl glucosinolate
Glucoraphasatin
Group 2
2−(3−Methylthio)propylmalic acid
2−(4−Methylthio)butylmalic acid
I
N-Methanesulfonyl imidazole
Rutalexin
Fumaric acid
Succinate
Pantothenic acid
Glutathione
Group 2
Malic acid
Citric acid
J
2-Butenoic acid, 2-hydroxy-4-(1-methyl-1H-indol-3-yl)-4-oxo
4-Hydroxyindole-3-carbaldehyde
Tryptophan
N-Acetyl-L-tryptophan
N,O-Diacetyl-L-tyrosine
Phenylalanine
Tyrosine
K
Hydroxyferulic acid
Sinapic acid
Group 1
Sinapaldehyde
Kaempferol
Quercetin
p-Coumaric acid
Chorismic acid
Sinapyl alcohol
Feruloyl glycerol
Isorhamnetin
Esculetin
L
Vomifoliol 9-O-?-D-glucopyranoside
Blumenol A malonyl hexoside
Dihydroxyphaseic acid 1-O-glucoside
Vanillic acid glucoside
Group 3
Butyl acetyl hexoside
Km 3,7-di-O-glucoside
Feruloyl loganic acid
Apodanthoside
Ascorbic acid
Lariciresinol

FIGURE 2
Tanimoto similarity-based clustering tree representing the chemical complexity of the obtained metabolic profiles. The distance matrix is clustered using the complete linkage method. The classification of chemical compounds, according to the Dictionary of Natural Products (https://dnp.chemnetbase.com), is represented by the color code. Twelve major compound cluster groups are marked by letters A to L. Metabolites exhibiting identical fingerprints are represented by numbered groups from 1 to 32. Members of all groups are listed in the proximity of their original position.

238.02 (loss of Glc, −162.05 amu), *m/z* 204.03 (loss of SGlc, −196.04 amu) and *m/z* 158.06 (loss of Glc + SO3, −242.01 amu). Therefore, this long-chain methylsulfinylalkyl glucosinolate was tentatively identified as glucohesperin (Supplementary Table S1). Glucohesperin was reported in *Arabidopsis thaliana* (van de Mortel et al., 2012), while it was detected here in *B. rapa* for the first time. In total, 33 glucosinolates were putatively identified, consisting of 24 aliphatic glucosinolates, 4 aromatic glucosinolates, and 5 indole glucosinolates, including most of the glucosinolates reported earlier in *B. rapa* (Liu et al., 2020; Zou et al., 2021). To our knowledge, 16 of the detected glucosinolates were found in *B. rapa* leaves for the first time.

Glucosinolates are well-known for their roles in plant defenses against herbivores and pathogens (Chhajed et al., 2020). In addition, previous studies demonstrated that aliphatic glucosinolates were predominant in *B. rapa*, with gluconapin and glucobrassicanapin being the most abundant (Klopsch et al., 2018). In the present study, gluconapin and glucobrassicanapin were present in all investigated *B. rapa* genotypes. Moreover, yellow sarson accession #82, (here and further in the text "#" denotes genotype number in Supplementary Table S2) and turnip #89 were found to contain the highest contents of gluconapin and glucobrassicanapin, respectively. Meanwhile, the lowest contents of them were found in Chinese cabbage #57 and pak choi #14, respectively. Among all genotypes, the relative contents of gluconapin and glucobrassicanapin had 19,085- and 1,463-fold differences between the highest and the lowest values, respectively, indicating a large variation in glucosinolates. This is in line with previous studies that demonstrated extensive variation in glucosinolates in 113 turnip varieties (Padilla et al., 2007), 91 different *B. rapa* genotypes (Klopsch et al., 2018) and 82 *B. rapa* varieties (Yang and Quiros, 2010). We found various glucosinolates accumulation patterns among genotypes. For example, epiglucobarbarin and glucohesperin were presented in most genotypes, while glucocleomin and glucolesquerellin were highly accumulated only in accession savoy #20, and at lower levels in all mizuna, turnip, and purple caitai accessions. Upon cell disruption, glucosinolates are hydrolyzed to various breakdown products, which possess a wide range of health-promoting properties. Sulforaphane, the active hydrolysis product of glucoraphanin, has attracted attention due to its significant anticancer properties (Haq et al., 2021). We found that its precursor glucoraphanin was highly enriched in yellow sarson #81 and #82 and rapid cycling #79. Also, indole-3-carbinol, derived from the breakdown of glucobrassicin, showed diverse biological properties with anti-atherogenic, antioxidant, anti-carcinogenic, and anti-inflammatory activities (Kim and Park, 2018). The precursor glucobrassicin was found at the highest level in Chinese cabbage #60, which is in line with earlier reports (Padilla et al., 2007; Yang and Quiros, 2010). In the case of aromatic glucosinolates, gluconasturtiin and glucotropaeolin have been reported to be hydrolyzed by the plant enzyme myrosinase to yield phenethyl isothiocyanate and benzyl isothiocyanate, which have anti-cancer and antimicrobial activities (Cao et al., 2021). In this study, we found that accessions savoy #20, komatsuna #83, and pak choi #5 exhibited relatively higher levels of gluconasturtiin as compared with other genotypes (highest in komatsuna). In addition, glucotropaeolin was mainly accumulated in Chinese cabbage #33. Therefore, these *B. rapa* genotypes with high levels of glucosinolates might be used in future health-related applications.

3.3.2 Flavonols

Flavonols are the predominant phenolic compounds in *B. rapa*. Identification of flavonol glycosides was based on their fragmentation pattern (Ferreres et al., 2004; Harbaum et al., 2007; Lin et al., 2011). The breakdown of the O-glyosidic bond is a typical fragmentation of flavonol glycosides. Previous studies indicated that the O-glyosidic bond at the 7-position was the weakest glycosidic linkage in the flavonols molecule (Ferreres et al., 2004). Thus, the first loss usually was the glycose or acyl-glycose moiety at the 7-position, and then the loss of glycose or acyl moieties at position 3. For acylated flavonol glycosides, neutral loss information was used to characterize acyl groups by the losses of 42.01, 146.04, 162.03, 176.05, 178.03, 192.04, and 206.06 amu for acetyl, p-coumaroyl, caffeoyl, feruloyl, hydroxycaffeoyl, hydroxyferuloyl, and sinapoyl, respectively. In addition, losses of 180.06, 162.05, and 120.04 amu from interglycosidic fragmentations suggested sophoroside or sophorotrioside with 1→2 glycosidic linkage (Ferreres et al., 2004; Lin et al., 2011). As an example, compounds 205–208 were found with the same deprotonated molecular ion at *m/z* 977.26 and aglycone ions (*m/z* 285.04 and *m/z* 284.03), suggesting that they were the isomeric kaempferol glycosides. First, for compounds 205 and 206, the fragment ion at *m/z* 815.20 as a base peak was observed due to the loss of a glucosyl moiety (−162.05 amu) at the 7-O position. Another fragment ion at *m/z* 609.15 was due to the further loss of sinapoyl moiety at the 3-position (−206.06 amu). After the loss of a diglucosyl moiety at the 3-position (−324.11 amu), kaempferol aglycone ions were detected. Moreover, compound 205 showed the fragments at *m/z* 489.11 (−120.04 amu) and *m/z* 429.08 (−180.06 amu) to confirm the sophorosyl moiety. Thus, compounds 205 and 206 were putatively identified as kaempferol 3-O-sinapoylsophoroside-7-O-glucoside and kaempferol 3-O-sinapoyldiglucoside-7-O-glucoside, respectively. For the other two isomers (compound 207 and 208), a fragment ion at *m/z* 609.15 was detected as a base peak due to the simultaneous loss of a glucosyl moiety (−162.05 amu) and a sinapoyl moiety (−206.06 amu) at the 7-position. Further loss of diglucosyl moiety at the 3-position gave rise to the kaempferol aglycone ions. Together with the characteristic fragments of sophoroside, compounds 207 and 208 were putatively identified as kaempferol 3-O-sophoroside-7-O-sinapoylglucoside and kaempferol 3-O-diglucoside-7-O-sinapoylglucoside, respectively (Supplementary Table S1). In this study, a total of 105 flavonols, including 66 kaempferol derivatives, 28 quercetin derivatives, and 11 isorhamnetin derivatives were identified. To the best of our knowledge, 50 *B. rapa* flavonols are reported here for the first time. Among flavonols, three aglycons (kaempferol, quercetin, and isorhamnetin), 16 non-acylated glycosides, 61 monoacylated, and 25 diacylated glycosides were detected. *B. rapa* leaves showed complex flavonols conjugate with different glycosylation and acylation patterns. Some flavonols possess molecular weight above 1,000 Da, and this increased the complexity of metabolite identification. In one example, detailed identification of compound 263, a diacylated quercetin tetra-glycoside with *m/z* 1,317.34, is presented in the Supplementary Material. In the case of non-acylated flavonol glycosides, mono-, di-, and tri-glycosides of isorhamnetin and quercetin as well as mono- to tetra-glycosides of kaempferol were found. Moreover, 86 flavonol glycosides were acylated with acetic, *p*-coumaric, caffeic, sinapic, ferulic, hydroxyferulic, and hydroxycaffeic acids. We found that mono-acylated glycosides widely existed as kaempferol, quercetin, and isorhamnetin glycosides. However, diacylated glycosides were only present as quercetin tetraglycosides as well as tri-, tetra-, and pentaglycosides of kaempferol. In good agreement with previous reports (Chihoub et al., 2019; Wiesner-Reinhold et al., 2021), kaempferol glycosides were the most diverse flavonols derivatives in *B. rapa,* with 7 non-acylated, 39 mono-acylated, and 19 di-acylated glycosides, respectively.

Remarkable variations of flavonols levels were observed among all genotypes, especially for kaempferol and quercetin glycosides. We found kaempferol 3-O-disinapoylsophorotrioside-7-O-glucoside (compound 235) and kaempferol 3-O-sinapoylsophorotrioside-7-O-glucoside (compound 225) exhibited extremely different levels with 13,123- and 10,432-fold differences between the highest and the lowest values among the tested genotypes, respectively. Similarly, quercetin 3-O-triglucoside-7-O-sinapoylglucoside (compound 258) and quercetin 3-O-triglucoside-7-O-feruloylglucoside (compound 257) displayed 5,577- and 3,549-fold change among all genotypes, respectively. Previous studies have shown that kaempferol derivatives were the most abundant flavonols in Chinese cabbage, pak choi, turnip, and mizuna (Soengas et al., 2018; Dejanovic et al., 2021; Kyriacou et al., 2021; Wiesner-Reinhold et al., 2021), with kaempferol-3,7-di-O-glucoside (compound 176), kaempferol 3-O-caffeoylsophoroside-7-O-glucoside (compound 194), kaempferol 3-O-hydroxyferuloylsophoroside-7-O-glucoside (compound 199), kaempferol 3-O-feruloylsophoroside-7-O-glucoside (compound 201) and kaempferol 3-O-sinapoylsophoroside-7-O-glucoside (compound 205) being the most abundant kaempferol derivatives. These compounds have been reported as antioxidants with high free radical scavenging activity and antimicrobials with effective inhibition of Gram-positive and -negative bacteria (Favela-González et al., 2020; Abellán et al., 2021). In the present study, we found that accessions Chinese cabbage #48, pak choi #10, turnip #88, and mizuna #85 contained the highest levels of these compounds. In addition, Chinese cabbage #48 and pak choi #11 also exhibited high concentrations of isorhamnetin-3,7-di-O-glucoside (compound 168) and isorhamnetin-3-O-glucoside (compound 112). These metabolites have been demonstrated to be active compounds in *Salicornia herbacea* (Lee et al., 2021) and mustard leaf (*Brassica juncea*) (Yokozawa et al., 2002), affecting insulin secretion and blood glucose levels.

The wide variation of flavonols in *B. rapa* determines diverse and important biological functions. Quercetin, kaempferol, and isorhamnetin and their derivatives have diverse bioactivities including antioxidant, antimicrobial, antifungal, and antiviral potentials (Barreca et al., 2021). We found that Chinese cabbage #33 contained the highest amount of isorhamnetin and kaempferol. The highest amount of quercetin was found in another Chinese cabbage accession #35. Previous studies revealed that the caffeoyl moiety due to the O-dihydroxy structure could enhance radical scavenging ability (Braca et al., 2003). In our study, we found 18 caffeoyl kaempferol and quercetin glycosides and one hydroxycaffeoyl kaempferol glycoside. However, glycosylation has been reported to decrease the scavenging activity of flavonoids (De Winter et al., 2015). Considering caffeoyl moiety and glycosylation, kaempferol 3-O-caffeoylsophoroside (compound 182), kaempferol 3-O-caffeoyldiglucoside (compound 183), and quercetin 3-O-caffeoyldiglucoside (compound 244) were expected to be strong antioxidants in *B. rapa* leaves. The highest amounts of compounds 182 and 183 were found in pak choi #11, while Chinese cabbage #63 showed the highest level of compound 244. Therefore, pak choi #11 and Chinese cabbage #63 may be excellent sources of strong antioxidants in *B. rapa*.

3.3.3 Hydroxycinnamic acid derivatives

In *B. rapa*, hydroxycinnamic acid derivatives represent another prominent class of phenolic compounds. The fragmentation of hydroxycinnamic acid glycosides showed the loss of glycosyl and hydroxycinnamoyl moiety to produce hydroxycinnamic acid ions. In the case of hydroxycinnamoyl diglycosides in *Brassica* vegetables, the diglycosyle moiety was mainly characterized as a gentiobiose unit (1→6 glycosidic linkage) (Harbaum et al., 2007; Olsen et al., 2009). For example, compounds 133–136 were detected with a deprotonated molecular ion at *m/z* 739.21. The fragment ion at *m/z* 515.14 was formed by the loss of the sinapoyl (−224.07 amu). The hydroxyferulic acid ion at *m/z* 209.04 was formed by successive loss of gentiobiose moiety (−306.09 amu). Further loss of H2O (−18.01 aum) from hydroxyferulic acid resulted in a fragment ion at *m/z* 191.03. Thus, they were putatively identified as isomers of sinapoyl hydroxyferuloyl gentiobiose. Notably, many isomers of hydroxycinnamic acid derivatives were detected in *B. rapa* leaves (Supplementary Table S1). These isomers could be the result of a different linkage position of the hydroxycinnamoyl group. Some isomers could be distinguished using authentic standards. For example, four caffeoylquinic acid isomers were characterized based on their molecular ion (*m/z* 353.09) and predominant fragment ions (*m/z* 191.06 and 173.04 for quinic acid; *m/z* 179.03, 161.04, and

135.04 for caffeic acid). Finally, three isomers were identified with high confidence as 3-O-caffeoylquinic acid (chlorogenic acid), 5-O-caffeoylquinic acid (neochlorogenic acid), and 4-O-caffeoylquinic acid or 1-O-caffeoylquinic acid (Supplementary Table S1). According to authentic standards and the previously described fragmentation patterns (Harbaum et al., 2007; Lin et al., 2011; Sun et al., 2013), in total 93 hydroxycinnamic acid derivatives were identified in this study, including 4 hydroxycinnamic acids, 4 glycerol and shikimic acid esters, 10 malic acid esters, 14 quinic acid esters, and 61 glycosides. The main derivatives were hydroxycinnamic acid glycosides, including mono-, di-, or triglucose integrated with one, two, or three hydroxycinnamoyl units.

Hydroxycinnamic acid derivatives displayed high variability among the different *B. rapa* studied here. Generally, hydroxycinnamoyl quinic acids accumulated to high levels in all accessions of Chinese cabbage, savoy, pak choi, taicai, and caixin, while they were present at low levels in broccoletto, rapid cycling, yellow sarson, and wild cabbage. Similarly, low levels of hydroxycinnamoyl malic acids were also detected in rapid cycling and yellow sarson. In contrast, hydroxycinnamoyl glycosides were highly abundant in rapid cycling and yellow sarson, as well as in Chinese cabbage and komastuna, but low in caixin, broccoletto, and wild cabbage. According to previous reports (Soengas et al., 2018; Dejanovic et al., 2021), sinapic acid derivatives were the major hydroxycinnamic acid derivatives in *B. rapa*, including 1,2-disinapoyl gentiobiose (compound 147–149) and 1-sinapoyl-2-feruloyl gentiobiose (compound 124–128). These were reported to exhibit antioxidant and anti-inflammatory effects in human plasma and human peripheral blood mononuclear cells (Olszewska et al., 2020). In our study, Chinese cabbage #31 contained the highest amounts of disinapoyl gentiobiose and sinapoyl feruloyl gentiobiose. In another study with pak choi (Heinze et al., 2018) and mizuna (Wiesner-Reinhold et al., 2021), sinapoyl malate was a major hydroxycinnamic acid derivative. Previous studies showed that sinapoyl malate together with other hydroxycinnamoyl malic acids may play an important role in *B. rapa* jasmonate-mediated defense response (Liang et al., 2006). The highest amount of sinapoyl malate was found in mizuna #85. Free hydroxycinnamic acids have been reported to act as powerful antioxidants (Coman and Vodnar, 2020). It was found that hydroxycinnamates work as effective UV-B protectants in *Arabidopsis* (Landry et al., 1995). In this study, four hydroxycinnamic acids were detected in all tested genotypes. The highest amounts of p-coumaric acid and hydroxyferulic acid were found in turnip #92, while pak choi #16 contained the highest levels of sinapic acid and ferulic acid.

3.3.4 Anthocyanins

Anthocyanins are important water-soluble pigments in plants. In the negative ion mode, anthocyanins exhibited a unique doublet of ions [M-2H]$^-$ and [M-2H + H2O]$^-$ for their molecular ion, which could be used to identify anthocyanins and differentiate them from other polyphenols (Sun et al., 2012). In addition, doubly charged ions were observed for pelargonin and cyanidin glycosides, in some cases as the base peak (Sun et al., 2012). The MS fragmentation of anthocyanins occurred mainly at the glycosidic bonds between the flavylium ring and sugar moieties as well as ester bonds between the sugar moieties and acyl groups (Wu and Prior, 2005). For example, compound 271 was an anthocyanin with the highest level in purple caitai #101. Characteristic doublet ions [M-2H]$^-$ and [M-2H + H2O]$^-$ were observed at *m/z* 1,239.31 and 1,257.31, respectively. In addition, doubly charged ions at *m/z* 619.14 [M-2H]$^{2-}$ and *m/z* 628.15 [M-2H + H2O]$^{2-}$ were found as the major peaks. The MS fragmentation showed a double-charged ion at *m/z* 597.15 and a single-charged ion at *m/z* 1,195.32 by loss of a carboxyl residue (43.99 amu). In addition, two fragment ions at *m/z* 1,153.30 and *m/z* 991.25 were observed by loss of malonyl residue (86.00 amu) and malonylglucoside moiety (248.05 amu), indicating the presence of a malonylglucoside moiety at the 5-position. Furthermore, successive loss of a feruloyl residue (176.05 amu) and a sinapoyl residue (206.06 amu) gave rise to the fragment ions at *m/z* 609.14, revealing the presence of a feruloyl-sinapoyl residue at the 3-position. Finally, the loss of a diglucose moiety (324.11 amu) from the 3-position produced the cyanidin aglycone ions at *m/z* 285.04 and 284.03. Based on earlier reports (Guo et al., 2015; Song et al., 2020), compound 271 was putatively identified as cyanidin 3-feruloylsinapoylsophoroside-5-malonylglucoside.

All accessions of purple caitai and purple turnip were the only accessions exhibiting purple color due to the presence of anthocyanins. Interestingly, purple caitai only contained cyanidin derivatives, while purple turnip contained exclusively pelargonidin derivatives, indicating different biosynthetic pathways of anthocyanins. In purple caitai, all nine anthocyanins were acylated cyanidin-3-sophoroside-5-glucoside derivatives, as previously shown in *B. rapa* (Guo et al., 2015; Song et al., 2020). However, information on anthocyanin composition in purple turnips was limited so far. In this study, five acylated pelargonidin-3-O-diglucoside-5-O-malonoylglucoside derivatives were putatively identified in purple turnip, which were similar to the anthocyanins reported in red radish (Wu and Prior, 2005; Jing et al., 2014). Anthocyanins have been demonstrated to possess antioxidant activity and preventive activities against cardiovascular disease, metabolism disease, diabetes, and obesity (Ghareaghajlou et al., 2021). The chemical structures of anthocyanins determine their stability, color intensity, and potential biological activity. Previous studies reported that diacylated anthocyanins were characterized by higher antioxidant capacity than monoacylated anthocyanins, while the latter had higher antioxidant capacity than nonacylated forms (Wiczkowski et al., 2013). In addition, acylation with sinapic acid leads to higher antioxidant capacity than with ferulic acid, followed by *p*-

coumaric acid (Wiczkowski et al., 2013). In this study, we found that most of the anthocyanins identified in purple caitai and purple turnip contained diacylation with sinapic acid and ferulic acid, modifications that will contribute to good stability and high antioxidant capacity.

3.3.5 Monolignol and oligolignol derivatives

Lignin, an aromatic biopolymer found in plant cell walls, is essential for water transport and mechanical support, and plays an important role in plant defense (Chantreau et al., 2014). Lignin is derived from the combinatorial coupling of monolignol radicals. The MS fragmentation pattern of lignin oligomers has been described previously (Morreel et al., 2010a; Morreel et al., 2010b; Morreel et al., 2014). In this study, we characterized glycosylation and esterification groups, monolignol units, and linkage types. First, for oligolignol glycosides and malate esters, MS fragmentation occurred by loss of glycosyl moiety (324.11 and 162.05 amu) or the malyl moiety (116.01 amu). Second, small neutral losses provide information on the three types of linkages. Third, the first product ions resulting from the cleavage of the linkage yielded the information on the units. For example, compound 298 was detected as a deprotonated ion at *m/z* 581.19. The base peak in MS fragmentation was at *m/z* 419.13 indicating a hexose loss (−162.05 aum). Furthermore, a fragment ion at *m/z* 371.11 was observed that likely resulted from the *β*-aryl ether, a combined loss of water and formaldehyde (−48.02 amu). Fragmentations yield ions at *m/z* 223.06 and 195.06, representing the units derived from sinapic acid and coniferyl alcohol. Furthermore, fragment ions at *m/z* 208.04 and 165.06 indicated a further methyl radical loss from sinapic acid and formaldehyde loss from the G unit, respectively. Therefore, compound 298 was characterized as G(8-O-4)sinapic acid ester hexoside (Supplementary Table S1). In this study, we identified 51 monolignol and oligolignol derivatives in *B. rapa* leaves, including 20 monolignol, 21 lignans and neolignans, and 10 trimeric oligolignols derivatives. To the best of our knowledge, 46 monolignol and oligolignol derivatives are reported here in *B. rapa* leaves for the first time; eight of them were identified with high-confidence levels.

In *B. rapa* leaves, monolignol and oligolignols were mainly composed of guaiacyl (G) and syringyl (S) units that are derived from coniferyl alcohol and sinapyl alcohol, respectively. Various inter-monomeric linkages were observed, including *β*-aryl ether linkage (8-O-4), resinol linkage (8-8), and phenylcoumaran linkage (8-5). In the case of the 8-8 linkage, lignans belonging to different classes were identified, including lariciresinol, pinoresinol, secoisolariciresinol, syringaresinol, and dehydrodiconiferyl alcohol derivatives. Due to the free hydroxyl group from G and S units, most monolignols and oligolignols were glycosylated with one or two hexoses. In addition, monolignol, lignans, and neolignans were conjugated with ferulic acid or sinapic acid, which were further esterified by malate. Recently, a wide range of monolignol and oligolignol derivatives have been found in seed coats of pomegranate (Qin et al., 2020) and arabidopsis leaves (Dima et al., 2015). This indicated that monolignols not only incorporated into lignin polymer biosynthesis/assembly but also participated in other metabolic pathways to form diverse metabolites.

Sinapyl alcohol, the only free monolignol detected in *B. rapa* leaves, exhibited significantly higher amounts in all accessions of oil cabbage, up to 60-fold higher as compared to other genotypes. Monolignol derivatives were detected in higher amounts in all accessions of mizuna and turnip, while the lower amount in oil cabbage, rapid cycling, and yellow sarson. Lignan and neolignan derivatives exhibited a similar accumulation pattern across the accessions, with higher amounts detected in all accessions of mizuna, turnip, and yellow sarson. Trimeric oligolignols derivatives exhibited the highest levels in all accessions of mizuna, turnip, yellow sarson, and broccoletto, while the lowest levels were found in all accessions of oil cabbage. A recent study demonstrated that lignans possess antimicrobial, anti-inflammatory, and antioxidant activities (Hano et al., 2021). Lariciresinol, pinoresinol glucoside (symplocosin), and pinoresinol diglucoside have been proven to possess considerable antioxidant potential in different *in vitro* assays (Gülçin et al., 2006; Soleymani et al., 2020). Here, we found that Chinese cabbage accession #48 contained the highest amount of lariciresinol, while two turnip accessions #87 and #92 contained the highest amount of pinoresinol glucoside and pinoresinol diglucoside, respectively. A previous study showed that lariciresinol glycoside exhibited potent anti-inflammatory activity through the NF-κB signaling pathway (Bajpai et al., 2018). The highest amounts of lariciresinol glycoside were found in Chinese cabbage #48 and rapid cycling #79. Syringaresinol glucoside, an effective regulator of lipogenesis and glucose consumption (Wang et al., 2017), was mainly abundant in mizuna #85 and oil cabbage #78. Finally, some lignans and their glycosides, including secoisolariciresinol, pinoresinol, and lariciresinol, are the precursors for enterolignans with phytoestrogen activity (Hano et al., 2021). Enterolignans are characterized by various biologic activities, including tissue-specific estrogen receptor activation, together with anti-inflammatory and apoptotic effects (Senizza et al., 2020). In this study, the secoisolariciresinol, pinoresinol, and lariciresinol glycoside characterized in *B. rapa* may also be the precursors for the formation of enterolignans (compounds, formed by the action of gut microflora on lignans).

3.3.6 Organic acids and other metabolites

Malic acid, citric acid, and ascorbic acid have been reported as the predominant organic acids in *B. rapa* (Arias-Carmona et al., 2014). Here, we putatively identified ten organic acids in *B. rapa* leaves and they were common in all genotypes. It is well known that malic acid and citric acid contribute to the sensory characteristics due to their sour taste, while ascorbic acid is an

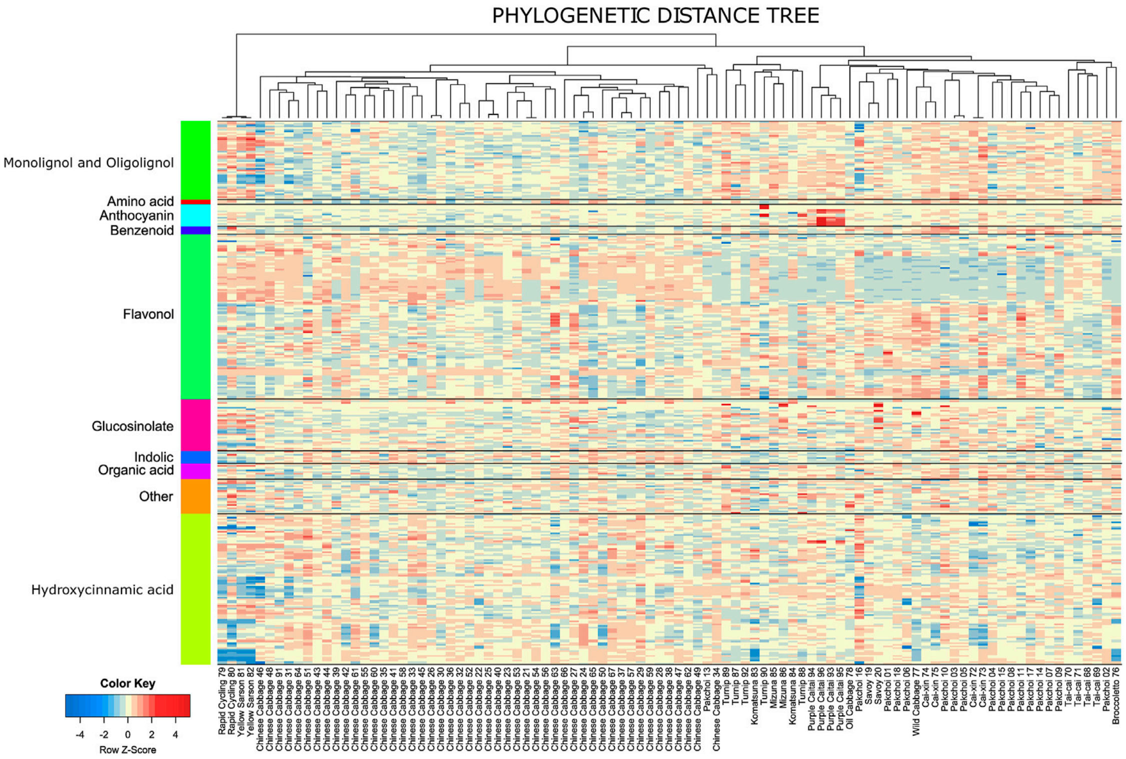


FIGURE 3
Heatmap of the relative metabolite abundance across the accessions. The heatmap is scaled row-wise (Z-scores from the log values of the metabolite intensity); thus, the colors represent the deviation from the average value obtained for a metabolite. The columns are ordered according to the midpoint-rooted phylogenetic tree.

important enzyme cofactor, radical scavenger, and donor/acceptor in electron transport (Davey et al., 2000). Malic acid, citric acid, and ascorbic acid showed similar distribution among all genotypes with 2.2-, 6.7-, and 6.1-fold variations between the lowest and the highest levels, respectively. The highest amounts of malic acid, citric acid, and ascorbic acid were detected in Chinese cabbage accession #33, pak choi #16, and a second pak choi accession #10, respectively. The amino acids phenylalanine, tyrosine, and tryptophan were detected in all accessions with 6.1-, 7.8-, and 19.8-fold variations between the lowest and the highest amounts, and the highest amounts were all found in Chinese cabbage #26. Roseoside vomifoliol 9-O-β-D-glucopyranoside (compound 78) was identified with high confidence in *B. rapa* for the first time. This compound characterized before in *Leea aequata* L. showed anticancer activity due to the induction of apoptosis (Rahim et al., 2021). A recent report showed potent anti-inflammatory, antiallergic, and COVID-19 protease inhibitory activities of roseosides (Ebada et al., 2020). In this study, Chinese cabbage #43 had the highest level of vomifoliol 9-O-β-D-glucopyranoside, followed by two additional Chinese cabbage accessions #41 and #52, and turnip #92. Finally, Bn-NCC-1 and Bn-NCC-2, compounds belonging to the tetrapyrroles, were considered potential biomarkers of *Brassica* plants (according to the Dictionary of Natural Products version 30.2 https://dnp.chemnetbase.com) and were both detected in all *B. rapa* genotypes.

3.4 Metabolic diversity of the *B. rapa* subspecies and their accessions

All except four measured metabolites exhibited significant variation across the genotypes according to ANOVA (FDR-adjusted p-value ≤ 0.05; Supplementary Table S2). The only non-significant metabolites included rutin (compound 240), N,O-diacetyl-L-tyrosine (compound 25), and two low-abundant anthocyanins: pg 3-p-coumaroyldiglucoside-5-malonoylglucoside (compound 275) and one of its isomers (compound 277). At the same time, the total contribution of the between-replicate variance was 12% across all measured metabolites. This indicated high specificity of the specialized metabolites composition in measured samples. It appeared that the similarity between metabolic profiles across the *B. rapa* genotypes of the same subspecies is much higher than the similarity between the subspecies (Figure 3). This concerns practically all measured metabolite classes, but most

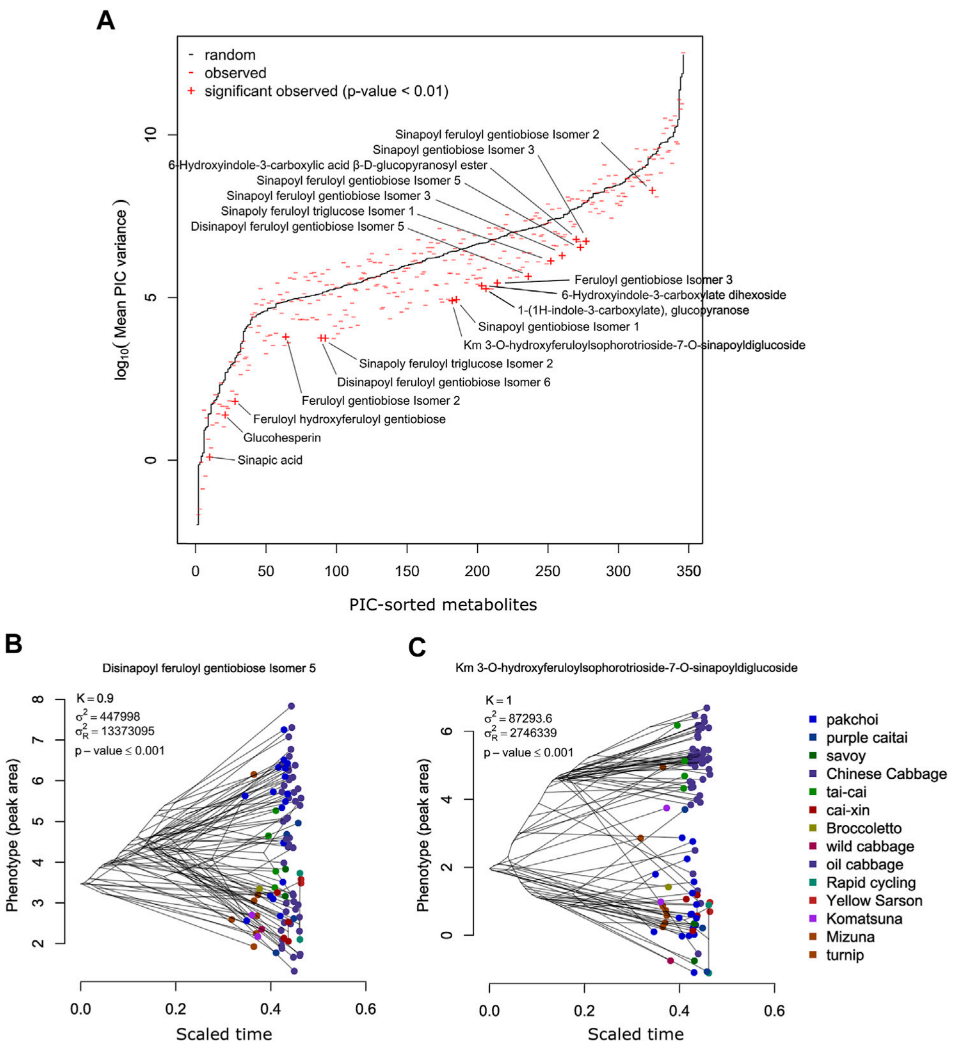


FIGURE 4
(A) Comparison of the PIC values expected from the random Brownian motion model (black) and PIC observed for the measured metabolites. Metabolites exhibiting significantly lower PIC (p-value ≤ 0.01) are labeled. **(B)** Phenogram of the disinapoyl feruloyl gentiobiose isomer 5—an example of a metabolite exhibiting high and significant phylogenetic signal. **(C)** Phenogram of the km 3-O-hydroxyferuloylsophorotrioside-7-O-sinapoyldiglucoside—another example of a metabolite exhibiting high and significant phylogenetic signal.

remarkably anthocyanins, flavonols, and hydroxycinnamic acid derivatives. For example, anthocyanins are found in high levels in all Purple caitai accessions. Several flavonols are specifically accumulated in all Chinese cabbage accessions but were found at low levels in pak choi, caixin, savoy, turnip, and mizuna accessions. Yellow sarson and rapid cycling accessions on the other hand accumulated a group of lignans and several specific hydroxycinnamic acid derivatives.

Knowing the phylogenetic relationship and quantitative phenotype traits, across a population, it is possible to quantify how much the value of a certain trait is related to the phylogeny. In the case of the analyzed *B. rapa* subspecies and their accessions, the evolutionary process is represented by agronomical trait selection and the traits of interest are levels of measured biochemical compounds. We explored this phenomenon in a systematic way enumerating the trait-phylogeny relationship.

Phylogenetic signal, a quantitative measure of the trait-phylogeny relationship (Hillis and Huelsenbeck, 1992), has been estimated for the metabolic profiles using the phylogenetic neighbor-joining tree calculated from all annotated polymorphisms. Specifically, we applied a well-established method of Blomberg et al. (2003), using a Brownian motion model to simulate the evolution of the traits along the branches of the phylogenetic tree. The distribution of the metabolic traits, for example, relative accumulation of each

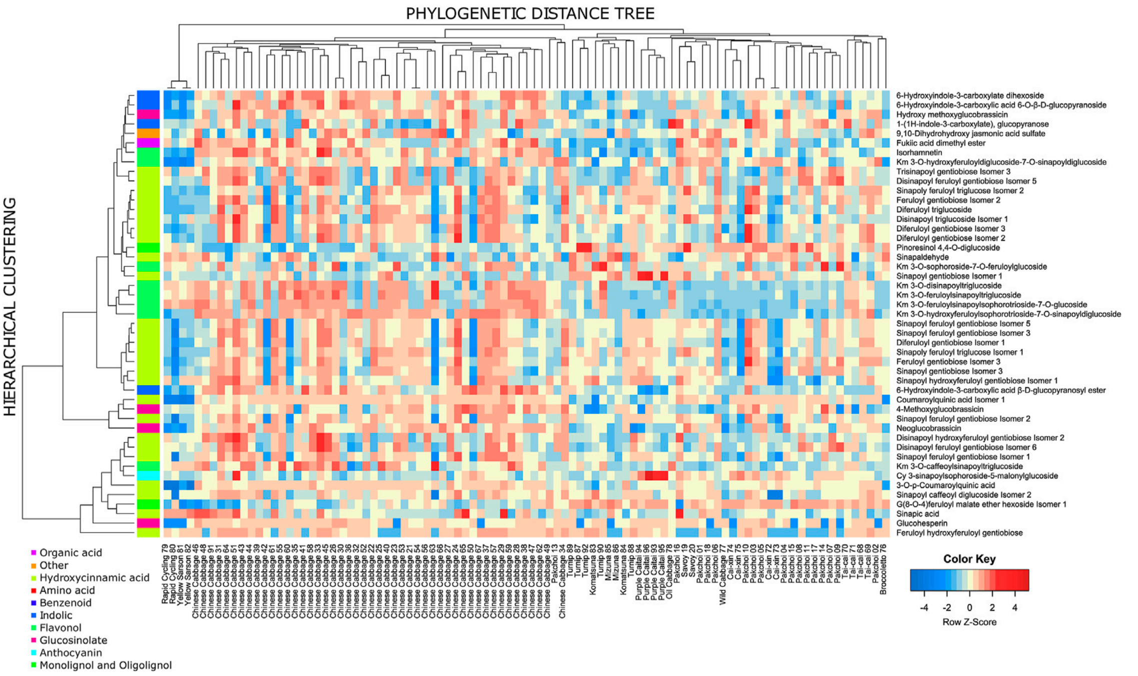


FIGURE 5
Heatmap of compounds exhibiting significant phylogenetic signal (p-value ≤ 0.05). The side color sidebar represents the classification of measured compounds to 10 biochemical classes. The row-wise clustering tree is based on the Euclidean distance between Z-transformed metabolite levels, and the average linkage method for cluster agglomeration. The column tree is a midpoint-rooted phylogenetic tree.

metabolite, compared with the simulated model provides an informative statistical output in terms of comparable K statistic values. The method is adequate for rough phylogenetic relatedness estimation using the SNP-based neighbor-joining tree, as it was shown to be robust against errors likely emerging in branch length estimation (Münkemüller et al., 2012). For a test of significance estimation, both the theoretical and empirical p-values have been computed. Due to the sensitivity of the K statistics to the differences in the trait values distribution, here we used only empirical p-values derived from the 999-fold permutation test (results attached in Supplementary Table S3).

A comparison between observed and randomized data is shown in Figure 4A. PIC variance value (standardized phylogenetic independent contrast scaled by the branch length) is a measure reflecting how the independence of the trait values decreases with the decreasing phylogenetic distance. The PIC for randomized data without scaling is affected by the total trait variance and thus individual tests were performed for each metabolite. In Figure 4A, results for individual metabolites were sorted according to the average PIC obtained for 999 random permutations. In general, observed PIC values are shifted towards lower values with respect to the mean value obtained in 1,000 random permutations (red marks are mostly below the black line); however, only some of them deviate significantly. In Figure 4A, 18 metabolites with an empirical p-value ≤ 0.01 have been highlighted.

Among the 346 measured metabolites, 18 metabolites exhibited a significant phylogenetic signal with a p-value ≤ 0.01 and 47 with a p-value ≤ 0.05. To visualize the connection between metabolite level and phylogeny, we show phylograms of two highly significant metabolites: disinapoyl feruloyl gentiobiose isomer 5 (compound 160) and kaempferol 3-O-hydroxyferuloylsophorotrioside-7-O-sinapoyldiglucoside (compound 239) (Figures 4B, C). In a phylogram, branches of a phylogenetic tree are organized according to their phenotype value (y-axis) and standardized time of the modeled evolutionary process (x-axis). We observed that the phylogenetic branches of Chinese cabbage are shifted towards higher levels of both metabolites, whereas, for example, pak choi and yellow sarson are much lower. There are also differences between both metabolites, whereas for compound 160, pak choi accessions exhibited a wide range of metabolite accumulation, overlapping with the Chinese cabbage; in the case of compound 239, pak choi accumulated much lower levels than the Chinese cabbage. The 47 compounds that have been selected as exhibiting significant phylogenetic signals with an empirical p-value ≤ 0.05 are enriched in hydroxycinnamic acid derivatives and indolics (Figure 5; Fisher's exact test p-values ≤ 0.01). Most of the significant metabolites are described by the differential accumulation in four major phylogenetic branches: 1) the Chinese cabbage, 2) the yellow sarson and rapid cycling, 3) the mizuna-komatsuna-turnip-caitai branch, and 4) the rest of the genotypes. This separation highlights the major metabolic effects of the selection pressure, leading to the development of modern *B. rapa* subspecies and their individual accessions. It is also an indication that the stepwise changes in specialized metabolism during *B. rapa* selective breeding processes are observable and can be reconstructed from metabolomics data. Finally, it is important to note that while the estimated phylogeny stays in concordance with the population structure (Supplementary Figure S6), it is only a rough approximation of the evolutionary process leading to the emergence of the analyzed genotypes. At this point, analysis of the evolution of particular metabolic traits and identification of evolutionary events and loci associated with the accumulation of specific metabolites requires the inclusion of more genotypes and the association mapping with higher statistical power.

4 Conclusion

This study presents comprehensive metabolite profiling of *B. rapa* leaves from 102 different genotypes. By this approach, a total of 346 metabolites were identified. Among them, 36 metabolites were identified in high confidence, and 184 metabolites were reported in *B. rapa* leaves for the first time. HCA and phylogenetic analysis were applied to reveal metabolite diversity and accumulation patterns as well as to identify species-specific metabolites. This work expanded the current information on *B. rapa* metabolites. It provides valuable information for developing new *B. rapa* accessions with high levels of selected metabolites possessing health-promoting activity or desired physiological function. The analysis also exemplified how selective pressure in agriculture might utilize the native biosynthetic capacity of the species to achieve highly divergent metabolic phenotypes.

Data availability statement

Original datasets with respective metadata and methods are available in a publicly accessible e!DAL repository (Arend et al. 2016): https://doi.org/10.5447/ipk/2022/27.

Author contributions

SZ performed the metabolomic analysis and wrote the manuscript, JS performed data analysis and wrote selected sections of the manuscript, NS assisted in metabolomic data processing, SMa supervised metabolomic analysis, SMe assisted in the metabolomic analysis, XW provided plant material, AA initiated and supervised the study, and IR supervised the study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.953189/full#supplementary-material

References

Arend, D., Junker, A., Scholz, U., Schüler, D., Wylie, J., and Lange, M. (2016). PGP repository: A plant phenomics and genomics data publication infrastructure. *Database*. doi:10.1093/database/baw033

Abellán, Á., Domínguez-Perles, R., García-Viguera, C., and Moreno, D. A. (2021). *In vitro* evidence on bioaccessibility of flavonols and cinnamoyl derivatives of cruciferous sprouts. *Nutrients* 13, 4140. doi:10.3390/nu13114140

Arias-Carmona, M. D., Romero-Rodríguez, M. Á., and Vázquez-Odériz, M. L. (2014). Determination of organic acids in Brassica rapa L. leaves (turnip greens and turnip tops) regulated by the protected geographical indication "Grelos De Galicia". *J. Food Nutr. Res.* 2, 786–791. doi:10.12691/JFNR-2-11-5

Bajpai, V. K., Alam, M. B., Quan, K. T., Ju, M.-K., Majumder, R., Shukla, S., et al. (2018). Attenuation of inflammatory responses by (+)-syringaresinol via MAP-Kinase-mediated suppression of NF-κB signaling *in vitro* and *in vivo*. *Sci. Rep.* 8, 9216. doi:10.1038/s41598-018-27585-w

Barreca, D., Trombetta, D., Smeriglio, A., Mandalari, G., Romeo, O., Felice, M. R., et al. (2021). Food flavonols: Nutraceuticals with complex health benefits and functionalities. *Trends Food Sci. Technol.* 117, 194–204. doi:10.1016/j.tifs.2021.03.030

Blomberg, S. P., Garland, T., Jr., and Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution* 57, 717–745. doi:10.1111/j.0014-3820.2003.tb00285.x

Braca, A., Fico, G., Morelli, I., De Simone, F., Tomè, F., and De Tommasi, N. (2003). Antioxidant and free radical scavenging activity of flavonol glycosides from different Aconitum species. *J. Ethnopharmacol.* 86, 63–67. doi:10.1016/s0378-8741(03)00043-6

Cao, Q., Wang, G., and Peng, Y. (2021). A critical review on phytochemical profile and biological effects of turnip (*Brassica rapa L.*). *Front. Nutr.* 8, 721733. doi:10.3389/fnut.2021.721733

Chantreau, M., Portelette, A., Dauwe, R., Kiyoto, S., Crônier, D., Morreel, K., et al. (2014). Ectopic lignification in the flax lignified bast fiber1 mutant stem is associated with tissue-specific modifications in gene expression and cell wall composition. *Plant Cell.* 26, 4462–4482. doi:10.1105/tpc.114.130443

Cheng, F., Wu, J., Cai, C., Fu, L., Liang, J., Borm, T., et al. (2016). Genome resequencing and comparative variome analysis in a Brassica rapa and *Brassica oleracea* collection. *Sci. Data* 3, 160119. doi:10.1038/sdata.2016.119

Chhajed, S., Mostafa, I., He, Y., Abou-Hashem, M., El-Domiaty, M., and Chen, S. (2020). Glucosinolate biosynthesis and the glucosinolate–myrosinase system in plant defense. *Agronomy* 10, 1786. doi:10.3390/agronomy10111786

Chihoub, W., Dias, M. I., Barros, L., Calhelha, R. C., Alves, M. J., Harzallah-Skhiri, F., et al. (2019). Valorisation of the green waste parts from turnip, radish and wild cardoon: Nutritional value, phenolic profile and bioactivity evaluation. *Food Res. Int.* 126, 108651. doi:10.1016/j.foodres.2019.108651

Coman, V., and Vodnar, D. C. (2020). Hydroxycinnamic acids and human health: Recent advances. *J. Sci. Food Agric.* 100, 483–499. doi:10.1002/jsfa.10010

Davey, M. W., Montagu, M. V., Inzé, D., Sanmartin, M., Kanellis, A., Smirnoff, N., et al. (2000). Plant L-ascorbic acid: Chemistry, function, metabolism, bioavailability and effects of processing. *J. Sci. Food Agric.* 80, 825–860. doi:10.1002/(sici)1097-0010(20000515)80:7<825::aid-jsfa598>3.0.co;2-6

De Winter, K., Dewitte, G., Dirks-Hofmeister, M. E., De Laet, S., Pelantová, H., Křen, V., et al. (2015). Enzymatic glycosylation of phenolic antioxidants: Phosphorylase-mediated synthesis and characterization. *J. Agric. Food Chem.* 63, 10131–10139. doi:10.1021/acs.jafc.5b04380

Dejanovic, G. M., Asllanaj, E., Gamba, M., Raguindin, P. F., Itodo, O. A., Minder, B., et al. (2021). Phytochemical characterization of turnip greens (Brassica rapa ssp. rapa): A systematic review. *PLoS One* 16, e0247032. doi:10.1371/journal.pone.0247032

Dima, O., Morreel, K., Vanholme, B., Kim, H., Ralph, J., and Boerjan, W. (2015). Small glycosylated lignin oligomers are stored in Arabidopsis leaf vacuoles. *Plant Cell.* 27, 695–710. doi:10.1105/tpc.114.134643

Ebada, S. S., Al-Jawabri, N. A., Youssef, F. S., El-Kashef, D. H., Knedel, T.-O., Albohy, A., et al. (2020). Anti-inflammatory, antiallergic and COVID-19 protease inhibitory activities of phytochemicals from the Jordanian hawksbeard: Identification, structure–activity relationships, molecular modeling and impact on its folk medicinal uses. *RSC Adv.* 10, 38128–38141. doi:10.1039/d0ra04876c

Fabre, N., Poinsot, V., Debrauwer, L., Vigor, C., Tulliez, J., Fouraste, I., et al. (2007). Characterisation of glucosinolates using electrospray ion trap and electrospray quadrupole time- of-flight mass spectrometry. *Phytochem. Anal.* 18, 306–319. doi:10.1002/pca.983

Farris, J. S. (1972). Estimating phylogenetic trees from distance matrices. *Am. Nat.* 106, 645–668. doi:10.1086/282802

Faulon, J.-L., Visco, D. P., and Pophale, R. S. (2003). The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* 43, 707–720. doi:10.1021/ci020345w

Favela-González, K. M., Hernández-Almanza, A. Y., and De La Fuente-Salcido, N. M. (2020). The value of bioactive compounds of cruciferous vegetables (Brassica) as antimicrobials and antioxidants: A review. *J. Food Biochem.* 44, e13414. doi:10.1111/jfbc.13414

Felsenstein, J. (2004). *PHYLIP (phylogeny inference package) version 3.6*. Distributed by the author. Available at: http://www.evolution.gs.washington.edu/phylip.html.

Ferreres, F., Llorach, R., and Gil-Izquierdo, A. (2004). Characterization of the interglycosidic linkage in di-tri-tetra- and pentaglycosylated flavonoids and differentiation of positional isomers by liquid chromatography/electrospray ionization tandem mass spectrometry. *J. Mass Spectrom.* 39, 312–321. doi:10.1002/jms.586

Fligner, M. A., Verducci, J. S., and Blower, P. E. (2002). A modification of the jaccard–tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics* 44, 110–119. doi:10.1198/004017002317375064

Francisco, M., Moreno, D. A., Cartea, M. E., Ferreres, F., García-Viguera, C., and Velasco, P. (2009). Simultaneous identification of glucosinolates and phenolic compounds in a representative collection of vegetable Brassica rapa. *J. Chromatogr. A* 1216, 6611–6619. doi:10.1016/j.chroma.2009.07.055

Gascuel, O. (1997). BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* 14, 685–695. doi:10.1093/oxfordjournals.molbev.a025808

Ghareaghajlou, N., Hallaj-Nezhadi, S., and Ghasempour, Z. (2021). Red cabbage anthocyanins: Stability, extraction, biological activities and applications in food systems. *Food Chem.* 365, 130482. doi:10.1016/j.foodchem.2021.130482

Guha, R. (2007). Chemical informatics functionality in R. *J. Stat. Softw.* 18, 1–16. doi:10.18637/jss.v018.i05

Gülçin, İ., Elias, R., Gepdiremen, A., and Boyer, L. (2006). Antioxidant activity of lignans from fringe tree (Chionanthus virginicus L.). *Eur. Food Res. Technol.* 223, 759–767. doi:10.1007/s00217-006-0265-5

Guo, N., Wu, J., Zheng, S., Cheng, F., Liu, B., Liang, J., et al. (2015). Anthocyanin profile characterization and quantitative trait locus mapping in zicaitai (Brassica rapa L. ssp. chinensis var. purpurea). *Mol. Breed.* 35, 113. doi:10.1007/s11032-015-0237-1

Hano, C. F., Dinkova-Kostova, A. T., Davin, L. B., Cort, J. R., and Lewis, N. G. (2021). Editorial: Lignans: Insights into their biosynthesis, metabolic engineering, analytical methods and health benefits. *Front. Plant Sci.* 11, 630327. doi:10.3389/fpls.2020.630327

Haq, I. U., Khan, S., Awan, K. A., and Iqbal, M. J. (2021). Sulforaphane as a potential remedy against cancer: Comprehensive mechanistic review. *J. Food Biochem.* 46, e13886. doi:10.1111/jfbc.13886

Harbaum, B., Hubbermann, E. M., Wolff, C., Herges, R., Zhu, Z., and Schwarz, K. (2007). Identification of flavonoids and hydroxycinnamic acids in pak choi varieties (Brassica campestris L. ssp. chinensis var. communis) by HPLC–ESI-MS n and NMR and their quantification by HPLC–DAD. *J. Agric. Food Chem.* 55, 8251–8260. doi:10.1021/jf071314+

Heinze, M., Hanschen, F. S., Wiesner-Reinhold, M., Baldermann, S., Gräfe, J., Schreiner, M., et al. (2018). Effects of developmental stages and reduced UVB and low UV conditions on plant secondary metabolite profiles in Pak Choi (Brassica rapa subsp. chinensis). *J. Agric. Food Chem.* 66, 1678–1692. doi:10.1021/acs.jafc.7b03996

Hillis, D. M., and Huelsenbeck, J. P. (1992). Signal, noise, and reliability in molecular phylogenetic analyses. *J. Hered.* 83, 189–195. doi:10.1093/oxfordjournals.jhered.a111190

Ihaka, R., and Gentleman, R. (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Statistics* 5, 299–314. doi:10.1080/10618600.1996.10474713

Jeon, J., Lim, C. J., Kim, J. K., and Park, S. U. (2018). Comparative metabolic profiling of green and purple pakchoi (Brassica rapa subsp. chinensis). *Molecules* 23, E1613. doi:10.3390/molecules23071613

Jing, P., Song, L.-H., Shen, S.-Q., Zhao, S.-J., Pang, J., and Qian, B.-J. (2014). Characterization of phytochemicals and antioxidant activities of red radish brines during lactic acid fermentation. *Molecules* 19, 9675–9688. doi:10.3390/molecules19079675

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., et al. (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26, 1463–1464. doi:10.1093/bioinformatics/btq166

Kim, J. K., and Park, S. U. (2018). Current results on the biological and pharmacological activities of Indole-3-carbinol. *EXCLI J.* 17, 181–185. doi:10.17179/excli2017-1028

Klopsch, R., Witzel, K., Artemyeva, A., Ruppel, S., and Hanschen, F. S. (2018). Genotypic variation of glucosinolates and their breakdown products in leaves of Brassica rapa. *J. Agric. Food Chem.* 66, 5481–5490. doi:10.1021/acs.jafc.8b01038

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., and Neumann, S. (2012). CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* 84, 283–289. doi:10.1021/ac202450g

Kyriacou, M. C., El-Nakhel, C., Pannico, A., Graziani, G., Zarrelli, A., Soteriou, G. A., et al. (2021). Ontogenetic variation in the mineral, phytochemical and yield attributes of brassicaceous microgreens. *Foods* 10, 1032. doi:10.3390/foods10051032

Landry, L. G., Chapple, C. C., and Last, R. L. (1995). Arabidopsis mutants lacking phenolic sunscreens exhibit enhanced ultraviolet-B injury and oxidative damage. *Plant Physiol.* 109 (4), 1159–1166. doi:10.1104/pp.109.4.1159

Lee, D., Park, J. Y., Lee, S., and Kang, K. S. (2021). *In vitro* studies to assess the α-glucosidase inhibitory activity and insulin secretion effect of isorhamnetin 3-O-glucoside and quercetin 3- O-glucoside isolated from Salicornia herbacea. *Processes* 9, 483. doi:10.3390/pr9030483

Liang, Y.-S., Choi, Y. H., Kim, H. K., Linthorst, H. J. M., and Verpoorte, R. (2006). Metabolomic analysis of methyl jasmonate treated Brassica rapa leaves by 2-dimensional NMR spectroscopy. *Phytochemistry* 67, 2503–2511. doi:10.1016/j.phytochem.2006.08.018

Lin, L.-Z., Sun, J., Chen, P., and Harnly, J. (2011). UHPLC-PDA-ESI/HRMS/MSn analysis of anthocyanins, flavonol glycosides, and hydroxycinnamic acid derivatives in red mustard greens (Brassica juncea coss variety). *J. Agric. Food Chem.* 59, 12059–12072. doi:10.1021/jf202556p

Liu, Y., Rossi, M., Liang, X., Zhang, H., Zou, L., and Ong, C. N. (2020). An integrated metabolomics study of glucosinolate metabolism in different Brassicaceae genera. *Metabolites* 10, 313. doi:10.3390/metabo10080313

Managa, M. G., Sultanbawa, Y., and Sivakumar, D. (2020). Effects of different drying methods on untargeted phenolic metabolites, and antioxidant activity in Chinese cabbage (Brassica rapa L. subsp. chinensis) and nightshade (solanum retroflexum dun.). *Molecules* 25, 1326. doi:10.3390/molecules25061326

Mandrich, L., and Caputo, E. (2020). Brassicaceae-derived anticancer agents: Towards a green approach to beat cancer. *Nutrients* 12, 868. doi:10.3390/nu12030868

Morreel, K., Dima, O., Kim, H., Lu, F., Niculaes, C., Vanholme, R., et al. (2010a). Mass spectrometry-based sequencing of lignin oligomers. *Plant Physiol.* 153, 1464–1478. doi:10.1104/pp.110.156489

Morreel, K., Kim, H., Lu, F., Dima, O., Akiyama, T., Vanholme, R., et al. (2010b). Mass spectrometry-based fragmentation as an identification tool in lignomics. *Anal. Chem.* 82, 8095–8105. doi:10.1021/ac100968g

Morreel, K., Saeys, Y., Dima, O., Lu, F., Van De Peer, Y., Vanholme, R., et al. (2014). Systematic structural characterization of metabolites in Arabidopsis via candidate substrate-product pair networks. *Plant Cell.* 26, 929–945. doi:10.1105/tpc.113.122242

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., et al. (2012). How to measure and test phylogenetic signal. *Methods Ecol. Evol.* 3, 743–756. doi:10.1111/j.2041-210X.2012.00196.x

Olsen, H., Aaby, K., and Borge, G. I. A. (2009). Characterization and quantification of flavonoids and hydroxycinnamic acids in curly kale (*Brassica oleracea* L. Convar. Acephala var. Sabellica) by HPLC-DAD-ESI-MSn. *J. Agric. Food Chem.* 57, 2816–2825. doi:10.1021/jf803693t

Olszewska, M. A., Granica, S., Kolodziejczyk-Czepas, J., Magiera, A., Czerwińska, M. E., Nowak, P., et al. (2020). Variability of sinapic acid derivatives during germination and their contribution to antioxidant and anti-inflammatory effects of broccoli sprouts on human plasma and human peripheral blood mononuclear cells. *Food Funct.* 11, 7231–7244. doi:10.1039/d0fo01387k

Padilla, G., Cartea, M. E., Velasco, P., De Haro, A., and Ordás, A. (2007). Variation of glucosinolates in vegetable crops of Brassica rapa. *Phytochemistry* 68, 536–545. doi:10.1016/j.phytochem.2006.11.017

Paul, S., Geng, C.-A., Yang, T.-H., Yang, Y.-P., and Chen, J.-J. (2019). Phytochemical and health- beneficial progress of turnip (Brassica rapa). *J. Food Sci.* 84, 19–30. doi:10.1111/1750-3841.14417

Qin, G., Liu, C., Li, J., Qi, Y., Gao, Z., Zhang, X., et al. (2020). Diversity of metabolite accumulation patterns in inner and outer seed coats of pomegranate: Exploring their relationship with genetic mechanisms of seed coat development. *Hortic. Res.* 7, 10. doi:10.1038/s41438-019-0233-4

Rahim, A., Mostofa, M. G., Sadik, M. G., Rahman, M. a. A., Khalil, M. I., Tsukahara, T., et al. (2021). The anticancer activity of two glycosides from the leaves of Leea aequata L. *Nat. Prod. Res.* 35, 5867–5871. doi:10.1080/14786419.2020.1798661

Raiola, A., Errico, A., Petruk, G., Monti, D. M., Barone, A., and Rigano, M. M. (2018). Bioactive compounds in Brassicaceae vegetables with a role in the prevention of chronic diseases. *Molecules* 23, 15. doi:10.3390/molecules23010015

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47. doi:10.1093/nar/gkv007

Salehi, B., Quispe, C., Butnariu, M., Sarac, I., Marmouzi, I., Kamle, M., et al. (2021). Phytotherapy and food applications from Brassica genus. *Phytother. Res.* 35, 3590–3609. doi:10.1002/ptr.7048

Senizza, A., Rocchetti, G., Mosele, J. I., Patrone, V., Callegari, M. L., Morelli, L., et al. (2020). Lignans and gut microbiota: An interplay revealing potential health implications. *Molecules* 25, 5709. doi:10.3390/molecules25235709

Shahaf, N., Rogachev, I., Heinig, U., Meir, S., Malitsky, S., Battat, M., et al. (2016). The WEIZMASS spectral library for high-confidence metabolite identification. *Nat. Commun.* 7, 12423. doi:10.1038/ncomms12423

Smith, C. A., Want, E. J., O'maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 78, 779–787. doi:10.1021/ac051437y

Soengas, P., Cartea, M. E., Velasco, P., and Francisco, M. (2018). Endogenous circadian rhythms in polyphenolic composition induce changes in antioxidant properties in Brassica cultivars. *J. Agric. Food Chem.* 66, 5984–5991. doi:10.1021/acs.jafc.8b01732

Soleymani, S., Habtemariam, S., Rahimi, R., and Nabavi, S. M. (2020). The what and who of dietary lignans in human health: Special focus on prooxidant and antioxidant effects. *Trends Food Sci. Technol.* 106, 382–390. doi:10.1016/j.tifs.2020.10.015

Song, B., Xu, H., Chen, L., Fan, X., Jing, Z., Chen, S., et al. (2020). Study of the relationship between leaf color formation and anthocyanin metabolism among different purple pakchoi lines. *Molecules* 25, 4809. doi:10.3390/molecules25204809

Sun, J., Lin, L.-Z., and Chen, P. (2012). Study of the mass spectrometric behaviors of anthocyanins in negative ionization mode and its applications for characterization of anthocyanins and non- anthocyanin polyphenols. *Rapid Commun. Mass Spectrom.* 26, 1123–1133. doi:10.1002/rcm.6209

Sun, J., Xiao, Z., Lin, L.-Z., Lester, G. E., Wang, Q., Harnly, J. M., et al. (2013). Profiling polyphenols in five Brassica species microgreens by UHPLC-PDA-ESI/HRMSn. *J. Agric. Food Chem.* 61, 10960–10970. doi:10.1021/jf401802n

Van De Mortel, J. E., De Vos, R. C. H., Dekkers, E., Pineda, A., Guillod, L., Bouwmeester, K., et al. (2012). Metabolic and transcriptomic changes induced in Arabidopsis by the rhizobacterium Pseudomonas fluorescens SS101. *Plant Physiol.* 160, 2173–2188. doi:10.1104/pp.112.207324

Wang, S., Wu, C., Li, X., Zhou, Y., Zhang, Q., Ma, F., et al. (2017). Syringaresinol-4-O-β-d-glucoside alters lipid and glucose metabolism in HepG2 cells and C2C12 myotubes. *Acta Pharm. Sin. B* 7, 453–460. doi:10.1016/j.apsb.2017.04.008

Wiczkowski, W., Szawara-Nowak, D., and Topolska, J. (2013). Red cabbage anthocyanins: Profile, isolation, identification, and antioxidant activity. *Food Res. Int.* 51, 303–309. doi:10.1016/j.foodres.2012.12.015

Wiesner-Reinhold, M., Dutra Gomes, J. V., Herz, C., Tran, H. T. T., Baldermann, S., Neugart, S., et al. (2021). Subsequent treatment of leafy vegetables with low doses of UVB-radiation does not provoke cytotoxicity, genotoxicity, or oxidative stress in a human liver cell model. *Food Biosci.* 43, 101327. doi:10.1016/j.fbio.2021.101327

Wu, X., and Prior, R. L. (2005). Identification and characterization of anthocyanins by high-performance liquid Chromatography–Electrospray Ionization–Tandem mass spectrometry in common foods in the United States: Vegetables, nuts, and grains. *J. Agric. Food Chem.* 53, 3101–3113. doi:10.1021/jf0478861

Yang, B., and Quiros, C. F. (2010). Survey of glucosinolate variation in leaves of Brassica rapa crops. *Genet. Resour. Crop Evol.* 57, 1079–1089. doi:10.1007/s10722-010-9549-5

Yeo, H. J., Baek, S.-A., Sathasivam, R., Kim, J. K., and Park, S. U. (2021). Metabolomic analysis reveals the interaction of primary and secondary metabolism in white, pale green, and green pak choi (Brassica rapa subsp. chinensis). *Appl. Biol. Chem.* 64, 3. doi:10.1186/s13765-020-00574-2

Yokozawa, T., Kim, H. Y., Cho, E. J., Choi, J. S., and Chung, H. Y. (2002). Antioxidant effects of isorhamnetin 3, 7-di-O-β-d-glucopyranoside isolated from mustard leaf (Brassica juncea) in rats with streptozotocin-induced diabetes. *J. Agric. Food Chem.* 50, 5490–5495. doi:10.1021/jf0202133

Zou, L., Tan, W. K., Du, Y., Lee, H. W., Liang, X., Lei, J., et al. (2021). Nutritional metabolites in Brassica rapa subsp. chinensis var. parachinensis (choy sum) at three different growth stages: Microgreen, seedling and adult plant. *Food Chem.* 357, 129535. doi:10.1016/j.foodchem.2021.129535

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 21 December 2022
DOI 10.3389/fmolb.2022.1042231

Exercise blood-drop metabolic profiling links metabolism with perceived exertion

Tobias Opialla[1,2,3†], Benjamin Gollasch[4†], Peter H. J. L. Kuich[1†], Lars Klug[4], Gabriele Rahn[4], Andreas Busjahn[4,5], Simone Spuler[2], Michael Boschmann[4], Jennifer A. Kirwan[3], Friedrich C. Luft[4]* and Stefan Kempa[1,3]*

[1]Department of Proteomics and Metabolomics Max-Delbrück-Center for Molecular Medicine Berlin, Berlin Institute for Medical Systems Biology, Berlin, Germany, [2]Muscle Research Unit, Experimental and Clinical Research Center, A Joint Collaboration Between Max-Delbr ück-Center and Charité Universitätsmedizin Berlin, Berlin, Germany, [3]Berlin Institute of Health Metabolomics Platform, Charite Universitätsmedizin Berlin, Berlin, Germany, [4]Experimental and Clinical Research Unit, Joint collaboration between Max-Delbr ück-Center and Charité Universitätsmedizin Berlin, Berlin, Germany, [5]HealthTwiSt GmbH, Berlin, Germany

OPEN ACCESS

EDITED BY
Wolfram Weckwerth,
University of Vienna, Austria

REVIEWED BY
Georgios Theodoridis,
Aristotle University of Thessaloniki, Greece
Fulvio Mattivi,
Fondazione Edmund Mach, Italy

*CORRESPONDENCE
Stefan Kempa,
stefan.kempa@mdc-berlin.de
Friedrich C. Luft,
friedrich.luft@charite.de

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 12 September 2022
ACCEPTED 09 December 2022
PUBLISHED 21 December 2022

CITATION
Opialla T, Gollasch B, Kuich PHJL, Klug L, Rahn G, Busjahn A, Spuler S, Boschmann M, Kirwan JA, Luft FC and Kempa S (2022), Exercise blood-drop metabolic profiling links metabolism with perceived exertion.
Front. Mol. Biosci. 9:1042231.
doi: 10.3389/fmolb.2022.1042231

Background: Assessing detailed metabolism in exercising persons minute-to-minute has not been possible. We developed a “drop-of-blood” platform to fulfill that need. Our study aimed not only to demonstrate the utility of our methodology, but also to give insights into unknown mechanisms and new directions.

Methods: We developed a platform, based on gas chromatography and mass spectrometry, to assess metabolism from a blood-drop. We first observed a single volunteer who ran 13 km in 61 min. We particularly monitored relative perceived exertion (RPE). We observed that 2,3-bisphosphoglycerate peaked at RPE in this subject. We next expanded these findings to women and men volunteers who performed an RPE-based exercise protocol to RPE at $Fi\ O_2$ 20.9% or $Fi\ O_2$ 14.5% in random order.

Results: At 6 km, our subject reached his maximum relative perceived exertion (RPE); however, he continued running, felt better, and finished his run. Lactate levels had stably increased by 2 km, ketoacids increased gradually until the run’s end, while the hypoxia marker, 2,3 bisphosphoglycerate, peaked at maximum relative perceived exertion. In our normal volunteers, the changes in lactate, pyruvate, *ß* hydroxybutyrate and *a* hydroxybutyrate were not identical, but similar to our model proband runner.

Conclusion: Glucose availability was not the limiting factor, as glucose availability increased towards exercise end in highly exerted subjects. Instead, the tricarboxylic acid→oxphos pathway, lactate clearance, and thus and the oxidative capacity appeared to be the defining elements in confronting maximal exertion. These ideas must be tested further in more definitive studies. Our preliminary work suggests that our single-drop methodology could be of great utility in studying exercise physiology.

KEYWORDS

gas chromatography, blood drop sampling, relative perceived exertion, hypoxia, metabolomics

1 Introduction

Physical exercise is healthy. (Castillo-Garzón et al., 2006; Warburton et al., 2006) The benefits are the same whether the exercise is recreational or occupational. (Lear et al., 2017) There are numerous assessment signs, including maximal oxygen consumption, muscle strength, muscular endurance, responses in heart rate, and others. (Nindl et al., 2015) Exercisers experience regular cycles of physiological stress accompanied by transient inflammation, oxidative stress, and immune perturbations. (Herrmann et al., 2015; Loprinzi, 2015; Palacios et al., 2015) The relevance of such findings to normal healthy individuals is not always clear. Furthermore, combining these diverse variables into an understandable paradigm is difficult.

Metabolomics is the scientific study of chemical processes involving metabolites, including an assessment of the unique chemical fingerprints that specific cellular processes leave behind, following a metabolic event. (Bassini & Cameron, 2014; Gong et al., 2017) The metabolome represents the collection of all metabolites in a cell, tissue, organ or organism, which are the end products of cellular processes. Gas chromatography–mass spectrometry (GC–MS) based metabolomics, as well as other technologies, now enable us to vastly increase our panoramic inspection of these processes. (Gong et al., 2017)

The field of exercise metabolomics is at its beginning. Klein and others reviewed a number metabolomics studies of bio-fluids and describe analytical platforms (Klein et al., 2021). The most comprehensive analysis of molecular changes post exercise was published recently by Contrepois and others (Contrepois et al., 2020). In a number of individuals metabolites, proteins and mRNA expression was studied post-exercise and all measured parameters were correlated to insulin resistance or sensitivity. Studies so far usually describe metabolic changes post exercise in plasma or serum.

Concurrently, we were interested in developing an analytical strategy that allows a simplified (dropwise) blood sampling that can be generally applied clinically or even to monitor subjects at or in the field. We have used a liquid–liquid sampling method for whole blood sampling that stabilizes the metabolome instantly and is optimal to monitor a broad spectrum of intermediates from central metabolism that represents the energy providing machinery. Our methods brings metabolomics into the practicable clinical arena.

Applicability commonly results after initial clinical observations. We performed initial observations in a model proband runner (MPR) who recorded his relative perceived (admittedly subjective) exertion (RPE). (Borg, 1982) Nonetheless, we had obtained an "*n*-of-one" dataset. We therefore aimed to investigate the practical utility of metabolomics. We extracted nine key metabolic features that seem to describe the feeling states observed in RPE on a personalized level. Many of these nine metabolic features are directly involved with oxygenation status. To test our observation that the subjective state, RPE, is reflected in the metabolome in a wider population, we designed a follow-up study in normal, recreationally active, women and men across a broad fitness and age spectrum, to inspect metabolomics outcomes.

We chose alterations of oxygen availability as additional stressor corresponding to normobaric-altitude sea level *versus* simulated 3,000 m ($\mathit{Fi}\,\mathrm{O}_2$ 14.5%) in a randomized setting, which is also employed in terms of fitness strategies. (Hobbins et al., 2017) We found that RPE is reflected in the metabolome in a wider population and underscored the ratios for liver metabolism previously established in 1967 by KREBS and associates. (Krebs, 1967; Williamson et al., 1967)

2 Results

Our MPR (Figure 1 and Table S1) ran six laps (about 13 km in total) over an irregular terrain. He reported his perceptions of energy availability on a qualitative exertion scale (relative perceived exertion) RPE scale (Figures 1A–E). (Borg, 1982) Our subject was confronted with exhaustion on lap 3. He recovered, and remained approximately at the same performance level till the end of the run. In the drops of blood, we annotated 276 separate peaks of which 93 were identified using a comparison mix of commercial standards analyzed in the same batch (Table S2) (see methods and Opialla et al., 2020) giving a general central carbon metabolism coverage (Supplementary Figure S1). Lactate levels increased early in our subject, glucose remained flat, acetoacetate and *ß*-hydroxybutyrate increased progressively, while 2,3-biphosphoglycerate (2,3-BPG) peaked in lap 3 (Figures 1B–E). Succinate and other TCA intermediates also rose early, except for citrate, which only increased concurrently with 2,3-BPG levels. When we performed unsupervised hierarchical clustering, we were surprised that the samples clustered according to feeling state ("I am done"/"I feel ok"): most related to baseline and recovery; namely relative perceived exertion (RPE, Figure 1F), and not as one might expect according to exercise status.

We explored the co-behavior of metabolic changes and subjective feeling states of our MPR. Using a factor analysis we identified nine key metabolites that possessed the highest eigen-values. Based on these nine key metabolites, we were able to reproduce the clustering (Figure 1G, Supplementary Table S1) of our subject's "feeling" states. We assigned physiological states to the metabolic features and set the dynamic changes in

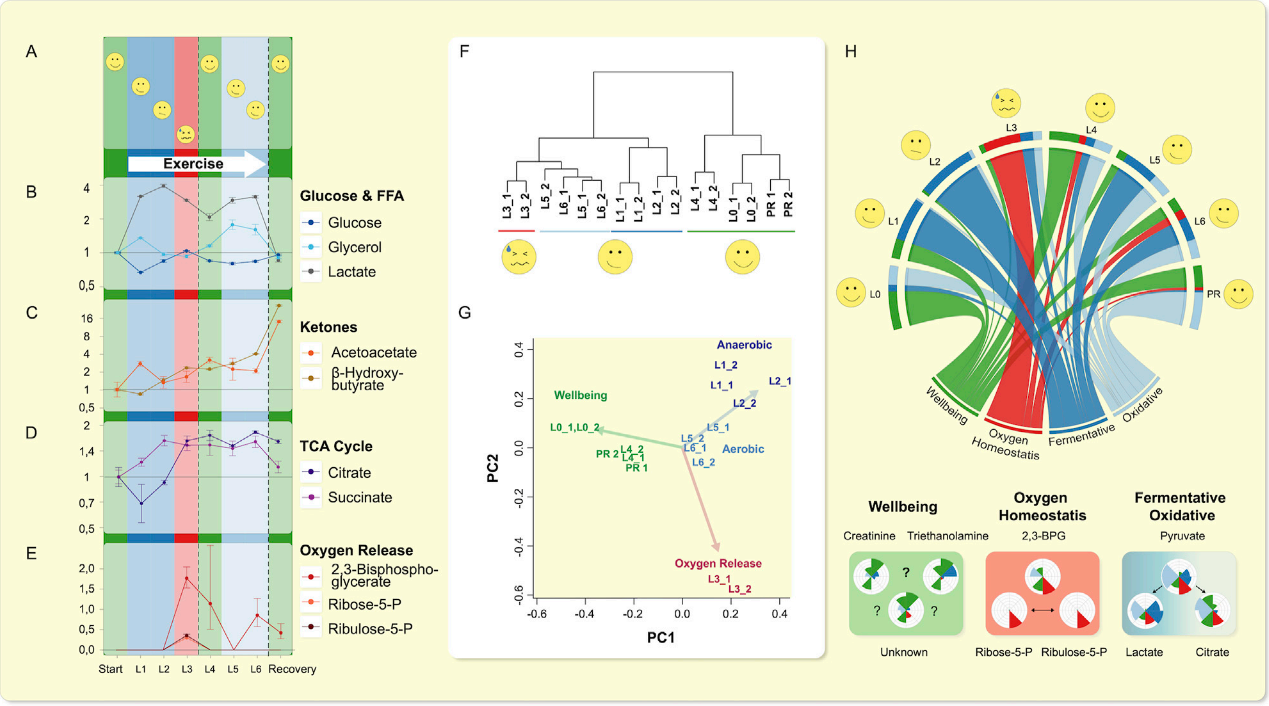


FIGURE 1

Single drop-blood analysis from an exerciser. **(A–E)** The rating of perceived exercise (RPE) variable is displayed as a subjective scale. Lactate (as opposed to glucose) increases similar to TCA-intermediates (Supplemental Figure S14), except for citrate. **(D)** Citrate only increased when oxygen release was promoted as indicated by rise of 2,3-BPG and PPP-intermediates Ribulose-5P and Ribose-5P above LOQ. **(E)** Glycerol showed two phases (corroborated by fatty acids, Supplementary Figure S15) of release and reaches baseline value, while **(B)** ketone bodies accumulate. **(F)** Hierarchical clustering of the technical replicates throughout the run on metabolic features is shown. Clustering, using all polar metabolites, grouped the sampling time-points in accordance with the volunteer's self-perception. Lap 3 (L3), in which exhaustion occurred, was least related to beginning recovery (PR), and the most at maximum RPE (L4, Lap 4), while these three timely very diverse laps were closest related. **(G)** Visualization of a nine-metabolite principle-component analysis found by factor analysis representing the explanatory components of all data (relative activity of anaerobic and TCA cycle activity (pyruvate/lactate vs. pyruvate/citrate), oxygen release, and oxidative stress, as well as markers of "feeling energetic" are shown. These metabolites reproduced the principal relationships of metabolic states and transitions as observed within all data. **(H)** The visualization of the progression through the exercise regime is summarized in a *circos* plot (Krzywinski et al., 2009) using the nine explanatory metabolic features and relationships depicted in NIGHTINGALE plots below. The *circos* plot shows progression of each of the four factors determined in **(G)**, both for each factor (lower "root") and through each lap (start at left "L0"). The NIGHTINGALE plots below show the progression of the nine key metabolites (see main text). Beginning at rest, the subject entered an initial anaerobic phase, which was followed by an energy crisis caused by insufficient oxygen availability. Resolution in Lap 4 was accompanied by a transient increase in indicators of feeling "energetic" (i.e. low RPE) and immediately followed the successful transition to oxidative TCA cycle driven energy supply that remained elevated until after exercise completion.

relationship to each other and show them in a *circos* plot (Krzywinski et al., 2009) (Figure 1H). Beginning at rest, the subject entered an initial anaerobic phase, which was followed by an episode likely caused by insufficient oxygen availability. The resolution was accompanied by a transient increase in indicators of feeling more robust and immediately followed the successful transition to oxidative KREBS' tricarboxylic acid cycle (TCA)-driven energy supply, that remained elevated until after exercise completion (Figure 1H). We observed that creatine and triethanolamine were elevated when perceived maximum effort was "overcome". We interpreted the 2,3-bisphosphoglycerate, ribose-5-phosphate, and ribulose-5-phosphate levels as indicating changes in oxygen homeostasis, and pyruvate, lactate, citrate ratios, as a switch between anaerobic and aerobic metabolism. Since 2,3-bisphosphglycerate is involved in shifting the oxygen-hemoglobin-saturation curve rightward (Benesch & Benesch, 1967; Chiba & Sasaki, 1978), we connected the results into our subsequent research plan.

To study metabolism at exercise further (Figure 2) and the influence of oxygen availability, we next recruited 26 normal women and men volunteers across a broad age and fitness-level spectrum (Supplementary Table S3), who were randomized (cross-over) to perform at oxygen levels ($Fi\ O_2$) at sea level or at a simulated 3,000 m altitude and at a running speed corresponding to 65% of maximum power output according to JONES *et al.* (Jones et al., 1985) (Figure 2A). The subjects all lived in the area of Berlin, Germany (≈50 m NN), reported various degrees of compliance to accepted healthy life styles

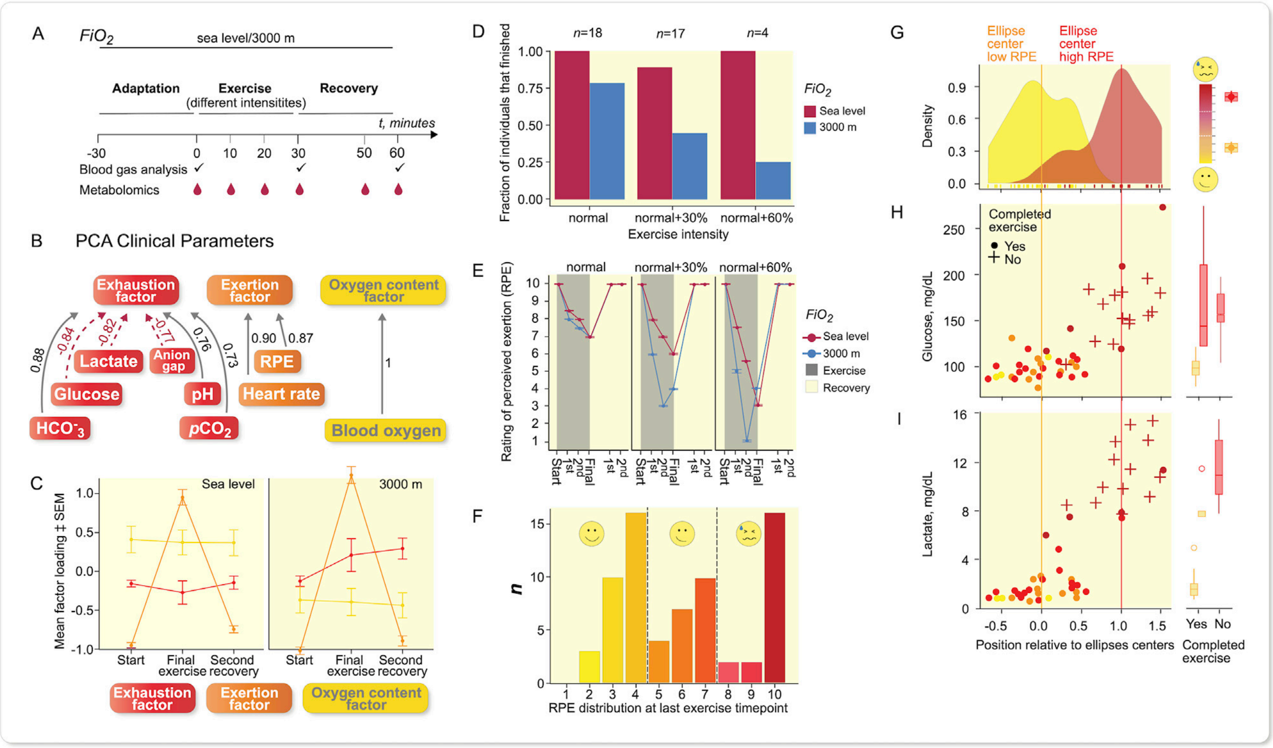


FIGURE 2
Results from recreational (normal) volunteers. Sampling scheme for larger panel study of recreational athletes. **(A)** Study design outline. **(B)** PCA of clinical parameters revealed three different items, namely exhaustion, exertion, and oxygen. **(C)** When plotting mean factor loadings of PCA of the three principle components, the data indicate that the oxygen content factor was influenced by *Fi* O $_2$ albeit constant throughout the experiment. Exertion was high at last exercise time-point and reaches baseline values during recovery, while the exhaustion increases under more hypoxic conditions and remains elevated during recovery, error bars: SEM, *n* = 39 at each oxygenation level. **(D)** The fraction of subjects being able to complete the exercise bouts reflects that *Fi* O $_2$ had a stronger influence on that with increasing power output (running speed), *n* given in the figure refers to each oxygenation level. **(E)** Individual RPE values were more different between power output levels; lower *Fi* O $_2$ led to more marked exertion according to rating of perceived exertion (RPE), error bars: SEM, n-numbers as in **(D)**. **(F)** Histogram of RPE-distribution at last exercise time-point revealed a trimodal distribution, which was not explained by power output or *Fi* O $_2$ alone (see Supplementary Figure S4). **(G)** PCA of metabolite level-fold changes final exercise time-point vs. baseline from experiments leading to distinct levels of RPE, revealed a separation of less exerted individuals from highly exerted individuals; density plot shows distribution along vector between confidence ellipses' centers (Supplementary Figure S5). Glucose **(H)** and lactate **(I)** were major contributors to the separation. A + indicates that exercise was not completed, ● shows exercise completed, color reflects RPE at last exercise time-point. **(G–I)**: low RPE: *n* = 29, high RPE: *n* = 20.

and a broad gamut of physical fitness activities ranging from very little training to dedicated daily fitness schedules. The "normal" running speed often was not appropriately challenging to some subjects. As we were interested in a state of high RPE ("exhaustion"), we encouraged repetition of the experiment on another day with increased running speed at 30% or even 60% faster (Supplementary Table S4).

The individual and mean exertional-related effects (Supplementary Table S5) show a substantial load on most subjects and demonstrated, that a 30% increase in effort and/or hypoxia were successful challenges (Figures 2D,E). Principal component analysis (PCA) shows the clinical variables in relation to exhaustion, exertion, and oxygen partial pressure (Figure 2B). A Spearman ranking of the variables is given (Figure S2), and a comparison between sea level and 3,000 m (Figure 2C). The exhaustion factor was certainly influenced by altitude as was the oxygen factor. The change in individual variables with exercise at two performance levels and two altitudes are most visible in the arterial blood gases, pCO$_2$, anion gap (AG), and lactate values (Supplementary Table S5, Supplementary Figure S18). The fraction of persons completing the run decreased at the different exercise levels (Figure 2D), indicating that our normal recruits also commonly experienced their limits (Figure 2E). Obviously, there was a relationship between rating of perceived exertion (RPE), running speed, and oxygenation. Since the formula according to Jones *et al.* (Jones et al., 1985) does not accommodate for the subjects' training status, we looked deeper into RPE and found that differences in *Fi* O $_2$ only made a substantial difference at 30% running speed and a higher level (Figure 2E).

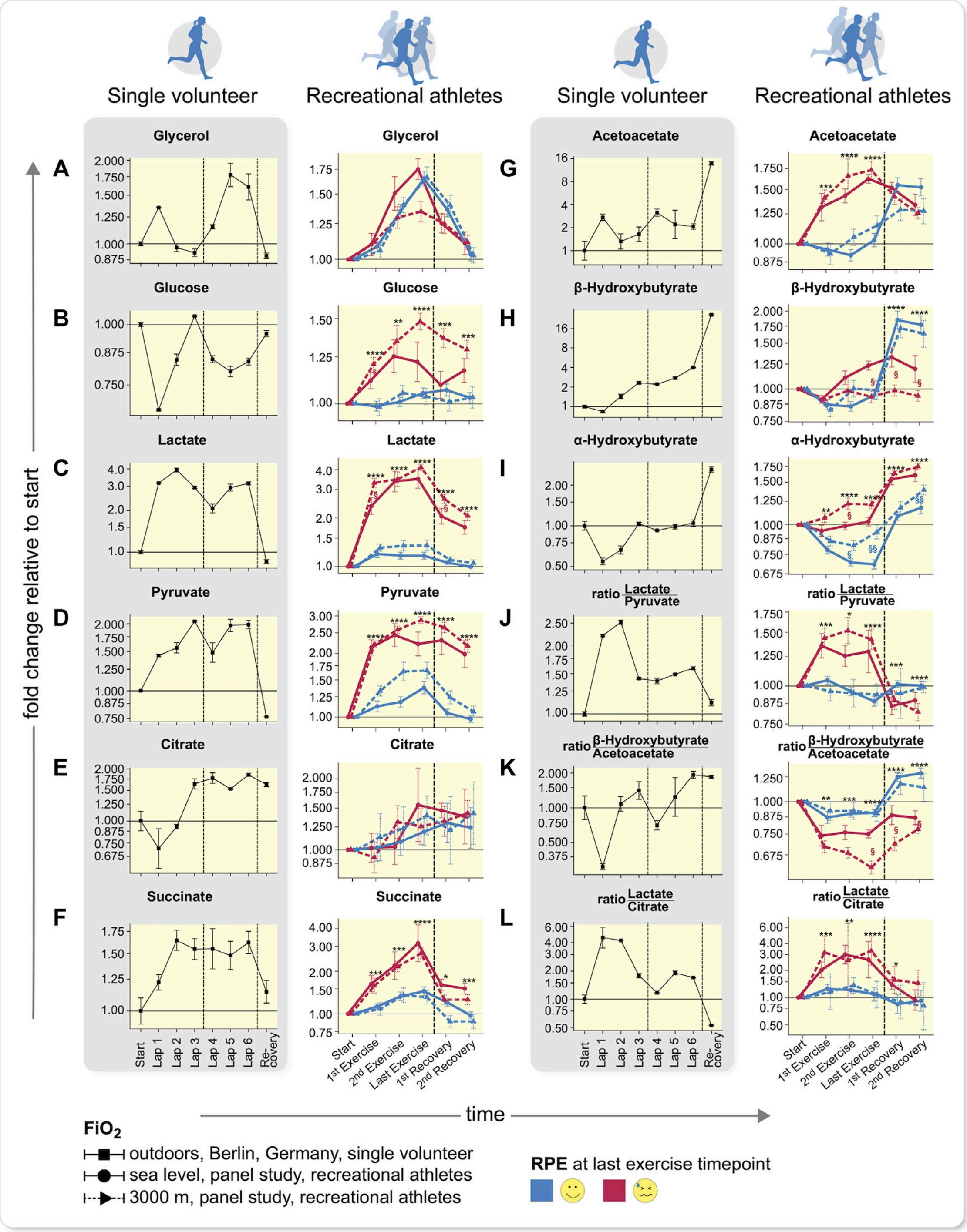


FIGURE 3
Time-profiles throughout exercise to recovery comparing single volunteer to recreational athletes. **(A–F)** Carbohydrate-based metabolite values, **(G–I)** ketone bodies, and **(J–L)** metabolite ratios in a single volunteer (left, error bars demonstrate deviations of duplicate measurments to anticipate the measurement accuracy) and recreational athletes (right) are shown (mean ± SEM). Difference between two technical replicate measurements are grouped according to $Fi\,O_2$ (solid: sea level, dotted: 3,000 m) and RPE at last exercise time-point. Dashed lines encompass comparable time intervals in different setups. In panels on the left only one recovery time-point was measured, while in the RPE right panels exertion was not overcome; however, exercise was discontinued after 30 min. The samples between the vertical lines in the shaded, left plots depict a state, after overcoming subjective exhaustion that could not be compared to samples from the larger cohort study (*p*-values from two sided WILCOXON-rank-sum (BENJAMINI-HOCHBERG FDR-corrected), *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ***$p < 0.0001$ of all values within one RPE-group, for clarity only significance vs. lowest RPE group is shown, but all comparisons were accounted for *p*-value correction, $^{\S}p < 0.05$; $^{\S\S}p < 0.01$ between hypoxia and normoxia of lowest RPE group. Explanations see main text. Low RPE: $n = 29$, high RPE: $n = 20$).

We found a trimodal distribution in RPE at final exercise time-point based on 77 runs (Figure 2F). These results were not based on performance level or FiO_2 alone (Supplementary Figures S3,4). Since RPE is somewhat subjective and because we wanted to investigate reflection of feeling state (RPE) in the blood metabolome, we concentrated on the runs finishing at low RPE and high RPE in the further analysis (55 runs total). After removing strong outliers according to HOTELLING's-T^2-test we were left with 46 samples. Grouping of the data revealed a separation of subjects with highest RPE-values from those with lowest RPE-values (Figure 2G, Supplementary Figure S5), indicated by very little overlap of 95% confidence ellipses. Glucose and lactate values showed corresponding increases (Figures 2H,I). These unsupervised analyses were performed on fold changes between baseline and final exercise time-point in the metabolomics data set (122 identified peak species across all samples). We extracted the contributions of individual metabolites along the vector between the 95% confidence ellipse's centers, similar to a dogleg plot of principal components. We could not underscore a role of 2,3-bisphosphoglycerate, but interpreted these data as indicating that glucose availability was not among rate-limiting factors.

We next more closely examined the main contributing metabolites and plotted their time-profiles of intensity-fold changes relative to exercise start from start to recovery (Figure 3). The data of our single runner and recreational volunteers are clearly delineated. We were interested in dissecting the mechanism involved in self-perceived maximal exercise. Since we observed a separation in the PCA according to RPE (Figure 2G, Supplementary Figure S5) and to a lesser extent to FiO_2 (Figure S7B), we separated the time-profiles (Figure 3) according to RPE-group (red/blue) and FiO_2 (solid/dashed lines). Significance values are indicated by asterisks (*) and significance represent FDR-adjusted (BENJAMINI–HOCHBERG) *p*-values from WILCOXON-tests between RPE-groups, § denotes false-discovery rate (FDR)-adjusted *p*-values from WILCOXON-tests between FiO_2 levels within the respective RPE-groups. The separation observed in PCA (Supplementary Figures S5,6) was mostly caused by glucose (Figure 3B), lactate (Figure 3C), and pyruvate (Figure 3D), while citrate values (Figure 3E) appeared less so. The ketone bodies, acetoacetate (Figure 3G) and *a*-hydroxybutyrate (Figure 3I), other sugars and polyols such as mannose, fructose, threonate, sorbitol, as well as alanine (Supplementary Figure S19), tri-ethanolamine-phosphate, TCA-intermediates, such as succinate (Figure 3F and Supplementary Figure S6) and glucose-6-phosphate (Supplementary Figure S20) also appeared discriminatory. Some separation according to FiO_2 was observed (Supplementary Figure S7B); however, exercise bouts under hypoxia, in which where RPE was low, group well with normoxia samples where RPE in general was less challenged. We interpret this result as indicating that exercise under hypoxia tends to lead to higher RPE. Other factors such as sex (Supplementary Figure S7C) or training state (Supplementary Figure S7D) showed no separation. However, when examining the main separator in our recreational volunteers, namely glucose, the data showed a strong separation between FiO_2 levels in individuals with high RPE.

Some individuals with higher RPE were not clearly separated from the less exhausted group. We filtered out those in highest RPE and already separated, and performed PCA again while keeping all RPE groups (Supplementary Figures S8–S10). Those individuals in the highest RPE, but not separated entirely from the lower exertion groups, were now separated from the subjects in the lower exertion groups along PC1 (Supplementary Figure S11). These remaining samples in the high RPE-group exhibited a similar but less pronounced phenotype for glucose and lactate (Supplementary Figure S11).

In accordance with our data with increasing exercise intensity, the amount of fat oxidized remained constant, while the additional energy is derived from glycogen and glucose. (van Loon et al., 2001) The glycerol values in our studies serve as a marker for fatty acid mobilization and were similar across all groups identified by RPE at the last exercise time-point (Figure 3A and Supplementary Figure S13). In hypoxia under high RPE less glycerol was mobilized, indicating a lower ability to oxidize fatty acids. In our athlete (MPR) (Figures 3A–C), an initial increase in glycerol and slight decrease in glucose was observed; however, when he reached his maximum RPE (Lap 3), these values had returned to baseline. In a second phase, while the TCA cycle was running (Supplementary Figure S14), glycerol again increased together with an increase in fatty acids (Supplementary Figure S15) and fatty acid-derived ketone bodies acetoacetate (Figure 3G), as well as *ß*-hydroxybutyrate (Figure 3H). His lactate level quadrupled but was actually decreasing when he reached his maximum RPE. In our normal volunteers (Figures 3A–C), glucose levels increased, compared to those values in the less-exhausted subjects. We want to point out, that the increase in glucose was much more prominent in our exercisers under highest RPE and more pronounced at 3,000 m than at sea level. Under hypoxia, the values in the exercisers remained elevated, whereas under normoxia, the concentrations decreased coinciding with the accumulation/formation of *ß*-hydroxybutyrate, most pronounced during recovery. Lactate and pyruvate increased with RPE in our exercising subjects (Figures 3C,D right), while the difference between oxygen levels decreased with increasing RPE. (Supplementary Figure S13) In our MPR, lactate and pyruvate concentrations showed a profile quite similar to the highly exhausted subjects. We therefore focused our attention on TCA cycle that is downstream to glycolysis and generates more ATP.

For citrate (Figure 3E), we observed similar profiles in both studies, albeit with high variability between the different subjects. Succinate and other TCA cycle intermediates (Figure 3F, Supplementary Figure S16) also increased with RPE but less at

3,000 m than at sea level. In a few selected individuals, we noted a marked increase of citrate, similar as during RPE in our MPR (Lap 3). The data suggests that not only single metabolites account for the changes in feeling state, but instead their interrelationships and ratios to one another. These observations would be in accord with those of KREBS' findings of metabolite ratios in liver (see below), that show oxygenation status according to lactate, pyruvate and ketone-bodies. (Krebs, 1967)

We were particularly interested in ketone bodies. Acetoacetate (Figure 3G) concentrations increased both under normoxia and hypoxia, more commonly in the subjects arriving at maximal RPE. The *ß*-hydroxybutyrate values increased as well (Figure 3H), but not for the highest RPE group at 3,000 m where no increase was observed. In our MPR, acetoacetate increased already at the first time-point but decreased thereafter (Figure 3G), while in the larger subject panel at higher exhaustion acetoacetate increased and remained elevated (Figure 3G). The *ß*-hydroxybutyrate profile was similar in the MPR as in exercisers who reached a high exhaustion state at sea level but not in those under maximal RPE at 3,000 m. After the exercise session, ketone bodies accumulated as in the lower exhaustion-level groups, but to a much greater extent (16-fold) in the single MPR. In our volunteers, these values *versus* doubled in lower RPE group (Figures 3G–H).

The *a*-hydroxybutyrate concentration (Figure 3I) is a marker for early-onset insulin resistance. (Gall et al., 2010) The values increased according to RPE levels during exercise. The higher levels coincided with runs leading to elevated blood glucose levels and were already above baseline levels at the first exercise time-point in the respective runs. The general profile shape was similar in all RPE groups, while glucose and *a*-hydroxybutyrate exhibit a correlation (Supplementary Figure S17) consistent with the findings of GALL *et al.* (Gall et al., 2010) The data from our MPR underscore our technical approach. Furthermore, the data suggest that oxygen availability could be the energy-limiting factor. The fact that the glucose values increased, suggested that lack of glucose was not responsible for arriving at RPE. We therefore chose to explore energy availability. For continuous exercise, most energy is derived from oxidative-phosphorylation, thus oxygen availability is a likely highly influential factor.

We next studied the ratios of lactate/pyruvate and *ß*-hydroxybutyrate/acetoacetate which reflect the NAD^+/NADH ratio (redox-potential) in cytosol and mitochondria respectively according to KREBS (Krebs, 1967) (Figures 3J,K). Our observations are in accordance with these ratios that were initially established in liver: in high RPE lower *Fi* O $_2$ led to a more anaerobic and less aerobic metabolism, according to KREBS' ratios. This state-of-affairs was too low to satisfy energy needs from oxidative metabolism. The TCA cycle was apparently not running as fast as necessary (in relation to glycolysis) and the NAD^+ required for glycolysis was regenerated by lactate formation. Therefore, we observed a decrease in pH (Supplementary Figure S18). This conclusion was also underscored by the ratio of lactate/citrate. Relative lactate concentrations increased similarly in our MPR, who was able to overcome his discomfort, and in our subjects, who exerted themselves to a maximal degree to about a 4-fold increase (Figures 3C,D). The ratio of lactate/citrate increased similarly in our MPR and in this highly exerted subject group. However, in our MPR, the ratio decreased during exercise when he continued, while in the high RPE group the ratio remained. These findings suggest a higher citrate synthesis rate in our MPR after his exhaustive episode, while the recreational subjects running towards high RPE show many symptoms of metabolic acidosis as one might expect in type-1 diabetes (low pCO_2. Low bicarbonate, low pH, and high anion gap, but also high levels of lactate, acetoacetate, and glucose). During intense exercise and low oxygen availability, we observed higher glucose levels. High lactate-pyruvate ratios (Figure 3J) and increased glucose at the same time suggest that metabolism was not able to utilize the available glucose through glycolysis and that lactate clearance was fully engaged. Integrating these results, we suggest that mitochondrial metabolism was insufficient to process the resulting pyruvate. Alanine levels that were a separator between the RPE groups (Supplementary Figure S19), coincide with increased pyruvate levels and indicate a higher reliance on the CORI and CAHILL cycle.

The *a*-hydroxybutyrate (αOHB) and glucose levels were directly correlated (Figures 3B,I and Supplementary Figure S17). The αOHB concentration has been implicated as a marker sensitive to changes in glucose levels in type 2 diabetes-prone patients. (Gall et al., 2010) We observed an increase early during exercise in those subjects who later developed the highest glucose levels during exercise. Overall, the formation of lactate was similar within RPE groups; however, lactate clearance was lower at the simulated 3,000 m altitude. When similar exertional levels were achieved under different *Fi* O $_2$ levels, we observed that glucose accumulated while the buffer systems in the blood were maximally challenged (Supplementary Figure S18). This observation suggests that lactate cannot be further metabolized while glucose is being funneled into the blood stream under high-energy demand conditions. Overall, empty glucose stores do not appear

To explain exhaustion, while oxygen availability and mitochondrial capacity would appear to be primarily responsible.

3 Discussion

We conclude that our translational experiment had utility. From single blood drops during exercising individuals, we can elucidate what is going on, better than singlularly measuring the current parameters. We believe that the most important finding in this study is that a single drop of capillary blood is useful in evaluating metabolism during exercise in contrast to metabolic studies done so far. We initially studied a serious hobby MPR, who led the way and then women and men volunteers who subjected themselves to an exercise protocol designed to address their perceived performance levels. The more strenuous exercise in terms of oxidative capacity the more glucose is used. (van Loon et al., 2001) We observed that glucose availability appeared not to be the limiting factor, but rather implicate the tricarboxylic acid→oxidative phosphorylation pathway. We were able to reduce the metabolomics dataset from a single volunteer to nine key metabolites and assessed these variables as the defining elements for the individual RPE. (Borg, 1982) The findings suggest that energy state in our setting is more dependent on oxygen than on fuel (glucose) availability. No elite athletes were represented here; however, more than 40% of marathon runners experience severe and performance-limiting depletion of physiologic reserves. The phenomenon has been attributed to carbohydrate depletion and thousands of runners drop out before reaching the finish lines. (Rapoport, 2010) This interpretation has been questioned and exercise-induced muscle damage has been suggested as being responsible. (Venhorst et al., 2018) Muscle damage can best be studied invasively; however, since exhaustion subjects recover to go on, we reasoned metabolic causes were responsible.

We did not study trained marathon or similar runners. However, we believe our findings have relevance to the personally perceived RPE value. Each and every individual must determine the exhaustion level. Although we picked the extreme RPE groups found in our dataset for mechanistic interpretation, the individuals with median RPE-levels show profiles and PCA-grouping in-between the two more extreme groups (Supplementary Figures S8–S10 and S13). Comparing samples obtained under hypoxic and normoxic conditions alone did not lead to interpretable results. Only when we grouped the samples according to RPE at final exercise time point did we find meaningful insights.

We observed known metabolic changes throughout exercise, such as an initial reliance on glucose as the main fuel source, the subsequent activity of the CORI cycle and CAHILL cycle, as well as fatty-acid mobilization as indicated by increases in glycerol and free fatty acids (Figure 3A, Supplementary Figure S13). This observation makes us confident that our data reflect true metabolite behavior. The potentially novel mechanism responsible for the limit was identified by combining known facts about single metabolites and pathways, such as 2,3-BPG that changes the binding affinity of hemoglobin to blood oxygen. Thus, by not only summarily considering the orchestrated interplay of different tissues reflected in the blood metabolome, but also by considering the fact that new insights might arise from truly novel relationships, we accrued new insights. We believe that these indicators could explain why the MPR "felt badly" during exercise - most likely due to an insufficient ATP supply stemming from oxygen shortage. Succinate, which influences the carbon routing towards TCA-cycle *versus* anaerobic metabolism increased about 2-fold in our MPR, while it rose much higher in highly exhausted subjects. This state-of-affairs might indicate that in our MPR who was accustomed to exhaustion, the oxidative capacity, namely the TCA-intermediates' basal level, was so much higher. The "mitochondrial ratio" according to KREBS shows an initial dip in both the MPR (here much stronger) and the subjects that are running towards exhaustion. However, in the MPR we observed a recovery of this ratio and, probably due to his higher oxidative capacity, the MPR was able to achieve an even higher ratio than at beginning of exercise and towards the end an even higher ratio than those subjects who were less exhausted by the exercise.

We succeeded in distilling the entire dataset of our single MPR into nine key metabolites and their interrelationships. Together, these nine metabolites accounted for four metabolic states and their three transitions. This insight was possible by the quantification of metabolites of different classes and pathways that are not measured in any clinical panel, let alone a single measurement. Erythrocyte-specific metabolites were especially crucial, as for example 2,3-BPG reflects oxygenation status. The diagnostic potential of erythrocytes is almost entirely ignored by the near exclusive investigation of serum and plasma. Although we do not have sufficient time resolution to determine in which order the switches in metabolism occur, we do have the necessary time resolution level to describe these for the first time. The observations made in our cohort under high RPE exhibited decreased pH, lowered pCO_2 and reduced bicarbonate, while the anion gap increased (Supplementary Figure S18). The subjects' lactate, acetoacetate, and glucose values were elevated.

TheLUEBERING-RAPOPORT shunt is a metabolic pathway in mature erythrocytes involving the formation of 2,3-bisphosphoglycerate (2,3-BPG), which regulates oxygen release from hemoglobin and delivery to tissues. 2,3-BPG, the reaction product of the Luebering-Rapoport pathway. Through the Luebering–Rapoport pathway, bisphosphoglycerate mutase catalyzes the transfer of a phosphoryl group from C1 to C2 of 1,3-BPG, giving 2,3-BPG. 2,3-bisphosphoglycerate, the most concentrated organophosphate in the erythrocyte, forms 3-PG by the action of bisphosphoglycerate phosphatase. The concentration of 2,3-BPG varies proportionately with the pH, since it is inhibitory to catalytic action of bisphosphoglyceromutase. We have strong reason to believe that this pathway played a role in our results and should be a topic of intense future investigation.

We used a metabolomics methodology that allows the quantitative determination of a large number of central metabolites and have optimized the method to allow such analysis from a single drop of full blood. Metabolomics samples were taken in combination with the recording of clinical parameters to characterize the impact of exercise on the individuals. Because full blood also includes the hematocrit, the values encompass all cellular components of the blood. As there are some substantial differences in sample handling and quenching of metabolism, we have not compared full blood against serum or plasma. Our approach quenches metabolism immediately as all cells are lysed, enzymes denatured and the extracts cooled immediately to ca. −80°C. We employed a sampling strategy that is potentially available including outside of a clinical laboratory. From our data, we conclude that we are able to detect drastic metabolic changes and that there are additional features that are exclusively measurable in whole blood. While we could not obtain absolute amounts for all metabolites, our findings rely on changes relative to baseline and on ratios of these changes. Samples of every subject under one exercise condition and at two oxygen availability level were kept in sets throughout extraction and measurement, but measurement order was randomized within the sets. Glucose and lactate, two metabolites on which are conclusions are based, were in good agreement between well-established clinical analyzers and GC–MS based measurements (Supplementary Figure S21). We determined intermediates of glycolysis and pentose phosphate pathway that were reflecting the metabolic switch when exhausted. These specific markers were only measurable from full blood during and after crisis in our MPR and were observed only in the highest intensity exercise in our second experiment. Thus they were deemed not to be general markers for RPE at the levels of analytical sensitivity we could achieve. However, the interrelationships from all measured metabolites point towards the influence of oxygen availability. The less trained subjects might have shown similar markers of oxygen release at their maximum.

There are clear limitations in our study. Our initial observations were based on a single MPR, whose fitness level was self-reported rather than measured directly. He gave a subjective RPE report and his values were measured under outdoor "field" conditions. Better would have been to test a homogeneous group of serious athletes in a common protocol to determine whether or not the results of our MPR could be repeated. Circumstances and our desire for generality dictated otherwise. We recruited a very heterogeneous group volunteers whose fitness levels were also not documented. These persons ran in a controlled setting at fixed speeds at two levels of oxygen availability. They were not required to exercise to RPE. We measured our variables in capillary blood. Our MPR was studied in the summer and our volunteer cohort was exercised at room temperatures. Under these conditions, our samples are close to, but not identical to arterial samples. Finally, we are aware that lactate kinetics, clearance, uptake, release, and turnover cannot be completely deduced from whole blood measurements. Thus, we are not able to analyze lactate as a "fulcrum of metabolism". (Brooks, 2020)

The factor most reflected in the metabolome was RPE which in turn seems to be reflected in oxygen availability in the tissue. We suggest that exhaustion concerns an insufficiency of the tricarboxylic acid cycle and oxidative capacity. We did not begin our analysis with the goal to find reflections of RPE in the metabolome, but rather to understand the processes involved. Nevertheless, simple unsupervised data reduction technique (PCA and hierarchical clustering) reflected a connection; namely, we can perceive our metabolic status. While we refrain from postulating general biomarkers for RPE, the key to bringing metabolomics into clinical medicine is to have each person act as an own control. Longer-term observation will allow for preventive medicine and we presented here a relatively simple tool to achieve this end. Non-etheless, we now have a technology available to address these questions.

4 Methods

4.1 Study design, sample collection

4.1.1 Observational study

After due procedures and written informed consent, a preliminary study was performed by a member of our laboratory. Our MPR was 26-year-old, 86 kg, 1.87 m man who views himself as competent athlete and scientist. He arrived in the laboratory at 08:00 after a 12 h overnight fast (but drank water *ad libitum*) to provoke exhaustion state, which individuals often try to ameliorate by "carbohydrate loading". He then ran cross-country at a rate estimated <4 min/km. Each lap consisted of about 2.2 km. The few seconds necessary for sampling were accompanied by a "self-perception" RPE. (Borg, 1982) The MPR described strong exhaustion, similar to "hitting the wall". (Rapoport, 2010) Thereafter, recovery with euphoria termed "runners' high" has been reported. (Kozinc & Sarabon, 2017) At baseline, after each of 6 laps and after 20 min recovery, we obtained 10 µL of capillary full blood from our MPR. The samples were immediately quenched in 1 ml cold MCW (5:2: 1 methanol-|chloroform|water), containing cinnamic acid as internal standard. One round, where a breakdown of performance was felt, the lap was cut short by 200 m in order not to miss this crucial observation-point. Samples were shaken and stored on dry ice. Samples were extracted as lined out below on the same day. In the framework of the subsequent study below, the Charité institutional review board allowed us to continue these investigations further. As a follow-up study, we conceived of a metabolomics study in normal volunteers.

4.1.2 Prospective trial

The ethical committee of the Charité approved the study and written informed consent was obtained. The study was duly registered: ClinicalTrials.gov Identifier: NCT03121885, https://clinicaltrials.gov/ct2/show/NCT03121885 (first posted 20/04/2017). Subjects were recruited by advertisement. Men and non-pregnant women >18 years were recruited who were healthy and ingesting no medications. We purposely did not focus on fitness parameters or abilities. Athletes were not excluded but were purposely not specifically recruited. Some of the subjects were very fit and we cannot exclude the possibility that a few might have even been better than our MPR. Thirteen men and twelve women aged 18–74 years participated in the study. (ACSM et al., 2009) Please refer also to Supplementary Table S4 for an overview over subjects and their individual characterisation.

The subjects arrived in our Clinical Research Center after 12 h fasting (but drank water *ad libitum*) and underwent history and physical examinations. Body composition estimates were performed with BodPod (Life Measurement Inc. Concord, CA, United States), Bioimpedance, and a 3D Body Scanner (Human Solutions GmbH, Kaiserslautern, Germany). Venous blood was obtained for baseline, routine tests (Radiometer ABL800 Flex, Copenhagen, Denmark) and a resting electrocardiogram was performed. Blood pressure was measured oscillometrically and anthropometric data were obtained. The subjects were questioned as to exercise habits and rendered an assessment of their fitness levels.

We determined the performance levels, (Jones et al., 1985), as adapted to treadmill exercise according to normal standards as estimated from ergometer testing indoor. We aimed for an estimated eight metabolic equivalent of task (MET) performance for 30 min. If this task was insufficient to exhaust the subjects, the test was repeated with a 30% increment and in some very fit individuals a 60% increment was performed. To determine RPE, we relied on a 1-through-10 modified BORG scale. (Borg, 1982) Subjects were randomized to order of exercise at sea level ($Fi\ O_2$ 20.9%) or to normobaric hypoxia (altitude 3,000 m, $Fi\ O_2$ 14.5%). They were unaware of the regimens provided, as our chamber was used for all studies. Respective to performance-ability, baseline, 10 min, 20 min, and 30 min samples (or when RPE was experienced) as well as 10 min (recovery 1) and 20 min (recovery 2) after exercise of 20 µL capillary full blood was taken from the ear-lobe and immediately quenched in 1 ml cold MCW (5:2:1 methanol|chloroform|water) containing cinnamic acid as internal standard, shaken, and stored on dry ice Samples were stored at −80°C until further extraction.

At blood drawing, subjects were asked to estimate their performance stress on a scale of 1–10, similar to the one established by BORG. (Borg, 1982) Samples were collected in ice cold methanol | chloroform | water (5:2:1) containing cinnamic acid, immediately quenching metabolic activity. Capillary blood 10 µL from the earlobe was obtained at baseline, at 30 min or at exhaustion, and 30 min after exercise. In these samples, we measured blood gases for pH, pO_2, pCO_2, HCO_3^-, Na^+, K^+, Cl^-, Ca^{2+}, and anion gap (Radiometer). Glucose and lactate were measured separately with routine chemical analysis. Blood pressure, heart rate, and pulse oximetry were determined at baseline, 10 min, 20 min, 30 min or at exhaustion, and 20 min and 30 min of recovery.

4.2 Metabolomics

4.2.1 Sample extraction

Samples were removed from the freezer in batches and kept at 4°C throughout the extraction process. 500 µL water (in the prospective trial also containing isotopically labeled internal standards for normalization) were added to induce phase separation. Samples were shaken (Eppendorf 1,000 rpm) for 20 min to ensure phase equilibration. After 10 min centrifugation, polar (upper) and lipid phase (lower) were obtained. Lipid extracts were dried under nitrogen stream and stored at −80°C until measurement. Polar phase extracts were dried in a rotational vacuum concentrator (Martin Christ, Germany) without heating in <4 h. Samples were stored at −80°C (observational study, −20°C) until derivatization.

4.2.2 Standardization

For substance identification across batches, we used mixtures of 100 substances to compare RI and mass spectra. For 69 substances we measured 8-point calibration curves to check linearity (Pietzke et al., 2014). Sample intensities within one set of experiments were standardized by cinnamic acid added to the extraction solvent (MCW). Furthermore, we added 2 stably labeled isotopomers during extraction (see also Quantification/Normalization). As such, we standardized by cinnamic acid and used fully labeled lactate and glucose to assess our quantification against the established clinical methods. (Supplementary Figure S21)

4.2.3 Derivatization

For derivatization, extracts were thawed in a rotational vacuum concentrator (Martin Christ, Germany) without heating for 20 min 10 µL of 40 mg methoxyamine hydrochloride/mL pyridine were added, samples were incubated for 90 min at 30°C. Next 30 µL of MSTFA containing 200 µg/ml *n*-alkanes (C_{10}, C_{12}, C_{15}, C_{17}, C_{19}, C_{22}, C_{28}, C_{32}, C_{36}) as retention index markers were added as previously described (Pietzke et al., 2014). Derivatization was carried out simultaneously for every sample in a single measurement batch.

4.2.4 Randomization

Samples were randomized as follows: To ensure highest level of comparability among one subject's samples blocks from one subject's performances at a certain exercise intensity at different $Fi\ O_2$ were formed giving blocks of ≤12 samples (2 $Fi\ O_2$ levels and maximum 6 time-points, depending on ability). These blocks were kept throughout extraction, derivatization and measurement. Extraction batches (*n* = 10) consisted of ≤4 blocks, measurement batches (*n* = 10) also contained ≤4 blocks but consisted of different sets of randomly selected blocks. Measurement order within one block was randomized.

4.2.5 Gas chromatography-mass spectrometry measurement

Gas chromatography-mass spectrometry was carried out using a previously published method using a Pegasus IV GC-ToF MS (Leco, United States) (Pietzke et al., 2014). Scan rates of 20 Hz and a mass range of 70–600 Th were used. Ionization energy was set to 70 eV. Gas chromatographic separation of compounds was performed on an Agilent 6890N (Agilent, Santa Clara, CA, United States) equipped with a VF-5ms column of 30 m length (Varian, Palo Alto, CA, United States). The initial temperature was held at 67.5°C for 2°min, followed by a temperature gradient of 5°C min^{-1} until 120°C, then 7°C min^{-1} until 200°C, followed by 12°C min^{-1} until 320°C with a hold time of 6 min. The transfer line was kept at 250°C throughout. A cold injection system was used with a matching baffled deactivated liner (CIS4, Gerstel, Mülheim an der Ruhr, Germany), operating in split mode (split 1:5, injection volume 1 µL), with the following temperature gradient applied: hold of the initial temperature of 80°C for 0.25°min, followed by a temperature increase of 12°C s^{-1} to 120°C, followed by a temperature increase of 7°C s^{-1} to 300°C with a hold time of 2 min.

4.2.6 Peak picking, annotation

Data was smoothed and baseline corrected using ChromaTOF (vendor software). Peaks were picked using ChromaTOF with a signal to noise threshold of 20. Given the rather small sample size, formal tests for Gaussian distribution and linearity would be underpowered and not informative. Accordingly, we decided to use Spearman rank correlations to produce the correlation matrix underlying the PCA.

Annotation was performed using an in-house version of a published software (Kuich et al., 2014), as well as using manual inspection with proprietary software (ChromaTOF, LECO). This approach allowed us to inspect the mass spectra of each peak from all measurements individually. We matched peaks stepwise against i) standard mixes included at the beginning of every batch (library size = 137) (Opialla et al., 2020), as well as ii) an in-house library (library size = 12) of compounds individually measured on our machines, and iii) a subset of the Golm-metabolome database (Kopka et al., 2005). For the observational study, we also annotated and reported unidentified but consistently occurring peaks. Lipid compounds were matched against an in-house library (library size = 36).

4.2.7 Quantification, normalization

Metabolites were quantified using the top 5 mass traces according to intensity, excluding masses if adjacent peaks had same nominal mass and mass traces originating from derivatization agents (e.g. 73 Th, 147 Th). Also, characteristic masses were included purposefully (e.g. 299 Th for phosphates). The scans along the peaks were summed up to give AUC without interpolation. Glucose and lactate were also measured with clinically approved methods, so we used the included u-^{13}C-labeled substances added during extraction, to compare clinical measurements, our top-5 approach and the current gold standard: heavy labeled internal standards. For glucose we used the ion pairs 319/323 Th and 217/220 Th, for lactate 117/119 Th and 190/193 Th. Samples were normalized using cinnamic acid included in the extraction solvent at sample collection.

4.2.8 Missing values

As with any MS-dataset, several metabolites have missing values. Except for clear oxygenation markers accordingly with *Fi* O $_2$ and effort level (ribose-5-phosphate and ribulose-5-phosphate, Supplementary Figures S22A, 23) and iso-aminobutyrate in females (Supplementary Figure S22B), no compound was significantly missing more in one condition. After careful manual curation we found, that with generally lower intensity also more missing values occur (NMAR). We therefore treated the missing values as not missing at random values (NMAR). Values were imputed using QRILC (Lazaar, 2015) on a per metabolite basis on the normalized values, we allowed generally up to 20% missing values. If fraction of missing values was higher, metabolites were excluded from multivariate statistics in the prospective trial.

4.2.9 Statistical analyses, time-profiles

Statistical analysis was carried out using R and tidyverse (https://www.tidyverse.org/packages); visualizations except where noted, were created using ggplot2 and inkscape (https://inkscape.org/release/inkscape-0.92.4).

Since some of the subjects in the prospective trial were not able to complete the exercise bout (resulting in very low RPE values) we re-encoded data collected at last exercise time-point as 30 min exercise value. As not every individual was able to complete the exercise, we sometimes obtained less than three samples from exercise. For plotting time-profiles we shifted the samples in time in a way that all values obtained from final exercise time-point have the same time coordinate, as this reflects the most similar state possible, when dealing with such a heterogeneous group of performance levels as in our study. Normally distributed clinical data were statistically analyzed by repeated-measures analysis of variance with appropriate adjustments. For PCA we removed outliers according to HOTELLING's-*T* criteria. For line-plots all samples were included. From a statistics point of view, principal components are unobservable higher-order traits covering a wider range of observable measures. Naming those components is inevitably arbitrary, and we deduced the main shared feature from the underlying highly correlated traits.

The authors confirm that all methods were carried out in accordance with relevant guidelines and regulations.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Committee Charité Berlin Board Affiliation: Ethics Committee 1 at the Campus Charité - Mitte Phone: 030/450-517222 Email: ethikkommission@charite.de—Charitéplatz 1, 10117 Berlin. The patients/participants provided their written informed consent to participate in this study.

Author contributions

TO, BG, PK, FL, and SK conceptualized the study. TO, BG, PK, LK, GR, AB, and MB took care of data curation. TO, PK, and AB performed formal analysis. SS, JK, FL, and SK acquired funding. TO, BG, PK, LK, GR, MB, FL, and SK conducted experiments. TO, BG, PK, FL, and SK developed methodology. TO, BG, PK, LK, GR, MB, JK, FL, and SK performed project administration. TO, BG, PK, LK, GR, AB, SS, MB, JK, FL, and SK provided resources. TO, PK, and AB wrote software and scripts. TO, BG, PK, SS, JK, FL, and SK supervised experiments and analysis. TO, BG, PK, AB, FL, and SK validated the results. TO, PK, and AB visualized the data. TO, PK, FL, and SK wrote the original draft. TO, BG, PK, SS, JK, FL, and SK reviewed and edited the manuscript.

Funding

The study was supported by the DfG (LU 435/13-1) and the Helmholtz Fundation (ZT-0026) by the AMPRO consortium. The Berlin Institute for Medical Systems Biology, the Berlin Institute of Health provided general support. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We thank Jenny Grobe for excellent technical assistance and all the runners for giving their utmost best.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.1042231/full#supplementary-material

References

American College of Sports MedicineChodzko-Zajko, W. J., Proctor, D. N., Fiatarone Singh, M. A., Minson, C. T., Nigg, C. R., et al. (2009). American College of Sports Medicine position stand. Exercise and physical activity for older adults. *Med. Sci. Sports Exerc* 41, 1510–1530. doi:10.1249/MSS.0b013e3181a0c95c

Bassini, A. (2014). Sportomics: Building a new concept in metabolic studies and exercise science. *Biochem. Biophys. Res. Commun.* 445, 708–716. doi:10.1016/j.bbrc.2013.12.137

Benesch, R., and Benesch, R. E. (1967). The effect of organic phosphates from the human erythrocyte on the allosteric properties of hemoglobin. *Biochem. Biophys. Res. Commun.* 26, 162–167. doi:10.1016/0006-291x(67)90228-8

Borg, G. A. (1982). Psychophysical bases of perceived exertion. *Med. Sci. Sports Exerc* 14, 377–381. Available At: OR. doi:10.1249/00005768-198205000-00012https://europepmc.org/article/MED/7154893 https://journals.lww.com/acsm-msse/Abstract/1982/05000/Psychophysical_bases_of_perceived_exertion_.12.aspx.

Brooks, G. A. (2020). Lactate as a fulcrum of metabolism. *Redox Biol.* 35, 101454. doi:10.1016/j.redox.2020.101454

Castillo-Garzón, M. J., Ruiz Jonatan, R., Ortega, F. B., and Gutierrez, A. (2006). Anti-aging therapy through fitness enhancement. *Clin. Interv. Aging* 1, 213–220. Available At: OR. doi:10.2147/ciia.2006.1.3.213https://europepmc.org/article/MED/18046873 https://www.dovepress.com/anti-aging-therapy-through-fitness-enhancement-peer-reviewed-article-CIA.

Chiba, H., and Sasaki, R. (1978). Functions of 2, 3-bisphosphoglycerate and its metabolism. *Curr. Top. Cell Regul.* 14, 75–116. doi:10.1016/b978-0-12-152814-0.50007-1

Contrepois, K., Wu, S., Moneghetti, K. J., Hornburg, D., Ahadi, S., Tsai, M. S., et al. (2020). Molecular choreography of acute exercise. *Cell* 181, 1112–1130. doi:10.1016/j.cell.2020.04.043

Gall, W. E., Beebe, K., Lawton, K. A., Adam, K. P., Mitchell, M. W., Nakhle, P. J., et al. (2010). RISC Study Group *et al.* Alpha-Hydroxybutyrate Is an Early Biomarker of Insulin Resistance and Glucose Intolerance in a Nondiabetic Population. *PLoS One* 5, e10883. doi:10.1371/journal.pone.0010883

Gong, Z. G., Hu, J., Wu, X., and Xu, Y. J. (2017). The recent developments in sample preparation for mass spectrometry-based metabolomics. *Crit. Rev. Anal. Chem.* 47, 325–331. doi:10.1080/10408347.2017.1289836

Haug, K., Cochrane, K., Nainala, V. C., Williams, M., Chang, J., Jayaseelan, K. V., et al. (2019). MetaboLights: A resource evolving in response to the needs of its scientific community. *Nucleic Acids Res.* D1, D440–D444. doi:10.1093/nar/gkz1019

Herrmann, D., Pohlabeln, H., Gianfagna, F., Konstabel, K., Lissner, L., Mårild, S., et al.Consortium IDEFICS (2015). Association between bone stiffness and nutritional biomarkers combined with weight-bearing exercise, physical activity, and sedentary time in preadolescent children. A case-control study. *Bone* 78, 142–149. doi:10.1016/j.bone.2015.04.043

Hobbins, L., Hunter, S., Gaoua, N., and Girard, O. (2017). Normobaric hypoxic conditioning to maximize weight loss and ameliorate cardio-metabolic Health in obese populations: A systematic review. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 313, R251–R264. doi:10.1152/ajpregu.00160.2017

Jones, N. L., Makrides, L., Hitchcock, C., Chypchar, T., and McCartney, N. (1985). Normal standards for an incremental progressive cycle ergometer test. *Am. Rev. Respir. Dis.* 131, 700–708. Available At: OR. doi:10.1164/arrd.1985.131.5.700https://www.atsjournals.org/doi/abs/10.1164/arrd.1985.131.5.700 https://europepmc.org/article/MED/3923878.

Klein, D. J., Anthony, T. G., and McKeever, K. H. (2021). Metabolomics in equine sport and exercise. *J. Anim. Physiol. Anim. Nutr.* 105, 140–148. doi:10.1111/jpn.13384

Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmüller, E., et al. (2005). GMD@CSB.DB: The Golm metabolome database. *Bioinformatics* 21, 1635–1638. doi:10.1093/bioinformatics/bti236

Kozinc, Ž., and Šarabon, N. (2017). Effectiveness of movement therapy interventions and training modifications for preventing running injuries: A meta-analysis of randomized controlled trials. *J. Sports Sci. Med.* 16, 421–428. Available At: OR https://europepmc.org/article/MED/28912661https://www.jssm.org/hf.php?id=jssm-16-421.xml.

Krebs, H. A. (1967). The redox state of nicotinamide adenine dinucleotide in the cytoplasm and mitochondria of rat liver. *Adv. Enzyme Regul.* 5, 409–434. doi:10.1016/0065-2571(67)90029-5

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: An information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi:10.1101/gr.092759.109

Kuich, P. H. J. L., Hoffmann, N., and KempaMaui-Via, S. (2014). Maui-VIA: A user-friendly software for visual identification, alignment, correction, and quantification of gas chromatography-mass spectrometry data. *Front. Bioeng. Biotechnol.* 2, 84. doi:10.3389/fbioe.2014.00084

Lazar, C. (2015). *imputeLCMD: A collection of methods for left-censored missing data imputation*. Available At: https://CRAN.R-project.org/package=imputeLCMD.

Lear, S. A., Hu, W., Rangarajan, S., Gasevic, D., Leong, D., Iqbal, R., et al. (2017). The effect of physical activity on mortality and cardiovascular disease in 130000 people from 17 high-income, middle-income, and low-income countries: The PURE study. *Lancet* 390, 2643–2654. doi:10.1016/S0140-6736(17)31634-3

Loprinzi, P. D. (2015). Dose-response association of moderate-to-vigorous physical activity with cardiovascular biomarkers and all-cause mortality: Considerations by individual sports, exercise and recreational physical activities. *Prev. Med.* 81, 73–77. doi:10.1016/j.ypmed.2015.08.014

Nindl, B. C., Jaffin, D. P., Dretsch, M. N., Cheuvront, S. N., Wesensten, N. J., Kent, M. L., et al. (2015). Human performance optimization metrics: Consensus findings, gaps, and recommendations for future research. *J. Strength Cond. Res.* 29 (11), S221–S245.Suppl. doi:10.1519/JSC.0000000000001114

Opialla, T., Kempa, S., and Pietzke, M. (2020). Towards a more reliable identification of isomeric metabolites using pattern guided retention validation. *Metabolites* 10, 457. doi:10.3390/metabo10110457

Palacios, G., Pedrero-Chamizo, R., Palacios, N., Maroto-ánchez, S. B., Aznar, S., and González-Gross, M.E.X.E.R.N.E.T. (2015). Biomarkers of physical activity and exercise. *Nutr. Hosp.* 31 (3), 237–244. Suppl. doi:10.3305/nh.2015.31.sup3.8771

Pietzke, M., Zasada, C., Mudrich, S., and Kempa, S. (2014). Decoding the dynamics of cellular metabolism and the action of 3-bromopyruvate and 2-deoxyglucose using pulsed stable isotope-resolved metabolomics. *Cancer Metab.* 2, 9. doi:10.1186/2049-3002-2-9

Rapoport, B. I. (2010). Metabolic factors limiting performance in marathon runners. *PLoS Comput. Biol.* 6, e1000960. doi:10.1371/journal.pcbi.1000960

van Loon, L. J., Greenhaff, P. L., Constantin-Teodosiu, D., Saris, W. H., and Wagenmakers, A. J. (2001). The effects of increasing exercise intensity on muscle fuel utilisation in humans. *J. Physiol.* 536, 295–304. doi:10.1111/j.1469-7793.2001.00295.x

Venhorst, A., Micklewright, D., and Noakes, T. D. (2018). Modelling the process of falling behind and its psychophysiological consequences. *Br. J. Sports Med.* 52, 1523–1528. doi:10.1136/bjsports-2017-097632

Warburton, D. E. R., Nicol, C. W., and Bredin, S. S. D. (2006). Health benefits of physical activity: The evidence. *Can. Med. Assoc. J.* 174, 801–809. doi:10.1503/cmaj.051351

Williamson, D. H., Lund, P., and Krebs, H. A. (1967). The redox state of free nicotinamide-adenine dinucleotide in the cytoplasm and mitochondria of rat liver. *Biochem. J.* 103, 514–527. doi:10.1042/bj1030514

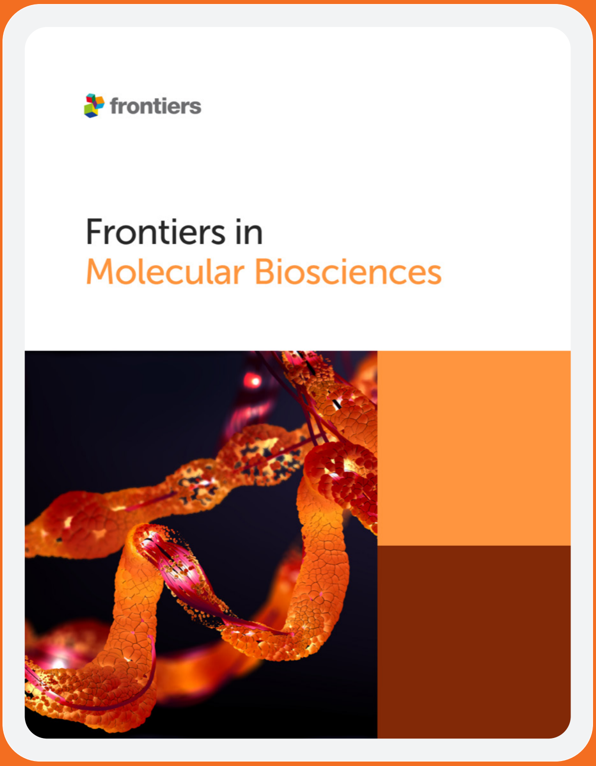
Frontiers in
Molecular Biosciences

Explores biological processes in living organisms on a molecular scale

Focuses on the molecular mechanisms underpinning and regulating biological processes in organisms across all branches of life.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Molecular Biosciences

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact