frontiers | Research Topics

Insights in computational genomics 2022

Edited by
Richard D. Emes, Mehdi Pirooznia, Quan Zou
and Marco Pellegrini

Published in
Frontiers in Genetics

FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-3173-0
DOI 10.3389/978-2-8325-3173-0

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 computational genomics: 2022

Topic editors

Richard D. Emes — Nottingham Trent University, United Kingdom
Mehdi Pirooznia — Johnson & Johnson, Washington, D.C., United States
Quan Zou — University of Electronic Science and Technology of China, China
Marco Pellegrini — Institute of Informatics and Telematics, Department of Engineering, ICT and Technology for Energy and Transport, National Research Council (CNR), Italy

Citation

Emes, R. D., Pirooznia, M., Zou, Q., Pellegrini, M., eds. (2023). *Insights in computational genomics: 2022*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-3173-0

Table of contents

TYPE Editorial
PUBLISHED 24 July 2023
DOI 10.3389/fgene.2023.1256011

OPEN ACCESS

EDITED AND REVIEWED BY
Mensur Dlakic,
Montana State University, United States

*CORRESPONDENCE
Marco Pellegrini,
✉ marco.pellegrini@iit.cnr.it

RECEIVED 10 July 2023
ACCEPTED 17 July 2023
PUBLISHED 24 July 2023

CITATION
Emes RD, Pirooznia M, Zou Q and Pellegrini M (2023), Editorial: Insights in computational genomics: 2022. *Front. Genet.* 14:1256011.
doi: 10.3389/fgene.2023.1256011

COPYRIGHT
© 2023 Emes, Pirooznia, Zou and Pellegrini. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Editorial: Insights in computational genomics: 2022

Richard D. Emes[1], Mehdi Pirooznia[2,3], Quan Zou[4] and Marco Pellegrini[5]*

[1]Nottingham Trent University, Nottingham, United Kingdom, [2]School of Medicine, Johns Hopkins University, Baltimore, MD, United States, [3]Pharmaceutical Data Sciences, R&D Johnson & Johnson, Boston, MA, United States, [4]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China, [5]Consiglio Nazionale delle Ricerche, Pisa, Italy

KEYWORDS

computational genomics, cancer studies, single cell RNA-seq technology, human genomic variation, deep learning

Editorial on the Research Topic
Insights in computational genomics: 2022

The goal of this Research Topic is to shed light on the progress made in the past decade in the Computational Genomics field, to gauge its future challenges, and to provide a thorough overview of the field's current status. We hope that this article Research Topic will inform, inspire and provide guidance to researchers in the field. Foundational Research Topic ranging from still unsolved evolutionary mechanisms at the genomic level and the challenges posed by human genomic variation are powerful drivers of current and future research in Computational Genomics. The challenge that genomics poses to computational theory is to be highlighted, ranging from the role of Artificial Intelligence and Deep Learning (DL) in this context to the likely impact of emerging Single Cell RNA Sequencing (scRNA-Seq) methodologies in transcriptomics. Steady progress in tools for automated managing of large heterogeneous genomic and biological data is likely to bring good dividends in the near future. Specific applications of computational genomics support cancer studies for tasks such as drug repositioning and finding the role of immune system genes in cancer. Also, computational genomics is a key helper to plant science in the effort to cope with the effects of climate changes in the long run, with global food security as a goal.

Here is an overview of the issues presented in this Research Topic.

The evolution of genomes and codon encodings is a source of key fundamental questions still needing an answer. In this area, Belinky et al. highlight major differences between prokaryotes and eukaryotes regarding the double substitutions of nucleotides in codon encodings.

Dong et al. provide experimental evidence on the performance of recently developed DL methods compared to more traditional flavors of Machine Learning (ML) when applied to the prediction of risk in cancer. Using a very large cohort of patients and three cancer test types, they give interesting hints for further research on this Research Topic.

The Human Genome Project (HGP) lasted from 1990 to 2003 and has brought about a scientific revolution in genomics of the type the philosopher Thomas S. Kuhn has described. As with every scientific revolution, its long-lasting value is that of posing new questions. Singh et al. argue that the Human Pangenome Project is the next logical step on the road opened by the HGP, allowing us to reach new heights in genomic research in the near future.

Single Cell RNA Sequencing is one of the key new technologies in transcriptomics, enabling a finer view of the diverse roles of single cells (or cell types) within a sample. New technologies often need new algorithmic insights and robust statistical filtering methodologies. Li et al. propose in their article a novel method for detecting a limited number of representative cells within a pool of thousands (up to hundreds of thousands) of cells typically analyzed in a single scRNA-Seq run. Carangelo et al. give an overview of the scRNA-Seq technologies with particular emphasis on the scalability of the algorithmic analytic tools needed to cope with the ever-increasing rate of data generation. Interestingly the field of scRNA-Seq research is poised to move from a niche sector to a broader role in a clinical setting for human health.

Modern metabolomics assays produce vast volumes of complex data. Thus, there is a growing opportunity for the application of Machine Learning to analyze such data, recognize new patterns, and build models across multiple levels (from the genomic to the metabolic aspects). Galal et al. give an overview of this burgeoning area of research with emphasis on its potential for leading to new disease classifications and to uncovering key aspects of the disease onset and progression.

Gene duplication and gene transfer are important evolutionary mechanisms still needing further research and attention. Here Zhang et al. give an overview of the current trends and resources in the fascinating subject of intra-species detection of gene duplications, which is often a key strategy in solving the intra-species functional gene annotation problem. Nayar et al. study the phenomenon of horizontal gene transfer mediated by conjugation, which is considered an important evolutionary mechanism of bacteria. With the new proposed methodology (ggMOB) they found that over half of the bacterial genomes contained one or more known conjugation features that matched exactly to at least one other genome. Science and technology often require a constant and valuable background activity of standardization and unification of procedures and data. Maia et al. describe a computational workflow (AnnotaPipeline) for the annotation of eukaryotic proteins using multi-omics data aiming at overcoming problems due to the variety of sequencing platforms that generate increasing amounts of data, thus making manual annotation no longer feasible.

Cancer studies have been and still are at the forefront of the application of genomics to human health. In this context, we report two studies: Ai et al. identify genes involved in the immune response to colorectal cancer for the purpose of cancer prognosis and potential impact on future immune therapies, while Bennett et al. compare the gene expression profiles of disease-states with the perturbation on gene-expression profiles by a given drug and are thus able to identify 24 existing drugs with potential beneficial effects for patients of Esophageal Cancer (EC).

As climate and climate-related phenomena are gaining continuous attention in public discourse and long-term planning, promoting plant genomics studies is strategic for the future of food security. Here we report two exemplary applicative studies in plant genomics. Jiang et al. analyze the transcriptional dynamics of filling stage Tartary buckwheat seeds in order to provide a theoretical basis for improving the yield of Tartary buckwheat. Pan et al. investigate wheat genotypes and differential gene expression in several climate-related conditions to highlight and clarify the cold-resistance mechanisms leading to potentially higher yields in uncertain climatic conditions.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This work has been partially supported with co-funding of the European Union—Next-Generation EU, in the context of The National Recovery and Resilience Plan, Investment 1.5 Ecosystems of Innovation, Project Tuscany Health Ecosystem (THE), CUP: B83C22003920001.

Acknowledgments

We thank the reviewers and editorial team for support in developing this Research Topic.

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.

METHODS
published: 18 July 2022
doi: 10.3389/fgene.2022.954024

Selecting Representative Samples From Complex Biological Datasets Using K-Medoids Clustering

Lei Li[1,2]*, Linda Yu-Ling Lan*[1,2]*, Lei Huang*[3]*, Congting Ye*[4]*, Jorge Andrade*[3,5] *and Patrick C. Wilson*[1,2]*

[1]University of Chicago Department of Medicine, Section of Rheumatology, University of Chicago, Chicago, IL, United States, [2]Knapp Center for Lupus and Immunology Research, University of Chicago, Chicago, IL, United States, [3]Center for Research Informatics, University of Chicago, Chicago, IL, United States, [4]Key Laboratory of the Ministry of Education for Coastal and Wetland Ecosystems, College of the Environment and Ecology, Xiamen University, Xiamen, China, [5]Department of Pediatrics, University of Chicago, Chicago, IL, United States

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Weksi Budiaji,
Sultan Ageng Tirtayasa University, Indonesia
Sammed Mandape,
University of North Texas Health Science Center, United States

***Correspondence:**
Patrick C. Wilson
pcw4001@med.cornell.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 26 May 2022
Accepted: 20 June 2022
Published: 18 July 2022

Citation:
Li L, Lan LY-L, Huang L, Ye C, Andrade J and Wilson PC (2022) Selecting Representative Samples From Complex Biological Datasets Using K-Medoids Clustering. Front. Genet. 13:954024. doi: 10.3389/fgene.2022.954024

Rapid growth of single-cell sequencing techniques enables researchers to investigate almost millions of cells with diverse properties in a single experiment. Meanwhile, it also presents great challenges for selecting representative samples from massive single-cell populations for further experimental characterization, which requires a robust and compact sampling with balancing diverse properties of different priority levels. The conventional sampling methods fail to generate representative and generalizable subsets from a massive single-cell population or more complicated ensembles. Here, we present a toolkit called Cookie which can efficiently select out the most representative samples from a massive single-cell population with diverse properties. This method quantifies the relationships/similarities among samples using their Manhattan distances by vectorizing all given properties and then determines an appropriate sample size by evaluating the coverage of key properties from multiple candidate sizes, following by a k-medoids clustering to group samples into several clusters and selects centers from each cluster as the most representatives. Comparison of Cookie with conventional sampling methods using a single-cell atlas dataset, epidemiology surveillance data, and a simulated dataset shows the high efficacy, efficiency, and flexibly of Cookie. The Cookie toolkit is implemented in R and is freely available at https://wilsonimmunologylab.github.io/Cookie/.

Keywords: single cell, sampling, k-medoids, R, antibody candidate selection

INTRODUCTION

Single-cell sequencing techniques grew extensively by developing higher cell throughput, improved sensitivity, better reliability, and more modalities in the last decade (Tang et al., 2009; Peterson et al., 2017; Svensson et al., 2018; Stuart and Satija, 2019). Among all biological topics and contexts, the immune system contains a massive amount of highly diverse cells in phenotype and function, and therefore has benefited enormously from the application of novel single-cell RNA sequencing (scRNA-seq) in order to investigate the development and activation of immune cells (Bendall et al., 2014; Goldstein et al., 2019; Winkels et al., 2018; Zhang et al., 2019). In detail, people are able to characterize diverse properties, for example, transcriptome expression, B cell repertoire (BCR), and surface protein expression, for a massive amount of single immune cells in a single experiment (Peterson et al., 2017; Goldstein et al., 2019; Li et al., 2021). This gives people immense power to comprehensively scan a whole population of immune cells in order to identify candidates for further experimental characterization (e.g., neutralizing and antibody binding) (Dugan et al., 2021). Since experimental characterizations are usually resource and human labor intensive, the number of candidates is usually limited by budget. Therefore, a sampling strategy that is capable to effectively select compact and representative samples from a massive population with diverse properties is highly demanded.

The selection of representative samples to reflect the properties and maxima proportion of a large population is a common problem (McCarty et al., 1997; Tominaga, 1998; Siddiqui et al., 2006; Chen et al., 2016). Compared to conventional sampling problems, it imposes even more challenges when selecting samples from a massive biological dataset, for example, single-cell atlas dataset, as biological sample selections are often size sensitive and have diverse properties with different types and importance, and all properties need to be balanced in the selection. More specifically, novel biological data, represented by single-cell atlas data, proposed three specific requirements to representative sampling. First, the selected samples should be able to maximally represent the distribution of original population. Second, the sample size should be as compact as possible in order to save human labor and reagents. Third, randomness of selected samples is not preferred in those cases because subsequent experimental design requires robust and repeatable results. In general, a sampling strategy that can achieve the balance between scientific sufficiency and expense economy with high efficiency is preferred, which can effectively address the contradiction between growing detection capabilities and limited experimental capabilities.

Sampling from a large population has been well studied, and multiple probability and nonprobability sampling methods, including simple random sampling, systematic sampling, cluster sampling, stratified sampling, quota sampling, and snowball sampling, have been proposed for practical sampling problems (Cochran, 2007; Fricker, 2008). Two implementations of probability sampling methods, R package "sampling" and "survey," have been developed and widely used in the community (Tillé and Matei, 2006; Tillé et al., 2016; Lumley and Lumley, 2019). Those conventional methods do not or rarely use data structure in sampling; therefore, they fail to maximally balance the given properties. Some minor groups maybe ignored causing samples in those groups being rejected. Furthermore, the randomness in the results of probability sampling methods is not preferred or even strictly prohibited in candidates sampling of single-cell atlas data and some other contexts (e.g., influenza surveillance) because robust and repeatable results of each step are crucial for these studies. In addition, a group of Markov chain Monte Carlo (MCMC)–based sampling methods, for example, Metropolis–Hastings sampling and Gibbs Sampling, were proposed to solve sampling problem from high-dimensional population (Geman and GemanHastings, 1970;, 1984). These MCMC-based methods select samples by using data distribution on multiple properties of whole population, and therefore can generate much more representative results than conventional sampling methods. However, MCMC-based sampling methods are usually used to estimate parameters of unknown distribution by constructing a big stochastic process from a given population, or to generate representative samples from a known probability distribution. For single-cell datasets, the joint probability distributions of multiple properties are usually unknown and incalculable, which makes MCMC-based sampling unavailable. Moreover, after algorithms reach a convergence, MCMC-based methods prefer to select more samples for better estimation, which contradicts the requirement of compatibility of single-cell data selection. In practice, compatibility, stability, and representativeness on massive population are three priorities that may not be easily achieved by existing sampling methods. Meanwhile, a systematic approach to determine an appropriate sample size is required.

To overcome these challenges, we developed a k-medoids clustering-based sampling strategy. This method achieves both stable and representative results and allows users to determine an optimized sample size by evaluating the coverage of key properties. We have made Cookie available on a public repository for users worldwide: https://wilsonimmunologylab.github.io/Cookie/.

MATERIALS AND METHODS

Datasets

Simulated dataset: We generated a simulated dataset with 10,000 samples and five factors. We generated three-character type factors (Factors 1–3) and two numerical type factors (Factors 4 and 5). Factor 1 is a character factor with levels from 1—20; Factor 2 is a character factor with levels from 1—50; Factor 3 is a character factor with levels from group 1—group 9; Factor 4 is a numerical factor with integer values within the range of 1–20; and Factor 5 is a numerical factor with floating number values that follow a normal distribution (mean = 0, standard deviation = 1). There are a total of 10,000 records in this dataset, and we also extract different size subsets (1,000, 2,500, and 5,000) from this dataset to test the efficiency of our method on different data sizes.

Single-cell B cell dataset: In the vaccine clinical trial, we applied Cookie to unbiasedly select representative monoclonal antibodies for expression/characterization from 1,937 antibodies from 19 subjects, seven transcriptional clusters, four isotypes (IgA, IgG, IgM, and IgD), various V locus gene usages, and various CDR3 peptide lengths. We generated these monoclonal antibodies using single-cell B cell receptor cloning of a pair of the heavy chain and light chain genes followed by *in vitro* expression to further characterize mAb specificity and function to evaluate the vaccine response.

Human influenza H1N1 surveillance viral sample dataset: We downloaded all data records of human influenza H1N1 viruses collected between August 1, 2018 and August 1, 2019 from GISAID database (https://www.gisaid.org/) (Shu and McCauley, 2017). A total of 8,449 viruses were retained after removing the redundant records. By comparing the sequences to the WHO recommended H1N1 vaccine strain A/Michigan/45/2015 (H1N1) (https://www.cdc.gov/flu/season/flu-season-2018-

2019.htm), we calculated mutation numbers of the HA1 protein for all H1N1 viruses. Mutation numbers of H1 epitopes was also calculated for all H1N1 viruses. The protein sequences were aligned using MAFFT v7.427 (Katoh and Standley, 2013). Positions of five epitopes of H1 protein were adopted from the literature (Li et al., 2020). In this dataset, there are four factors: month, continent, mutations, and mutations on epitopes. The month factor is a character factor with 12 levels (2018-08 to 2019-07); the continent factor is a character factor with six levels (Africa, Asia, Europe, North America, Oceania, and South America); mutations is a numerical factor with integer values within the range of 0–14; and mutation on epitope is a numerical factor with integer values within the range of 0–3. The dataset was downloaded on August 29, 2019.

Data Vectorization

Each sample was represented by a vector, and dimensions of the vector are factors from the original data (e.g., subject, transcriptional cluster or just "cluster," gender, and antibody isotype). All factors can be divided into two groups: character factors and numerical factors. Character factors usually have multiple (two or more) discrete values, representing clusters, subjects, groups, batches, and so on. Numerical factors have continuous numerical values with either integers or floating numbers, and different levels can be quantified by the difference of these values. To clarify, character factors also have numerical levels. The difference between numerical factors and character factors is that the levels of character factors are none-quantifiable labels and the levels of numerical factors are quantifiable values. The difference between any two levels can be quantified by the difference of their values. For example, for a character factor (e.g., a cluster) which has three levels, the difference between levels 1 and 2 is equal to that between levels 1 and 3. For a numerical factor (e.g., number of mutations) which has three levels, the difference between levels 1 and 2 is smaller than that between levels 1 and 3.

Linearization

In biological datasets, logarithmic values are a commonly used data type (e.g., HI titers in Influenza hemagglutination inhibition assays, https://www.cdc.gov/flu/about/professionals/antigenic.htm). In order to compare values within a factor, values of all the numerical factors should be linear (Sun et al., 2013). All of the nonlinear factors should be linearized in advance of further normalization. A logarithm will transfer logarithmic values into linear values. Users should choose the base number of the logarithm according to their dataset. For example, original HI titers are equal to $HI = 10 \times 2^n$, where n is the number of dilutions, so the base of logarithm will be two for HI data; thus, the linearized HI titer should be: $HI' = log_2(HI/10)$.

Normalization

Data normalization is essential for numerical factors in order to be comparable with other factors (Hancock et al., 1988; Singh and Singh, 2020). Here, we adopted a min–max normalization method to scale a numerical factor such that all values are within the range of [0,1]. The normalized value x' can be calculated by the following equation:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}},$$

where x, x_{min}, and x_{max} are the original values, the minimum value of original samples, and the maximum value of original samples, respectively.

Distance Calculation

For any two samples, we calculated pairwise distances following a two-step strategy, that is, (1) compute the differences between two samples on individual factors and 2) calculate the overall distance by integrating the differences from all factors. We applied binary distance coding to represent the difference among character factors (0 for equal and 1 for difference). For numerical factors, the distance is equal to the absolute value of the difference. Then, overall distance D was calculated by integrating differences on all factors d_i using a weighted L_1 norm (summary of their absolute values with weights). w_i denotes weight of the i-th factor:

$$D = w_i d_{i1} = \sum_{i=1}^{n} \left| w_i d_i \right|.$$

In case of missing values in the dataset, we consider the difference between missing values and any other value as 0. This strategy prevents introducing biases from comparing missing values with real values.

Embedding

To visualize the sampling results, we utilized two state-of-the-art nonlinear dimensional reduction methods, that is, uniform manifold approximation and projection (UMAP) and t-distributed stochastic neighbor embedding (t-SNE) (McInnes et al., 2018; Van der Maaten and Hinton, 2012). Both embedding methods accept pairwise distances as input and render a 2D projection of the samples.

Roles of Factors

In this workflow, we designed three different roles for factors which are as follows: prime factor, important factors, and regular factors. All factors contribute to the distance calculation. The prime factor and important factor are optional in a sampling. The prime factor is unique in a dataset, and the representative samples were selected evenly from each element of prime factor (e.g., subject and animal) instead of selecting from the entire dataset. Important factor indicates a 100% coverage requirement and can be multiple. Regular factors contribute to the distance calculation as other factors do but without any specific requirement to the sampling. The determination of prime factor and important factor is up to users own choice. Users can determine each factor from their dataset to any role (prime, important, or regular) according to their sampling needs and domain knowledge. For example, in the sampling from our single-cell B cell dataset, we would like to select samples from each subject (donors), so that "subject" was set as the prime factor. We would also like to investigate all transcriptional clusters, so that "cluster" was set as an important factor. The rest of the factors were set as regular factors.

Clustering-Based Two-step Sampling Strategy

To achieve a high sampling coverage with good representativeness, we designed a two-step sampling strategy. The first step is to select N samples from the entire dataset or from each subject if the prime factor was set using a k-medoids clustering method. The cost function in k-medoids algorithm is given as (Kaufman et al., 1987)

$$c = \sum_{C_i} \sum_{P_i \in C_i} \left| P_i - C_i \right|,$$

where C_i denotes the medoid and P_i denotes the sample. The most common implementation of k-medoids clustering is the partitioning around medoids (PAM) algorithm. More specifically, samples can be clustered into multiple evenly distributed clusters using the k-medoids clustering method. The medoid for each cluster can be considered as the most representative samples of the corresponding cluster. This step guarantees the representativeness of selected samples.

The second step is to investigate the coverage rate of all important factors (defined by users) from the representative candidates picked by k-medoids clustering from last step. If any important factor has a coverage rate lower than 100%, then an additional selection will be performed to pick the proper samples from the unpicked population to cover all the levels/categories of the important factor. The strategy for adding qualified samples is as follows: for a category of an important factor that has not been covered by samples selected in step 1, if there is more than one candidate, we select the one that has the largest local distances with all selected samples in the first step. We define local distance as

$$D_{Local} = \min_{i \in S} D_i,$$

where D_i denotes the distance between the current sample and the i-th selected samples. S denotes the set of selected samples.

Evaluation of Sampling

The quality of sampling can be evaluated and quantified by coverage rate on each single factor. Here, for character factors, we define the coverage rate as

$$\text{Coverage rate } = \frac{number\ of\ levels\ in\ selected\ samples}{number\ of\ levels\ in\ the\ original\ population}.$$

To be consistent, for numerical factors, since they have been scaled into [0,1], we assigned them to ten evenly divided bins ([0, 0.1] [0.1, 0.2], ... [0.9, 1]); and then the coverage rate of numerical factors can be defined as

$$\text{Coverage rate } = \frac{number\ of\ levels\ in\ selected\ samples}{number\ of\ levels\ in\ the\ original\ population}.$$

Of note, a statistical test between original population and selected population can also be used to evaluate the sampling quality for a numerical factor.

Using the quantified coverage rate on each single factor, users can determine an optimized sample size that balances both factor coverage and cost.

Users can also check the distribution of selected samples on each factor. For example, if the distribution of selected samples is identical to that of the original samples, it indicates that the sampling is of high quality. The similarity of distribution on each factor between the original population and selected samples can also be approximately quantified by Pearson correlation coefficient.

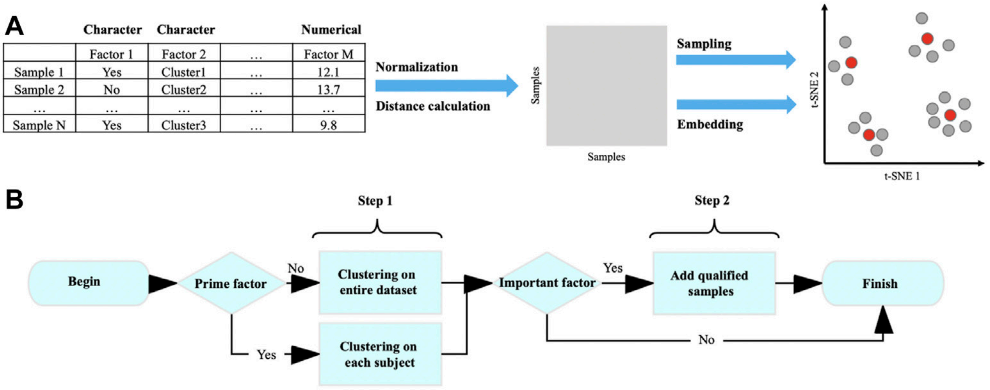
$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}},$$

where n is the number of levels in this factor and x_i and y_i are the number of samples of the i-th level of this factor for two sets of samples; $\bar{x} = \frac{1}{2} \sum_{i=1}^{n} x_i$ and analogously for $\bar{y}$.

RESULTS

Cookie: Representative Sample Selection From a Massive Population Using K-Medoids Clustering

Here, we present Cookie, a user-friendly toolkit, to select representative samples from massive populations (especially single-cell sequencing data). The prime idea of this method is quantifying and vectorizing all samples in order to quantify their dissimilarity by their Manhattan distances, and then samples can be classified into several clusters by k-medoids clustering according to their dissimilarity and centers of those clusters are representative samples (**Figure 1**). In detail, each sample is presented as a numerical vector, and the elements of the vector are attributes of the original data (e.g., subject, transcriptional cluster, and gender). The relationships/dissimilarity among all samples were quantified by calculating a pairwise Manhattan distance matrix. Based on that, a two-step sampling strategy was performed as follows: 1) classify samples into k clusters by k-medoids clustering and select centers of the k clusters and 2) add proper samples to qualify the coverage requirement on specified factors. This method is composed of four steps: normalization, distance calculation, sampling, and embedding (**Figure 1A**). In this toolkit, we defined three roles of factors, prime, important, and regular, to help users better describe their sampling goal. To achieve better representativeness, we designed a two-step sampling strategy (**Figure 1B**). The first step is to select k samples using k-medoids method from the entire population or from each subject of prime factor (Kaufman et al., 1987; Schubert and Rousseeuw, 2018). The second step is to add qualified samples to cover all the categories/levels of important factors (see *Results* for details). Cookie calculated the summary of distances between candidates and selected samples and always picks the one with largest distance if there is more than one candidate.



	Character Factor 1	Character Factor 2	…	Numerical Factor M
Sample 1	Yes	Cluster1	…	12.1
Sample 2	No	Cluster2	…	13.7
…	…	…	…	…
Sample N	Yes	Cluster3	…	9.8

FIGURE 1 | Workflow of k-medoids–based sampling. **(A)** Workflow of Cookie pipeline and **(B)** selecting representative samples using k-medoid clustering method.

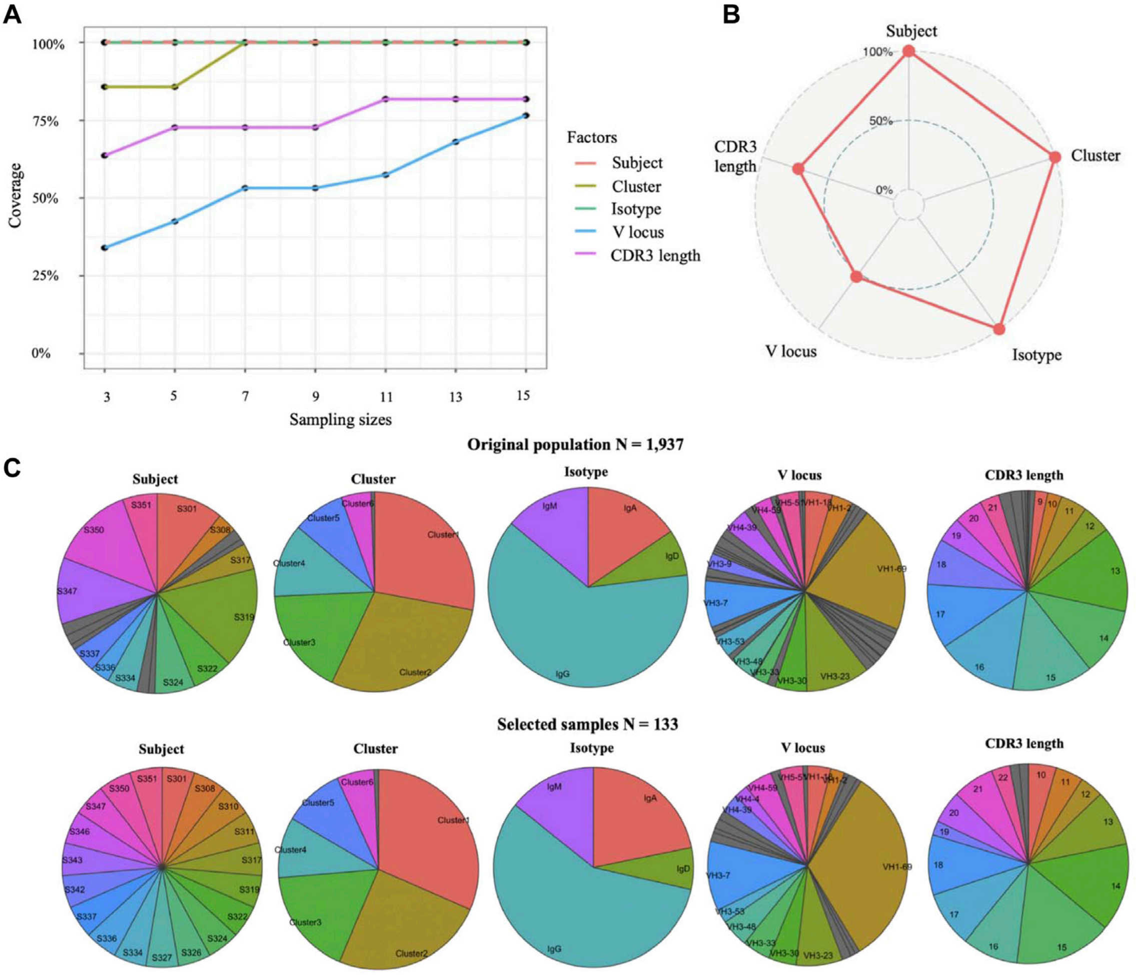


FIGURE 2 | Select representative samples from a large single-cell population. **(A)** Determine appropriate sample size by quantifying coverage on all factors. Line of subject factor was indicated by dash line to avoid overlap. **(B)** Coverage on each factor of selected samples. **(C)** Compare distributions on each factor between original population and select samples.

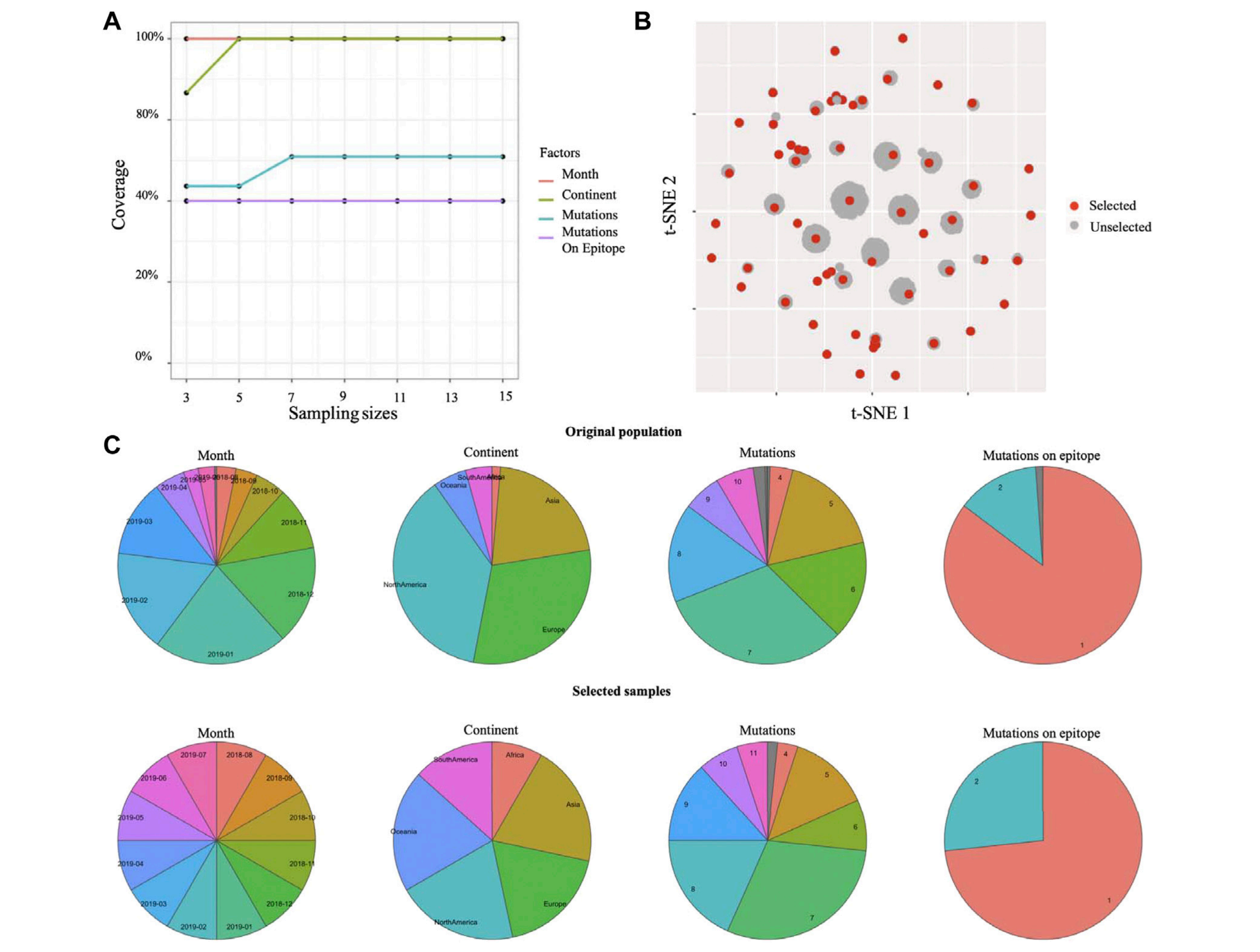


FIGURE 3 | Selected representative samples from human Influenza H1N1 surveillance data. **(A)** Testing coverage rate on each factor of different sample sizes. **(B)** Selected samples and unselected samples on 2D visualization (t-SNE). **(C)** Distributions of each factor of original population and selected samples.

TABLE 1 | Runtime of major steps of Cookie pipeline on different population sizes. All the tests were performed on a simulated dataset using a 2015 Apple MacBook Pro (Core i5, 2.7GHZ, 8 GB DDR3 memory). N denotes population size.

Processing Step		Runtime (seconds)			
		N = 1,000	N = 2,500	N = 5,000	N = 10,000
Preprocess	Create object	0.001	0.004	0.004	0.05
	Normalization	0.003	0.005	0.012	0.027
	Distance calculation	0.347	1.77	7.375	32.511
	Nonlinear reduction (t-SNE)	6.759	13.731	50.701	142.668
Prime factor mode* (PAM algorithm)	Sample size test	0.279	1.414	9.097	54.703
	Sampling	0.04	0.195	1.523	6.134
Prime factor mode* (FastPAM algorithm)	Sample size test	0.663	1.036	4.276	43.047
	Sampling	0.123	0.154	0.578	3.089
Nonprime factor mode** (PAM algorithm)	Sample size test	26.258	301.05	1750.68	>3,000
	Sampling	3.517	36.251	261.614	>3,000
Non-prime factor mode** (FastPAM algorithm)	Sample size test	4.403	27.656	115.79	598.854
	Sampling	0.493	3.31	13.522	68.22

**A prime factor is determined in this run. Algorithms for k-medoids clustering are indicated in the brackets.*

***No prime factor is determined in this run. Algorithms for k-medoids clustering are indicated in the brackets.*

TABLE 2 | Coverage rates of k-medoids sampling on different population sizes. All the tests were performed on a simulated dataset using a 2015 Apple MacBook Pro (Core i5, 2.7GHZ, 8 GB DDR3 memory). N denotes population size. Tests were generated using the Cookie package with the FastPAM method. The sample size for prime factor mode is set to 10 (from each level of prime factor) and that for no-prime factor mode is set to 100.

Factor		Coverage Rate (%)			
		N = 1,000	N = 2,500	N = 5,000	N = 10,000
Prime factor	Factor 1	84.00	82.00	78.00	80.00
	Factor 2	100.00	100.00	100.00	100.00
	Factor 3	100.00	100.00	100.00	100.00
	Factor 4	90.91	100.00	90.91	90.91
	Factor 5	81.82	72.7	72.7	72.73
Nonprime factor	Factor 1	92.00	86.00	88.00	84.00
	Factor 2	100.00	100.00	100.00	100.00
	Factor 3	100.00	100.00	100.00	100.00
	Factor 4	90.91	100.00	90.91	90.91
	Factor 5	81.82	72.73	72.7	72.73

Application of Cookie on Single-Cell Atlas Data to Select Candidates of Monoclonal Antibody for Further Experimental Characterization

We applied this method to select candidates of monoclonal antibody from the isolated genes for 1,937 antibodies for laborious protein expression and downstream analysis. The single-cell atlas dataset consists of seven transcriptional clusters, 19 subjects, a variety of V gene usages, four major isotypes, and a variety of complementarity-determining region 3 (CDR3) lengths among the genes for 1,937 antibodies. Our goal was to 1) select representative samples for laborious protein expression from the genes of 1,937 antibodies and 2) determine the optimized sample size that can balance sampling coverage on all factors and economy. For this dataset, we wanted to evenly select samples from each subject, and a 100% coverage is required for transcriptional clusters. We set “Subject” as the prime factor and “transcriptional cluster” as an important factor. As shown in **Figure 2A**, coverage of all factors is positively correlated with sample sizes from each subject, and N = 7 is the optimal sample size from each subject because 100% coverage of three key factors (subject, cluster, and isotype) and high coverage of other two factors have been achieved. After determining the sample size to seven per subject, we selected 133 samples from 1,937 antibodies with a 100% coverage on subject, cell cluster, and isotype (**Figure 2B**). We observed highly similar distributions between selected 133 samples and the original population by comparing the distribution of five factors (**Figure 2C**). Moreover, the total runtime of sample size determination and sampling is less than ten seconds. In conclusion, results on real single-cell B cell dataset showed that Cookie toolkit is effective and efficient in selecting candidate antibodies for further experimental characterization from massive single-cell population.

Application of Cookie on Human Influenza Virus Surveillance Data

Beside single-cell sequencing data, Cookie toolkit is also compatible with more biological applications. Here, we examined the flexibility of our method on a different type of biological dataset. Influenza virus has a highly mutable replication process allowing it to escape from immunity, often on an annual basis (Kosikova et al., 2018). In order to control this escape, each influenza season, tens of thousands of samples of influenza viruses are collected from surveillance programs across six continents (Lackenby et al., 2018). Identifying antigenic variants from those viral samples is the key to a successful vaccine strain selection to generate a vaccine protective against the most common viral variants (Koel et al., 2013). The main challenge is that people can only investigate antigenic profiles for a small proportion of all viral samples using HI assay, which is time- and labor-intensive. An efficient sampling method that can balance samples with genetic variations, locations, and times of sampling (month) is required. The k-medoids sampling method proposed in this study is capable of addressing this problem. We performed the k-medoids sampling on a human H1N1 influenza dataset with 8,449 viral samples (see dataset section for details) using Cookie toolkit. To identify the earliest antigenic variant, we set “Month” as a prime factor to balance samples from different time periods. The sample size test indicates that a sample size of five (setting sample size to seven will slightly increase the coverage of mutation, if budget allows) is an appropriate choice for this dataset (**Figure 3A**). With the sample size of five, the selected samples covered all the clusters, and therefore are able to represent all of the genetic-temporal-spatial combined variances (**Figure 3B**). The distribution of each factor also shows that the selected samples have a highly similar distribution as the original population (**Figure 3C**). In general, results on two real datasets showed that Cookie specializes in solving contradiction between large detective capabilities and limited experimental capabilities and is compatible with multiple biological contexts.

Evaluation of Cookie Performance Using Simulated Data of Different Population Sizes

To evaluate the efficiency and compatibility of this method, we generated a simulated dataset (see datasets section for details)

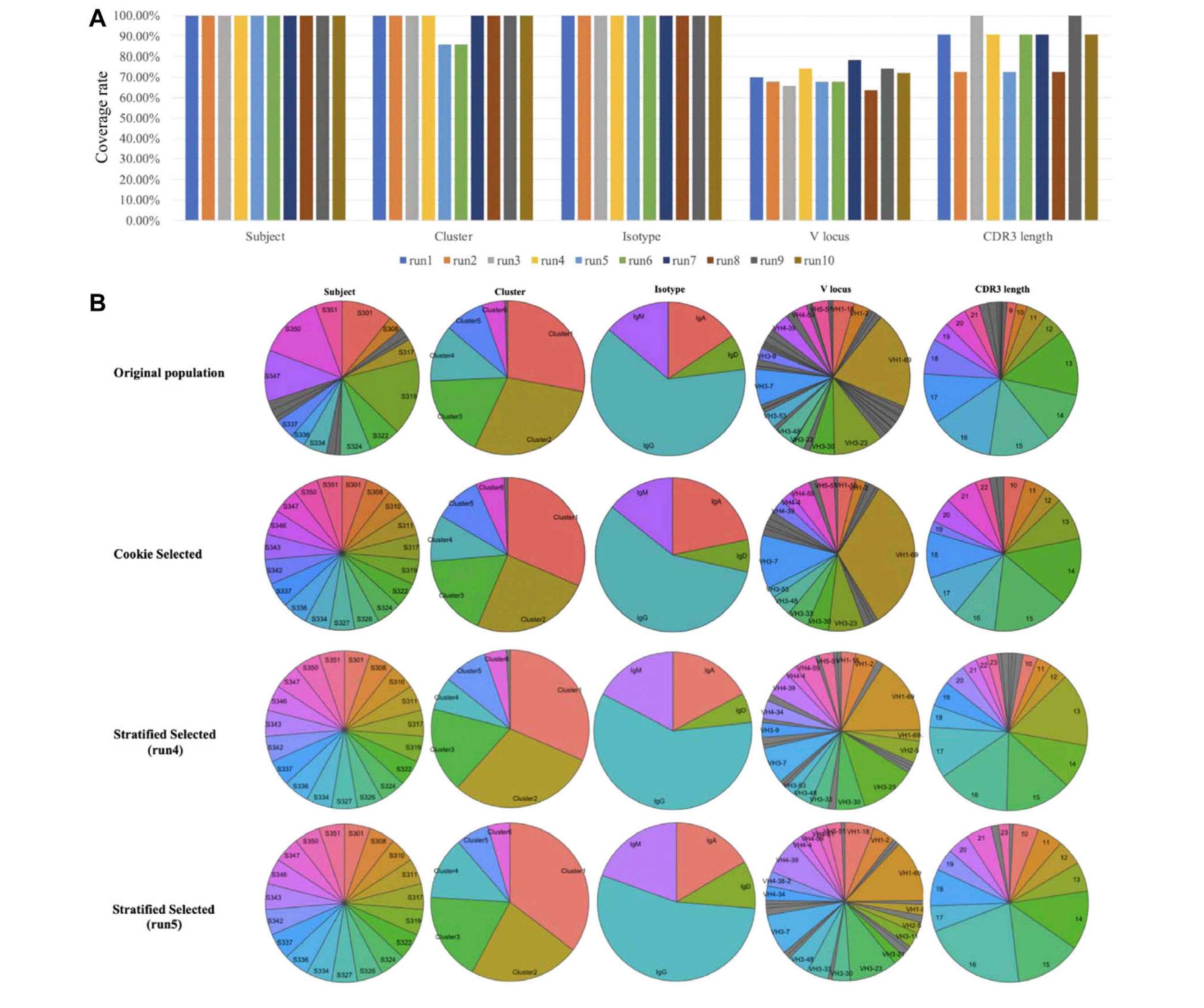


FIGURE 4 | Compare k-medoid sampling with probability sampling method (stratified sampling). **(A)** Coverage rate on each factor of ten independent runs of stratified sampling. **(B)** Distributions of each factor of original population and samples selected by k-medoid sampling and stratified sampling.

and tested our method on this simulated dataset. We compared the runtime of our method on four different data sizes: 1,000, 2,500, 5,000, and 10,000 samples. As shown in **table 1**, with increasing population sizes, the runtime of the four major steps (distance calculation, nonlinear reduction, sample size test, and sampling) increased exponentially. In addition, sampling from the levels of prime factor is much faster than sampling from the entire population, especially for large populations. That is because the runtime complexity of the k-medoids clustering algorithm (also called as PAM algorithm) is $O(k(n-k)^2)$, and the runtime is proportional to population size n and cluster number k. A recent study proposed an optimized k-medoids clustering algorithm called FastPAM that reduces the runtime complexity to $O(n^2)$ (Schubert and Rousseeuw, 2018). By adopting the FastPAM algorithm, the runtime was largely reduced (**Table 1**). Of note, conventional probability sampling methods are much faster than k-medoids sampling because such methods do not (or rarely) use data structure and distribution of the original population as seen in k-medoids. In conclusion, our results indicate that k-medoids sampling is able to effectively and efficiently select representative samples from large populations (**Tables 1**, **2**). These results also demonstrate that sampling from levels of a prime factor or using algorithm acceleration (FastPAM) could significantly reduce the sampling time.

Comparison With a Conventional Probability Sampling Method

The randomness of probability sampling methods is not preferred in antibody selection from single-cell data and some other biological studies. In these cases, distributions and importance of factors are well known. The top priority of sampling is to select the most representative samples based on those factors. Randomness will help less to establish representativeness and may result in inconvenience for further experimental design. Another issue with probability sampling is that the results from two independent probability samplings may be different. Nevertheless, we compared our method to probability sampling methods. We used stratified sampling, the most suitable method for this single-cell dataset among all probability sampling methods, as an example of probability sampling methods. This comparison was performed on our single-cell dataset (see dataset section for details). As shown in **Figure 4**, we compared our method to the stratified sampling method with the same sampling size (select 133 samples from 1,937 cells). Samples were stratified according to “Subject” in stratified sampling. “Subject” was set as the prime factor and “Cluster” was set as an important factor in our method. We performed ten independent runs of stratified sampling on the single-cell dataset, and the results showed that the coverage rates of each factor among ten runs vary (**Figure 4A**), with two runs not even covering all cell clusters (run5 and run6). We picked two from the ten runs (run4 the best and run5 the worst) and compared the results to Cookie selection and the original population (**Figure 4B**). The results show that both k-medoids clustering selection and run4 of stratified sampling are able to represent the original population while run5 fails (fail to select any sample from a small cluster, “Cluster 7”). The results prove that the k-medoids clustering method is not only effective for the selection of representative samples but also able to avoid potential bias caused by the randomness of probability sampling.

DISCUSSION

Based on a k-medoids clustering strategy, we developed a method to select representative samples from a large population. A similar approach for geographical sampling using a k-means clustering method was developed in a prior study (Walvoort et al., 2010). Their results also proved the representativeness and practicability of application of clustering methods in sampling. Of note, their method requires an existing distance measurement among the original samples. It limited the application range of the method since most of biological/clinical datasets do not satisfy the requirement. By developing a workflow consisting of data vectorization and distance calculation steps, our method normalizes different types of factors into the same scale and quantifies the distances among samples based on those normalized factors. This workflow can quantify relationships among samples for all the populations with multiple numerical and non-numerical factors and greatly expand the range of application of our method. Compared to the previous clustering-based sampling approach, our method is advantageous for single-cell populations with complicated structures (multiple factors with different types and priority levels) and compatible with most of the biological datasets.

Conventional probabilistic/nonprobabilistic sampling methods do not or rarely use data structure in sampling. While it highly improves efficiency of sampling process by not using data structure however, the representativeness of samples through random selections usually cannot be guaranteed. By contrast, our method uses the entire data structure when selecting samples. It generates pairwise distance matrix by considering all factors with different priority levels to quantify relationships among samples. Then our method selects samples using k-medoids clustering method by dividing entire population into k clusters. Since the clustering results are subject to pairwise distance that considers all factors, factors of selected samples are therefore maximumly balanced. In other words, the representativeness of selected samples is achieved by balancing all factors of original population. Of note, considering all details in data structure will result in inefficiency, especially for large populations due to the exponential growth of running time as the sample number increases. By introducing a recent proposed method FastPAM, the runtime complexity was greatly reduced. Simulation results showed that Cookie toolkit is capable for robust and efficient sample selection from large populations.

In practice, the number of candidates to be experimentally characterized is usually limited; therefore, selected sample size should be optimized to balance the representativeness and economy. Conventional methods usually do not offer an effective method to determine an appropriate sample size. Furthermore, the representativeness of a sample selection is usually difficult to evaluate. To overcome this challenge, Cookie implemented coverage rate of factors to quantify and evaluate the representativeness of a sample selection. The method also allows users to determine an appropriate sampling size by comparing coverage rates of different sample sizes. In addition, our framework is highly modularized and extendable; other evaluation metrics, for example, Pearson’s correlation coefficient, can also be incorporated to the evaluation process. By evaluating on different population sizes using both experimental data and simulated data, our method was proven to be effective and efficient.

In conclusion, we proposed a sampling method that achieved representativeness, stability, economy, and universality. The method is implemented in an R package Cookie and is freely accessible on GitHub. We hope this toolkit (package) will help biologists select representative samples in an unbiased manner from large-scale datasets.

Limitations

There are two major limitations of this workflow. First, there is only one distance metric (Manhattan distance) in current model. Since different distance measurements can highly affect the clustering results, therefore affecting the final sampling results, investigating effects of different distance measurements is promising to improve the clustering and sampling in the future work. The second limitation is that the time complex of calculating the pairwise distance matrix increases exponentially as the sample number increases. It limits the application of this method on future massive datasets (e.g., datasets have more than 50,000 samples). Furthermore, in current model, we approximately quantify the differences between any two levels of a character factor as the same. A more precise strategy for differences quantification of character factors is also needed to improve the sampling results.

Code Availability

The method is implemented in R and is freely available on GitHub https://github.com/WilsonImmunologyLab/Cookie. The source is also available at Zenodo: https://zenodo.org/record/6639035#.YqdqBRPMIvo. Tutorials and documents are available at https://wilsonimmunologylab.github.io/Cookie/.

The package has been tested under 1) macOS Mojave version 10.14.6 with R version 3.6.0 and RStudio Version 1.2.1335 and 2) ubuntu 18.04 64bit with R version 3.5.

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

LL designed the computational model, wrote the pipeline, performed the analysis, and wrote the manuscript. LY-LL generated the single-cell data and performed the analysis. LH revised the pipeline. CY revised the manuscript. JA revised the manuscript. PW supervised the work and wrote the manuscript.

FUNDING

This project was funded in part by the National Institute of Allergy and Infectious Disease, National Institutes of Health grant numbers 2P01AI097092-06A1, U19AI109946, and U19AI057266, the NIAID Centers of Excellence for Influenza Research and Surveillance (CEIRS) grant numbers HHSN272201400005C, and Bill and Melinda Gates Foundation (OPP1084518). The Center for Research Informatics was funded by the Biological Sciences Division at the University of Chicago with additional funding provided by the Institute for Translational Medicine/Clinical and Translational Award (NIH5UL1TR002389-02) and the University of Chicago Comprehensive Cancer Center Support Grant (NIH P30CA014599).

REFERENCES

Bendall, S. C., Davis, K. L., Amir, E.-A. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., et al. (2014). Single-cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell.* 157, 714–725. doi:10.1016/j.cell.2014.04.005

Chen, W.-R., Yun, Y.-H., Wen, M., Lu, H.-M., Zhang, Z.-M., and Liang, Y.-Z. (2016). Representative Subset Selection and Outlier Detection via Isolation Forest. *Anal. Methods* 8, 7225–7231. doi:10.1039/c6ay01574c

Cochran, W. G. (2007). *Sampling Techniques*. Hoboken, NJ, United States: John Wiley & Sons.

Dugan, H. L., Stamper, C. T., Li, L., Changrob, S., Asby, N. W., Halfmann, P. J., et al. (2021). Profiling B Cell Immunodominance after SARS-CoV-2 Infection Reveals Antibody Evolution to Non-neutralizing Viral Targets. *Immunity* 54, 1290–1303. e7. doi:10.1016/j.immuni.2021.05.001

Fricker, R. D. (2008). "Sampling Methods for Web and E-Mail Surveys," in *The SAGE Handbook of Online Research Methods*. London: SAGE Publications Ltd, 195–216. doi:10.4135/9780857020055.n11

Geman, S., and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-6, 721–741. doi:10.1109/tpami.1984.4767596

Goldstein, L. D., Chen, Y. J., Wu, J., Chaudhuri, S., Hsiao, Y. C., Schneider, K., et al. (2019). Massively Parallel Single-Cell B-Cell Receptor Sequencing Enables Rapid Discovery of Diverse Antigen-Reactive Antibodies. *Commun. Biol.* 2, 304–310. doi:10.1038/s42003-019-0551-y

Hancock, A. A., Bush, E. N., Stanisic, D., Kyncl, J. J., and Lin, C. T. (1988). Data Normalization before Statistical Analysis: Keeping the Horse before the Cart. *Trends Pharmacol. Sci.* 9, 29–32. doi:10.1016/0165-6147(88)90239-8

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* 57, 97–109. doi:10.2307/2334940

Katoh, K., and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 30, 772–780. doi:10.1093/molbev/mst010

Kaufman, L., Rousseeuw, P., and Dodge, Y. (1987). *Clustering by Means of Medoids in Statistical Data Analysis Based on the. L1 Norm*. North-Holland, Amsterdam.

Koel, B. F., Burke, D. F., Bestebroer, T. M., van der Vliet, S., Zondag, G. C. M., Vervaet, G., et al. (2013). Substitutions Near the Receptor Binding Site Determine Major Antigenic Change during Influenza Virus Evolution. *Science* 342, 976–979. doi:10.1126/science.1244730

Kosikova, M., Li, L., Radvak, P., Ye, Z., Wan, X.-F., and Xie, H. (2018). Imprinting of Repeated Influenza A/H3 Exposures on Antibody Quantity and Antibody Quality: Implications for Seasonal Vaccine Strain Selection and Vaccine Performance. *Clin. Infect. Dis.* 67, 1523–1532. doi:10.1093/cid/ciy327

Lackenby, A., Besselaar, T. G., Daniels, R. S., Fry, A., Gregory, V., Gubareva, L. V., et al. (2018). Global Update on the Susceptibility of Human Influenza Viruses to Neuraminidase Inhibitors and Status of Novel Antivirals, 2016-2017. *Antivir. Res.* 157, 38–46. doi:10.1016/j.antiviral.2018.07.001

Li, L., Chang, D., Han, L., Zhang, X., Zaia, J., and Wan, X. F. (2020). Multi-task Learning Sparse Group Lasso: a Method for Quantifying Antigenicity of Influenza A(H1N1) Virus Using Mutations and Variations in Glycosylation of Hemagglutinin. *BMC Bioinforma.* 21, 182. doi:10.1186/s12859-020-3527-5

Li, L., Dugan, H. L., Stamper, C. T., Lan, L. Y.-L., Asby, N. W., Knight, M., et al. (2021). Improved Integration of Single-Cell Transcriptome and Surface Protein Expression by LinQ-View. *Cell. Rep. Methods* 1, 100056. doi:10.1016/j.crmeth.2021.100056

Lumley, T., and Lumley, M. T. (2019). Package 'survey'.

McCarty, C., Bernard, H. R., Killworth, P. D., Shelley, G. A., and Johnsen, E. C. (1997). Eliciting Representative Samples of Personal Networks. *Soc. Netw.* 19, 303–323. doi:10.1016/s0378-8733(96)00302-4

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.

Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., et al. (2017). Multiplexed Quantification of Proteins and Transcripts in Single Cells. *Nat. Biotechnol.* 35, 936–939. doi:10.1038/nbt.3973

Schubert, E., and Rousseeuw, P. J. (2018). Faster K-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. *arXiv preprint arXiv: 1810.05691*.

Shu, Y., and McCauley, J. (2017). GISAID: Global Initiative on Sharing All Influenza Data - from Vision to Reality. *Eurosurveillance* 22, 30494. doi:10.2807/1560-7917.es.2017.22.13.30494

Siddiqui, S., Okasha, T. M., Funk, J. J., and Al-Harbi, A. M. (2006). Improvements in the Selection Criteria for the Representative Special Core Analysis Samples. *SPE Reserv. Eval. Eng.* 9, 647–653. doi:10.2118/84302-pa

Singh, D., and Singh, B. (2020). Investigating the Impact of Data Normalization on Classification Performance. *Appl. Soft Comput.* 97, 105524. doi:10.1016/j.asoc.2019.105524

Stuart, T., and Satija, R. (2019). Integrative Single-Cell Analysis. *Nat. Rev. Genet.* 20, 257–272. doi:10.1038/s41576-019-0093-7

Sun, H., Yang, J., Zhang, T., Long, L. P., Jia, K., Yang, G., et al. (2013). Using Sequence Data to Infer the Antigenicity of Influenza Virus. *MBio* 4, e00230–13. doi:10.1128/mBio.00230-13

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential Scaling of Single-Cell RNA-Seq in the Past Decade. *Nat. Protoc.* 13, 599–604. doi:10.1038/nprot.2017.149

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq Whole-Transcriptome Analysis of a Single Cell. *Nat. Methods* 6, 377–382. doi:10.1038/nmeth.1315

Tillé, Y., Matei, A., Matei, M. A., and Imports, M. A. S. S. (2016). Package 'sampling'. Survey Sampling. *Kasutatud* 23, 2017.

Tillé, Y., and Matei, A. (2006). The R Package Sampling, a Software Tool for Training in Official Statistics and Survey Sampling, 1473–1482.

Tominaga, Y. (1998). Representative Subset Selection Using Genetic Algorithms. *Chemom. Intelligent Laboratory Syst.* 43, 157–163. doi:10.1016/s0169-7439(98)00085-9

Van der Maaten, L., and Hinton, G. (2012). Visualizing Non-metric Similarities in Multiple Maps. *Mach. Learn.* 87, 33–55.

Walvoort, D. J. J., Brus, D. J., and de Gruijter, J. J. (2010). An R Package for Spatial Coverage Sampling and Random Sampling from Compact Geographical Strata by K-Means. *Comput. Geosciences* 36, 1261–1267. doi:10.1016/j.cageo.2010.04.005

Winkels, H., Ehinger, E., Vassallo, M., Buscher, K., Dinh, H. Q., Kobiyama, K., et al. (2018). Atlas of the Immune Cell Repertoire in Mouse Atherosclerosis Defined by Single-Cell RNA-Sequencing and Mass Cytometry. *Circ. Res.* 122, 1675–1688. doi:10.1161/circresaha.117.312513

Zhang, L., Dong, X., Lee, M., Maslov, A. Y., Wang, T., and Vijg, J. (2019). Single-cell Whole-Genome Sequencing Reveals the Functional Landscape of Somatic Mutations in B Lymphocytes across the Human Lifespan. *Proc. Natl. Acad. Sci. U.S.A.* 116, 9014–9019. doi:10.1073/pnas.1902510116

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.

Copyright © 2022 Li, Lan, Huang, Ye, Andrade and Wilson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 22 August 2022
DOI 10.3389/fgene.2022.990412

Check for updates

OPEN ACCESS

EDITED BY
Quan Zou,
University of Electronic Science and Technology of China, China

REVIEWED BY
Chen Cao,
Nanjing Medical University, China
Xiaoning Cao,
Shanxi Agricultural University, China

*CORRESPONDENCE
Dabing Xiang,
xiangdabing@cdu.edu.cn
Zhibin Lv,
lvzhibin@pku.edu.cn

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 09 July 2022
ACCEPTED 18 July 2022
PUBLISHED 22 August 2022

CITATION
Jiang L, Liu C, Fan Y, Wu Q, Ye X, Li Q, Wan Y, Sun Y, Zou L, Xiang D and Lv Z (2022), Dynamic transcriptome analysis suggests the key genes regulating seed development and filling in Tartary buckwheat (Fagopyrum tataricum Garetn.). *Front. Genet.* 13:990412. doi: 10.3389/fgene.2022.990412

COPYRIGHT
© 2022 Jiang, Liu, Fan, Wu, Ye, Li, Wan, Sun, Zou, Xiang and Lv. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Dynamic transcriptome analysis suggests the key genes regulating seed development and filling in Tartary buckwheat (*Fagopyrum tataricum* Garetn.)

Liangzhen Jiang[1], Changying Liu[1], Yu Fan[1], Qi Wu[1], Xueling Ye[1], Qiang Li[1], Yan Wan[1], Yanxia Sun[2], Liang Zou[1], Dabing Xiang[1]* and Zhibin Lv[3]*

[1]Key Laboratory of Coarse Cereal Processing, Ministry of Agriculture and Rural Affairs, Sichuan Engineering & Technology Research Center of Coarse Cereal Industralization, College of Food and Biological Engineering, Chengdu University, Chengdu, China, [2]College of Tourism and Culture Industry, Chengdu University, Chengdu, China, [3]Department of Medical Instruments and Information, College of Biomedical Engineering, Sichuan University, Chengdu, China

Tartary buckwheat is highly attractive for the richness of nutrients and quality, yet post-embryonic seed abortion greatly halts the yield. Seed development is crucial for determining grain yield, whereas the molecular basis and regulatory network of Tartary buckwheat seed development and filling is not well understood at present. Here, we assessed the transcriptional dynamics of filling stage Tartary buckwheat seeds at three developmental stages by RNA sequencing. Among the 4249 differentially expressed genes (DEGs), genes related to seed development were identified. Specifically, 88 phytohormone biosynthesis signaling genes, 309 TFs, and 16 expansin genes participating in cell enlargement, 37 structural genes involved in starch biosynthesis represented significant variation and were candidate key seed development genes. *Cis*-element enrichment analysis indicated that the promoters of differentially expressed expansin genes and starch biosynthesis genes are rich of hormone-responsive (ABA-, AUX-, ET-, and JA-), and seed growth-related (MYB, MYC and WRKY) binding sites. The expansin DEGs showed strong correlations with DEGs in phytohormone pathways and transcription factors (TFs). In total, phytohormone ABA, AUX, ET, BR and CTK, and related TFs could substantially regulate seed development in Tartary buckwheat through targeting downstream expansin genes and structural starch biosynthetic genes. This transcriptome data could provide a theoretical basis for improving yield of Tartary buckwheat.

KEYWORDS

Tartary buckwheat, seed development, transcriptome, expression analysis, expansin, phytohormone pathways, starch biosynthesis

1 Introduction

Seed is the most important organ of crop. Seed development is an essential and complex process, which involves seed size change and nutrients accumulation (Kong et al., 2015). Broadly, seed development can be divided into two important phases: embryogenesis and maturation. In embryogenesis phase, cells divide and expand to establish the tissues and organelles, while in the maturation phase, resources are allocated to synthesize storage compounds (Ruuska et al., 2002). Kernel weight of seeds plays a vital role in the yield of cereal crops and is determined by the duration and rate of grain filling. Thus, improving grain filling increases the grain weight and cereal yield (Savadi, 2018). Grain filling is predominantly regulated by genetic factors and is also greatly influenced by physiological pathways and environmental factors. Phytohormones, including auxin (AUX), cytokinin (CTK), gibberellin (GA), brassinosteroid (BR), abscisic acid (ABA), and ethylene (ET) contribute to seed development (Jameson and Song, 2016; Savadi, 2018; Li N. et al., 2019; Kozaki and Aoyanagi, 2022). Various genes involved in different mechanisms control the process of seed development/grain filling (Tzafrir et al., 2004). Recently, much research was conducted to investigate the specific genes or pathways controlling seed development in barley (Bian et al., 2019), rice (Bian et al., 2019), maize (Chen et al., 2014; Li et al., 2014), wheat (Cantu et al., 2011; Li et al., 2013), soybean (Jones and Vodkin, 2013; Lu et al., 2016; Du et al., 2017), chickpea (Garg et al., 2017), and *Brassica napus* (Basnet et al., 2013; Zhou et al., 2017). These genes include TFs, hormone related genes, genes involved in seed size regulation, seed storage proteins (SSPs), starch and lipid biosynthesis. However, key seed developmental genes and their regulatory networks in plants were far more than elucidated, especially in those non-model crops.

Tartary buckwheat (*Fagopyrum tataricum* Garetn.), as an edible and medicinal non-model crop, is becoming highly attractive for the high-quality proteins and pharmaceutical ingredients, such as flavonoids, polyphenols, and D-chiro-inositol in the seeds (Li and Zhang, 2001). However, the yield of Tartary buckwheat is only about 1500 kg/ha, which is remarkably lower than staple crops, such as rice or wheat. The production is hard to be broken-through by laborious agricultural strategies only (Xiang et al., 2016; Xiang et al., 2019; Xiang et al., 2020). Seed development is crucial for determining grain yield, and elucidation of the molecular mechanism of seed development could potentially improve yield through molecular breeding. Nevertheless, poor post-embryonic grain filling always occurred in Tartary buckwheat seed development process, which greatly hinders the grain yield improvement. Therefore, it is of considerable interest to identify key genes and dissect the molecular mechanisms of seed development in Tartary buckwheat.

Recently, *FtARF2* was reported to promote Tartary buckwheat fruit enlargement by prolonging the cycle of embryonic development and increasing the cycle of cell division (Liu et al., 2018b). Cytochrome P450 monooxygenase superfamily participating in the synthesis of flavonoids, plant growth and development in Tartary buckwheat were clarified (Sun et al., 2020). Five members of *FtCYP78A* family were suggested to be candidate genes that regulate seed size (Sun et al., 2020). Other studies of the Tartary buckwheat seed development by transcriptome analysis mainly focused on the molecular foundation of nutrients accumulation, such as flavonoid (Gao et al., 2017; Huang et al., 2017; Liu et al., 2018a; Li H. Y. et al., 2019). Yet other mechanisms involved in the molecular basis and regulatory network of seed development, especially those governing post-embryonic grain filling process, such as cell enlargement and starch accumulation, were not clarified. Thus, we carried out global transcriptional expression profiling at three stages spanning important developmental stages of seed development to identify the potential regulators involved in these processes in Tartary buckwheat seeds.

2 Materials and methods

2.1 Plant materials and growth conditions

Seeds of Tatary buckwheat (*F. tataricum* cv. Xiqiao No.1) was used as the experimental materials in this study, which were obtained from the Key Laboratory of Coarse Cereal Processing, Ministry of Agriculture and Rural Affairs, Chengdu, Sichuan Province, China. The seeds were sown and grown in plastic pots (25 cm in diameter, 20 cm in height) at the density of 8 seeds for each pot, and the seedlings were thinned to three at the early vegetative stage (cotyledons). Each pot contained 15 kg of air-dried soil (remove large stones, plant roots, and other litter), and the soil was sandy soil in texture and alkaline (pH = 7.88) with 48.3, 20.7, and 31.1 mg kg^{-1} available N, P, and K, respectively; 0.76, 0.49, and 12.8 g kg^{-1} total N, P, and K, respectively; and 10.2 g kg^{-1} organic matter. The soil physical properties (0–0.2 m) were determined according to the method proposed by Pang et al. (2006). The pot was placed in the field to keep consistent with field production under normal agricultural management.

At the beginning of anthesis stage, we labeled and recorded the time of flowering, as to determine the time of anthesis and developmental stage of grain. The time of anthesis and developmental stage of Tartary buckwheat grain were determined as previously described (Song et al., 2016). We selected the grain of Tartary buckwheat at the stage 1, stage 2 and stage 3 (Stage 1, Seed formation started; Stage 2, Milk-ripe stage, the endosperm is solidifying; Stage 3, The seeds matured, and pericarp was totally black; Figure 1) to collect the sample, and three biological replicates were sampled. For each sample, seeds were collected directly into liquid nitrogen by decorticating the

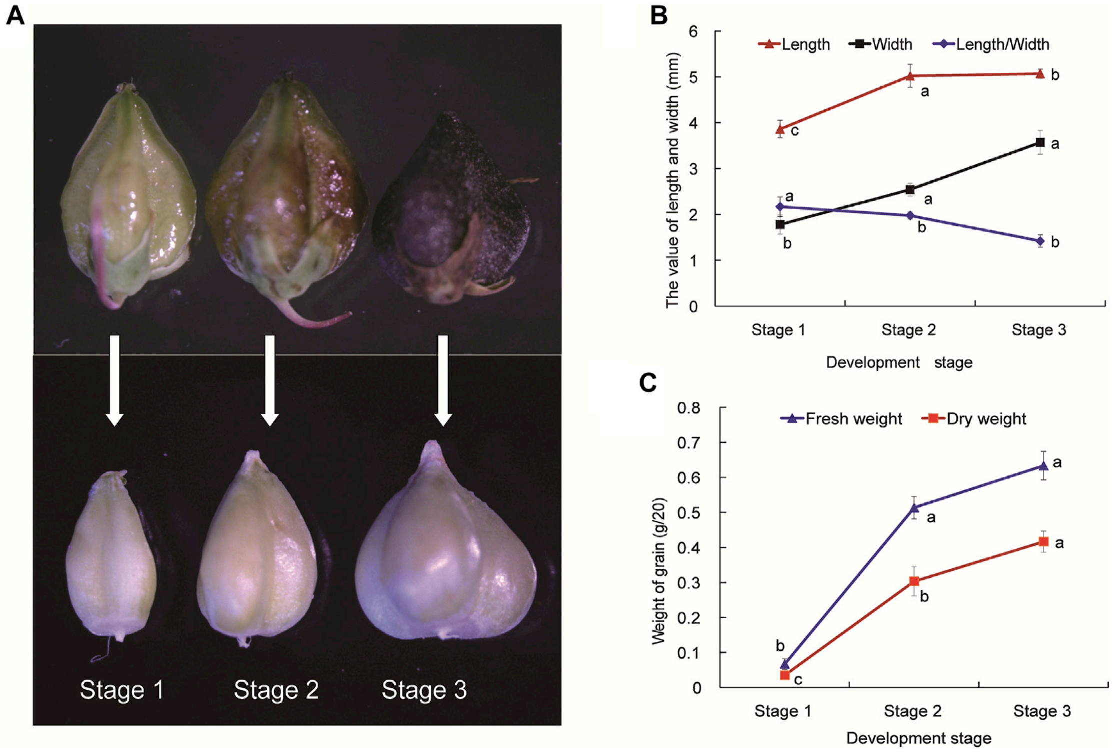


FIGURE 1
The developmental stages **(A)**, agronomic traits **(B)** and weight **(C)** of Tartary buckwheat seeds. Stage 1 (S1), Seed formation started, torpedo embryo; Stage 2 (S2), Milk-ripe stage, the endosperm is solidifying, initial mature embryo; Stage 3 (S3), The seed matured, mature embryo and pericarp was totally black. Data are presented as the mean ± SD of three biological replicates. Grains were divided into different developmental stages with 20 grains measured at each stage. Different letters denote significant differences ($p < 0.05$).

grain from 20 to 30 individual plants for the purposes of homogeneity.

2.2 Analysis of seed morphology and weight

The twenty grains per replicate were sampled and measured at stage 1, 2 and 3 (Figure 1), and three biological replicates were sampled. The length, width and length/width of grain, fresh and dry weight of grain were measured. The dry weight was oven-dried at 65°C to constant weight and then measured.

2.3 Sample collection and preparation for RNA sequencing

The purity, concentration and integrity of RNA samples are tested using advanced molecular biology equipment to ensure the use of qualified samples for transcriptome sequencing. A total amount of 1 μg RNA per sample was used as input material for the RNA sample preparations. Sequencing libraries were generated using NEBNext UltraTM RNA Library Prep Kit from Illumina (NEB, United States) following the manufacturer's recommendations and index codes were added to attribute sequences to each sample.

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v4-cBot-HS (Illumia) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina platform and paired-end reads were generated.

2.4 Data analysis

The adaptor sequences and low-quality sequence reads were removed from the data sets. Raw sequences were transformed into clean reads after data processing. These clean reads were then mapped to the reference genome sequence of the Tartary buckwheat (Pinku1) genome (http://www.mbkbase.org/Pinku1/) using TopHat (v2.0.12) (Zhang et al., 2017). Only reads with a perfect match or one mismatch were further analyzed and annotated based on the reference genome. Hisat2 tools soft were used to map with reference genome.

Differential expression analysis of two conditions/groups was performed using the DEseq. DEseq provide statistical routines for determining differential expression in digital gene expression data using a model based on the negative binomial distribution. The resulting p values were adjusted using the Benjamini and Hochberg's approach for controlling the false discovery rate. Genes with an adjusted p-value < 0.01 found by DEseq were assigned as differentially expressed. To gain insight into the function of DEGs, KEGG enrichment analysis of the DEGs was implemented by using the KOBAS software (Kanehisa et al., 2008). Pathways with correlated p value less than 0.05 were assigned as significantly enriched items.

K-means clustering analysis of DEGs using the k-means function in R, where k = 12 within the cluster package by Euclidean distance. Heat maps were drawn using the P heatmap package in R and were clustered using Pearson correlation distance. The search for orthologous genes of seed size-related genes from other species in Tartary buckwheat (Pinku1) genome was performed *via* TB tools (Chen et al., 2020). The promoter sequences (about 1.5k bps upstream of the transcription start site) of DEGs were extracted from Tartary buckwheat (Pinku1) genome by TB tools. The *cis*-regulatory elements enrichment analysis of the promoter area of DEGs involved in seed development was performed by the online prediction tool Plantcare (Lescot et al., 2002) (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/), with defaulted parameters for TF family assignment and thresholds. The illustration of the predicting TF-binding sites was drawn by TB tools.

3 Results

3.1 Shape and weight change of seeds during Tartary buckwheat development

To explore the molecular mechanisms of grain filling process during Tartary buckwheat seed development, three stages (Stage 1, Seed formation started, torpedo embryo; Stage 2, Milk-ripe stage, initial mature embryo, the endosperm is solidifying; Stage 3, The seeds matured, mature embryo, the pericarp was totally black; Figure 1A) of seeds were selected. Several agronomic traits including length, width, and length/width, fresh and dry weight were measured. As shown in Figure 1A, outer fruit shape changed gently with the seed development (upper picture), with both the length and width of seed coat reaching a maximum size in S3 phase, which sets an upper limit to final size of a grain. Yet significant sharper physiological changes were seen among the dehulling seeds without coats, which continuously expanded until the maturation phase in Tartary buckwheat (Figure 1A, lower picture). As kernel weight of seeds plays a vital role in the yield of cereal crops, dehulling seeds were then analyzed for the grain filling process in the following research. The seed length and width, especially width, increased significantly (p < 0.05) from stage 1 to stage 3 (Figure 1B). The ratio of seed length to width was substantially the highest at stage 1 among three stages, and the grains became conical gradually after stage 2. The weight of grains changed significantly (p < 0.05) during seed maturation (Figure 1C). Accompany with the shape change, both the fresh and dry weight of Tartary buckwheat seed considerably elevated with the seed growth, with the highest in stage 3 (0.63 and 0.42 g/20 grains, respectively).

3.2 Transcriptional profiles of filling stage seeds

The dynamics of mRNA abundance at three pivotal stages (Figure 1) of grain development in Tartary buckwheat were assessed. Totally, 40.67 Gb of raw data was obtained for all of the samples, with the average about 22.6 million pair-end reads with 150 bp in size for each sample. After removing the low-quality reads and adaptor sequences from reads, in total we obtain 203 million clean reads from S1, S2 and S3, which yielded 7.16 billion, 6.46 billion and 6.63 billion nucleotides, respectively (Table 1). These clean reads could represent 94.07% of the raw data. With respect to GC content, the S1, S2 and S3 library reached 48.78, 47.55 and 46.23%, respectively (Table 1). After quality filtering, all of the clean reads were mapped to the reference genome of Tartary buckwheat (Zhang et al., 2017). The results showed that on average 45.19 million paired-end reads (93.59%) could be mapped to the reference genome. The average number uniquely mapped to the reference genome at three stages was 39.70 million paired-end reads (87.77%), with the range from 38.26 (88.54%) to 40.03 (90.18%) (Table 1).

Gene expression pattern was calculated by the fragments per kilobase of exon per million mapped reads (FPKM) method. Based on the gene expression levels, Pearson correlation coefficient between different samples was calculated. The results indicated that three biological replicates of all samples demonstrated consistent determinations of transcript abundance with a coefficient (R^2) greater than 0.842 (Figure 2). Simultaneously, the correlation between gene expression and different developmental stages was compared. We found that the coefficients of S1 and S2 were 0.624 (R^2 < 0.63), while those of S2 and S3 were even lower, with the value <0.317. PCA (Principle Component Analysis) grouping of different sample expression profiles displayed that all the samples were separated into three groups, which is consistent with the correlation results above (Figure 3). Venn diagram analysis showed that 15,301 genes were ubiquitously expressed in all samples, and 388, 264 and 1,941 genes showed specific expression in S1, S2 and S3,

TABLE 1 Characteristics of generated read data and results of sequence mapped to the reference genome.

Item	Grain development stage		
	S1	S2	S3
Total Reads	47,963,131	43,215,621	44,398,329
Clean reads	23,981,565	21,607,810	22,199,164
Clean base number	7,163,455,810	6,457,976,261	6,631,570,818
GC content (%)	48.78	47.55	46.23
Q30 percentage (%)	94.21	93.57	94.42
Mapped Reads	44,285,248 (92.09%)	40,878,284 (94.59%)	41,774,374 (94.10%)
Uniq Mapped Reads	40,802,372 (84.60%)	38,259,936 (88.54%)	40,027,825 (90.18%)
Multiple Map Reads	3,482,875 (7.49%)	2,618,348 (6.05%)	1,746,549 (3.92%)
Reads Map to "+"	21,874,547 (45.46%)	20,157,291 (46.65%)	20,732,040 (46.70%)
Reads Map to "–"	21,977,813 (45.68%)	20,322,772 (47.03%)	20,779,812 (46.81%)

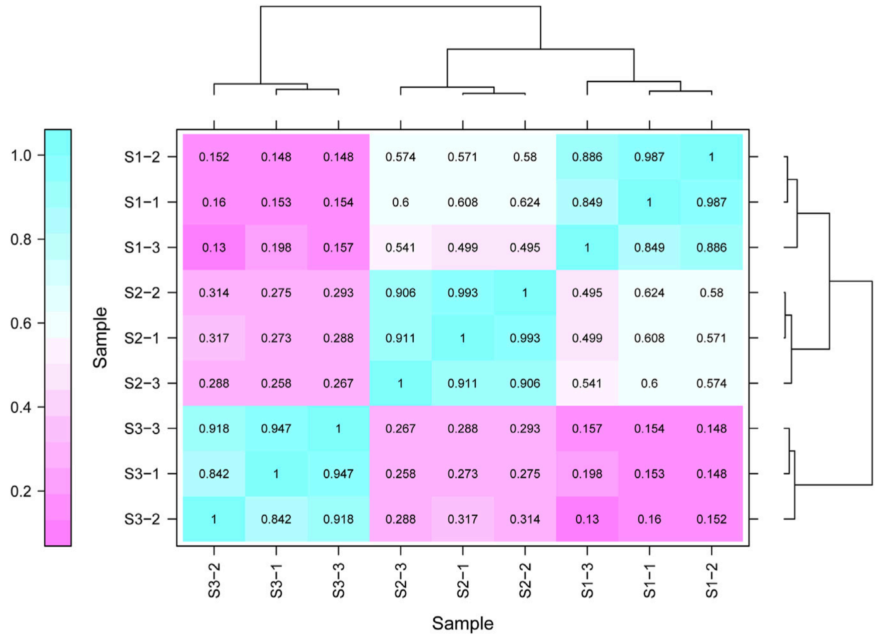


FIGURE 2
The correlation data sets between the gene expression and three growth stages of Tartary buckwheat with three biological duplicates.

respectively (Figure 4). S3 harbored more distinctly-expressed genes than S2 and S1.

3.3 Analysis of differentially expressed genes among the seeds at three developmental stages

With an adjusted p-value (P) < 0.01 and fold change (FC) > 2, 10065 significant differentially expressed genes (DEGs) were identified totally at three stages of seed development. As shown in Supplementary Figure S1, compared with S1, 1771 and 1330 genes increased and decreased the expression respectively in S2, while the number of genes being up-regulated and down-regulated specifically in S3 was 4500 and 4135 respectively (Supplementary Figure S1). To acquire the information of key genes involved in seed filling/development in Tartary buckwheat, a cutoff of FC > 4 and p < 0.01 was applied for gene analysis. In brief, a total of 4249 DEGs were identified by pair-wised comparison of three stages of seed development,

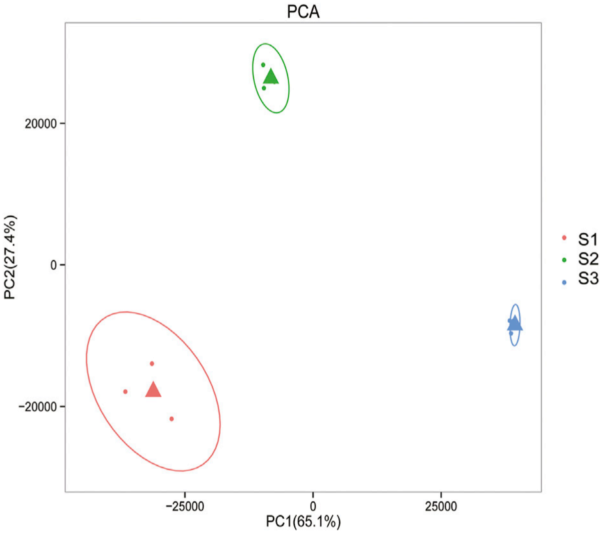


FIGURE 3
Principal component analysis of gene expression profiles indicated that all the samples were divided into three distinct groups.

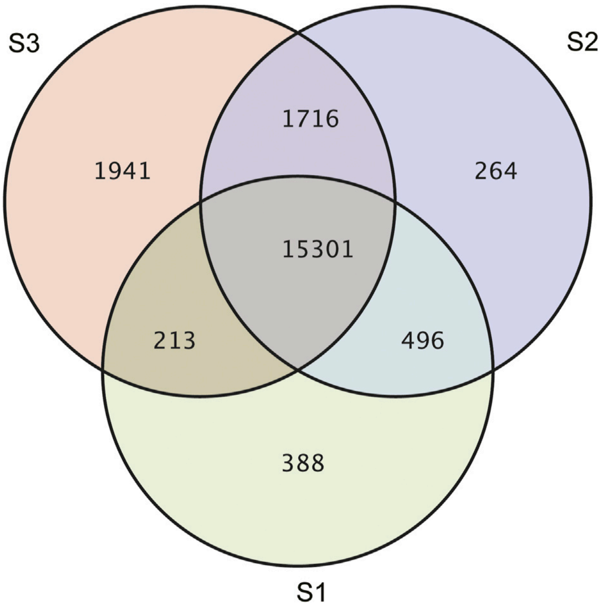


FIGURE 4
A Venn diagram showing the specifically or commonly expressed genes in different groups.

among which 376 were new genes. These 4249 DEGs were selected for subsequent further analysis. 793, 1891 and 3874 genes were differentially expressed in comparison of S1 vs. S2, S2 vs. S3 and S1 vs. S3, respectively (Supplementary Figure S2). Compared with S1, 406 and 278 DEGs were commonly up-regulated and down-regulated respectively both in S2 and S3 (Figures 5A,B). Moreover, 78 and 31 DEGs were specifically up-regulated and down-regulated in S2 respectively by comparing with S1, while 1549 and 1641 DEGs respectively were specifically up-regulated and down-regulated in S3. 133 up-regulated genes and 166 down-regulated genes respectively were identified particularly in comparison S2 vs. S3. This together with the PCA result indicated that sharp transitions in gene expression occurred along with the seed development in Tartary buckwheat, especially in the transition from S2 to S3, which is consistent with the shape and weight changes of seeds in Figure 1 and previous reports that significant transcriptional and physiological changes occurred with seed growth in Tartary buckwheat (Liu et al., 2018a).

To disclose the expression pattern of the DEGs during seed development, the gene expression profile clustering was conducted by the K-means method. All DEGs were assigned to 12 different kinetic clusters with similar expression patterns (Figure 6). Though with different extent of variation, Cluster 1, 7 and 9 showed continuously rising expression patterns, while Cluster 3, 4 and 8 displayed oppositely declined expression trend successively along with the seed maturation. Among them, Cluster 7 (107) and 9 (282) displayed the most vigorous continuous upward trend, as Cluster 4 (68) was the top continuous downward cluster. In Cluster 5, 138 DEGs tend to increase the transcription specifically in transition from S1 to S2. Compared with S1, the DEGs in Cluster 2 (196) declined the expression only in S3, while DEGs in Cluster 11 (413) seemed to increase the expression more sharply in S3 than in S2.

3.4 KEGG and GO analysis of differentially expressed genes

To gain more insight into the biological function of these genes, KEGG and GO analysis were conducted and the functional enrichment results were obtained as shown in Supplementary Figures S3–S7. Among the upward trend clusters, Cluster 7 was significantly associated with ubiquitin mediated proteolysis (3) by KEGG analysis. By GO enrichment, Cluster 7 was more associated with DNA binding (28) and protein heterodimer activity (20) in molecular function, and nucleosome assembly (21) and cell proliferation (6) in biological process. Cluster 9 was dominantly enriched with ribosome (11), starch and sucrose metabolism (8), carbon metabolism (8) and carbon fixation in photosynthetic organisms (7) by KEGG analysis. The GO analysis of Cluster 9 showed enrichment of DNA binding (20), structural constituent of ribosome (10) and protein heterodimer activity (8) in molecular function, and nucleosome assembly (10), regulation of cell cycle (8), and cell proliferation (7) in biological process. The downward trend clusters were

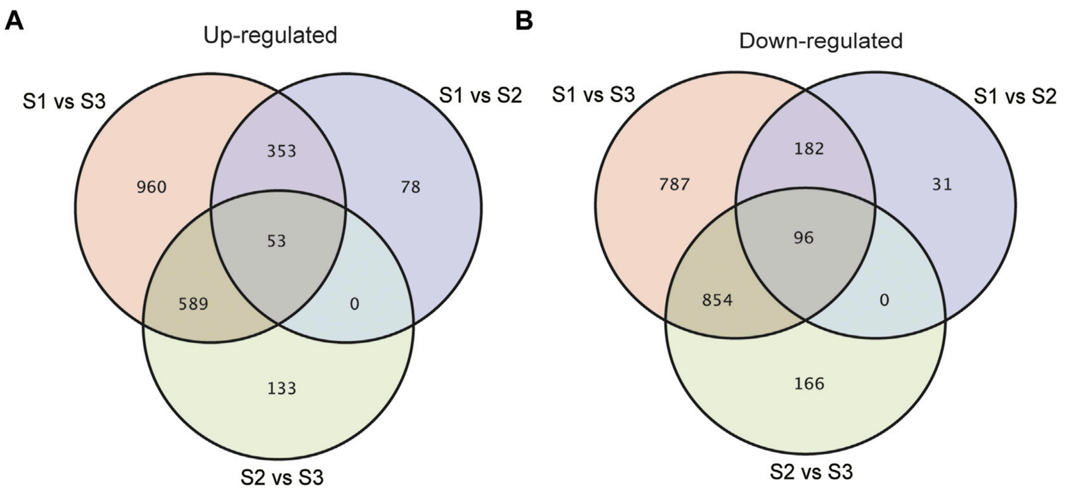


FIGURE 5
Venn diagrams showing the up-regulated **(A)** or down-regulated **(B)** DEGs in pair-wised comparison groups with $p < 0.01$ and FC > 4.

particularly enriched into protein processing in endoplasmic reticulum (9) in Cluster 4, diterpenoid biosynthesis (4) and glutathione metabolism (4) in Cluster 3 by KEGG analysis. Yet by GO analysis of Cluster 4, protein folding (8), response to hydrogen peroxide (7), response to high light intensity (7) in biological process were enriched. Cluster 3 was enriched with oxidation-reduction process (16), response to oxidative stress (4) in biological process, and iron ion binding (13), heme binding (10) and transcription factor activity (8) in molecular function. Cluster 5 (138) was particularly enriched with plant hormone transduction (4) and starch and sucrose metabolism (3) by KEGG analysis, and oxidation-reduction process (13) in biological process and copper ion binding (4) in molecular function by GO enrichment. For Cluster 11, ribosome (22), ribosome biosynthesis in eukaryote (12), RNA transport (7) and cysteine and methionine metabolism (6) were particularly associated by KEGG analysis, and translation (23), carbohydrate metabolic process (11) and nucleosome assembly (10) in biological process and DNA binding (31), structural constituent of ribosome (26), protein heterodimer activity (10) and heme binding (9) in molecular function were enriched by GO analysis. Cluster 2 was significantly associated with protein processing in endoplasmic reticulum (9) by KEGG analysis, unfolded protein binding (6) and nutrient reservoir (5) in molecular function, response to stress (10) and protein folding (7) in biological process by GO enrichment. In summary, these results together indicated that remarkable dynamics of chromatin structure, active ribosome biosynthesis and cell proliferation took place along with the seed maturation, which was accompanied with considerable elevated protein synthesis, DNA binding and carbohydrate metabolism. Yet, with the filling process, the seeds went through lower level of protein processing and secondary metabolite synthesis.

3.5 Analysis of key genes involved in the seed development of Tartary buckwheat

3.5.1 Dynamic transcriptome analysis of phytohormone signaling pathway genes involved in seed development

Phytohormones play notable roles in seed development in rice, maize and Arabidopisis (Santner et al., 2009). Several DEGs were assigned to "plant hormone signal transduction" (ko04075) in the transcriptome data by KEGG enrichment analysis. Genes involved in phytohormone biosynthesis and signaling pathways were thus elaborately analyzed. For ABA biosynthesis (Figure 7A), 3 *NCED*, which encode enzymes catalyzing the rate-limiting step in ABA biosynthesis, and one *AAO*, displayed a descending tendency of expression in maturation phase. Among the ABA signaling genes, 4 *PP2C* gradually down-regulated the expression with the seed growth, while the transcription of 2 *PYL* decreased significantly in S2, and the other *PYL* decreased in S3. Together these results were in support of active ABA signaling in the early seed development. For AUX biosynthesis (Figure 7B), the expression level of one *TAR2* was up-regulated in S3 specifically, with 2 *YUCCA* (*FtPinG0006446600.01*, *FtPinG0007888900.01*) being up-regulated since S2, and one *YUCCA* (*FtPinG0000848200.01*) being up-regulated along the seed growth. Additionally, the AUX catabolic gene, *GH3.5* (*FtPinG0003743200.01*) was down-regulated, which together indicated AUX accumulation in the grain filling seeds. Consistently, one *LAX* (*FtPinG0007643300.01*), 2 *PIN* (*FtPinG0009541700.01*, *FtPinG0009310400.01*) and one *BIG*

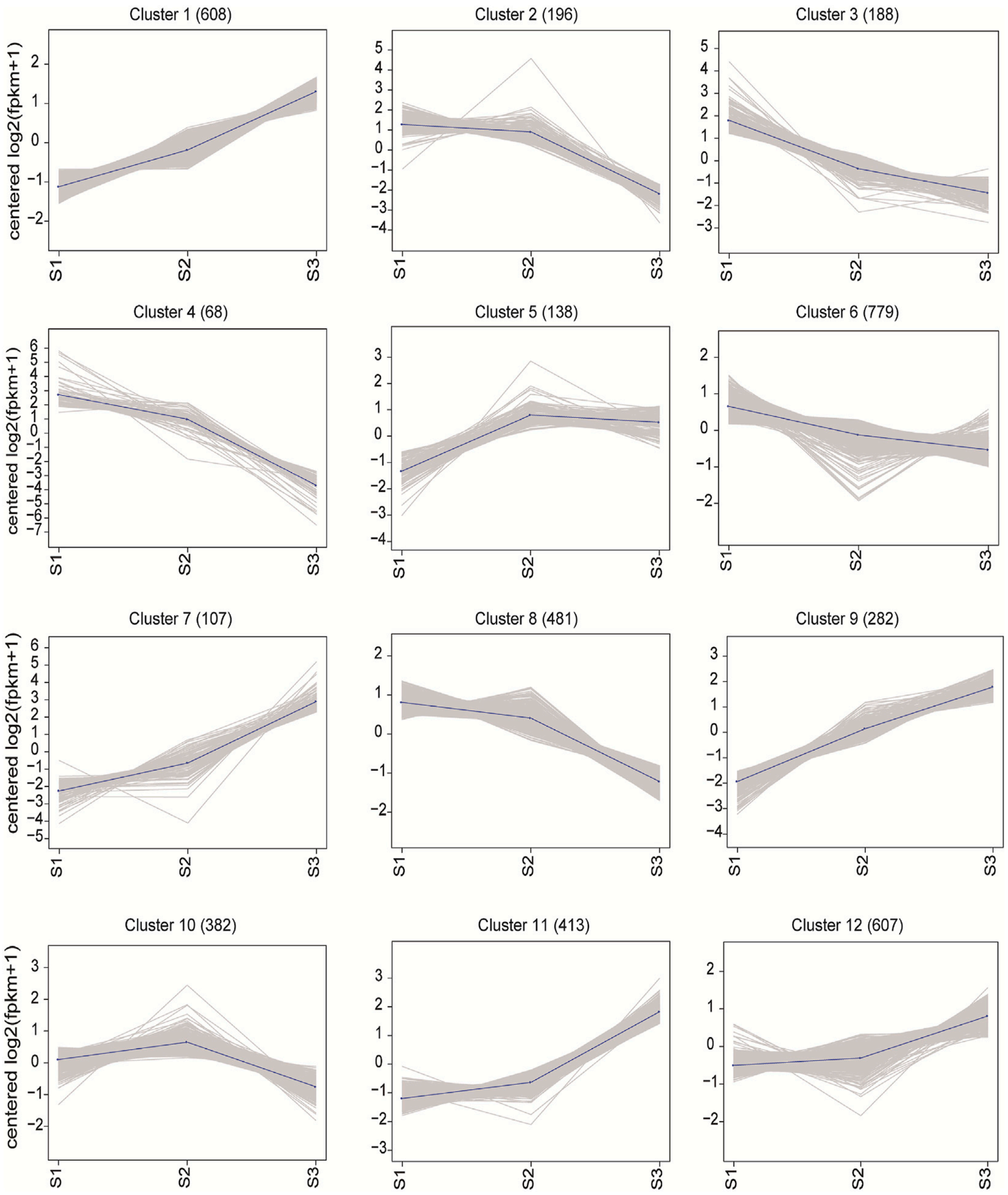


FIGURE 6
Clusters of co-expressed genes and their kinetic patterns during the seed development.

responsible for AUX transport gradually enhanced the transcription mildly. For the signal transduction, the expression of the AUX receptor-encoding gene *TIR1* (*FtPinG0004641600.01*) significantly increased with the seed growth. Together these results suggested active AUX signaling in the filling process of Tartary buckwheat seeds. Several genes in the ET biosynthesis and signaling displayed differential expression (Figure 7E), as one *SAM*, one *ACO* (*FtPinG0004699200.01*) and 2 *ACS* involved in ET biosynthesis, one *EIN2*, one *EIN3*, one *EIN4*, 2 *ETR1* and one *EBF* involved in ET signaling pathway, strengthened the transcription especially in S3 phase, yet one *ETR2* (*FtPinG0006853500.01*) and one *EIL4* (*FtPinG0008254600.01*) genes involved in signaling transduction decreased the expression significantly with the seed growth. As in Figure 7C, several genes involved in BR biosynthesis and signaling pathways were identified to be differentially regulated. One *CYP724B1* gene (*FtPinG0003995500.01*), homologous to *D11* in rice (Tanabe et al., 2005), involved in BR biosynthesis increased its transcription in maturation phase. 2 DEGs (*FtPinG0005616500.01*,

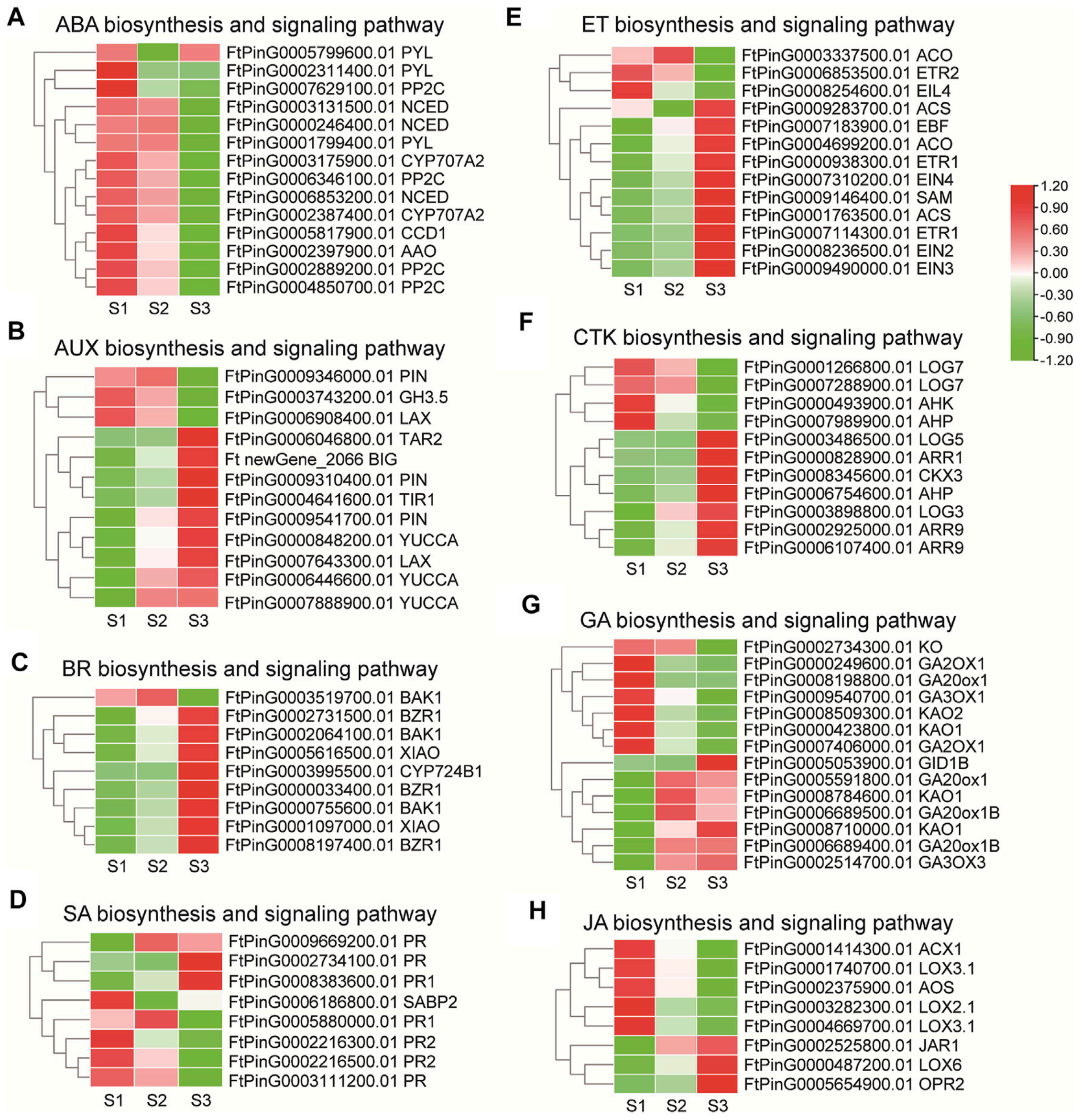


FIGURE 7
The transcriptome dynamic of genes involved phytohormone biosynthesis and signaling pathways during grain filling in Tartary buckwheat **(A)**, ABA; **(B)**, AUX; **(C)**, BR; **(D)**, SA; **(E)**, ET; **(F)**, CTK; **(G)**, GA; **(H)**, JA.

FtPinG0001097000.01) homologous to *XIAO* in rice (Jiang et al., 2012) were moderately up-regulated with the seed growth, while 2 *BAK1* (*FtPinG0002064100.01*, *FtPinG0000755600.01*) and 3 *BZR1* raised their expression significantly with seed growth, providing insight of positive roles of BR during seed filling. Genes in the CTK pathways displayed variated expression (Figure 7F). The expression of 2 *LOG* (*FtPinG0003486500.01*, *FtPinG0003898800.01*) involved in biosynthesis, and one *AHP* (*FtPinG0006754600.01*), 3 *ARR* involved in signaling pathway were up-regulated particularly in S3 phase, while 2 *LOG* (*FtPinG0001266800.01*, *FtPinG0007288900.01*), one *AHK* and one *AHP* (*FtPinG0007988900.01*) declined the expression especially in S3 phase indicating regulation by CTK in the early or late filling stages. For genes involved the GA biosynthesis (Figure 7G), one *KO* decreased the expression in maturation phase. 2 *KAO* (*FtPinG0008509300.01*, *FtPinG0000423800.01*), one *GA2OX1* (*FtPinG0008198800.01*), one *GA3OX1* (*FtPinG0009540700.01*) decreased the expression since middle filling stage. Yet the expression of 2 *KAO* (*FtPinG0008710000.01*, *FtPinG0008784600.01*), 3 *GA20ox1* (*FtPinG0005591800.01*, *FtPinG0006689400.01*, *FtPinG0006689500.01*), and one *GA3OX3* (*FtPinG0002514700.01*) were up-regulated since middle filling phase. The *GID1B* encoding the GA receptor in the signaling

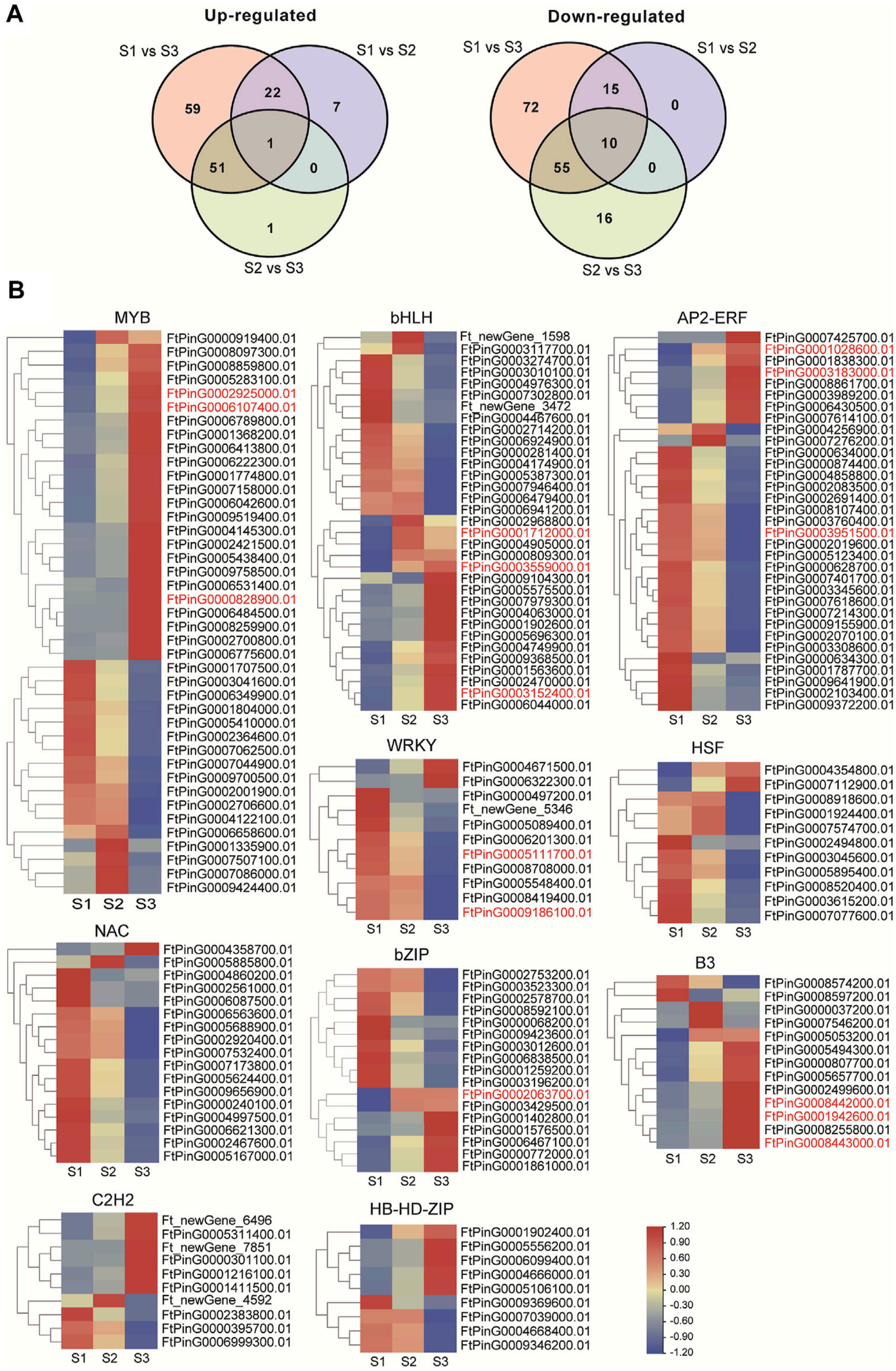


FIGURE 8
Transcriptome dynamic of transcription factors during seed development. **(A)**. Venn map showing the differentially up-regulated or down-regulated TFs. **(B)**. Heatmap illustration of top 10 family TF DEGs clustered by gene families (MYB; bHLH; AP2/ERF; NAC; bZIP; B3; HSF; WRKY; C2H2; and HD-HB-ZIP). The TFs involved in phytohormone signaling or with high similarity with characterized genes involved in seed size control in other plants were marked in red color.

transduction was moderately up-regulated in S3 (Figure 7G). These results suggested complicated GA signaling transduction during seed maturation. Among the DEGs involved in JA biosynthesis and signaling (Figure 7H), one *LOX6*, one *OPR2* genes increased the expression significantly, while one *LOX2*, 2 *LOX3*, one *AOS* and one *ACX1* gene declined the expression with the seed maturation. No specific DEG was identified to be involved in JA signaling pathway. One *SABP2* (*FtPinG0008383600.01*) and 3 *PR* (*FtPinG0002734100.01*, *FtPinG0008383600.01*, *FtPinG0009669200.01*) involved in SA signaling raised the expression with the seed filling (Figure 7D), as the transcription of *PR* (*FtPinG0002216300.01*, *FtPinG0002216500.01*, *FtPinG0003111200.01, FtPinG0005880000.01*) declined with the seed growth. No DEGs involved in SA biosynthesis was discovered. These results together pointed to less participation of JA and SA signaling pathways in seed filling.

3.5.2 Dynamic transcriptome analysis of transcription factors during seed development

Transcription factors (TFs) could participate in many aspects of cellular processes in seed development, thus the expression dynamics of total TF genes involved in seed development of Tartary buckwheat were carefully investigated. According to the RNA-seq data, a total of 1077 TFs were identified as expressed in at least one developmental stage. An overview of the transcription factors that were differentially regulated was shown in Supplementary Figure S8. In total, 309 TFs were taken as DEGs, with 7 being identified as new TF genes. Additionally, 87 DEGs encoding *RLK-Pelle* family kinases and 45 DEGs transcription repressors involved in transcription regulation showed remarkable expression variation. Briefly, the top 10 TF families with the largest numbers were MYB (42), bHLH (33), AP2/ERF (33), NAC (17), bZIP (17), B3 (13), HSF (11), WRKY (11), C2H2(10), and HD-HB-ZIP (10) (Supplementary Figure S8). Both up- and down-regulation of these TF genes occur in the process of seed maturation. Compared with S1, 23 TF genes were commonly up-regulated in S2 and S3, with 7 TF genes being up-regulated specifically to S2, and 59 TF DEGs being up-regulated specifically to S3 (Figure 8A). Additionally, compared with S1, 25 TF genes were commonly down-regulated in S2 and S3, with 72 TF genes being down-regulated specifically to S3. The expression of the top 10 family TF DEGs was displayed in Figure 8B. 9 out of 11 DEGs in WRKY and NAC families were down-regulated with seed growth, with the other 2 being up-regulated in S2 or S3 phase, which indicated WRKY and NAC DEGs may function mainly in the early stage of seed development or negatively during grain filling. Special attention was paid to the TFs involved in phytohormone signaling or with high similarity with characterized genes involved in seed size control in other plants (marked in red color). The 3 *ARF* (*FtPinG0008443000.01*, *FtPinG0001942600.01*, *FtPinG0008442000.01*) of B3 family and 4 *DOF* (*FtPinG0000702500.01*, *FtPinG0008252900.01*, *FtPinG0001221400.01*, *FtPinG0006052400.01*) of C2C2 family involved in AUX signaling pathway increased the expression either in middle filling stage or the maturation stage. Moreover, the *ABI5* (*FtPinG0002063700.01*) of bZIP family in ABA signaling and the *ARR* (*FtPinG0000828900.01*, *FtPinG0002925000.01*, *FtPinG0006107400.01*) of MYB family involved in CTK signaling displayed ascending transcription pattern with the seed growth. The *ERF* (*FtPinG0003951500.01*, *FtPinG0003183000.01*, *FtPinG0001028600.01*) of AP2/ERF family, homologous to *ANT* (Meng et al., 2015) and *AP2* (Jofuku et al., 2005) in rice, and the *bHLH* (*FtPinG0003559000.01*, *FtPinG0003152400.01*, *FtPinG0001712000.01*), homologous to *Awn-1* (Luo et al., 2013) in Arabidopsis showed remarkable variation in S3 phase in Tartary buckwheat seeds, whereas the *WRKY* (*FtPinG0009186100.01*, *FtPinG0005111700.01*), homologous to *WRKY53* (Tian et al., 2017) involved in seed size were mainly down-regulated with the seed growth.

3.5.3 Dynamic transcriptome analysis of expansin family proteins during seed development

Seed size changed distinctly with the development of growth stages in Tartary buckwheat seeds (Figure 1B), with the highest length and width at maturation phase. Plant expansin genes belong to a group of loosening proteins located in the cell wall, and was an important component for cell expansion (Cosgrove, 2015). Previously, expansin was reported to function fundamentally in seed size and yield determination in many plants (Cosgrove, 2015). To explore the roles of expansin family genes in Tartary buckwheat, the transcription dynamics during the grain filling stage were mined here. As in Figure 9A, out of the genome wide 37 expansin genes, 19 were expressed in seeds (with FPKM>1). Among them, the most abundant gene in dehulling seeds was *FtEXPA12*. *FtEXPA12*, *FtEXPA8* and *FtEXLA1* were constantly expressed. As 14 expansin genes substantially displayed an increased trend of expression along with the seed growth, only 2 genes (*FtEXPA6* and *FtEXPA26*) declined their expression notably. Genes with variated expression were analyzed subsequently. As in Figure 10A, among the 16 differentially expressed genes, 14 were classified into subfamily-A, with 2 being classified into subfamily-B (*FtEXPB4*, *FtEXPB1*). Further correlation analysis showed that most of the up-regulated expansin genes were strongly correlated with each other (Supplementary Figure S9). Among the up-regulated expansin genes, the transcription level of *FtEXPA5*, *FtEXPA11*, *FtEXPA15*, *FtEXPA21*, *FtEXPA27*, *FtEXPA28*, *FtEXPA29* and *FtEXPB4* reached the top in seed maturation stage, while *FtEXPA19* was transcribed most in middle filling stage and slightly decreased in seed maturation stage. Nevertheless, *FtEXPA5*, *FtEXPA15*, *FtEXPA21*, *FtEXPA27*, *FtEXPA28*, *FtEXPA29* and *FtEXPB4* displayed the most significant transcriptional variations along with the seed growth (Figure 9A). Interestingly, these genes were phylogenetically quite close, as *FtEXPA5*, *FtEXPA28*, *FtEXPA29* and *FtEXPA27* were classified in the same branch in the previous report (Sun et al., 2021).

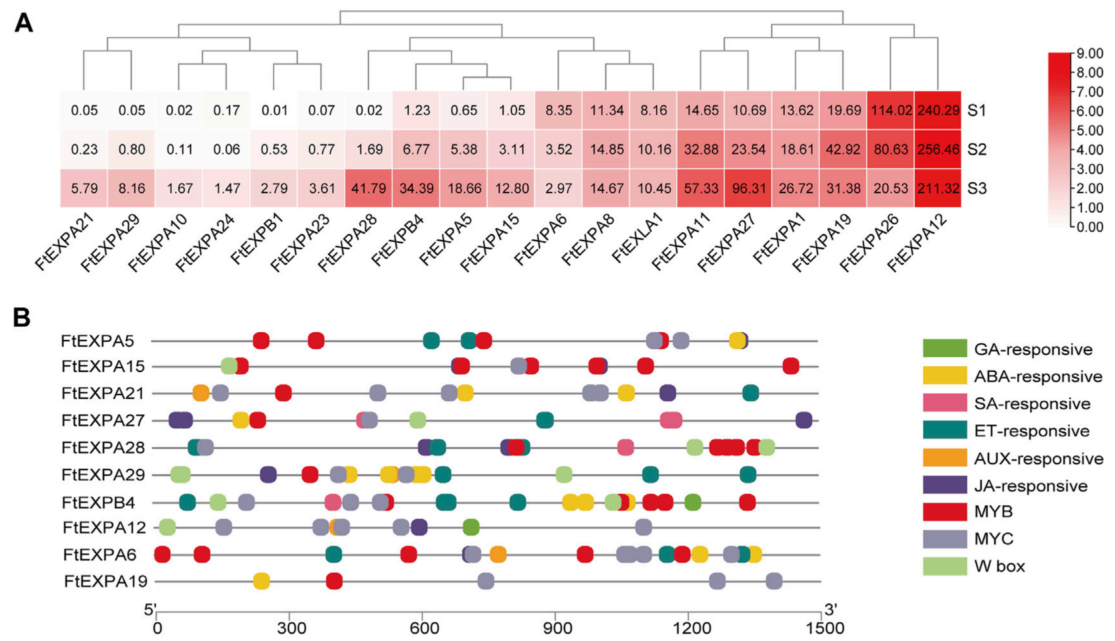


FIGURE 9
Transcriptome dynamics of expansin family proteins in the seed development. **(A)**. Heatmap representation of the transcriptomic dynamic of differentially expressed expansin genes during seed maturation. The values in the heatmap indicate the FPKM value of the DEGs. **(B)**. Illustration of seed development associated *cis*-acting elements in the promoter region of differentially expressed expansin genes.

Downstream effecting genes involved in seed development were usually regulated by upstream phytohormone signals and transcription regulators. To explore the transcriptional regulation of expansin genes in Tartary buckwheat seeds, the upstream 1.5 Kb promoter sequences of expansin genes were subsequently subjected to Plantcare for *cis*-regulatory element analysis. As shown in Figure 9B, aside from the consensus eukaryotic promoter elements, such as the TATA-box and CAAT-box, light-responsive elements and stress-responsive elements (*cis*-acting regulatory element essential for the anaerobic induction, MBS in drought-inducibility, and WUN-motif involved in wound-responsiveness) were also discovered in the promoters. To be noted, *cis*-regulatory elements, such as ERE, ABRE, TGACG-motif, CGTCA-motif, P-box and TCA-element responsive to phytohormone ET, ABA, JA, SA, GA, and AUX were notably over-represented in the promoter region of expansin DEGs. Transcription binding sites of MYB, MYC and WRKY TFs were also frequently dispersed. Among the most transcribed genes, the significant differentially expressed genes *FtEXPA5*, *FtEXPA15*, *FtEXPA27*, *FtEXPA28* and *FtEXPB4* were analyzed. Particularly, the promoter region of *FtEXPA27* was richer in SA- and JA-responsive elements, while the *FtEXPA28* promoter bared more ET- and JA-responsive elements and MYB and WRKY binding sites. In the *FtEXPA5* promoter, 4 MYB and 2 MYC binding sites and 2 ET-responsive elements were found, while 6 MYB binding sites and 2 JA-responsive elements were discovered in the *FtEXPA15* promoter region. The B-type expansin *FtEXPB4* promoter seemed to have relatively more TF binding sites than A-type expansin genes and was quite rich in ET- and ABA-responsive elements, and MYB, MYC and WRKY binding sites, with an additional GA-responsive *cis*-acting element. *FtEXPA12* had 5 MYC binding sites and one GA-responsive element in the promoter. MYC is important component of the JA signaling pathway. Thus, the above data provided clues that phytohormone ABA, ET, JA, GA and SA, and TF MYB, MYC, and WRKY probably participated in expansin gene regulation during seed development in Tartary buckwheat.

To further explore the relationship of phytohormone with expansin genes in the process of seed maturation, the DEGs in the ABA, ET, GA, AUX, JA and SA phytohormone signaling pathways were then subjected to correlation analysis with the variated expansin genes. From our research, BR was found to promote Tartary buckwheat grain filling rate (Wei, 2021), thus DEGs in BR biosynthesis and signaling pathway were also included for correlation analysis. As shown in Supplementary Figure S10, most up-regulated expansin genes, except *FtEXPA19*, showed strong positive correlations with DEGs in ET (Supplementary Figure S10A), BR (Supplementary Figure S10E) and AUX (Supplementary Figure S10F) signaling pathways. Yet the declined expansin

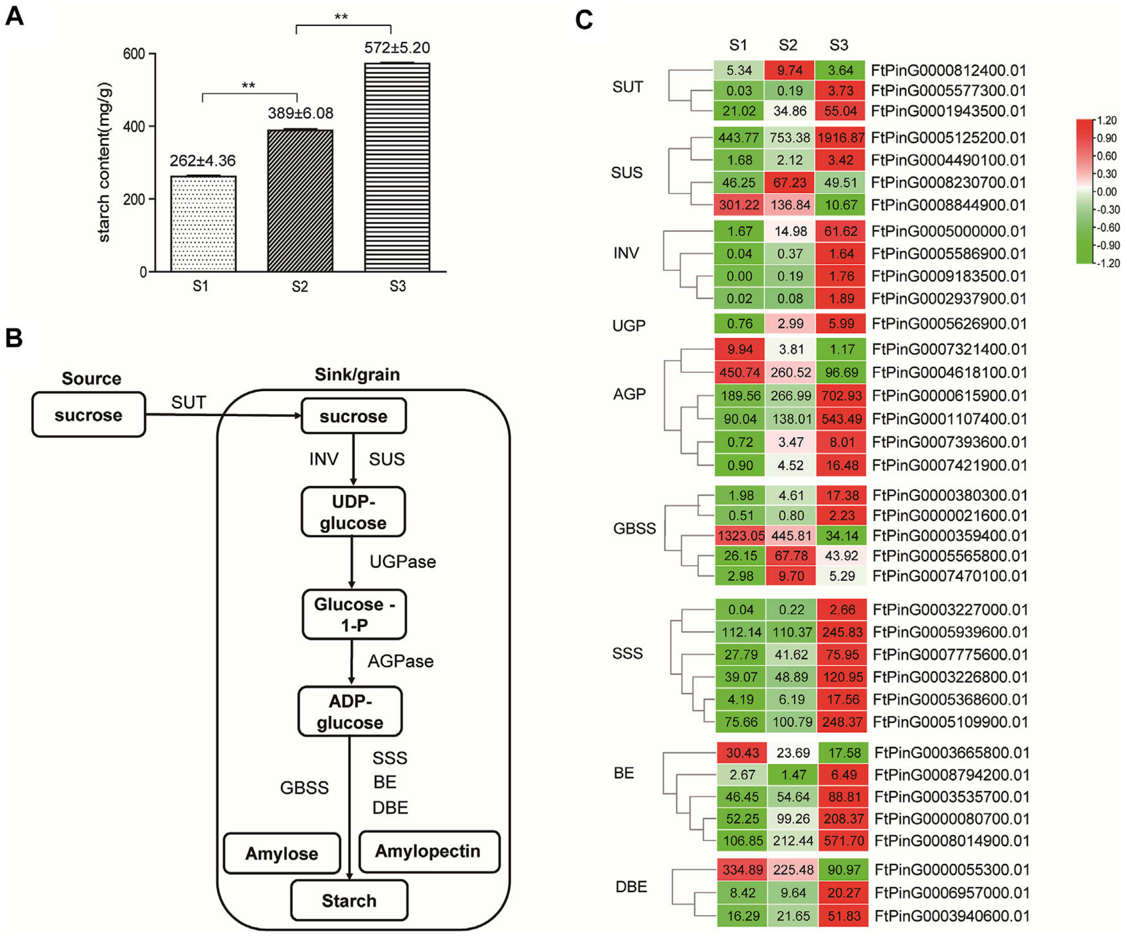


FIGURE 10
Starch synthesis in Tartary buckwheat seeds. **(A)**. The amount of total starch in Tartary buckwheat seeds. **(B)**. Schematic representation of starch biosynthesis in the seeds. **(C)**. Expression pattern of starch biosynthesis genes that were differentially expressed during seed development (The genes with FPKM≥1 in at least one stage are shown, and the original FPKM value were marked in the heatmap).

gene *FtEXPA26* was in strong negative correlation with most differentiated genes in ET (Supplementary Figure S10A), BR (Supplementary Figure S10E) and AUX (Supplementary Figure S10F) signaling pathways. Moreover, most up-regulated expansin genes, except *FtEXPA19*, displayed strong negative correlations with DEGs in ABA signaling pathways, while *FtEXPA26* was in strong positive correlation (Supplementary Figure S10B). Nevertheless, strong positive or negative correlations were only found between up-regulated expansin genes (except *FtEXPA19*) and 2 *PR* in the SA signaling pathway, and *OPR2* (*FtPinG0005654900.01*) and *LOX6* (*FtPinG0000487200.01*) in JA signaling pathway, and *GID1B* (*FtPinG0005053900.01*) and *KAO1* (*FtPinG0008710000.01*) in the GA signaling pathway. Together, these results preferred regulation of expansin family genes by phytohormone ET, ABA, AUX and BR more potentially than GA, JA and SA, in the process of grain filling. Additionally, the above TFs that were homologs of genes involved in seed size determination in other plants (Figure 9, red marked TFs) were subjected to correlation analysis with variated expansin genes. As in Supplementary Figures S11, 3 *ARF* and 2 *DOF* (*FtPinG0000702500.01*, *FtPinG0008252900.01*) involved in AUX signaling pathway, 2 *ARR* in CTK signaling pathway, one *bHLH* (*FtPinG0003152400.01*) homologous to *Awn-1* in Arabidopsis and 2 *ERF* (*FtPinG0003183000.01*, *FtPinG0001028600.01*) homologous to *ANT* and *AP2* in rice displayed strong correlation with differentially expressed expansin genes, except *FtEXPA19*, *FtEXPA6* and *FtEXPA26*. Yet, the *Awn-1* homologous *bHLH* (*FtPinG0001712000.01*) correlated strongly with *FtEXPA19*. These results suggested possible positive transcriptional regulation of *expansin* by these TFs during Tartary buckwheat grain filling. Nevertheless, the 2 *WRKY*, homologous to *WRKY53*, and the *ERF* (*FtPinG0003951500.01*) involved in seed size represented strong negative correlation with most up-regulated *expansin*, indicating negative regulatory roles.

3.5.4 Analysis of starch biosynthesis gene expression in seed development

With the remarkable change of seed shape along with the development, both the fresh and dry weight of grains companionly elevated significantly ($p < 0.05$) (Figure 1C), reaching the top at stage 3. Starch was the major form of carbohydrates accumulated at mature seeds and the main nutrient that make the seeds and other storage organs expand and enlarge (Saripalli and Gupta, 2015). Therefore, the amount of starch largely determined the final grain yield in plants. The level of total starch in the grain filling stage seeds of Tartary buckwheat was subsequently measured. As shown in Figure 10A, the starch amount indeed increased remarkably along with the process of seed maturation from 262 ± 4.36 mg/g to 572 ± 5.20 mg/g, in accordance with severe change of seeds both in width and dry weight (Figure 1). According to above KEGG analysis, "Starch and sucrose metabolism" were dominantly enriched in significantly variated co-expression clusters (Cluster 2, 5, 9) (Supplementary Figures S4, S5, S7). So, genes involved in starch and sucrose metabolism were proposed to significantly affect the grain filling process and final yield in Tartary buckwheat. Thus, the transcriptional profile of genes involved in starch and sucrose metabolism were thoroughly explored. Among them, 61 genes were significantly differentially expressed, with 34 belonging to glycosyl hydrolases family (Supplementary Figure S12). Among the 7 DEGs encoding glucan endo-1,3-β-glucosidase, the expression of *FtPinG0000387400.01* increased only in middle filling stage and *FtPinG0002200200.01* specifically decreased in late filling stage, as the expression of the other 5 DEGs gradually increased with the seed growth. 10 β-glucosidase-encoding genes were identified, with 5 being continuously up-regulated with grain filling, 2 being significantly down-regulated in S3 phase, and 3 being down-regulated with the seed growth. Both 2 DEGs encoding α-glucosidase increased their expression in S3 phase. Notably, 2 out of the 4 genes encoding pectinesterase identified were remarkable down-regulated in S3. 3 DEGs encoding amylase displayed notable decline along with the seed filling, indicating low level of starch hydrolysis. Moreover, 4 genes encoding trehalose-phosphate phosphatase displayed significant transcriptional dynamic. Several genes involved in starch biosynthesis were identified noticeably and were discussed below.

In plants, sucrose was transported from source, such as leaves, to the sink for the synthesis of starch by sucrose transport protein (SUT) (Figure 10B). In our transcriptome data, 3 *SUT* displayed differential expression pattern (Figure 10C), with *FtPinG0001943500.01* being mostly transcribed and continuously up-regulated moderately with seed growth. Starch biosynthesis in seeds initiated with the transition of sucrose to UDG-glucose by sucrose synthase (SUS) and invertase (INV). As in Figure 10C, among the 4 differentially expressed sucrose synthase genes, *FtPinG0005125200.01* represented the highest expression level and continuously upward trend along with the seed maturation, as *FtPinG0008230700.01* were up-regulated in S2 phase only and *FtPinG0008844900.01* was down-regulated with the seed growth. 4 *INV* increased the expression significantly, as *FtPinG0005000000.01* and *FtPinG0008078700.01* were insoluble cell wall invertases, and the other 2 were soluble vacuolar invertases. Among them, *FtPinG0005000000.01* displayed the highest expression level and an increased pattern with the seed maturation. Only one *UGP* (*FtPinG0005626900.01*) was differentially expressed in seeds and showed an up-regulated tendency with the grain filling. The formation of ADP-glucose from glucose-1-phosphate by AGP, is considered as the committed rate-limiting step of starch biosynthesis (Saripalli and Gupta, 2015). 6 *AGP* displayed differential expression along with the seed growth, with 4 being gradually up-regulated and 2 being gradually down-regulated. Among them, *FtPinG0000615900.01* and *FtPinG0001107400.01* were transcribed at the highest level in mature seeds, as *FtPinG00004618100.01* displayed the highest expression in the early phase. 5 *GBSS* were differentially expressed in filling stage seeds, with 2 (*FtPinG0005565800.01*, *FtPinG0007470100.01*) being transcribed most at middle filling stage, and 2 (*FtPinG0000380300.01*, *FtPinG0000021600.01*) gradually increased the expression until the seed maturation. The *GBSS* (*FtPinG0000359400.01*) was the exception, which showed the highest expression out of the 5 in early filling stage. 6 *SSS* displayed differentially expression, and all reached the top expression level at maturation phase. 4 out 5 differentiated *BE* genes increased their expression with the seed growth, yet *FtPinG0000080700.01* and *FtPinG0008014900.01* were most transcribed in mature seeds. *FtPinG0006957000.01* and *FtPinG0003940600.01* in the *DBE* DEGs reached the highest expression levels in maturation phase. These above results indicated that up-regulated genes involved in starch biosynthesis may participate in the starch accumulation process during Tartary buckwheat grain filling, while down-regulated genes may function in the early stage of seed development.

To clarify the regulation of structural genes involved in starch biosynthesis, *cis*-regulatory element analysis of the 1.5 Kb promoter region was also carried out by Plantcare. As shown in Supplementary Figures S13, S14, similar with expansin genes, aside from the consensus eukaryotic promoter elements, such as the TATA-box and CAAT-box, light-responsive elements, and stress-responsive elements (*cis*-acting regulatory element essential for the anaerobic induction, MBS in drought-inducibility, and WUN-motif involved in wound-responsiveness), hormone-responsive *cis*-regulatory elements, such as ERE, ABRE, TGACG-motif, CGTCA-motif and TCA-element responsive to phytohormone ET, ABA, JA, SA, GA, and

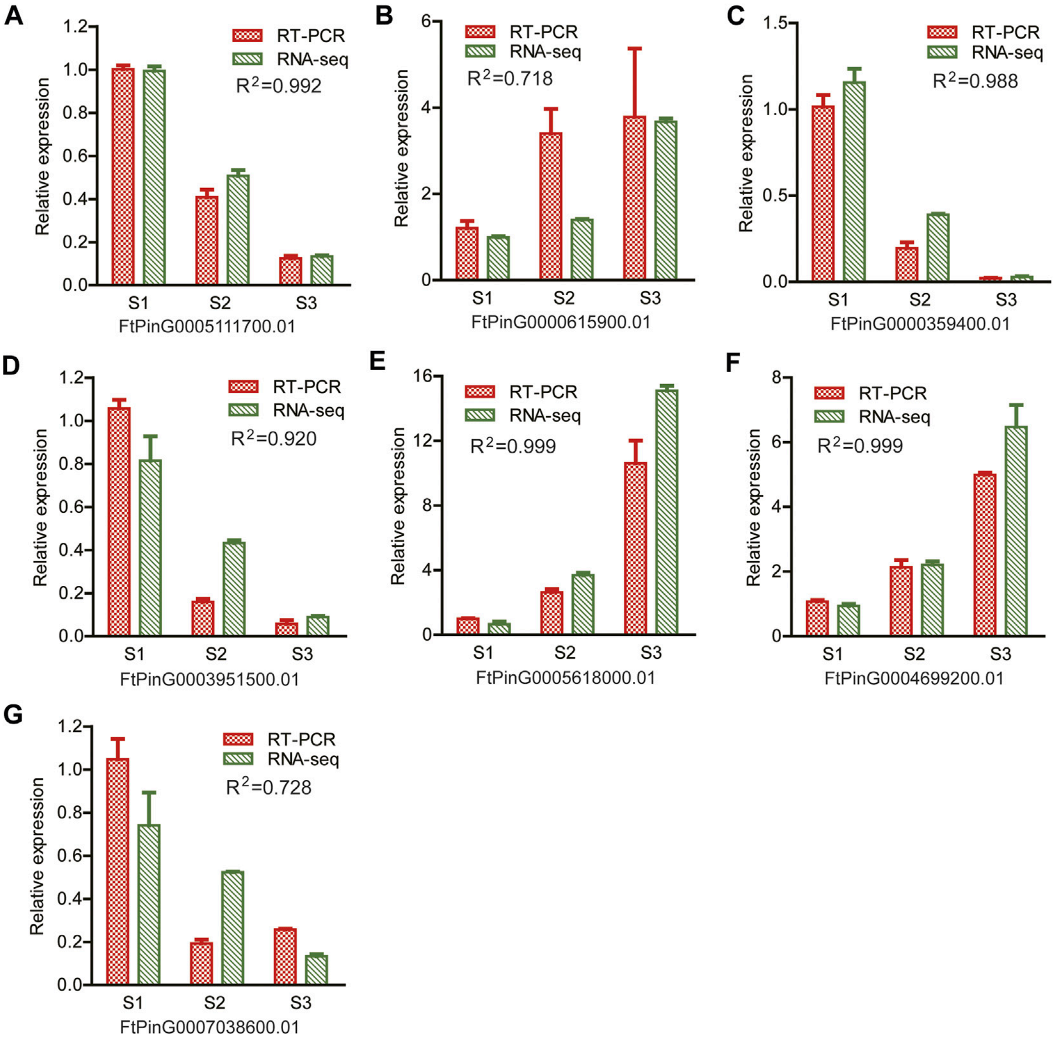


FIGURE 11
Illustration of the relative expression of 7 randomly selected genes for RT-PCR verification. R^2 value indicated the correlation coefficient of the relative expression levels between the RT-PCR results and the RNA-seq by Pearson correlation analysis **(A)** FtPinG0005111700.01; **(B)** FtPinG0000615900.01; **(C)** FtPinG0000359400.01; **(D)** FtPinG0003951500.01; **(E)** FtPinG0005618000.01; **(F)** FtPinG0004699200.01; **(G)** FtPinG0007038600.01.

AUX, and transcription binding sites of MYB, MYC and WRKY TFs were also over-represented. These results indicated the candidate regulation of starch biosynthesis by these phytohormone and TFs.

3.5.5 Verification of RNA-Seq gene expression by RT-PCR

A quantitative RT-PCR was performed to verify the reliability of the RNA-seq data. Seven DEGs were randomly selected, as displayed in Figure 11. Among them, *ACO* (*FtPinG0004699200.01*), *AGP* (*FtPinG0000615900.01*), *ERF* (*FtPinG0003951500.01*) and *WRKY* (*FtPinG0005111700.01*) were included as representatives. The expression of these DEGs were consistent with the results of the RNA-seq data (R^2 > 0.7), confirming the reproducibility of the transcriptome data.

4 Discussion

Various nutritional and pharmacological effects of Tartary buckwheat have been extensively studied (Li and Zhang, 2001; Huda et al., 2021; Zhang et al., 2021). Nevertheless, poor post-embryonic grain filling in Tartary buckwheat largely halts the grain yield improvement. Few reports were about the gene regulatory network governing the physiological changes during the filling stage of Tartary buckwheat seeds. Thus, in this study we tried to sort out the key seed development genes based on our transcriptome data from different developmental stages of the dehulling filling stage seeds of Tartary buckwheat (*F. tataricum* cv. Xiqiao No.1). Sharp transitions in gene expression occurred along with the grain filling (Figures 2–4), consistent with significant physiological changes with seed growth in Tartary buckwheat (Figure 1). 4249 DEGs (with FC > 4) were identified totally that may be key genes related to seed development.

Phytohormones play indispensable roles in seed development in rice, maize and Arabidopisis (Santner et al., 2009). AUX and ABA function notably in controlling the embryogenesis pattern and promoting the accumulation of storage products during the subsequent filling stage (Schussler et al., 1984; Moller and Weijers, 2009; Chandrasekaran et al., 2014). Liu et al. (2018a) found the ascending tendency of ABA in the process of Tartary buckwheat fruit maturation and the variation of GA and AUX. Here, significant variation of genes in the signal transduction pathways of phytohormones was observed (Figure 7). Yet genes involved in ABA biosynthesis and signaling displayed a descending pattern of expression, indicating active ABA signaling in the early seed development (Figure 7A). Increase of AUX biosynthesis genes (*TAR* and *YUCCA*) and decrease of catabolic genes (*GH3.5*) (Figure 7A) suggested AUX accumulation during grain filling. The AUX transport genes (*LAX*, *PIN* and *BIG*) strengthened the expression in Tartary buckwheat kernel. These results strongly implied positive regulatory roles of AUX along with grain filling process in Tartary buckwheat. As the balance of AUX and ABA was proposed to be key factors regulating the cell division rate in the early seed development of buckwheat (Liu et al., 2018b), and evidence indicated synergistic regulation of cell expansion *via* both AUX and GAs (Fenn and Giovannoni, 2021), the nexus of ABA, AUX, GA and other hormones during grain filling required further elucidation. Yet variated expression of genes in the GA and CTK biosynthesis and signalling pathways suggested complicated roles by these phytohormones in seed filling process. Overexpression of OsBZR1, a BR-signaling TF, resulted in higher grain yield in rice (Zhu et al., 2015). Here, the BR biosynthesis related *CYP724B1*, *BZR1* and coreceptor BAK1 homologs were up-regulated gradually with the seed growth. As we previously found, appropriate spraying of BR on the leaves of Tartary buckwheat could significantly improve the seed setting rate, grain filling rate, and yield, and reduce the abortion rate of grains notably (Wei, 2021). Thus, it is speculated that BR signaling play positive roles in grain filling. Ethylene can promote fruit ripening, and ethylene signaling plays an important role in fruit development (Persak and Pitzschke, 2014; Fenn and Giovannoni, 2021). Here, genes involved in ET biosynthesis (*SAM*, *ACO*, *ACS*), and signaling (*EIN2*, *EIN3*, *EIN4*, *ETR1*, *EBF*) enhanced the expression with the seed growth (Figure 7E), indicating consistent positive function of ethylene in filling stage Tartary buckwheat.

Transcriptional regulators were important factors controlling seed size in plants, including TF, transcriptional coactivators, and regulators involved in chromatin modification. The significant enrichment of up-regulated genes in DNA binding, nucleosome assembly and cell proliferation in Cluster 7 and 9 by GO analysis indicated remarkable change of chromatin structure and dynamics along with the seed maturation in Tartary buckwheat, in which transcriptional regulators could be involved. In the maturation process of Tartary buckwheat seeds, 309 TF DEGs were identified with MYB, bHLH, AP2/ERF, NAC, bZIP, B3, HSF, WRKY, C2H2, and HD-HB-ZIP as the top 10 families. MYB, NAC, WRKY, bHLH, MADS and AP2/ERF TFs were reported to regulate of fruit development/maturation (Jakoby et al., 2002; Persak and Pitzschke, 2014; Zinsmeister et al., 2016; Zhang et al., 2018; Liu et al., 2019a; Liu et al., 2019b; Liu et al., 2019c; Ma et al., 2019; Liu et al., 2020). In Tartary buckwheat, several MYB (FtMYB6, FtMYB116, FtMYB3, et al.) were reported to be positively or negatively involved in flavonoid biosynthesis (Yao et al., 2020; Wang et al., 2022), yet no specific TF gene has been characterized to participate in grain filling process directly. Since CTKs are the key drivers of seed yield (Jameson and Song, 2016), the up-regulated *ARR*s involved in CTK signaling could be candidate MYB genes involved in grain filling in Tartary buckwheat. Similarly, with the importance of AUX and ABA in seed maturation in Tartary buckwheat (Liu et al., 2018a), the *ARF* and *DOF*, *ABI5* DEGs involved in AUX and ABA signaling pathways could be targets to improve grain filling. In wheat developing endosperm, NAC019-A1 was a negative regulator of starch synthesis (Liu et al., 2020). Therefore, the decreased NAC DEGs (Figure 8B) were worth for further characterization of whether regulating starch biosynthesis or not. In addition, distinct variation of TF genes homologous to *ANT* and *AP2* in rice, and *Awn-1* in Arabidopsis controlling seed size (Figure 8) were supposed to regulate seed development process either.

Cell walls provide essential plasticity for plant cell division and defense, which are often conferred by the expansin superfamily with cell wall-loosening functions. Recently, expansin family genes were reported to participate in many aspects of plant growth and development processes, such as root hair growth, germination, leaf growth, and grain yield in different plants (Cosgrove, 2015; Calderini et al., 2021). Cotton plants overexpressing GhRDL1 and GhEXPA1 proteins produced strikingly more fruits, larger seed size and doubled seed mass (Xu et al., 2013). Transgenic over-expression of sweet potato expansin gene (*IbEXP1*) in Arabidopsis produced larger seeds, accumulated more protein and starch in each seed, and produced more inflorescence stems and siliques than control plants (Bae et al., 2014; Chen et al., 2016). Targeted expression of *TaExpA6* in the young seed lead to a significant increase in grain size without a negative effect on grain number, and a final yield boost by 10% in wheat under field conditions (Calderini et al., 2021), which provided an opportunity to overcome a common bottleneck to yield improvement across many crops (Cosgrove, 2021). During grain filling of Tartary buckwheat seeds, significant transcriptional variation of expansin genes were discovered during grain filling. Strong correlation of the up-regulated expansin genes indicated the possible common regulation. Phytohormone exerts variated effects on *EXP* expression. Repressive effect by AUX on the expression of *FaEXP2* and *FaEXP5* in Chilean strawberry fruit (Figueroa et al., 2009) and by GA on *CDK-Exp3* transcription in persimmon, and positive regulation of *MiExpA1* expression by ET within a short time in mango (Sane et al., 2005), during fruit softening were reported. GA mediated expansion of floral organs *via* expansins prior to anthesis (Azeez et al., 2010). Transgenic overexpressed wheat *TaEXPB23* in tobacoo, involved in the abiotic stress response, were upregulated by exogenous JA and salt stress, but downregulated by exogenous GA, ET, IAA and alpha-naphthlcetic acid (NAA) (Han et al., 2012). A regulatory module controlling GA-mediated endosperm cell expansion involving spatiotemporal control of the cell expansion gene *AtEXPA2* is critical for seed germination in Arabidopsis (Sanchez-Montesino et al., 2019). In this study, phytohormone (ABA, ET, AUX, JA, SA and GA) responsive elements and MYB, MYC and WRKY binding sites were found in the promoters of the remarkable increased expansin gene. Strong correlation of these expansin DEGs with BR, JA, AUX and ET signaling (Supplementary Figure S10) and the AUX and CTK related TFs (ARF, DOF, ARR) or the seed size related TFs (bHLH, ERF and WRKY) (Supplementary Figure S11) was discovered. Combining these results, we proposed that phytohormone AUX, ABA, ET and BR and their responsive TFs were candidate important regulators of grain filling through targeting expansin genes in Tartary buckwheat. The functions of significantly variated expansin genes were worth for further study which could be candidate targets to improve crop yield in Tartary buckwheat.

As shown in Figure 1, the length and width of seed coat reached a maximum size in S3 phase, which sets an upper limit to final size of a grain (Figure 1A). However, the size of the decorticated seeds continuously expanded until the maturation phase in Tartary buckwheat (Figure 1A, lower picture). As the primary nutrient of buckwheat fruit, starch accounts for 70% of the total substance content (De Bock et al., 2021). Thus, the amount of starch in the filling seeds largely determined the final yield. Consequently, genes involved in starch and sucrose metabolism were proposed to noticeably affect the grain filling process and final yield in Tartary buckwheat. Down-regulation of amylase genes during seed filling depicted low starch hydrolysis, and up-regulation of *SUT* either in the middle filling seeds or the early maturation seeds, supported efficient sucrose translocation from the source to the sink (seed) for starch synthesis during grain filling in Tartary buckwheat (Figure 10). Indeed, remarkable starch accumulation was found during seed development (Figure 10A). In rice and Arabidopisis, genes in the starch biosynthesis pathway have been reported (Saripalli and Gupta, 2015). However, the buckwheat starch biosynthesis genes remained largely uncharacterized, except for *GBSS* (Wang et al., 2014). Through mining the transcriptome data (Cao et al., 2022), 37 candidate genes covering all steps of starch biosynthesis displayed differential expression (Figure 10C). Up-regulation of biosynthetic genes was found in each step with seed maturation. The abundantly transcribed and enhanced genes in the filling stage, such as *FtPinG0005125200.01* of *SUS*, *FtPinG0005000000.01* of *INV*, *FtPinG0000615900.01* and *FtPinG0001107400.01* of *AGP*, *FtPinG0005565800.01*, *FtPinG0007470100.01* of *GBSS*, *FtPinG0005109900.01*, *FtPinG0005939600.01* and *FtPinG0003226800.01* of *SSS*, and *FtPinG0000080700.01* and *FtPinG0008014900.01* of *BE*, are worth for further study to elucidate the precise chemical mechanisms of these enzymes in starch synthesis in Tartary buckwheat. In view of the increased trend of starch biosynthesis genes until the maturation phase, it is speculated that without the size limit by the seed coat, starch biosynthesis would possibly continue to expand the seed kernel and raise the final grain yield. Thus, procedures to improve the seed coat size, such as by ectopic expression of specific expansin genes (Cosgrove, 2021), may be effective breeding strategy in Tartary buckwheat.

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: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA857365/.

Author contributions

Conceptualization, DX, LJ, and ZL; methodology, LJ, CL, QW, and ZL; software, LJ, QL, CL, and XY; validation, LJ and YF; resources, DX, YF, LZ; data curation, XY, YW, LJ, and DX; writing—original draft preparation, LJ and CL; writing—review and editing, LJ, DX, ZL, QW, CL, YW, YS, and LZ. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the National Natural Science Foundation of China (No. 31771716 and 62001090), the Earmarked Fund for China Agriculture Research System (No. CARS-07-B-1), the Sichuan Science and Technology Program (No. 2022NSFSC1706 and 2022NSFSC1725), the Scientific and Technological Innovation Research and Development Project of Chengdu Science and Technology Bureau (No. 2019-YF05-02421-SN), the National Key R&D Program of China (No. 2019YFD1001302 and 2019YFD1001300), and the Fundamental Research Funds for the Central Universities of Sichuan University (No.YJ2021104).

Acknowledgments

We appreciated the help of Rui Li (College of Food and Biological Engineering, Chengdu University) for his excellent technical assistance and Huiling Yan (College of Food and Biological Engineering, Chengdu University) for her support of the experimental materials.

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/fgene.2022.990412/full#supplementary-material

References

Azeez, A., Sane, A. P., Tripathi, S. K., Bhatnagar, D., and Nath, P. (2010). The gladiolus GgEXPA1 is a GA-responsive alpha-expansin gene expressed ubiquitously during expansion of all floral tissues and leaves but repressed during organ senescence. *Postharvest Biol. Technol.* 58, 48–56. doi:10.1016/j.postharvbio.2010.05.006

Bae, J. M., Kwak, M. S., Noh, S. A., Oh, M. J., Kim, Y. S., and Shin, J. S. (2014). Overexpression of sweetpotato expansin cDNA (IbEXP1) increases seed yield in Arabidopsis. *Transgenic Res.* 23, 657–667. doi:10.1007/s11248-014-9804-1

Basnet, R. K., Moreno-pachon, N., Lin, K., Bucher, J., Visser, R. G., Maliepaard, C., et al. (2013). Genome-wide analysis of coordinated transcript abundance during seed development in different *Brassica rapa* morphotypes. *BMC Genomics* 14, 840. doi:10.1186/1471-2164-14-840

Bian, J., Deng, P., Zhan, H., Wu, X., Nishantha, M., Yan, Z., et al. (2019). Transcriptional dynamics of grain development in barley (*Hordeum vulgare* L.). *Int. J. Mol. Sci.* 20, E962. doi:10.3390/ijms20040962

Calderini, D. F., Castillo, F. M., Arenas-m, A., Molero, G., Reynolds, M. P., Craze, M., et al. (2021). Overcoming the trade-off between grain weight and number in wheat by the ectopic expression of expansin in developing seeds leads to increased yield potential. *New Phytol.* 230, 629–640. doi:10.1111/nph.17048

Cantu, D., Pearce, S. P., Distelfeld, A., Christiansen, M. W., Uauy, C., Akhunov, E., et al. (2011). Effect of the down-regulation of the high Grain Protein Content (GPC) genes on the wheat transcriptome during monocarpic senescence. *BMC Genomics* 12, 492. doi:10.1186/1471-2164-12-492

Cao, C., Wang, J., Kwok, D., Cui, F., Zhang, Z., Zhao, D., et al. (2022). webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. *Nucleic Acids Res.* 50, D1123–D1130. doi:10.1093/nar/gkab957

Chandrasekaran, U., Xu, W., and Liu, A. (2014). Transcriptome profiling identifies ABA mediated regulatory changes towards storage filling in developing seeds of castor bean (*Ricinus communis* L.). *Cell Biosci.* 4, 33. doi:10.1186/2045-3701-4-33

Chen, J., Zeng, B., Zhang, M., Xie, S. J., Wang, G. K., Hauck, A., et al. (2014). Dynamic transcriptome landscape of maize embryo and endosperm development. *Plant Physiol.* 166, 252–264. doi:10.1104/pp.114.240689

Chen, Y., Han, Y., Zhang, M., Zhou, S., Kong, X., and Wang, W. (2016). Overexpression of the wheat expansin gene TaEXPA2 improved seed production and drought tolerance in transgenic tobacco plants. *PLoS One* 11, e0153494. doi:10.1371/journal.pone.0153494

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi:10.1016/j.molp.2020.06.009

Cosgrove, D. J. (2015). Plant expansins: Diversity and interactions with plant cell walls. *Curr. Opin. Plant Biol.* 25, 162–172. doi:10.1016/j.pbi.2015.05.014

Cosgrove, D. J. (2021). Expanding wheat yields with expansin. *New Phytol.* 230, 403–405. doi:10.1111/nph.17245

De Bock, P., Daelemans, L., Selis, L., Raes, K., Vermeir, P., Eeckhout, M., et al. (2021). Comparison of the chemical and technological characteristics of wholemeal flours obtained from amaranth (*Amaranthus* sp.), quinoa (*Chenopodium quinoa*) and buckwheat (*fagopyrum* sp.) seeds. *Foods* 10, 651. doi:10.3390/foods10030651

Du, J., Wang, S. D., He, C. M., Zhou, B., Ruan, Y. L., and Shou, H. X. (2017). Identification of regulatory networks and hub genes controlling soybean seed set and size using RNA sequencing analysis. *J. Exp. Bot.* 68, 1955–1972. doi:10.1093/jxb/erw460

Fenn, M. A., and Giovannoni, J. J. (2021). Phytohormones in fruit development and maturation. *Plant J.* 105, 446–458. doi:10.1111/tpj.15112

Figueroa, C. R., Pimentel, P., Dotto, M. C., Civello, P. M., Martinez, G. A., Herrera, R., et al. (2009). Expression of five expansin genes during softening of *Fragaria chiloensis* fruit: effect of auxin treatment. *Postharvest Biol. Technol.* 53, 51–57. doi:10.1016/j.postharvbio.2009.02.005

Gao, J., Wang, T. T., Liu, M. X., Liu, J., and Zhang, Z. W. (2017). Transcriptome analysis of filling stage seeds among three buckwheat species with emphasis on rutin accumulation. *PLoS One* 12, e0189672. doi:10.1371/journal.pone.0189672

Garg, R., Singh, V. K., Rajkumar, M. S., Kumar, V., and Jain, M. (2017). Global transcriptome and coexpression network analyses reveal cultivar-specific molecular signatures associated with seed development and seed size/weight determination in chickpea. *Plant J.* 91, 1088–1107. doi:10.1111/tpj.13621

Han, Y. Y., Li, A. X., Li, F., Zhao, M. R., and Wang, W. (2012). Characterization of a wheat (*Triticum aestivum* L.) expansin gene, TaEXPB23, involved in the abiotic stress response and phytohormone regulation. *Plant Physiol. biochem.* 54, 49–58. doi:10.1016/j.plaphy.2012.02.007

Huang, J., Deng, J., Shi, T. X., Chen, Q. J., Liang, C. G., Meng, Z. Y., et al. (2017). Global transcriptome analysis and identification of genes involved in nutrients accumulation during seed development of rice tartary buckwheat (*Fagopyrum Tararicum*). *Sci. Rep.* 7, 11792. doi:10.1038/s41598-017-11929-z

Huda, M. N., Lu, S., Jahan, T., Ding, M., Jha, R., Zhang, K., et al. (2021). Treasure from garden: bioactive compounds of buckwheat. *Food Chem.* 335, 127653. doi:10.1016/j.foodchem.2020.127653

Jakoby, M., Weisshaar, B., Droge-Laser, W., Vicente-Carbajosa, J., Tiedemann, J., Kroj, T., et al. (2002). bZIP transcription factors in Arabidopsis. *Trends Plant Sci.* 7, 106–111. doi:10.1016/s1360-1385(01)02223-3

Jameson, P. E., and Song, J. (2016). Cytokinin: a key driver of seed yield. *J. Exp. Bot.* 67, 593–606. doi:10.1093/jxb/erv461

Jiang, Y. H., Bao, L., Jeong, S. Y., Kim, S. K., Xu, C. G., Li, X. H., et al. (2012). XIAO is involved in the control of organ size by contributing to the regulation of signaling and homeostasis of brassinosteroids and cell cycling in rice. *Plant J.* 70, 398–408. doi:10.1111/j.1365-313X.2011.04877.x

Jofuku, K. D., Omidyar, P. K., Gee, Z., and Okamuro, J. K. (2005). Control of seed mass and seed yield by the floral homeotic gene APETALA2. *Proc. Natl. Acad. Sci. U. S. A.* 102, 3117–3122. doi:10.1073/pnas.0409893102

Jones, S. I., and Vodkin, L. O. (2013). Using RNA-seq to profile soybean seed development from fertilization to maturity. *PLoS One* 8, e59270. doi:10.1371/journal.pone.0059270

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480–D484. doi:10.1093/nar/gkm882

Kong, L., Guo, H., and Sun, M. (2015). Signal transduction during wheat grain development. *Planta* 241, 789–801. doi:10.1007/s00425-015-2260-1

Kozaki, A., and Aoyanagi, T. (2022). Molecular aspects of seed development controlled by gibberellins and abscisic acids. *Int. J. Mol. Sci.* 23, 1876. doi:10.3390/ijms23031876

Lescot, M., Dehais, P., Thijs, G., Marchal, K., Moreau, Y., Van De Peer, Y., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Res.* 30, 325–327. doi:10.1093/nar/30.1.325

Li, S. Q., and Zhang, Q. H. (2001). Advances in the development of functional foods from buckwheat. *Crit. Rev. Food Sci. Nutr.* 41, 451–464. doi:10.1080/20014091091887

Li, H. Z., Gao, X., Li, X. Y., Chen, Q. J., Dong, J., and Zhao, W. C. (2013). Evaluation of assembly strategies using RNA-seq data associated with grain development of wheat (*Triticum aestivum* L.). *PLoS One* 8, e83530. doi:10.1371/journal.pone.0083530

Li, G. S., Wang, D. F., Yang, R. L., Logan, K., Chen, H., Zhang, S. S., et al. (2014). Temporal patterns of gene expression in developing maize endosperm identified through transcriptome sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 111, 7582–7587. doi:10.1073/pnas.1406383111

Li, H. Y., Lv, Q. Y., Ma, C., Qu, J. T., Cai, F., Deng, J., et al. (2019a). Metabolite profiling and transcriptome analyses provide insights into the flavonoid biosynthesis in the developing seed of tartary buckwheat (*Fagopyrum tataricum*). *J. Agric. Food Chem.* 67, 11262–11276. doi:10.1021/acs.jafc.9b03135

Li, N., Xu, R., and Li, Y. H. (2019b). Molecular networks of seed size control in plants. *Annu. Rev. Plant Biol.* 70 (70), 435–463. doi:10.1146/annurev-arplant-050718-095851

Liu, M., Ma, Z., Zheng, T., Sun, W., Zhang, Y., Jin, W., et al. (2018a). Insights into the correlation between Physiological changes in and seed development of Tartary buckwheat (*Fagopyrum tataricum* Gaertn.). *BMC Genomics* 19, 648. doi:10.1186/s12864-018-5036-8

Liu, M., Ma, Z., Zheng, T., Wang, J., Huang, L., Sun, W., et al. (2018b). The potential role of auxin and abscisic acid balance and FtARF2 in the final size determination of Tartary buckwheat fruit. *Int. J. Mol. Sci.* 19, E2755. doi:10.3390/ijms19092755

Liu, M., Fu, Q., Ma, Z., Sun, W., Huang, L., Wu, Q., et al. (2019a). Genome-wide investigation of the MADS gene family and dehulling genes in tartary buckwheat (*Fagopyrum tataricum*). *Planta* 249, 1301–1318. doi:10.1007/s00425-019-03089-3

Liu, M., Ma, Z., Sun, W., Huang, L., Wu, Q., Tang, Z., et al. (2019b). Genome-wide analysis of the NAC transcription factor family in Tartary buckwheat (*Fagopyrum tataricum*). *BMC Genomics* 20, 113. doi:10.1186/s12864-019-5500-0

Liu, M., Sun, W., Ma, Z., Zheng, T., Huang, L., Wu, Q., et al. (2019c). Genome-wide investigation of the AP2/ERF gene family in Tartary buckwheat (*Fagopyum Tataricum*). *BMC Plant Biol.* 19, 84. doi:10.1186/s12870-019-1681-6

Liu, Y., Hou, J., Wang, X., Li, T., Majeed, U., Hao, C., et al. (2020). The NAC transcription factor NAC019-A1 is a negative regulator of starch synthesis in wheat developing endosperm. *J. Exp. Bot.* 71, 5794–5807. doi:10.1093/jxb/eraa333

Lu, X., Li, Q. T., Xiong, Q., Li, W., Bi, Y. D., Lai, Y. C., et al. (2016). The transcriptomic signature of developing soybean seeds reveals the genetic basis of seed trait adaptation during domestication. *Plant J.* 86, 530–544. doi:10.1111/tpj.13181

Luo, J., Liu, H., Zhou, T., Gu, B., Huang, X., Shangguan, Y., et al. (2013). An-1 encodes a basic helix-loop-helix protein that regulates awn development, grain size, and grain number in rice. *Plant Cell* 25, 3360–3376. doi:10.1105/tpc.113.113589

Ma, Z., Liu, M., Sun, W., Huang, L., Wu, Q., Bu, T., et al. (2019). Genome-wide identification and expression analysis of the trihelix transcription factor family in Tartary buckwheat (*Fagopyrum tataricum*). *BMC Plant Biol.* 19, 344. doi:10.1186/s12870-019-1957-x

Meng, L. S., Wang, Z. B., Yao, S. Q., and Liu, A. (2015). The ARF2-ANT-COR15A gene cascade regulates ABA-signaling-mediated resistance of large seeds to drought in Arabidopsis. *J. Cell Sci.* 128, 3922–3932. doi:10.1242/jcs.171207

Moller, B., and Weijers, D. (2009). Auxin control of embryo patterning. *Cold Spring Harb. Perspect. Biol.* 1, a001545. doi:10.1101/cshperspect.a001545

Pang, X. Y., Bao, W. K., and Zhang, Y. M. (2006). Evaluation of soil fertility under different Cupressus chengiana forests using multivariate approach. *Pedosphere* 16, 602–615. doi:10.1016/s1002-0160(06)60094-5

Persak, H., and Pitzschke, A. (2014). Dominant repression by Arabidopsis transcription factor MYB44 causes oxidative damage and hypersensitivity to abiotic stress. *Int. J. Mol. Sci.* 15, 2517–2537. doi:10.3390/ijms15022517

Ruuska, S. A., Girke, T., Benning, C., and Ohlrogge, J. B. (2002). Contrapuntal networks of gene expression during Arabidopsis seed filling. *Plant Cell* 14, 1191–1206. doi:10.1105/tpc.000877

Sanchez-Montesino, R., Bouza-Morcillo, L., Marquez, J., Ghita, M., Duran-Nebreda, S., Gomez, L., et al. (2019). A regulatory module controlling GA-mediated endosperm cell expansion is critical for seed germination in Arabidopsis. *Mol. Plant* 12, 71–85. doi:10.1016/j.molp.2018.10.009

Sane, V. A., Chourasia, A., and Nath, P. (2005). Softening in mango (Mangifera indica cv. Dashehari) is correlated with the expression of an early ethylene responsive, ripening related expansin gene, MiExpA1. *Postharvest Biol. Technol.* 38, 223–230. doi:10.1016/j.postharvbio.2005.07.008

Santner, A., Calderon-Villalobos, L. I. A., and Estelle, M. (2009). Plant hormones are versatile chemical regulators of plant growth. *Nat. Chem. Biol.* 5, 301–307. doi:10.1038/nchembio.165

Saripalli, G., and Gupta, P. K. (2015). AGPase: its role in crop productivity with emphasis on heat tolerance in cereals. *Theor. Appl. Genet.* 128, 1893–1916. doi:10.1007/s00122-015-2565-2

Savadi, S. (2018). Molecular regulation of seed development and strategies for engineering seed size in crop plants. *Plant Growth Regul.* 84, 401–422. doi:10.1007/s10725-017-0355-3

Schussler, J. R., Brenner, M. L., and Brun, W. A. (1984). Abscisic Acid and its relationship to seed filling in soybeans. *Plant Physiol.* 76, 301–306. doi:10.1104/pp.76.2.301

Song, C., Xiang, D. B., Yan, L., Song, Y., Zhao, G., Wang, Y. H., et al. (2016). Changes in seed growth, levels and distribution of flavonoids during Tartary buckwheat seed development. *Plant Prod. Sci.* 19, 518–527. doi:10.1080/1343943x.2016.1207485

Sun, W. J., Ma, Z. T., and Liu, M. Y. (2020). Cytochrome P450 family: Genome-wide identification provides insights into the rutin synthesis pathway in Tartary buckwheat and the improvement of agricultural product quality. *Int. J. Biol. Macromol.* 164, 4032–4045. doi:10.1016/j.ijbiomac.2020.09.008

Sun, W. J., Yu, H. M., Liu, M. Y., Ma, Z. T., and Chen, H. (2021). Evolutionary research on the expansin protein family during the plant transition to land provides new insights into the development of Tartary buckwheat fruit. *BMC Genomics* 22, 252. doi:10.1186/s12864-021-07562-w

Tanabe, S., Ashikari, M., Fujioka, S., Takatsuto, S., Yoshida, S., Yano, M., et al. (2005). A novel cytochrome P450 is implicated in brassinosteroid biosynthesis via the characterization of a rice dwarf mutant, dwarf11, with reduced seed length. *Plant Cell* 17, 776–790. doi:10.1105/tpc.104.024950

Tian, X., Li, X., Zhou, W., Ren, Y., Wang, Z., Liu, Z., et al. (2017). Transcription factor OsWRKY53 positively regulates brassinosteroid signaling and plant architecture. *Plant Physiol.* 175, 1337–1349. doi:10.1104/pp.17.00946

Tzafrir, I., Pena-Muralla, R., Dickerman, A., Berg, M., Rogers, R., Hutchens, S., et al. (2004). Identification of genes required for embryo development in Arabidopsis. *Plant Physiol.* 135, 1206–1220. doi:10.1104/pp.104.045179

Wang, X., Feng, B., Xu, Z., Sestili, F., Zhao, G., Xiang, C., et al. (2014). Identification and characterization of granule bound starch synthase I (GBSSI) gene of tartary buckwheat (*Fagopyrum tataricum* Gaertn.). *Gene* 534, 229–235. doi:10.1016/j.gene.2013.10.053

Wang, L., Deng, R., Bai, Y., Wu, H., Li, C., Wu, Q., et al. (2022). Tartary buckwheat R2R3-MYB gene FtMYB3 negatively regulates anthocyanin and proanthocyanin biosynthesis. *Int. J. Mol. Sci.* 23, 2775. doi:10.3390/ijms23052775

Wei, W. (2021). *Effect of brassinolide on grain filling and yield of Tartary buckwheat*. Chengdu: Chengdu University. Master's thesis.

Xiang, D. B., Zhao, G., Wan, Y., Tan, M. L., Song, C., and Song, Y. (2016). Effect of planting density on lodging-related morphology, lodging rate, and yield of Tartary buckwheat (*Fagopyrum tataricum*). *Plant Prod. Sci.* 19, 479–488. doi:10.1080/1343943x.2016.1188320

Xiang, D. B., Song, Y., Wu, Q., Ma, C. R., Zhao, J. L., Wan, Y., et al. (2019). Relationship between stem characteristics and lodging resistance of Tartary buckwheat (*Fagopyrum tataricum*). *Plant Prod. Sci.* 22, 202–210. doi:10.1080/1343943x.2019.1577143

Xiang, D. B., Wei, W., Ouyang, J. Y., Le, L. Q., Zhao, G., Peng, L. X., et al. (2020). Nitrogen alleviates seedling stage drought stress response on growth and yield of Tartary buckwheat. *Int. J. Agr. Biol.* 24, 1167–1177. doi:10.17957/IJAB/15.1546

Xu, B., Gou, J. Y., Li, F. G., Shangguan, X. X., Zhao, B., Yang, C. Q., et al. (2013). A cotton BURP domain protein interacts with alpha-expansin and their co-expression promotes plant growth and fruit production. *Mol. Plant* 6, 945–958. doi:10.1093/mp/sss112

Yao, P., Huang, Y., Dong, Q., Wan, M., Wang, A., Chen, Y., et al. (2020). FtMYB6, a light-induced SG7 R2R3-MYB transcription factor, promotes flavonol biosynthesis in Tartary buckwheat (*Fagopyrum tataricum*). *J. Agric. Food Chem.* 68, 13685–13696. doi:10.1021/acs.jafc.0c03037

Zhang, L. J., Li, X. X., Ma, B., Gao, Q., Du, H. L., Han, Y. H., et al. (2017). The tartary buckwheat genome provides insights into rutin biosynthesis and abiotic stress tolerance. *Mol. Plant* 10, 1224–1237. doi:10.1016/j.molp.2017.08.013

Zhang, K., Logacheva, M. D., Meng, Y., Hu, J., Wan, D., Li, L., et al. (2018). Jasmonate-responsive MYB factors spatially repress rutin biosynthesis in *Fagopyrum tataricum*. *J. Exp. Bot.* 69, 1955–1966. doi:10.1093/jxb/ery032

Zhang, L. L., He, Y., Sheng, F., Hu, Y. F., Song, Y., Li, W., et al. (2021). Towards a better understanding of *Fagopyrum dibotrys*: a systematic review. *Chin. Med.* 16, 89. doi:10.1186/s13020-021-00498-z

Zhou, L. H., Wang, H. Y., Chen, X., Li, Y. L., Hussain, N., Cui, L. B., et al. (2017). Identification of candidate genes involved in fatty acids degradation at the late maturity stage in *Brassica napus* based on transcriptomic analysis. *Plant Growth Regul.* 83, 385–396. doi:10.1007/s10725-017-0305-0

Zhu, X., Liang, W., Cui, X., Chen, M., Yin, C., Luo, Z., et al. (2015). Brassinosteroids promote development of rice pollen grains and seeds by triggering expression of Carbon Starved Anther, a MYB domain protein. *Plant J.* 82, 570–581. doi:10.1111/tpj.12820

Zinsmeister, J., Lalanne, D., Terrasson, E., Chatelain, E., Vandecasteele, C., Vu, B. L., et al. (2016). ABI5 is a regulator of seed maturation and longevity in legumes. *Plant Cell* 28, 2735–2754. doi:10.1105/tpc.16.00470

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 09 September 2022
DOI 10.3389/fgene.2022.991249

No evidence for widespread positive selection on double substitutions within codons in primates and yeasts

Frida Belinky[1], Anastassia Bykova[2], Vyacheslav Yurchenko[2]* and Igor B. Rogozin[1]*

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, United States, [2]Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava, Czech Republic

OPEN ACCESS

EDITED BY
Mehdi Pirooznia,
Johnson & Johnson, United States

REVIEWED BY
Yunyun Lv,
Neijiang Normal University, China
Lyu Xueliang,
Huazhong Agricultural University, China
Pawel Mackiewicz,
University of Wrocław, Poland

*CORRESPONDENCE
Vyacheslav Yurchenko,
vyacheslav.yurchenko@osu.cz
Igor B. Rogozin,
rogozin@ncbi.nlm.nih.gov

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 11 July 2022
ACCEPTED 29 August 2022
PUBLISHED 09 September 2022

CITATION
Belinky F, Bykova A, Yurchenko V and Rogozin IB (2022), No evidence for widespread positive selection on double substitutions within codons in primates and yeasts.
Front. Genet. 13:991249.
doi: 10.3389/fgene.2022.991249

COPYRIGHT
© 2022 Belinky, Bykova, Yurchenko and Rogozin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Nucleotide substitutions in protein-coding genes can be divided into synonymous (S) and non-synonymous (N) ones that alter amino acids (including nonsense mutations causing stop codons). The S substitutions are expected to have little effect on function. The N substitutions almost always are affected by strong purifying selection that eliminates them from evolving populations. However, additional mutations of nearby bases can modulate the deleterious effect of single N substitutions and, thus, could be subjected to the positive selection. This effect has been demonstrated for mutations in the serine codons, stop codons and double N substitutions in prokaryotes. In all abovementioned cases, a novel technique was applied that allows elucidating the effects of selection on double substitutions considering mutational biases. Here, we applied the same technique to study double N substitutions in eukaryotic lineages of primates and yeast. We identified markedly fewer cases of purifying selection relative to prokaryotes and no evidence of codon double substitutions under positive selection. This is consistent with previous studies of serine codons in primates and yeast. In general, the obtained results strongly suggest that there are major differences between studied pro- and eukaryotes; double substitutions in primates and yeasts largely reflect mutational biases and are not hallmarks of selection. This is especially important in the context of detection of positive selection in codons because it has been suggested that multiple mutations in codons cause false inferences of lineage-specific site positive selection. It is likely that this concern is applicable to previously studied prokaryotes but not to primates and yeasts where markedly fewer double substitutions are affected by positive selection.

KEYWORDS

natural selection, tandem mutations, short-term evolution, neutral evolution, double substitutions, positive selection, negative selection, purifying selection

Introduction

In classic population genetics, co-localized substitutions are assumed to occur one at a time, independently of one another. However, clustering of mutations, in particular, those occurring in adjacent sites (multiple nucleotide mutations) has been documented in many diverse organisms (Averof et al., 2000; Drake et al., 2005; Drake, 2007; Schrider et al., 2011; Stone et al., 2012; Terekhanova et al., 2013; Harris and Nielsen, 2014; Besenbacher et al., 2016). Double substitutions within the same codon in protein-coding genes have also been claimed to be driven by positive selection. This conclusion stemmed from comparisons of the observed frequencies of double substitutions to those expected from the frequencies of single substitutions: if the frequency of a double substitution is significantly greater than the product of the frequencies of the respective single substitutions, positive selection is inferred (Bazykin et al., 2004; Rogozin et al., 2016; Belinky et al., 2018). This observation is consistent with the possibility that a prevalence of positively selected double nucleotide mutations is compensation for the first deleterious mutation through subsequent positive selection acting on the second substitution (Bazykin et al., 2004; Rogozin et al., 2016; Belinky et al., 2018). Positive selection affecting double substitutions has been detected as a general trend in the rodent lineage (Bazykin et al., 2004). Similarly, signatures of positive selection have been found for double substitutions in stop codons in prokaryotes (UAG → UGA and UGA → UAG), which could be attributed to the deleterious non-stop intermediate, UGG (Belinky et al., 2018) and double substitutions in two disjoint series of codons for serine (Rogozin et al., 2016). Thus, multiple nucleotide mutations in codons potentially could originate from selection, mutational biases including clusters of mutations (Averof et al., 2000; Drake et al., 2005; Drake, 2007; Schrider et al., 2011; Stone et al., 2012; Terekhanova et al., 2013; Harris and Nielsen, 2014; Besenbacher et al., 2016) or a combination of both these factors.

Previously, we assessed the selection that affects double substitutions within codon in prokaryotes (Belinky et al., 2019). Briefly, we compared the frequency of each such double substitution to the frequency of a double synonymous substitution in adjacent codons with the same base composition (Belinky et al., 2019). Although it is well known that transition (A:T ↔ G:A) and transversion (A:T ↔ T:A, A:T ↔ C:G, G:C ↔ C:G) rates differ substantially, the differences between different combinations of specific transitions and transversions are less thoroughly characterized, and it is not clear to what extent adjacency of mutations is modulated by base composition. We thus compared all codon double substitutions to their respective double synonymous substitutions with the same nucleotide changes. In many cases, it was found that a codon double substitution has a significantly higher double/single ratio, compared to the same double synonymous substitution, suggesting that these are true cases of positive selection that acts on the second substitution and brings it to fixation in prokaryotes (Belinky et al., 2019).

In this paper the same methodology was applied for analyses of selection in yeasts and primates (including human). No signs of wide-spread positive selection were detected. This result suggests major differences in selection modes between prokaryotes (Belinky et al., 2019) and two studied eukaryotic lineages (primates and yeasts). This is likely to be important for inference of lineage-specific site positive selection.

Materials and methods

Datasets

To reconstruct mutations in protein-coding DNA under the parsimony principle, we inferred and analyzed single and double substitutions in triplets of closely related primates and yeasts as previously described (Rogozin et al., 2016). In brief, the parsimony principle implies that mutations occur along the thick branches in the trees (Figure 1A) assuming that there is no mutation or one mutation per each position. Whole-genome alignments of three yeast species (*Saccharomyces cerevisiae*, *S. paradoxus*, and *S. mikatae*) were downloaded from the *Saccharomyces* Genome Database (SGD, www.yeastgenome.org/). Local alignments of protein-coding regions were extracted using the SGD orthology assignments (Rogozin et al. , 2016). Protein-coding sequences for primates (*Homo sapiens*, *Callithrix jacchus* and *Otolemur garnettiiwere*) and their orthology assignments were obtained from Ensembl databases as previously described (Belinky et al., 2018). Briefly, protein-coding sequences were downloaded for each species from the Ensembl database, as well as orthology assignments from Ensembl mart (Kersey et al., 2016). Genes with ‘one-to-one’ orthology were aligned using MAFFT with the -linsi algorithm (Katoh et al., 2005). In total, 15,234 primate and 4,100 yeast gene alignments were used for further analyses.

Analysis of codon double substitutions

Details of analyses of double substitutions in codons are described in (Belinky et al., 2019). Here, we provide a brief description of the methodology. For each codon change (Figure 1B), the frequency of change to any other codon was the number of changes divided by the number of ancestral reconstructions of this codon based on the parsimony principal. For each double substitution the double/single ratio was the observed double substitution frequency divided by the cumulative single substitution frequency. For example, for the change AAA→GGA the double/single ratio was the observed frequency of AAA→GGA divided by the cumulative counts of

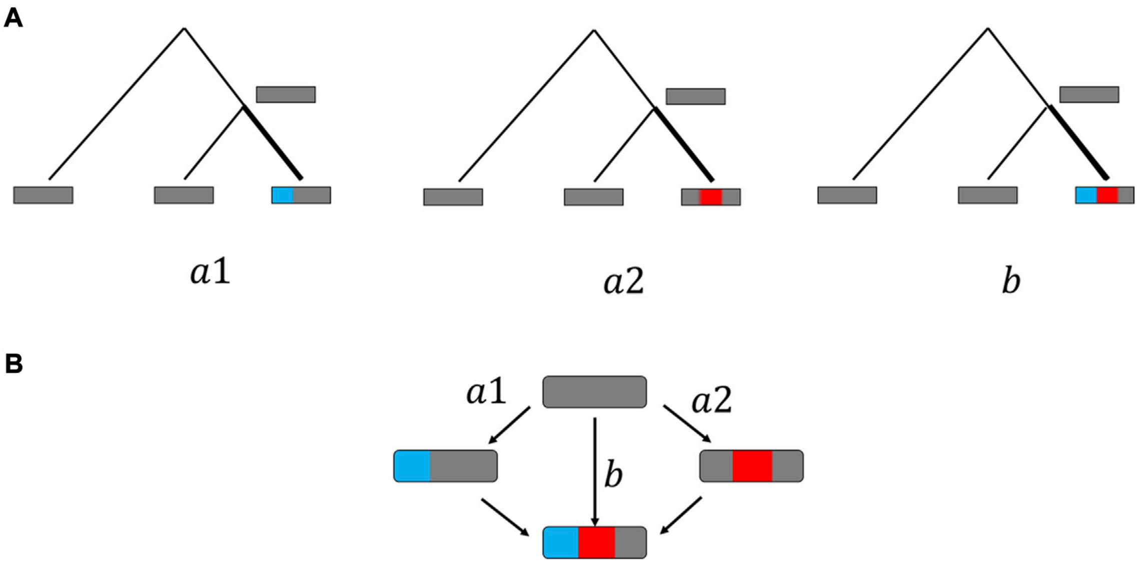


FIGURE 1
Conceptual scheme of double substitution analysis. **(A)**Single or double substitutions are inferred from the genomic data by construction of genomes triplets and relying on parsimony principle (see Material and Methods). **(B)** Point mutations are assumed to appear one at a time, such that observed double substitutions **(B)** occur through intermediate single substitutions states. For each double substitution, there are two possible single substitution pathways **(a1, a2)**. The double fraction DF is calculated as the ratio between the number of double substitutions **(b)** and the sum of relevant single (a1+a2) and double **(b)** substitutions.

AAA→AGA, AAA→GAA and AAA→GGA. Thus, for each double substitution (Figure 1) the following data were collected and estimated:

1) The double substitution count (b in the Figure 1).
2) The single substitution count (which is the summation of the two single counts (a1 and a2 in the Figure 1).

We used double fractions (DFs) as a measure of selection. The DF is calculated as the observed double substitution count (b in the Figure 1) divided by the sum of the single (a1 and a2 in the Figure 1) and double substitution counts:

$$DF = b/(a1 + a2 + b)$$

The selection on double substitutions was analyzed by comparing DF for within-codon double substitutions to two null models described below.

Analysis of double synonymous substitutions in adjacent codons—null models

For double synonymous substitutions in adjacent codons, we collected the same data as for codon double substitutions in codon-like 3-base sequences with three possible configurations (Figure 2):

A. An invariant 2nd codon positions followed by a 4-fold degenerate site in the 3rd codon positions, that is, followed by a 2-fold degenerate site in the 1st codon position of the next codon (the 231 configuration, Figure 2B).

B. A 4-fold degenerate site in the 3rd codon positions, that is, followed by a 2-fold degenerate site in the 1st codon position of the next codon, that is, followed by an invariant base in the 2nd codon position of the second codon (the 312 configuration, Figure 2C).

C. A 4-fold degenerate site in the 3rd codon positions, that is, followed by an invariant 1st codon position in the second codon of which the 2nd position is disregarded and followed by a 4-fold degenerate site in the 3rd codon position (Figure 2D).

The first codon in configurations A-B can be any of the 4-fold degenerate codons, i.e, codons for L, V, S, P Y, A, R and G, and the second codon of configurations A-B can be either a codon for R or L which are the only two amino acids that have a degenerate 1st codon position. An additional restriction for configurations A-B is that the ancestral state of the 3rd codon position of the 2nd codon is a purine (A/G) since only then the 1st codon substitution can be synonymous. Similarly, the 1st and 2nd codons configuration C can be any of the 4-fold degenerate codons.

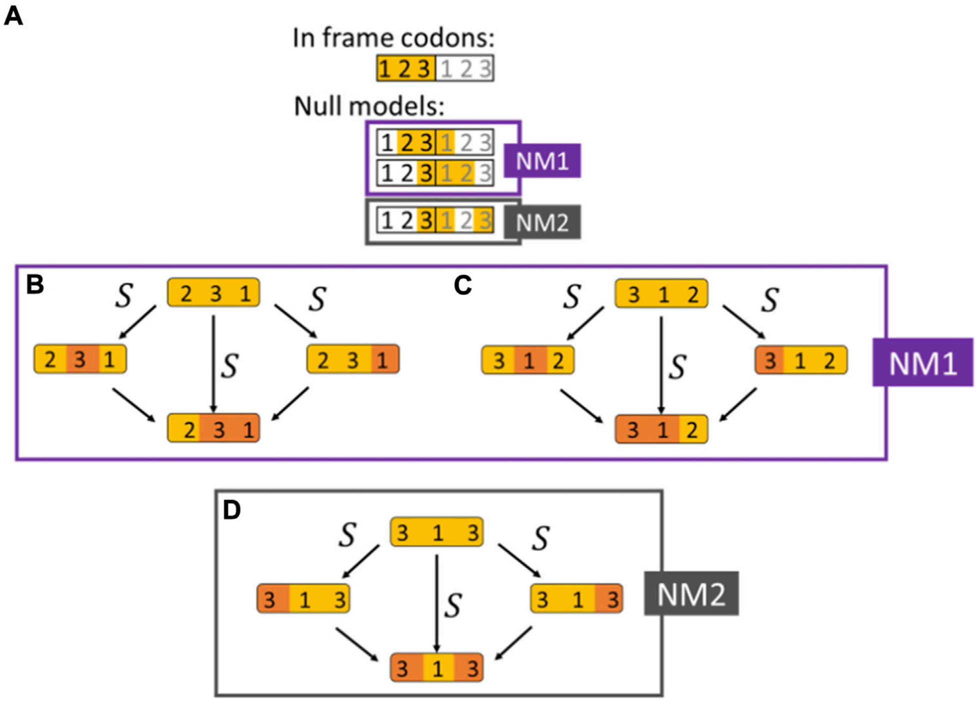


FIGURE 2
Double synonymous substitutions in adjacent codons used as null models. **(A)** The selection on double substitutions inferred by comparing the DF for codons and their respective null models shown in orange (NM1 and NM2). Two adjacent codons are illustrated, and the nucleotide position within the codon is indicated according to the reading frame. The three null models are artificial codons constructed by considering positions from two adjacent codons. **(B)** Null model NM1 (the 321 configuration). An invariant 2nd codon positions in the first codon, followed by a 4-fold degenerate site in the 3rd positions of the first codon, that is, followed by a 2-fold degenerate site in the 1st codon position of the 2nd codon. **(C)** Null model NM1 (the 312 configuration). A 4-fold degenerate site in the 3rd codon position followed by a 2-fold degenerate site in the 1st codon position of the second codon, that is, followed by an invariant base in the 2nd codon position of the second codon. **(D)** Null model NM2. A 4-fold degenerate site in the 3rd position of the 1st codon followed by an invariant 1st position in the second codon and by a 4-fold degenerate site in the 3rd codon position (skipping the 2nd position of the 2nd codon).

Assignment of codon double substitution types

For each codon double substitution there are two distinct paths to get from the ancestral state codon to the final (derived) codon state, with each step in the path having a single substitution to reach an intermediate state codon (Figure 2). Each step can be either synonymous or non-synonymous, and the ancestral vs. final codon could be either non-synonymous or synonymous. Some codon substitution could have a stop as an intermediate codon in one of the paths, these cases were disregarded in the current analysis. In this analysis we assigned the combination type to each codon double substitution based on the synonymy of the ancestral to the intermediate codons, and the synonymy of the ancestral vs. the final codon state (Figure 3, left panels and Supplementary Figure S2). NS denotes codon double substitutions in which (at least) one of the intermediates is non-synonymous while the final codon is synonymous compared to the ancestral codon (Figure 3A and Supplementary Figures S1D S2). SS denotes codon double substitutions in which both intermediates and the final codon are all synonymous codons (Figure 3B and Supplementary Figure S2). SN denotes codon double substitutions in which (at least) one intermediate is synonymous while the final codon is nonsynonymous compared to the ancestral codon (Figure 3 and Supplementary Figure S2). NN denotes codon double substitutions in which both intermediates are nonsynonymous, and the final codon is also nonsynonymous compared to the ancestral one (Figure 3D and Supplementary Figure S2).

Statistical testing

Fisher's exact test was used to compare the number of double codon substitutions to single cumulative substitutions, to test for significant differences in DF between codon double substitutions and the comparable null models. An example of the comparison of the non-adjacent codon double substitution

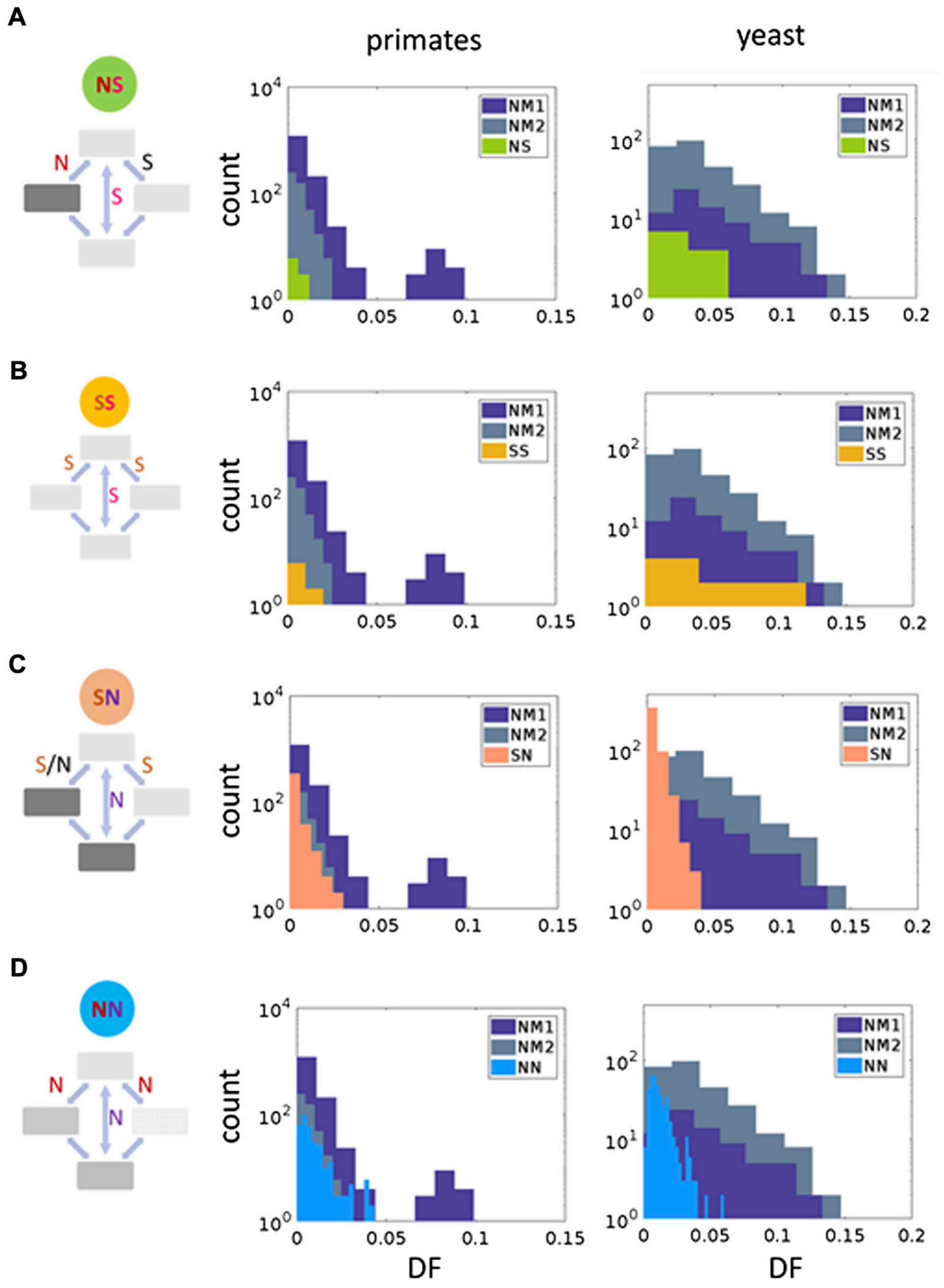


FIGURE 3
Selective regimes of the codon double substitutions in primates and yeasts. Right panels show a classification of codon double substitution based on the synonymy of the ancestral vs. the final codon state, and the synonymy of the ancestral to the intermediate codons. Two left panels show comparisons of DF for each codon double substitution class to the double synonymous null models (NM1 and NM2) using the Mann–Whitney U test. **(A)** NS, one non-synonymous intermediate, synonymous final codon. Primates: NM1 *p*-value = 0.77, NM2 *p*-value = 0.72. Yeasts: NM1 *p*-value = 0.06, NM2 *p*-value = 0.25. **(B)** SS, double synonymous codon substitutions. Primates: NM1 *p*-value = 0.42, NM2 *p*-value = 0.17. Yeasts: NM1 *p*-value = 0.82, NM2 *p*-value = 0.45. **(C)** SN, at least one synonymous intermediate codon, non-synonymous final codon. Primates: NM1 *p*-value = 2.38×10^{-63}, NM2 *p*-value = 8.73×10^{-34}. Yeasts: NM1 *p*-value = 4.28×10^{-38}, NM2 *p*-value = 5.64×10^{-98}. **(D)** NN—both intermediates and the final codon are non-synonymous to the ancestral. Primates: NM1 *p*-value = 0.059, NM2 *p*-value = 2.53×10^{-5}. Yeasts: NM1 *p*-value = 7.16×10^{-27}, NM2 *p*-value = 5.61×10^{-56}.

CTT→TTA is shown in the Supplementary Figure S1D. The Mann–Whitney U test was used to compare the DF values between each of the codon double substitution types (SS, SN, NS, NN) and each of the null models (NM1 and NM2). The Bonferroni correction was applied to correct for multiple testing.

Results

Different types of codon double substitutions in primates and yeasts

Representing all within-codon double substitutions in the general form, "ancestral-intermediate-final", we define the following 4 combinations of codons: 1) SS is "S intermediate—S final" codons, 2) SN is "S intermediate—N final" codons, 3) NS is "N intermediate—S final" codons, 4) NN is "N intermediate—N final" codons (Figure 3, left panels and Supplementary Figure S2) (Rogozin et al., 2016; Belinky et al., 2018; Belinky et al., 2019).

Similar to our previous study of double substitutions in prokaryotes (Belinky et al., 2019), we consider three types of codon-like double synonymous substitutions that were used as null models for the double substitutions in codons (Supplementary Figure S1). The selection pressure on each codon double substitution is assessed by comparing the double/single substitution ratio DF (that is, the ratio of the frequency of a double substitution to the sum of the frequencies of the single and double substitutions in the respective codon positions) to that for double synonymous substitutions (Supplementary Figure S1). The DF is assumed to be mostly affected by the substitution rate at the second step (from intermediate codons to final codons, Figure 1B). Thus, a significantly lower DF compared to that of the corresponding double synonymous substitution will be indicative of purifying selection, and conversely, a higher ratio will point to positive selection.

Comparisons of double mutation DF values with null models NM1 and NM2 (Figure 3, central and right panels) suggested that the dominant mode of selection is purifying selection. In all eight studied cases in primates and yeasts the mean DF values is smaller than DF values for null models (Figure 3). These differences are statistically significant for NN and SN values (Figure 3). The NM1 model tends to produce wider distributions compared to NM2 model (Figure 3). This is likely to be due to a higher frequency of tandem mutations compared to mutations separated by one nucleotide (Averof et al., 2000; Drake et al., 2005; Drake, 2007; Schrider et al., 2011; Stone et al., 2012; Terekhanova et al., 2013; Harris and Nielsen, 2014; Besenbacher et al., 2016).

Modes of selection in specific codon double substitution classes in primates

We analyzed four types of double substitutions in more detail. To characterize the modes of selection that affect each codon double substitution in greater detail, the frequency of each codon double substitution was compared to the same codon-like substitution pattern in a double synonymous null model (Figure 2). Each codon double substitution is compared to either NM1 or NM2 depending on the distance between the substituted bases (Supplementary Table S1). In total, of the 716 codon double substitutions compared (Supplementary Table S1), only <1% (2 cases after Bonferroni correction) had significantly higher DF compared to the equivalent double synonymous substitutions (Supplementary Table S1), which is compatible with positive selection, and 15% (104 cases after the Bonferroni correction) had significantly lower DF, compatible with purifying selection (Figure 4A and Supplementary Table S1). This result suggests that positive selection affects a negligible fraction of double substitutions in codons although these cases may be false positives. A substantial fraction of double substitutions is subject to purifying selection (Figure 4A).

For NS and SS double substitutions no signs of positive or negative selection were detected (Figure 4A). A significant trend of purifying selection on codon double substitutions is evident in combination SN (Figure 4A), in which double substitutions have significantly lower DF compared to the double synonymous DF (Figure 4A). Combination NN (312 instances) has 2 cases with codon under positive selection and 4 cases compatible with purifying selection, thus neutrality cannot be rejected for the entire group (Figure 4A). The individual cases in combination NN that are compatible with positive selection are TTT → GGT (F → G) and TTT → GCT (F → A) (Supplementary Table S1).

Modes of selection in specific codon double substitution classes in yeasts

Highly similar results were obtained for yeasts (Figure 4B). In total, of the 317 codon double substitutions compared (Supplementary File S1), only 1% (4 cases after Bonferroni correction) had significantly higher DF compared to the equivalent double synonymous substitutions (Supplementary Table S1), which is compatible with positive selection. This result suggests that positive selection affects a negligible fraction of double substitutions in codons although these cases may be false positives. 34% of studied (108 cases after the Bonferroni correction) had significantly lower DF, which is compatible with purifying selection. A substantial fraction of double substitutions is likely to be subject to purifying selection (Figure 4B).

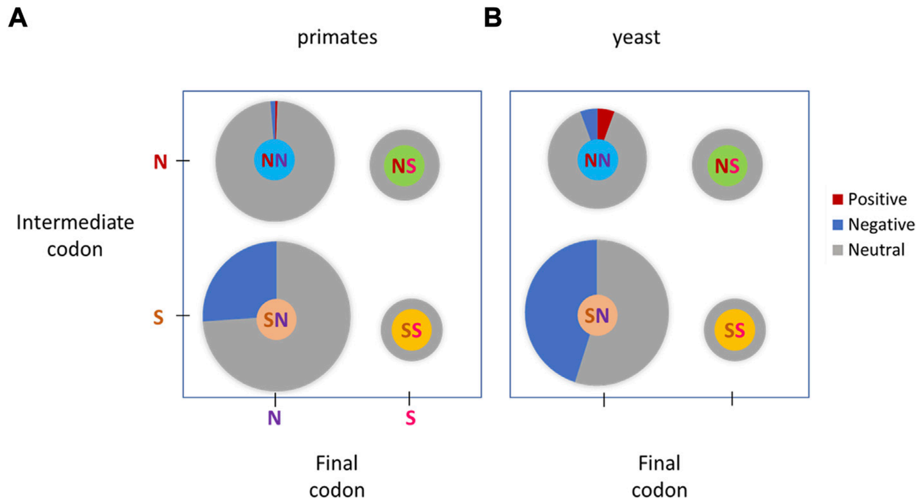


FIGURE 4
Selective pressure in different codon double substitutions classes. Positive, combinations compatible with positive selection, where a codon double substitution has a significantly higher DF than the corresponding DF of a null model (NM1 or NM2). Negative, combinations compatible with purifying selection, where a codon double substitution has a significantly lower DF than the corresponding DF of a null model. Neutral, combinations, compatible with neutral evolution, where the codon DF was not significantly different from that of the corresponding DF of a null model. **(A)**, primates; **(B)**, yeasts.

For NS and SS double substitutions no signs of positive or negative selection were detected (Figure 4B). A significant trend of purifying selection on codon double substitutions is evident in combination SN (Figure 4A), in which many double substitutions have significantly lower DF compared to the double synonymous DF (Figure 4A). Combination NN contains only 4 cases with codon under positive selection and 4 cases compatible with purifying selection. Thus, neutrality cannot be rejected for the entire group (Figure 4A). The individual cases in combination NN that are compatible with positive selection are ACT → GTT (T → V), CCT → TTT (P → F), TCT → CTT (S → L), and TTT → CCT (F → P) (Supplementary Table S1).

Discussion

Multiple mutations within the same codon have been claimed to be driven by positive selection (Bazykin et al., 2004; Rogozin et al., 2016; Belinky et al., 2018). This claim is consistent with the possibility that a prevalence of positively selected double nucleotide mutations is a compensation for the first deleterious mutation through subsequent positive selection (Bazykin et al., 2004; Rogozin et al., 2016; Belinky et al., 2018). The main goals of this work were to consider the mutational biases in the inference of selection in codon double substitutions and to understand whether codon double substitutions in yeasts and primates were under any type of selection compared to double synonymous substitutions. Just a few cases of elevated DF (<1 and 1% for human and yeast, accordingly) were detected for the combination NN. Such cases are compatible with previously reported positive selection on multiple nucleotide substitutions (Bazykin et al., 2004). Analysis of individual cases in primates and yeasts suggested that codons TTT (encoding phenylalanine) and CCT (encoding proline) are most frequent in terms of positively selected double substitutions (Supplementary Table S1).

Distributions of DF values for NS and SS double substitutions are not statistically different from NM1 and NM2 distributions (Figure 3), whereas SN and NN had significantly lower DF values suggesting that purifying selection substantially influences these classes of double substitutions in both primates and yeasts (Figure 3). In total, 15 and 34% double substitutions in primates and yeasts had significantly lower DF (after the Bonferroni correction), compatible with purifying selection. This result suggests that purifying selection affects a substantial fraction of double substitutions in codons. However, it is evident that in all four categories neutrality is the dominant mode of evolution (Figure 4).

We used synonymous sites as a control. Selection on synonymous sites have been previously shown in prokaryotes as well as in eukaryotes (Chamary and Hurst, 2005; Zhou et al., 2010; Gu et al., 2012; Lawrie et al., 2013; Shabalina et al., 2013; Long et al., 2018), while the reason behind this selection is not completely clear and could be contributed to stability of the DNA and staking effects (Goncearenco and Berezovsky, 2014), translational accuracy (Stoletzki and Eyre-Walker, 2007), and importance of secondary structure (Chamary and Hurst, 2005;

Shabalina et al., 2013). Possible factors at the protein level are protein folding/structure (Oresic and Shalloway, 1998; Pechmann and Frydman, 2013) and a general selection at the amino acid level interacting with nucleotide replacements (Morton, 2001; Blazej et al., 2017). Although synonymous positions can be under some level of purifying selection, the same mutational forces are expected to influence codon non-synonymous double substitutions of the same bases, e.g., mutation rates that are influenced by specific bases would be similarly affected whether the mutation is synonymous or non-synonymous.

Previously, we assessed the selection that affects double substitutions within codons in prokaryotes (Belinky et al., 2019) using the same approach described in this paper. In many cases, it was found that codon double substitutions have significantly higher double/single ratios, compared to the same double synonymous substitutions (14%), suggesting that these are true cases of positive selection that acts on the second substitution and brings it to fixation in prokaryotes (Belinky et al., 2019). In primates and yeasts, we found just a few cases of putative positive selection (~1%). Overall, the fraction of neutrally evolving codons is dramatically different: 11% in prokaryotes (Belinky et al., 2019) vs. 75% in primates and 65% in yeasts.

Recently it has been claimed that positive selection is overestimated by the branch-site test (BST), since most of the sites supporting positive selection are due to multinucleotide mutations (MNS) (Venkat et al., 2018). Phylogenetic tests of adaptive evolution, such as the widely used BST (branch-site test), assume that nucleotide substitutions occur independently. However, recent research has shown that errors at adjacent sites often occur during DNA repair/replication (Drake et al., 2005; Drake, 2007; Schrider et al., 2011; Stone et al., 2012; Terekhanova et al., 2013; Harris and Nielsen, 2014; Besenbacher et al., 2016), and the resulting MNS are overwhelmingly likely to be nonsynonymous (Venkat et al., 2018). Simulations under conditions derived from human and fly sequence alignments without positive selection show that realistic rates of MNS cause a systematic bias towards false inferences of selection (Venkat et al., 2018). This concern is certainly consistent with the observed substantial fraction of positively evolving double substitutions observed in prokaryotes (Belinky et al., 2019). However, the conclusion of the Venkat and co-workers (Venkat et al., 2018) requires a lot of caution, when applied to studied eukaryotes (primates and yeasts), where markedly fewer double substitutions are under positive selection (Figure 4).

The observed difference between pro- and eukaryotes (primates and yeasts) was observed previously for serine codons (Rogozin et al., 2016). Here, in the analyzed two eukaryotic lineages (yeast and primates), the difference of the DF of codon double substitutions over DF of the double synonymous in null models was much smaller than in prokaryotes (Belinky et al., 2019). This is consistent with the fundamental population-genetic theory (Lynch, 2007; Charlesworth, 2009; Loewe and Hill, 2010), whereby eukaryotes have substantially smaller effective population sizes than prokaryotes, and the consequent decrease in the power of selection most likely cause weaker pressure for restoration of amino acids that are under positive selection in prokaryotes, but not in studied eukaryotes (primates and yeasts). This hypothesis is also consistent with the observed larger fraction of positively and negatively selected double substitutions for yeasts compared to primates (Figure 4), which have much smaller population sizes.

The observed low fraction of deleterious intermediates associated with further positive selection (Figure 3) could be also due to various compensatory mechanisms at the RNA or protein level (Ellis, 1990; Fink, 1999; El-Brolosy and Stainier, 2017). For example, one reason for the higher complexity of eukaryotes compared to prokaryotes is the increased number of domain combinations found in eukaryotes, where, for example, binding domains have been added to existing catalytic proteins (Bjorklund et al., 2005). Thus, compensatory mechanisms at the level of interactions between proteins and domains within multidomain proteins are expected to be more abundant in eukaryotes compared to prokaryotes (Ekman et al., 2006; Bhaskara and Srinivasan, 2011). It should be noted that involvement of other non-trivial compensatory mechanisms in eukaryotes cannot be excluded. Future analyses of the impact of various compensatory mechanisms are likely to provide a clearer picture of eukaryote-specific trends of evolution.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: www.yeastgenome.org https://www.ensembl.org/index.html.

Author contributions

Formal analysis: FB and AB. Supervision: VY and IR. Original draft writing: FB and IR. Text editing: all authors.

Funding

This research was supported in part by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health (FB and IR). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Work in VY lab is supported by the European Regional Development Funds (CZ.02.1.01/0.0/0.0/16_019/0000759). AB was supported by the grant SGS/PřF/2022 from the University of Ostrava. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Acknowledgments

IR thanks Youri Pavlov and Eugenia Poliakov for the useful discussion.

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/fgene.2022.991249/full#supplementary-material

References

Averof, M., Rokas, A., Wolfe, K. H., and Sharp, P. M. (2000). Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* 287, 1283–1286. doi:10.1126/science.287.5456.1283

Bazykin, G. A., Kondrashov, F. A., Ogurtsov, A. Y., Sunyaev, S., and Kondrashov, A. S. (2004). Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. *Nature* 429, 558–562. doi:10.1038/nature02601

Belinky, F., Babenko, V. N., Rogozin, I. B., and Koonin, E. V. (2018). Purifying and positive selection in the evolution of stop codons. *Sci. Rep.* 8, 9260. doi:10.1038/s41598-018-27570-3

Belinky, F., Sela, I., Rogozin, I. B., and Koonin, E. V. (2019). Crossing fitness valleys via double substitutions within codons. *BMC Biol.* 17, 105. doi:10.1186/s12915-019-0727-4

Besenbacher, S., Sulem, P., Helgason, A., Helgason, H., Kristjansson, H., Jonasdottir, A., et al. (2016). Multi-nucleotide de novo Mutations in Humans. *PLoS Genet.* 12, e1006315. doi:10.1371/journal.pgen.1006315

Bhaskara, R. M., and Srinivasan, N. (2011). Stability of domain structures in multi-domain proteins. *Sci. Rep.* 1, 40. doi:10.1038/srep00040

Bjorklund, A. K., Ekman, D., Light, S., Frey-Skott, J., and Elofsson, A. (2005). Domain rearrangements in protein evolution. *J. Mol. Biol.* 353, 911–923. doi:10.1016/j.jmb.2005.08.067

Blazej, P., Mackiewicz, D., Wnetrzak, M., and Mackiewicz, P. (2017). The impact of selection at the amino acid level on the usage of synonymous codons. *G3 (Bethesda)* 7, 967–981. doi:10.1534/g3.116.038125

Chamary, J. V., and Hurst, L. D. (2005). Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* 6, R75. doi:10.1186/gb-2005-6-9-r75

Charlesworth, B. (2009). Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* 10, 195–205. doi:10.1038/nrg2526

Drake, J. W., Bebenek, A., Kissling, G. E., and Peddada, S. (2005). Clusters of mutations from transient hypermutability. *Proc. Natl. Acad. Sci. U. S. A.* 102, 12849–12854. doi:10.1073/pnas.0503009102

Drake, J. W. (2007). Too many mutants with multiple mutations. *Crit. Rev. Biochem. Mol. Biol.* 42, 247–258. doi:10.1080/10409230701495631

Ekman, D., Light, S., Bjorklund, A. K., and Elofsson, A. (2006). What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae*? *Genome Biol.* 7, R45. doi:10.1186/gb-2006-7-6-r45

El-Brolosy, M. A., and Stainier, D. Y. R. (2017). Genetic compensation: a phenomenon in search of mechanisms. *PLoS Genet.* 13, e1006780. doi:10.1371/journal.pgen.1006780

Ellis, R. J. (1990). The molecular chaperone concept. *Semin. Cell Biol.* 1, 1–9.

Fink, A. L. (1999). Chaperone-mediated protein folding. *Physiol. Rev.* 79, 425–449. doi:10.1152/physrev.1999.79.2.425

Goncearenco, A., and Berezovsky, I. N. (2014). The fundamental tradeoff in genomes and proteomes of prokaryotes established by the genetic code, codon entropy, and physics of nucleic acids and proteins. *Biol. Direct* 9, 29. doi:10.1186/s13062-014-0029-2

Gu, W., Wang, X., Zhai, C., Xie, X., and Zhou, T. (2012). Selection on synonymous sites for increased accessibility around miRNA binding sites in plants. *Mol. Biol. Evol.* 29, 3037–3044. doi:10.1093/molbev/mss109

Harris, K., and Nielsen, R. (2014). Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Res.* 24, 1445–1454. doi:10.1101/gr.170696.113

Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33, 511–518. doi:10.1093/nar/gki198

Kersey, P. J., Allen, J. E., Armean, I., Boddu, S., Bolt, B. J., Carvalho-Silva, D., et al. (2016). Ensembl genomes 2016: More genomes, more complexity. *Nucleic Acids Res.* 44, D574–D580. doi:10.1093/nar/gkv1209

Lawrie, D. S., Messer, P. W., Hershberg, R., and Petrov, D. A. (2013). Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genet.* 9, e1003527. doi:10.1371/journal.pgen.1003527

Loewe, L., and Hill, W. G. (2010). The population genetics of mutations: good, bad and indifferent. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 1153–1167. doi:10.1098/rstb.2009.0317

Long, H., Sung, W., Kucukyildirim, S., Williams, E., Miller, S. F., Guo, W., et al. (2018). Evolutionary determinants of genome-wide nucleotide composition. *Nat. Ecol. Evol.* 2, 237–240. doi:10.1038/s41559-017-0425-y

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl. Acad. Sci. U. S. A.* 104, 8597–8604. doi:10.1073/pnas.0702207104

Morton, B. R. (2001). Selection at the amino acid level can influence synonymous codon usage: implications for the study of codon adaptation in plastid genes. *Genetics* 159, 347–358. doi:10.1093/genetics/159.1.347

Oresic, M., and Shalloway, D. (1998). Specific correlations between relative synonymous codon usage and protein secondary structure. *J. Mol. Biol.* 281, 31–48. doi:10.1006/jmbi.1998.1921

Pechmann, S., and Frydman, J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* 20, 237–243. doi:10.1038/nsmb.2466

Rogozin, I. B., Belinky, F., Pavlenko, V., Shabalina, S. A., Kristensen, D. M., and Koonin, E. V. (2016). Evolutionary switches between two serine codon sets are driven by selection. *Proc. Natl. Acad. Sci. U. S. A.* 113, 13109–13113. doi:10.1073/pnas.1615832113

Schrider, D. R., Hourmozdi, J. N., and Hahn, M. W. (2011). Pervasive multinucleotide mutational events in eukaryotes. *Curr. Biol.* 21, 1051–1054. doi:10.1016/j.cub.2011.05.013

Shabalina, S. A., Spiridonov, N. A., and Kashina, A. (2013). Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Res.* 41, 2073–2094. doi:10.1093/nar/gks1205

Stoletzki, N., and Eyre-Walker, A. (2007). Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol. Biol. Evol.* 24, 374–381. doi:10.1093/molbev/msl166

Stone, J. E., Lujan, S. A., Kunkel, T. A., and Kunkel, T. A. (2012). DNA polymerase zeta generates clustered mutations during bypass of endogenous DNA lesions in *Saccharomyces cerevisiae*. *Environ. Mol. Mutagen.* 53, 777–786. doi:10.1002/em.21728

Terekhanova, N. V., Bazykin, G. A., Neverov, A., Kondrashov, A. S., and Seplyarskiy, V. B. (2013). Prevalence of multinucleotide replacements in evolution of primates and Drosophila. *Mol. Biol. Evol.* 30, 1315–1325. doi:10.1093/molbev/mst036

Venkat, A., Hahn, M. W., and Thornton, J. W. (2018). Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nat. Ecol. Evol.* 2, 1280–1288. doi:10.1038/s41559-018-0584-5

Zhou, T., Gu, W., and Wilke, C. O. (2010). Detecting positive and purifying selection at synonymous sites in yeast and worm. *Mol. Biol. Evol.* 27, 1912–1922. doi:10.1093/molbev/msq077

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 23 September 2022
DOI 10.3389/fgene.2022.992070

Deep learning methods may not outperform other machine learning methods on analyzing genomic studies

Yao Dong[1,2,3], Shaoze Zhou[2], Li Xing[4], Yumeng Chen[1,3], Ziyu Ren[1,3], Yongfeng Dong[1,3]* and Xuekui Zhang[2]*

[1]School of Artifcial Intelligence, Hebei University of Technology, Tianjin, China, [2]Department of Mathematics and Statistics, University of Victoria, Victoria, BC, Canada, [3]Hebei Province Key Laboratory of Big Data Computing, Tianjin, China, [4]Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, Saskatoon

Deep Learning (DL) has been broadly applied to solve big data problems in biomedical fields, which is most successful in image processing. Recently, many DL methods have been applied to analyze genomic studies. However, genomic data usually has too small a sample size to fit a complex network. They do not have common structural patterns like images to utilize pre-trained networks or take advantage of convolution layers. The concern of overusing DL methods motivates us to evaluate DL methods' performance versus popular non-deep Machine Learning (ML) methods for analyzing genomic data with a wide range of sample sizes. In this paper, we conduct a benchmark study using the UK Biobank data and its many random subsets with different sample sizes. The original UK Biobank data has about 500k participants. Each patient has comprehensive patient characteristics, disease histories, and genomic information, i.e., the genotypes of millions of Single-Nucleotide Polymorphism (SNPs). We are interested in predicting the risk of three lung diseases: asthma, COPD, and lung cancer. There are 205,238 participants have recorded disease outcomes for these three diseases. Five prediction models are investigated in this benchmark study, including three non-deep machine learning methods (Elastic Net, XGBoost, and SVM) and two deep learning methods (DNN and LSTM). Besides the most popular performance metrics, such as the F1-score, we promote the hit curve, a visual tool to describe the performance of predicting rare events. We discovered that DL methods frequently fail to outperform non-deep ML in analyzing genomic data, even in large datasets with over 200k samples. The experiment results suggest not overusing DL methods in genomic studies, even with biobank-level sample sizes. The performance differences between DL and non-deep ML decrease as the sample size of data increases. This suggests when the sample size of data is significant, further increasing sample sizes leads to more performance gain in DL methods. Hence, DL methods could be better if we analyze genomic data bigger than this study.

Check for updates

OPEN ACCESS

EDITED BY
Mehdi Pirooznia,
Johnson & Johnson, United States

REVIEWED BY
Xiangyu Luo,
Renmin University of China, China
Rahim Alhamzawi,
University of Al-Qadisiyah, Iraq

*CORRESPONDENCE
Yongfeng Dong,
dongyf@hebut.edu.cn
Xuekui Zhang,
xuekui@uvic.ca

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 12 July 2022
ACCEPTED 04 August 2022
PUBLISHED 23 September 2022

CITATION
Dong Y, Zhou S, Xing L, Chen Y, Ren Z, Dong Y and Zhang X (2022), Deep learning methods may not outperform other machine learning methods on analyzing genomic studies.
Front. Genet. 13:992070.
doi: 10.3389/fgene.2022.992070

COPYRIGHT
© 2022 Dong, Zhou, Xing, Chen, Ren, Dong and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

KEYWORDS

deep learning, machine learning, genomic analysis, disease prediction, imbalance data, hit curve

1 Introduction

Machine Learning (ML) has been widely applied in genomic analysis and disease prediction. ML is considered an objective and reproducible method that integrates multiple quantitative variables to improve diagnostic accuracy (Wang et al., 2021). There are many successful applications. In disease prediction, Deberneh and Kim (2021) presented an ML model for predicting the T2D (type 2 diabetes) occurrence in the following year (Y+1) using variables in the current year (Y). The model's performance proved to be reasonably good at forecasting the occurrence of T2D in the Korean population. Park and Lee (2021) constructed a disease recurrence prediction model using ML techniques. Their study compared the performance of 5 ML models (decision tree, random forest, eXtreme Gradient Boosting [XGBoost], LightGBM, and Stacking models) related to recurrence prediction based on accuracy, and the Decision Tree model showed the best accuracy at 95%. In another study, Hussain et al. (2021) proposed a voting ensemble classifier with 24 features to identify the severity of chronic obstructive pulmonary disease (COPD) patients. Five ML classifiers were applied, namely random forests (RF), support vector machine (SVM), gradient boosting machine (GBM), XGBoost, and K-nearest neighbour (KNN) in their study. These classifiers were trained with a set of 24 features. After that, they combined the results with a soft voting ensemble (SVE) method. The results showed that the SVE classifier outperforms conventional ML-based methods for patients with COPD. In addition, ML-based methods for genetic analysis have also been reported in multiple studies (Rowlands et al., 2019; Placek et al., 2021), such as ML approaches for the prioritization of genomic variants impacting Pre-mRNA splicing; ML suggests the polygenic risk for cognitive dysfunction in amyotrophic lateral sclerosis and so on.

Deep Learning (DL) is a subset of ML, and it goes beyond non-deep ML by creating more complex multi-layered models to mimic how humans function. DL is known to work well in big data applications. Still, DL has been used in disease prediction primarily based on publicly available medical image data, which have common structural patterns to utilize pre-trained networks or take advantage of convolution layers. For example, Chao et al. (2021) presented a DL CVD risk prediction model, which was trained with 30,286 LDCTs from the National Lung Cancer Screening Trial. As a result, the model obtained an area under the curve (AUC) of 0.871 on a separate test set of 2085 subjects and was able to identify patients at high risk of CVD mortality (AUC of 0.768). Zhou et al. (2020) proposed a DL model to classify the HCM genotypes based on a non-enhanced four-chamber view of cine images. Lin et al. (2020) developed and validated a DL algorithm for detecting coronary artery disease (CAD) based on facial photos. Jin et al. (2021) presented a multi-task deep learning approach that allows simultaneous tumour segmentation and response prediction. Their approach to capturing dynamic information in longitudinal images may be broadly used for screening, treatment response evaluation, disease monitoring, and surveillance.

However, compared with image data, genomic data has less structure information to train a DL model. Moreover, building an accurate DL model usually requires immense amounts of data, which is often difficult to find in biological studies with a limited number of participants. Therefore, we are motivated to investigate the effectiveness of DL in genomic analysis and the amount of genomic sample size fitting for the DL model.

Our study explores and compares three non-deep ML and two DL methods in genomic analysis, including elastic net, XGBoost, SVM, long short-term memory (LSTM), and deep neural network (DNN). These methods are applied to the UK Biobank study, which includes a wide array of genotypic and phenotypic information from 502,524 participants. Coupled with the current impact of COVID-19, lung diseases have attracted widespread attention. We choose three specific lung diseases from UK Biobank, combined with SNPs and other relevant covariates to build prediction models with these five typical non-deep ML and DL algorithms. Large-scale computation works are conducted using high-performance computing servers provided by Compute Canada. To investigate how DL and non-deep ML methods perform in genomic analysis on various sample sizes, we generate random subsets of original data with 10 different levels of sample size and evaluate the prediction performance of each method using multiple metrics, including F1 score, precision, recall, and the hit curve. Besides comparing DL and non-deep ML methods, we also investigated the relation between performance change and other important factors, such as sample sizes increase and the imbalanced ratio (defined as the proportion of samples in the number of a control group to the number of case group (Sun et al., 2019).

The rest of the paper is organized as follows. Section 2 provides detailed processing and summary statistics of the dataset from UK Biobank, and five DL and non-deep ML methods are discussed in detail. In Section 3, experiment results are presented and compared. Concluding remarks are given in Section 4.

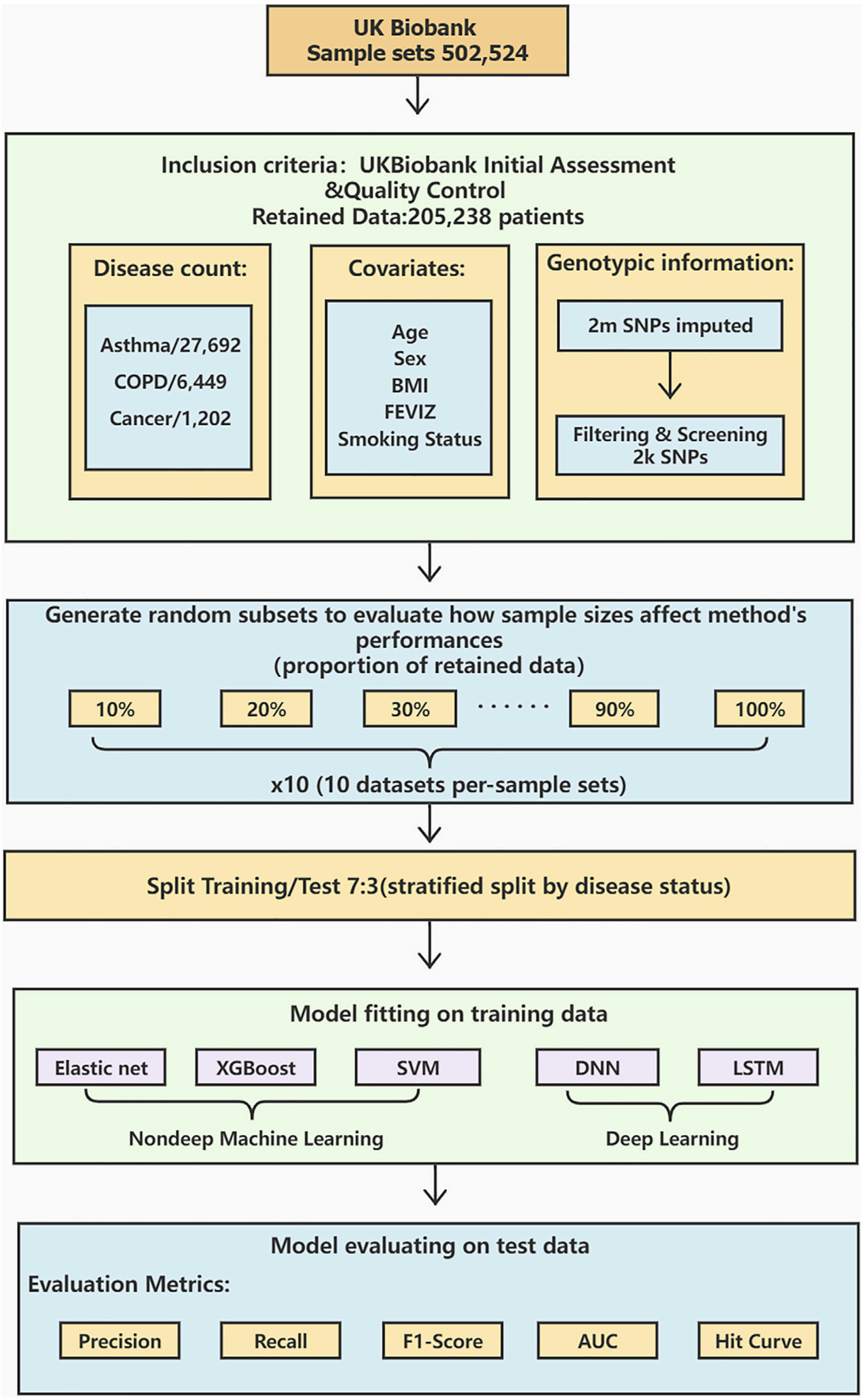

UK Biobank
Sample sets 502,524
Inclusion criteria: UKBiobank Initial Assessment
&Quality Control
Retained Data:205,238 patients
Disease count:
Asthma/27,692
COPD/6,449
Cancer/1,202
Covariates:
Age
Sex
BMI
FEVIZ
Smoking Status
Genotypic information:
2m SNPs imputed
Filtering & Screening
2k SNPs
Generate random subsets to evaluate how sample sizes affect method's performances
(proportion of retained data)
10%
20%
30%
......
90%
100%
x10 (10 datasets per-sample sets)
Split Training/Test 7:3(stratified split by disease status)
Model fitting on training data
Elastic net
XGBoost
SVM
DNN
LSTM
Nondeep Machine Learning
Deep Learning
Model evaluating on test data
Evaluation Metrics:
Precision
Recall
F1-Score
AUC
Hit Curve

FIGURE 1
A workflow diagram of the study process. We perform data preprocessing on 502,524 sample sets from UK Biobank. After the initial assessment and quality control, the data is retained for 205,238 cases with detailed procedures in. There are 27,692 asthma cases, 6,449 COPD cases, and 1,202 lung cancer cases. Age, sex, BMI, FEVIZ, and smoking status are covariates. 2,000 SNPs are retained after filtering and screening the original 2 million SNPs. The retained dataset was divided into ten subsets per-sample sets from 10 to 100%. We split the data by disease status into 70% as training and 30% as testing sets. This study uses three non-deep ML models (Elastic net, XGBoost, and SVM) and two DL models (DNN and LSTM) to construct the prediction models. Finally, the model performance is evaluated by the metrics, such as precision, recall, F1-score, AUC, and hit curve.

TABLE 1 Descriptive statistics of the dataset. This table gives the relationships between smoking status and other covariates, i.e., age, sex, BMI, FEV1Z score, asthma status, COPD status, and lung cancer status.

Covariates	Never smoked	Previously smoked	Currently smokes
Age			
< 55 years	47,137 (42.1%)	22,112 (29.3%)	8,269 (46.6%)
≥55 years	64,826 (57.9%)	53,414 (70.7%)	9,480 (53.4%)
Sex			
Male	69,300 (61.9%)	39,670 (52.5%)	8,912 (50.2%)
Female	42,663 (38.1%)	35,856 (47.5%)	8,837 (49.8%)
BMI_mean	27.00 (±4.67)	27.83 (±4.68)	26.93 (±4.65)
FEV1Z_mean	0.31 (±1.05)	0.44 (±1.10)	0.85 (±1.17)
Asthmastatus	15,110 (13.5%)	10,343 (13.7%)	2,239 (12.6%)
COPDstatus	1,350 (1.2%)	3,338 (4.4%)	1,761 (9.9%)
Cancerstatus	185 (0.17%)	627 (0.83%)	390 (2.2%)

2 Data and methods

The workflow diagram is shown in Figure 1.

2.1 Data

With the rapid spread of COVID-19, lung diseases have attracted widespread social attention. It was suggested that the presence of lung diseases, in general, may contribute to severe COVID-19 symptoms. About 600 million people have asthma, and lung cancer and COPD are the first and the third leading cause of death worldwide. Genetic variants such as single nucleotide polymorphisms (SNPs) have been focused on in lung disease research.

The dataset we use in our study is the release of the 2018 UK Biobank. The original dataset has collected a wide array of phenotypic and phenotypic information from 502,524 participants. We only select three specific lung diseases (i.e. asthma, COPD, and lung cancer), combined with participants' SNPs, sex, body mass index (BMI), age, smoking status, and Z-score of the forced expiratory volume in one second (FEV1Z).

2.1.1 Genotype and quality control procedure

Quality control and imputation were performed centrally by UK Biobank. We exclude the following participants from our analyses: 1) participants not of white British ancestry either by self-report or principal component analysis conducted by UK Biobank, 2) participants with more than 10% missing genotype data, 3) participants with putative sex-chromosome aneuploidy, 4) participants where the self-reported sex does not match the genetically-inferred sex, 5) participants that UK Biobank has flagged for having high heterozygosity/missingness and 6) participants with at least ten putative 3rd-degree relatives. Further, we remove SNPs with imputation information score <0.1, minor allele frequency <0.001, more than 5% missing genotype data, p-value $<10^{-6}$ in the Hardy-Weinberg Equilibrium test, and SNPs that fail UK Biobank quality control in at least one batch. After sample filtering and SNP screening, we are left with a sample size of 205,238 participants and 2,000 SNPs.

2.1.2 Data statistics

The average age of subjects is 56.5 years, with an age range of 40–69 and a sex ratio (females/males) of 1.35. The selected features are BMI, sex, age, Smoking status, FEV1Z, and 2,000 SNPs information. The summary of data is shown in Table 1. To explore the model performance and the prediction effect of DL and non-deep ML in the case of large and small data, we randomly generate ten subsets from 10 to 100% and repeat it ten times. The detailed subset information is shown in the Supplementary Material (Supplementary Tables S1–S4).

2.2 Methods

2.2.1 Elastic net

In general, the elastic net is the regularized linear regression method (Zou and Hastie, 2005). It is a middle ground between ridge regression and lasso regression. The penalty term is a simple mix of ridge and lasso's penalties, and the mix ratio can be controlled. The estimates from the elastic net method are defined by

$$\hat{\beta} = \arg\min_{\beta}\left(\|y - X\beta\|^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1\right), \quad (1)$$

where $\lambda_1\|\beta\|_1$ and $\lambda_2\|\beta\|_2^2$ are the L_1 norm and L_2 norm, respectively, y is the response variable vector, and X is covariates vector. The relationship between λ_1 and λ_2 can be written as

$$\lambda_1 = \alpha\lambda, \tag{2}$$

and

$$\lambda_2 = \frac{(1-\alpha)}{2}\lambda. \tag{3}$$

When the mix ratio α approaches 0, the elastic net is equivalent to ridge regression, and as the ratio α goes to 1, it is equal to lasso regression. As a result of balancing the L1 norm and L2 norm, the computational cost of the elastic net is expensive. However, it reduces the impact of different features while not eliminating all of the features to improve the model performance.

In this study, Elastic net models are implemented by R. The parameters α and λ are tuned and chosen by function cv. glmnet ().

2.2.2 XGBoost

XGBoost is an optimized distributed gradient boosting library designed to be efficient, flexible, and portable. It implements the algorithms in the Gradient Boosting framework, which integrates many weak classifiers to form a strong classifier (Ma et al., 2021). The weak classifiers compensate each other to improve the performance of the strong classifier.

Unlike the traditional integrated decision tree algorithm, XGBoost adds a regular term in the loss function to control the complexity of the model while preventing the model from overfitting. The objective function is defined by

$$F(x) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{4}$$

where $l(y_i, \hat{y}_i)$ is the model's loss function, $\Omega(f_k)$ is the regular term, n is the number of samples, and K is the number of the CART tree. After that, a second-order Taylor expansion approximation is applied to the loss function, and the objective function is optimized to approach the actual value and improve the prediction accuracy.

GridSearchCV function is used to find the optimal parameters. The parameter max_depth of the XGBoost model is set to 5. The larger the max_depth, the more specific and local samples the model learns. The min_child_weight determines the minimum sum of instance weight needed in a child, and its value is 4. The parameter subsample is 0.8, which controls the proportion of random samples for each tree. The parameter colsample_bytree is used to manage the percentage of columns sampled per randomly sampled tree (each column is a feature), and its value is 0.8. The objective parameter defines the loss function that needs to be minimized. Reg_alpha and reg_lambada are the L1 regularization terms of the weights and the L2 regularization terms of the weights, respectively. These two parameters help reduce overfitting, and their values are 60 and 2, respectively.

2.2.3 SVM

SVM is a supervised learning algorithm. The learning strategy uses supporting vectors and margins to find the optimal segmentation hyperplane to classify the data (Fan et al., 2021). SVM can be used for classification and regression analysis. As a training algorithm, SVM has a highly accurate and strong generalization ability.

This study uses the LinearSVC module in SVM. LinearSVC implements a linear classification support vector machine and can choose a variety of penalty parameters and loss functions. Normalization also works well when the number of training set instances is large.

We add the regularization term L1 norm to reduce the impact of overfitting. The parameter C of the LinearSVC model is 1.0.

2.2.4 LSTM

LSTM is a recurrent neural network (RNN). It can solve the problem of gradient disappearance and gradient explosion in traditional RNN. LSTM consists of a forget gate, an input gate, and an output gate (Elsheikh et al., 2021). The input vector and output vector of the hidden layer of LSTM are x_t and h_t, and the forward propagation process can be used in Equations 5-9.

The input gate is mainly used to control how many values of the current input will flow directly to a memory unit, defined as follows:

$$i_t = \sigma\left(W_{x_i}x_t + W_{h_i}h_{t-1} + b_i\right). \tag{5}$$

The forget gate is an essential component of the LSTM memory cell, which controls the retention and forgetting of information to avoid gradient disappearance and gradient explosion caused by the backward propagation of gradients over time. The value of the forget gate f_t and the value of the memory cell c_t are expressed as:

$$f_t = \sigma\left(W_{x_f}x_t + W_{h_f}h_{t-1} + b_f\right) \tag{6}$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tanh\left(W_{x_c}x_t + W_{h_c}h_{t-1} + b_c\right). \tag{7}$$

The role of the output gate is to effectively control the effect of a memory processing unit on the input and output values in these messages. The value of the output gate o_t and the output h_t of LSTM at moment t are expressed as:

$$o_t = \sigma\left(W_{x_o}x_t + W_{h_o}h_{t-1} + b_o\right) \tag{8}$$

$$h_t = o_t \otimes \tanh(c_t). \tag{9}$$

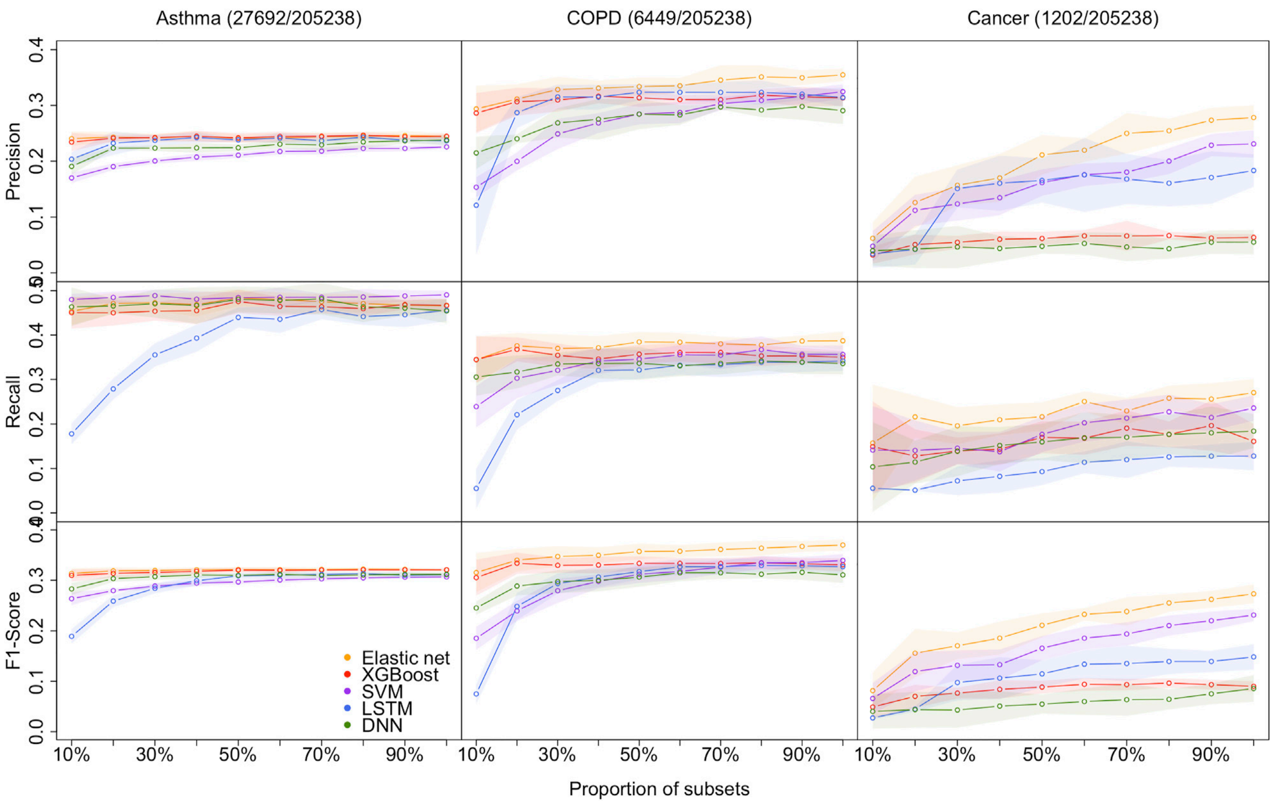


FIGURE 2
Models performance of the 5 methods with the 10 different sample size for predicting asthma, COPD, and lung cancer, respectively. Performances are shown by precision, recall, and F1-score. the shaded parts are the 1 standard error confidence bounds.

We construct a network structure with three LSTM layers and one dense layer and use sigmoid as the activation function and binary cross-entropy as the loss function. We set batch size and epoch as 128 and 100, respectively, and the learning rate is 0.0005.

2.2.5 DNN

A deep neural network (DNN) is a framework of deep learning. It is a neural network with at least one hidden layer (Ye et al., 2021), which can also be called a multi-layer perceptron.

For the DNN model, We divide it into the input layer, hidden layer, and output layer. Since we are exploring the classification and prediction of these three diseases, we choose $binary_crossentropy$ as our loss function. Secondly, we put three total connection layers into the hidden layer. The number of neurons in the hidden layer is set to 64. Each neuron in the top connection layer is fully connected with all neurons in the previous layer, which can integrate the local information with category differentiation in each layer. To improve the network performance of DNN, we applied the ReLU function to the activation function of each neuron.

Meanwhile, we found through experiments that when the batch size was set to 128, the model's accuracy could be effectively improved, and the model could converge more accurately towards the direction where the extreme value was. Moreover, when the epoch was 200 iterations, the training results tended to be stable basically. Although the model performance is improved, it is more prone to overfitting due to many parameters. Therefore, we added a regularization term L1 norm to constrain training parameters by adding a penalty norm for training parameters to the loss function to prevent model overfitting.

3 Results

In this study, five disease prediction models based on non-deep ML and DL models, i.e. elastic net, XGBoost, SVM, LSTM, and DNN, are constructed. The original dataset is randomly selected into ten sets of 10–100% datasets (shown in Supplementary Tables S1–S4). In the modelling process, we perform ten cross-validations on each set of 10–100%

datasets to find the optimal threshold for prediction and apply it to the test set. Finally, the mean and the standard deviation of the accumulated AUC, precision, recall, F1-score values, and hit curve plot are used as evaluation metrics. The detailed statistics are described in Supplementary Material (Supplementary Tables S5–S7).

As shown in Figure 2, the proposed models are evaluated by the standard metrics of precision, recall, and F1-score, and an increasing trend is generally discovered. Precision is the proportion of positive predictions that are actually correct. Recall is the proportion of actual patients identified correctly. The F1-score is the harmonic mean of precision and recall, and is often used to interpret imbalanced data. However, AUC is not sensitive to imbalanced data (its results are shown in the Supplementary Material [Supplementary Tables S5–S7])). Hence, we are more interested in precision, recall, and F1-score due to the imbalanced data structure.

3.1 Performance on small-sized datasets

Asthma status prediction. For the 10% dataset, the highest precision and F1-score are 0.2404 (±0.0127) and 0.3135 (±0.0098), respectively, obtained by the elastic net model. SVM beats other models' recall value, which is 0.4800 (±0.0126). LSTM has the lowest performance on recall and F1-score, which are 0.1780 (±0.0234) and 0.1891 (±0.0142), respectively. The lowest precision value generated from SVM is 0.1701 (±0.0092). For the 20% dataset, LSTM significantly improved recall and F1-score, but still lower than the other four models. Despite that, the performances of all models remained the same.

COPD status prediction. For the 10% dataset, elastic net models' results are better than that of other models. Its precision, recall, and F1-score are 0.2938 (±0.0415), 0.3446 (±0.0524), and 0.3153 (±0.0386), respectively. The results of XGBoost are very close to those of the elastic net model. However, LSTM has poor performance in this case, and its precision, recall, and F1-score are 0.1210 (±0.0886), 0.0550 (±0.0443), and 0.0749 (±0.0191), respectively. For the 20% dataset, the optimum values of each indicator are also derived from the elastic net. LSTM has a decent improvement in precision performance. And its precision is 0.2871 (±0.0274), while the elastic net has precision value of 0.3115 (±0.0264).

Cancer status prediction. All metrics of two DL models underperform that of three non-deep ML models for 10 and 20% datasets. The top F1-score of the two DL models is 0.0402 for the 10% dataset, which is evaluated from the DNN model, whereas the lowest F1-score from non-deep ML methods (XGBoost) is 0.0088 higher.

In summary, it is clear that on a small dataset, the performance of non-deep ML models is superior to that of DL models.

3.2 Overall model performance on DL and non-deep ML

As the size of the dataset increases, the overall model performances increase, and the gap between non-deep ML and DL decreases.

Asthma status prediction. The F1-score of elastic net, XGBoost, SVM, LSTM, and DNN for 50% dataset are 0.3214 (±0.0047), 0.3201 (±0.0047), 0.2966 (±0.0043), 0.3088 (±0.0060), and 0.3098 (±0.0032), respectively. As the data volume rises to 100%, the performances of the five models do not change a lot.

COPD status prediction. When the dataset size increases to 50%, LSTM improves its performance rapidly. The F1-score of LSTM has grown three times from 0.0749 (±0.0191) to 0.3171 (±0.0154). When the dataset size expands from 50 to 100%, the optimal F1-score is 0.3699 (±0.0110) from the elastic net. The F1-scores of XGBoost, SVM, LSTM and DNN become 0.3307 (±0.0130), 0.3394 (±0.0125), 0.3269 (±0.0145), and 0.3106 (±0.0157), respectively.

Cancer status prediction. On 50% of the dataset, the performance of all five models has improved. As the dataset grows to 100%, all models' performances are still climbing up.

In summary, DL models do not outperform non-deep ML models, even in extensive data with over 200k samples. The performance of all models improves when the sample size increases. The performance differences between DL and non-deep ML decrease as the sample size of data increases.

3.3 Impact of imbalanced data structure

In this study, the datasets are imbalanced, and the imbalanced rates (Control/Case) for asthma, COPD, and lung cancer are 6.5:1, 30.8:1, and 169.6:1, respectively. Model performances on cancer prediction are the lowest since the cancer dataset structure is highly imbalanced. For example, the F1-score of DNN for the 50% dataset is 0.3098 (±0.0032) for predicting asthma status, whereas it is 0.0547 (±0.0187) for predicting cancer status. Moreover, as the imbalanced rate increases, the confidence bands are getting wider. For instance, the width of the confidence band of XGBoost's F1-score for the 100% dataset is 0.0058 for predicting asthma; in contrast, it is 0.0202 for predicting lung cancer.

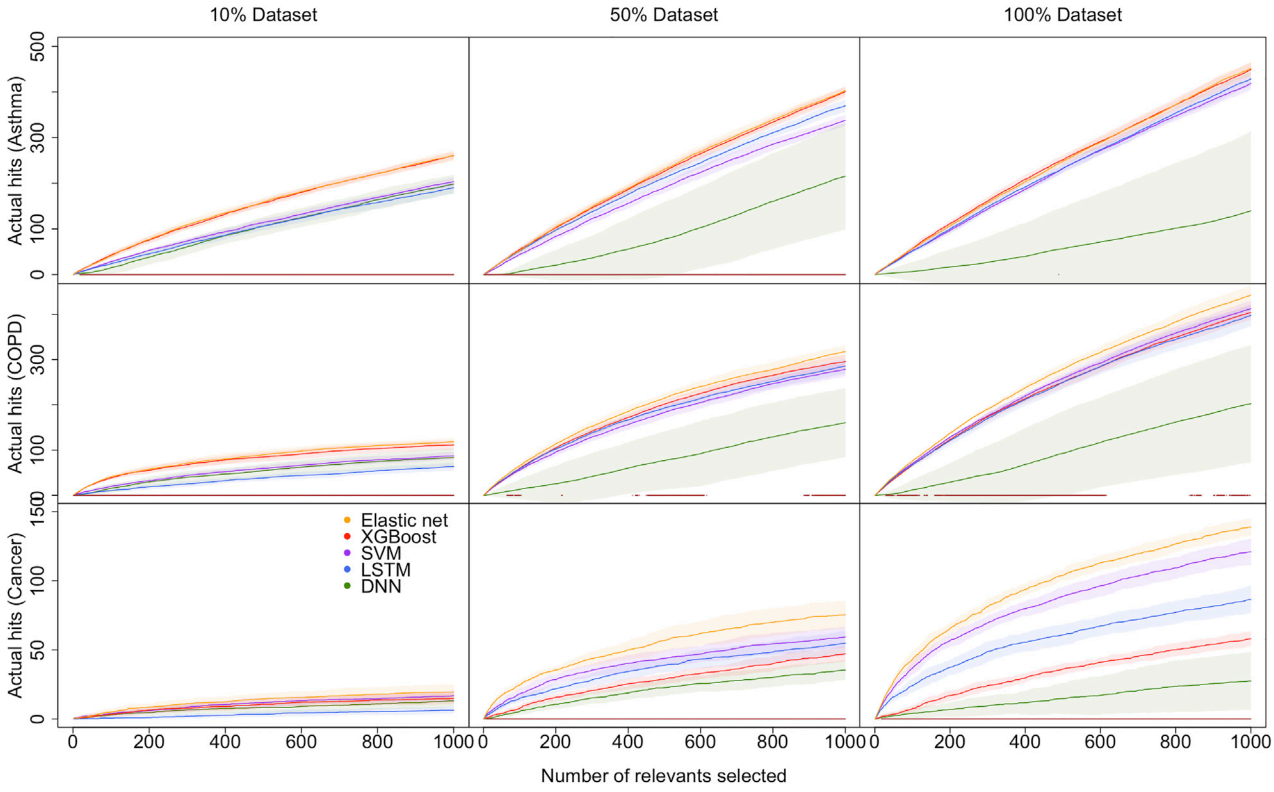


FIGURE 3
Hit curve graphs of AsthmaStatus, COPDStatus and CancerStatus classification by five models on 10–100% data sets. The x-axis represents the number of test subjects we selected by sorting the estimated probability up to down. The y-axis of the hit curve chart represents the number of subjects with certain conditions which are correctly diagnosed in the test set. The point (m_1, m_2) indicates there are m_2 patients in the first m_1 selected subjects are correctly predicted as diseased. The curves show the average hit curves of five models, and the shaded area denotes the confidence bounds constructed using 10-fold cross-validation (i.e. ± one standard error). The brown bar at the bottom means non-deep ML models are significantly better than DL models.

3.4 Promote hit curve as a particular visual tool

To summarize all the metric results we have found, a hit curve is promoted as a particular visual tool to compare the prediction models. In a biomedical study, it is impossible for a prediction model to accurately predict all cases, and a model can be effective without necessarily accurately predicting all cases. For example, in our research, a prediction model is considered to be doing an excellent job if it can choose a relatively small number of subjects, and correctly label the majority of the condition group. Therefore, hit curve is used to prioritize case. In this situation, cases with the largest prediction probabilities are chosen first. As we select cases according to the prediction probabilities, a “hit” occurs whenever the case is a success (people we selected are in a certain disease condition). Say we choose m_1 subjects and m_2 are diseased, and we can visually assess a prediction model by plotting m_2 against m_1, a so-called hit curve. A good prediction model will have m_2 increasing rapidly with m_1, as shown in Figure 3 (only the hit curve plots for 10, 50, and 100% of the dataset are shown here, and the result plots for the remaining percentage of the dataset are visible in the Supplementary Material [Supplementary Figures S1–S30]).

The elastic net curve and XGBoost curve are nearly identical, but they cross over each other at some points and are significantly higher than the others in predicting Asthma and COPD. For lung cancer condition prediction, XGBoost does not maintain a good performance. However, elastic net and SVM models are still superior to the LSTM model. DNN model is inferior to other models in all cases. Therefore, evidence supports that DL models often cannot overperform non-deep ML models. The brown bar appears in the 10 and 50% datasets on predicting asthma conditions. However, there are no brown bars in the 100% dataset plot. It implies the performance gap between DL and non-deep ML decreases as the sample size increases. And the difference will be trivial when the data sample size is as large as the biobank level. However, it is difficult to obtain such a large dataset. Hence, DL models often underperform non-deep ML models.

With lung cancer data's highly imbalanced data structure, none of the five models perform well when the data sample size is small. Their shaded areas are relatively broad, making the difference hard to tell. As the data sample size increases, their hit curves increase with different slopes. As a consequence, the performance differences become substantial. In other words, imbalanced data is also called weighted data. The effective sample size of a weighted data is smaller than its original sample size. It is almost impossible to evaluate those five models' performances due to a lack of a sufficient sample size. As the effective sample size gradually increases, the model performance differences become apparent. However, if the effective sample size reaches a certain large amount, the differences among all models are not significant again.

4 Discussion

This study evaluated the potential of DL models (DNN and LSTM) in predicting asthma, COPD, and lung cancer with various sample sizes from the UK Biobank dataset, compared with non-deep ML models (elastic net, XGBoost and SVM). Besides the most popular performance metrics, such as the F1-score, the hit curve, as a particular visual tool, is promoted to describe the performance of predicting rare events. The results suggest that we should not apply DL methods in most genomic studies, unless we have data with biobank-level sample sizes. We conclude not recommending standard deep learning methods for genomic studies based on the following two facts we observed in our study. First, the prediction performances of non-deep machine learning methods vastly outperform deep learning methods in small datasets (e.g., 10 and 20% random subsets of UK Biobank). Second, we observed that deep learning could not outperform non-deep methods in huge data like the entire cohort of UK Biobank (500k participants), although increasing sample size leads to the improvement of the deep learning method's performance, and its improvement is faster than non-deep methods. Therefore, we need more data than UK Biobank to prefer deep learning methods. However, the sample sizes of most publicly available genomic data cannot meet this requirement, which range from tens to few thousands.

Although deep learning methods achieved outstanding performance in image, video, and natural language analysis, we found their performance is not attractive in analyzing genomic studies. We believe this is the result of two characteristics of genomic data: 1) genomic data typically has small sample sizes to fit a complex network; and 2) genomic data lacks common structural patterns like images to use pre-trained networks or take advantage of convolution layers.

Besides comparing deep learning methods with non-deep methods, the following are other important messages we learned from this study and would like to share with the audience.

We found that cancer status is much harder to predict than the other two diseases. The results show that the uneven data structure also affects the model's performance. The control/case ratio is 6.5:1 for asthma, 30.8:1 for COPD, and 169.6:1 for lung cancer, respectively. We notice that all three disease conditions are imbalanced, and the imbalanced ratio of lung cancer conditions is particularly extreme, which leads to model overfitting and underperforming prediction. Therefore, the imbalanced rate between cases and controls is also a critical influencing factor. Although we operate by regulation, rare events are harder to predict. We would do the data augmentation to prevent the imbalance problem in the future.

Our predictions of disease status are based on genomic information but not the specific diagnostic tests of related diseases. Therefore, we don't expect high accuracy in the predictions. This prediction aims to segment the patients by their predicted risk of conditions and manage them differently (e.g., following up with a different visit frequency or using follow-up disease-specific diagnostic tests).

There are two types of classification mistakes: 1) incorrectly labeling a patient as low-risk or healthy; and 2) incorrectly labeling a healthy individual as a patient or high-risk. In our case, the first type of mistake is much more harmful than the second type. Follow-up diagnosis can fix the second mistake. The first mistake may cause a delay in treatment, while the timing of treatment can be the most critical factor in treating diseases like cancer. These two types of mistakes can be summarised by precision and recall, respectively. The most popular metric, F1-score, is the harmonic average of precision and recall, which regards these two prediction mistakes as costing equally. Fn-score can weigh two types of mistakes using user-defined weights. However, it is not easy to define weights objectively. Hence, we introduced our preferred metric, the hit curve, for rare event detection, which focus on detecting true positive rate. Different points on the curve correspond to different decision rules about who should be labelled as patients. Users can compare many decision rules between the two methods using their hit curves. Users can also use this visual tool to decide which decision rule is best (subjectively).

Data availability statement

The codes are available at (https://github.com/hebutdy/Evaluation-on-GS). The datasets for this study can be found in the UK Biobank (https://www.ukbiobank.ac.uk/).

Author contributions

YaD and XZ conceptualized this study and designed the experiment. YoD, SZ, ZR, and YC did the experiments and preparation of the first draft. All authors have developed drafts of the manuscript and approved the final draft of the manuscript. XZ supervised this project.

Funding

This work was supported by the Natural Sciences and Engineering Research Council Discovery Grants (RGPIN-2017-04722 PI:XZ, RGPIN-2021-03530 PI:LX), the Canada Research Chair (950-231363, PI:XZ), Chinese State Scholarship Fund (202108130108), Chinese National Key Research and Development Program (2019YFC1904601), Science, Technology Research Project of Higher Education in Hebei Province of China (QN2021213).

Acknowledgments

This research was enabled in part by support provided by WestGrid [https://www.westgrid.ca/] and Compute Canada [https://www.computecanada.ca/].

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/fgene.2022.992070/full#supplementary-material

References

Chao, H., Shan, H., Homayounieh, F., Singh, R., Khera, R., Guo, H., et al. (2021). Deep learning predicts cardiovascular disease risks from lung cancer screening low dose computed tomography. *Nat. Commun.* 12, 2963. doi:10.1038/s41467-021-23235-4

Deberneh, H. M., and Kim, I. (2021). Prediction of type 2 diabetes based on machine learning algorithm. *Int. J. Environ. Res. Public Health* 18, 3317. doi:10.3390/ijerph18063317

Elsheikh, A. H., Saba, A. I., Elaziz, M. A., Lu, S., Shanmugan, S., Muthuramalingam, T., et al. (2021). Deep learning-based forecasting model for Covid-19 outbreak in Saudi Arabia. *Process Saf. Environ. Prot.* 149, 223–233. doi:10.1016/j.psep.2020.10.048

Fan, S., Zhao, Z., Zhang, Y., Yu, H., Zheng, C., Huang, X., et al. (2021). Probability calibration-based prediction of recurrence rate in patients with diffuse large b-cell lymphoma. *BioData Min.* 14, 38. doi:10.1186/s13040-021-00272-9

Hussain, A., Choi, H.-E., Kim, H.-J., Aich, S., Saqlain, M., and Kim, H.-C. (2021). Forecast the exacerbation in patients of chronic obstructive pulmonary disease with clinical indicators using machine learning techniques. *Diagnostics* 11, 829. doi:10.3390/diagnostics11050829

Jin, C., Yu, H., Ke, J., Ding, P., Yi, Y., Jiang, X., et al. (2021). Predicting treatment response from longitudinal images using multi-task deep learning. *Nat. Commun.* 12, 1851. doi:10.1038/s41467-021-22188-y

Lin, S., Li, Z., Fu, B., Chen, S., Li, X., Wang, Y., et al. (2020). Feasibility of using deep learning to detect coronary artery disease based on facial photo. *Eur. Heart J.* 41, 4400–4411. doi:10.1093/eurheartj/ehaa640

Ma, B., Yan, G., Chai, B., and Hou, X. (2021). Xgblc: An improved survival prediction model based on XGBoost. *Bioinformatics* 38, 410–418. doi:10.1093/bioinformatics/btab675

Park, Y. M., and Lee, B.-J. (2021). Machine learning-based prediction model using clinico-pathologic factors for papillary thyroid carcinoma recurrence. *Sci. Rep.* 11, 4948. doi:10.1038/s41598-021-84504-2

Placek, K., Benatar, M., Wuu, J., Rampersaud, E., Hennessy, L., Van Deerlin, V. M., et al. (2021). Machine learning suggests polygenic risk for cognitive dysfunction in amyotrophic lateral sclerosis. *EMBO Mol. Med.* 13, e12595. doi:10.15252/emmm.202012595

Rowlands, C. F., Baralle, D., and Ellingford, J. M. (2019). Machine learning approaches for the prioritization of genomic variants impacting pre-mrna splicing. *Cells* 8, 1513. doi:10.3390/cells8121513

Sun, Y., Milne, S., Jaw, J. E., Yang, C., Xu, F., Li, X., et al. (2019). Bmi is associated with fev1 decline in chronic obstructive pulmonary disease: A meta-analysis of clinical trials. *Respir. Res.* 20, 236. doi:10.1186/s12931-019-1209-5

Wang, G., Zhang, Y., Li, S., Zhang, J., Jiang, D., Li, X., et al. (2021). A machine learning-based prediction model for cardiovascular risk in women with preeclampsia. *Front. Cardiovasc. Med.* 8, 736491. doi:10.3389/fcvm.2021.736491

Ye, J., Wang, S., Yang, X., and Xianjun, T. (2021). Gene prediction of aging-related diseases based on dnn and mashup. *BMC Bioinforma.* 22, 597. doi:10.1186/s12859-021-04518-5

Zhou, H., Li, L., Liu, Z., Zhao, K., Chen, X., Lu, M., et al. (2020). Deep learning algorithm to improve hypertrophic cardiomyopathy mutation prediction using cardiac cine images. *Eur. Radiol.* 31, 3931–3940. doi:10.1007/s00330-020-07454-9

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 67, 301–320. doi:10.1111/j.1467-9868.2005.00503.x

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 28 September 2022
DOI 10.3389/fgene.2022.991842

Check for updates

OPEN ACCESS

EDITED BY
Marco Pellegrini,
National Research Council (CNR), Italy

REVIEWED BY
Manuela Petti,
Sapienza University of Rome, Italy
Teng Mao,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Kei Hang Katie Chan,
katie.kh.chan@cityu.edu.hk

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 12 July 2022
ACCEPTED 12 September 2022
PUBLISHED 28 September 2022

CITATION
Bennett AN, Huang RX, He Q, Lee NP, Sung W-K and Chan KHK (2022), Drug repositioning for esophageal squamous cell carcinoma.
Front. Genet. 13:991842.
doi: 10.3389/fgene.2022.991842

COPYRIGHT
© 2022 Bennett, Huang, He, Lee, Sung and Chan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Drug repositioning for esophageal squamous cell carcinoma

Adam N. Bennett[1], Rui Xuan Huang[2], Qian He[3], Nikki P. Lee[4], Wing-Kin Sung[5] and Kei Hang Katie Chan[2,3,6]*

[1]Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Kowloon, Hong Kong SAR, China, [2]Department of Electrical Engineering, City University of Hong Kong, Hong Kong, Hong Kong SAR, China, [3]Department of Biomedical Sciences, City University of Hong Kong, Hong Kong, Hong Kong SAR, China, [4]Department of Surgery, The University of Hong Kong, Pokfulam, Hong Kong SAR, China, [5]Department of Computer Sciences, National University of Singapore, Singapore, Singapore, [6]Department of Epidemiology, Centre for Global Cardiometabolic Health, Brown University, Providence, RI, United States

Esophageal cancer (EC) remains a significant challenge globally, having the 8th highest incidence and 6th highest mortality worldwide. Esophageal squamous cell carcinoma (ESCC) is the most common form of EC in Asia. Crucially, more than 90% of EC cases in China are ESCC. The high mortality rate of EC is likely due to the limited number of effective therapeutic options. To increase patient survival, novel therapeutic strategies for EC patients must be devised. Unfortunately, the development of novel drugs also presents its own significant challenges as most novel drugs do not make it to market due to lack of efficacy or safety concerns. A more time and cost-effective strategy is to identify existing drugs, that have already been approved for treatment of other diseases, which can be repurposed to treat EC patients, with drug repositioning. This can be achieved by comparing the gene expression profiles of disease-states with the effect on gene-expression by a given drug. In our analysis, we used previously published microarray data and identified 167 differentially expressed genes (DEGs). Using weighted key driver analysis, 39 key driver genes were then identified. These driver genes were then used in Overlap Analysis and Network Analysis in Pharmomics. By extracting drugs common to both analyses, 24 drugs are predicted to demonstrate therapeutic effect in EC patients. Several of which have already been shown to demonstrate a therapeutic effect in EC, most notably Doxorubicin, which is commonly used to treat EC patients, and Ixazomib, which was recently shown to induce apoptosis and supress growth of EC cell lines. Additionally, our analysis predicts multiple psychiatric drugs, including Venlafaxine, as repositioned drugs. This is in line with recent research which suggests that psychiatric drugs should be investigated for use in gastrointestinal cancers such as EC. Our study shows that a drug repositioning approach is a feasible strategy for identifying novel ESCC therapies and can also improve the understanding of the mechanisms underlying the drug targets.

KEYWORDS

drug repositioning, drug repurposing, esophageal squamous cell carcinoma, ESCC treatment, cancer biology

Introduction

There are two major subtypes of Esophageal cancer (EC), esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma (EAC) (Ye et al., 2021). In China, more than 90% of esophageal cancer cases are ESCC (Zhang X. et al., 2021). EC as a whole remains a significant challenge globally, having the 8th highest incidence and the 6th highest mortality worldwide killing over 500,000 people in 2020 (Sung et al., 2021). A major driver of the high mortality rate is likely due to the fact that there are very few effective therapeutic options for EC patients. In recent years, there has been a significant increase in survival for many cancers, largely due to the availability of targeted therapies. For EC, however, targeted therapies are yet to make a significant impact on patient survival. Consequently, patients are often relying on more traditional therapies such as chemotherapy and surgical resection. In-order-to increase patient survival, novel therapeutic strategies for EC patients must be devised. Unfortunately, the development of novel drugs also presents its own significant challenges as most novel drugs do not make it to market due to lack of efficacy or safety concerns. Therefore, it is more time and cost effective to identify existing drugs, that have already been approved for treatment of other diseases, which can be repurposed to treat EC patients. This can be achieved using a drugs gene signature, the alterations in gene expression as a result of exposure to the drug. The gene signature of a drug indicates the underlying biological pathways and mechanisms that are involved in the therapeutic effect of the drug. With this knowledge, we can then identify candidate drugs which have gene signatures capable of reversing aberrant gene expression patterns observed in disease-states to those observed in normal cells. This gene signature-based approach has been adopted by previous research to identify drugs that can be repositioned to treat a variety of diseases including, but not limited to, cancer, Alzheimer's, hyperlipidaemia, hypertension, and inflammatory disease (Corbett et al., 2012; Hall et al., 2014; Guney et al., 2016; Subramanian et al., 2017; Cheng et al., 2018; Carvalho et al., 2021; Wu et al., 2022). To date, drug repositioning to target gene signatures has primarily involved identifying directly overlapping drug genes and disease genes (herein referred to as overlap analysis) (Subramanian et al., 2017; Wang et al., 2018; Chen et al., 2022). More recently, network analysis has been greatly employed in this area as it offers distinct advantages over more traditional statistical methods. This is due to the fact that the models that can be built with this methodology are an excellent way to capture a molecules relationship with other molecules. In particular, nodes can be used to represent multiple entities such as genes, molecules, proteins, etc, and the edges can also represent a vast array of information such as mode-of-actions (MoAs), underlying mechanisms, or functional similarities (Jarada et al., 2020) Hence, network-based methods can accurately represent the biological mechanisms which are driving diseases (Barabási et al., 2011). As a result, network-based drug repositioning can identify drugs which target the underlying biology of the disease. It is worth noting, however, that other methods of computational drug repositioning have also been adopted, such as Data Mining and Machine Learning. An excellent review of the different methodologies, as well as their advantages and disadvantages has recently been published (Jarada et al., 2020). Due to the success of drug repositioning overall, and the absence of effective treatments for ESCC, it has been proposed that this method be used to identify novel treatment strategies for ESCC. However,

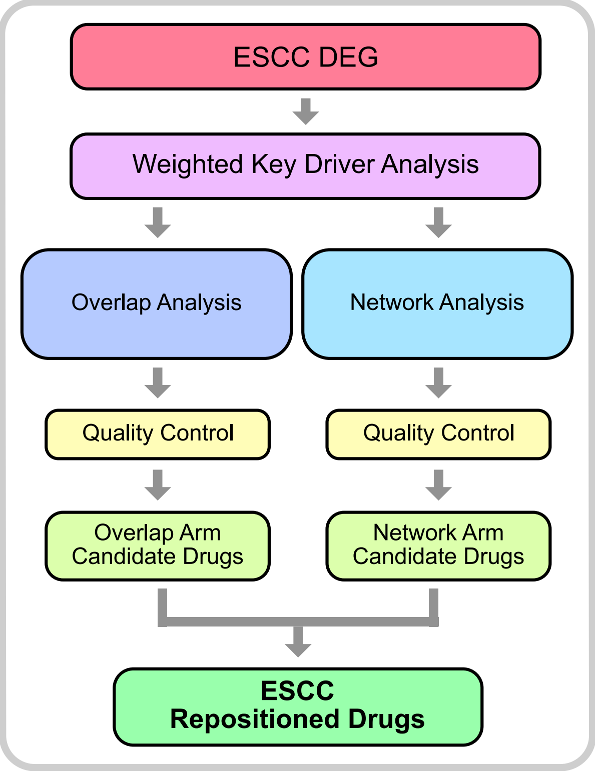


FIGURE 1
Drug Repositioning Analysis Methodology. The initial step of the analysis included differential gene expression on previously published array data from GEO (accession: GSE23400). DEGs were then used to identify key driver genes in a weighted key driver analysis. The key driver genes were then used as input in 2 arms; overlap analysis and network analysis. Qualify control was performed to filter out erroneous results and identify candidate drugs. Drugs which were common to both arms were considered robust and considered ESCC Repositioned Drugs.

TABLE 1 Top 25 up-regulated genes in differential gene expression analysis comparing cancer tissue with adjacent tissue in ESCC patients.

Gene	logFC	Adj. P-value
MMP1	4.443	8.65×10^{-29}
SPP1	3.187	3.38×10^{-23}
POSTN	3.066	2.03×10^{-22}
COL1A1	2.990	5.89×10^{-32}
JUP	2.831	1.68×10^{-16}
COL1A2	2.698	2.95×10^{-26}
COL11A1	2.405	1.64×10^{-20}
CDH11	2.370	1.14×10^{-21}
MMP12	2.240	5.17×10^{-19}
MAGEA6	2.226	2.40×10^{-09}
PTHLH	2.213	7.27×10^{-13}
MAGEA3	2.213	1.71×10^{-09}
VCAN	2.204	2.11×10^{-20}
SNAI2	2.202	2.62×10^{-25}
MMP10	2.193	8.12×10^{-11}
COL3A1	2.164	7.24×10^{-22}
SULF1	2.125	1.69×10^{-22}
ECT2	2.112	2.30×10^{-31}
COL5A2	2.087	1.59×10^{-20}
TOP2A	2.004	2.93×10^{-23}
PLAU	1.994	4.17×10^{-27}
CKS2	1.968	1.90×10^{-22}
INHBA	1.904	2.28×10^{-15}
ISG15	1.870	8.29×10^{-14}
CEP55	1.846	5.48×10^{-26}

TABLE 2 Top 25 down-regulated genes in differential gene expression analysis comparing cancer tissue with adjacent tissue in ESCC patients.

Gene	logFC	Adj. P-value
CRISP3	−4.247	8.53×10^{-21}
MAL	−3.968	8.65×10^{-20}
CRNN	−3.654	3.31×10^{-16}
SCEL	−3.496	4.16×10^{-17}
CLCA4	−3.425	2.59×10^{-18}
TGM3	−3.329	5.45×10^{-19}
CRCT1	−3.175	2.76×10^{-15}
TMPRSS11E	−3.106	3.69×10^{-15}
SLURP1	−2.952	1.03×10^{-17}
CLIC3	−2.913	7.72×10^{-17}
ENDOU	−2.774	3.52×10^{-21}
IL1RN	−2.769	2.06×10^{-22}
PPP1R3C	−2.750	5.02×10^{-24}
SPINK5	−2.745	8.77×10^{-17}
HPGD	−2.647	5.22×10^{-24}
RHCG	−2.628	7.00×10^{-13}
KRT4	−2.606	5.59×10^{-14}
FLG	−2.432	2.72×10^{-15}
KLK13	−2.353	1.73×10^{-20}
ECM1	−2.351	8.47×10^{-17}
KRT13	−2.305	3.20×10^{-10}
CEACAM6	−2.291	8.44×10^{-13}
ADH1B	−2.288	3.47×10^{-20}
PSCA	−2.260	$2.25 \times 10–{}^{15}$
HOPX	−2.233	7.07×10^{-15}

these studies have largely, though not completely, been limited to testing existing cancer drugs *in vitro* with drug screening methods (Xie et al., 2020; Li Y. et al., 2021). Herein, we adopt both a network-based and overlap-based drug repositioning methodology, to identify existing drugs that can specifically target the aberrant expression profile of ESCC and impede oncogenesis. To do this, we used previously published data for *in-silico* drug repositioning analysis utilising the PharmOmics webserver (Chen et al., 2022). The repositioning analysis consisted of two arms, the 'overlap analysis' arm and the 'network analysis' arm (Figure 1), which utilise two methods of drug repositioning.

Results

Identification of DEGs

The dataset GSE23400 was downloaded using the GEOquery R package function getGEO. In total, 167 DEGs were identified between ESC and normal samples (Details of the differential gene expression analysis can be found in the methods section). Of which, 65 were upregulated and 102 were downregulated. The top 5 most upregulated genes are *MMP1, SPP1, POSTN, COL1A1,* and *JUP*. The top 5 most downregulated genes are *CRISP3*, *MAL*, *CRNN*, *SCEL*, *CLCA4*. The top 25 up-regulated genes can be observed in Table 1, whereas the top 25 down-regulated genes can be observed in Table 2.

Functional and pathway enrichment analyses

Functional and pathway enrichment analyses were performed used the 'clusterProfiler' R package. Gene Set Enrichment Analysis (GSEA) was performed with Gene Ontology (GO) (hereafter referred to as GSEA-GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway (hereafter referred to as GSEA-KEGG). GSEA-GO analysis was performed with the gene set categories Biological Process (BP), Cellular Component (CC), and Molecular Function (MF), which identified 253, 36, 25 enriched gene

TABLE 3 ESCC repositioned drugs.

Drug	Study	z-score	Jaccard score	Odds ratio	Adj. P-value	Within species rank
Erlotinib	*In Vitro*	−8.806362317	1.59×10^{-2}	2.16×10^{1}	5.66×10^{-5}	0.956
Palbociclib	*In Vitro*	−8.090451972	1.56×10^{-2}	2.11×10^{1}	6.18×10^{-5}	0.953
Doxorubicin	*In Vitro*	−7.851164741	3.41×10^{-2}	5.35×10^{1}	5.13×10^{-9}	0.993
Methotrexate	PharmOmics meta	−7.504929239	1.35×10^{-2}	1.83×10^{1}	7.80×10^{-4}	0.930
Crizotinib	*In Vitro*	−7.50277149	2.08×10^{-2}	2.95×10^{1}	1.61×10^{-6}	0.980
Vinblastine	PharmOmics meta	−6.871294272	5.14×10^{-2}	9.04×10^{1}	2.10×10^{-17}	0.998
Gemcitabine	*In Vitro*	−5.585120295	2.43×10^{-2}	3.58×10^{1}	3.91×10^{-10}	0.987
Daunorubicin	*In Vitro*	−5.155171903	2.94×10^{-2}	4.53×10^{1}	2.07×10^{-7}	0.991
Venlafaxine	*In Vitro*	−5.053770712	1.53×10^{-2}	2.07×10^{1}	6.74×10^{-5}	0.950
Ethanol	PharmOmics meta	−4.264304125	2.33×10^{-2}	3.41×10^{1}	5.61×10^{-10}	0.985
Tamoxifen	PharmOmics meta	−4.072907051	1.79×10^{-2}	2.51×10^{1}	3.24×10^{-5}	0.969
Arsenic trioxide	PharmOmics meta	−3.980019706	4.67×10^{-2}	7.95×10^{1}	2.10×10^{-11}	0.997
Dasatinib	*In Vitro*	−3.747277559	2.08×10^{-2}	2.95×10^{1}	1.61×10^{-6}	0.980
Ixazomib	PharmOmics meta	−3.730099165	5.73×10^{-2}	1.07×10^{2}	5.33×10^{-21}	0.999
Penicillamine	PharmOmics meta	−3.248848376	4.13×10^{-2}	6.75×10^{1}	1.53×10^{-13}	0.996
Nefazodone	*In Vitro*	−3.176922914	1.15×10^{-2}	1.51×10^{1}	1.35×10^{-3}	0.893
Leflunomide	PharmOmics meta	−2.888272698	4.65×10^{-2}	7.91×10^{1}	1.83×10^{-15}	0.997
Fulvestrant	*In Vitro*	−2.792994137	2.35×10^{-2}	3.41×10^{1}	8.07×10^{-7}	0.985
Azithromycin	*In Vitro*	−2.53558291	2.79×10^{-2}	4.16×10^{1}	2.17×10^{-8}	0.990
Hydrocortisone	PharmOmics meta	−2.37861135	3.20×10^{-2}	4.89×10^{1}	5.37×10^{-10}	0.992
Etanercept	PharmOmics meta	−2.333538798	1.50×10^{-2}	2.32×10^{1}	3.88×10^{-3}	0.948
Acetaminophen	PharmOmics meta	−2.196753074	3.65×10^{-2}	5.90×10^{1}	2.93×10^{-14}	0.994
Lapatinib	*In Vitro*	−2.141804897	2.63×10^{-2}	3.93×10^{1}	4.09×10^{-7}	0.989
Niacin	PharmOmics meta	−2.092485943	2.25×10^{-2}	3.24×10^{1}	1.03×10^{-6}	0.983
Anastrozole	PharmOmics meta	−2.073331977	3.75×10^{-2}	6.08×10^{1}	2.19×10^{-14}	0.994

sets, respectively. Numerous BP gene sets identified by the analysis are related to extracellular matrix and cell differentiation. Ranking BP analysis by adjusted p-value, the top 5 most enriched gene sets are cellular component organization, cellular component organization or biogenesis, extracellular matrix organization, extracellular structure organization, multicellular organism development. According to adjusted p-value, the top 5 most enriched CC category are endoplasmic reticulum lumen, external encapsulating structure, extracellular matrix, fibrillar collagen trimer, banded collagen fibril. Furthermore, the top 5 categories in the MF analysis identified extracellular matrix structural constituent, extracellular matrix structural constituent conferring tensile strength, protein-containing complex binding, cell adhesion molecule binding, glycosaminoglycan binding. The full results for BP, CC, MF can be observed in Supplementary Tables S1–S3, respectively. Gene Set Enrichment Analysis of KEGG (GSEA-KEGG) identified 12 enriched gene sets (Supplementary Table S4), including those previously identified as ESCC-related, such as Focal adhesion, ECM-receptor interaction, PI3K-Akt signalling pathway.

Weighted key driver analysis

Weighted Key Driver Analysis (wKDA) was performed using Mergeomics webserver. In this analysis, genes which possess a local network neighbourhood that have a significant enrichment of genes that are ESCC-associated are considered key drivers (KDs) (Ding et al., 2021). The analysis identified 89 key driver genes which were then filtered to select those which possessed an FDR <0.05, ensuring that only the strongest KDs are used in subsequent analyses. This resulted in 39 key driver genes (Supplementary Table S5). The top 10 key driver genes are *NCAPG*, *PLG*, *NUSAP1*, *COL17A1*, *ASPM*, *TOP2A*, *ITGB3*, *P4HB*, *TTK*, and *COL7A1*.

Repositioned drugs

Drug repositioning analysis was performed using both the Overlap Drug Repositioning and the Network Drug Repositioning modules from PharmOmics (Chen et al., 2022). The potential drugs from the analysis, were then filtered to

TABLE 4 Current use of ESCC Repositioned Drugs.

Drug	Standard treatment for ESCC/Clinical trial	Clinical trial remarks	Reference
Erlotinib	Yes	Limited activity in EC overall but response was observed in ESCC (Only 2/13 participants were ESCC)	Ilson et al. (2011)
		Promising results if combined with radiotherapy	Zhao et al. (2016)
Palbociclib	Yes	Not promising result in clinic trials. However, authors claim that the drug could be useful in combination with other drugs	Karasic et al. (2020)
Doxorubicin	Yes	Used successfully in combination with other drugs (cisplatin and fluorouracil combination therapy)	Honda et al. (2010)
Methotrexate	Yes	Used for palliative care in combination with other drugs	DUSI et al. (2020)
Crizotinib	No	—	—
Vinblastine	Yes	Phase 2 Clinical Trial - Promising results	Conroy et al. (1996)
Gemcitabine	Yes	Phase 1 Clinical Trial - Promising results	Oettle et al. (2002)
Daunorubicin	No	—	—
Venlafaxine	No	—	—
Ethanol	Yes	Used for palliative care. Evidence of use for unresectable in case report with combination with chemotherapy	Wadleigh et al. (2006)
Tamoxifen	No	—	—
Arsenic trioxide	No	—	—
Dasatinib	No	—	—
Ixazomib	No	—	—
Penicillamine	No	—	—
Nefazodone	No	—	—
Leflunomide	No	—	—
Fulvestrant	No	—	—
Azithromycin	No	—	—
Hydrocortisone	No	—	—
Etanercept	No	—	—
Acetaminophen	No	—	—
Lapatinib	No	—	—
Niacin	No	—	—
Anastrozole	No	—	—

identify robust ESCC repositioned drugs. The repositioning analysis identified 25 drugs that are strong candidates for ESCC treatment (Table 3). The top 10 repositioned drugs are Erlotinib, Palbociclib, Doxorubicin, Methotrexate, Crizotinib, Vinblastine, Gemcitabine, Daunorubicin, Venlafaxine, and Ethanol. We predicted that drugs which interact with EGFR (Erlotinib, Crizotinib, and Lapatinib), estrogen signalling (Tamoxifen, Fulvestrant, Hydrocortisone, and Anastrozole) and TRAIL-mediated apoptosis (Azithromycin and Anastrozole) pathways have potential for treating ESCC.

Drug validation

To validate our findings, we performed a literature search to determine whether any of the drugs identified by our analysis are currently used in ESCC treatment (Table 4). We found that 7 of the top 10 repositioned drugs, according to z-score, are already used to treat ESCC or have been shown to demonstrate efficacy in clinical trials. Candidate drugs were then validated using Binding DB (Gilson et al., 2016). Each drug was searched in the database to ascertain whether they bind to proteins known to be involved in ESCC. We found that 21 out of 25 repositioned drugs have a strong binding affinity to proteins that have been associated with ESCC in some manner previously (Table 5). Additionally, we performed a literature search to assess whether there is any biological evidence (*in vitro* or *in vivo*) that demonstrates efficacy or establishes a plausible mechanism by which the novel repositioned drugs could be beneficial for ESCC patients (Table 6). We found that all of our novel ESCC drugs, except for Venlafaxine, target pathways or proteins which have been demonstrated to drive oncogenesis in several cancers, including ESCC. Therefore, these drugs should be able to target the underlying biological processes driving oncogenesis

TABLE 5 Binding DB Target Validation. Repositioned drugs were investigated using Binding DB to determine whether the proteins that the drugs have strong affinity to have been previously shown to be associated with ESCC.

Drug	Protein binding in homo sapiens	Binding protein ESCC-Associated	Reference
Erlotinib	Epidermal growth factor receptor (EGFR)	Yes	Kashyap and Abdel-Rahman, (2018)
Palbociclib	CDK9	Yes	Tong et al. (2017)
	CDK1	Yes	Zhang et al. (2021a)
	CDK2	Yes	Zhou et al. (2021)
	CDK4	Yes	Huang et al. (2021)
Doxorubicin	Androgen Receptor	Yes	Sukocheva et al. (2015)
Methotrexate	Dihydrofolate reductase	Yes - Indirectly through MDM2	Maguire et al. (2008)
	MMP7	Yes	Tanioka et al. (2003)
Crizotinib	Epidermal growth factor receptor (EGFR)	Yes	Kashyap and Abdel-Rahman, (2018)
	FLT3	Yes	Zhu et al. (2021)
Vinblastine	—	—	—
Gemcitabine	Equilibrative nucleoside transporter 1	Yes - Indirectly through mIR-1269	Xie et al. (2022)
Daunorubicin	Multidrug resistance protein 1	Yes	Zhang et al. (2016)
Venlafaxine	Sodium-dependent dopamine transporter	Yes	Guo et al. (2018)
Ethanol	—	—	—
Tamoxifen	17-beta-hydroxysteroid dehydrogenase type 3	No	—
Arsenic trioxide	—	—	—
Dasatinib	Tyrosine- and threonine-specific cdc2-inhibitory kinase	Yes (and also via CDK1)	Zhang et al. (2019)
Ixazomib	Proteasome component C5	No	—
Penicillamine	Bile salt export pump	Yes	Bernstein et al. (2009)
Nefazodone	Alpha-1A adrenergic receptor	Yes	Zhang et al. (2018)
	5-hydroxytryptamine receptor 2A	Yes	Wei et al. (2022)
Leflunomide	matrix metalloproteinase 1	Yes	Pang et al. (2016)
	Dihydroorotate dehydrogenase	Yes	Qian et al. (2020)
Fulvestrant	Estrogen receptor	Yes	Zhang et al. (2017)
Azithromycin	Cytochrome P450 3A4	Yes	Bergheim et al. (2007)
Hydrocortisone	Corticosteroid-binding globulin (SERPINA6)	Yes	Ma et al. (2019)
Etanercept	—	—	—
Acetaminophen	Carbonic anhydrase 12	Yes	Ochi et al. (2015)
	Dipeptidyl peptidase 3	Yes	Liu et al. (2021)
Lapatinib	Epidermal growth factor receptor (EGFR)	Yes	Kashyap and Abdel-Rahman, (2018)
	Receptor tyrosine-protein kinase erbB-2 (HER2 or ERBB2)	Yes	Rong et al. (2020)
Niacin	Hydroxycarboxylic acid receptor 2	No	—
	Xanthine dehydrogenase/oxidase	Yes	Li et al. (2021a)
Anastrozole	Cytochrome P450 19A1	Yes	Bergheim et al. (2007)

in ESCC and inhibit proliferation and/or initiate apoptosis in ESCC.

Discussion

ESCC is one of the most common malignancies and possess a significant mortality rate worldwide. This is largely due to late diagnosis and scarcity of efficacious treatment strategies upon being diagnosed (Feng et al., 2021). To address this, we performed a disease-based drug repositioning analysis with previously published ESCC gene expression data from paired patient samples. Differential gene expression analysis data identified 167 differentially expressed genes (DEGs) which were then used in wKDA and identified 39 key driver genes (KDGs). The genes with the highest absolute logFC identified by our differential gene expression analysis are *MMP1*, *CRISP3*, *MAL*, *CRNN*, *SCEL*. The most upregulated gene, *MMP1*, encodes a protein involved in the breakdown of the extracellular matrix (ECM) by cleaving collagens and other molecules. The most downregulated gene, *CRISP3*, encodes a protein located in the ECM and

TABLE 6 Potential mechanism of action for novel drugs.

Drug	Potential mechanism of action	Additional remarks	Citation(s)
Crizotinib	Protein kinase inhibitor (inc. HGFR)	Acts as an inhibitor against anaplastic lymphoma kinase. Crizotinib is an inhibitor of c-Met and could be used to target HGF pathway	Digklia and Voutsadakis, (2013)
Daunorubicin	Intercalates with DNA and interrupts cell proliferation	—	Niaki et al. (2020)
Venlafaxine	—	Has been used for managing hot flashes during breast cancer therapy	Biglia et al. (2005)
Tamoxifen	Selective estrogen receptor modulator (SERM)/partial agonist of ER	Evidence of efficacy in cell and animal models. Preliminary evidence in adenocarcinoma of enhancing chemo therapy effect	Due et al., (2016); Huang et al., 2019; Wang et al., 2020
Arsenic trioxide	Induces programmed cell death	Evidence of DNA damage-mediated cyclin D1 degradation in ESCC cell lines	Zhu et al. (2020)
Dasatinib	Tyrosine kinase inhibitor	Dasatinib increases ESCC cell lines sensitivity to cisplatin	Chen et al. (2015)
Ixazomib	Inhibits the protein proteasome subunit beta type-5 (PSMB5)	Supresses proliferation in Esophageal squamous cell carcinoma in cell lines through c-Myc/NOXA pathway. *In vivo* evidence of efficacy in non-small cell lung cancer	Chattopadhyay et al., 2015; Wang et al., 2021b
Penicillamine	Radio-chemo-sensitisation involving H_2O_2-mediated oxidative stress	Enhances breast and lung cancer response to radiation and carboplatin via H_2O_2-mediated oxidative stress	Sciegienka et al. (2017)
Nefazodone	Disrupts mitochondrial function	Demonstrates anticancer properties in multiple cell lines	Varalda et al. (2020)
Leflunomide	Dihydroorotase dehydrogenase (DHODH) and/or Tyrosine kinase inhibition	Potential anticancer drug through disruption of pyrimidine synthesis and EGFR signalling. *In vitro* and *in vivo* evidence for inducing apoptosis in neuroblastoma	Zhu et al., 2013; Zhang and Chu, (2018)
Fulvestrant	Estrogen receptor antagonist	Results in complete inhibition of estrogen signalling through the ER	Nathan and Schmid, (2017)
Azithromycin	Apoptosis induction via TRAIL	Efficacy *in vitro* and *in vivo* in colon cancer by TRAIL autophagy	Qiao et al. (2018)
Hydrocortisone	Binds glucocorticoid receptor to inhibit inflammatory transcription factors	Evidence to suggest BRCA1 downregulation in breast cancer	Antonova and Mueller, (2008)
Etanercept	Tumour necrosis factor (TNF) inhibitor	Prolonged disease stabilisation was observed in EC used in combination with chemotherapy	Monk et al., 2006; Shirmohammadi et al., 2020
Acetaminophen	Apoptosis induction	Promising results used in combination with chemotherapy in lung cancer	Lee et al. (2019)
Lapatinib	tyrosine kinase inhibitor/EGFR/HER1 and HER2 receptors	ESCC cell and patient-derived xenograft model	HOU et al., 2013; Saito et al., 2015
Niacin	Modulation of NAD + levels	Evidence of TRAIL mediated autophagy in colon cancer	Kim et al. (2015)
Anastrozole	Aromatase Inhibition	Has been used with Anti-Fibroblast growth factor receptor 1 (FGFR1) drug in breast cancer. Evidence that FGFR1 can be used as a independent prognosis marker in ESCC and anti-FGFR1 decreases proliferation via MEK-ERK downstream pathways	Milani et al., 2009; Chen et al., 2017

thought to be involved in cellular matrix remodelling (Ribeiro et al., 2011). The wKDA identified 39 significant driver genes for ESCC. Amongst the top 10 most significant KDGs, *NUSAP1*, *COL17A1*, *ITGB3* and *COL7A1* are involved in ECM maintenance. For example, the 4th most significant key driver gene, *COL17A1*, encodes a protein involved in cell-matrix adhesion (Jones et al., 2020). Taken together, these results suggest that alterations in ECM are an important driver of ESCC oncogenesis (Chen et al., 2016). Furthermore, KEGG analysis found both Focal adhesion and ECM-receptor interaction to be the 3rd and 4th most enriched term, respectively. This is in line with previous research that indicates higher levels of Serum human relaxin 2 (H2 RLN), a protein involved in ECM, collagen, and matrix metalloproteinase is associated with worse prognosis, including higher clinical stage and poorer survival (Ren et al., 2013; Napier et al., 2014).

Using 39 KDGs in an ESCC drug repositioning analysis, we identified 25 drugs that are predicted to have therapeutic effect in ESCC. Of which, 7 are either currently used in the clinic or have been used in clinical trials and 2 have shown efficacy *in vitro* or *in vivo*. Importantly, those which have been used in clinical trials have demonstrated efficacy particularly when used in combination with other drugs, such as chemotherapy. This is not surprising, however, as combination therapy has long been a standard practice in cancer therapy, including for ESCC where the current first-line treatment regimen is a combination of 5-fluorouracil and cisplatin (Hiramoto et al., 2018; Hirano and Kato, 2019). Each repositioned drug was validated *in-silico* using the drug binding database BindingDB, to identify which drug targets have previously been associated with ESCC (Table 5). Significantly, 21 of the 25 repositioned drugs have targets that have previously been associated with ESCC in some manner, which demonstrate the robustness of our findings. To further validate our findings, we performed a literature search on the novel repositioned drugs to examine whether there is an underlying biological mechanism which would justify the drugs appearance in the results (Table 6). We found that almost all of the repositioned drugs have been shown to demonstrate anti-cancer effects in multiple cancers, most notably breast cancers and non-small cell lung carcinoma (NSCLC). Interestingly, many of the repositioned drugs target specific pathways; EGFR (Erlotinib, Crizotinib, and Lapatinib), estrogen signalling (Tamoxifen, Fulvestrant, Hydrocortisone, and Anastrozole) and TRAIL-mediated apoptosis (Azithromycin and Anastrozole) pathways, suggesting that these pathways are key drivers of ESCC. This is in line with previous research which identified the EGFR AND ER pathways as drivers of ESCC oncogenesis and metastasis and have also been associated with patient outcome (Maron et al., 2020). Crucially, some of these drugs have been shown to have therapeutic potential *in vitro*. For example, Lapatinib, which acts through EGFR and HER2 has been shown to be efficacious in ESCC patient-derived xenografts (Rong et al., 2020). The potential mechanisms by which novel drugs identified by our study can be observed in Table 6. It is also worth noting that there are 2 anti-depressants present in our results, Venlafaxine and Nefazodone. These results are particularly interesting as it has recently been shown that psychiatric drugs offer potential as anti-cancer therapeutics (Loehr et al., 2021). Moreover, a recent review has specifically addressed the need to investigate psychiatric drugs for treatment of gastrointestinal cancers (Avendaño-Félix et al., 2020). We hypothesise that Crizotinib, Lapatinib, and Dasatinib are amongst the drugs with the most potential. Particularly Crizotinib and Lapatinib are of note as they target the EGFR pathway which is already targeted in ESCC treatment with Erlotinib. Dasatinib also has high potential due to targeting Tyrosine- and threonine-specific cdc2-inhibitory kinase and CDK1, proteins known to be involved in ESCC, and also due to displaying efficacy in cell lines (Chen et al., 2015; Zhang et al., 2019).

There are several limitations to our study, however, most notably that due to limited data availability, the sample size of patient samples was relatively small. We were unable to stratify patients according to subtype of ESCC. This means that the analysis is focussed on ESCC as whole and does not take into consideration specific subtypes. Moreover, as multiple drugs identified in our study are more efficacious when in combination with another drug, it would be beneficial to know what other drugs should be used in combination with the novel therapeutics identified. However, this analysis does not predict drug combinations that would be effective in treating ESCC.

On the other hand, our study has several strengths. To our knowledge, this is the first study to adopt a primarily computational approach to perform drug repositioning analysis in ESCC. Particularly, there are studies that have a computational component, but they do not use patient samples to identify drugs based on network analysis of differentially expressed genes (Li et al., 2022). Moreover, this is also the first to adopt Pharmomics unique network analysis to perform the analysis on ESCC. Furthermore, as Pharmomics contains >18000 species/tissue-specific gene signatures for 941 drugs and chemicals, it provides a larger scope of potential drugs compared to other studies in ESCC. Another strength of the study is that we used a two armed approach to ensure robust findings as each repositioning methodology has its own strengths. The overlap-based repositioning allows us to identify drugs which target the KDGs whereas the network-based repositioning allows for insights into the molecular and mechanistic therapeutic effects of the drugs. As this specific form of network-based repositioning is unique to PharmOmics, our study can provide valuable insights into the underlying molecular mechanisms driving ESCC. Another strength of our study is the consistency of our results with previously published literature. DEGs which displayed the highest absolute logFC were consistent with previously published literature including *MMP1*, *SPP1*, *COL1A2*, and *COL1A1* amongst the top upregulated genes and *CRISPR3*, *MAL*, *TMPRSS11E*, and *CRNN* amongst the top downregulated genes (Feng et al., 2021; Song et al., 2021). Indeed, the gene with the highest absolute logFC, *MMP1*, is already known to be associated with ESCC oncogenesis (Chen et al., 2016). Additionally, higher *MMP1* is associated with poorer prognosis (Feng et al., 2021). Moreover, the most significant key driver genes (KDGs) identified by our wKDA are consistent with previously published studies (Li et al., 2020; Yu-jing et al., 2020; Wang M. et al., 2021). Significantly, multiple drugs identified by our analysis target key pathways known to be involved in ESCC oncogenesis and metastasis.

Conclusion

Herein we utilised *in silico* disease-based drug repositioning to identify novel therapeutics for esophageal squamous cell carcinoma. Amongst 25 potential repositioned drugs identified in our study, 9 are currently used in the clinic or have shown promising results in clinical trials in combination with other treatments. Crucially, we identified 16 novel therapeutic strategies which possess a strong biological rationale for use in ESCC patients. Our study shows that drug repositioning approach is a feasible strategy in ESCC therapies and can improve the understanding of the mechanisms of the drug targets.

Materials and methodology

Data acquisition and identification of DEGs

Dataset was acquired from the Gene Expression Omnibus under accession code GSE23400 using the GEOquery R package function getGEO. The dataset consists of 53 paired patient samples from Esophageal squamous cell carcinoma (ESCC) patients. Additionally, 14,335 genes were in the dataset. Differentially expressed genes (DEGs) between paired tumour and non-tumour samples were identified using the limma package. The log-fold change (logFC) was calculated for DEGs. Genes with absolute logFC >1.5 and adjusted p-value < 0.01 were considered significant and used in subsequent analyses.

Functional and pathway enrichment analyses

The DEGs identified above were analysed using the clusterProfiler R package in order to identify biological annotations from the Gene ontology (GO) functional enrichment and Kyoto Encyclopaedia of Genes and Genomes (KEGG). The GO analysis was performed for biological process (BP), cellular component (CC) and molecular function (MF). An adjusted p-value < 0.05 was considered as statically significant for all analyses.

Weighted key driver analysis

Weighted key driver analysis (wKDA) was performed on DEGs using Mergeomics webserver to identify key driver genes (KDGs). wKDA has higher accuracy than standard key driver analysis as it considers edge weight information. The network used in the analysis was STRING PPI Network and default parameters were used (Search depth of 1, Undirected Edges, Min Hub Overlap of 0.33, and edge factor of 0.0). Genes which had an FDR <0.05 were considered as significant KDGs and used in subsequent analyses.

Drug repositioning analysis

By the Pharmomics webserver, Drug Repositioning Analyses was performed using genes obtained from the wKDA analysis. The analysis consisted of two arms: the Overlap Drug Repositioning and ADR Analysis (Overlap-DR) arm and the Meta-Signature Network Drug Repositioning and ADR Analysis (Meta-Net-DR) arm. The network analysis adopted by Pharmomics uses a network proximity measure between drug DEGs and disease-related genes that has been adopted previously for protein-network-based analysis. Specifically, tissue-specific Bayesian gene regulatory networks (BNs) are used and then the mean shortest distance between drug DEGs and disease genes are tested. Hence, it combines species and tissue specific *in vivo* drug signatures with gene networks to identify connections between disease genes and known drug targets. On the other hand, overlap analysis adopted by Pharmomics is largely similar to that which has been adopted previously, and assesses direct overlap between input genes and drug gene signatures. To do so, the Jaccard score, gene overlap fold enrichment, and Fisher's exact test p values as measures of direct gene overlap are calculated. This analysis is based upon the premise that if disease and drug signatures target similar pathways then they would more than likely have gene overlaps and/or connect extensively in a gene network. Meta-Signature Network Drug Repositioning and ADR Analysis was performed using the multi-tissue network. In Overlap-DR, Jaccard score was used to measure the similarity between the 39 KDG's gene networks and the drug target gene networks. In Meta-Net-DR, the connectivity of the gene network between drug signatures from PharmOmics and the KDs is used. The z-score of each drug is calculated which represents the distance between the KD network and the PharmOmics drug network. The smaller the z-score, the closer the distance between the networks. The output from these analyses were considered as possible repositioned drugs and were then filtered in Drug Candidate Selection to identify ESCC repositioned drugs.

Drug candidate selection

Repositioned drugs from both Pharmomics analyses were used as candidates to identify potential drugs for ESCC. Candidate drugs from Overlap-DR results were filtered to keep drugs with an adjusted P-value < 0.05, species equal to *Homo sapiens,* and a within species rank >0. The mean Jaccard score was then calculated and drugs with a Jaccard score less than the mean were removed. Subsequently, the drugs were sorted according to 'Drug Name', 'Within Species Rank', 'Jaccard Score', and 'P-value' and duplicate drugs were removed, keeping only the highest-ranking occurrence of each drug. Candidate drugs from Meta-Net-DR were filtered to keep drugs with adjusted p-value < 0.05. Candidate drugs were then sorted according to 'Drug Name' and 'Rank' and then duplicate drugs were removed, keeping only the highest-ranking occurrence of each drug. The filtered results from Overlap-DR and Meta-Net-DR were then compared to extract candidate drugs common to both arms of analysis. Drugs common to both arms were considered ESCC Repositioned Drugs. 'Study', 'Jaccard Score', 'Odds Ratio', 'Adj. P-Value', 'Within Species Rank' data from the Overlap-DR analysis and 'z-score' from Meta-Net-DR was used to construct the final ESCC Repositioned Drugs table. ESCC

Repositioned Drugs were then sorted according to z-score (Table 3).

Drug candidate validation

In order to validate the repositioned drugs that were identified by the analysis, we performed a literature search to ascertain whether the drugs have previously been used in ESCC treatment (Table 4). Each drug was then investigated using the drug binding database Binding DB. For each ESCC Repositioned Drug, we identified which proteins they display a high binding affinity to. We then performed a literature search on these proteins, using Google Scholar and PubMed, to ascertain whether or not they have previously been shown to be ESCC-related *in vitro* or *in vivo* (Table 5). Finally, we performed a literature search on novel drugs identified by our analysis to elucidate the underlying biological processes and causal mechanisms which would explain why it is predicted to have therapeutic utility (Table 6).

Data availability Statement

Publicly available datasets were analyzed in this study. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Ethics statement

The studies involving human participants were reviewed and approved by Human Subjects Ethics Sub-Committee at City University of Hong Kong (Reference number 2-11-201810_02). The patients/participants provided their written informed consent to participate in this study.

Author contributions

NL, W-KS, and KC envisioned and directed the project, assisted in writing manuscript, and performed data interpretation. AB and RH developed the analytical pipeline and interpreted data. AB performed the analysis and wrote the manuscript. QH assisted in analytical pipeline development and contributed towards the manuscript. KC coordinated the development of the analysis and revised the manuscript.

Funding

This project was funded by The Jockey Club College of Veterinary Medicine and Life Sciences (JCC) Interdisciplinary PhD Programme at City University of Hong Kong in Collaboration with Cornell University. This work was supported by the City University of Hong Kong New Research Initiatives/Infrastructure Support from Central (APRC; grant number 9610401).

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/fgene.2022.991842/full#supplementary-material

SUPPLEMENTARY FIGURE S1
This figure shows the enriched gene set categories for the Gene Set Enrichment Analysis for Biological Processes.

SUPPLEMENTARY FIGURE S2
This figure shows the enriched gene set categories for the Gene Set Enrichment Analysis for Cellular Component.

SUPPLEMENTARY FIGURE S3
This figure shows the enriched gene set categories for the Gene Set Enrichment Analysis for Molecular Function.

References

Antonova, L., and Mueller, C. R. (2008). Hydrocortisone down-regulates the tumor suppressor gene BRCA1 in mammary cells: A possible molecular link between stress and breast cancer. *Genes Chromosom. Cancer* 47, 341–352. doi:10.1002/gcc.20538

Avendaño-Félix, M., Aguilar-Medina, M., Bermudez, M., Lizárraga-Verdugo, E., López-Camarillo, C., and Ramos-Payán, R. (2020). Refocusing the use of psychiatric drugs for treatment of gastrointestinal cancers. *Front. Oncol.* 10, 1452. doi:10.3389/fonc.2020.01452

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68. doi:10.1038/nrg2918

Bergheim, I., Wolfgarten, E., Bollschweiler, E., Hölscher, A.-H., Bode, C., and Parlesak, A. (2007). Cytochrome P450 levels are altered in patients with esophageal squamous-cell carcinoma. *World J. Gastroenterol.* 13, 997–1002. doi:10.3748/wjg.v13.i7.997

Bernstein, H., Bernstein, C., Payne, C. M., and Dvorak, K. (2009). Bile acids as endogenous etiologic agents in gastrointestinal cancer. *World J. Gastroenterol.* 15, 3329–3340. doi:10.3748/wjg.15.3329

Biglia, N., Torta, R., Roagna, R., Maggiorotto, F., Cacciari, F., Ponzone, R., et al. (2005). Evaluation of low-dose venlafaxine hydrochloride for the therapy of hot flushes in breast cancer survivors. *Maturitas* 52, 78–85. doi:10.1016/j.maturitas.2005.01.001

Carvalho, R. F., Canto, L. M. do, Cury, S. S., Hansen, T. F., Jensen, L. H., and Rogatto, S. R. (2021). Drug repositioning based on the reversal of gene expression signatures identifies TOP2A as a therapeutic target for rectal cancer. *Cancers* 13, 5492. doi:10.3390/cancers13215492

Chattopadhyay, N., Berger, A. J., Koenig, E., Bannerman, B., Garnsey, J., Bernard, H., et al. (2015). KRAS genotype correlates with proteasome inhibitor Ixazomib activity in preclinical *in vivo* models of colon and non-small cell lung cancer: Potential role of tumor metabolism. *Plos One* 10, e0144825. doi:10.1371/journal.pone.0144825

Chen, B., Liu, S., Gan, L., Wang, J., Hu, B., Xu, H., et al. (2017). FGFR1 signaling potentiates tumor growth and predicts poor prognosis in esophageal squamous cell carcinoma patients. *Cancer Biol. Ther.* 19, 76–86. doi:10.1080/15384047.2017.1394541

Chen, J., Lan, T., Zhang, W., Dong, L., Kang, N., Fu, M., et al. (2015). Dasatinib enhances cisplatin sensitivity in human esophageal squamous cell carcinoma (ESCC) cells via suppression of PI3K/AKT and Stat3 pathways. *Arch. Biochem. Biophys.* 575, 38–45. doi:10.1016/j.abb.2014.11.008

Chen, Y.-K., Tung, C.-W., Lee, J.-Y., Hung, Y.-C., Lee, C.-H., Chou, S.-H., et al. (2016). Plasma matrix metalloproteinase 1 improves the detection and survival prediction of esophageal squamous cell carcinoma. *Sci. Rep.* 6, 30057. doi:10.1038/srep30057

Chen, Y.-W., Diamante, G., Ding, J., Nghiem, T. X., Yang, J., Ha, S.-M., et al. (2022). PharmOmics: A species- and tissue-specific drug signature database and gene-network-based drug repositioning tool. *Iscience* 25, 104052. doi:10.1016/j.isci.2022.104052

Cheng, F., Desai, R. J., Handy, D. E., Wang, R., Schneeweiss, S., Barabási, A.-L., et al. (2018). Network-based approach to prediction and population-based validation of *in silico* drug repurposing. *Nat. Commun.* 9, 2691. doi:10.1038/s41467-018-05116-5

Conroy, T., Etienne, P. L., Adenis, A., Wagener, D. J., Paillot, B., François, E., et al. (1996). Phase II trial of vinorelbine in metastatic squamous cell esophageal carcinoma. European organization for research and treatment of cancer gastrointestinal treat cancer cooperative group. *J. Clin. Oncol.* 14, 164–170. doi:10.1200/jco.1996.14.1.164

Corbett, A., Pickett, J., Burns, A., Corcoran, J., Dunnett, S. B., Edison, P., et al. (2012). Drug repositioning for Alzheimer's disease. *Nat. Rev. Drug Discov.* 11, 833–846. doi:10.1038/nrd3869

Digklia, A., and Voutsadakis, I. A. (2013). Targeted treatments for metastatic esophageal squamous cell cancer. *World J. Gastrointest. Oncol.* 5, 88–96. doi:10.4251/wjgo.v5.i5.88

Ding, J., Blencowe, M., Nghiem, T., Ha, S., Chen, Y.-W., Li, G., et al. (2021). Mergeomics 2.0: A web server for multi-omics data integration to elucidate disease networks and predict therapeutics. *Nucleic Acids Res.* 49, W375–W387. doi:10.1093/nar/gkab405

Due, S. L., Watson, D. I., Bastian, I., Ding, G. Q., Sukocheva, O. A., Astill, D. St. J., et al. (2016). Tamoxifen enhances the cytotoxicity of conventional chemotherapy in esophageal adenocarcinoma cells. *Surg. Oncol.* 25, 269–277. doi:10.1016/j.suronc.2016.05.029

Dusi, V. S., C, O. R., Attili, S. V., and Palanki, S. D. (2020). Gefitinb along with methotrexate as palliative therapy in PS 3 and above in metastatic esophagus squamous cell carcinoma with focus on Q-TWIST. *J. Clin. Oncol.* 38, 355. doi:10.1200/jco.2020.38.4_suppl.355

Feng, Z., Qu, J., Liu, X., Liang, J., Li, Y., Jiang, J., et al. (2021). Integrated bioinformatics analysis of differentially expressed genes and immune cell infiltration characteristics in Esophageal Squamous cell carcinoma. *Sci. Rep.* 11, 16696. doi:10.1038/s41598-021-96274-y

Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J. (2016). BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* 44, D1045–D1053. doi:10.1093/nar/gkv1072

Guney, E., Menche, J., Vidal, M., and Barábasi, A.-L. (2016). Network-based *in silico* drug efficacy screening. *Nat. Commun.* 7, 10331. doi:10.1038/ncomms10331

Guo, J., Huang, J., Zhou, Y., Zhou, Y., Yu, L., Li, H., et al. (2018). Germline and somatic variations influence the somatic mutational signatures of esophageal squamous cell carcinomas in a Chinese population. *Bmc Genomics* 19, 538. doi:10.1186/s12864-018-4906-4

Hall, C. J., Wicker, S. M., Chien, A.-T., Tromp, A., Lawrence, L. M., Sun, X., et al. (2014). Repositioning drugs for inflammatory disease – fishing for new anti-inflammatory agents. *Dis. Model. Mech.* 7, 1069–1081. doi:10.1242/dmm.016873

Hiramoto, S., Kato, K., Shoji, H., Okita, N., Takashima, A., Honma, Y., et al. (2018). A retrospective analysis of 5-fluorouracil plus cisplatin as first-line chemotherapy in the recent treatment strategy for patients with metastatic or recurrent esophageal squamous cell carcinoma. *Int. J. Clin. Oncol.* 23, 466–472. doi:10.1007/s10147-018-1239-x

Hirano, H., and Kato, K. (2019). Systemic treatment of advanced esophageal squamous cell carcinoma: Chemotherapy, molecular-targeting therapy and immunotherapy. *Jpn. J. Clin. Oncol.* 49, 412–420. doi:10.1093/jjco/hyz034

Honda, M., Miura, A., Izumi, Y., Kato, T., Ryotokuji, T., Monma, K., et al. (2010). Doxorubicin, cisplatin, and fluorouracil combination therapy for metastatic esophageal squamous cell carcinoma. *Dis. Esophagus* 23, 641–645. doi:10.1111/j.1442-2050.2010.01070.x

Hou, W., Qin, X., Zhu, X., Fei, M., Liu, P., Liu, L., et al. (2013). Lapatinib inhibits the growth of esophageal squamous cell carcinoma and synergistically interacts with 5-fluorouracil in patient-derived xenograft models. *Oncol. Rep.* 30, 707–714. doi:10.3892/or.2013.2500

Huang, C.-M., Huang, C.-S., Hsu, T.-N., Huang, M.-S., Fong, I.-H., Lee, W.-H., et al. (2019). Disruption of cancer metabolic SREBP1/miR-142-5p suppresses epithelial–mesenchymal transition and stemness in esophageal carcinoma. *Cells* 9, 7. doi:10.3390/cells9010007

Huang, J., Wang, X., Zhang, X., Chen, W., Luan, L., Song, Q., et al. (2021). CDK4 amplification in esophageal squamous cell carcinoma associated with better patient outcome. *Front. Genet.* 12, 616110. doi:10.3389/fgene.2021.616110

Ilson, D. H., Kelsen, D., Shah, M., Schwartz, G., Levine, D. A., Boyd, J., et al. (2011). A phase 2 trial of erlotinib in patients with previously treated squamous cell and adenocarcinoma of the esophagus. *Cancer* 117, 1409–1414. doi:10.1002/cncr.25602

Jarada, T. N., Rokne, J. G., and Alhajj, R. (2020). A review of computational drug repositioning: Strategies, approaches, opportunities, challenges, and directions. *J. Cheminform.* 12, 46. doi:10.1186/s13321-020-00450-7

Jones, V. A., Patel, P. M., Gibson, F. T., Cordova, A., and Amber, K. T. (2020). The role of collagen XVII in cancer: Squamous cell carcinoma and beyond. *Front. Oncol.* 10, 352. doi:10.3389/fonc.2020.00352

Karasic, T. B., O'Hara, M. H., Teitelbaum, U. R., Damjanov, N., Giantonio, B. J., d'Entremont, T. S., et al. (2020). Phase II trial of Palbociclib in patients with advanced esophageal or gastric cancer. *Oncologist* 25, e1864–e1868. doi:10.1634/theoncologist.2020-0681

Kashyap, M. K., and Abdel-Rahman, O. (2018). Expression, regulation and targeting of receptor tyrosine kinases in esophageal squamous cell carcinoma. *Mol. Cancer* 17, 54. doi:10.1186/s12943-018-0790-4

Kim, S.-W., Lee, J.-H., Moon, J.-H., Nazim, U. M. D., Lee, Y.-J., Seol, J.-W., et al. (2015). Niacin alleviates TRAIL-mediated colon cancer cell death via autophagy flux activation. *Oncotarget* 7, 4356–4368. doi:10.18632/oncotarget.5374

Lee, S.-M., Park, J. S., and Kim, K.-S. (2019). Improving combination cancer therapy by acetaminophen and romidepsin in non-small cell lung cancer cells. *Biomed. Sci. Lett.* 25, 293–301. doi:10.15616/bsl.2019.25.4.293

Li, J., Xie, Y., Wang, X., Jiang, C., Yuan, X., Zhang, A., et al. (2020). Identification of hub genes associated with esophageal cancer progression using bioinformatics analysis. *Oncol. Lett.* 20, 214. doi:10.3892/ol.2020.12077

Li, X., Zhao, L., Wei, M., Lv, J., Sun, Y., Shen, X., et al. (2021a). Serum metabolomics analysis for the progression of esophageal squamous cell carcinoma. *J. Cancer* 12, 3190–3197. doi:10.7150/jca.54429

Li, Y., Xu, F., Chen, F., Chen, Y., Ge, D., Zhang, S., et al. (2021b). Transcriptomics based multi-dimensional characterization and drug screen in esophageal squamous cell carcinoma. *Ebiomedicine* 70, 103510. doi:10.1016/j.ebiom.2021.103510

Li, Z., Zou, L., Xiao, Z.-X., and Yang, J. (2022). Transcriptome-based drug repositioning identifies TPCA-1 as a potential selective inhibitor of esophagus squamous carcinoma cell viability. *Int. J. Mol. Med.* 49, 75. doi:10.3892/ijmm.2022.5131

Liu, J.-K., Abudula, A., Yang, H.-T., Xu, L.-X., Bai, G., Tulahong, A., et al. (2021). DPP3 expression promotes cell proliferation and migration *in vitro* and tumor growth *in vivo* that associates with poor prognosis of esophageal carcinoma. doi:10.21203/rs.3.rs-1078211/v1

Loehr, A. R., Pierpont, T. M., Gelsleichter, E., Galang, A. M. D., Fernandez, I. R., Moore, E. S., et al. (2021). Targeting cancer stem cells with differentiation agents as an alternative to genotoxic chemotherapy for the treatment of malignant testicular germ cell tumors. *Cancers* 13, 2045. doi:10.3390/cancers13092045

Ma, X.-L., Yao, H., Wang, X., Wei, Y., Cao, L.-Y., Zhang, Q., et al. (2019). ILK predicts the efficacy of chemoradiotherapy and the prognosis of patients with esophageal squamous cell carcinoma. *Oncol. Lett.* 18, 4114–4125. doi:10.3892/ol.2019.10768

Maguire, M., Nield, P. C., Devling, T., Jenkins, R. E., Park, B. K., Polański, R., et al. (2008). MDM2 regulates dihydrofolate reductase activity through monoubiquitination. *Cancer Res.* 68, 3232–3242. doi:10.1158/0008-5472.can-07-5271

Maron, S. B., Xu, J., and Janjigian, Y. Y. (2020). Targeting EGFR in esophagogastric cancer. *Front. Oncol.* 10, 553876. doi:10.3389/fonc.2020.553876

Milani, M., Jha, G., and Potter, D. A. (2009). Anastrozole use in early stage breast cancer of post-menopausal women. *Clin. Med. Ther.* 1, 141–156. doi:10.4137/cmt.s9

Monk, J. P., Phillips, G., Waite, R., Kuhn, J., Schaaf, L. J., Otterson, G. A., et al. (2006). Assessment of tumor necrosis factor Alpha blockade as an intervention to improve tolerability of dose-intensive chemotherapy in cancer patients. *J. Clin. Oncol.* 24, 1852–1859. doi:10.1200/jco.2005.04.2838

Napier, K. J., Scheerer, M., and Misra, S. (2014). Esophageal cancer: A review of epidemiology, pathogenesis, staging workup and treatment modalities. *World J. Gastrointest. Oncol.* 6, 112–120. doi:10.4251/wjgo.v6.i5.112

Nathan, M. R., and Schmid, P. (2017). A review of fulvestrant in breast cancer. *Oncol. Ther.* 5, 17–29. doi:10.1007/s40487-017-0046-2

Niaki, E. F., Acker, T. V., Imre, L., Nánási, P., Tarapcsák, S., Bacsó, Z., et al. (2020). Interactions of cisplatin and Daunorubicin at the chromatin level. *Sci. Rep.* 10, 1107. doi:10.1038/s41598-020-57702-7

Ochi, F., Shiozaki, A., Ichikawa, D., Fujiwara, H., Nakashima, S., Takemoto, K., et al. (2015). Carbonic anhydrase XII as an independent prognostic factor in advanced esophageal squamous cell carcinoma. *J. Cancer* 6, 922–929. doi:10.7150/jca.11269

Oettle, H., Arnold, D., Kern, M., Hoepffner, N., Settmacher, U., Neuhaus, P., et al. (2002). Phase I study of gemcitabine in combination with cisplatin, 5-fluorouracil and folinic acid in patients with advanced esophageal cancer. *Anticancer. Drugs* 13, 833–838. doi:10.1097/00001813-200209000-00008

Pang, L., Li, Q., Li, S., He, J., Cao, W., Lan, J., et al. (2016). Membrane type 1-matrix metalloproteinase induces epithelial-to-mesenchymal transition in esophageal squamous cell carcinoma: Observations from clinical and *in vitro* analyses. *Sci. Rep.* 6, 22179. doi:10.1038/srep22179

Qian, Y., Liang, X., Kong, P., Cheng, Y., Cui, H., Yan, T., et al. (2020). Elevated DHODH expression promotes cell proliferation via stabilizing β-catenin in esophageal squamous cell carcinoma. *Cell Death Dis.* 11, 862. doi:10.1038/s41419-020-03044-1

Qiao, X., Wang, X., Shang, Y., Li, Y., and Chen, S. (2018). Azithromycin enhances anticancer activity of TRAIL by inhibiting autophagy and up-regulating the protein levels of DR4/5 in colon cancer cells *in vitro* and *in vivo*. *Cancer Commun.* 38, 43. doi:10.1186/s40880-018-0309-9

Ren, P., Yu, Z.-T., Xiu, L., Wang, M., and Liu, H.-M. (2013). Elevated serum levels of human relaxin-2 in patients with esophageal squamous cell carcinoma. *World J. Gastroenterol.* 19, 2412–2418. doi:10.3748/wjg.v19.i15.2412

Ribeiro, F. R., Paulo, P., Costa, V. L., Barros-Silva, J. D., Ramalho-Carvalho, J., Jerónimo, C., et al. (2011). Cysteine-rich secretory protein-3 (CRISP3) is strongly up-regulated in prostate carcinomas with the TMPRSS2-ERG fusion gene. *Plos One* 6, e22317. doi:10.1371/journal.pone.0022317

Rong, L., Wang, B., Guo, L., Liu, X., Wang, B., Ying, J., et al. (2020). HER2 expression and relevant clinicopathological features in esophageal squamous cell carcinoma in a Chinese population. *Diagn. Pathol.* 15, 27. doi:10.1186/s13000-020-00950-y

Saito, S., Morishima, K., Ui, T., Hoshino, H., Matsubara, D., Ishikawa, S., et al. (2015). The role of HGF/MET and FGF/FGFR in fibroblast-derived growth stimulation and lapatinib-resistance of esophageal squamous cell carcinoma. *Bmc Cancer* 15, 82. doi:10.1186/s12885-015-1065-8

Sciegienka, S. J., Solst, S. R., Falls, K. C., Schoenfeld, J. D., Klinger, A. R., Ross, N. L., et al. (2017). D-penicillamine combined with inhibitors of hydroperoxide metabolism enhances lung and breast cancer cell responses to radiation and carboplatin via H2O2-mediated oxidative stress. *Free Radic. Biol. Med.* 108, 354–361. doi:10.1016/j.freeradbiomed.2017.04.001

Shirmohammadi, E., Ebrahimi, S.-E. S., Farshchi, A., and Salimi, M. (2020). The efficacy of etanercept as anti-breast cancer treatment is attenuated by residing macrophages. *Bmc Cancer* 20, 836. doi:10.1186/s12885-020-07228-y

Song, Y., Wang, X., Wang, F., Peng, X., Li, P., Liu, S., et al. (2021). Identification of four genes and biological characteristics of esophageal squamous cell carcinoma by integrated bioinformatics analysis. *Cancer Cell Int.* 21, 123. doi:10.1186/s12935-021-01814-1

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A next generation connectivity map: L1000 platform and the first 1, 000, 000 profiles. *Cell* 171, 1437–1452. e17. doi:10.1016/j.cell.2017.10.049

Sukocheva, O. A., Li, B., Due, S. L., Hussey, D. J., and Watson, D. I. (2015). Androgens and esophageal cancer: What do we know? *World J. Gastroenterol.* 21, 6146–6156. doi:10.3748/wjg.v21.i20.6146

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca. Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Tanioka, Y., Yoshida, T., Yagawa, T., Saiki, Y., Takeo, S., Harada, T., et al. (2003). Matrix metalloproteinase-7 and matrix metalloproteinase-9 are associated with unfavourable prognosis in superficial oesophageal cancer. *Br. J. Cancer* 89, 2116–2121. doi:10.1038/sj.bjc.6601372

Tong, Z., Chatterjee, D., Deng, D., Veeranki, O., Mejia, A., Ajani, J. A., et al. (2017). Antitumor effects of cyclin dependent kinase 9 inhibition in esophageal adenocarcinoma. *Oncotarget* 8, 28696–28710. doi:10.18632/oncotarget.15645

Varalda, M., Antona, A., Bettio, V., Roy, K., Vachamaram, A., Yellenki, V., et al. (2020). Psychotropic drugs show anticancer activity by disrupting mitochondrial and lysosomal function. *Front. Oncol.* 10, 562196. doi:10.3389/fonc.2020.562196

Wadleigh, R. G., Abbasi, S., and Korman, L. (2006). Palliative ethanol injections of unresectable Advanced esophageal carcinoma combined with chemoradiation. *Am. J. Med. Sci.* 331, 110–112. doi:10.1097/00000441-200602000-00022

Wang, M., Liu, D., Huang, Y., Jiang, Z., Wu, F., Cen, Y., et al. (2021a). Identification of key genes related to the prognosis of esophageal squamous cell carcinoma based on chip Re-annotation. *Appl. Sci. (Basel).* 11, 3229. doi:10.3390/app11073229

Wang, T., Zhang, P., Chen, L., Qi, H., Chen, H., Zhu, Y., et al. (2021b). Ixazomib induces apoptosis and suppresses proliferation in esophageal squamous cell carcinoma through activation of the c-Myc/NOXA pathway. *J. Pharmacol. Exp. Ther.* 380, 15–25. JPET-AR-2021-000837. doi:10.1124/jpet.121.000837

Wang, X., Li, K., Cheng, M., Wang, G., Han, H., Chen, F., et al. (2020). Bmi1 severs as a potential tumor-initiating cell marker and therapeutic target in esophageal squamous cell carcinoma. *Stem Cells Int.* 2020, 8877577. doi:10.1155/2020/8877577

Wang, Z., Lachmann, A., Keenan, A. B., and Ma'ayan, A. (2018). L1000FWD: Fireworks visualization of drug-induced transcriptomic signatures. *Bioinformatics* 34, 2150–2152. doi:10.1093/bioinformatics/bty060

Wei, Y., Wu, W., Jiang, Y., Zhou, H., Yu, Y., Zhao, L., et al. (2022). Nuplazid suppresses esophageal squamous cell carcinoma growth *in vitro* and *in vivo* by targeting PAK4. *Br. J. Cancer* 126, 1037–1046. doi:10.1038/s41416-021-01651-z

Wu, P., Feng, Q., Kerchberger, V. E., Nelson, S. D., Chen, Q., Li, B., et al. (2022). Integrating gene expression and clinical data to identify drug repurposing candidates for hyperlipidemia and hypertension. *Nat. Commun.* 13, 46. doi:10.1038/s41467-021-27751-1

Xie, Y., Zhang, J., Lu, B., Bao, Z., Zhao, J., Lu, X., et al. (2020). Mefloquine inhibits esophageal squamous cell carcinoma tumor growth by inducing mitochondrial autophagy. *Front. Oncol.* 10, 1217. doi:10.3389/fonc.2020.01217

Xie, Z., Zhong, C., and Duan, S. (2022). miR-1269a and miR-1269b: Emerging carcinogenic genes of the miR-1269 family. *Front. Cell Dev. Biol.* 10, 809132. doi:10.3389/fcell.2022.809132

Ye, B., Fan, D., Xiong, W., Li, M., Yuan, J., Jiang, Q., et al. (2021). Oncogenic enhancers drive esophageal squamous cell carcinogenesis and metastasis. *Nat. Commun.* 12, 4457. doi:10.1038/s41467-021-24813-2

Yu-jing, T., Wen-jing, T., and Biao, T. (2020). Integrated analysis of hub genes and pathways in esophageal carcinoma based on NCBI's gene expression Omnibus (GEO) database: A bioinformatics analysis. *Med. Sci. Monit.* 26, 9239344–e924020. doi:10.12659/msm.923934

Zhang, C., and Chu, M. (2018). Leflunomide: A promising drug with good antitumor potential. *Biochem. Biophys. Res. Commun.* 496, 726–730. doi:10.1016/j.bbrc.2018.01.107

Zhang, H., Chen, G., Chen, S., Fu, Z., Zhou, H., Feng, Z., et al. (2021a). Overexpression of cyclin-dependent kinase 1 in esophageal squamous cell carcinoma and its clinical significance. *Febs Open Bio* 11, 3126–3141. doi:10.1002/2211-5463.13306

Zhang, Q., Zhao, X., Zhang, C., Wang, W., Li, F., Liu, D., et al. (2019). Overexpressed PKMYT1 promotes tumor progression and associates with poor survival in esophageal squamous cell carcinoma. *Cancer Manag. Res.* 11, 7813–7824. doi:10.2147/cmar.s214243

Zhang, S., Cao, W., Yue, M., Zheng, N., Hu, T., Yang, S., et al. (2016). Caveolin-1 affects tumor drug resistance in esophageal squamous cell carcinoma by regulating expressions of P-gp and MRP1. *Tumour Biol.* 37, 9189–9196. doi:10.1007/s13277-015-4778-z

Zhang, X., Peng, L., Luo, Y., Zhang, S., Pu, Y., Chen, Y., et al. (2021b). Dissecting esophageal squamous-cell carcinoma ecosystem by single-cell transcriptomic analysis. *Nat. Commun.* 12, 5291. doi:10.1038/s41467-021-25539-x

Zhang, Y., Xu, Y., Li, Z., Zhu, Y., Wen, S., Wang, M., et al. (2018). Identification of the key transcription factors in esophageal squamous cell carcinoma. *J. Thorac. Dis.* 10, 148–161. doi:10.21037/jtd.2017.12.27

Zhang, Z., He, Q., Fu, S., and Zheng, Z. (2017). Estrogen receptors in regulating cell proliferation of esophageal squamous cell carcinoma: Involvement of intracellular Ca2+ signaling. *Pathol. Oncol. Res.* 23, 329–334. doi:10.1007/s12253-016-0105-2

Zhao, C., Lin, L., Liu, J., Liu, R., Chen, Y., Ge, F., et al. (2016). A phase II study of concurrent chemoradiotherapy and erlotinib for inoperable esophageal squamous cell carcinoma. *Oncotarget* 7, 57310–57316. doi:10.18632/oncotarget.9809

Zhou, Y., He, X., Jiang, Y., Wang, Z., Yu, Y., Wu, W., et al. (2021). Benzydamine, a CDK2 kinase inhibitor, suppresses the growth of esophageal squamous cell carcinoma *in vitro* and *in vivo*. doi:10.21203/rs.3.rs-464145/v1

Zhu, G., Li, X., Li, J., Zhou, W., Chen, Z., Fan, Y., et al. (2020). Arsenic trioxide (ATO) induced degradation of Cyclin D1 sensitized PD-1/PD-L1 checkpoint inhibitor in oral and esophageal squamous cell carcinoma. *J. Cancer* 11, 6516–6529. doi:10.7150/jca.47111

Zhu, S., Yan, X., Xiang, Z., Ding, H.-F., and Cui, H. (2013). Leflunomide reduces proliferation and induces apoptosis in neuroblastoma cells *in vitro* and *in vivo*. *Plos One* 8, e71555. doi:10.1371/journal.pone.0071555

Zhu, Z., Song, J., Gu, J., Xu, B., Sun, X., and Zhang, S. (2021). FMS-related tyrosine kinase 3 ligand promotes radioresistance in esophageal squamous cell carcinoma. *Front. Pharmacol.* 12, 659735. doi:10.3389/fphar.2021.659735

frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 03 October 2022
DOI 10.3389/fgene.2022.994069

Check for updates

OPEN ACCESS

EDITED BY
Quan Zou,
University of Electronic Science and Technology of China, China

REVIEWED BY
Yanming Di,
Oregon State University, United States
Fu Yang,
Duke University, United States

*CORRESPONDENCE
Roberto Semeraro,
roberto.semeraro@unifi.it

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 14 July 2022
ACCEPTED 15 September 2022
PUBLISHED 03 October 2022

CITATION
Carangelo G, Magi A and Semeraro R (2022), From multitude to singularity: An up-to-date overview of scRNA-seq data generation and analysis.
Front. Genet. 13:994069.
doi: 10.3389/fgene.2022.994069

COPYRIGHT
© 2022 Carangelo, Magi and Semeraro. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

From multitude to singularity: An up-to-date overview of scRNA-seq data generation and analysis

Giulia Carangelo[1], Alberto Magi[2] and Roberto Semeraro[3]*

[1]Department of Experimental and Clinical Biomedical Sciences "Mario Serio", University of Florence, Florence, Italy, [2]Department of Information Engineering, University of Florence, Florence, Italy, [3]Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy

Single cell RNA sequencing (scRNA-seq) is today a common and powerful technology in biomedical research settings, allowing to profile the whole transcriptome of a very large number of individual cells and reveal the heterogeneity of complex clinical samples. Traditionally, cells have been classified by their morphology or by expression of certain proteins in functionally distinct settings. The advent of next generation sequencing (NGS) technologies paved the way for the detection and quantitative analysis of cellular content. In this context, transcriptome quantification techniques made their advent, starting from the bulk RNA sequencing, unable to dissect the heterogeneity of a sample, and moving to the first single cell techniques capable of analyzing a small number of cells (1–100), arriving at the current single cell techniques able to generate hundreds of thousands of cells. As experimental protocols have improved rapidly, computational workflows for processing the data have also been refined, opening up to novel methods capable of scaling computational times more favorably with the dataset size and making scRNA-seq much better suited for biomedical research. In this perspective, we will highlight the key technological and computational developments which have enabled the analysis of this growing data, making the scRNA-seq a handy tool in clinical applications.

KEYWORDS

single cell, RNA sequencing, transcriptomics, spatial transcriptomics, biomedical applications, technological evolution

1 Introduction

For many years researchers have tried to comprehend the complexity of tissues, organs and organisms (Grizzi and Chiriva-Internati, 2005). In order to gain this understanding, many studies have focused on cell characterization, redefining the cell as not only the structural but also the functional unit of life (Arendt et al., 2016).

Traditionally, cells have been classified by their morphology or by the expression of certain proteins in functionally distinct settings, but the advent of NGS techniques paved the way for the detection and quantitative analysis of cellular content (Mosmann et al., 1986; Orkin, 2000; Poulin et al., 2016). The high amount of data generated in modern genomics and transcriptomics experiments permitted to better characterize the architecture of genomes and the complexity of the molecular mechanisms underlying cellular activity, allowing an increasingly more accurate and in-depth depiction of cell plasticity in dynamic processes such as development, differentiation and disease evolution (Sedlazeck et al., 2018; Stark et al., 2019).

Modern cellular and molecular biology knowledge is largely derived from RNA sequencing (RNA-seq) experiments. Over the last 20 years, the transcriptome quantification has shaped our understanding of mechanisms responsible for phenomena, such as the alternativeness of the mRNA splicing process, the regulation of gene expression by non-coding and enhancer RNAs respectively and the drug resistance in some types of cancer, becoming a common and powerful technology suitable for biomedical research (Wang et al., 2008; Morris and Mattick, 2014; Li et al., 2016; Marco-Puche et al., 2019).

The adaptation and evolution of RNA-seq has been driven by technological developments and resulted in a progressive increase of the analysis resolution. Starting from the so called "bulk" RNA-seq, capable of measuring the average gene expression levels of ensembles of millions of cells, we moved to the scRNA-seq that, by allowing to profile the transcriptome of single cells, has revealed rare cellular properties and biologically meaningful cell-to-cell variability, laying the groundwork for heterogeneity-oriented studies (Svensson et al., 2018; Li and Wang, 2021).

As experimental protocols have improved rapidly, computational workflows for processing the data have also been refined, taking into account the increased throughput of scRNA-seq experiments (Andrews et al., 2021). The current "standard" analysis pipeline consists of two main sections: preprocessing, including all the steps necessary to clean the data matrix from unwanted sources of information (quality control, normalization, data correction, feature selection and dimensionality reduction) and cell- and gene-level downstream analysis, used to extract biological insights and describe the underlying biological system. For each of these steps, computational biologists developed a range of methods which perform better in different tasks and settings, making the creation of generalizable workflows for single cell experiments analysis challenging.

In this perspective, we will present an overview of the computational workflow, arguing the tools available to proceed in each step and highlighting the key technological developments which have enabled the analysis of this ever-increasing amount of data, making the scRNA-seq a handy tool in biomedical research.

2 Single cell sequencing

The first studies of single cells date back to the early 90s and were motivated by incoming discoveries which highlighted cell plasticity in dynamic processes and the different functionality based on localization (Eberwine et al., 1992). The advent of NGS techniques opened up to the era of quantitative analysis of cellular content, although first transcriptomics techniques (bulk RNA-seq) were not able to survey the diversity of cell types in a sample (Hong et al., 2020). The scaling of technologies to profile large numbers of cells in parallel has been the key to driving single cell transcriptomics forward (see Figure 1).

2.1 Technical evolution

The first example of single cell transcriptomics is the study of a handful of mouse primordial germ cells by Tang et al. (2009). By manual modification of cDNA amplification protocols previously employed in microarray analyses, he captured and quantified for the first time the full-length cDNAs for 64% of the expressed genes of a single cell, without affecting the accuracy of the protocol, which was however very time consuming and limited to small numbers of atypically large cells.

In the wake of Tang et al., new different approaches were developed including the so-called tag sequencing methods. For instance, in 2011, Islam et al. quantified the transcriptome of 85 cells by means of single cell tagged reverse transcription (STRT) (Islam et al., 2011). In brief, the authors settled single cells into the wells of a 96-well PCR plate preloaded with lysis buffer and then added reverse transcription (RT) reagents to generate a first-strand cDNA. Next, a unique template-switching oligo (TSO) with a specific sequence (six-base) on its 3′ end and a universal primer sequence on the 5′ was added to each well triggering the RT template-switching mechanism which produces a cDNA molecule incorporating the sequence at the 3′ of the TSO.

The introduction of these "barcode" sequences allowed, for the first time, to assay many cells in parallel *via* multiplexed unbiased RNA-seq, although, in the STRT-seq method, full-length cDNA is amplified by template-switching, but only the 5′ end fragment is captured and sequenced. To overcome thislimitation, the full-length SMART-seq (Ramsköld et al., 2012) and SMART-seq2 (Picelli et al., 2013) protocols were developed by Ramsköld et al. and Picelli et al., in 2011 and 2013 respectively. Compared with existing tag based methods, SMART-seq has improved read coverage across transcripts, promoting a detailed analyses of alternative transcript isoforms and identification of single-nucleotide polymorphisms (SNP).

In sight of this, it is therefore necessary to clarify that it is possible to profile the transcriptome through full-length transcript analysis or by digital counting of either 3′ or 5′ ends. While the two methods carry similar levels of reproducibility, the latter methods consist in a cost-effective solution to quantify a high amount of transcripts at the expense of a large loss of information for each of these,

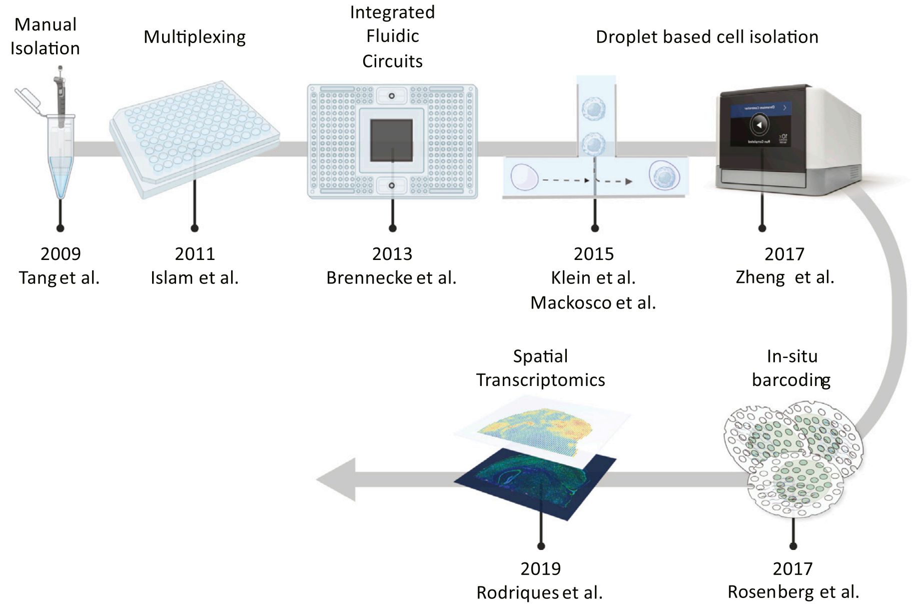


FIGURE 1
Noteworthy technologies that have allowed to profile large numbers of cells in parallel. Starting from manual isolation methods, a jump to ~100 cells was enabled by sample multiplexing and than the development of integrated fluidic circuits increased these numbers to an order of magnitude. Next, the introduction of nanodroplet technologies increased throughput even further to hundreds of thousands of cells, as for *in situ* barcoding which favoured the development of spatial methods.

contrary to the former which, by taking advantage of full-length transcripts entirety, allows the detection of splice variants and alternative transcripts, as well as genetic alterations in the transcribed fraction but for a lower number of cells.

A further application of SMART-seq2 protocols, although with some modifications (Egidio et al., 2014), is found also in the work of Brennecke et al. (2013). By means of an integrated fluidic circuit (IFC) method, implemented in the Fluidigm C1 system, they studied 96 cells isolated into individual reaction chambers and subjected to automatic staining, lysis, and sequencing in extraordinarily fast times and in a "passive" manner never seen before. In fact, the key feature of this technology is the design of microfluidics devices (or chips) that allow the sequential delivery of very small and precise volumes into tiny reaction chambers. However, a major limitation derives from the number of these chambers (96) which restrict the analysis to an equivalent number of cells, as for Brennecke in 2013. Some following large-scale studies made use of a large number of IFCs to create big data sets (Zeisel et al., 2015).

In 2015, the advent of microfluidic platforms bypassed this drawback thanks to the usage of nanoliter microreactor droplets which can encapsulate cells with no physical, and therefore numerical, restraints. The inDrop (Klein et al., 2015) and the Drop-seq (Macosko et al., 2015) protocols enter the scene with related commercial systems that allow to randomly capture cells in beads containing lysis buffer, RT reagents and barcoded oligonucleotide primers, so that mRNA is released from each cell and remains trapped in the bead to be barcoded during synthesis of cDNA. The two methods mainly differ in barcoding strategy and amplification technique, since the inDrop protocol uses hydrogel beads bearing poly(T) primers with defined barcodes and, after pooling, initiates linear amplification (IVT), contrary to Drop-seq which uses beads with random barcodes and amplifies through PCR. The random isolation of cells, however, comes with inherent limitations. Poisson statistics of cell capture to ensure that mostly single cells are isolated means there will always be large inefficiencies in terms of cell isolation, and the pool of barcodes will always have to be substantially larger than the number of cells captured to avoid barcode duplication. A large number of barcodes means the usage of very long and therefore expensive oligos. To reduce their synthesis costs, two different strategies are adopted by both methods: the combination of multiple shorter designed barcodes (e.g., 8–10 bases) into longer barcodes (e.g., 8 bases +22-base linker +10 bases = 40 bases), as for InDrop, or the synthesis of very long (e.g., 12 bases) random barcodes, as for DropSeq. This second procedure is simpler than the first and does not require any synthesized oligos for the barcodes. However, in the first approach barcodes can be designed to avoid biases and ensure that each sequence will be distinct.

The need for a large number of oligos was mitigated in 2017, through the advent of the combinatorial *in situ* barcoding methods, when Rosenberg et al. introduced the split-pool ligation-based transcriptome sequencing (SPLiT-seq), a low-cost, scRNA-seq method that enables transcriptional profiling of hundreds of thousands of fixed cells or nuclei in a single experiment (Rosenberg et al., 2018). In brief, a suspension of formaldehyde-fixed cells or nuclei passes through four rounds of combinatorial barcoding. At the first round, cells are distributed in a 96-wells plate and labelled with a specific tag. Next, cells are pooled and subjected to another label-expanding round. So, in the third round, another portion is added, carrying with it a unique molecular identifier (UMI) specific for each transcript and also used in other tag-based methods, such as STRT-seq, InDrop and Drop-seq, to better quantify the native, unamplified transcript levels (Islam et al., 2014; Stegle et al., 2015). Finally, sequencing adapters are introduced by PCR and, subsequently, each transcriptome is assembled by combining reads containing the same four-barcode combination.

Along with SPLiT-seq, one of the most vastly used methods makes its entry. The 10x Genomics company presents a new system called Chromium, based on an inDrop-seq variant. Specifically, single cells, RT reagents, Gel Beads containing barcoded oligonucleotides, and oil are combined onto a microfluidic chip to form reaction vesicles called Gel Beads in Emulsion, or GEMs. GEMs are formed in parallel within the 8 microfluidic channels of the chip, allowing the user to process hundreds to hundreds of thousands of single cells in a single 7-min run, with a ~65% of capture efficiency (Zheng et al., 2017). Within each GEM reaction vesicle, a single cell is lysed, the Gel Bead is dissolved to free the identically barcoded RT oligonucleotides into solution, and reverse transcription of polyadenylated mRNA occurs. As a result, all cDNAs from a single cell will have the same barcode, allowing the sequencing reads to be mapped back to their single cells of origin. The scalability and robustness of the system has favored the rapid diffusion of this device and its acquisition by many research laboratories in the medical field. Another contribution to this field comes from the so-called spatial RNA sequencing (spRNA-seq). Introduced in 2019 to enable the understanding of how tumor cells can communicate with each other, escape the immune system, develop drug resistance and metastasize, it combines the strengths of the global transcriptional analysis of bulk RNA-seq and *in situ* hybridization, providing whole transcriptome data with spatial information. Two technologies are currently available by 10x Genomics and Nanostring Technologies, both using proprietary spatial gene expression slides on which to fix fresh-frozen or Formalin-Fixed Paraffin-Embedded (FFPE) tissue. The two technologies differ for slide functionalization. The 10x device contains oligo capture probes, similar to those coating the gel beads, and once the tissue is fixed, stained and imaged, it is permeabilized to release the RNA, captured by probes and subjected to on-slide cDNA synthesis (Ståhl et al., 2016; Rodriques et al., 2019). The Nanostring system, uses barcode-labeled probes and fluorescent markers to hybridize to mRNA targets and to establish tissue "geography" respectively. After the regions-of-interest (ROIs) are selected, the barcodes are released *via* UV exposure and collected from the ROIs on the tissue (Moses and Pachter, 2022).

The labeled RNAs, for both technologies, are then sequenced through standard NGS procedures.

The spRNA-seq is still in its early stages and there are several common challenges that limit its applications, including non-single cell resolution, relatively low sensitivity, high cost and labor-intensive process, but given its capacity to dissect intercellular subpopulations sensitively and spatially, it will inevitably become a fundamental area of research in both discovery and therapeutics.

2.2 Bioinformatic analysis

2.2.1 General information and workflow

The rapid technological evolution that allowed the parallel analysis of thousands of cells, promoting the spread of scRNA-seq techniques, was accompanied by the development of new data analysis pipelines capable of managing such a large amount of data. The mathematical representation of these massive datasets is an "expression" matrix, defined by the number of detected genes and observed cells respectively. The process aimed at its generation starts with the read quality check. The FastQ files outputted from the sequencer are evaluated by means of quality check tools, like FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), to undergo de-multiplexing, adapter trimming, alignment and count. Tailored pipelines such as Cell Ranger (Zheng et al., 2017), UMI-tools (Smith et al., 2017), scPipe (Tian et al., 2018) and zUMIs (Parekh et al., 2018), were developed to carry out these preliminary steps. Alternatively, researchers can build their own workflows by combining individual methods that address each of the aforementioned tasks (see Table 1). For instance, the STAR (Dobin et al., 2013) aligner implements the STARsolo algorithm suited to trim, align and count this kind of data in a very fast way (Brüning et al., 2022).

Moreover, if reads are UMI-tagged, only cell barcodes that represent intact individual cells are kept. The most unambiguous approach to assess emptiness is to calculate a dataset-specific threshold of the minimum number of UMIs required to consider a barcode as a cell (Zheng et al., 2017). Alternatively tools, such as EmptyDrops (Lun et al., 2016a), identify cell barcodes that significantly deviate from background levels of RNA present in empty wells. The resulting cells still show unwanted biases. All processes involved in bias removal define the so called "preprocessing" which consists in quality control, normalization, batch correction, feature selection and dimensionality reduction. All these steps are preparatory for the following expression analysis, used to extract biological insights and describe the underlying biological system (see Figure 2).

TABLE 1 Raw data processing tools.

	Name	Alignment	QC	Count	CC	PL	References
Pipelines	CellRanger	x	x	x	x	R/Python	Zheng et al. (2017)
	UMI-tools	x	x	x	x	Python	Smith et al. (2017)
	scPipe	x	x	x	x	C++/R	Tian et al. (2018)
	zUMIs	x	x	x	x	R/Perl	Parekh et al. (2018)
	dropEst	x	x	x	x	C++	Petukhov et al. (2018)
	Optimus	x	x	x	x	Python/C++	
Tools	STAR	x	x	x	x	C/C++	Dobin et al. (2013)
	HISAT2	x	-	-	-	C/C++	Kim et al. (2015)
	kallisto	-	-	x	-	C/C++	Bray et al. (2016)
	FastQC	-	x	-	-	Java	
	HTSeq	-	x	x	-	Python	Putri et al. (2022)
	featureCount	-	-	x	-	C	Liao et al. (2014)
	EmptyDrops	-	-	-	x	R	Lun et al. (2019)

QC, quality check; CC, cell calling; PL, programming language.

Also in this context, tailored pipelines and individual tools are available to perform each operation. Toolboxes, such as Scanpy (Wolf et al., 2018), SCell (Diaz et al., 2016), Seurat (Hao et al., 2021) and scater (McCarthy et al., 2017) allow to complete multiple tasks bypassing problems related to data format conversions, making the analysis simpler. On the other hand, it is important to remember that it is difficult for a tool with many functions to continue to represent the state of the art in all of them.

In this perspective, we will present an overview of the computational workflow, arguing the tools available to proceed in each step (see Table 2).

2.2.1.1 Quality control

Before analyzing the expression matrix, we must assess the uniqueness of each barcode and cell viability. To this end, it is important to keep in mind that some droplets might contain more than 1 cell or no cell at all, making it a doublet, multiplet or an empty droplet. Furthermore, cells can be dying or damaged during isolation, misrepresenting the sample composition. So, we need to filter out them.

A possible solution is to identify these cells by evaluating three aspects of the data: the number of counts per cell/barcode (count depth), the number of genes per cell/barcode, and the fraction of counts from mitochondrial genes per cell/barcode. The thresholds for these covariates are arbitrary based on the general characteristics of the data itself, but they allow us to filter out cells with low count depths, few detected genes and/or high fraction of mitochondrial counts, as those are considered damaged cells, and at the same time they allow to filter out cells with too high counts which are indicative of doublets or multiplets (Ilicic et al., 2016).

However, a misinterpretation of these covariates could lead to wrong filtering, since in some cases a deviation in one of these values may be related to a particular cell condition, such as heavy respiration (high mitochondrial counts), quiescence (low counts, few genes) and a larger size (high counts). Therefore, they should be considered jointly when univariate thresholding decisions are made, and these thresholds should be set as permissive as possible to avoid filtering out viable cell populations unintentionally.

For doublet detection, more precise methods were developed (Xiong et al., 2022). For instance, scrublet (Wolock et al., 2019) is able to discern “embedded” (same cell type) from “neotypic” (different cell types) doublets, assuming that among all observed transcriptomes, multiplets are relatively rare events and that all cell states contributing to doublets are also present as single cells elsewhere in the data.

Quality control can also include a gene filtering step, since genes expressed in few cells are non-informative of the cellular heterogeneity. The threshold is again arbitrary, but in principle it should scale with the number of cells in the dataset and the intended downstream analysis, because, based on that choice, for example, it could limiti the identification of small clusters that might actually carry valuable information about less represented cell population.

2.2.1.2 Normalization

By means of quality control we removed sources of unwanted and inaccurate information. However, the dataset is still affected by multiple biases due to technical and biological variability. Sources responsible for such events could be, for example, capture efficiency, amplification and incomplete library sequencing. The consequence is an alteration in the counts which make cells incomparable (Macosko et al., 2015).

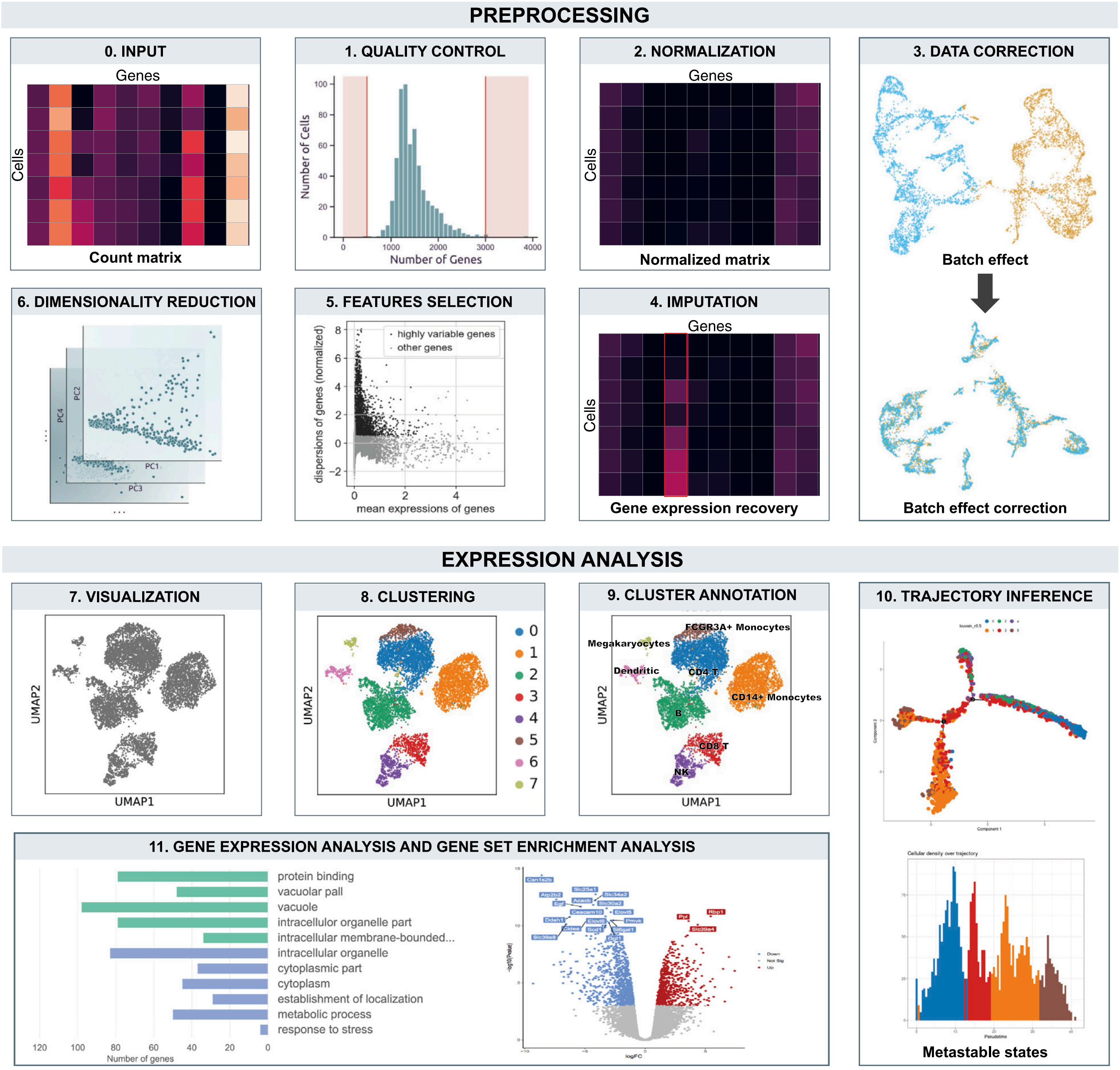


FIGURE 2
Overview of the workflow. The count matrix undergo preprocessing and expression analysis. Boxes are ordered according data analysis flow.

Normalization addresses this issue by e.g., scaling count data to obtain correct relative gene expression abundances between cells. Available methods can be linear or non-linear: a linear approach involves the estimation of size factors based on a linear regression over genes, while non-linear methods usually apply parametric modelling on count data and correlate technical and biological sources of variability to correct both (Lytal et al., 2020).

The most common normalization approach is the count depth scaling by “counts per million” (CPM), which operates by dividing gene counts by the total number of mapped reads per sample and multiplying by 1×10^6. CPM falls within linear global scaling normalization methods and assumes that all cells in the dataset initially contained an equal number of mRNA molecules (10^6) and count depth differences arise only due to sampling. Variations of this method scale the size factors with different factors of 10, or by the median count depth per cell in the dataset. Tools such as scran (Lun et al., 2016b) and Scanpy implement extensions of CPM approach. The former was proven to perform better than others in order to proceed with differential expression (DE) analysis (Vieth et al., 2019).

TABLE 2 Analysis tools.

		Preprocessing				Expression analysis				PL	
	Name	QC	N	BC	DR	V	C	DE	TI		References
Pipelines	CellRanger	x	x	x	x	x	x	x	-	R/Python	Zheng et al. (2017)
	Scanpy	x	x	x	x	x	x	x	x	Python	Wolf et al. (2018)
	Seurat	x	x	x	x	x	x	x	-	R	Hao et al. (2021)
	SCell	x	x	x	x	x	x	x	x	Matlab	Diaz et al. (2016)
	scater	x	x	x	x	x	x	x	x	R	McCarthy et al. (2017)
	Pagoda2	x	x	x	x	x	x	x	-	R	Lopez et al. (2018)
Tools	Doublet Finder	x	-	-	-	-	-	-	-	R	McGinnis et al. (2019)
	Scrublet	x	-	-	-	-	-	-	-	Python	Wolock et al. (2019)
	scds	x	-	-	-	-	-	-	-	R	Bais and Kostka (2020)
	scran	x	x	-	-	-	-	-	-	R	Bray et al. (2016)
	SCnorm	-	x	-	-	-	-	-	-	R	
	bioinfokit	-	x	-	-	-	-	-	-	R	Putri et al. (2022)
	ComBat	-	-	x	-	-	-	-	-	R	Johnson et al. (2007)
	mnnCorrect	-	-	x	-	-	-	-	-	R	Haghverdi et al. (2018)
	Harmony	-	-	x	-	-	-	-	-	R	Korsunsky et al. (2019)
	BBKNN	-	-	x	-	-	-	-	-	Python	Polański et al. (2020)
	SAUCIE	-	-	x	x	x	x	-	-	Python	Amodio et al. (2019)
	scVI	-	-	x	x	-	-	x	-	Python	Boyeau et al. (2019)
	PCA	-	-	-	x	-	-	-	-	Python	Pedregosa et al. (2011)
	t-SNE	-	-	-	x	x	-	-	-	Python/R	Van der Maaten and Hinton (2008)
	UMAP	-	-	-	x	x	-	-	-	Python/R	McInnes et al. (2018)
	Louvain	-	-	-	-	-	x	-	-	Python/R	Blondel et al. (2008)
	Leiden	-	-	-	-	-	x	-	-	Python/R	Traag et al. (2019)
	MAST	-	-	-	-	-	-	x	-	R	Finak et al. (2015)
	scCODE	-	-	-	-	-	-	x	-	R	Zou et al. (2022)
	Slingshot	-	-	-	-	-	-	-	x	R	Street et al. (2018)
	DPT	-	-	-	-	-	-	-	x	Python	Haghverdi et al. (2016)
	Whishbone	-	x	-	x	x	-	x	x	Python	Setty et al. (2016)
	Monocle2	-	x	x	x	x	x	x	x	R	Trapnell et al. (2014)
	Monocle3	-	x	x	x	x	x	x	x	R	Cao et al. (2019)
	velocyto	-	x	x	x	x	x	x	x	Python/R	La Manno et al. (2018)
	scVelo	-	x	x	x	x	x	x	x	Python	Bergen et al. (2020)

QC, quality check; N, normalization; BC, batch correction; DR, dimensionality reduction; V, visualization; C, clustering; DE, differential expression; TI, trajectory inference; PL, programming language.

For datasets with strong batch effects, non-linear methods were proven to be more reliable, particularly for plate-based scRNA-seq data, usually affected by batch effect between plates (Svensson et al., 2017).

For full-length sequencing protocols, methods which consider the gene length are more suitable. The most common is “transcripts per million” method (TPM), implemented, for example, in the bioinfokit toolbox (http://doi.org/10.5281/zenodo.3698145) (Putri et al., 2022).

Another crucial factor for normalization is the presence of synthetic spike-ins or UMIs as a means to correct for amplification bias. By adding known concentrations of external transcripts, called spike-ins, it is possible to evaluate the presence of technical artifacts, looking for differences between their observed and expected expression. By calculating a cell-specific factor that adjusts for the differences, and by applying that factor to endogenous genes, normalized expression estimates can be obtained. In spite of the promise, there are many challenges in getting spike-ins to work well, which can result in inconsistent detections (Grün et al., 2014). Contrary to spike-ins, UMIs are easier to handle since they are attached to individual transcripts prior to PCR, making each molecule unique and allowing an absolute molecular count (Kivioja et al., 2011).

Also, genes can be normalized to make them comparable between cells. Gene counts can be scaled to have a zero mean and a unit variance (z-score), making genes equally weighted. The scaling is currently not a routine because sometimes it could be useful to give genes the same weight and sometimes not, due to the effect produced by an expression magnitude difference.

Normalized data should be log (x+1)-transformed for use with following analysis methods that assume data are normally distributed. Three main effects derive from this transformation: log values represent log fold changes (unit to measure expression), they become normally distributed, reducing the skewness of the data and finally, the mean-variance relationship typical of single cell data is mitigated (Brennecke et al., 2013).

2.2.1.3 Batch Correction

Through the normalization, we mitigated the sources of technical variability responsible for gene counts alterations. However, the dataset may still contains unwanted signals of technical and biological nature. In the latter category falls for e.g., the cell cycle effect, while in the former, the batch effect deriving from different experimental protocols or/and different plates.

In order to get rid of these biases, it is possible to proceed with data and batch correction. Currently, several tools can accomplish these tasks with different approaches (Chu et al., 2022). For example, in development-oriented studies regressing out the cell cycle effect could uncover the desired biological signals (Vento-Tormo et al., 2018; Büttner et al., 2019). To this end, methods such as Scanpy and Seurat implement functions to score the cell cycle phases and regress linearly their biological effect. Alternatively, tailored tools based on complex models, like f-scLVM (Buettner et al., 2017), are available. Sometimes, also the count bias produced by differences in cell size, if not enough corrected through normalization, could be further mitigated to emphasize development-related signals. In this situation, regressing both covariates at the same time could be the best solution to account for dependence between them.

Correcting for biological biases, however, it is not always necessary or useful, since they can be avoided through pondered experimental design or because they can relate to the biological process of interest. The same observation is in part valid also for those of technical nature. In fact, even in this case a clever experimental design allows to reduce their influence but, if present, they have no correlation with the biological signals, so they must be mitigated. This process, named batch correction, can be conducted between samples and cells of the same experiment through linear models, or among different datasets derived from multiple experimental settings through non-linear models.

One of the most common linear methods is ComBat (Johnson et al., 2007) which take into account the batch effect on mean and variance of the dataset, performing very well in most settings (Büttner et al., 2019).

If the differences in the datasets are more pronounced, linear models could confound the intra- and inter-technical and biological biases, and in this circumstances non-linear models implemented in tools such as Canonical Correlation Analysis (CCA) (Butler et al., 2018), Mutual Nearest Neighbors (MNN) (Haghverdi et al., 2018), Batch balanced kNN (BBKNN) (Polański et al., 2020) and Harmony (Korsunsky et al., 2019) have been proved to overcome the same issue and smooth out unwanted and misleading differences.

2.2.1.4 Imputation

The information stored in a single cell dataset has a very sparse nature. In mathematical terms, it translates into a matrix full of zeros. Many normalization approaches do not remove them, assuming that they represent missing values to account in calculations. However, reducing their number could reduce the noise, improving the estimation of gene-gene correlations (van Dijk et al., 2018).

Currently, many tools are available to achieve this task, and the best performing ones are mainly based on deep learning algorithms (Bao et al., 2022). In this category fall DeepImpute (Arisdakessian et al., 2019) and Deep Count Autoencoder network (DCA) (Eraslan et al., 2019). The first one uses highly correlated genes of the target genes to impute the missing values, while the second can capture the nonlinear gene-gene correlation. Their application proved to improve the performance in cell clustering, DE analysis and trajectory inference.

However, when applying expression recovery, one should take into consideration that no method is perfect. Thus, any method may over- or under-correct noise in the data. Indeed, false correlation signals have been reported as a result of expression recovery (Andrews and Hemberg, 2018).

In light of this, it is hard to assess if imputation will succeed in a particular application. A reasonable approach would be to impute for visualization and avoid it to generate hypothesis during exploratory data analysis.

2.2.1.5 Feature selection and dimensionality reduction

After proceeding with the "data cleaning" steps, a human scRNA-seq dataset can still contain up to 15,000 genes. Such a big and multidimensional object is, however, hard to manage and visualize. For these reasons, it is subjected to dimensionality reduction.

To go through this process it is important to keep in mind that many residual genes do not represent the data variability, which is a key feature to explore the heterogeneity of the sample, and so that we can consider them uninformative and ignorable. This process is called feature selection. A common way to reach this result is to look for highly variable genes (HVGs) by binning them by their mean expression and preserving the ones with the highest mean-to-variance ratio in each bin (Brennecke et al., 2013). Methods such as Scanpy and Cell Ranger implement functions to define the HVGs starting from log-transformed data, while others like Seurat work on the raw counts. Typically, between 1,000 and 5,000 HVGs are selected to proceed with robust downstream analysis (Klein et al., 2015).

Their identification is crucial also to proceed with the following dimensionality reduction. Indeed, common methods like the Principal Component Analysis (PCA) (Pearson, 1901; Pedregosa et al., 2011) benefit from using HVGs to define the reduced components used to summarize the dataset features in a low-dimensional space. This is possible through a linear approach which transforms a set of correlated variables into a smaller number of uncorrelated variables, called principal components (PCs), preserving as much of the data's variation as possible. To determine the N most informative PCs, "elbow" heuristics or the permutation-test-based jackstraw method can be used (Chung and Storey, 2015; Macosko et al., 2015).

The PCA is a technique that comes from the field of linear algebra and can be used as a data preparation technique to create a projection of a dataset prior to fitting a model. Indeed, for complex datasets whose structure could not be captured by two or three PCs, non-linear combination methods such as t-distributed stochastic neighbour embedding (t-SNE) (Van der Maaten and Hinton, 2008) and Uniform Approximation and Projection (UMAP) (McInnes et al., 2018) perform better, taking advantage of PCA data.

2.2.1.6 Visualization and clustering

Non-linear methods are commonly used to create a two-dimensional plot summarizing an scRNA-seq dataset from a larger number of significant components. t-SNE and UMAP are two typical solutions to achieve this task and are implemented in almost all scRNA-seq data processing toolbox. t-SNE takes a high dimensional data set and reduces it to a low dimensional graph focusing on capturing local similarity at the expense of global structure. UMAP, instead, tends to favour fully connected representations of the dataset using a cell-cell nearest-neighbour network to then estimates a low dimensional embedding of the data. The latter is largely replacing the former, although different representations could give different insights. In this perspective, it is good to know that also diffusion maps and partition-based methods exists to visualize complex data in different manners and for different applications, e.g., diffusion maps are good to make inferences in trajectory analyses, while partition-based methods approximate the topology of the data using clusters to produce a simplified "coarse-grain" visualization of the data, useful with very large datasets.

The clustering is commonly performed with the Louvain (Blondel et al., 2008) and the Leiden (Traag et al., 2019) algorithms.

The aim of this step is to define groups of cells with similar expression profiles, because these groups could represent cell types, intermediate cell states or other interesting aspects of the data.

Both methods are based on K-Nearest Neighbour approach (KNN graph) where cells are represented as nodes in a graph, each connected to its K most similar cells, obtained using Euclidean distances on the PC-reduced expression space, so that densely sampled regions of expression space will be represented as densely connected regions in the plot (Zappia and Oshlack, 2018).

Clustering can also be performed at multiple resolutions to inspect data at different levels of detail (i.e., more clusters of smaller dimensions). Moreover, the resulting groups can be iteratively subclustered to allow the identification of cell states captured within the same cluster.

2.2.1.7 Cluster annotation

Once clusters have been defined, it is time to identify the represented cell populations. This can be done by defining their gene signatures through the identification of marker genes. To this end, DE testings are usually applied between two groups representing the cluster and the rest of the dataset. Next, simple statistical tests such as the Wilcoxon rank-sum test or the *t*-test are used to rank the derived genes by their difference in expression. The top-ranked genes from the respective test statistic are regarded as marker genes.

Clusters can be also annotated by comparing marker genes from the dataset with those from reference datasets *via* enrichment tests, the Jaccard index or other overlap statistics. Indeed, reference databases such as the mouse brain atlas (Zeisel et al., 2018) or the Human Cell Atlas (HCA) (Regev et al., 2017) are increasingly becoming available, facilitating cell identity annotation. Also automated methods like single cell NET (Tan and Cahan, 2019) are available to accomplish this step and speedup the annotation process, although a manual revision is always suggested due to the plasticity of cell states which sometimes could be confused with others.

2.2.1.8 Trajectory analysis and metastable states

Cell clustering sometimes is not the appropriate strategy to study a dataset. Many biological processes, characterizing a dataset, cannot be described through discrete classification but rather in a more continuous way (Tanay and Regev, 2017). To achieve this result we need to apply gene dynamic models capable of ordering cells along an axis defining the time process, also known as pseudotime. This type of approach is commonly used to study processes such as development and differentiation, and it is called Trajectory analysis.

Several methods are currently available to infer trajectories of increasing complexity, from simple linear or bifurcating paths to complex graphs, trees, or more intricated trajectories.

Usually, these algorithms take the reduced or corrected data as input in order to minimize technical variation and capture only the biological one, taking advantage also of HVGs, which are used to define the consecutive states derived from transcriptional distances from a root cell. None of the available methods has been shown to overperform the others for all kinds of trajectories, although different approaches benefit different ends, as shown in previous comparative studies (Saelens et al., 2019).

For instance, the tool Slingshot (Street et al., 2018) proved to perform better when inferring linear or multifurcating trajectories, contrary to the current state-of-the-art, Monocle2 (Trapnell et al., 2014), which gives best results in more complex and branched situations, along with its later version Monocle3 (Cao et al., 2019) and the Diffusion Pseudotimes (DPT) implemented in Scanpy (Haghverdi et al., 2016).

The aforementioned python toolbox offers also the chance to reconcile the information derived from clustering and trajectory inference, by means of the Partition-based graph abstraction (PAGA) algorithm (Wolf et al., 2019). In detail, using a statistical model for cell cluster interactions, PAGA places an edge between cluster nodes whose cells are more similar than expected, generating a map representing the static and dynamic nature of the data.

As trajectory inference deals with the way the cells in our sample change according to a pseudotime, it becomes possible to define the "preferential" transcriptomic states of the process evaluating the region density. Dense regions of cells represent the so called "metastable states" which can be visualized through histograms.

Unfortunately, few of the aforementioned methods include an evaluation of uncertainty in their model, so the predicted results should be confirmed with alternative approaches to avoid method bias (Griffiths et al., 2018). A common way to achieve this goal is to infer time dynamics by measuring relative abundances of exonic and intronic reads, representing spliced and unspliced transcripts. The change of their abundance, termed RNA velocity, allows to infer the direction in which each cell is moving in expression space along with an estimate of the rate of change, unlocking new ways to study cellular dynamics by granting access to not only the descriptive state of a cell, but also to its direction and speed of movement.

Currently, two modeling approaches exist, the originally proposed "steady-state" model adopted by velocyto (La Manno et al., 2018) and the subsequently extended dynamical model implemented in scVelo (Bergen et al., 2020). The former estimates velocities as the deviation of the observed ratio of unspliced to spliced mRNA from an inferred steady-state ratio, by leading sometimes to predicition errors if the central assumptions of a common splicing rate and the observation of the full splicing dynamics with steady-state mRNA levels are violated. The latter overcomes these limitations by generalizing velocity estimation to transient systems through the application of a likelihood-based dynamical model which solves the full transcriptional dynamics of splicing kinetics.

2.2.1.9 Gene expression analysis

Once the nature of each cluster is assessed, focusing on gene expression can give us a much broader idea on processes and mechanisms that differ among them. In this perspective, tools such as DE analysis and gene set enrichment analysis (GSEA) can help us investigate the molecular variability deriving from different experimental (medical treatment) or biological (different cell lines) conditions.

DE methods originate with bulk sequencing data analysis, where a few samples were compared to understand the molecular consequences of different experimental conditions. In single cell settings, the variables at stake increase as the number of cells under examination increases, due to cell-to-cell variability and biases such as dropout (Vallejos et al., 2017; Hicks et al., 2018). Tailored tools like MAST (Finak et al., 2015) or scCODE (Zou et al., 2022) are available to handle these features and perform DE on large single cell datasets in reasonable times, however, bulk DE tools, like DESeq2 (Love et al., 2014) and EdgeR (Robinson et al., 2010), have been proved to outperform some single cell counterparts if properly calibrated, but taking long times (Van den Berge et al., 2018). Uncorrected data are preferred for these applications, so it is crucial to account for confounding factors to perform a robust estimation of differentially expressed genes.

The testing result consists in a long list of genes differentially expressed between two or more conditions, sometimes hard to interpret in a meaningful way. To overcome this limitation, we can analyze them by grouping into sets based on shared characteristics, e.g., biological process and matabolic pathway. This approach, called GSEA, tests whether these characteristics are overrepresented in the candidate gene list and relies on the usage of curated databases such as the Gene Ontology (Ashburner et al., 2000; The Gene Ontology Consortium, 2017), KEGG (Kanehisa et al., 2017), String (https://string-db.org) and Reactome (Gillespie et al., 2022). Tools like gseapy (https://gseapy.readthedocs.io/en/latest/) and biomaRt (Durinck et al., 2009) are available to accomplish this task through multiple tests, querying the mentioned databases. Furthermore, novel algorithms (Vento-Tormo et al., 2018) allowed to proceed with paired ligand-receptor analyses which inspect the interaction between cell clusters.

2.3 Experimental design considerations

scRNA-seq has opened new avenues for the characterization of heterogeneity in a large variety of cellular systems, allowing to obtain transcriptome-wide data from individual cells. Although gene-expression profiling at single cell level has revealed an unprecedented variety of cell types and subpopulations that were invisible with traditional experimental techniques, it introduced new challenges due to the intrinsic nature of the data.

Indeed, scRNA-seq datasets show increased variability, complex expression distributions and an abundance of zeros compared to those produced in "bulk" experiments, making challenging to create broadly applicable experimental designs. In light of this, each experiment requires the user to make informed decisions before to proceed with a pondered design, which have to satisfy three principles formalized by R. A. Fisher in 1935: replication, randomization and blocking (Box, 1980).

To prepare an experiment, respecting such principles, it is good to start with a balanced block design in which samples collected from multiple conditions are evenly distributed across plates and lanes of the sequencer in order to reduce technical variation and not confound it with the biological one (Baran-Gale et al., 2018). On the opposite, processing samples separately, isolating cells from each sample onto separate plates (one for sample) and sequencing them on separate lanes (one for sample), produces a confounded design affected by additional sources of technical variation associated with batch preparation of libraries or sequencing. In this context, balanced design allow to bypass the batch correction step in the computational analysis, reducing computational times and user intervention on data.

Experimental design considerations will also be affected by the various protocols and platforms available for scRNA-seq. For instance, full-length capture or 3′ methods offer different way to explore sample characteristics.

As example, in an observational study setting, working with high numbers of cells could be the best solution to get insights on the transcriptional heterogeneity of the sample. To this aim, 3′ methods represent the best solution allowing to capture higher amounts of cells (100–1,00,000) and quantify their transcriptomes in a more simple and precise way, thanks to the usage of UMIs. On the other hand, to conduct more "in depth" observations or study genetic alterations (SNPs, structural variants) in the transcribed fraction, full-length approaches are more well suited, benefiting from a higher capture efficiency and a more precise information, but at cost of a minor number of cells (96–384). Therefore, more reads will be required for more refined tasks (Pollen et al., 2014; Wu et al., 2014), such as fully characterizing transcript structure, estimating the expression of rare isoforms, or distinguishing cells on the basis of subtle differences, while fewer reads but larger cell numbers may be preferred when mapping out a large population, searching for rare but distinct cell types, or pooling cells *in silico* to obtain average gene-expression clusters. According to this, if we design an experiment to search for a rare cell population, we have to take into account the number of cells that need to be sequenced to get such a population. This parameter can be estimated based on the expected heterogeneity of all cells in a sample, the minimum frequency expected of the rare cell type within the sample and the minimum number of cells of each type desired in the resulting data set.

In case no prior knowledge about the heterogeneity of the cell population is available, a practical solution is to perform the study with a high cell number and lower sequencing depth, and then perform pre-purification of the interested cells by fluorescence-activated cell sorting (FACS) with in-depth sequencing.

Another relevant difference between the two protocols relates to the UMIs usage. Indeed, full-length approaches make the inclusion of UMIs difficult, as each full-length transcript is fragmented following reverse transcription, and each fragment would need to be linked to the single UMI for that transcript. On the other hand, 3′ methods, like the 10x Genomics system, include a 10/12 bp UMI in each read at the beginning of the protocol, facilitating the molecule counting and the evaluation of sequencing saturation through the analysis of UMI duplicates. Moreover, the use of UMI has an impact on normalization procedure, since they are a consistent means to correct for amplification bias.Overall, several factors need to be considered before choosing a method for scRNA-seq. Whatever the design, it is always beneficial to record and retain information on as many factors as possible to facilitate downstream diagnostics.

3 Biomedical applications

Modern cellular and molecular biology knowledge is largely derived from RNA-seq experiments which allowed to understand the complexity of the dynamics responsible for metabolic alterations, fueling much discovery and innovation in the field of medicine over recent years.

The evolution of such techniques was driven by the development of protocols and devices capable of extracting transcriptomic information from an ever increasing number of single cells, laying the groundwork for heterogeneity-oriented studies.

The chance to dissect a sample in its composing cell lines opened up new perspectives in clinical studies oriented to the discovery of rare cell populations involved in the onset and evolution of diseases such as tumors. A proof of this assertion comes from Ramsköld et al., in 2012 and Patel et al., in 2014, which studied, for the first time (Ramsköld et al., 2012; Patel et al., 2014), the compositional architecture of melanoma and glioblastoma samples at single cell level. In the wake of them, an increasing number of studies and researchers have started exploiting the technique to successfully characterize cell populations in a variety of tumors (Dago et al., 2014; Ting et al., 2014; Puram et al., 2017; Zhao et al., 2020; Pal et al., 2021; Tian et al., 2022), defining their role into the disease process and their identity through the assignment of gene signatures (Young et al., 2018; Peired et al., 2020). Other contributions to the field comes from the integration of scRNA-seq and Copy Number Variant (CNV) detection. Tirosh et al., in 2016, successfully applied this technique to get new insights on intra- and interindividual, spatial, functional and genomic heterogeneity in melanoma cells, as well as details related to the tumor microenvironment and the cells populating it, validating the presence of a dormant drug-resistant population (Tirosh et al., 2016).

Similarly, in 2018, Fan et al. took advantage of CNVs and Loss of Heterozygosis (LOH) to identify and characterize the transcriptional programs which drive the distinct genetic subclones in a tumor sample (Fan et al., 2018).

Also in the neurological field, the scRNA-seq succeeded, revealing the heterogenous nature of brain cells involved in Alzheimer's disease and the different outcomes related to their different gene expression patterns (Mathys et al., 2019). In this contest, Lodato et al. exploited single cell sequencing to identify Single Nucleotide Variants (SNVs) in neuronal cells, demonstrating how somatic mutations can be used to reconstruct the developmental lineage of neurons, which live for decades in a postmitotic state accumulating mutations responsible for the creation of nested lineage trees and the relative polyclonal architecture (Lodato et al., 2015).

While, for blood, liver and heart samples, the introduction of trajectory analyses have provided new insights on differentiation processes, allowing to trace the fate of progenitor cells revealing the plasticity of their transcriptome through the identification of new transitional cell states (Jia et al., 2018; Popescu et al., 2019; Liang et al., 2022). However, the regulatory networks driving these processes are more complex and characterized by confounding factors like redundancy and nonlinear cross talk between pathways, e.g., developmental and signaling factors in the immune system. An unbiased approach to elucidate such a circuits and their alterations are the perturbation studies, which, by making use of the massive parallelism of single cell technologies merged with CRISPR-mediated editing, allow to knockout multiple target genes simultaneously producing different cell responses useful to clarify the function of multiple factors and their interactions in tens of thousands of cells (Adamson et al., 2016; Dixit et al., 2016; Jaitin et al., 2016). To extend this application to the analysis of multiple unrelated individuals, new methods that harness natural genetic variation were developed. Tools like demuxlet (Kang et al., 2018) determine the sample identity of each droplet, using genotyping data (SNPs), to characterize inter-individual variation and cell-type-specific genetic control of gene expression. Similarly, Van der Wijst et al. used SNP data to characterize alterations of gene co-expression pathways, focusing also on celltype-specific expression quantitative trait loci (eQTLs) (van der Wijst et al., 2018), promoting a new way to identify genetic variants that impact regulatory networks.

Another hot topic is damage recovery, since a better understanding of these mechanisms could allow us to identify the players involved in success or fail of such processes, offering new hints in the development of better diagnostic tools, prognostic biomarkers and signaling pathways amenable to therapeutic targeting (Kirita et al., 2019; Melica et al., 2022).

4 Future perspectives and conclusion

Single cell RNA sequencing was proven to be a cutting-edge technology in life sciences over the past decade. This field is developing remarkably rapidly and numerous easily accessible commercial solutions capable of characterizing hundreds of thousands of cells in parallel in reasonable times at competitive costs are currently available, making scRNA-seq much better suited for biomedical research and for clinical applications.

The spread of these devices fueled much discovery and innovation also in the computational biology field, promoting the development of novel approaches to extract information from the data produced by such technologies, and algorithms capable of analyzing them, scaling computational times more favorably with the dataset size. Moreover, along with RNA profiling, single cell technologies are currently employed to acquire information about multiple types of molecules in parallel, promoting the so-called "multimodal profiling". In fact, today it is possible to integrate information related to chromatin accessibility (Cusanovich et al., 2015), methylation state (Angermueller et al., 2016), cell-surface proteins (Stoeckius et al., 2017), to reveal the full-scale complexity of biological systems. Also, the developmental trajectories can be studied in a more precise way by matching the single cell technologies with CRISPR-Cas9 based genome editing. Methods such as scGESTALT (Raj et al., 2018) and LINNAEUS (Spanjaard et al., 2018) allow to simultaneously characterize molecular identities and lineage histories of thousands of cells during development and disease through the analysis of lineage barcodes, generated by genome editing.

However, high-throughput techniques come with the expense of decreased molecule capture rates, and future methods need to better balance cell numbers with cell resolution. Furthermore, with the future development of new and better bioinformatic tools, the individual tool recommendations presented here will require updates, yet the general considerations regarding the stages of data processing should remain the same.

Spatial dimension of single cell transcriptomics also represents an exciting field because, although novel and more precise technologies are becoming available (Eng et al., 2019), it presents several common challenges that limit its applications, including non-single cell resolution, relatively low sensitivity, high cost and labor-intensive process.

In conclusion, we have presented a brief and concise overview of single cell RNA sequencing technology and its applications. The continuous development of the technology will broaden its adoption in clinical and personalized medicine.

Author contributions

GC and RS wrote the manuscript and organized the figures. AM supervised the manuscript. All authors read and approved the final manuscript.

Funding

This work has been supported by the Associazione Italiana per la Ricerca sul Cancro (AIRC Investigator Grant 20307, “Third Generation Cancer Genomics”).

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.

References

Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., et al. (2016). A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell* 167, 1867–1882.e21. doi:10.1016/j.cell.2016.11.048

Amodio, M., van Dijk, D., Srinivasan, K., Chen, W. S., Mohsen, H., Moon, K. R., et al. (2019). Exploring single-cell data with deep multitasking neural networks. *Nat. Methods* 16, 1139–1145. doi:10.1038/s41592-019-0576-7

Andrews, T. S., and Hemberg, M. (2018). False signals induced by single-cell imputation. *F1000Res.* 7, 1740. doi:10.12688/f1000research.16613.2

Andrews, T. S., Kiselev, V. Y., McCarthy, D., and Hemberg, M. (2021). Tutorial: guidelines for the computational analysis of single-cell rna sequencing data. *Nat. Protoc.* 16, 1–9. doi:10.1038/s41596-020-00409-w

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* 13, 229–232. doi:10.1038/nmeth.3728

Arendt, D., Musser, J. M., Baker, C. V. H., Bergman, A., Cepko, C., Erwin, D. H., et al. (2016). The origin and evolution of cell types. *Nat. Rev. Genet.* 17, 744–757. doi:10.1038/nrg.2016.127

Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., and Garmire, L. X. (2019). Deepimpute: an accurate, fast, and scalable deep neural network method to impute single-cell rna-seq data. *Genome Biol.* 20, 211. doi:10.1186/s13059-019-1837-6

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. the gene ontology consortium. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Bais, A. S., and Kostka, D. (2020). scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics* 36, 1150–1158. doi:10.1093/bioinformatics/btz698

Bao, S., Li, K., Yan, C., Zhang, Z., Qu, J., and Zhou, M. (2022). Deep learning-based advances and applications for single-cell RNA-sequencing data analysis. *Brief. Bioinform.* 23, bbab473. doi:10.1093/bib/bbab473

Baran-Gale, J., Chandra, T., and Kirschner, K. (2018). Experimental design for single-cell RNA sequencing. *Brief. Funct. Genomics* 17, 233–239. doi:10.1093/bfgp/elx035

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. (2020). Generalizing rna velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* 38, 1408–1414. doi:10.1038/s41587-020-0591-3

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008, P10008. doi:10.1088/1742-5468/2008/10/p10008

Box, J. F. (1980). Ra fisher and the design of experiments, 1922–1926. *Am. Stat.* 34, 1–7. doi:10.2307/2682986

Boyeau, P., Lopez, R., Regier, J., Gayoso, A., Jordan, M. I., and Yosef, N. (2019). Deep generative models for detecting differential expression in single cells. *bioRxiv.* doi:10.1101/794289

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic rna-seq quantification. *Nat. Biotechnol.* 34, 525–527. doi:10.1038/nbt.3519

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., et al. (2013). Accounting for technical noise in single-cell rna-seq experiments. *Nat. Methods* 10, 1093–1095. doi:10.1038/nmeth.2645

Brüning, R. S., Tombor, L., Schulz, M. H., Dimmeler, S., and John, D. (2022). Comparative analysis of common alignment tools for single-cell RNA sequencing. *Gigascience* 11, giac001. doi:10.1093/gigascience/giac001

Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C., and Stegle, O. (2017). f-sclvm: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.* 18, 212. doi:10.1186/s13059-017-1334-8

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi:10.1038/nbt.4096

Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., and Theis, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* 16, 43–49. doi:10.1038/s41592-018-0254-1

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566, 496–502. doi:10.1038/s41586-019-0969-x

Chu, S.-K., Zhao, S., Shyr, Y., and Liu, Q. (2022). Comprehensive evaluation of noise reduction methods for single-cell RNA sequencing data. *Brief. Bioinform.* 23, bbab565. doi:10.1093/bib/bbab565

Chung, N. C., and Storey, J. D. (2015). Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics* 31, 545–554. doi:10.1093/bioinformatics/btu674

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., et al. (2015). Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348, 910–914. doi:10.1126/science.aab1601

Dago, A. E., Stepansky, A., Carlsson, A., Luttgen, M., Kendall, J., Baslan, T., et al. (2014). Rapid phenotypic and genomic change in response to therapeutic pressure in prostate cancer inferred by high content analysis of single circulating tumor cells. *PLoS One* 9, e101777. doi:10.1371/journal.pone.0101777

Diaz, A., Liu, S. J., Sandoval, C., Pollen, A., Nowakowski, T. J., Lim, D. A., et al. (2016). Scell: integrated analysis of single-cell RNA-seq data. *Bioinformatics* 32, 2219–2220. doi:10.1093/bioinformatics/btw201

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., et al. (2016). Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167, 1853–1866.e17. doi:10.1016/j.cell.2016.11.038

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). Star: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nat. Protoc.* 4, 1184–1191. doi:10.1038/nprot.2009.97

Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., et al. (1992). Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci. U. S. A.* 89, 3010–3014. doi:10.1073/pnas.89.7.3010

Egidio, C., Ooi, A., Holcomb, I., Ruff, D., Boutet, S., Wang, J., et al. (2014). A method for detecting protein expression in single cells using the c1^{TM} single-cell auto prep system (tech2p.874). *J. Immunol.* 192, 135.5.

Eng, C.-H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., et al. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqfish. *Nature* 568, 235–239. doi:10.1038/s41586-019-1049-y

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell rna-seq denoising using a deep count autoencoder. *Nat. Commun.* 10, 390. doi:10.1038/s41467-018-07931-2

Fan, J., Lee, H.-O., Lee, S., Ryu, D.-E., Lee, S., Xue, C., et al. (2018). Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.* 28, 1217–1227. doi:10.1101/gr.228080.117

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., et al. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16, 278. doi:10.1186/s13059-015-0844-5

Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., et al. (2022). The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* 50, D687–D692. doi:10.1093/nar/gkab1028

Griffiths, J. A., Scialdone, A., and Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* 14, e8046. doi:10.15252/msb.20178046

Grizzi, F., and Chiriva-Internati, M. (2005). The complexity of anatomical systems. *Theor. Biol. Med. Model.* 2, 26. doi:10.1186/1742-4682-2-26

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nat. Methods* 11, 637–640. doi:10.1038/nmeth.2930

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* 13, 845–848. doi:10.1038/nmeth.3971

Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36, 421–427. doi:10.1038/nbt.4091

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., 3rd, Zheng, S., Butler, A., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell* 184, 3573–3587.e29. doi:10.1016/j.cell.2021.04.048

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19, 562–578. doi:10.1093/biostatistics/kxx053

Hong, M., Tao, S., Zhang, L., Diao, L.-T., Huang, X., Huang, S., et al. (2020). RNA sequencing: new technologies and applications in cancer research. *J. Hematol. Oncol.* 13, 166. doi:10.1186/s13045-020-01005-x

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., et al. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* 17, 29. doi:10.1186/s13059-016-0888-1

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., et al. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* 21, 1160–1167. doi:10.1101/gr.110882.110

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., et al. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* 11, 163–166. doi:10.1038/nmeth.2772

Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., et al. (2016). Dissecting immune circuits by linking crispr-pooled screens with single-cell RNA-seq. *Cell* 167, 1883–1896.e15. doi:10.1016/j.cell.2016.11.039

Jia, G., Preussner, J., Chen, X., Guenther, S., Yuan, X., Yekelchyk, M., et al. (2018). Single cell RNA-seq and atac-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nat. Commun.* 9, 4877. doi:10.1038/s41467-018-07307-6

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* 8, 118–127. doi:10.1093/biostatistics/kxj037

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

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., et al. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* 36, 89–94. doi:10.1038/nbt.4042

Kim, D., Langmead, B., and Salzberg, S. L. (2015). Hisat: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi:10.1038/nmeth.3317

Kirita, Y., Chang-Panesso, M., and Humphreys, B. D. (2019). Recent insights into kidney injury and repair from transcriptomic analyses. *Nephron* 143, 162–165. doi:10.1159/000500638

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., et al. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9, 72–74. doi:10.1038/nmeth.1778

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., et al. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187–1201. doi:10.1016/j.cell.2015.04.044

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., et al. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods* 16, 1289–1296. doi:10.1038/s41592-019-0619-0

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., et al. (2018). RNA velocity of single cells. *Nature* 560, 494–498. doi:10.1038/s41586-018-0414-6

Li, X., and Wang, C.-Y. (2021). From bulk, single-cell to spatial RNA sequencing. *Int. J. Oral Sci.* 13, 36. doi:10.1038/s41368-021-00146-0

Li, W., Notani, D., and Rosenfeld, M. G. (2016). Enhancers as non-coding RNA transcription units: Recent insights and future perspectives. *Nat. Rev. Genet.* 17, 207–223. doi:10.1038/nrg.2016.4

Liang, Y., Kaneko, K., Xin, B., Lee, J., Sun, X., Zhang, K., et al. (2022). Temporal analyses of postnatal liver development and maturation by single-cell transcriptomics. *Dev. Cell* 57, 398–414.e5. doi:10.1016/j.devcel.2022.01.004

Liao, Y., Smyth, G. K., and Shi, W. (2014). featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi:10.1093/bioinformatics/btt656

Lodato, M. A., Woodworth, M. B., Lee, S., Evrony, G. D., Mehta, B. K., Karger, A., et al. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* 350, 94–98. doi:10.1126/science.aab1785

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nat. Methods* 15, 1053–1058. doi:10.1038/s41592-018-0229-2

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8

Lun, A. T. L., Bach, K., and Marioni, J. C. (2016a). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17, 75. doi:10.1186/s13059-016-0947-7

Lun, A. T. L., McCarthy, D. J., and Marioni, J. C. (2016b). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. *F1000Res.* 5, 2122. doi:10.12688/f1000research.9501.2

Lun, A. T. L., Riesenfeld, S., Andrews, T., Dao, T. P., Gomes, T., and Marioni, J. C.participants in the 1st Human Cell Atlas Jamboree (2019). Emptydrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* 20, 63. doi:10.1186/s13059-019-1662-y

Lytal, N., Ran, D., and An, L. (2020). Normalization methods on single-cell RNA-seq data: An empirical survey. *Front. Genet.* 11, 41. doi:10.3389/fgene.2020.00041

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi:10.1016/j.cell.2015.05.002

Marco-Puche, G., Lois, S., Benítez, J., and Trivino, J. C. (2019). RNA-seq perspectives to improve clinical diagnosis. *Front. Genet.* 10, 1152. doi:10.3389/fgene.2019.01152

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., et al. (2019). Single-cell transcriptomic analysis of alzheimer's disease. *Nature* 570, 332–337. doi:10.1038/s41586-019-1195-2

McCarthy, D. J., Campbell, K. R., Lun, A. T. L., and Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in r. *Bioinformatics* 33, 1179–1186. doi:10.1093/bioinformatics/btw777

McGinnis, C. S., Murrow, L. M., and Gartner, Z. J. (2019). Doubletfinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* 8, 329–337.e4. doi:10.1016/j.cels.2019.03.003

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *J. Open Source Softw.* 3, 861. doi:10.21105/joss.00861

Melica, M. E., Antonelli, G., Semeraro, R., Angelotti, M. L., Lugli, G., Landini, S., et al. (2022). Differentiation of crescent-forming kidney progenitor cells into podocytes attenuates severe glomerulonephritis in mice. *Sci. Transl. Med.* 14, eabg3277. doi:10.1126/scitranslmed.abg3277

Morris, K. V., and Mattick, J. S. (2014). The rise of regulatory RNA. *Nat. Rev. Genet.* 15, 423–437. doi:10.1038/nrg3722

Moses, L., and Pachter, L. (2022). Museum of spatial transcriptomics. *Nat. Methods* 19, 534–546. doi:10.1038/s41592-022-01409-2

Mosmann, T. R., Cherwinski, H., Bond, M. W., Giedlin, M. A., and Coffman, R. L. (1986). Two types of murine helper t cell clone. i. definition according to profiles of lymphokine activities and secreted proteins. *J. Immunol.* 136, 2348–2357.

Orkin, S. H. (2000). Diversification of haematopoietic stem cells to specific lineages. *Nat. Rev. Genet.* 1, 57–64. doi:10.1038/35049577

Pal, B., Chen, Y., Vaillant, F., Capaldo, B. D., Joyce, R., Song, X., et al. (2021). A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *EMBO J.* 40, e107333. doi:10.15252/embj.2020107333

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., and Hellmann, I. (2018). zumis - a fast and flexible pipeline to process RNA sequencing data with umis. *Gigascience* 7, giy059. doi:10.1093/gigascience/giy059

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 1396–1401. doi:10.1126/science.1254257

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin philosophical Mag. J. Sci.* 2, 559–572. doi:10.1080/14786440109462720

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.

Peired, A. J., Antonelli, G., Angelotti, M. L., Allinovi, M., Guzzi, F., Sisti, A., et al. (2020). Acute kidney injury promotes development of papillary renal cell adenoma and carcinoma from renal progenitor cells. *Sci. Transl. Med.* 12, eaaw6003. doi:10.1126/scitranslmed.aaw6003

Petukhov, V., Guo, J., Baryawno, N., Severe, N., Scadden, D. T., Samsonova, M. G., et al. (2018). dropest: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *Genome Biol.* 19, 78. doi:10.1186/s13059-018-1449-6

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* 10, 1096–1098. doi:10.1038/nmeth.2639

Polański, K., Young, M. D., Miao, Z., Meyer, K. B., Teichmann, S. A., and Park, J.-E. (2020). Bbknn: fast batch alignment of single cell transcriptomes. *Bioinformatics* 36, 964–965. doi:10.1093/bioinformatics/btz625

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32, 1053–1058. doi:10.1038/nbt.2967

Popescu, D.-M., Botting, R. A., Stephenson, E., Green, K., Webb, S., Jardine, L., et al. (2019). Decoding human fetal liver haematopoiesis. *Nature* 574, 365–371. doi:10.1038/s41586-019-1652-y

Poulin, J.-F., Tasic, B., Hjerling-Leffler, J., Trimarchi, J. M., and Awatramani, R. (2016). Disentangling neural cell diversity using single-cell transcriptomics. *Nat. Neurosci.* 19, 1131–1141. doi:10.1038/nn.4366

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* 171, 1611–1624.e24. doi:10.1016/j.cell.2017.10.044

Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E., and Zanini, F. (2022). Analysing high-throughput sequencing data in python with htseq 2.0. *Bioinformatics* 38, 2943–2945. doi:10.1093/bioinformatics/btac166

Raj, B., Wagner, D. E., McKenna, A., Pandey, S., Klein, A. M., Shendure, J., et al. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* 36, 442–450. doi:10.1038/nbt.4103

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., et al. (2012). Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* 30, 777–782. doi:10.1038/nbt.2282

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., et al. (2017). The human cell atlas. *Elife* 6, e27041. doi:10.7554/eLife.27041

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., et al. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 363, 1463–1467. doi:10.1126/science.aaw1219

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 360, 176–182. doi:10.1126/science.aam8999

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* 37, 547–554. doi:10.1038/s41587-019-0071-9

Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* 19, 329–346. doi:10.1038/s41576-018-0003-4

Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., et al. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* 34, 637–645. doi:10.1038/nbt.3569

Smith, T., Heger, A., and Sudbery, I. (2017). Umi-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.* 27, 491–499. doi:10.1101/gr.209601.116

Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., et al. (2018). Simultaneous lineage tracing and cell-type identification using crispr-cas9-induced genetic scars. *Nat. Biotechnol.* 36, 469–473. doi:10.1038/nbt.4124

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353, 78–82. doi:10.1126/science.aaf2403

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nat. Rev. Genet.* 20, 631–656. doi:10.1038/s41576-019-0150-2

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16, 133–145. doi:10.1038/nrg3833

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., et al. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14, 865–868. doi:10.1038/nmeth.4380

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., et al. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 19, 477. doi:10.1186/s12864-018-4772-0

Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., et al. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* 14, 381–387. doi:10.1038/nmeth.4220

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* 13, 599–604. doi:10.1038/nprot.2017.149

Tan, Y., and Cahan, P. (2019). Singlecellnet: A computational tool to classify single cell RNA-seq data across platforms and across species. *Cell Syst.* 9, 207–213.e2. doi:10.1016/j.cels.2019.06.004

Tanay, A., and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. *Nature* 541, 331–338. doi:10.1038/nature21350

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382. doi:10.1038/nmeth.1315

The Gene Ontology Consortium (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338. doi:10.1093/nar/gkw1108

Tian, L., Su, S., Dong, X., Amann-Zalcenstein, D., Biben, C., Seidi, A., et al. (2018). scpipe: A flexible r/bioconductor preprocessing pipeline for single-cell RNA-sequencing data. *PLoS Comput. Biol.* 14, e1006361. doi:10.1371/journal.pcbi.1006361

Tian, Y., Carpp, L. N., Miller, H. E. R., Zager, M., Newell, E. W., and Gottardo, R. (2022). Single-cell immunology of sars-cov-2 infection. *Nat. Biotechnol.* 40, 30–41. doi:10.1038/s41587-021-01131-y

Ting, D. T., Wittner, B. S., Ligorio, M., Vincent Jordan, N., Shah, A. M., Miyamoto, D. T., et al. (2014). Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.* 8, 1905–1918. doi:10.1016/j.celrep.2014.08.029

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., 2nd, Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352, 189–196. doi:10.1126/science.aad0501

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Sci. Rep.* 9, 5233. doi:10.1038/s41598-019-41695-z

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386. doi:10.1038/nbt.2859

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* 14, 565–571. doi:10.1038/nmeth.4292

Van den Berge, K., Perraudeau, F., Soneson, C., Love, M. I., Risso, D., Vert, J.-P., et al. (2018). Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* 19, 24. doi:10.1186/s13059-018-1406-4

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.* 9, 2579–2605.

van der Wijst, M. G. P., Brugge, H., de Vries, D. H., Deelen, P., Swertz, M. A., Franke, L., et al. (2018).Single-cell RNA sequencing identifies celltype-specific cis-eqtls and co-expression qtls. *Nat. Genet.* 50, 493–497. doi:10.1038/s41588-018-0089-9

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell* 174, 716–729.e27. doi:10.1016/j.cell.2018.05.061

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., et al. (2018). Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* 563, 347–353. doi:10.1038/s41586-018-0698-6

Vieth, B., Parekh, S., Ziegenhain, C., Enard, W., and Hellmann, I. (2019). A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.* 10, 4667. doi:10.1038/s41467-019-12266-7

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476. doi:10.1038/nature07509

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol.* 19, 15. doi:10.1186/s13059-017-1382-0

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., et al. (2019). Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* 20, 59. doi:10.1186/s13059-019-1663-x

Wolock, S. L., Lopez, R., and Klein, A. M. (2019). Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* 8, 281–291.e9. doi:10.1016/j.cels.2018.11.005

Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., et al. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* 11, 41–46. doi:10.1038/nmeth.2694

Xiong, K.-X., Zhou, H.-L., Lin, C., Yin, J.-H., Kristiansen, K., Yang, H.-M., et al. (2022). Chord: an ensemble machine learning algorithm to identify doublets in single-cell RNA sequencing data. *Commun. Biol.* 5, 510. doi:10.1038/s42003-022-03476-9

Young, M. D., Mitchell, T. J., Vieira Braga, F. A., Tran, M. G. B., Stewart, B. J., Ferdinand, J. R., et al. (2018). Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* 361, 594–599. doi:10.1126/science.aat1699

Zappia, L., and Oshlack, A. (2018). Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *Gigascience* 7. doi:10.1093/gigascience/giy083

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., et al. (2015). Brain structure. cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142. doi:10.1126/science.aaa1934

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., et al. (2018). Molecular architecture of the mouse nervous system. *Cell* 174, 999–1014.e22. doi:10.1016/j.cell.2018.06.021

Zhao, J., Guo, C., Xiong, F., Yu, J., Ge, J., Wang, H., et al. (2020). Single cell RNA-seq reveals the landscape of tumor and infiltrating immune cells in nasopharyngeal carcinoma. *Cancer Lett.* 477, 131–143. doi:10.1016/j.canlet.2020.02.010

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8, 14049. doi:10.1038/ncomms14049

Zou, J., Deng, F., Wang, M., Zhang, Z., Liu, Z., Zhang, X., et al. (2022). sccode: an r package for data-specific differentially expressed gene detection on single-cell RNA-sequencing data. *Brief. Bioinform.* doi:10.1093/bib/bbac180

frontiers | Frontiers in Genetics

TYPE Mini Review
PUBLISHED 13 October 2022
DOI 10.3389/fgene.2022.1012788

OPEN ACCESS

EDITED BY
Mehdi Pirooznia,
Johnson & Johnson, United States

REVIEWED BY
Martijn Derks,
Wageningen University and Research,
Netherlands

*CORRESPONDENCE
Xi Zhang,
xi.zhang@dal.ca
David Roy Smith,
dsmit242@uwo.ca

SPECIALTY SECTION
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

RECEIVED 05 August 2022
ACCEPTED 20 September 2022
PUBLISHED 13 October 2022

CITATION
Zhang X and Smith DR (2022), An overview of online resources for intra-species detection of gene duplications. *Front. Genet.* 13:1012788.
doi: 10.3389/fgene.2022.1012788

COPYRIGHT
© 2022 Zhang and Smith. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

An overview of online resources for intra-species detection of gene duplications

Xi Zhang[1,2]* and David Roy Smith[3]*

[1]Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS, Canada, [2]Institute for Comparative Genomics, Dalhousie University, Halifax, NS, Canada, [3]Department of Biology, Western University, London, ON, Canada

Gene duplication plays an important role in evolutionary mechanism, which can act as a new source of genetic material in genome evolution. However, detecting duplicate genes from genomic data can be challenging. Various bioinformatics resources have been developed to identify duplicate genes from single and/or multiple species. Here, we summarize the metrics used to measure sequence identity among gene duplicates within species, compare several computational approaches that have been used to predict gene duplicates, and review recent advancements of a Basic Local Alignment Search Tool (BLAST)-based web tool and database, allowing future researchers to easily identify intra-species gene duplications. This article is a quick reference guide for research tools used for detecting gene duplicates.

KEYWORDS

comparative genomics, genome duplication, genome evolution, gene duplication, paralogous genes

Introduction

Gene duplication can generate new genetic functions, a phenomenon which has been widely evidenced across the eukaryotic Tree of Life (Conant and Wolfe, 2008). There exist various models and mechanisms to explain the formation and retention of duplicated genes within genomes (Koonin, 2005; Innan and Kondrashov, 2010). For example, neutral processes can contribute to the evolution of gene duplication *via* genetic drift (Lynch, 2007; Brunet and Doolittle, 2018). Various adaptive hypotheses are available to explain how duplicate genes can be retained within species, such as the gene dosage hypothesis (Qian and Zhang, 2008) and the "escape from adaptive conflict" model (Des Marais and Rausher, 2008). There are five broad classes of mechanisms for generating gene duplicates, including whole-genome duplication (WGD) events, tandem duplications, transposon-mediated duplications, segmental duplications (also known as highly homologous sequence elements), and retroduplications, resulting from the "copy and paste" mechanism during reverse transcription (Panchy et al., 2016). In some instances, environmental conditions can impact the rate of fixation/loss of gene duplicates. For example, studies were carried out on the retention of duplicated genes involved in stress response, sensory functions, transport, and/or metabolism given specific environmental conditions (Kondrashov, 2012). Likewise, the yeast genomes *Saccharomyces cerevisiae* and

Schizosaccharomyces pombe were explored to evidence gene duplication in organismal adaptation (Qian and Zhang, 2014). A large-scale genomic analysis of land plants was carried out to support gene duplication assisting the evolution of novel functions, such as the production of specific floral structures and disease resistance (Panchy et al., 2016). Regarding algae, it was discovered that gene dosage might play a role in the survival of the Antarctic green alga *Chlamydomonas* sp. UWO241 (renamed *Chlamydomonas priscuii*) *via* the retention of highly similar duplicate genes (HSDs) (Cvetkovska et al., 2018; Zhang et al., 2021a; Stahl-Rommel et al., 2022).

Multidomain protein structures, functional redundancy, and/or extensive small-scale duplication events are some of the major challenges in detecting gene duplicates (Li et al., 2001; Prince and Pickett, 2002; Li et al., 2003b). Moreover, when trying to identify duplicate genes within or across species it is often difficult to distinguish between orthologs vs. paralogs. The latter are homologous genes descended from a common ancestor via duplication events, while the former are homologs derived by speciation events (Lallemand et al., 2020). When identifying homologous genes within species, it is common practice to identify paralogs using similarity assessment metrics. When exploring homologous genes across multiple species, it becomes more challenging to differentiate paralogs and orthologs, especially among more distantly related species. However, there are some publicly available genome databases providing the classification and identification of paralogs and orthologs, such as NCBI (Pruitt et al., 2005) and Ensembl (Birney et al., 2004; Howe et al., 2021). The former allows users to select and compare gene orthologs in closely related species, while the latter allows researchers to analyze the submitted sequences in a tree-based pipeline (https://useast.ensembl.org/info/genome/compara/homology_method.html) where the gene trees are reconciled against species trees to distinguish duplication and speciation events (i.e., paralogues and orthologues). Besides, there are various available methods for identifying orthologous genes by building orthologous groups in multispecies (Kuzniar et al., 2008; Altenhoff and Dessimoz, 2012). For example, tree-based methods usually recognize groups of genes based on the inferred types of relationship ahead of building a phylogenetic tree, such as TreeFam (Schreiber et al., 2014) and PhylomeDB (Huerta-Cepas et al., 2014); however, the multispecies, graph-based methods need to form the homology graph first and then build sets of genes dependent on the types of suggested relationships, such as OrthoMCL (Li et al., 2003a) and OrthoFinder (Emms and Kelly, 2019).

Here, we focus on gene duplication detection resources for intra-species analyses and review recent advancements in this area. We first summarize the metrics used to measure the similarity of gene duplicates within species, then compare several computational approaches that have been used to predict and collect gene duplicates within a particular genome. In addition, we review the recent development of a Basic Local Alignment Search Tool (BLAST)-based web tool (HSDFinder) (Zhang et al., 2021b; Zhang et al., 2021c) and database (HSDatabase) (Zhang et al., 2022). Using these two bioinformatics resources, a comparative platform can be built to understand the role of gene duplication in genome evolution.

Metrics for measuring sequence similarity of gene duplicates

Measuring duplicated genes within species typically involves the gene structure method and/or sequence similarity method. For example, three metrics are usually applied to evaluate the sequence similarity of the paralogous relationships in genes, such as aligned length, sequence identity and E-value (Lallemand et al., 2020). Other kinds of metrics are also available, but they are not necessarily as straightforward to measure (e.g., bit-score). Sequence similarity and alignment length of genes can be rapidly quantified by many tools, including DIAMOND (Buchfink et al., 2015) and BLAST (Kent, 2002). When identifying gene duplicates, the amino acid sequence is typically preferred over the nucleotide sequence as the former is more evolutionarily conserved providing more reliable sequence alignments as compared to DNA sequences. This is also why many gene duplication detection tools have the input files running from BLASTP or BLASTX (Kent, 2002). Furthermore, the timescale of the gene duplicates can greatly impact the selection of different metrics in the alignment software. Filtering recent gene duplicates usually requires more restrictive thresholds and vice versa. The metrics used to define the paralogs in a BLAST all-against-all amino acids sequence search usually include a smaller E-value cut-off (e.g., ≤ 1e-5), a higher identity score (e.g., ≥ 30%), and a longer aligned length (e.g., ≥ 150 amino acids) (Sander and Schneider, 1991; Maere et al., 2005; Panchy et al., 2016).

To overcome the limitations of similarity-based assessments, efforts have been made in developing various similarity-based metrics. For example, the homology-derived secondary structures of proteins (HSSP) method (Sander and Schneider, 1991) creates a formula to help researchers quantify genetic paralogous relationships (Rost, 1999; Li et al., 2001). Many databases have been developed to collect the conserved domains and pathways, which can be used to infer gene similarity (Lallemand et al., 2020), such as Pfam database (El-Gebali et al., 2019), InterPro pattern (Mitchell et al., 2019), and KEGG pathway (Kanehisa and Goto, 2000). But it should be noted that the quality of the genome assembly and annotation can play a key role in the accuracy of gene similarity assessement analyses. For example, 'duplicate' contigs from different

TABLE 1 Estimation of the amount of duplicated genes in different species. Adapted from (Lallemand et al., 2020) under the creative commons attribution license.

Species	No. of median gene count	No. of estimated gene copies	Percentage of estimated gene copies	Duplicated gene types	References
Arabidopsis thaliana	25,557	11,937	46.7	Not specified, all paralogous pairs were searched[a]	Blanc and Wolfe, (2004)
	22,810	21,622	94.8	WGD, tandem, proximal, DNA based transposed, retrotransposed, and dispersed duplications[b]	Wang et al. (2011), Lee et al. (2012)
	27,558	12,761	46.3	Not specified, genes families were obtained[c]	Maere et al. (2005)
	27,560	14,225	51.6	All paralogous pairs were searched[d]	Zhang et al. (2022)
Homo sapiens (human)	19,727	12,981	65.8	Gene families (tandem duplications searched among families)[e]	Shoja and Zhang, (2006)
	20,415	15,569	76.3	WGD and SSD[f]	Singh et al. (2015)
	22,447	11,740	~52.3	WGD and SSD[g]	Acharya and Ghosh, (2016)
	19,531	6,352	32.5	All paralogous pairs were searched[d]	Zhang et al. (2022)
Mus musculus (mouse)	21,305	14,043	65.9	Gene families (tandem duplications searched for among families)[f]	Singh et al. (2015)
	27,736	16,091	~58.0	Ensembl family database and genes >300 nt. Tandem duplications were then searched for among families[h]	Pan and Zhang, (2008)
	30,736	8,855	28.8	All paralogous pairs were searched[d]	Zhang et al. (2022)
Rattus norvegicus (rat)	18,468	12,466	67.5	Gene families (tandem duplications searched for among families)[e]	Singh et al. (2015)
	27,194	16,446	~60.5	Gene families (tandem duplications searched for among families)[h]	Pan and Zhang, (2008)
	22.219	8,757	39.4	All paralogous pairs were searched[d]	Zhang et al. (2022)
Oryza sativa (rice)	18,562	9,149	49.3	Not specified, all paralogous pairs were searched[i]	Blanc and Wolfe, (2004)
	27,910	21,461	76.9	WGD, tandem, proximal, DNA based transposed, retrotransposed, and dispersed duplications[b]	Wang et al. (2011); Lee et al. (2012)
	28,735	14,704	51.2	All paralogous pairs were searched[d]	Zhang et al. (2022)
Zea mays (maize)	~62,000	~43,000	~69.0	All paralogous pairs were searched[j]	Panchy et al. (2016)
	34,328	22,499	65.5	All paralogous pairs were searched[d]	Zhang et al. (2022)

[a] All-against-all nucleotide sequence similarity searches using BLASTN, among the transcribed sequences. Sequences aligned over >300 bp and showing at least 40% identity were defined as pairs of paralogs.

[b] All-against-all protein sequence similarity search using BLASTP (top five non-self protein matches with E-value of 1e-10 were considered). Genes without hits that met a threshold of E-value 1e-10 were deemed singletons. Single gene duplications were derived by excluding pairs of WGD, duplicates from the population of gene duplications. Tandem duplications were defined as being adjacent to each other on the same chromosome. Proximal duplications were defined as non-tandem genes on the same chromosome with no more than 20 annotated genes between each other. Single gene transposed-duplications were searched for from the remaining single gene duplications using syntenic blocks within and between 10 species to determine the ancestral locus. If the parental copy had more than two exons and the transposed copy was intronless, the pair of duplicates was classified as coming from a retrotransposition. Other cases of single gene-transposed duplications were classified as DNA-based transpositions. Dispersed duplications corresponded to the remaining duplications not classified as WGD, tandem, proximal, or transposed duplications (Lee et al., 2012; Wang et al., 2011).

[c] All-against-all protein sequence similarity search using BLASTP (E-value cutoff of 1e-10). Sequences alignable over a length of 150 amino acids with an identity of 30% were defined as paralogs. Gene families were built through single-linkage clustering.

[d] A combination of thresholds was used to acquire a larger dataset of HSD candidates (Zhang et al., 2022). All-against-all protein sequence similarity search using BLASTP (E-value cutoff of ≤1e−10) filtered via the criteria with in certain amino acid length differences and larger than certain amino acid pairwise identities. HSD candidates were added one after another at different similarity assessment metrics (i.e., HSDs identified at more relaxed thresholds were treated more strictly than those found using more conservative thresholds). For example, HSDs identified at a threshold of 90%_30aa were added on to those identified at a threshold of 90%_10aa (denoted as "90%_30aa+90%_10aa"); any redundant HSD candidates picked out at this combination threshold were removed if the more relaxed threshold (i.e., 90%_30aa) had the identical genes or contained the same gene copies from the stricter cutoff (i.e., 90%_10aa). Moreover, any HSD candidates pinpointed at the combination threshold (90%_30aa+90%_10aa) were removed if the minimum gene copy length was less than half of the maximum gene copy length for each HSD, or if HSD candidates had gene copies with incomplete conserved domains (i.e., different number of Pfam domains). After filtering the combination threshold at (90%_30aa+90%_10aa), a more relaxed threshold 90%_50aa was added on [i.e., 90%_50aa+(90%_30aa+90%_10aa)] and then carried out the same HSD candidate removal/filtering process. To minimize the redundancy and to acquire a larger dataset of HSD candidates, each selected species was proceeded with the following combination of thresholds: E + {D + [C + (B + A)]}. A = 90%_100aa+{90%_70aa+[90%_50aa+(90%_30aa+90%_10aa)]}; B = 80%_100aa+{80%_70aa+[80%_50aa+(80%_30aa+80%_10aa)]}; C = 70%_100aa+{70%_70aa+[70%_50aa+(70%_30aa+70%_10aa)]}; D = 60%_100aa+{60%_70aa+[60%_50aa+(60%_30aa+60%_10aa)]}; E = 50%_100aa+{50%_70aa+[50%_50aa+(50%_30aa+50%_10aa)]}.

[e] All-against-all protein sequence similarity search using BLASTP, with the BLOSUM62 matrix and the SEG filter, TribeMCL, with the default parameters. Tandem duplications were then searched for among families.

[f] Pooling of different datasets from Singh et al. (2015) and all-against-all protein sequence similarity search using BLASTP. WGD refers to whole genome duplication, SSD refers to small-scale duplication.

[g] Ensembl version 77, >50% sequence identity, and high confidence for paralogy.

[h] Ensembl family database and genes >300 nt. Tandem duplications were then searched for among families.

[i] All-against-all nucleotide sequence similarity searches using BLASTN, were done among the transcribed sequences. Sequences aligned over >300 bp and showing at least 40% identity were defined as pairs of paralogs.

[j] A gene is regarded as duplicated if it is significantly similar to another gene in a BLAST search (identity ≥30%, aligned region ≥150 amino acids, E-value cutoff of ≤1e−5).

TABLE 2 Summary of the characteristics of different existing tools for identifying gene duplicates.

Name	Input	Output Text	Output Plots	Main algorithm	Specificities	Other information	Resource links	Programming languages	Interface	References
Duplicated Gene Database (DGD)	Protein sequences and gene annotations data from Ensembl Flicek et al. (2013)	Tabulated txt	None	DGD defines groups of duplicated using Rost's Blast Rost (1999) parameters analysis	Using a maximum genomic distance of 2.5 MB between two putative duplicated genes	Not updated any new species since 2012	Web: http://dgd.genouest.org	No requirements	Web user interface	Ouedraogo et al. (2012)
Plant Genome Duplication Database (PGDD)	Coding DNA sequences, protein sequences and general feature format (GFF) file	Tabulated txt	Graphical visualization	Providing genome alignments from a single resource based on uniform standards that have been validated	Providing synteny information in terms of colinearity between chromosomes Wang et al. (2013)	The web link from the publication is no longer working	Web: http://chibba.agtec.uga.edu/duplication/	No requirements	Web user interface	(Lee et al., 2012; Lee et al., 2017)
DupGen_finder; PlantDGD	Pre-computed BLAST results (-outfmt 6) and gene location information (GFF file)	Tabulated txt	None	Each duplicate gene was assigned to a unique mode after all of the duplicated gene pairs were classified into different gene duplication types	Including duplicate genes derived from whole-genome, tandem, proximal, transposed, and dispersed duplication that was identified using uniform standards	MCScanX algorithm Wang et al. (2013) was incorporated in this pipeline	GitHub: https://github.com/qiao-xin/DupGen_finder; Web: http://pdgd.njau.edu.cn:8080	Perl	Web user interface and command line	(Wang et al., 2011; Qiao et al., 2019)
PTGBase	Coding DNA sequence file, protein sequence file and general feature format file	Tabulated txt	None	Using in-house scripts to look at phylogenetic relationship, location of gene models, and tandem duplicated arrays	Functional annotation of tandem duplicated genes including InterPro and Gene Ontology (GO)	The web link from the publication seems not working at the day of writing (20 June 2022)	Web: http://ocri-genomics.org/PTGBase/	No requirements	Web user interface	Yu et al. (2015)
RetrogeneDB	All sequences of all species were downloaded from Ensembl 73 Flicek et al. (2013) and Ensembl Plants 30 Kersey et al. (2016)	Tabulated txt	Graphical visualization	Using the LAST program Kielbasa et al. (2011) by the translated protein sequence alignment to the hard-masked reference genome sequence	Genes that contain a reverse transcriptase domain were excluded from the set.	The database has updated to a secondary version	Web: http://yeti.amu.edu.pl/retrogenedb; http://rhesus.amu.edu.pl/retrogenedb	No requirements	Web user interface	(Kabza et al., 2014; Rosikiewicz et al., 2017)

(Continued on following page)

TABLE 2 (*Continued*) Summary of the characteristics of different existing tools for identifying gene duplicates.

Name	Input	Output Text	Output Plots	Main algorithm	Specificities	Other information	Resource links	Programming languages	Interface	References
HSDFinder	BLASTP output of protein sequence and associated InterProScan and KEGG annotation file	Tabulated txt	Graphical visualization (heatmap plot)	Collecting by all-against-all BLAST and grouping by a simple transitive link between remaining genes	Highly relied on the similarity metrics provided	Can identify all pairs of gene duplicates only if they are satisfied with certain criteria	GitHub: https://github.com/zx0223winner/HSDFinder; Web: http://hsdfinder.com	Python, Bash	Web user interface and command line	(Zhang et al., 2021b; Zhang et al., 2021c)
HSDatabase	A list of species name, gene name and the respective HSDFinder results	Tabulated txt	Graphical visualization (Genome browser and sequence alignment)	A series of combination thresholds were applied to collect and curate HSDs from a diversity of species	To find paralogs that are highly similar in sequence and thus likely carry out the same function	Can be used on its own for comparative analyses of gene duplicates or in conjunction with HSDFinder	Web: http://hsdfinder.com/database/	No requirements	Web user interface and command line	Zhang et al. (2022)

haplotypes can remain in the final genome assemblies (especially for heterozygous genomes), potentially leading to false detection of gene duplicates; although this has been improved considerably with long-read sequencing technologies. In Table 1, various species were assessed for gene similarity showing that the observed numbers of gene duplicates can be distinct with each given threshold and assembled genome.

Bioinformatics approaches to identify gene duplicates

Researchers have been studying gene duplication for years, which has led to the development of various bioinformatics databases and tools for within and/or among genomes/species analyses (Lallemand et al., 2020). It is important to know how these tools function in order to choose the correct one for studying gene duplicates. There are a few factors for future researchers to consider, such as genome structure (e.g., diploid or haploid; plant or animal; eukaryotic or prokaryotic), the specific questions being asked (e.g., WGD genes or retrogenes; tandem duplicates or segmental duplicates), and the bioinformatics skills needed (e.g., command line environment or graphical user interface). Also, as noted above, the challenges associated with distinguishing orthologs from paralogs increase when exploring homologous genes between distant species. But there are still some tools available, such as the graph-based duplication prediction software OrthoMCL (Li et al., 2003a), which has a built-in Markovian Cluster algorithm, and the popular orthologous protein-coding genes database OrthoDB (Zdobnov et al., 2017). Besides, researchers developed an efficient and simple-to-use tool OrthoFinder (Emms and Kelly, 2015; Emms and Kelly, 2019) aimed at detecting the relationship of orthologous groups between/among species, especially one-to-many and many-to-many relationships between orthologues. This allows unique orthologous genes can to be collected using a reciprocal best hits (RBH) approach, which gets more complex as the number of gene duplication events increases. OrthoFinder can detect these relationships and provide comprehensive statistics for comparative genomic analyses via protein sequence files (one per species) in FASTA format. Despite the convenience of these tools, there is still an increasing need for bioinformatics tools and databases for studying specific types of gene duplications within a particular genome.

There are many web tools and databases devoted to within-species gene duplication analysis, some of which are no longer maintained. Table 2 presents the different types of algorithms used in these software/databases with a focus on those that are recently developed and/or actively maintained. For example, co-localized gene duplicates were collected from nine species in an early developed database named Duplicated Gene Database (DGD); however, it appears that no new species have been added since 2012 (Ouedraogo et al., 2012). Two genes are treated as co-localized relationship in the DGD only when they fit in the 100 gene window of all-against-all BLAST results and meet the following criteria ($I' = I\ x\ Min(n_1/L_1,n_2/L_2)$; $I' \geq 30\%$ if $L \geq 150$ amino acids; I is the sequence identity, L_i is the length of sequence, n_i is the number of amino acids in the aligned region) and formula ($I \geq 0.01n+4.8L^{-0.32(1+exp(-L/1000))}$) (Li et al., 2001). The database RetrogeneDB provides detailed data on retrogene duplicates, which must have at least 50% amino acid identity and coverage to the location from which they initially arose from, and be at least 150 bp long (Kabza et al., 2014; Rosikiewicz et al., 2017). PTGBase is built as an integrated database focusing on tandemly duplicated genes in plants; the tandem duplicates were collected by looking at if two or more genes from the same orthologous group are next to each other in the target genome (Yu et al., 2015). Similarly, gene and genome duplication of representative plant genomes were collected in the Plant Genome Duplication Database (PGDD) (Lee et al., 2012; Lee et al., 2017). More recently, Wang and colleagues developed a duplication events detection pipeline, called DupGen_finder, which has the built-in algorithm of MCScanX (Wang et al., 2013) and can identify duplicates of different type, such as tandem, whole-genome, transposed, proximal, or dispersed duplications (Wang et al., 2011; Qiao et al., 2019).

Recent advancement of a BLAST-based web tool and database

The psychrophilic, Antarctic green alga *Chlamydomonas priscuii* was recently shown to contain hundreds of highly similar duplicate genes, which may be helping this species survive extreme conditions via a gene dosage effect (Cvetkovska et al., 2018; Zhang et al., 2021a; Stahl-Rommel et al., 2022). A novel HSD detection tool, called HSDFinder, was developed for analyzing gene duplicates in *C. priscuii* (Zhang et al., 2021b; Zhang et al., 2021c). This tool has now been applied to many other eukaryotic genomes, the results of which are available in a online database called HSDatabase, housing 117,864 HSDs arising from 40 eukaryotic species (Zhang et al., 2022). HSDatabase contains an assortment of user-friendly features allowing users to glean important information on HSDs, including alignment length and percentage identify, and it provides external links to NCBI's genome browser, Pfam protein domains, and KEGG pathways. Furthermore, HSDatabase has a built-in BLAST tool for users to search genes of interest.

With this newly developed tool, BLAST all-against-all amino acid sequences can be used as the input file for the web server - HSDFinder (Zhang et al., 2021b) to furtherly explore sequence similarity. With a user-friendly interface, amino acid length variance and sequence identity can be conveniently submitted as similarity assessment metrics. By using these metrics, duplicate genes are grouped by a simple transitive link between remaining genes. There is an online heatmap option for users to compare intra-species gene duplicates under different thresholds. The KEGG pathway framework is used to categorize the detected duplicates in the heatmap.

A combination of thresholds (relaxed ones added onto stricter ones) was developed to acquire a larger dataset of HSD candidates in HSDatabase (Zhang et al., 2022). Also, any HSD candidates were screened out if the minimum length of gene copy was less than half of the maximum length of gene copy for every HSD group. Incomplete or unequal conserved protein family domains of HSD candidates will also result in the removal of the HSD group. But due to the limitation of this strategy, it should be noted that there are some large groups of HSD candidates in the database that likely diverged in function from one another. In the database, those putatively diverged HSD groups were labelled as "candidate HSDs" and a warning note was added that users should proceed with caution when working with these types of datasets.

Concluding perspectives

There is no stand-alone software that can detect all types of gene duplicates within and across species. There are many factors that can influence the choice of tools being used for gene duplication detection. These include, for instance, the kinds of questions being asked and the genomes being analyzed as well as the bioinformatics skills of the user. For developers, a lot of features and statistics can be added to assist future researchers, such as the rates of synonymous and nonsynonymous substitutions (dN/dS rates) and differential expression levels in different gene duplicates. One of the big challenges moving forward is how to properly help users select an appropriate threshold for their given dataset/genome and provide them with the freedom to fine-tune specific metrics. In the future, it is likely that users will be aided by species-specific gene threshold values for gene duplication detection tools. With more and more genomes being sequenced and re-sequenced, gene duplicate data from highly polished model genomes will broaden our understanding of the role of gene duplication in genome evolution and adaptation to extreme environments.

Author contributions

The study was conceptualized by XZ and DS. XZ wrote the initial draft and performed the data analysis. DS contributed to the manuscript editing. All authors commented to produce the manuscript for peer review.

Funding

The authors gratefully acknowledge funding of Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Acknowledgments

We appreciate the constructive suggestions from all the reviewers. Due to the length of the review, we apologize for not citing all relevant literatures.

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.

References

Acharya, D., and Ghosh, T. C. (2016). Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. *BMC genomics* 17, 71–14. doi:10.1186/s12864-016-2392-0

Altenhoff, A., and Dessimoz, C. (2012). *Inferring orthology and paralogy in: Evolutionary genomics*. Totowa, New Jersey, United States: Humana Press.

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., et al. (2004). An overview of Ensembl. *Genome Res.* 14, 925–928. doi:10.1101/gr.1860604

Blanc, G., and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16, 1667–1678. doi:10.1105/tpc.021345

Brunet, T., and Doolittle, W. F. (2018). The generality of constructive neutral evolution. *Biol. Philos.* 33, 2–25. doi:10.1007/s10539-018-9614-6

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi:10.1038/nmeth.3176

Conant, G. C., and Wolfe, K. H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nat. Rev. Genet.* 9, 938–950. doi:10.1038/nrg2482

Cvetkovska, M., Szyszka-Mroz, B., Possmayer, M., Pittock, P., Lajoie, G., Smith, D. R., et al. (2018). Characterization of photosynthetic ferredoxin from the Antarctic alga *Chlamydomonas* sp. UWO241 reveals novel features of cold adaptation. *New Phytol.* 219, 588–604. doi:10.1111/nph.15194

Des Marais, D. L., and Rausher, M. D. (2008). Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* 454, 762–765. doi:10.1038/nature07092

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432. doi:10.1093/nar/gky995

Emms, D. M., and Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238–314. doi:10.1186/s13059-019-1832-y

Emms, D. M., and Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16, 157–214. doi:10.1186/s13059-015-0721-2

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., et al. (2013). Ensembl 2013. *Nucleic Acids Res.* 41, D48–D55. doi:10.1093/nar/gks1236

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., et al. (2021). Ensembl 2021. *Nucleic Acids Res.* 49, D884–D891. doi:10.1093/nar/gkaa942

Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L. P., Marcet-Houben, M., and Gabaldon, T. (2014). PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.* 42, D897–D902. doi:10.1093/nar/gkt1177

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: Classifying and distinguishing between models. *Nat. Rev. Genet.* 11, 97–108. doi:10.1038/nrg2689

Kabza, M., Ciomborowska, J., and Makałowska, I. (2014). RetrogeneDB—A database of animal retrogenes. *Mol. Biol. Evol.* 31, 1646–1648. doi:10.1093/molbev/msu139

Kanehisa, M., and Goto, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Kent, W. J. (2002). BLAT—The BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi:10.1101/gr.229202

Kersey, P. J., Allen, J. E., Armean, I., Boddu, S., Bolt, B. J., Carvalho-Silva, D., et al. (2016). Ensembl genomes 2016: More genomes, more complexity. *Nucleic Acids Res.* 44, D574–D580. doi:10.1093/nar/gkv1209

Kielbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493. doi:10.1101/gr.113985.110

Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. Biol. Sci.* 279, 5048–5057. doi:10.1098/rspb.2012.1108

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 39, 309–338. doi:10.1146/annurev.genet.39.073003.114725

Kuzniar, A., van Ham, R. C., Pongor, S., and Leunissen, J. A. (2008). The quest for orthologs: Finding the corresponding gene across genomes. *Trends Genet.* 24, 539–551. doi:10.1016/j.tig.2008.08.009

Lallemand, T., Leduc, M., Landès, C., Rizzon, C., and Lerat, E. (2020). An overview of duplicated gene detection methods: Why the duplication mechanism has to be accounted for in their choice. *Genes* 11, 1046. doi:10.3390/genes11091046

Lee, T.-H., Kim, J., Robertson, J. S., and Paterson, A. H. (2017). "Plant genome duplication database," in *Plant genomics databases*. Editor D. Aalt (Berlin, Germany: Springer).

Lee, T.-H., Tang, H., Wang, X., and Paterson, A. H. (2012). PGDD: A database of gene and genome duplication in plants. *Nucleic Acids Res.* 41, D1152–D1158. doi:10.1093/nar/gks1104

Li, L., Stoeckert, C. J., and Roos, D. S. (2003a). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi:10.1101/gr.1224503

Li, W.-H., Gu, Z., Cavalcanti, A. O., and Nekrutenko, A. (2003b). Detection of gene duplications and block duplications in eukaryotic genomes. *J. Struct. Funct. Genomics* 3, 27–34. doi:10.1023/a:1022644628861

Li, W.-H., Gu, Z., Wang, H., and Nekrutenko, A. (2001). Evolutionary analyses of the human genome. *Nature* 409, 847–849. doi:10.1038/35057039

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl. Acad. Sci. U. S. A.* 104, 8597–8604. doi:10.1073/pnas.0702207104

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., et al. (2005). Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5454–5459. doi:10.1073/pnas.0501102102

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., et al. (2019). InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* 47, D351–D360. doi:10.1093/nar/gky1100

Ouedraogo, M., Bettembourg, C., Bretaudeau, A., Sallou, O., Diot, C., Demeure, O., et al. (2012). The duplicated genes database: Identification and functional annotation of co-localised duplicated genes across genomes. *PloS one* 7, e50653. doi:10.1371/journal.pone.0050653

Pan, D., and Zhang, L. (2008). Tandemly arrayed genes in vertebrate genomes. *Comp. Funct. Genomics* 2008, 1–11. doi:10.1155/2008/545269

Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. (2016). Evolution of gene duplication in plants. *Plant Physiol.* 171, 2294–2316. doi:10.1104/pp.16.00523

Prince, V. E., and Pickett, F. B. (2002). Splitting pairs: The diverging fates of duplicated genes. *Nat. Rev. Genet.* 3, 827–837. doi:10.1038/nrg928

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI reference sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 33, D501–D504. doi:10.1093/nar/gki025

Qian, W., and Zhang, J. (2008). Gene dosage and gene duplicability. *Genetics* 179, 2319–2324. doi:10.1534/genetics.108.090936

Qian, W., and Zhang, J. (2014). Genomic evidence for adaptation by gene duplication. *Genome Res.* 24, 1356–1362. doi:10.1101/gr.172098.114

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., et al. (2019). Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* 20, 38–23. doi:10.1186/s13059-019-1650-2

Rosikiewicz, W., Kabza, M., Kosiński, J. G., Ciomborowska-Basheer, J., Kubiak, M. R., and Makałowska, I. (2017). RetrogeneDB–a database of plant and animal retrocopies. *Database (Oxford).* 2017, bax038. doi:10.1093/database/bax038

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* 12, 85–94. doi:10.1093/protein/12.2.85

Sander, C., and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9, 56–68. doi:10.1002/prot.340090107

Schreiber, F., Patricio, M., Muffato, M., Pignatelli, M., and Bateman, A. (2014). TreeFam v9: A new website, more species and orthology-on-the-fly. *Nucleic Acids Res.* 42, D922–D925. doi:10.1093/nar/gkt1055

Shoja, V., and Zhang, L. (2006). A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. *Mol. Biol. Evol.* 23, 2134–2141. doi:10.1093/molbev/msl085

Singh, P. P., Arora, J., and Isambert, H. (2015). Identification of ohnolog genes originating from whole genome duplication in early vertebrates, based on synteny comparison across multiple genomes. *PLoS Comput. Biol.* 11, e1004394. doi:10.1371/journal.pcbi.1004394

Stahl-Rommel, S., Kalra, I., D'Silva, S., Hahn, M. M., Popson, D., Cvetkovska, M., et al. (2022). Cyclic electron flow (CEF) and ascorbate pathway activity provide constitutive photoprotection for the photopsychrophile, Chlamydomonas sp. UWO 241 (renamed Chlamydomonas priscuii). *Photosynth. Res.* 151, 235–250. doi:10.1007/s11120-021-00877-5

Wang, Y., Li, J., and Paterson, A. H. (2013). MCScanX-transposed: Detecting transposed gene duplications based on multiple colinearity scans. *Bioinformatics* 29, 1458–1460. doi:10.1093/bioinformatics/btt150

Wang, Y., Wang, X., Tang, H., Tan, X., Ficklin, S. P., Feltus, F. A., et al. (2011). Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. *PloS one* 6, e28150. doi:10.1371/journal.pone.0028150

Yu, J., Ke, T., Tehrim, S., Sun, F., Liao, B., and Hua, W. (2015). PTGBase: An integrated database to study tandem duplicated genes in plants. *Database*. 2015, bav017. doi:10.1093/database/bav017

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simao, F. A., Ioannidis, P., et al. (2017). OrthoDB v9. 1: Cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* 45, D744–D749. doi:10.1093/nar/gkw1119

Zhang, X., Cvetkovska, M., Morgan-Kiss, R., Hüner, N. P., and Smith, D. R. (2021a). Draft genome sequence of the Antarctic green alga Chlamydomonas sp. UWO241. *iScience* 24, 102084. doi:10.1016/j.isci.2021.102084

Zhang, X., Hu, Y., and Smith, D. R. (2022). HSDatabase—a database of highly similar duplicate genes from plants, animals, and algae. *Database* 2022, baac086. doi:10.1093/database/baac086

Zhang, X., Hu, Y., and Smith, D. R. (2021b). HSDFinder: A BLAST-based strategy for identifying highly similar duplicated genes in eukaryotic genomes. *Front. Bioinform.* 1, 803176. doi:10.3389/fbinf.2021.803176

Zhang, X., Hu, Y., and Smith, D. R. (2021c). Protocol for HSDFinder: Identifying, annotating, categorizing, and visualizing duplicated genes in eukaryotic genomes. *Star. Protoc.* 2, 100619. doi:10.1016/j.xpro.2021.100619

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 20 October 2022
DOI 10.3389/fgene.2022.1015673

Computational genomics insights into cold acclimation in wheat

Youlian Pan[1]*, Yifeng Li[1,2], Ziying Liu[1], Jitao Zou[3] and Qiang Li[3,4]*

[1]Digital Technologies, National Research Council Canada, Ottawa, ON, Canada, [2]Department of Computer Science, Department of Biological Science, Brock University, St. Catharines, ON, Canada, [3]Aquatic and Crop Research and Development, National Research Council Canada, Saskatoon, SK, Canada, [4]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, Hubei, China

Check for updates

OPEN ACCESS

EDITED BY
Richard D Emes,
University of Nottingham,
United Kingdom

REVIEWED BY
Gábor Galiba,
Centre for Agricultural Research,
Hungary
Muhammad Ahsan Asghar,
Centre for Agricultural Research,
Hungary

*CORRESPONDENCE
Youlian Pan,
youlian.pan@nrc-cnrc.gc.ca
Qiang Li,
qli@mail.hzau.edu.cn

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 10 August 2022
ACCEPTED 03 October 2022
PUBLISHED 20 October 2022

CITATION
Pan Y, Li Y, Liu Z, Zou J and Li Q (2022), Computational genomics insights into cold acclimation in wheat.
Front. Genet. 13:1015673.
doi: 10.3389/fgene.2022.1015673

COPYRIGHT
© 2022 His Majesty the King in Right of Canada. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Development of cold acclimation in crops involves transcriptomic reprograming, metabolic shift, and physiological changes. Cold responses in transcriptome and lipid metabolism has been examined in separate studies for various crops. In this study, integrated computational approaches was employed to investigate the transcriptomics and lipidomics data associated with cold acclimation and vernalization in four wheat genotypes of distinct cold tolerance. Differential expression was investigated between cold treated and control samples and between the winter-habit and spring-habit wheat genotypes. Collectively, 12,676 differentially expressed genes (DEGs) were identified. Principal component analysis of these DEGs indicated that the first, second, and third principal components (PC1, PC2, and PC3) explained the variance in cold treatment, vernalization and cold hardiness, respectively. Differential expression feature extraction (DEFE) analysis revealed that the winter-habit wheat genotype Norstar had high number of unique DEGs (1884 up and 672 down) and 63 winter-habit genes, which were clearly distinctive from the 64 spring-habit genes based on PC1, PC2 and PC3. Correlation analysis revealed 64 cold hardy genes and 39 anti-hardy genes. Cold acclimation encompasses a wide spectrum of biological processes and the involved genes work cohesively as revealed through network propagation and collective association strength of local subnetworks. Integration of transcriptomics and lipidomics data revealed that the winter-habit genes, such as *COR413-TM1*, *CIPKs* and *MYB20*, together with the phosphatidylglycerol lipids, PG(34:3) and PG(36:6), played a pivotal role in cold acclimation and coordinated cohesively associated subnetworks to confer cold tolerance.

KEYWORDS

cold acclimation, transcriptomics, lipidomics, phosphatidylglycerol lipid, differential expression feature extraction, RNA-seq, wheat

Introduction

Wheat is the second most-produced cereal crop in the world; its yield and quality are severely affected by abiotic stress such as cold. During exposure to low but non-freezing temperature, plants increase their freezing tolerance in a process termed cold acclimation. Cold acclimation is a multi-genic processes, involves reprogramming of the transcriptome, proteome, lipidome, and metabolome, affects signaling between subcellular organelles, and induces significant changes in physiological processes and morphology (Li et al., 2018; Fürtauer et al., 2019; Li et al., 2021).

In response to cold stress, genetic and molecular analyses have identified dehydration-responsive element-binding protein 1/C-repeat binding factors (DREB1s/CBFs) as master transcription factors that regulate expression of cold regulated genes (CORs) during cold acclimation (Maruyama et al., 2009; Shi et al., 2018; Kidokoro et al., 2022). Many transcription factors regulate the cold-inducible expression of *DREB1* gene in the very complex manner (Thomashow, 2010; Kidokoro et al., 2020). In the downstream, *DREB1/CBF* transcription factors upregulate many cold-responsive genes (CORs). Multiple COR genes are identified as CBF regulon (Tchagang et al., 2017; Song Y. et al., 2021; Liu et al., 2021) with respect to multiple stresses such as cold, heat, drought, and salt. The expression of the COLD REGULATED 314 THYLAKOID MEMBRANE 1 (*COR413-TM1*) correlates with cold tolerance (Breton et al., 2003). Overexpression of *DREB1A* (*CBF3*) improves stress tolerance to both freezing and dehydration in transgenic plants. Under cold and dehydration conditions, the expression of many genes encoding starch-degrading enzymes changes dynamically; many monosaccharides, disaccharides, trisaccharides, and sugar alcohols accumulate in Arabidopsis (Maruyama et al., 2009).

Winter habit plants require prolonged exposure to cold, such as winter, to promote flowering in spring through a process known as vernalization (Chouard, 1960; Amasino, 2005; Kim et al., 2009). The two important evolutionarily adaptive mechanisms, cold acclimation for winter hardiness and vernalization, are thus initiated within the same time frame upon low temperature exposure (Limin and Fowler, 2002; Danyluk et al., 2003; Li et al., 2018). Studies in Arabidopsis have shown that epigenetic regulation of *FLC* (FLOWERING LOCUS C) plays an important role in the vernalization (Crevillen et al., 2014); whereas, the FLC genes in cereal plants appear to be implicated in many other aspects of plant growth and development in addition to vernalization (Kennedy and Geuten, 2020).

In wheat, *VRN1*, together with *VRN2* and *VRN3*, forms a pivotal regulatory module for its vernalization process (Oliver et al., 2009; Chen et al., 2018). Genetic studies revealed that the two loci on chromosome 5A, Frost Resistance-1 (*FR-1*) and *FR-2* affect freezing tolerance and winter hardiness of the temperate cereal plants (Knox et al., 2010; Fowler et al., 2016). *FR-1* is believed to be a pleiotropic effect of *VRN-A1* (Brule-Babel and Fowler, 1988). The *FR-2* QTL loci spanning on chromosome 5A contains a number of genes including a cluster of 21 genes encoding CBFs which are involved in cold acclimation (Vágújfalvi et al., 2003). *VRN-A1* appears to down-regulate the expression of COR genes in the CBF regulon adjacent to the *FR-2* locus in cold acclimated winter cereals (Limin and Fowler, 2006) indicating an interaction between *VRN-A1* and *FR-2* loci (Zhu et al., 2014). Low temperature induces the expression of *VRN1s*, while genes in cold pathways including CBFs and CORs are repressed (Danyluk et al., 2003; Li et al., 2018). On the other hand, CBF proteins are believed to directly bind the promoter of the *VRN1s* to repress flowering by negatively regulating its expression in cereals (Dhillon et al., 2010; Deng et al., 2015).

Changes in membrane fluidity, cytoskeleton rearrangement, and calcium influxes are among the earliest events taking place in plants upon exposure to low temperatures (Browse and Xin, 2001; Kidokoro et al., 2022). Membrane lipid unsaturation has been well documented for its role in low temperature adaptation in plants (Wolf et al., 2001; Zheng et al., 2021), while a lower membrane unsaturation level is favored under high temperature (Murakami et al., 2000). In addition, the level of desaturated Phosphatidylglycerol (PG) which contains a combination of 16:0, 18:0 and 16:1-*trans* fatty acids in PG is related to low temperature adaptability of plants (Murata and Los, 1997). Moreover, the reduction of *trans*-Δ3 hexadecenoic acid (t16:1) has been shown to be correlated with freezing tolerance, especially in cereal crops such as wheat (Huner et al., 1989; Li et al., 2021). Studies have also proposed that adjustment in lipid redistribution between the two glycerolipid pathways as well as lipid exchanges between the ER and chloroplast is critical for temperature adaptation in plants (Li et al., 2015).

Our previous study investigated the interactions between vernalization and cold acclimation pathways in the crown tissue (Li et al., 2018). Further analysis in leaf tissue revealed a mechanistic role of *trans*-16:1 in PG as a specific metabolite marker for screening freezing tolerance in wheat and genes in lipid pathways were specifically investigated (Li et al., 2021). However, the complexity of gene regulatory networks involved in mediating cold responses as well as lipid metabolism in leaves has not been fully explored. In this paper, we employed four wheat genotypes, winter habit Norstar (N), spring habit Manitou (M), and their near isogenic lines (NILs), winter Manitou (WM) and spring Norstar (SN) with the *VRN-A1* alleles swapped (Limin and Fowler, 2002), to study the cold acclimation process in leaves through computational pattern recognition, principal component analysis, and genes and lipids association networks. Genes associated with cold acclimation are identified and characterized.

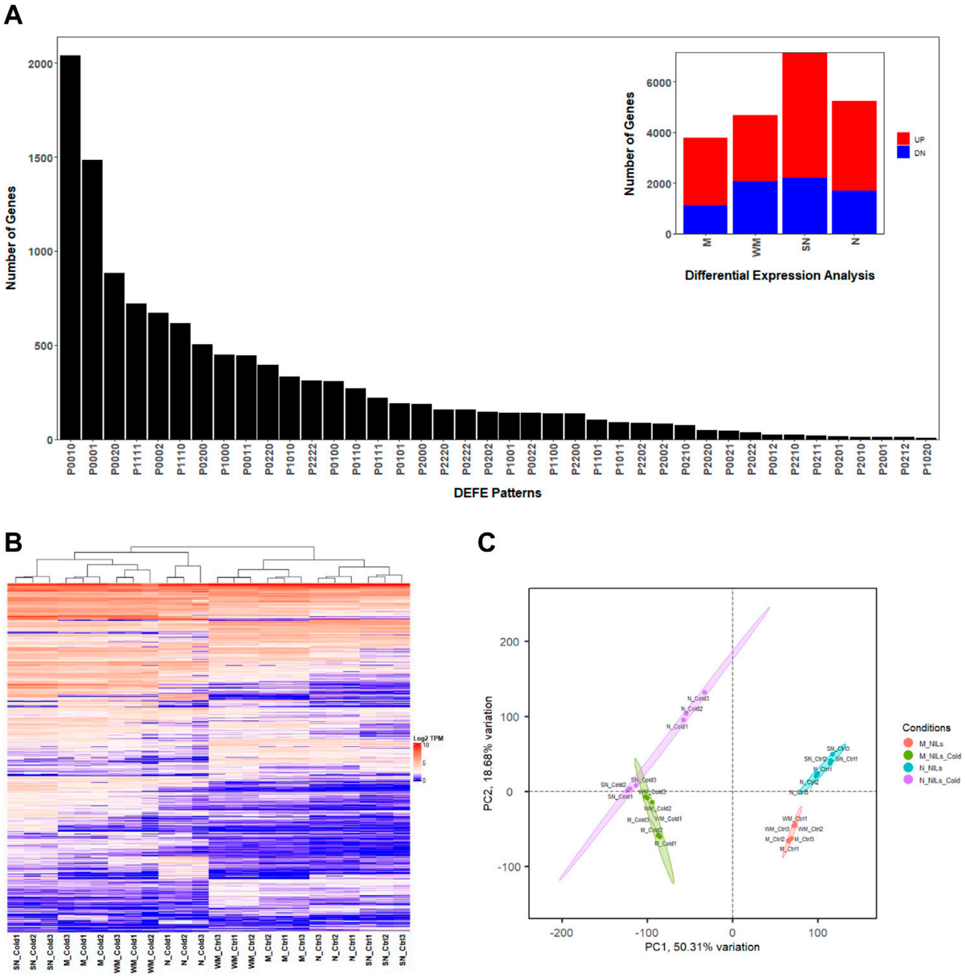


FIGURE 1
Transcriptome overview based on the 12,676 DEGs. **(A)** Frequency distribution (insert) and top 40 patterns of P series of DEFE analysis; **(B)** Heatmap; **(C)** Principal component analysis, where PC1 and PC2 are principal components 1 and 2, respectively.

Results

Transcriptome overview

In this study, RNA-seq data were obtained from leaves grown under low temperature treatment of four wheat genotypes with different LT50s (temperatures at which 50% of a population survives in an artificial freeze test), including Norstar (N, LT50 = −21.7°C), Manitou (M, LT50 = −8.3°C) and their near isogenic lines (NILs) spring Norstar (SN, LT50 = −13°C) and winter Manitou (WM, LT50 = −13.2°C) with the *VRN-A1* alleles swapped (Li et al., 2021). On average, 93% of the 26 million reads per sample that satisfied filtering criteria were mapped to the 106,914 known high confident wheat genes in the IWGSC RefSeq v2.1 genome assembly, from which, 12,676 differentially expressed genes (DEGs, Supplementary File S1) were identified based on criteria provided in the Materials and Method Section (|log2 fold-change| ≥ 2, adjusted *p*-value ≤ 0.01, and the maximum number of transcripts per million reads in a pair of compared samples ≥2, Figure 1). Between the cold treated samples and controls, the number of DEGs were spring Norstar (SN) > Norstar (N) > winter Manitou (WM) > Manitou (M) (Table 1). When the winter-habit genotype was compared with its respective near isogenic line (NIL) of spring-habit genotype, the difference between N and SN was more than three times as many DEGs as between WM and M in the cold treated samples, but the difference was only at about 1.4 times in the control samples (Table 1).

Three series of differential expression feature extraction (DEFE, Pan et al., 2022) analyses were performed. The P series [the cold treated samples compared with their respective controls, P(M, WM, SN, N); *see* Material and

TABLE 1 Number of DEGs in each pair-wise comparison.

	Cold/Control				Cold treated		Control	
	M	WM	N	SN	WM/M	N/SN	WM/M	N/SN
Up	2633	2588	3522	4932	360	1515	522	856
Down	1127	2079	1702	2210	719	2085	407	431

Method Section for details] revealed the numbers of DEGs unique to each genotype were SN (2920, P0010 = 2036, P0020 = 884) > N (2156, P0001 = 1484, P0002 = 672) > WM (811, P0100 = 308, P0200 = 503) > M (637, P1000 = 450, P2000 = 187, Figure 1A). There were 1035 common DEGs (DEFE patterns: P1111 = 722, P2222 = 313) among the four wheat genotypes (Figure 1A). Across all four genotypes, there were more DEGs up-regulated than down-regulated when subjected to cold treatment (Table 1). The top DEFE expression pattern was unique up-regulation to SN (P0010 = 2036), which were followed by those unique up-regulation to N (P0001 = 1484). Among the number of uniquely down-regulated genes, SN had the highest number (P0020 = 884) and followed by N (P0002 = 672). Gene Ontology enrichment analyses of the genes unique to each genotype and common DEGs are available in Supplementary File S2.

Principal component analysis indicated that over 78% of variance were explained collectively by the first three principal components (PC1, PC2, and PC3). PC1 explained 50% variance related to cold treatment and clearly separated cold treated samples from the controls (Figure 1C and Supplementary Figure S1A). PC2 explains over 18% of variance in the differences between the two pairs NILs, and to some degree the between winter-habit and spring-habit as well (Figure 1C and Supplementary Figure S1B). PC3 explained 9% variance mainly associated with difference between winter-habit and spring-habit (Supplementary Figure S1).

Within each pair of NILs, we sought to understand the difference between genotypes of winter-habit (WM and N) and spring-habit (M and SN), and also the similarity and difference in gene expression profiles between the two pairs of NILs (N vs. SN; WM vs. M). Under cold treatment, we identified 1515 up-modulated genes and 2085 down-modulated between the winter-habit Norstar as compared to its spring-habit counterpart spring Norstar (C*1, C*2, Supplementary Figure S2A), the majority of which were unique to the N and SN pair (C01 = 1284, 85%; C02 = 1883, 90%). In comparison, the contrast between winter Manitou and Manitou was smaller (C1* = 360, C2* = 719; C10 = 171, 48%; C20 = 475, 66%). The disparity in the number of DEGs between these two pairs of NILs appeared to be related to the difference in freezing tolerance (delta LT50) between winter-habit and spring habit genotypes in each NIL. The delta LT50 is 8.7 between N and SN, but 4.9 between WM and M. Between the two pairs of NILs under cold treatment, they shared 189 genes up- and 202 down-modulated genes of winter-habit genotypes *versus* their respective spring-habit counterparts (C11, C22). In the control samples, the difference between the winter-habit and the spring-habit genotypes within each pair of NILs were not as drastic as those under cold treatment (Supplementary Figure S2B). Collectively, between the winter-habit and spring-habit genotypes, we found 4246 DEGs when treated with cold (C**), but 1898 DEGs in the controls (K**).

Genes specific to winter-habit and spring habit

With regard to the winter-habit specific genes, we were particularly interested in those that were commonly differentially expressed in both winter-habit genotypes (N and WM), but not in either of the spring-habit genotypes (SN and M) when they were subjected to cold treatment. These genes could be represented by DEFE patterns P0101 (191 genes) and P0202 (144 genes) for up- and down-regulation, respectively. The genes up- or down-regulated in spring-habit, but not winter-habit genotypes as a result of cold treatment were represented by P1010 (335 genes) and P2020 (48 genes). Under cold treatment, up- or down-modulated in the winter-habit when compared with their spring-habit NIL pair (C11 = 189, C22 = 202), but not in the controls (K00), could be considered as supporting evidence of functional significance in low temperature adaptation. Integrating these three series of DEFE patterns, 63 genes were found to be up-regulated by cold, specific to both winter-habit genotypes (P0101∩C11∩K00 = 63, Table 2), while seven genes were down-regulated (P0202∩C22∩K00 = 7). On the contrary, 64 genes were found to be up-regulated by cold, specific to both spring-habit genotypes (P1010∩C22∩K00 = 64), while two genes were down-regulated (P2020∩C11∩K00 = 2) (*see* Supplementary File S3 tab Lists). These four groups of genes were distinctive in the three dimensional space represented by the first three principal components (PC1, PC2, and PC3 (Figure 2A and Supplementary Table S1).

We named the 63 genes up-regulated specifically in both winter-habit genotypes (P0101∩C11∩K00) as ***winter-habit genes (WHGs)*** and the 64 genes up-regulated specifically in both

TABLE 2 Winter-habit genes.

Gene_ID	Gene name	Gene description
TraesCS5B03G0571900		AAA-ATPase At5g57480
TraesCS4B03G0940000		acid phosphatase 1-like
TraesCS2D03G0746300		amino acid transporter AVT1I-like
TraesCS7A03G1216800		Basic helix-loop-helix dimerisation region bHLH domain containing protein
TraesCS4D03G0229100		Basic-leucine zipper (BZIP) transcription factor family protein
TraesCS3D03G0865700	BGLU42	Beta-glucosidase 42
TraesCS2D03G1058400	CHL	chloroplastic lipocalin-like
TraesCS1B03G1168700	COR413-TM1	Cold acclimation protein COR413-TM1, Cold-regulated 413 inner membrane protein 1, chloroplastic
TraesCS5A03G1113800		Cytochrome P450 family protein
TraesCS4D03G0748700	DEFL8	Defensin-like protein 1
TraesCS6D03G0772500	DHN3	dehydrin DHN4-like
TraesCS4D03G0668900		embryonic protein DC-8-like isoform X1
TraesCS2A03G0994800	ERF039	Ethylene-responsive transcription factor ERF039
TraesCS6D03G0772300		filaggrin-2-like
TraesCS6B03G0877600		galactan beta-1,4-galactosyltransferase GALS1-like
TraesCS2D03G1214900		geraniol 8-hydroxylase-like
TraesCS2A03G0593000		high mobility group nucleosome-binding domain-containing protein 5-like
TraesCS2D03G0347900		Hypothetical conserved gene
TraesCS7D03G0087400		late embryogenesis abundant protein 6-like
TraesCS4A03G0856100		leucine-rich repeat receptor-like protein kinase PEPR1
TraesCS3A03G0832100		Lipase, GDSL domain containing protein
TraesCS1D03G0421900		low-temperature-induced 65 kDa protein-like isoform X1
TraesCS7B03G0198800	LYP6	Lysin motif-containing protein, Pattern recognition receptor, Peptidoglycan and chitin perception in innate immunit
TraesCS1A03G0908000		non-specific lipid-transfer protein 2-like
TraesCS1B03G1066700		non-specific lipid-transfer protein 2-like
TraesCS3D03G0964600		Non-specific serine/threonine protein kinase
TraesCS3A03G1036100		Non-specific serine/threonine protein kinase
TraesCS5A03G0796200		noroxomaritidine synthase 2-like
TraesCS5A03G0796300		noroxomaritidine synthase 2-like
TraesCS5B03G0828400		noroxomaritidine synthase 2-like
TraesCS5B03G0828500		noroxomaritidine synthase 2-like
TraesCS5D03G0752600		noroxomaritidine synthase 2-like
TraesCS1D03G0066300	OEP161	Outer envelope pore protein 16-1, chloroplastic
TraesCS2A03G0069900		Pectinesterase inhibitor domain containing protein
TraesCS2B03G0102400		Pectinesterase inhibitor domain containing protein
TraesCS1B03G0841700		Phosphatidylethanolamine-binding protein PEBP domain containing protein
TraesCS5D03G1149800		phytosulfokine receptor 1-like
TraesCS4A03G0858900		phytosulfokine receptor 2
TraesCS5D03G1149600		phytosulfokine receptor 2-like
TraesCS7A03G0380300		probable apyrase 3
TraesCS5A03G1073800		probable lactoylglutathione lyase, chloroplastic
TraesCS2B03G0580500		Protein of unknown function DUF1218 family protein
TraesCS2D03G0389100	DHFR	putative anthocyanidin reductase
TraesCS7D03G0803900		Seed maturation protein domain containing protein
TraesCS3B03G1352700	GER8	Similar to Germin-like protein 1–3
TraesCS2D03G0826700		Similar to gibberellin receptor GID1L2
TraesCS1B03G0752200		Similar to Glutathione S-transferase GST 41 (EC 2.5.1.18)
TraesCS7D03G0446000		Similar to Pyruvate dehydrogenase E1 alpha subunit (EC 1.2.4.1)

(Continued on following page)

TABLE 2 (*Continued*) Winter-habit genes.

Gene_ID	Gene name	Gene description
TraesCS7A03G0517200		TB2/DP1 and HVA22 related protein family protein
TraesCS2B03G0488200	GL7	TON1 RECRUIT MOTIF (TRM)-containing protein, Regulation of grain size and shape
TraesCS2A03G0367000	GL7	TON1 RECRUIT MOTIF (TRM)-containing protein, Regulation of grain size and shape
TraesCS5A03G0532400	MYB20	Transcription factor MYB20
TraesCS4B03G0828000		uncharacterized protein LOC123093546 isoform X1
TraesCS5D03G0225200		uncharacterized protein LOC123124437
TraesCS7B03G0888400		uncharacterized protein LOC123162191
TraesCS7D03G0109900		uncharacterized protein LOC123168984
TraesCS2A03G0862400		Zinc finger, RING/FYVE/PHD-type domain containing protein
TraesCS2A03G1086200		
TraesCS3D03G0047400		
TraesCS4D03G0738600		
TraesCS5A03G0028100		
TraesCS5A03G0564900		
TraesCS5A03G1156200		

spring-habit genotypes as ***spring-habit genes (SHGs)***. Gene Ontology enrichment analysis indicated that WHGs were highly represented by genes with functions in cold acclimation, embryo development ending in seed dormancy, regulation of monopolar cell growth, response to abscisic acid, response to lipid, oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, and heme binding among others (*see* Supplementary File S3 tab GO_P0101∩C11∩K00). On the other hand, four of the seven genes suppressed by cold treatment in the two winter-habit genotypes had GO annotations that were enriched with calcium-dependent phospholipid binding (TraesCS1B03G0711800), passive transmembrane transporter activity (TraesCS5B03G0835100), fatty acid biosynthetic process (TraesCS7D03G0081000), and glucosidase activity (TraesCS7A03G0020800) that includes sucrose alpha-glucosidase activity (GO:0004575) and beta-fructofuranosidase activity (GO:0004564) (*see* Supplementary File S3 tab GO_P0202∩C22∩K00).

Under cold treatment, the 64 up-regulated genes specific in the two spring-habit genotypes (P1010∩C22∩K00) were enriched with phosphatidylethanolamine binding, photoperiodism, flowering, oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, S-adenosylmethioninamine biosynthetic process, RNA polymerase II transcription regulatory region sequence-specific DNA binding, and inositol 3-alpha-galactosyltransferase activity among others (*see* Supplementary File S3 tab GO_P1010∩C22∩K00). Genes down-regulated in the spring-habit genotypes were enriched with cell redox homeostasis, and protein-disulfide reductase activity (*see* Supplementary File S3 tab GO_P2020∩C11∩K00).

Both WHGs and SHGs groups were enriched with genes encoding oxidoreductase enzyme activities. In this regard, the 63 WHGs included genes encoding a geraniol 8-hydroxylase-like (TraesCS2D03G1214900), an indole-2-monooxygenase-like isoform X1 (TraesCS5A03G1113800), and five noroxomaritidine synthase 2-like (TraesCS5A03G0796200, TraesCS5D03G0752600, TraesCS5B03G0828500, TraesCS5A03G0796300, and TraesCS5B03G0828400); these seven genes acted on paired donors (GO:0016705). Whereas, the 64 SHGs include genes encoding a lipoxygenase (TraesCS5D03G0104900, EC:1.13.11.12), a linoleate 9S-lipoxygenase (TraesCS6B03G0405500, EC:1.13.11.58), and two uncharacterized proteins both involved in oxidoreductase activity and metal ion binding (TraesCS6D03G0269100 and TraesCS6D03G0269200); these four genes acted on single donors (GO:0016701). We thus further looked into the redox pathway and uncovered four peroxiredoxin genes for the DEG list (Figure 3). They were all highly expressed in the most winter hardy Norstar under cold treatment, but the three gene encoding peroxiredoxin-2E-2 were down-regulated in spring Norstar.

Genes associated with cold hardiness

We scaled cold hardiness of each genotype based on their LT50 value (Li et al., 2021) according to the following formula (Table 3):

$$H = LT50/-25 \tag{1}$$

where, H is termed as ***cold hardiness index***, LT50 is the half lethal temperature of a genotype, −25°C is the temperature below which most wheat genotypes would perish (Skinner and Garland-

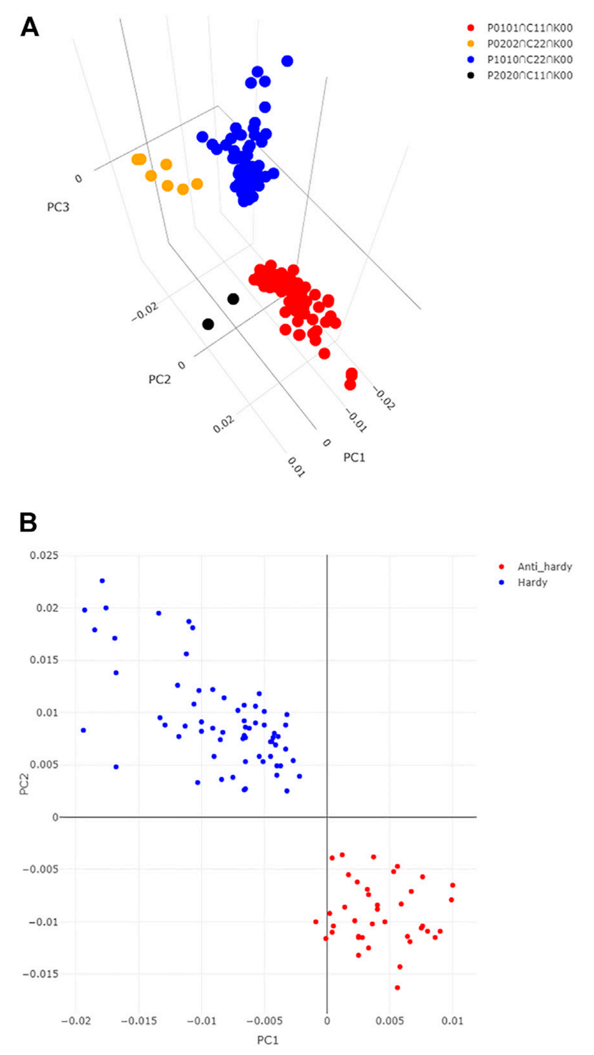


FIGURE 2
Distinction of genes associated with cold acclimation from the others. **(A)** Distinction of the 63 WHGs from the other three groups of DEGs identified in DEFE analysis as revealed by their scores of the first three principal components. **(B)** Distinction of the 64 cold hardy genes from the 39 anti-hardy genes revealed by their scores of the first two principal components. Where, PC1, PC2, and PC3 are the principal components 1, 2, and 3.

Campbell, 2008; Li et al., 2021). We considered a gene to be associated with cold hardiness, and therefore defined as ***cold hardy gene*** when the log2 fold change values and the expression values of the four genotypes under cold treatment were both significantly correlated ($p \leq 0.05$) with the defined cold hardiness (Table 3); in addition, the expression values in all four genotypes under cold treatment were higher than their controls. Conversely, a gene would be considered ***anti-hardy*** when *1*) it was significantly down-regulated by cold in the extreme hardy genotype Norstar, and *2*) both log2FC and expression values were negatively correlated with the defined cold hardiness index among the four genotypes (Table 3). From the 12,676 DEG, 64 emerged as cold-hardy genes (Table 4) and the anti-hardy genes accounted 39 (Supplementary File S4; also in Supplementary File S1, “1” and “−1” in tab DEGs col AH). These two group of genes had distinctive variance distribution in the expression profiles as revealed in PC1 and PC2 two dimensional space (Figure 2B).

The differences in cold hardiness index between N and SN (0.348, or delta LT50 = 8.7) was 1.8 time of that between winter Manitou and Manitou (0.196 or delta LT50 = 4.9). We compared the same difference in fold changes of differential expression between the two pairs of NILs and found that 29 of the 64 cold hardy genes had the same or larger extent of difference in their differential expression between the two pairs. In contrast, none in the 39 anti-cold-hardy genes had such an extent of difference. Orthology search against *Brachypodium distachyon*, *Oryza sativa*, and *Arabidopsis thaliana* indicate that the 64 cold hardy genes include genes encoding auxin responsive protein IAA31-like and auxin-binding protein 4; an early nodulin, *OSENOD93B*; high-affinity nitrate transporter, *NRT2.4* that involves both nitrate transport and auxin signalling; HR-like lesion-inducer family protein, lysin motif-containing protein, *LYP6*; no apical meristem (NAM) protein domain containing protein and others (Table 4 and Supplementary File S4).

There were four cold hardy genes (6.25%) among the 63 WHGs, and the percentage of total 64 cold hardy genes in the entire list of 12,676 DEGs was 0.50%. Therefore, the WHGs contained over 10 time enrichment of cold hardy genes as compared to the entire list of DEGs. These four genes included a galactan beta-1,4-galactosyltransferase (TraesCS6B03G0877600, EC:2.4.1.-), a salt-induced YSK2 dehydrin 3 (*DHN3*, TraesCS6D03G0772500), and a pyruvate dehydrogenase E1 component subunit alpha 2, mitochondrial isoform (TraesCS7D03G0446000, EC:1.2.4.1) and an unknown gene (TraesCS3D03G0047400). Gene ontology enrichment analysis indicated that the 64 cold hardy genes were enriched with cellular response to water deprivation and cold, chloroplast mRNA processing, kinase inhibitor activity, cysteine-type endopeptidase inhibitor activity, auxin binding among others detailed in Supplementary File S4 tab GO_Hardy.

Gene-lipid association network analyses

Since membrane lipids are known to be altered in response to cold stress and in cold acclimation processes (Li et al., 2015; Li et al., 2021), we combined the 12,676 DEGs with 224 lipid traits in association network and clustering analyses to explore associations between transcriptome and lipidome.

Correlation analyses between all 12,676 DEGs and 224 lipid traits together with five experimental conditions (cold treatment, winter-habit, spring-habit, winter-habit genotypes treated with cold, spring-habit genotypes treated with cold) indicated that majority (58 and 62 genes, respectively) of the 63 WHGs and

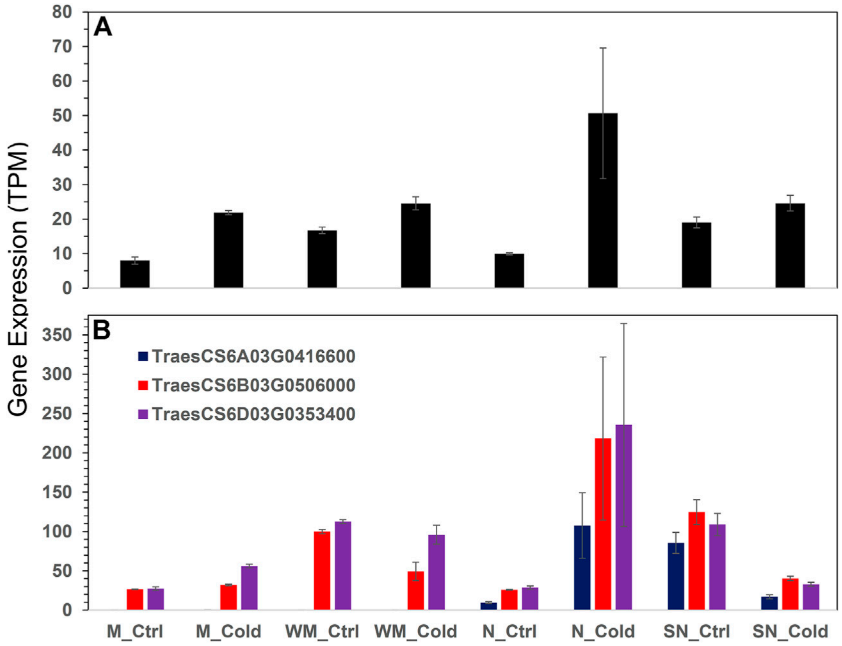


FIGURE 3
Expression of peroxiredoxins across all samples. **(A)** peroxiredoxin-2F, mitochondrial isoform X1, **(B)** peroxiredoxin-2E-2, chloroplastic-like. Error bars are one standard error of the mean of three replicates.

TABLE 3 The cold hardiness indices of the four genotypes in this study.

	M	WM	SN	N
Index	0.332	0.528	0.52	0.868

64 SHGs were positively correlated with the respectively designated experimental conditions (Supplementary File S3 tab geneTraitCor_R). These two groups of genes correlate with distinctive lipidomics profiles (Table 5). For example, the WHGs were positively correlated with phosphatidylglycerol lipids PG(34:3), PG(34:2), and PG(36:6), most of monogalactosyldiacylglycerol (MGDG) and digalactosyldiacylglycerol (DGDG), total phosphatidylcholine (PC) and total phosphatidylethanolamine (PE), but negatively with PG(34:4), PG(34:1) and total PG. Whereas, high percentage the SHGs were significantly correlated with lysophosphatidylcholines (LPCs) and phosphatidylinositols (PIs).

By using topology overlap matrix in the WGCNA R package (Langfelder and Horvath, 2008), the network association degree and cluster membership of each gene were obtained and presented in Supplementary File S1. Correlation analyses were performed between each gene cluster and each lipid trait in addition to experimental design. Among the 50 clusters generated, six were significantly correlated (p < 0.05) with cold treatment to the two winter-habit genotypes, and together contained 71% of the WHGs (Supplementary File S1 tab ClusterTraitCor). The remaining 18 WHGs were in another large cluster less significant correlated with cold treatment to the two winter-habit genotypes (p = 0.067). Similarly, other eight clusters were significantly correlated to cold treatment to the two spring-habit genotypes and contained all 64 SHGs that were found through the aforementioned DEFE analysis. The cluster membership of these two groups was distinct. The association network analysis showed that none of the genes in the two groups were directly associated, through neither their immediate nor secondary neighboring nodes (Supplementary File S3 tabs WHGs_nodes and SHGs_nodes), thereby implying distinct functional space between the two groups of genes and lipids.

Similar analysis between the cold hardy genes and anti-hardy genes were conducted. These two groups were also well separated by distinctive lipidomic profiles, subnetworks, and cluster membership. Interestingly, all 64 cold hardy genes were positively collected with PG(34:3) and PG(36:6). In addition, the majority (64%) of the cold hardy genes were negatively correlated with PG(34:4) (Supplementary File S4 tab geneTraitCor_R).

As for network analysis, we extract the top 1% of the topology overlap matrix, which consists of 6,743 nodes connected with

TABLE 4 Cold-Hardy genes.

Gene_ID	Gene name	Gene description
TraesCS5A03G1126300		actin-depolymerizing factor 3
TraesCS7A03G0546400		alpha-1,3-arabinosyltransferase XAT3-like
TraesCS5A03G0248600		auxin-binding protein 4
TraesCS5D03G0169500		auxin-responsive protein IAA31-like
TraesCS4A03G0266600	CK1	CBL-interacting protein kinase 31-like isoform X2
TraesCS1A03G0304900		Conserved hypothetical protein
TraesCS7B03G1059600		Conserved hypothetical protein
TraesCS1D03G0737600	Oc2*	cysteine proteinase inhibitor
TraesCS3D03G0413100		cysteine proteinase inhibitor 12-like
TraesCS1B03G0882400	Oc2*	cysteine proteinase inhibitor-like
TraesCS3B03G0987700	P5CS2	delta-1-pyrroline-5-carboxylate synthase 2-like
TraesCS4B03G1005300		DIBOA-glucoside dioxygenase BX6-like
TraesCS7B03G0073300		early nodulin-93-like
TraesCS7D03G0286800	RAP2-9	ethylene-responsive transcription factor RAP2-9-like
TraesCS6B03G0877600		galactan beta-1,4-galactosyltransferase GALS1-like
TraesCS4A03G0679200		glucan endo-1,3-beta-glucosidase 7-like
TraesCS1D03G0221500		glutathione S-transferase 4-like
TraesCS7D03G0069600		Glutathione transferase
TraesCS1B03G0654800	SGPP	haloacid dehalogenase-like hydrolase domain-containing protein Sgpp
TraesCS4A03G0180900		Harpin-induced 1 domain containing protein
TraesCS7A03G1041100	NRT2.4	High-affinity nitrate transporter, Nitrate transport, Auxin signalin
TraesCS2A03G0054400		HR-like lesion-inducer family protein
TraesCS2B03G0082600		HR-like lesion-inducer family protein
TraesCS2D03G0055600		HR-like lesion-inducer family protein
TraesCS2D03G0812700	HST	Shikimate O-hydroxycinnamoyltransferase
TraesCS2B03G0514500		interferon-related developmental regulator 2-like
TraesCS2D03G0404700		interferon-related developmental regulator 2-like
TraesCS5A03G0564300		low molecular mass early light-inducible protein HV90, chloroplastic-like
TraesCS5A03G0864400		low temperature-induced protein lt101.2-like
TraesCS7A03G0398900	LYP6	lysM domain-containing GPI-anchored protein LYP6-like
TraesCS7D03G0383800	LYP6	lysM domain-containing GPI-anchored protein LYP6-like
TraesCS5D03G1068400		multiple inositol polyphosphate phosphatase 1
TraesCS3D03G0949700	ONAC041	NAC domain-containing protein 83-like
TraesCS3B03G0141200		non-specific lipid-transfer protein 4.1-like
TraesCS5B03G1309300		Nucleotide-diphospho-sugar transferase domain containing protein
TraesCS5B03G1309400		Nucleotide-diphospho-sugar transferase domain containing protein
TraesCS2A03G1236100		pectinesterase inhibitor 8-like
TraesCS4A03G1138500		probable glutathione S-transferase GSTU6
TraesCS1D03G0686300		probable membrane-associated kinase regulator 4
TraesCS2B03G0690600		probable receptor-like protein kinase At1g49730 isoform X1
TraesCS4D03G0736900		probable serine/threonine-protein kinase At1g01540
TraesCS5A03G0968800		protein AE7-like 1
TraesCS2D03G0911500		protein RKD5-like
TraesCS5D03G0561000		putative ripening-related protein 2
TraesCS7B03G0271100		pyruvate dehydrogenase E1 component subunit alpha-2, mitochondrial-like
TraesCS7D03G0446000		pyruvate dehydrogenase E1 component subunit alpha-2, mitochondrial-like
TraesCS1D03G0944000		ras-related protein RABA1f-like
TraesCS1B03G0360800		RING finger and transmembrane domain-containing protein 2-like

(Continued on following page)

TABLE 4 (*Continued*) Cold-Hardy genes.

Gene_ID	Gene name	Gene description
TraesCS6D03G0772500	DHN3	Salt-induced YSK2 dehydrin 3
TraesCS2B03G1308500	SAPK1	Serine/threonine protein kinase, Hyperosmotic stress respons
TraesCS5B03G1367500		serine/threonine-protein phosphatase PP1-like
TraesCS4B03G0915000		subtilisin-like protease 4
TraesCS1B03G0874000		transcription factor GAMYB-like isoform X1
TraesCS5B03G0149900		tuliposide A-converting enzyme 2, chloroplastic-like
TraesCS2D03G0234900		UDP-glucuronic acid decarboxylase 4-like
TraesCS7A03G0941200		uncharacterized membrane protein At1g16860-like
TraesCS7B03G0786000		uncharacterized membrane protein At1g16860-like
TraesCS7D03G0906100		uncharacterized membrane protein At1g16860-like
TraesCS5D03G0230900		uncharacterized protein LOC123119728
TraesCS2A03G0325900		uncharacterized protein LOC123187869
TraesCS3B03G0014400		
TraesCS3D03G0047400		
TraesCS5A03G0434000		
TraesCS7A03G0474100		

825,776 edges. For the purpose of this study, we focus on the following subnetworks relevant to cold acclimation. We first defined the association strength (AS) of a subnetwork by using average connection degree of all nodes in the sub-network normalized by total number of nodes in the subnet:

$$AS = \frac{\text{Average connection degree}}{\text{number of nodes in the subnet}} \quad (2)$$

WHG subnet

Fifty-one genes (81%) among the 63 WHGs were associated with at least one other gene in the group with an average connection degree of 33.5. Thus, the overall AS of the WHG subnet was 0.64. The top hub genes included a lipase, GDSL domain containing protein, orthologous to rice gene *OsGELP26* (Os01g0827700, connection degree = 44), a HVA22-like protein orthologous to *OsEnS-122* (Os08g0467500, 44), a galactan beta-1,4-galactosyltransferase (EC:2.4.1.-, 44), an AAA-ATPase orthologous to *At5g57480* (TraesCS5B03G0571900, 43), two pectinesterase inhibitor domain containing protein (Os04g0106000, 44 and 42), two non-specific serine/threonine protein kinases (TraesCS3D03G0964600, TraesCS3A03G1036100, both 42), a defensin (*DEFL8*, Os03g0130300, 42), pyruvate dehydrogenase E1 component subunit alpha (Os06g0246500, EC:1.2.4.1, 41), transcription factor *MYB20* (AT1G66230, 41), a Glutathione S-transferase *GST* (EC:2.5.1.18, 41), a pathogenesis-related transcriptional factor and ERF domain containing protein OsERF#034 (Os04g0550200, 40), and a cold acclimation protein *COR413-TM1* (Os05g0566800, 37). *COR413-TM1* was directly associated with all other hub genes mentioned above (Figure 4). A homoeolog of *COR413-TM1* on chromosome 1A was directly associated with PG(34:3). Both homoeologs of *COR413-TM1* were highly expressed in the two winter-habit genotypes (Figure 5). More details are available in Table 2 (Supplementary File S3 tab WHGs_nodes).

The associations among the 64 SHGs were very loose; only 27 genes (42%) have a direct neighbor within the group and spread over two subnets (Supplementary Figure S3). Collectively among the 27 genes, the association strength was 0.12 (Supplementary File S3 tab SHGs_nodes).

Cold hardy subnet

Similarly, we investigated the subnet of cold hardy *versus* anti-hardy genes. Among the 64 cold hardy genes, 61 (95%, Table 4 and Supplementary Figure S4A) were inter-associated with at least another gene within the subnet and have an association strength of 0.41. In addition, the cold hardy subnet had four points of contact with the WHGs subnet as described above. All the four points of contact were the hub nodes in both subnets. Phosphatidylglycerol lipids PG(34:3) and PG(36:6) were hub nodes in the cold hardy subnet with connection degrees of 22 and 10, respectively. Only 13 (33%) of the 39 anti-hardy genes have direct association with another gene within the group and they all were directly associated with PG(34:4) (Supplementary Figure S4B). More details are available in Supplementary File S4 tabs ColdHardy_nodes and AntiHardy_nodes). Two distinctive schools of network nodes were evident, one was represented by PG(34:3) and PG(36:6) and consisted of genes closely associated with cold hardiness, while the other was

TABLE 5 Number of winter-habit genes (left[a]) or spring-habit genes (right) correlated with respective lipid species.

Winter habit genes			Spring habit genes		
	Pos	neg		Pos	neg
DGDG(34:2)	0	25	PG(36:1)	17	0
DGDG(34:1)	0	13	LPG(18:2)	23	0
DGDG(36:2)	0	11	LPC(16:0)	53	0
DGDG(36:1)	0	12	LPC(18:3)	35	0
MGDG(34:1)	0	13	LPC(18:1)	9	0
PG(34:4)	0	6	LPC(20:1)	14	0
PG(34:3)	53	0	Total_LysoPC	38	0
PG(34:2)	49	0	PC(34:2)	0	20
PG(34:1)	0	15	PC(38:2)	0	34
PG(36:6)	51	0	PE(32:1)	47	0
Total_PG_All	0	50	PE(32:0)	45	0
PC(34:3)	42	0	PE(36:2)	10	0
PC(36:5)	9	0	PE(36:1)	28	0
PC(36:4)	49	0	PI(34:3)	28	0
PC(38:5)	14	0	PI(34:1)	46	0
PC(38:4)	36	0	PI(36:6)	49	0
PC(38:3)	51	0	PI(36:5)	38	0
PC(40:5)	15	0	PI(36:3)	43	0
PC(40:4)	19	0	PI(36:1)	36	0
PC(40:2)	6	0	Total_PI	27	0
Total_PC	40	0	PS(36:5)	10	0
PE(32:3)	38	0	DAG(16:0/16:0)	49	0
PE(36:6)	7	0	DAG(18:3/16:1)	0	41
PE(36:5)	9	0	TAG(50:4)_16:1_acyl_containing	53	0
PE(36:4)	37	0	TAG(52:7)_16:1_acyl_containing	44	0
PE(38:4)	20	0	TAG(52:6)_16:1_acyl_containing	45	0
PE(40:3)	39	0	TAG(52:5)_16:1_acyl_containing	27	0
PE(40:2)	41	0	Total_TAG_16:1_acyl_containing	39	0
PE(42:4)	32	0	TAG(52:8)_18:3_acyl_containing	21	0
PE(42:3)	23	0	TAG(52:6)_18:3_acyl_containing	17	0
PE(42:2)	8	0	TAG(52:5)_18:3_acyl_containing	21	0
Total_PE	7	0	TAG(52:4)_18:3_acyl_containing	44	0
PI(34:4)	0	12	MGDG(36:1)	24	0
PI(36:4)	0	15	LPC(18:0)	45	0
PS(34:3)	0	58	LPE(16:1)	17	0
PS(34:2)	0	16	LPC(16:1)	27	0
PS(36:6)	37	0	LPE(18:1)	28	0
PS(36:4)	0	40	TAG(48:1)_16:1_acyl_containing	9	0
PS(38:5)	0	50			
PS(38:2)	0	10			
PS(42:3)	38	0			
PS(42:2)	21	0			
Total_PS	0	13			
DAG(18:2/16:1)	0	38			
TAG(54:8)_18:3_acyl_containing	6	0			
TAG(54:7)_18:3_acyl_containing	6	0			
TAG(52:4)_18:2_acyl_containing	14	0			

(Continued on following page)

TABLE 5 (*Continued*) Number of winter-habit genes (left[a]) or spring-habit genes (right) correlated with respective lipid species.

Winter habit genes		Spring habit genes
TAG(52:3)_18:2_acyl_containing	12	0
PS(40:1)	0	21
DAG(18:2/16:3)	44	0
DAG(18:1/16:3)	45	0

[a]This table contains the lipid species having correlation with more than five genes. Otherwise, all data are available in Supplementary File S3 Tab geneTraitCor_R.

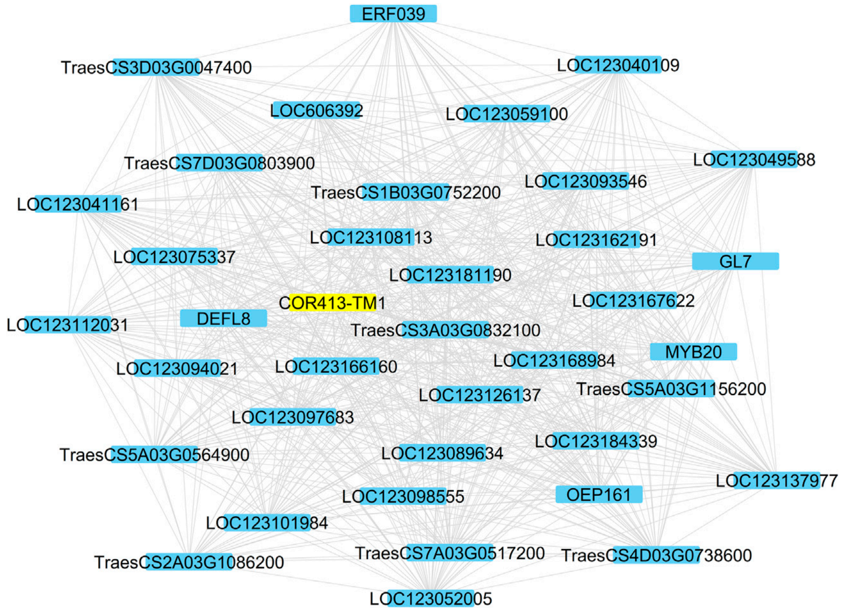


FIGURE 4
The cold acclimation genes directly associated with the Cold acclimation protein *COR413-TM1* highlighted. Details are available in Supplementary File S3.

represented by PG(34:4) and consisted of genes closely associated with anti-hardy (Supplementary File S5).

Vernalization subnet

The list of 12,676 DEGs included eight vernalization genes, three *VRN1* (TraesCS5A03G0935400, TraesCS5B03G0986000, and TraesCS5D03G0894800), four *VRN2* (TraesCS4B03G0958300, TraesCS4D03G0834500, TraesCS4D03G0834600, and TraesCS5A03G1265900), and one *VRN3* (TraesCS7B03G0031800) (Figure 6). The *VRN1* genes were highly up-regulated by cold in all four wheat genotypes, while their expression were higher in the two spring-habit genotypes (M and SN) than the winter-habit ones. The *VRN3* gene was up-regulated by cold treatment in the two spring-habit genotypes, while there was no effect in the two winter-habit genotypes. The *VRN2* were generally down-regulated by cold treatment. The *VRN2* and *VRN3* genes were not involved in the gene association network. Collectively, there were 214 genes in direct association with the three *VRN1*s, and they were interconnected to form a highly cohesive network with AS at 0.83. Nevertheless, these *VRN1* genes had no association even at the secondary neighborhood with either WHGs or cold hardy genes. From the 64 SHGs, one appeared in the *VRN1* immediate neighborhood and 42 other genes in the secondary neighborhood (Supplementary File S7). The *VRN-B1* (TraesCS5B03G0986000) had a direct association with the phospholipid-transporting ATPase (ALA1, EC:7.6.2.1, TraesCS4B03G0491700) as a single lipid gene in the immediate neighborhood of *VRN1* genes. There were other 19 lipid genes in the secondary neighborhood of *VRN1* genes.

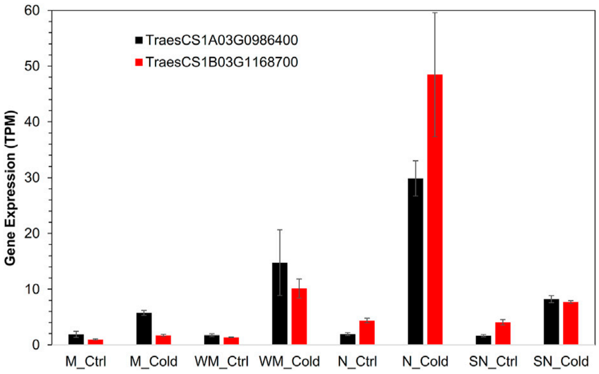


FIGURE 5
Expression of two homoeologs of *COR413-TM1* gene in different experimental conditions.

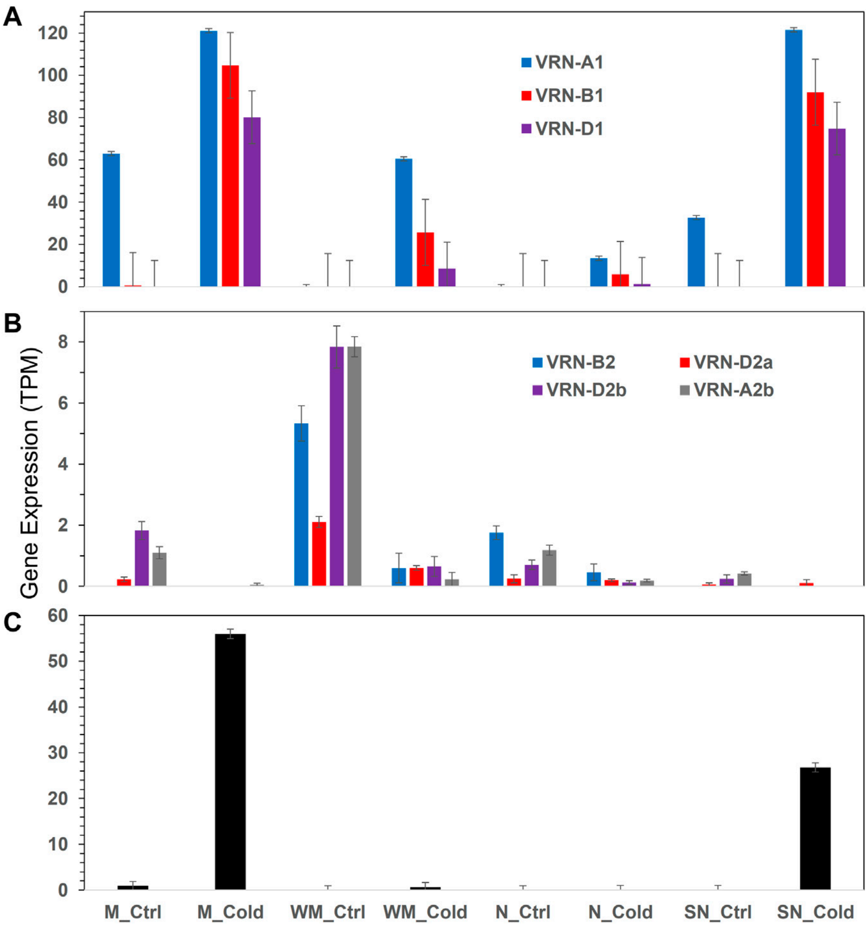


FIGURE 6
Expression of vernalization genes in different experimental conditions. **(A)** *VRN1*, **(B)** *VRN2*, **(C)** *VRN3*. Error bars are one standard error of the mean of three replicates. More details are available in Supplementary File S7.

Discussions

Overview—Integrative computational insights

Cold acclimation are investigated by integration of transcriptomics and lipidomics with various computational approaches including differential expression feature extraction, principal component analysis, correlation analysis, and gene-association network analyses. The differential expression feature extraction approach is a simple and effective pattern recognition method to find expression patterns in various conditions. Through integrating three differential expression feature extraction schemes, 63 winter-habit genes and 64 distinctive spring-habit genes are found. Correlation analysis reveals 64 cold hardy genes and 39 distinctive anti-hardy genes. The integration of transcriptomics and lipidomics analyses identifies two distinctive schools of network nodes (Supplementary File S5). The dimension reduction through principal component analysis is able to explain the majority of variance associated with cold treatment, cold hardiness, between winter-habit and spring-habit, and between two schools of network nodes in the reduced one, two, or three dimensional space represented by the first three principal components: PC1 for cold treatment, PC1+PC2 for cold hardiness and for two schools of network nodes, and PC1+PC2+PC3 for winter-habit and spring-habit. The distinction between the contrasting groups in each scenario is confirmed by integration of these methods. For example, the variance distribution with regard to the contrast between WHGs and SHGs is revealed by the differential expression feature extraction method and confirmed by principal component analysis, lipidomics association, and gene association network propagation. From gene association network perspective, the WHGs are highly associated among themselves as well as with others outside of the group. The association among the 64 SHGs, on the other hand, were very loose as indicated by the proportion of genes involved in the network and the network association strength. The same analogy is applied to the scenario between cold hardy and anti-hardy genes. These three scenarios of knowledge discovery indicate that there is no way of one-size-fit-all approach in the computational pipeline. Each case would have to be designed according to the characteristics of variance distribution in combination with domain knowledge. These pairwise contrasting analysis reveals that WHGs and cold hardiness are unique and yet they are inter-related to certain extent. They are two innate concerted efforts of plants to deal with cold stress.

Limitations and complements

The four genotypes of reciprocal NILs used in this study inspire significantly to the design for this and earlier experiments and indeed help achieving much progress in the field of cold tolerance research in cereal plants (e.g. Limin and Fowler, 2002; Li et al., 2015, 2018, 2021). Nevertheless, the success of computational investigation requires significant sample size, balanced distribution of sample types, and data consistency within each type of samples. The most obvious limitation to the methods and analysis in this study is the small sample size of this dataset, which creates high imbalance between the sizes of sample space and the number of genes, known as the curse of dimensionality. Principle component analysis is a typical method for dimension reduction and able to explain the main variance in this study in one, two, or three dimensional spaces and reveals the distinction between contrasting groups.

The limitation of small sample size is most obvious in network analysis of this study, the size of eight samples is below the conventional necessity for a successful systemic network study such as the AraNet, which comprises from many distinct types of interactions, and millions of experimental or computational observations from diverse data types over decades of studies in *Arabidopsis thaliana* (Lee et al., 2010; Lee and Lee, 2017). To complement this limitation, we take the top 1% from the topology overlap matrix to reduce the false positive. Taking such high stringency would certainly sacrifice information. For example, the three *VRN1* homoeologs have similar network connections and high correlation in expression profile between themselves and in the same cluster; technically, they should be directly inter-connected too. But actually, they are not under the current selection criteria. The homoeologs of *COR413-TM1* are also in similar situation; the one in B sub-genome (TraesCS1B03G1168700) is not directly associated with PG(34:3), while the one in A sub-genome (TraesCS1A03G0986400) is. Therefore, caution should be taken at the interpretation of the result. This is complemented by network propagation in this study and such complement enables discovery of the fact that they share similar neighborhood nodes. Applying the small world social circle theory in humanity research, similar to the backbone theory used in WAGNA topology overlap matrix (Langfelder and Horvath, 2008), achieves the overall success of network analysis in this study. Such result is further strengthened through integration of strengths of other methods applied.

The prime condition in this study is cold treatment, which is reflected in two contrasting pairs (WHGs *versus* SHGs and cold hardy *versus* anti-hardy) which are revealed by one computational method and supported by PCA and at least one more other method in this study. The cohesiveness of group membership is reflected by an association strength in the membership in each of these groups and is contributed by environmental, physiological, and/or genetic factors. Both groups of WHGs and cold hardy genes have higher membership involvement and association strengths than their respective counterparts. Finally, the vernalization subnetwork is highly relevant to the genetic factor contributed by the vernalization through the recessive allele *vrn1* associated with *VRN-A1* locus. The association strength of vernalization subnetwork is 0.83, i.e., each gene is directly associated with 83% in immediate neighborhood of *VRN1*, which indicates the extent of contribution by genetic contribution to over winter cold/freezing tolerance of wheat plants.

Insights into cold acclimation

Cold acclimation is a complex system and the 63 WHGs encompass a wide spectrum of biological processes. Gene association network analysis of WHGs subnet reveals that many hub genes in this group are directly associated collectively with over 70% of the genes in the group with an overall association strength of 0.64. This indicates that these genes coordinate in concerted manner to confer the common goal of cold tolerance.

Signaling of cold stress

In plants, the calcineurin B-like protein (CBL) family represents a group of calcium sensors and plays a pivotal role in decoding calcium transients by specifically interacting with and regulating a family of protein kinases (*CIPKs*). *CIPKs* is known to confer cold stress tolerance in cold acclimated *Arabidopsis thaliana* (Aslam et al., 2022), pepper, and tomato (Ma et al., 2022). Two CIPKs are among the hub genes of WHGs subnet in this study (connection degree = 42) and they are highly expressed in both winter-habit genotypes (WM and N) when treated with cold, but barely any expression in all other samples. They are directly associate with all hub genes in the subnet.

Involvement of carbohydrate metabolism in cold acclimation

Beta-glucosidases are the enzymes that catalyze the hydrolysis of terminal, non-reducing β-D-glucosyl residues from a variety of glucoconjugates which include glucosides, oligosaccharides, and 1-O-glucosyl esters (Godse et al., 2021). Beta-glucosidase is a rate-limiting enzyme that is involved in the hydrolysis of cellulose, affects cell wall structure, and plays a key role in cell adaptations to the physical deformations caused by cold stress (Sun et al., 2021). Beta-glucosidases hydrolyze inert precursors to release antioxidant substances under various abiotic stresses in rice (Opassiri et al., 2007). Expression of beta-glucosidase gene is induced in response to low temperature in chickpea (Khazaei et al., 2015). After cold acclimation, beta-glucosidase is require for freezing tolerance in *Arabidopsis thaliana* (Thorlby, 2004). In this study, the expression of the beta-glucosidase gene is upregulated to different extent in all four genotypes under cold stress and is much higher (>3 times) in both winter-habit genotypes than in the spring-habit genotypes.

Integrity of plasma membrane

Expression of lipocalins and lipocalin-like proteins in wheat (*Triticum aestivum*) is known to be associated with the plant's capacity to develop freezing tolerance, cold acclimation induces a high accumulation of temperature-induced lipocalin TaTIL-1 in an enriched plasma membrane fraction of cold-acclimated wheat but not in nuclei (Charron et al., 2005). The chloroplastic lipocalin AtCHL is known to prevent lipid peroxidation and protect Arabidopsis against oxidative stress (Levesque-Tremblay et al., 2009) and is required for sustained photoprotective energy dissipation (Malnoë et al., 2018). In this study, the chloroplastic lipocalin-like gene (CHL) is highly up-regulated by cold in both winter-habit genotypes, but not (or a minor extent of down-regulation) in the two spring-habit genotypes.

Resistance to oxidative stress and cellular detoxification

Plant adaptation to low temperature not only induces lipid desaturation in cellular membranes but also generation of reactive oxygen species (ROS) and changes in redox state (Murata and Los, 1997; Wallis and Browse, 2002). The multifunctional enzymes glutathione S-transferases (GSTs) participate in oxidative stress resistance and cellular detoxification and highly associated with cold stress of Hami melon (Song W. et al., 2021) and pumpkin (Abdul Kayum et al., 2018). There are 92 *GSTs* or *GSTs* like in the DEGs list, the majority of them, including a key hub gene in the WHGs subnetwork (TraesCS1B03G0752200), are significantly upregulated by cold, especially in Norstar and spring Norstar.

The pyruvate dehydrogenase E1 component subunit alpha-2, mitochondrial isoform (*PDH-E1a*, TraesCS7D03G0446000, EC: 1.2.4.1) appears to be a key hub gene in both WHGs and cold hardy subnets of this study. It's homoeolog in chromosome B (TraesCS7B03G0271100) is also a cold hardy gene and up regulated by cold in all four genotypes. The pyruvate dehydrogenase (*PDH*) complex catalyzes the oxidative decarboxylation of pyruvate with the formation of acetyl-CoA, CO2 and NADH. Much of the studies were done with animal in relation to the effect of cold. For example, *PDH* is associated with metabolic rate depression during freezing and anoxia of wood frogs (Al-Attar et al., 2019) and during hibernation of ground squirrel (Herinckx et al., 2017). In plants, *PDH* is found in both chloroplast and mitochondria. The two genes described above encoding mitochondrial isoform in this study is truly up-regulated by cold, related to respiration and anoxia. Whereas, the chloroplast isoform concerns fatty acid synthesis (Li et al., 2021) and photorespiration (Blume et al., 2013). There are three homoeolog genes in this study (TraesCS2A03G0021400, TraesCS2B03G0027300, and TraesCS2D03G0019300) encoding pyruvate dehydrogenase E1 component subunit alpha 3, the chloroplastic isoform (*PDHA1*); they are all down-regulated by cold in all four wheat genotypes (Supplementary File S1).

Upon the cold treatment, the WHGs and SHGs show oxidoreductase activity with incorporation of molecular oxygen. But WHGs act on paired donors (EC1.14.-.-) and are from a family of heme-binding and iron containing enzymes. They catalyze an oxidation-reduction (redox) reaction in which hydrogen or electrons are transferred from reduced flavin or flavoprotein and one other donor; one atom of oxygen is incorporated into one of the donors. This group consists of seven genes including one encoding geraniol 8-hydroxylase-like, an indole-2-monooxygenase-like isoform X1, and five noroxomaritidine synthase 2-like. Their expressions are significant in the two winter-habit genotypes but not otherwise. Whereas, as represented by two lipoxygenases (EC1.13.11.-), SHGs act on single donor (EC1.13.-.-) are from a family of non-heme iron containing enzymes, mostly catalyze the dioxygenation of polyunsaturated fatty acids. It has been shown that low temperature or cold stress induced reactive oxygen species (ROS) production is often accompanied by lipid peroxidation and oxidative damage to cellular membranes (Kim et al., 2013).

A recent study showed that a thylakoid-associated protein, peroxiredoxin Q, is required for the production of t16:1 in chloroplast and photosynthesis systems (Lamkemeyer et al., 2006; Horn et al., 2020), indicating a link between t16: 1 production and redox status. Three genes encoding the chloroplast peroxiredoxin-2E-2 were uniquely induced in Norstar (Figure 3) which might be related to the reduction in t16:1 levels. Also, the relationship between up-regulation of heme-binding proteins and stress tolerance in general, and specific with regard to cold tolerance. In *Arabidopsis thaliana*, the heme-associated protein *AtHAP5A* enhances freezing stress resistance and has significant effects on inhibiting cold-induced ROS accumulation and activating ABA-related genes' expression (Shi et al., 2014).

Transcriptome regulation

Myeloblastosis transcription factors *MYB20* is a key hub gene in the WHGs subnet directly associated with 41 other genes. In *Arabidopsis thaliana*, *MYB20* is well known to acts as a negative regulator of plant response to desiccation and cold stress and its expression is reduced to less than half (Gao et al., 2014). Another study shows transgenic plants overexpressing *AtMYB20* (AtMYB20-OX) enhance salt stress tolerance while repression lines (AtMYB20-SRDX) are more vulnerable to NaCl than wild-type plants (Cui et al., 2013). The expression level of *MYB20* in this study is near 100 folds in the two winter-habit genotypes (WM and N) as compared to respective controls, and also over 10 folds as compared to the two spring-habit genotypes under cold treatment. *MYB20* is involved in the transcriptional network regulating the secondary wall biosynthetic program (Zhong et al., 2008). In addition, MYB proteins activate transcriptional repressors that specifically inhibit flavonoid biosynthesis, which competes with lignin biosynthesis for the aromatic amino acid phenylalanine precursors (Geng et al., 2020)

The COLD REGULATED 314 THYLAKOID MEMBRANE 1 (*COR413-TM1*) is an integral component of chloroplast inner membrane and well-known in cellular responses of plant to cold, water deprivation, cold acclimation and abscisic acid. *COR413-TM1* is characterized to provide normal freezing tolerance in *Arabidopsis thaliana* (Okawa et al., 2008), *Brachypodium distachyon* (Colton-Gagnon et al., 2014) and wheat (Breton et al., 2003). There are two homoeologs of *COR413-TM1* gene among the 12,676 DEGs in this study. The one on A sub-genome (TraesCS1A03G0986400) is a member of School B and a key hub gene in the PG(34:3) subnet (Supplementary File S6). The other on B sub-genome (TraesCS1B03G1168700) is a major hub gene in WHGs subnet and directly associated to all major hub genes in the subnet. Both *COR413-TM1* homoeologs are associated with over 400 DEGs and lipid species in this study.

Vernalization, cold hardiness *etc.*

During the crossing process of the two wheat cultivars, the non-hardy spring wheat Manitou gained the *vrn-A1* loci and became winter Manitou, while the very cold hardy winter-habit Norstar, gained the dominant *Vrn-A1* locus and became spring Norstar. It is interesting to note that the LT50 value of the two NILs (WM and SN) are very close, but the change in LT50 is very different between the two pairs. Spring Norstar has a higher change (reduced by 8.7°C) in LT50 from Norstar, also has a highest number of DEGs as well as unique DEGs to the genotype when subjected to cold treatment. As a contrast, winter Manitou has a lower change (increased by 4.9°C) in LT50 from Manitou, also has a lower number of DEGs as well as unique DEGs to the genotype.

Cold acclimation and vernalization are two major mechanisms for winter survival in wheat (Li et al., 2018). Consistent with previous study on crown tissue, the *VRN1* genes including *VRN-A1*, *VRN-B1* and *VRN-D1* are induced at higher levels after cold treatment in Manitou and spring Norstar than that of in Norstar and winter Manitou. Vernalization requirement duration in winter wheat is controlled by *VRN-A1* at the protein level (Li et al., 2013). This is apparently relevant to the genetic background of vernalization genes in leaf and crown tissues as determined by the dominant allele *Vrn-A1* in Manitou and spring Norstar *versus* the recessive allele *vrn-A1* in winter Manitou and Norstar. As a result, integration of DEFE, PCA, gene association network analysis, and lipidomics analysis as discussed above, the 63 WHGs significantly expressed in winter-habit genotypes, winter Manitou and Norstar, are highly distinctive from the 64 SHGs. Such distinction is no doubt relevant to the genetic background of the four wheat genotypes.

The *VRN2* genes are known to be repressed by cold in cereal plants and the expression *VRN3* is subjected to negatively regulation by *VRN2* (Kim et al., 2009). Thus the down-regulation of *VRN2* in the leaf of this study permits the transient expression of *VRN3* gene. Also, the *VRN1* gene in cereals is known to plays a dual role of both a promoter of *VRN3* and a cold-activated repressor of *VRN2* (Kim et al., 2009). Our result is consistent in this regard.

Concluding remarks

Cold acclimation and vernalization are major strategies for winter survival in wheat. The differential expression feature extraction enables the discovery of a group of 63 WHGs that are significantly expressed in both vernalized winter-habit winter Manitou and Norstar, but not in either Manitou or spring Norstar. These genes are cohesively associated with one another in their local subnetwork and have a distinctive lipidomics association to achieve survival in the cold stress. They encompass a wide spectrum of transcriptional reprograming that involves signaling, maintenance of plasma membrane fluidity and rigidity, cell energy and redox homeostasis, and transcriptional regulation. The phosphatidylglycerol lipids, PG(34:3) and PG(36:6), appear to be well associated with majority of these WHGs including *COR413-TM1*, which play an integral role in chloroplast inner membrane and the well-known in cellular responses of plant to cold, water deprivation, cold acclimation and abscisic acid. The PG(34:3) and PG(36:6) play a master role in cold hardiness. The discovered WHGs and cold hardy genes are highly distinctive as confirmed by PCA, network propagation, and/or lipidomics profiles. The three *VRN1* genes are closely associated with their immediate neighborhood, which are highly cohesive.

Materials and methods

Plant materials

Plant materials as detailed in Limin and Fowler (2002) include four wheat (*Triticum aestivum* L.) genotypes (M: Manitou, WM: winter Manitou, N: Norstar and SN: spring Norstar). Briefly, a non-hardy spring wheat Manitou, determined by dominant *Vrn-A1* allele, and a very cold hardy winter-habit Norstar, determined by recessive *vrn-A1* allele, were crossed to produce the reciprocal near-isogenic lines (NILs) (Limin and Fowler, 2002). During the crossing process, the vernalization allele in Norstar (*vrn-A1*) was replaced by the spring-habit allele at *Vrn-A1* locus from Manitou to produce spring Norstar. Whereas, replacing the *Vrn-A1* allele of Manitou with the *vrn-A1* from Norstar made Manitou a vernalization-responsive winter-habit genotype (winter Manitou).

Briefly, for cold treatment under controlled environments, wheat plants were grown in chambers with 16-h-light (~120 μmol m^{-2} s^{-1}) and 8-h-dark at 23°C up to the stage of four leaves (3 weeks) and then transferred to 4°C chamber for 6 weeks. The third fully opened leaves from cold-treated and untreated plants were collected at around 10:00 a.m. and immediately frozen in liquid N_2. Samples were stored at −80°C until lipidomics and RNA-seq analyses. Each genotype under a condition has three independent biological replicates.

RNA sequencing and data quality control and mapping

The RNA-seq dataset in Li et al. (2021) was reanalyzed in this study. Briefly, total RNA was extracted from 0.1 g wheat leaf tissues for each of the 24 cold treated and un-treated samples using the Agilent Plant RNA isolation kit (Agilent Technologies) and sequenced as described in Li et al. (2021). In total, the RNA-seq dataset contains 24 wheat samples with an average of 34 million reads per sample and available at Gene Expression Omnibus (GEO, GSE156300). We trimmed adaptor sequence, discarded low-quality reads (Phred Score ≤20) and eliminated short reads (length ≤20 bps) using a software package FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). In average, 26 million reads remained and were aligned to the high confidence gene models in the IWGSC RefSeq Version 2.1 reference genome (Zhu et al., 2021) by using STAR (v2.7.10a, Dobin et al., 2013). From the BAM files generated by STAR, level of mRNA in each sample was quantified as transcript per million (TPM) by using RSEM (Li and Dewey, 2011).

DEG analysis

Recent studies in RNA-seq data analysis indicate that normalized expression data, such as TPM, FPKM or RPKM is not acceptable for DEG analysis (Zhao et al., 2020; Zhao et al., 2021). The read count data from STAR above were used to perform eight pairwise gene differential expression analyses using DESeq2 (Love et al., 2014). Each of the four genotypes were compared between cold-treated and un-treated control (WMC-WMK, MC-MK, NC-NK, and SNC-SNK, where, W = winter, S = spring, M = Manitou, N=Norstar, C = cold, K = control). Similarly, within each pair of NILs (winter Manitou and Manitou, Norstar and spring Norstar), we compared winter-habit genotype with its spring-habit counterpart in the cold treated samples (WMC-MC, NC-SNC) as well as in controls (WMK-MK, NK-SNK). The outputs from DESeq2 include log2 fold change values and associated statistical significance (*p*-values, and adjusted *p*-values).

Data reduction and partitioning

We applied the criteria of |log2FC| ≥ 2, adjusted $p \leq 0.01$ and the max(TPM of compared samples) ≥ 2 to identify differentially expressed genes (DEGs). Differential Expression Feature Extract (DEFE) method (Pan et al., 2022) was applied to partition the DEGs into groups of various expression profiles, whether they were consistent across all genotypes in response to cold treatment or specific to each or certain pairs of genotypes. Three series of DEFE analyses were performed. Firstly, for the comparison between cold treated samples in the four genotypes *versus* their respective controls, a series of DEFE patterns were identified with a prefix "P" and followed by four digits each representing a genotype in the order of M, WM, SN, and N. Among the four digits, "0" means not differentially expressed, "1" denotes up regulated and "2" down-regulated. For example, P0210 represents a group of genes that were not differentially expressed in Manitou and Norstar, down-regulated in winter Manitou and up-regulated in spring Norstar when treated with cold. Similarly, we were able to obtain groups which were either consistent between the two winter-habit genotypes in response to cold treatment as well as in control, or they were specific to one individual NIL. The pattern ID in these two series start with either "C" or "K" for cold treated or control samples, respectively, and followed by two digits, representing WM/M and N/SN, respectively.

Clustering, correlation, and gene association networks analyses

The WGCNA R package (Langfelder and Horvath, 2008) was used to cluster the normalized expression data of DEGs together with lipid traits based on the distance measure by topology overlap matrix (TOM). Hierarchical clustering was employed based on the similarity matrix to cluster genes as described in Pan et al. (2018). Briefly, the network connection weight was calculated based on TOM and the top 1% weight was used for network construction. The trait-trait, gene-trait, and cluster-trait correlation matrices were computed. Here, a trait refers to an experimental condition and a lipid species. Network visualization was performed by using Cytoscape (Shannon et al., 2003).

Gene orthologue, annotation and GO enrichment analysis

For the known IWGSC RefSeq 2.1 genes, we obtained their orthologues in *Arabidopsis thaliana*, *Brachypodium distachyon*, *Oryza sativa* Japonica, and gene names and descriptions from EnsemblPlants (http://plants.ensembl.org/Triticum_aestivum/Info/Index) through reciprocal best kit BlastP ($e \leq 10^{-5}$). The orthologues, annotations, cluster membership, and mapping of gene IDs with various previous genome assembly are available in the Supplementary File S1.

GOAL software (Tchagang et al., 2010) was used in the gene ontology (GO) enrichment analysis. The GO terms were recently updated from EnsemblPlants release 51 (http://plants.ensembl.org/Triticum_aestivum/Info/Index) and the gene-GO association file for this version of wheat genome are available in the Supplementary File S8. An updated version of GOAL software is available at https://github.com/DT-NRC/GOAL2.0.

Principal component analysis and visualization of data

Principal component analysis was performed by using PCAtools R package in Bioconductor (Blighe and Lun, 2022). The 12,676 DEGs were visualized in heatmap by using ComplexHeatmap R package in Bioconductor (Gu et al., 2016). Otherwise, R versions 4.2.1 were used in 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 in the article/Supplementary Material.

Author contributions

YP performed computational genomics analysis and led the manuscript preparation. QL and JZ designed the experiment and generated the data. YP, YL, and ZL performed RNA-seq data pre-processing, DEG analysis, and gene orthology analysis. All authors contributed intellectually in data analysis, writing and approved final version for submission.

Funding

This research was supported by grants A1-000455 to YP from the Canadian Wheat Improvement Program, National Research Council of Canada, that reviewed the grants and provided funding for operational costs and services, but otherwise did not contribute intellectually to conceiving of the idea, planning of experiments or interpreting the data.

Acknowledgments

The authors would like to thank Janet Condie at the National Research Council Canada (NRC), Saskatoon, SK for assistance with RNA-seq sequencing. This work was resulted from collaborations under Canadian Wheat Alliance including

NRC, Agriculture and Agri-Food Canada and University of Saskatchewan.

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/fgene.2022.1015673/full#supplementary-material

SUPPLEMENTARY FILE S1
Details of the 12676 differential expressed genes (DEGs) including their annotation, mapping of gene IDs to various previous genome versions, membership in various groups identified in this study. This data file contains three tabs: 1) DEGs: A list of 12676 DEGs, their annotation, orthologues, expression, DEFE pattern and their membership to various groups proposed; 2) DEFE Stats: Frequency statistics of all DEFE patterns; 3) ClusterTraitCor: Correlation between each of the 50 clusters with an experimental condition or a lipid species.

SUPPLEMENTARY FILE S2
Gene ontology (Biological Process) enrichment analyses for the genes uniquely up or down regulated by cold in one or all genotypes. This file contains 11 tabs leading by a "Summary" tab providing the statistics overview of the 10 groups of genes uniquely or commonly up or down regulated by cold in the four wheat genotypes. The subsequent tabs labelled by respective DEFE pattern ID followed by "BP" to provide details.

SUPPLEMENTARY FILE S3
Details of the group of 63 winter-habit genes (WHGs), spring-habit gene (SHGs) and others two groups presented in Figure 2A. This file contains 10 tabs: 1) Lists: membership in the four groups, their annotation, DEFE pattern, and the difference of fold changes between the two pairs of NILs (N/SN – WM/M); 2) geneTraitCor_R: Correlation coefficient of each gene with experimental conditions and with each lipid species; 3) - 6) Gene ontology enrichment analyses of the four groups labelled by their respective DEFE patterns: GO_P0101∩C11∩K00, GO_P0202∩C22∩K00, GO_P1010∩C22∩K00, GO_P2020∩C11∩K00; 7) - 10) Gene associate network analysis, nodes and edges of WHGs (P0101∩C11∩K00) and SHGs (P1010∩C22∩K00).

SUPPLEMENTARY FILE S4
Details of cold hardy genes and anti-hardy genes presented in Figure 2B. This file contains 8 tabs: 1) Lists: members in the two groups and their annotation; 2) geneTraitCor_R: Correlation coefficient of each gene with experimental conditions and with each lipid species; 3) - 4) Gene ontology enrichment analyses of the two groups labelled by their respective group names; 5) - 8) Gene associate network analysis of these two groups of genes including notes and edges.

SUPPLEMENTARY FILE S5
Description with supporting figures of the two schools of network nodes represented by phosphatidylglycerol lipids, PG(34:4) for School A, PG(34:3) and PG(36:6) for School B.

SUPPLEMENTARY FILE S6
Details of the two schools of network nodes presented in Supplementary File S5. This file contains 10 tabs: 1) School_A [PG(34_4)]_nodes: PG(34:4) and 371 genes in its immediate neighbourhood, their respective association degree in the entire network and in this subnetwork, and overall association strength; 2) PG(34_4)_edges: connection edges of this subnet; 3) PG(34_4)_GO: Gene ontology enrichment analysis (GOEA) of the 371 genes; 4) School_B_nodes: PG(34:3), PG(36:6) and 105 genes in their immediate neighbourhood, their respective association degree in the entire network and in respective subnetwork; 5) - 10) the nodes, edges and GOEA of the genes in PG(34:3) and PG(36:6) subnets.

SUPPLEMENTARY FILE S7
The network property of the 214 genes in the immediate neighbourhood of three homoeologs of the VRN1 gene (VRN-A1, VRN-B1, VRN-D1). This file contains three tabs: 1) The eight VRN genes; 2) the 217 nodes, 3) the 19570 edges.

SUPPLEMENTARY FILE S8
Gene-GO associations used for gene ontology enrichment analysis in this study. These associations were assemble from EnselblPlants release 51. The gene ID of IWGSC Refseq v1.2 was converted to IWGSC RefSeq v2.1 based on the ID mapping provided by Zhu et al., (2021).

SUPPLEMENTARY FILE S1
The third principal component (PC3) separates winter-habit genotypes from the spring-habit ones in each pair of the NILs when treated with cold. (A) PC3 vs. PC1, (B) PC3 vs. PC2.

SUPPLEMENTARY FILE S2
Frequency distribution of DEGs presented by DEFE patterns of the two pairs of NILs between winter-habit and spring-habit genotypes. (A) C series: cold-treated samples; (B) K series: control samples.

SUPPLEMENTARY FILE S3
Contrast in network association strength: (A) winter-habit genes were generally well associated within the group; (B) less than half of the spring-habit genes were loosely associated another gene in the group.

SUPPLEMENTARY FILE S4
Contrast between cold hardy and anti-hardy genes in network association perspective: (A) cold hardy genes were generally well associated within the group, the four WHGs and PG(34:3) and PG(36:6) are highlighted; (B) only 1/3 of the anti-hardy genes were loosely associated with another gene.

SUPPLEMENTARY TABLE S1
Range of PC scores of the genes up- or down-regulated specifically to winter-habit or spring-habit genotypes.

References

Abdul Kayum, M., Nath, U. K., Park, J. I., Biswas, M. K., Choi, E. K., Song, J. Y., et al. (2018). Genome-wide identification, characterization, and expression profiling of glutathione S-transferase (GST) family in pumpkin reveals likely role in cold-stress tolerance. *Genes (Basel)* 9, 84. doi:10.3390/genes9020084

Al-Attar, R., Wijenayake, S., and Storey, K. B. (2019). Metabolic reorganization in winter: Regulation of pyruvate dehydrogenase (PDH) during long-term freezing and anoxia. *Cryobiology* 86, 10–18. doi:10.1016/j.cryobiol.2019.01.006

Amasino, R. M. (2005). Vernalization and flowering time. *Curr. Opin. Biotechnol.* 16, 154–158. doi:10.1016/j.copbio.2005.02.004

Aslam, M., Greaves, J. G., Jakada, B. H., Fakher, B., Wang, X., and Qin, Y. (2022). AcCIPK5, a pineapple CBL-interacting protein kinase, confers salt, osmotic and cold stress tolerance in transgenic Arabidopsis. *Plant Sci.* 320, 111284. doi:10.1016/j.plantsci.2022.111284

Blighe, K., and Lun, A. (2022). *PCAtools: Everything principal components analysis*. Available at: https://bioconductor.org/packages/devel/bioc/vignettes/PCAtools/inst/doc/PCAtools.html.

Blume, C., Behrens, C., Eubel, H., Braun, H. P., and Peterhansel, C. (2013). A possible role for the chloroplast pyruvate dehydrogenase complex in plant glycolate and glyoxylate metabolism. *Phytochemistry* 95, 168–176. doi:10.1016/j.phytochem.2013.07.009

Breton, G., Danyluk, J., Charron, J. B., and Sarhan, F. (2003). Expression profiling and bioinformatic analyses of a novel stress-regulated multispanning transmembrane protein family from cereals and Arabidopsis. *Plant Physiol.* 132, 64–74. doi:10.1104/pp.102.015255

Browse, J., and Xin, Z. (2001). Temperature sensing and cold acclimation. *Curr. Opin. Plant Biol.* 4, 241–246. doi:10.1016/s1369-5266(00)00167-9

Brule-Babel, A. L., and Fowler, D. B. (1988). Genetic-control of cold hardiness and vernalization requirement in winter-wheat. *Crop Sci.* 28, 879–884. doi:10.2135/cropsci1988.0011183X002800060001x

Charron, J. B., Ouellet, F., Pelletier, M., Danyluk, J., Chauve, C., and Sarhan, F. (2005). Identification, expression, and evolutionary analyses of plant lipocalins. *Plant Physiol.* 139, 2017–2028. doi:10.1104/pp.105.070466

Chen, S., Wang, J., Deng, G., Chen, L., Cheng, X., Xu, H., et al. (2018). Interactive effects of multiple vernalization (*Vrn-1*)- and photoperiod (*Ppd-1*)-related genes on the growth habit of bread wheat and their association with heading and flowering time. *BMC Plant Biol.* 18, 374. doi:10.1186/s12870-018-1587-8

Chouard, P. (1960). Vernalization and its relations to dormancy. *Annu. Rev. Plant Physiol.* 11, 191–238. doi:10.1146/annurev.pp.11.060160.001203

Colton-Gagnon, K., Ali-Benali, M. A., Mayer, B. F., Dionne, R., Bertrand, A., Do Carmo, S., et al. (2014). Comparative analysis of the cold acclimation and freezing tolerance capacities of seven diploid *Brachypodium distachyon* accessions. *Ann. Bot.* 113, 681–693. doi:10.1093/aob/mct283

Crevillén, P., Yang, H., Cui, X., Greeff, C., Trick, M., Qiu, Q., et al. (2014). Epigenetic reprogramming that prevents transgenerational inheritance of the vernalized state. *Nature* 515, 587–590. doi:10.1038/nature13722

Cui, M. H., Yoo, K. S., Hyoung, S., Nguyen, H. T., Kim, Y. Y., Kim, H. J., et al. (2013). An Arabidopsis R2R3-MYB transcription factor, AtMYB20, negatively regulates type 2C serine/threonine protein phosphatases to enhance salt tolerance. *FEBS Lett.* 587, 1773–1778. doi:10.1016/j.febslet.2013.04.028

Danyluk, J., Kane, N. A., Breton, G., Limin, A. E., Fowler, D. B., and Sarhan, F. (2003). TaVRT-1, a putative transcription factor associated with vegetative to reproductive transition in cereals. *Plant Physiol.* 132, 1849–1860. doi:10.1104/pp.103.023523

Deng, W., Casao, M. C., Wang, P., Sato, K., Hayes, P. M., Finnegan, E. J., et al. (2015). Direct links between the vernalization response and other key traits of cereal crops. *Nat. Commun.* 6, 5882. doi:10.1038/ncomms6882

Dhillon, T., Pearce, S. P., Stockinger, E. J., Distelfeld, A., Li, C., Knox, A. K., et al. (2010). Regulation of freezing tolerance and flowering in temperate cereals: The VRN-1 connection. *Plant Physiol.* 153, 1846–1858. doi:10.1104/pp.110.159079

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). Star: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635

Fowler, D. B., N'Diaye, A., Laudencia-Chingcuanco, D., and Pozniak, C. J. (2016). Quantitative trait loci associated with phenological development, low-temperature tolerance, grain quality, and agronomic characters in wheat (*Triticum aestivum* L.). *PLoS One* 11, e0152185. doi:10.1371/journal.pone.0152185

Fürtauer, L., Weiszmann, J., Weckwerth, W., and Nägele, T. (2019). MYB20, MYB42, MYB43, and MYB85 regulate phenylalanine and lignin biosynthesis during secondary cell wall formation. *Plant Physiol.* 20, 5411. PMID: 31671650. doi:10.3390/ijms20215411

Gao, S., Zhang, Y. L., Yang, L., Song, J. B., and Yang, Z. M. (2014). AtMYB20 is negatively involved in plant adaptive response to drought stress. *Plant Soil* 376, 433–443. doi:10.1007/s11104-013-1992-6

Geng, P., Zhang, S., Liu, J., Zhao, C., Wu, J., Cao, Y., et al. (2020). MYB20, MYB42, MYB43, and MYB85 regulate phenylalanine and lignin biosynthesis during secondary cell wall formation. *Plant Physiol.* 182, 1272–1283. doi:10.1104/pp.19.01070

Godse, R., Bawane, H., Tripathi, J., and Kulkarni, R. (2021). Unconventional β-glucosidases: A promising biocatalyst for industrial biotechnology. *Appl. Biochem. Biotechnol.* 193, 2993–3016. doi:10.1007/s12010-021-03568-y

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32, 2847–2849. doi:10.1093/bioinformatics/btw313

Herinckx, G., Hussain, N., Opperdoes, F. R., Storey, K. B., Rider, M. H., and Vertommen, D. (2017). Changes in the phosphoproteome of Brown adipose tissue during hibernation in the ground squirrel, *Ictidomys tridecemlineatus*. *Physiol. Genomics* 49, 462–472. doi:10.1152/physiolgenomics.00038.2017

Horn, P. J., Smith, M. D., Clar, T. R., Froehlich, J. E., and Benning, C. (2020). PEROXIREDOXIN Q stimulates the activity of the chloroplast 16:1Δ3trans FATTY ACID DESATURASE4. *Plant J.* 102, 718–729. doi:10.1111/tpj.14657

Huner, N. P. A., Williams, J. P., Maissan, E. E., Myscich, E. G., Krol, M., Laroche, A., et al. (1989). Low temperature-induced decrease in trans-delta-hexadecenoic acid content is correlated with freezing tolerance in cereals. *Plant Physiol.* 89, 144–150. doi:10.1104/pp.89.1.144

Kennedy, A., and Geuten, K. (2020). The role of FLOWERING LOCUS C relatives in cereals. *Front. Plant Sci.* 11, 617340. doi:10.3389/fpls.2020.617340

Khazaei, M., Maali-Amiri, R., Talei, A. R., and Ramezanpour, S. (2015). Differential transcript accumulation of dhydrin and beta-glucosidase genes to cold-induced oxidative stress in chickpea. *J. Agr. Sci. Tech.* 17, 725–734.

Kidokoro, S., Kim, J. S., Ishikawa, T., Suzuki, T., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2020). DREB1A/CBF3 is repressed by transgene-induced DNA methylation in the Arabidopsis *ice1-1* mutant. *Plant Cell* 32, 1035–1048. doi:10.1105/tpc.19.00532

Kidokoro, S., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2022). Transcriptional regulatory network of plant cold-stress responses. *Trends Plant Sci.* 21, 922–935. doi:10.1016/j.tplants.2022.01.008

Kim, D. H., Doyle, M. R., Sung, S., and Amasino, R. M. (2009). Vernalization: Winter and the timing of flowering in plants. *Annu. Rev. Cell Dev. Biol.* 25, 277–299. doi:10.1146/annurev.cellbio.042308.113411

Kim, H. S., Oh, J. M., Luan, S., Carlson, J. E., and Ahn, S. J. (2013). Cold stress causes rapid but differential changes in properties of plasma membrane H+-ATPase of camelina and rapeseed. *J. Plant Physiol.* 170, 828–837. doi:10.1016/j.jplph.2013.01.007

Knox, A. K., Dhillon, T., Cheng, H., Tondelli, A., Pecchioni, N., and Stockinger, E. J. (2010). *CBF* gene copy number variation at *Frost Resistance-2* is associated with levels of freezing tolerance in temperate-climate cereals. *Theor. Appl. Genet.* 121, 21–35. doi:10.1007/s00122-010-1288-7

Lamkemeyer, P., Laxa, M., Collin, V., Li, W., Finkemeier, I., Schöttler, M. A., et al. (2006). Peroxiredoxin Q of *Arabidopsis thaliana* is attached to the thylakoids and functions in context of photosynthesis. *Plant J.* 45, 968–981. doi:10.1111/j.1365-313X.2006.02665.x

Langfelder, P., and Horvath, S. (2008). Wgcna: an R package for weighted correlation network analysis. *BMC Bioinforma.* 9, 559. doi:10.1186/1471-2105-9-559

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M., and Rhee, S. Y. (2010). Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nat. Biotechnol.* 28, 149–156. doi:10.1038/nbt.1603

Lee, T., and Lee, I. (2017). AraNet: A network biology server for *Arabidopsis thaliana* and other non-model plant species. *Methods Mol. Biol.* 1629, 225–238. doi:10.1007/978-1-4939-7125-1_15

Levesque-Tremblay, G., Havaux, M., and Ouellet, F. (2009). The chloroplastic lipocalin AtCHL prevents lipid peroxidation and protects Arabidopsis against oxidative stress. *Plant J.* 60, 691–702. doi:10.1111/j.1365-313X.2009.03991.x

Li, B., and Dewey, C. N. (2011). Rsem: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinforma.* 12, 323. PMID: 21816040; PMCID: PMC3163565. doi:10.1186/1471-2105-12-323

Li, G., Yu, M., Fang, T., Cao, S., Carver, B. F., and Yan, L. (2013). Vernalization requirement duration in winter wheat is controlled by *TaVRN-A1* at the protein level. *Plant J.* 76, 742–753. doi:10.1111/tpj.12326

Li, Q., Zheng, Q., Shen, W., Cram, D., Fowler, D. B., Wei, Y., et al. (2015). Understanding the biochemical basis of temperature-induced lipid pathway adjustments in plants. *Plant Cell* 27, 86–103. doi:10.1105/tpc.114.134338

Li, Q., Byrns, B., Badawi, M. A., Diallo, A. B., Danyluk, J., Sarhan, F., et al. (2018). Transcriptomic insights into phenological development and cold tolerance of wheat grown in the field. *Plant Physiol.* 176, 2376–2394. doi:10.1104/pp.17.01311

Li, Q., Shen, W., Mavraganis, I., Wang, L., Gao, P., Gao, J., et al. (2021). Elucidating the biochemical basis of trans-16:1 fatty acid change in leaves during cold acclimation in wheat. *Plant-Environment Interact.* 2, 101–111. doi:10.1002/pei3.10044

Limin, A. E., and Fowler, D. B. (2002). Developmental traits affecting low-temperature tolerance response in near-isogenic lines for the vernalization locus *Vrn-A1* in wheat (*Triticum aestivum* L. em Thell). *Ann. Bot.* 89, 579–585. doi:10.1093/aob/mcf102

Limin, A. E., and Fowler, D. B. (2006). Low-temperature tolerance and genetic potential in wheat (*Triticum aestivum* L.): Response to photoperiod vernalization, and plant development. *Planta* 224, 360–366. doi:10.1007/s00425-006-0219-y

Liu, Q., Ding, Y., Shi, Y., Ma, L., Wang, Y., Song, C., et al. (2021). The calcium transporter ANNEXIN1 mediates cold-induced calcium signaling and freezing tolerance in plants. *EMBO J.* 40, e104559. doi:10.15252/embj.2020104559

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8

Ma, X., Gai, W. X., Li, Y., Yu, Y. N., Ali, M., and Gong, Z. H. (2022). The CBL-interacting protein kinase *CaCIPK13* positively regulates defence mechanisms against cold stress in pepper. *J. Exp. Bot.* 73, 1655–1667. doi:10.1093/jxb/erab505

Malnoë, A., Schultink, A., Shahrasbi, S., Rumeau, D., Havaux, M., and Niyogi, K. K. (2018). The plastid lipocalin *lcnp* is required for sustained photoprotective energy dissipation in Arabidopsis. *Plant Cell* 30, 196–208. doi:10.1105/tpc.17.00536

Maruyama, K., Takeda, M., Kidokoro, S., Yamada, K., Sakuma, Y., Urano, K., et al. (2009). Metabolic pathways involved in cold acclimation identified by integrated analysis of metabolites and transcripts regulated by DREB1A and DREB2A. *Plant Physiol.* 150, 1972–1980. doi:10.1104/pp.109.135327

Murakami, Y., Tsuyama, M., Kobayashi, Y., Kodama, H., and Iba, K. (2000). Trienoic fatty acids and plant tolerance of high temperature. *Science* 287, 476–479. doi:10.1126/science.287.5452.476

Murata, N., and Los, D. A. (1997). Membrane fluidity and temperature perception. *Plant Physiol.* 115, 875–879. doi:10.1104/pp.115.3.875

Okawa, K., Nakayama, K., Kakizaki, T., Yamashita, T., and Inaba, T. (2008). Identification and characterization of Cor413im proteins as novel components of the chloroplast inner envelope. *Plant Cell Environ.* 31, 1470–1483. doi:10.1111/j.1365-3040.2008.01854.x

Oliver, S. N., Finnegan, E. J., Dennis, E. S., Peacock, W. J., and Trevaskis, B. (2009). Vernalization-induced flowering in cereals is associated with changes in histone methylation at the VERNALIZATION1 gene. *Proc. Natl. Acad. Sci. U. S. A.* 106, 8386–8391. doi:10.1073/pnas.0903566106

Opassiri, R., Pomthong, B., Akiyama, T., Nakphaichit, M., Onkoksoong, T., Cairns, M. K., et al. (2007). A stress-induced rice (*Oryza sativa* L.) beta-glucosidase represents a new subfamily of glycosyl hydrolase family 5 containing a fascin-like domain. *Biochem. J.* 408, 241–249. doi:10.1042/BJ20070734

Pan, Y., Liu, Z., Rocheleau, H., Fauteux, F., Wang, Y., McCartney, C., et al. (2018). Transcriptome dynamics associated with resistance and susceptibility against fusarium head blight in four wheat genotypes. *BMC Genomics* 19, 642. doi:10.1186/s12864-018-5012-3

Pan, Y., Surendra, A., Liu, Z., Ouellet, T., and Foroud, N. A. (2022). "Differential expression feature extraction (DEFE) – A case study in wheat FHB RNA-seq data analysis," in *Methods in molecular biology - plant pathogen interactions*. Editors N. Foroud and J. Neilson (Springer Nature book series). (accepted).

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, 2498–2504. doi:10.1101/gr.1239303

Shi, H., Ye, T., Zhong, B., Liu, X., Jin, R., and Chan, Z. (2014). AtHAP5A modulates freezing stress resistance in Arabidopsis through binding to CCAAT motif of *AtXTH21*. *New Phytol.* 203, 554–567. doi:10.1111/nph.12812

Shi, Y., Ding, Y., and Yang, S. (2018). Molecular regulation of CBF signaling in cold acclimation. *Trends Plant Sci.* 23, 623–637. doi:10.1016/j.tplants.2018.04.002

Skinner, D. Z., and Garland-Campbell, K. A. (2008). The relationship of LT50 to prolonged freezing survival in winter wheat. *Can. J. Plant Sci.* 88, 885–889. doi:10.4141/CJPS08007

Song, W., Zhou, F., Shan, C., Zhang, Q., Ning, M., Liu, X., et al. (2021a). Identification of glutathione S-transferase genes in Hami melon (*Cucumis melo* var. *saccharinus*) and their expression analysis under cold stress. *Front. Plant Sci.* 12, 672017. doi:10.3389/fpls.2021.672017

Song, Y., Zhang, X., Li, M., Yang, H., Fu, D., Lv, J., et al. (2021b). The direct targets of CBFs: In cold stress response and beyond. *J. Integr. Plant Biol.* 63, 1874–1887. doi:10.1111/jipb.13161

Sun, S., Lin, M., Qi, X., Chen, J., Gu, H., Zhong, Y., et al. (2021). Full-length transcriptome profiling reveals insight into the cold response of two kiwifruit genotypes (*A. arguta*) with contrasting freezing tolerances. *BMC Plant Biol.* 21, 365. doi:10.1186/s12870-021-03152-w

Tchagang, A. B., Gawronski, A., Bérubé, H., Phan, S., Famili, F., and Pan, Y. (2010). Goal: A software tool for assessing biological significance of genes groups. *BMC Bioinforma.* 11, 229. doi:10.1186/1471-2105-11-229

Tchagang, A. B., Fauteux, F., Tulpan, D., and Pan, Y. (2017). Bioinformatics identification of new targets for improving low temperature stress tolerance in spring and winter wheat. *BMC Bioinforma.* 18, 174. doi:10.1186/s12859-017-1596-x

Thomashow, M. F. (2010). Molecular basis of plant cold acclimation: Insights gained from studying the CBF cold response pathway. *Plant Physiol.* 154, 571–577. doi:10.1104/pp.110.161794

Thorlby, G., Fourrier, N., and Warren, G. (2004). The SENSITIVE TO FREEZING2 gene, required for FREEZING tolerance in *Arabidopsis thaliana*, encodes a beta-glucosidase. *Plant Cell* 16, 2192–2203. doi:10.1105/tpc.104.024018

Vágújfalvi, A., Galiba, G., Cattivelli, L., Dubcovsky, J., and VagujfAlvi, A. (2003). The cold-regulated transcriptional activator *Cbf3* is linked to the frost-tolerance locus *Fr-A2* on wheat chromosome 5A. *Mol. Genet. Genomics* 269, 60–67. doi:10.1007/s00438-003-0806-6

Wallis, J. G., and Browse, J. (2002). Mutants of Arabidopsis reveal many roles for membrane lipids. *Prog. Lipid Res.* 41, 254–278. doi:10.1016/s0163-7827(01)00027-3

Wolf, C., Koumanov, K., Tenchov, B., and Quinn, P. J. (2001). Cholesterol favors phase separation of sphingomyelin. *Biophys. Chem.* 89, 163–172. doi:10.1016/s0301-4622(00)00226-x

Zhao, S., Ye, Z., and Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA* 26, 903–909. doi:10.1261/rna.074922.120

Zhao, Y., Li, M. C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., et al. (2021). TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository. *J. Transl. Med.* 19, 269. doi:10.1186/s12967-021-02936-w

Zheng, Y., Yang, Y., Wang, M., Hu, S., Wu, J., and Yu, Z. (2021). Differences in lipid homeostasis and membrane lipid unsaturation confer differential tolerance to low temperatures in two Cycas species. *BMC Plant Biol.* 21, 377. doi:10.1186/s12870-021-03158-4

Zhong, R., Lee, C., Zhou, J., McCarthy, R. L., and Ye, Z. H. (2008). A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis. *Plant Cell* 20, 2763–2782. doi:10.1105/tpc.108.061325

Zhu, J., Pearce, S., Burke, A., See, D. R., Skinner, D. Z., Dubcovsky, J., et al. (2014). Copy number and haplotype variation at the *VRN-A1* and central *FR-A2* loci are associated with frost tolerance in hexaploid wheat. *Theor. Appl. Genet.* 127, 1183–1197. doi:10.1007/s00122-014-2290-2

Zhu, T., Wang, L., Rimbert, H., Rodriguez, J. C., Deal, K. R., De Oliveira, R., et al. (2021). Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome assembly. *Plant J.* 107, 303–314. doi:10.1111/tpj.15289

Glossary

ABA abscisic acid

AS association strength

CBF C-repeat binding factor

COR cold regulated gene

COR413-TM1 COLD REGULATED 314 THYLAKOID MEMBRANE 1

DEFE Differential Expression Feature Extraction

DEG differentially expressed gene

DGDG digalactosyldiacylglycerol

DREB1 dehydration-responsive element-binding protein 1

FC fold change

FLC FLOWERING LOCUS C

GDSL amino acid sequence motif consisting of Gly, Asp, Ser, and Leu around the active site Ser

GO Gene Ontology

IWGSC International Wheat Genome Sequencing Consortium

M Manitou

MGDG monogalactosyldiacylglycerol

MYB myeloblastosis domain containing transcription factor

N Norstar

NAM no apical meristem

NIL near-isogenic line

PC principal componentphosphatidylcholine

PC principal componentphosphatidylcholine

PCA principal component analysis

PE phosphatidylethanolamine

PG Phosphatidylglycerol

PI phosphatidylinositols

PPT palmitoyl-protein thioesterase

ROS reactive oxygen species

SA salicylic acid

SHG spring-habit gene

SN spring Norstar

TF transcription factor

TOM topology overlap matrix

VRN vernalization gene

WGCNA weighted gene correlation network analysis

WHG winter-habit gene

WM winter Manitou

TYPE Perspective
PUBLISHED 10 November 2022
DOI 10.3389/fgene.2022.1042550

Check for updates

OPEN ACCESS

EDITED BY
Quan Zou,
University of Electronic Science and Technology of China, China

REVIEWED BY
Wang Xi,
Harvard University, United States
Houxiang Zhu,
The University of Chicago, United States
Sammed Mandape,
University of North Texas Health Science Center, United States
Yan Nie,
Texas A & M University, United States

*CORRESPONDENCE
Anshu Bhardwaj,
anshu@imtech.res.in

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 12 September 2022
ACCEPTED 21 October 2022
PUBLISHED 10 November 2022

CITATION
Singh V, Pandey S and Bhardwaj A (2022), From the reference human genome to human pangenome: Premise, promise and challenge. *Front. Genet.* 13:1042550.
doi: 10.3389/fgene.2022.1042550

COPYRIGHT
© 2022 Singh, Pandey and Bhardwaj. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

From the reference human genome to human pangenome: Premise, promise and challenge

Vipin Singh[1], Shweta Pandey[2,3] and Anshu Bhardwaj[2,3]*

[1]University Institute of Biotechnology, Chandigarh University, Mohali, India, [2]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India, [3]Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh, India

The Reference Human Genome remains the single most important resource for mapping genetic variations and assessing their impact. However, it is monophasic, incomplete and not representative of the variation that exists in the population. Given the extent of ethno-geographic diversity and the consequent diversity in clinical manifestations of these variations, population specific references were developed overtime. The dramatically plummeting cost of sequencing whole genomes and the advent of third generation long range sequencers allowing accurate, error free, telomere-to-telomere assemblies of human genomes present us with a unique and unprecedented opportunity to develop a more composite standard reference consisting of a collection of multiple genomes that capture the maximal variation existing in the population, with the deepest annotation possible, enabling a realistic, reliable and actionable estimation of clinical significance of specific variations. The Human Pangenome Project thus is a logical next step promising a more accurate and global representation of genomic variations. The pangenome effort must be reciprocally complemented with precise variant discovery tools and exhaustive annotation to ensure unambiguous clinical assessment of the variant in ethno-geographical context. Here we discuss a broad roadmap, the challenges and way forward in developing a universal pangenome reference including data visualization techniques and integration of prior knowledge base in the new graph based architecture and tools to submit, compare, query, annotate and retrieve relevant information from the pangenomes. The biggest challenge, however, will be the ethical, legal and social implications and the training of human resource to the new reference paradigm.

KEYWORDS

human pangenome, reference human genome, graph based methods, variant discovery, precision medicine

1 Introduction

On 12 February 2001, The Human Genome Project Consortium announced the release of the first draft of the Reference Human Genome and the sequence was released into the public domain. Parallelly, Celera Genomics, a private initiative, also announced the release of the Alternate Human Genome assembly (Lander et al., 2001) (Venter et al., 2001). Considered as the "giant leap" in Biotechnology, this was an event no less celebrated than the landing of man on the moon, and divided modern Biotechnology into pre and post human genome era. This also marked the beginning of the Omics era - the study of something in totality and not in parts (Kiechle, Zhang, and Holland-Staley 2004). The Reference Human Genome remains the single most important resource for mapping human genetic variations and assessing their clinical impact. However, it was immediately realized that if we were to tap the full potential of the sequence information in terms of understanding genotype-phenotype correlation, mapping disease causing variations, in pharmacogenomics and in personalized medicine, a large number of individuals need to be sequenced in quick time and at an astronomically lower cost. The prohibitively high cost and time of sequencing genomes in 2001 lead researchers to explore alternate scale up technologies to Sanger Sequencing, so as to bring down the cost of sequencing to an affordable one thousand dollars. Technological advances in sequencing techniques, as also in information technology including high performance computing and development of novel algorithms lead to Second Generation short read sequencers. Given the short read lengths (100–400 bases depending on the sequencing platform) generated by the Second generation sequencers, accurate *de novo* assembly of the fragmented parts in large, complex and repeat rich genomes such as the human genome was improbable and a reference based assembly approach whereby the short reads were aligned and mapped on to the reference human genome was followed. Third generation long read sequencing techniques are now gaining attention as these, in combination with short read sequencers, allow almost error less *de novo* assembly of complex repeat rich genomes (Hu et al., 2021). As compared to three billion dollars and 10 years in 2001, a good quality, high coverage and accurate haplotype phased telomere-to-telomere assembly of the human genome can be obtained at about a 1000 dollars and in half a day today. Thus, technological advancements in data generation as well as analysis provide us with an opportune moment to gain further insights in human genetics and disease association.

In this article, we critically examine the limitations of the reference human genome and the need to redefine the reference *per se* - the human pangenome–a composite of multiple, haplotype resolved telomere-to-telomere assemblies. We assess the progress made in sequencing technologies since the release of the first draft of the reference human genome, which now make it possible to conceive the pan genome reference. We conclude with a discussion on the promises and challenges as we take definitive steps towards redefining the reference for human genetic studies.

1.1 The reference human genome and genomic variations

The working draft of the reference human genome was released in 2001 and the finished euchromatic genome was released in 2004 (International Human Genome Sequencing Consortium 2004) and has been revised several times since then. The current assembly GRCh38. p13/hg38 was released in December 2013. The reference human genome represents a linear coordinate system or grid facilitating the mapping and assembly of reads obtained from other sequencing experiments and serves as a standard for identifying the variations therein. It is the most extensively used resource for human medical genetics and genomics applications. Comparison of a human genome with the reference human genome allows identification of genomic variations which may associate with the observed phenotypes. Genomic variations have been studied extensively to understand their role in Mendelian and non-Mendelian disease association, diagnostics, prognostics, pharmacogenetics and pharmacogenomics. These genomic variations include the most commonly occurring–single nucleotide substitutions or Single Nucleotide Variations - SNVs, small (<50 base pairs) insertions/deletions, known as INDELS, large structural variations including large INDELS, segmental duplications - duplications of 1 kb or more, differences in copy numbers in tandem repeats, the presence/absence of transposon or mobile element insertions as well as large scale genomic rearrangements like translocations, inversions etc (Eichler 2019). A genomic variant occurring at a frequency of more than 1% in the population is referred to as polymorphic. Single Nucleotide Variation or Single Nucleotide Polymorphism is the most common type of variation found in the human genome. A typical genome differs from the reference human genome at 4.1 million to 5.0 million sites, suggesting that apart from the raw sequence data, one also needs to cater for 4.5 million variant sites if the comparison was done to reference human genome (Auton et al., 2015).

However, the reference human genome, which is used as the standard to elucidate the variations, is neither complete nor does it represent an exhaustive catalog of variations that may exist in the population. It represents a linear composite of merged haplotypes coming from predominantly European ancestry, with a single individual, of more than 20, contributing more than 70% of the reads used for the assembly (Ballouz, Dobin, and Gillis 2019). The reference human genome thus underrepresents and underestimates the full extent of variation that may exist in the population. In addition, due to limitations of read length offered by Sanger Sequencing technique, it is also not complete with gaps in centromeric, telomeric and other repeat rich regions. More than 50% of the gaps in the genome relate to complex Segmental Duplications. It is estimated that the use of short reads and reference based assembly may have resulted in non-reporting of more than 70% of the structural variations (Vollger et al., 2022). This results in a reference bias as well as an ascertainment bias confounding variant discovery, gene-disease association studies and inaccuracies in genetic analysis. The reference human genome is not ideally suited to serve as the "reference" (Chen et al., 2021).

2 Additional efforts to catalog and annotate human genomic variations

A need for exhaustive functional annotation of the genome was reflected in the ENCODE project (Dunham et al., 2012). A deeper cataloging of the variation that exists in the population was the motivation behind HapMap (Altshuler, Donnelly, and The International HapMap Consortium 2005), the 1000 genomes project and the 100000 genomes project besides others. The 1000 genomes Project reconstructed the genomes of 2,504 individuals from 26 populations using a combination of omics technologies including whole-genome sequencing at low coverage (average depth 7.4X), sequencing of the exome at high coverage (average depth 65.7X), and dense microarray genotyping and reported 84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants, all phased onto high-quality haplotypes. The structural variations catalogue comprises of 42,279 biallelic deletions, 6,025 biallelic duplications, 2,929 mCNVs (multi allelic copy-number variants), 786 inversions, 168 nuclear mitochondrial insertions (NUMTs), and 16,631 mobile element insertions (Auton et al., 2015).

With a decade of advancement in sequencing technologies that led to sequencing of a large number of genomes, the objective of the sequencing approach towards genetic screening or predisposition was further extended to target precision medicine. This extended vision demanded discovery and detailed annotation of disease associated variants including the rare variants, ushering in the era of population specific reference genomes. As more and more geographical regions sequenced indigenous populations, it became evident that the under-representation of the non-European samples in human genetic studies was limiting in capturing diversity of genomic datasets which significantly impact the clinical relevance of pathogenic variants identified in European samples to other datasets (Popejoy and Fullerton 2016). It was therefore realized that population specific reference genomes and Genome Wide Association Studies (GWAS) across diverse populations are required to capture the human genetic diversity which was otherwise missing and are critical to understanding disease biology. These efforts are also critical to annotate variants of unknown significance which were reported in large cohort studies but could not be discovered in underrepresented populations as well as in identifying false positive associations. Moving away from the persistent bias, several initiatives like the GenomeAsia100K project (Wall et al., 2019), H3Africa (Mulder et al., 2018), All of United States, IndiGenomes (Jain et al., 2021), etc were initiated to capture genetic variation, explore population structure, identify disease associations and map founder effects in diverse populations across the world. These projects also contributed to discovery of rare disease associated variants. A large number of such initiatives spawned overtime, making it imperative to aggregate these datasets for better understanding of population frequencies of variants, discover novel rare variants, identify novel disease associated genes and variants and prioritize the variants across different population groups. In this context, the Genome Aggregation Database (gnomAD) is the largest collection of harmonized population variation dataset (195,000 individuals as of now) (Gudmundsson et al., 2022). Based on the current version (v3) of gnomAD it is observed that on an average a single human exome carries 27 ± 13 novel unique coding variants. More importantly, the average of such unique novel variants vary across different population groups currently present in gnomAD with South Asians reported to have 38 ± 14 novel unique variants. It is proposed that this number is expected to be higher in population groups that are currently not well represented in gnomAD. Other large databases include NHLBI's TOPMed-BRAVO (Taliun et al., 2021) and DiscovEHR datasets (Wall et al., 2019). Furthermore, it has been shown that size of the datasets ancestral diversity increases the chances of discovery of rare variant. In this context, it is strongly recommended that the pathogenic status of already annotated pathogenic variants in databases like ClinVar (Landrum et al., 2018) needs to be revisited. A timeline of major milestones in the last 20 years of human genomics is illustrated in Figure 1 and an overview of approaches for biomedical applications of human genome studies is depicted in Figure 2.

3 Advancement in sequencing techniques - The three generation of sequencers

One of the primary drivers in the massive increase in sequence data and generation of diploid, phased, high coverage accurate assemblies and complete genomes, the primary prerequisites for generating the pangenome reference, is the evolution of sequencing technologies. Sanger Sequencing Technique was the only automated sequencing technique available at the time when the Human Genome Project was conceived in late 1980s. Considering its limited parallelization

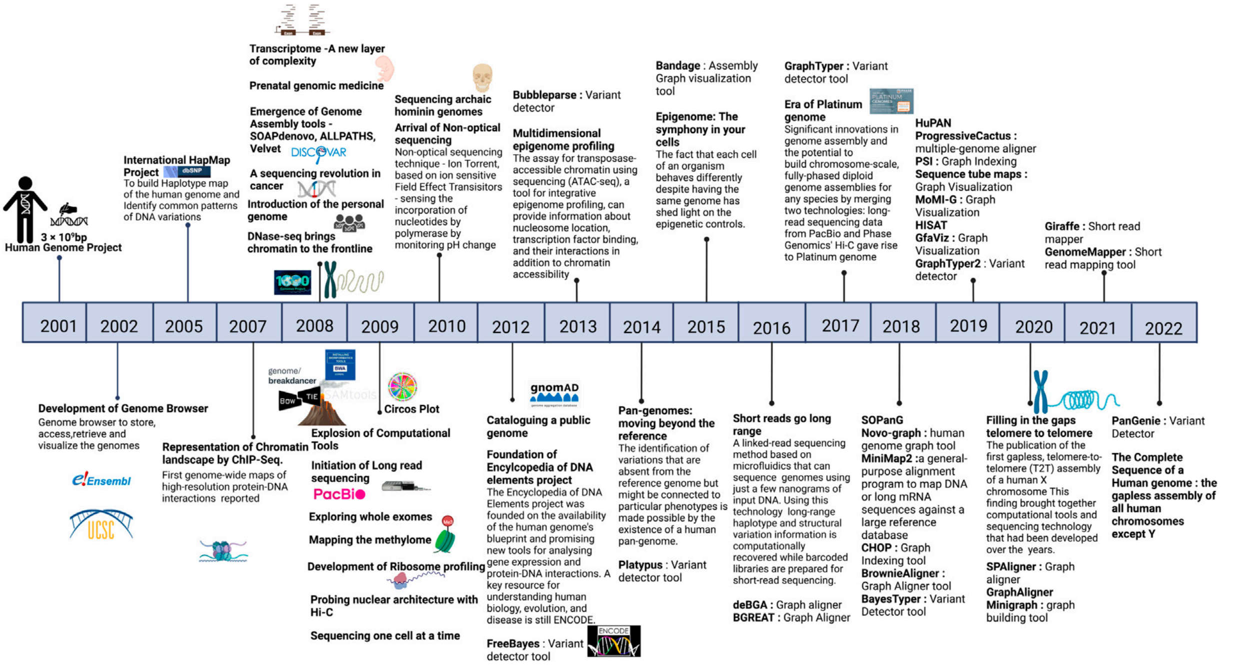

Transcriptome -A new layer of complexity
Prenatal genomic medicine
Emergence of Genome Assembly tools - SOAPdenovo, ALLPATHS, Velvet
DISCOVAR
A sequencing revolution in cancer
Introduction of the personal genome
DNase-seq brings chromatin to the frontline
International HapMap Project
dbSNP
To build Haplotype map of the human genome and Identify common patterns of DNA variations
3 × 10⁹bp
Human Genome Project
Sequencing archaic hominin genomes
Arrival of Non-optical sequencing
Non-optical sequencing technique - Ion Torrent, based on ion sensitive Field Effect Transisitors - sensing the incorporation of nucleotides by polymerase by monitoring pH change
Bubbleparse : Variant detector
Multidimensional epigenome profiling
The assay for transposase-accessible chromatin using sequencing (ATAC-seq), a tool for integrative epigenome profiling, can provide information about nucleosome location, transcription factor binding, and their interactions in addition to chromatin accessibility
Bandage : Assembly Graph visualization tool
Epigenome: The symphony in your cells
The fact that each cell of an organism behaves differently despite having the same genome has shed light on the epigenetic controls.
GraphTyper : Variant detector tool
Era of Platinum genome
PLATINUM GENOMES
Significant innovations in genome assembly and the potential to build chromosome-scale, fully-phased diploid genome assemblies for any species by merging two technologies: long-read sequencing data from PacBio and Phase Genomics' Hi-C gave rise to Platinum genome
HuPAN
ProgressiveCactus : multiple-genome aligner
PSI : Graph Indexing
Sequence tube maps : Graph Visualization
MoMI-G : Graph Visualization
HISAT
GfaViz : Graph Visualization
GraphTyper2 : Variant detector
Giraffe : Short read mapper
GenomeMapper : Short read mapping tool
2001
2002
2005
2007
2008
2009
2010
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Development of Genome Browser
Genome browser to store, access,retrieve and visualize the genomes
Ensembl
UCSC
Representation of Chromatin landscape by ChIP-Seq.
First genome-wide maps of high-resolution protein-DNA interactions reported
genome/ breakdancer
BWA
BOW TIE
SAMtools
Circos Plot
Explosion of Computational Tools
Initiation of Long read sequencing
PacBio
Exploring whole exomes
Mapping the methylome
Me3
Development of Ribosome profiling
Probing nuclear architecture with Hi-C
Sequencing one cell at a time
gnomAD
Cataloguing a public genome
Foundation of Encylcopedia of DNA elements project
The Encyclopedia of DNA Elements project was founded on the availability of the human genome's blueprint and promising new tools for analysing gene expression and protein-DNA interactions. A key resource for understanding human biology, evolution, and disease is still ENCODE.
FreeBayes : Variant detector tool
ENCODE
Pan-genomes: moving beyond the reference
The identification of variations that are absent from the reference genome but might be connected to particular phenotypes is made possible by the existence of a human pan-genome.
Platypus : Variant detector tool
Short reads go long range
A linked-read sequencing method based on microfluidics that can sequence genomes using just a few nanograms of input DNA. Using this technology long-range haplotype and structural variation information is computationally recovered while barcoded libraries are prepared for short-read sequencing.
deBGA : Graph aligner
BGREAT : Graph Aligner
SOPanG
Novo-graph : human genome graph tool
MiniMap2 :a general-purpose alignment program to map DNA or long mRNA sequences against a large reference database
CHOP : Graph Indexing tool
BrownieAligner : Graph Aligner tool
BayesTyper : Variant Detector tool
Filling in the gaps telomere to telomere
The publication of the first gapless, telomere-to-telomere (T2T) assembly of a human X chromosome This finding brought together computational tools and sequencing technology that had been developed over the years.
SPAligner : Graph aligner
GraphAligner
Minigraph : graph building tool
PanGenie : Variant Detector
The Complete Sequence of a Human genome : the gapless assembly of all human chromosomes except Y

FIGURE 1
Timeline of major milestones in the last twenty years of human genomics. Created with BioRender.com.

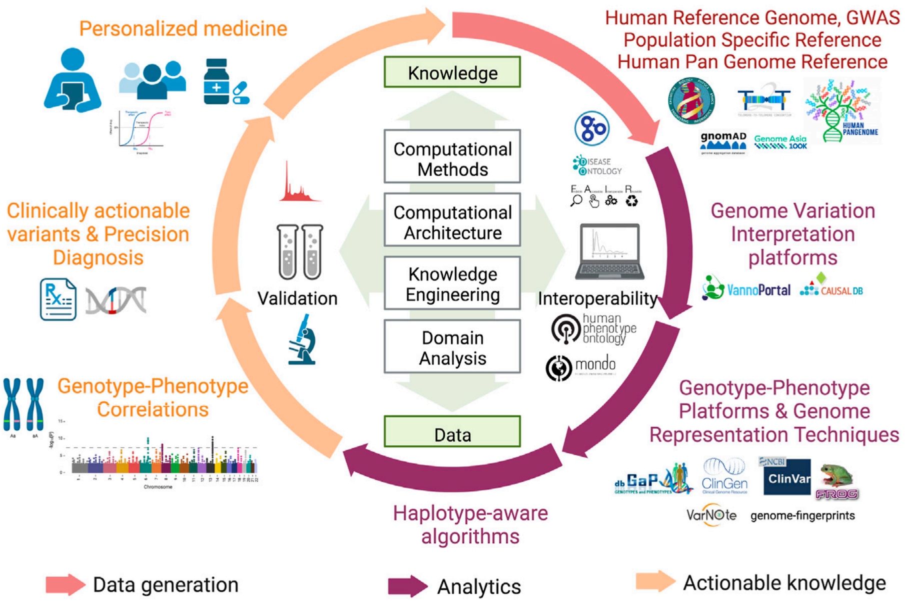

Personalized medicine
Human Reference Genome, GWAS
Population Specific Reference
Human Pan Genome Reference
Knowledge
Computational Methods
Computational Architecture
Knowledge Engineering
Domain Analysis
Data
Clinically actionable variants & Precision Diagnosis
Validation
Interoperability
DISEASE ONTOLOGY
human phenotype ontology
mondo
gnomAD
Genome Asia 100K
HUMAN PANGENOME
Genome Variation Interpretation platforms
VannoPortal
CAUSALDB
Genotype-Phenotype Correlations
Genotype-Phenotype Platforms & Genome Representation Techniques
dbGaP
ClinGen
ClinVar
FROG
VarNote
genome-fingerprints
Haplotype-aware algorithms
Data generation
Analytics
Actionable knowledge

FIGURE 2
Overview of approaches for biomedical applications of human genome studies. Adapted from "The Clinical Applications of Translational Bioinformatics", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates.

and intermediate read length of ~900 bases, and the size and complexity of the human genome (more than 50% repeat content), the Hierarchical shotgun sequencing strategy was devised by the Human Genome Project to ensure a reasonably accurate assembly (Lander et al., 2001). The success of Sanger technique in sequencing the human genome also proved to be its

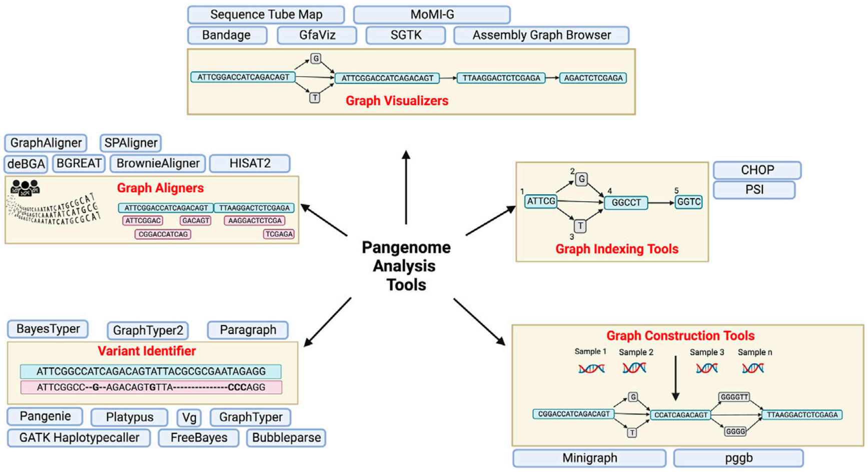


FIGURE 3
Pangenome applications and pipelines with brief description. Created with BioRender.com.

Waterloo as it exposed the limitations–three billion dollars and 10 years to generate the reference human genome. The quest for sequencing techniques that could sequence a human genome in less than a thousand dollars led to several novel approaches broadly classified as Second and Third Generation Sequencing techniques. The second generation sequencing techniques had shorter read lengths (~100 bases) than Sanger Sequencing but this was compensated by their massively parallel sequencing capabilities resulting in high throughput and high coverage quality data. However, the short reads posed an assembly challenge necessitating the use of a reference genome for alignment and mapping of short reads to assemble them into full genomes (Alkan, Sajjadian, and Eichler 2011). The third generation sequencing techniques largely consist of long read sequencers with an average read length >10 kb and facilitate *de novo* assembly of genomes. More significantly, they also enable mapping of some of the most intractable extensively repeat rich regions of the genome, not sequenced before, thus allowing for filling of the gaps and telomere to telomere assembly. Another advantage of the third generation sequencing techniques is that, as opposed to both first and second generation sequencing, where a complex bisulphite treatment step was involved, they allow a direct readout of the epigenetic state of nucleotides (Amarasinghe et al., 2020).

A shift from short read sequencers to long read sequencers which facilitates a more accurate, and complete and unbiased *de novo* assembly of genomes, including the intractable repeat rich regions as well as regions with high GC content is imminent (Pollard et al., 2018). With the recent release of the first telomere to telomere assembly of the human genome, T2T-CHM 13v2.0, adding nearly 200 million base pairs of novel DNA sequences, including 99 genes likely to encode for proteins and nearly 2,000 candidate genes that need further investigation (Nurk et al., 2022), an update of existing knowledge bases and resources and re-evaluation of prior comparisons has become imperative. This is a humongous task given the size and heterogeneity of genomic data. As more such complete genomes become available, a near complete catalogue of genomic variations and their functional impact may be unraveled, necessitating a multiple reference comparison.

4 The human pangenome

The plummeting costs of whole genome sequencing and the potential of long read sequencers to deliver complete, accurate, phased diploid assemblies with epigenomic status provides us with the most opportune moment to conceive a more informative, sophisticated complete reference human genome–the human pangenome - a collective whole genome sequence of multiple individuals capturing the maximum possible diversity that exists in the population. Towards this, the Human Pangenome Reference Consortium (HPRC) aims to create a graph based telomere to telomere representation of the global genomic diversity, to replace the current monophasic, incomplete reference human genome (Wang et al., 2022). To ensure that the pangenome is a true global representation of variations that exist in the human population, one of the goals of the HPRC is to identify individuals from diverse ethnic and biogeographical backgrounds and generate at least 350 reference quality telomere to telomere haplotype phased human diploid genomes - i.e. 700 haplotypes using long range sequencing techniques, trio binning and the use of haplotype aware algorithms. Initially, for high quality long read sequencing, the

HPRC has selected the individual cell lines in the 1000 genomes project which already offers a deep catalog of human variation from 26 populations (Auton et al., 2015). In disease context, the comparison may have to be done to several normal genomes so as to zoom into the variations responsible for the disease phenotype and subtract the silent ones. Pilot studies from China and Africa underline the importance of Pangenome studies in elucidating novel sequence and novel variations in the human genome.

The Chinese Pan Genome project used 486 deep-sequenced Han Chinese genomes and reports 276 Mbp of novel DNA sequences not reported in the reference human genome. The novel sequences belong to one of the two subcategories - individual-specific and shared common sequences. The common sequences, when used along with the reference human genome, improved the accuracy of mapping and variant calling (Li et al., 2021).

The African Pangenome Project used a deeply sequenced dataset of 910 individuals of African descent, to identify unique set DNA sequences present in the African population but not represented in the reference human genome. Unmapped reads to the human genome were assembled into contigs and re-examined (all-to-all comparisons), to derive unique sequences in the African population, not represented in the reference human genome. This dataset consists of 296,485,284 bp in 125,715 distinct contigs indicating that the African PanGenome is 10% larger than the reference human genome. The functional consequences of this extra sequence is under investigation. 387 of these novel contigs are found in 315 distinct protein coding genes (Sherman et al., 2019).

Recently, The Human Pangenome Reference Consortium (HPRC) published a first draft human pangenome reference. of 47 phased, diploid assemblies of genetically diverse individuals in bioRxiv. Covering more than 99% of the genome length, with an accuracy of more than 99%, the pangenome reports novel alleles at structurally complex loci, adds 119 million base pairs of euchromatic polymorphic sequence and 1,529 gene duplications relative to the existing reference, GRCh38. An additional 90 million base account for the structural variation (Liao et al., 2022).

4.1 Challenges in handling and evaluating large genomic datasets

Genome sequence data is poised to overtake the cumulative data on social media by 2025. The total number of individuals whose genome would be sequenced by then would be anywhere between 100 million to two billion. The data-storage demands for this alone could run to as much as 2–40 exabytes because the number of data that must be stored for a single genome is 30 times larger than the size of the genome itself. While the per-base cost of sequencing is dropping by about half every 5 months, the price of data storage falls by half every 14 months. Thus, it is evident that our capacity to generate data is going to far exceed our capacity to store and analyze this data indicating an imminent data management problem. Dealing with this data deluge in terms of storage, retrieval, analysis and exchange therefore indispensably requires novel, interoperable and scalable platforms (Stephens et al., 2015).

4.1.1 Genome variation discovery, annotation and representation

Variant discovery has largely relied on pairwise alignment with the reference human genome. However, as has already been mentioned, the reference human genome fails to capture the full array of variations in human population due to sampling errors and sequencing technology limitations of the times. Over the last 2 decade, there has been a tremendous increase in whole genome sequencing projects *via* Next Generation Sequencing technologies which has led to standardization of methods for variant discovery and generation of an equally vast array of genomic variants (DePristo et al., 2011). The methods of variant discovery are now evolving from reference based read alignment to graph based methods for capturing the complete diversity in human pangenome (Paten et al., 2017). Whole genome multiple assembly alignments with graph based dense representation of variations in the pangenome will facilitate a comprehensive and exhaustive detection of variations. Annotation of genes and other genomic features like regulatory elements including promoter, CpG elements, enhancers, boundary elements and repeats etc. will have to be overlaid on the pangenome. It is proposed that the pangenome will have both NCBI RefSeq and EMBL-EBI's Ensembl/Gencode gene set. In addition, other transcriptomics data will be mapped to individual haplotypes to improve the current annotation and identify novel genes. To understand how genomic variations influence genome function and the phenotype (genotype-phenotype correlation) experimental data from RNASeq, MethylC and ATACSeq experiments from major projects like Roadmap Epigenomics, ENCODE, 4DNucleome, IGVF etc (Wang et al., 2022) will also be integrated.

A whole new suite of user-friendly tools and analysis pipelines compatible with graph based architecture of the pangenome and maintaining organic continuity with the reference human genome will have to be developed for submission, alignment, visualization, analysis, format conversion, annotation and variant detection and sharing and retrieval of data. Some of these tools and pipelines already exist and will be improved overtime. These include graph building tools like minigraph (Li, Feng, and Chu 2020), graph aligners like deBGA (Liu et al., 2016), BGREAT (Limasset et al., 2016), HISTA2 (Kim et al., 2019) etc, tools for graph visualization including Bandage (Wick et al., 2015), AGB (Mikheenko, 2019) etc. and variant detection tools like PanGenie (Ebler et al., 2022), BayesTyper (Sibbesen et al., 2018), Paragraph (Chen et al., 2019), etc (listed in Table 1 and illustrated in Figure 3).

TABLE 1 List of pangenome applications and pipelines with brief description.

Tools	Description & variant identify	Year
FreeBayes Garrison and Marth (2012)	Identifies small variations - SNPs, MNPs, and INDELs (<50 bases). Uses the BAM files and reference genome as input. Detects haplotypes using Bayesian statistics	2012
Bubbleparse Leggett et al. (2013)	Uses Next Generation Sequencing (NGS) data to detect SNPs, without using the reference sequence	2013
Platypus Rimmer et al. (2014)	Detects small variations. Candidate variants are computed from read alignments, local *de novo* assembly, followed by local realignment and probabilistic haplotype estimation	2014
Bandage Wick et al. (2015)	Graph visualizer, the user can customize the graph by moving nodes, adding labels, altering colours, and extracting sequences while zooming in on particular regions of the graph. Does not support scaffold graphs	2015
deBGA Liu et al. (2016)	A graph-based aligner using the seed and extension approach. It organizes and indexes one or more reference genomes using de Bruijn graph (RdBG), and then aligns high-throughput sequencing (HTS) reads to those genomes	2016
BGREAT Limasset et al. (2016)	Assembly tool, uses a heuristic technique to map reads onto the branching path of de Bruijn graph (DBG)	2016
Graphtyper Eggertsson, (2017)	Used for identifying and genotype variations. Takes the reference genome as well as a list of known sequence variants in variant-call format (VCF) format as input, to construct a variant-aware pangenome graph. The seed-and-extend approach is then implemented to the read alignment	2017
BayesTyper Sibbesen, Maretty, and Krogh (2018)	A k-mer approach method that uses reference genome, sequence reads as well as the variant database of the candidate as input and constructs the variation graph. It genotypes all categories of variations (SNPs, indels and complex structural variants)	2018
BrownieAligner Heydari et al. (2018)	Alignment tool, uses seed and extend strategy to align short reads from Illumina to De Bruijn graph. Using high order Markov-model, it also resolves repeats in the graph	2018
GraphTyper2 Eggertsson et al. (2019)	Large Scale (tens of thousands of whole-genomes) identification of structural variants and small variants using pangenome graphs	2019
Paragraph Chen et al. (2019)	It is the graph-based genotyper that uses sequence graph for modeling structural variants (SVs) using short read sequence data	2019
HISAT2 Kim et al. (2019)	Quick and accurate algorithm used to align NGS data - DNA/RNA, to multiple human genomes and reference genome. It uses collection of small Graph FM (GFM) indexes that represent genome and with several alignment approaches, provides rapid alignment of reads	2019
GfaViz Gonnella, Niehus, and Kurtz (2019)	Visualization of sequence graph in graphical fragment assembly (GFA) format. Both command line and graphical user interface	2019
SGTK Kunyavskaya and Prjibelski (2019)	Enables the building and visualising the scaffold graphs using sequencing data	2019
Assembly Graph Browser Mikheenko (2019)	Visualization tool for large and complex assembly graph. Also helps in analysis of repeats and construction of assembly graph	2019
Sequence tube Map Beyer et al. (2019)	Visualization tool of genome graphs and displays variant information in tube format	2019
MoMI-G Yokoyama et al. (2019)	Web based visualization tool for genome graphs, identifies structural variants and hence useful for long reads analysis	2019
vgtoolkit Hickey et al. (2020)	It is used for SV genotyping by building variation graphs using either variant catalogs in the VCF format or assembly alignments	2020
GraphAligner Rautiainen and Marschall (2020)	Alignment tool for long reads. It supports both the GFA and variation graph (vg) formats and can operate with a variety of graphs, including those with overlapping and non-overlapping node sequences	2020
SPAligner Dvorkina et al. (2020)	It is multipurpose tool. It aligns nucleotide and protein sequence to assembly graph and effectively align reads from third generation sequencing	2020
CHOP Mokveld et al. (2020)	Indexing tool for population-graph that uses haplotype-level data to limit path indexing without requiring any pruning or heuristic filtering stages. This constrain eliminates the requirement to assess all k-paths	2020
Minigraph Li, Feng, and Chu (2020)	A sequence-to-graph mapper, that constructs the pangenome graphs using haplotype data and a minimap2 -like algorithm which is based on the seed-chain-align procedure	2020
GenomeMapper https://1001genomes.org/software/genomemapper.html	Short read mapper which align short reads either to reference genome or multiple genomes	2021
Pangenie Ebler et al. (2022)	It infers the genotype of the sample by taking a pangenome graph and short-read sequencing data as input and integrates the information with k-mer count using hidden Markov model (HMM). It enables the variants analysis of SNPs, INDELs, and SVs	2022
pggb https://github.com/pangenome/pggb	Toolkit to build the pangenome graph *via* the integration of various packages. Wfmash is used for pairwise sequence alignment, sequish is used for graph induction, and smoothxg and gfaffix are used	2022

(Continued on following page)

TABLE 1 (*Continued*) List of pangenome applications and pipelines with brief description.

Tools	Description & variant identify	Year
	for normalization of graph. Visualization can be done using Optimized Dynamic Genome/Graph Implementation (ODGI)	
ODGI Guarracino et al. (2022)	Toolkit that helps in the understanding of pangenome graphs. Provides tools for identifying complex regions within pan-genomic loci, and analyzing, manipulating, and visualizing the pangenome graph at the gigabase scale	2022

Assessing the clinical impact of genomic variations in the context of disease association is the ultimate goal of human genetic studies. Several annotation features like allele frequency, tissue or cell type specificity, phenotype association, heterozygosity, functional impact, etc are important to understand the context-dependent significance of genomic variations. Towards this, platforms like VannoPortal (Huang et al., 2022), CausalDB (Wang et al., 2020), DisGenNet (Piñero et al., 2017), PhenGenVar (Shin et al., 2022), etc have been developed. Data from these and other platforms can be utilized by disease-gene network resources like HumanNet v3 for exploring mechanistic insights (Kim et al., 2022). The true potential of these platforms lies in the fact that these variants can be compared across several individuals in a variety of different phenotypes. However, given the complexity of size and data representation methods, it becomes a daunting task to perform these analyses at scale and demand novel methods of data representation towards scalable data analysis.

A few groups have made attempts to develop novel data representation techniques in the form of fingerprints so as to simplify the task of comparison both from a computational as well as biological perspective. These tools include Fingerprinting Ontology of Genomic variation (FROG) (Abinaya, Narang, and Bhardwaj 2015), Ultrafast method (Glusman et al., 2017), Bitome (Lamoureux et al., 2020), VarNote (Huang et al., 2020), etc. To the best of our knowledge, FROG is the first method published in this direction and utilizes an ontology-based approach for representing genomic variation. FROG not only represents variation but also provides a comprehensive assessment of the impact of the variation at the levels of chromosome, DNA, RNA, protein or their interactions. Moreover, FROG ontologies are not dependent on genome data size, organism or on the diversity of ways in which impact of SNPs are reported. It also represents data in binary format to generate genome variation fingerprints for efficient computation, data compression and reducing dimensionality for comparison of the same across several individuals or populations. The Ultrafast method considers the position of the reference and the alternate alleles in an individual and applies a method of locality sensitive hashing for representation of genomic variants with the primary objective of landscaping population structure and not with the objective of variation interpretation. In fact, the method does not allow the variants to be traced back, making it a genome representation method suitable for managing datasets where privacy is a matter of concern. However, given the lack of variant interpretation aspects, this method is not suitable for data representation for genotype-phenotype correlations. Bitome, a method primarily developed to study prokaryotic genomes, represents genomic features at the level of base pairs and is shown to provide an overall profile of genomic features distributed across the genomes. One of the most recent methods, VarNote, performs rapid annotation of genome-scale variants and has been shown to prioritize causal regulatory variants for common diseases. This method utilizes parallel random-sweep searching algorithm and a novel indexing system for the same. However, all these methods are yet to be customized and developed to capture the complexity of the human pangenome and represent the impact of genome-wide variation in disease association studies.

5 Discussion

The current reference human genome assembly serves as a linear, coordinate system for sequence comparisons. While this is useful, differences from the reference are difficult to observe and, except for SNPs, confounding to describe by virtue of being exclusively present or absent in the reference.

Pangenomic methods allow all-to-all comparisons of multiple genomes and derive relations to each-other in the form of a pangenome graph. In this graph, sequences and variations therein are merged into a single coherent data structure. While still undergoing improvisations, broad parameters of sequence graph model, and the input and output data formats are reasonably well defined. Graphical Pangenome are usually represented in Graphical Fragment Assembly format (GFA). Graph Nodes are stored in sequence records (S), edges represented as link (L) records, and embedded sequences in path records. Mappings to GFA can be encoded in GAM (Graph Alignment Map format) or text based Graph Alignment Format (GAF).

Several Pangenome Graph tools for alignment, graph construction and genotyping of small (less than 50 bases–SNP, MNP and small INDELS), medium and large variations (structural variations- >1kb, inversions, balanced translocations, repeat polymorphisms etc) are already available and improvements and improvisations in terms of sensitivity, speed, space and memory utilization with emphasis on scaleup are ongoing. Some of these tools include the Pan Genome Graph Builder (pggb)–a pan genome Graph construction pipeline to create a pangenome graph of multiple genome sequences (https://github.com/pangenome/pggb), the variation graph toolkit -vg, a collection of computational methods for efficient mapping of reads on variation graphs using generalized compressed suffix arrays (Hickey et al., 2020). Assembly and graph visualization tools are also available - these include MoMI-G or Modular Multi-scale Integrated Genome graph browser (Yokoyama et al., 2019), GraphAligner (Rautiainen and Marschall 2020), Pantograph (Chen et al., 2019) and GfaViz (Gonnella, Niehus, and Kurtz 2019), Sequence Tube Map (Beyer et al., 2019), Bandage (Wick et al., 2015) etc. To ensure quality control, Pan Genome Graph Evaluator selects the best pangenome graph construction (https://github.com/pangenome/pgge). The compression of graph data can be achieved through GWBT which is based on Burrows-Wheeler Transform (https://github.com/jltsiren/gbwt).

5.1 Technical challenges

As discussed, a decent start and steady progress has been made with respect to generation, visualization and analysis of Pangenome data and is reflected in multiple aligners, graph generation, indexing and visualization tools (Table 1; Figure 3). Nevertheless, scale up remains the biggest challenge, and as more personal genomes data becomes available, the size of the beginning dataset -thousands of gigabse-scale genomes, is only going to grow exponentially demanding further time, memory and space efficient analysis algorithms and data representation formats.

These methods should not only cater to the linear reference genome, which has been at the core of reporting genomic variations, but also to the emerging datasets from the Human Pangenome studies. It is also important to mention here that standard ontologies are imperative for data interoperability and comparative genomics at scale and therefore the platforms thus developed should have built-in features for the same. Variation barcoding methods are recommended to represent genomic signatures both in the coding and the repeats regions of the genome. These barcodes may facilitate efficient comparison of genome-wide differences in the personal genomes (and also to human pangenome) making them more amenable for downstream analysis. The genomic signatures can be represented at various levels including the number of sites with specific signature, their location, functional annotation and distribution. It is imperative that genome-wide data is viewed at multiple-scales for better comprehension. Towards this, it is proposed that the data generated and analysed in the process may be coded as binary fingerprints, which not only needs less storage space but also makes the retrieval, analysis and sharing more scalable. Such platforms should also have features to annotate various genic signatures, identification of pathogenic variants and several repeat-associated genomic rearrangement signatures including but not limited to target site duplications, 3′ and 5′ flank transductions, insertion-mediated deletion, recombination mediated deletions, etc. (Singh and Mishra 2010). Towards this end, we propose a Personal Genomics Signature Platform (PGSP) - a standard ontology based platform to organize and classify the variation data and develop a universally applicable memory efficient language independent binary fingerprint for all variations that exist in the human genome. The Binary code allotted to each variant with respect to reference human pangenome is expected to facilitate easy classification, storage, retrieval and comparison of such variations across platforms. It will not only allow for data storage more efficiently but will also facilitate data interoperability. The Personal Genomics Signature Platform (PGSP) will allow for detailed annotation of the variation. This platform will also facilitate identification of genome-wide signatures among individual genomes. The algorithms thus generated will be applicable for better understanding of the role of genomic variations in inter-individual differences towards disease predisposition and drug-responses. PGSP should be developed as a universal, language independent, scalable binary digits based ontology for understanding the complex genotype-phenotype associations. The binary fingerprinting is likely to facilitate creation of a modular Minimal Code with Maximum Information (MCMI), shareable across different platforms and languages.

5.2 Ethical, legal, and social implications

As the data set expands from the currently proposed set of 700 haplotypes (350 individuals), one of the major challenges would be to ensure inclusivity. Linguistic, literacy, socio-economic barriers, coupled with the feeling of distrust among the racial-ethinc minorities and the aborigines restrict inclusivity in such projects (Dodson and Williamson 1999), Informed consent of participants requires that the participant be adequately educated about the project and its implications, which is a challenge in itself. Data privacy and protection in the era of open science only add to the ethical and legal complexity–the subjects need to be made consciously aware of how practices such as open science, data sharing and maintenance of electronic health records may impact their data and pose risks to privacy (Couzin-Frankel 2010). Lastly, the extent of information to be released to the subject, post the analysis and annotation of genomic data constitutes another layer of ethical dilemma. Legally, the subject is liable to complete information, but the impact this complete information and its interpretation can have on the subject's personal mental health, of the family and societal attitude towards the subjects provides room for reasonable constraints.

Fully aware of the challenges ahead, the HPRC is armed with a team of ELSI scholars working at the interface of genomics, biomedical ethics, law, social sciences, demography and community engagement. The HPRC is mobilizing Indigenous geneticists, leaders and community members for the outreach programs to ensure development of an authentic and truly representative global reference resource, guided by the FAIR and CARE principles (Carroll et al., 2021)

The scientific challenge of ensuring a transition of a whole generation of researchers from the conventional linear coordinate system based on the reference human genome to a graph-based system is daunting in itself and would demand massive outreach programs *via* physical and online workshops, development of SOPs and user-friendly GUIs.

6 Conclusion

To conclude, the first step towards a better understanding and interpretation of new genome data is to replace the reference human genome–a nonrepresentative, monolithic, monophasic, incomplete standard by a human pangenome–a more accurate, inclusive, representative and complete composite of high-quality multiple telomere to telomere assemblies, maximally capturing the variations that exist in the human population. This would also entail novel representation methods, a new data structure - a graph based architecture and downstream suite of tools to submit, query, retrieve and analyze the pangenome efficiently towards meaningful inferences in a time, memory and cost efficient manner. Built-in interoperability of these platforms must also be ensured so that data from one platform can easily be imported and directly input to the next pipeline facilitating comprehensive evaluation of the inter-individual genomic variations and their functional and clinical significance. As discussed, the biggest challenge, however, will be the ethical, legal and social implications and the training of human resource to the new reference paradigm.

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

SP performed literature review, analysis of computational methods and contributed to manuscript draft. VS and AB performed literature review, scoping and contributed to manuscript draft. All authors finalized the manuscript.

Funding

Authors acknowledge funding from SERB (SPG/2021/004761) and CSIR-IMTECH (OLP0172-20). SP acknowledges fellowship from CSIR (GATE31/0050(13092)/2022-EMR-I).

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.

References

Abinaya, E., Narang, P., and Bhardwaj, A. (2015). Frog - fingerprinting genomic variation ontology. *PLOS ONE* 10 (8), e0134693. doi:10.1371/journal.pone.0134693

Alkan, C., Sajjadian, S., and Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nat. Methods* 8 (1), 61–65. doi:10.1038/nmeth.1527

Altshuler, D., and Donnelly, P.The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* 437 (7063), 1299–1320. doi:10.1038/nature04226

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 21 (1), 30. doi:10.1186/s13059-020-1935-5

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., et al. (2015). A global reference for human genetic variation. *Nature* 526 (7571), 68–74. doi:10.1038/nature15393

Ballouz, S., Alexander, D., and Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biol.* 20 (1), 159. doi:10.1186/s13059-019-1774-4

Beyer, W., Adam, M. N. H., Chan, J., Tan, V., Paten, B., Zerbino, D. R., et al. (2019). Sequence Tube maps: Making graph genomes intuitive to commuters. *Bioinformatics* 35 (24), 5318–5320. doi:10.1093/bioinformatics/btz597

Carroll, S. R., Herczog, E., Hudson, M., Russell, K., and Stall, S. (2021). Operationalizing the CARE and FAIR principles for indigenous data futures. *Sci. Data* 8 (1), 108. doi:10.1038/s41597-021-00892-0

Chen, N-C., Solomon, B., Mun, T., Iyer, S., and Ben, L. (2021). Reference flow: Reducing reference bias using multiple population genomes. *Genome Biol.* 22 (1), 8. doi:10.1186/s13059-020-02229-3

Chen, S., Krusche, P., Dolzhenko, E., RachelSherman, M., Roman, P., Schlesinger, F., et al. (2019). Paragraph: A graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* 20 (1), 291. doi:10.1186/s13059-019-1909-7

Couzin-Frankel, J. (2010). Ethics. DNA returned to tribe, raising questions about consent. *Sci. (New York, N.Y.)* 328 (5978), 558. doi:10.1126/science.328.5978.558

DePristo, M. A., Banks, E., Poplin, R. E., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43 (5), 491–498. doi:10.1038/ng.806

Dodson, M., and Williamson, R. (1999). Indigenous peoples and the morality of the human genome diversity project. *J. Med. Ethics* 25 (2), 204–208. doi:10.1136/jme.25.2.204

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489 (7414), 57–74. doi:10.1038/nature11247

Dvorkina, T., Antipov, D., Korobeynikov, A., and Nurk, S. (2020). SPAligner: Alignment of long diverged molecular sequences to assembly graphs. *BMC Bioinforma.* 21 (12), 306. doi:10.1186/s12859-020-03590-7

Ebler, J., Ebert, P., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., et al. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* 54 (4), 518–525. doi:10.1038/s41588-022-01043-w

Eggertsson, H. P., Jonsson, H., Hjartarson, E., Kehr, B., Masson, G., Hakon, J., et al. (2017). Graphtyper Enables Population-Scale Genotyping Using Pangenome Graphs. *Nat. Genet.* 49 (11), 1654–1660. doi:10.1038/ng.3964

Eggertsson, H. P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Marteinn, T., et al. (2019). GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* 10 (1), 5402. doi:10.1038/s41467-019-13341-9

Eichler, E. E. (2019). Genetic variation, comparative genomics, and the diagnosis of disease. *N. Engl. J. Med.* 381 (1), 64–74. doi:10.1056/NEJMra1809315

Garrison, E., and Marth, G. (2012). Haplotype-Based variant detection from short-read sequencing. *arXiv.* doi:10.48550/arXiv.1207.3907

Glusman, G., Mauldin, D. E., Hood, L. E., and Robinson, M. (2017). Ultrafast comparison of personal genomes via precomputed genome fingerprints. *Front. Genet.* 8, 136. doi:10.3389/fgene.2017.00136

Gonnella, G., Niehus, N., and Kurtz, S. (2019). GfaViz: Flexible and interactive visualization of GFA sequence graphs. *Bioinformatics* 35 (16), 2853–2855. doi:10.1093/bioinformatics/bty1046

Guarracino, A., Simon, H., Nahnsen, S., Prins, P., and Garrison, E. (2022). Odgi: Understanding pangenome graphs. *Bioinformatics* 38 (13), 3319–3326. doi:10.1093/bioinformatics/btac308

Gudmundsson, S., Singer-Berk, M., Watts, N. A., Phu, W., Goodrich, J. K., Solomonson, M., et al. (2022). Variant interpretation using population databases: Lessons from gnomAD. *Hum. Mutat.* 43 (8), 1012–1030. doi:10.1002/humu.24309

Heydari, M., Miclotte, G., Yves Van de, P., and Jan, F. (2018). BrownieAligner: Accurate alignment of illumina sequencing data to de Bruijn graphs. *BMC Bioinforma.* 19 (1), 311. doi:10.1186/s12859-018-2319-7

Hickey, G., Heller, D., Jean, M., Sibbesen, J. A., Sirén, J., Jordan, E., et al. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* 21 (1), 35. doi:10.1186/s13059-020-1941-7

Hu, T., Chitnis, N., Monos, D., and Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Hum. Immunol.* 82 (11), 801–811. doi:10.1016/j.humimm.2021.02.012

Huang, D., Yi, X., Zhou, Y., Yao, H., Xu, H., Wang, J., et al. (2020). Ultrafast and scalable variant annotation and prioritization with big functional genomics data. *Genome Res.* 30 (12), 1789–1801. doi:10.1101/gr.267997.120

Huang, D., Zhou, Y., Yi, X., Fan, X., Wang, J., Yao, H., et al. (2022). VannoPortal: Multiscale functional annotation of human genetic variants for interrogating molecular mechanism of traits and diseases. *Nucleic Acids Res.* 50 (D1), D1408–D1416. doi:10.1093/nar/gkab853

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431 (7011), 931–945. doi:10.1038/nature03001

Jain, A., Bhoyar, R. C., Pandhare, K., Mishra, A., Sharma, D., Imran, M., et al. (2021). IndiGenomes: A comprehensive resource of genetic variants from over 1000 Indian genomes. *Nucleic Acids Res.* 49 (D1), D1225–D1232. doi:10.1093/nar/gkaa923

Kiechle, F. L., Zhang, X., and Holland-Staley, C. A. (2004). The -Omics era and its impact. *Arch. Pathol. Lab. Med.* 128 (12), 1337–1345. doi:10.5858/2004-128-1337-TOEAII

Kim, C. Y., Baek, S., Cha, J., Yang, S., Kim, E., Marcotte, E. M., et al. (2022). HumanNet v3: An improved database of human gene networks for disease research. *Nucleic Acids Res.* 50 (D1), D632–D639. doi:10.1093/nar/gkab1048

Kim, D., Paggi, J. M., Park, C., Bennett, C., and StevenSalzberg, L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37 (8), 907–915. doi:10.1038/s41587-019-0201-4

Kunyavskaya, O., and Prjibelski, A. D. (2019). Sgtk: A toolkit for visualization and assessment of scaffold graphs. *Bioinformatics* 35 (13), 2303–2305. doi:10.1093/bioinformatics/bty956

Lamoureux, C. R., Choudhary, K. S., King, Z. A., Sandberg, T. E., Gao, Y., Sastry, A. V., et al. (2020). The bitome: Digitized genomic features reveal fundamental genome organization. *Nucleic Acids Res.* 48 (18), 10157–10163.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409 (6822), 860–921. doi:10.1038/35057062

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., et al. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46 (D1), D1062–D1067. doi:10.1093/nar/gkx1153

Leggett, R. M., Ramirez-Gonzalez, R. H., Verweij, W., Kawashima, C. G., Iqbal, Z., and MacLean, D. (2013). Identifying and classifying trait linked polymorphisms in non-reference species by walking coloured de Bruijn graphs. *PLOS ONE* 8 (3), e60058. doi:10.1371/journal.pone.0060058

Li, H., Feng, X., and Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* 21 (1), 265. doi:10.1186/s13059-020-02168-z

Li, Q., Tian, S., Yan, B., Liu, C. M., Lam, T. W., Li, R., et al. (2021). Building a Chinese pan-genome of 486 individuals. *Commun. Biol.* 4 (1), 1016. doi:10.1038/s42003-021-02556-6

Liao, W-W., Asri, M., Jana, E., Doerr, D., Haukness, M., Hickey, G., et al. (2022). A draft human pangenome reference. *Prepr. Genomics.* doi:10.1101/2022.07.09.499321

Limasset, A., Cazaux, B., Rivals, E., and Peterlongo, P. (2016). Read mapping on de Bruijn graphs. *BMC Bioinforma.* 17 (1), 237. doi:10.1186/s12859-016-1103-9

Liu, B., Guo, H., Brudno, M., and Wang, Y. (2016). DeBGA: Read alignment with de Bruijn graph-based seed and extension. *Bioinforma. Oxf. Engl.* 32 (21), 3224–3232. doi:10.1093/bioinformatics/btw371

Mikheenko, H., and Kolmogorov, M. (2019). Alla, and mikhail KolmogorovAssembly graph browser: Interactive visualization of assembly graphs. *Bioinforma. Oxf. Engl.* 35 (18), 3476–3478. doi:10.1093/bioinformatics/btz072

Mokveld, T., Linthorst, J., Al-Ars, Z., Henne, H., and Reinders, M. (2020). CHOP: Haplotype-aware path indexing in population graphs. *Genome Biol.* 21 (1), 65. doi:10.1186/s13059-020-01963-y

Mulder, N., Abimiku, A., Adebamowo, S. N., de Vries, J., Matimba, A., Paul, O., et al. (2018). H3Africa: Current perspectives. *Pharmgenomics. Pers. Med.* 11, 59–66. doi:10.2147/PGPM.S141546

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Andrey, V., Vollger, M. R., et al. (2022). Bzikadze, alla MikheenkoThe complete sequence of a human genome. *Science* 376 (6588), 44–53. doi:10.1126/science.abj6987

Paten, B., Novak, A. M., Eizenga, J. M., and Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Res.* 27 (5), 665–676. doi:10.1101/gr.214155.116

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., et al. (2017). DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45 (D1), D833–D839. doi:10.1093/nar/gkw943

Pollard, M. O., Gurdasani, D., AlexanderMentzer, J., Porter, T., and Manjinder, S. S. (2018). Long reads: Their purpose and place. *Hum. Mol. Genet.* 27 (R2), R234–R241. doi:10.1093/hmg/ddy177

Popejoy, A. B., and Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature* 538 (7624), 161–164. doi:10.1038/538161a

Rautiainen, M., and Marschall, T. (2020). GraphAligner: Rapid and versatile sequence-to-graph alignment. *Genome Biol.* 21 (1), 253. doi:10.1186/s13059-020-02157-2

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Stephen, R. F. T., Andrew, O. M., et al. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* 46 (8), 912–918. doi:10.1038/ng.3036

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of african descent. *Nat. Genet.* 51 (1), 30–35. doi:10.1038/s41588-018-0273-y

Shin, J. M., Jeon, J., Jung, D., Kim, K., Kim, Y. J., Jeong, D-H., et al. (2022). PhenGenVar: A user-friendly genetic variant detection and visualization tool for precision medicine. *J. Pers. Med.* 12 (6), 959. doi:10.3390/jpm12060959

Sibbesen, J. A., Maretty, L., Anders, K., and Krogh, A. (2018). Accurate genotyping across variant classes and lengths using variant graphs. *Nat. Genet.* 50 (7), 1054–1059. doi:10.1038/s41588-018-0145-5

Singh, V., and Mishra, R. K. (2010). RISCI-Repeat induced sequence changes identifier: A comprehensive, comparative genomics-based, *in silico* subtractive hybridization pipeline to identify repeat induced sequence changes in closely related genomes. *BMC Bioinforma.* 11, 609. doi:10.1186/1471-2105-11-609

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Miles, J. E., et al. (2015). Big data: Astronomical or genomical? *PLoS Biol.* 13 (7), e1002195. doi:10.1371/journal.pbio.1002195

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., et al. (2021). Sequencing of 53, 831 diverse genomes from the NHLBI TOPMed program. *Nature* 590 (7845), 290–299. doi:10.1038/s41586-021-03205-y

VenterMyers, J. C., Mark, D. A., Eugene, W., PeterLiMural, W. R. J., Granger, G., Sutton, H. O. S., et al. (2001). The sequence of the human genome. *Science* 291 (5507), 1304–1351. doi:10.1126/science.1058040

Vollger, M. R., Guitart, X., Dishuck, P. C., Ludovica Mercuri, W. T. H., Gershman, A., Diekhans, M., et al. (2022). Segmental duplications and their variation in a complete human genome. *Science* 376 (6588), eabj6965. doi:10.1126/science.abj6965

Wall, J. D., Stawiski, E. W., Ratan, A., Kim, H. L., Kim, C., Gupta, R., et al. (2019). The GenomeAsia 100K project enables genetic discoveries across asia. *Nature* 576 (7785), 106–111. doi:10.1038/s41586-019-1793-z

Wang, J., Huang, D., Zhou, Y., Yao, H., Liu, H., Zhai, S., et al. (2020). CAUSALdb: A database for disease/trait causal variants identified using summary statistics of genome-wide association studies. *Nucleic Acids Res.* 48 (D1), D807–D816. doi:10.1093/nar/gkz1026

Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., et al. (2022). The human pangenome project: A global resource to map genomic diversity. *Nature* 604 (7906), 437–446. doi:10.1038/s41586-022-04601-8

Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage: Interactive visualization of de Novo genome assemblies. *Bioinforma. Oxf. Engl.* 31 (20), 3350–3352. doi:10.1093/bioinformatics/btv383

Yokoyama, T. T., Sakamoto, Y., Seki, M., Suzuki, Y., and Kasahara, M. (2019). MoMI-G: Modular multi-scale integrated genome graph browser. *BMC Bioinforma.* 20 (1), 548. doi:10.1186/s12859-019-3145-2

frontiers | Frontiers in Genetics

TYPE Brief Research Report
PUBLISHED 22 November 2022
DOI 10.3389/fgene.2022.1020100

Check for updates

OPEN ACCESS

EDITED BY
Marco Pellegrini,
National Research Council (CNR), Italy

REVIEWED BY
Guang-Zhong Wang,
Shanghai Institute of Nutrition and Health (CAS), China
Shang-Qian Xie,
University of Idaho, United States

*CORRESPONDENCE
Glauber Wagner,
glauber.wagner@ufsc.br

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 15 August 2022
ACCEPTED 11 November 2022
PUBLISHED 22 November 2022

CITATION
Maia GA, Filho VB, Kawagoe EK, Teixeira Soratto TA, Moreira RS, Grisard EC and Wagner G (2022), AnnotaPipeline: An integrated tool to annotate eukaryotic proteins using multi-omics data.
Front. Genet. 13:1020100.
doi: 10.3389/fgene.2022.1020100

COPYRIGHT
© 2022 Maia, Filho, Kawagoe, Teixeira Soratto, Moreira, Grisard and Wagner. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

AnnotaPipeline: An integrated tool to annotate eukaryotic proteins using multi-omics data

Guilherme Augusto Maia[1], Vilmar Benetti Filho[1], Eric Kazuo Kawagoe[1], Tatiany Aparecida Teixeira Soratto[1], Renato Simões Moreira[1,2], Edmundo Carlos Grisard[1,3] and Glauber Wagner[1,3]*

[1]Laboratório de Bioinformática, Universidade Federal de Santa Catarina (UFSC), Campus João David Ferreira Lima, Florianópolis, Brazil, [2]Instituto Federal de Santa Catarina (IFSC), Campus Lages, Lages, Brazil, [3]Laboratório de Protozoologia, Universidade Federal de Santa Catarina (UFSC), Campus João David Ferreira Lima, Florianópolis, Brazil

Assignment of gene function has been a crucial, laborious, and time-consuming step in genomics. Due to a variety of sequencing platforms that generates increasing amounts of data, manual annotation is no longer feasible. Thus, the need for an integrated, automated pipeline allowing the use of experimental data towards validation of *in silico* prediction of gene function is of utmost relevance. Here, we present a computational workflow named AnnotaPipeline that integrates distinct software and data types on a proteogenomic approach to annotate and validate predicted features in genomic sequences. Based on FASTA (i) nucleotide or (ii) protein sequences or (iii) structural annotation files (GFF3), users can input FASTQ RNA-seq data, MS/MS data from mzXML or similar formats, as the pipeline uses both transcriptomic and proteomic information to corroborate annotations and validate gene prediction, providing transcription and expression evidence for functional annotation. Reannotation of the available *Arabidopsis thaliana*, *Caenorhabditis elegans, Candida albicans, Trypanosoma cruzi,* and *Trypanosoma rangeli* genomes was performed using the AnnotaPipeline, resulting in a higher proportion of annotated proteins and a reduced proportion of hypothetical proteins when compared to the annotations publicly available for these organisms. AnnotaPipeline is a Unix-based pipeline developed using Python and is available at: https://github.com/bioinformatics-ufsc/AnnotaPipeline.

KEYWORDS

workflow, proteogenomics, genome annotation, functional annotation, hypothetical proteins

Introduction

Genome annotation involves a detailed description and understanding of the genome structure and assignment of biological functions to the genes (Stein, 2001). Structural annotation thus characterizes the physical structure of coding and non-coding regions on a given genome, resulting in a physical map of the genes' number and positioning. Along determination of the structure and organization of the protein-coding sequences (CDS) located within open reading frames (ORF) of each gene, annotation also includes a description of other genomic elements such as promoters and enhancers (Korf, 2004; Danchin et al., 2018). Several computational tools known as gene predictors, such as AUGUSTUS (Stanke and Waack, 2003) and GeneMark (Brůna, Lomsadze, and Borodovsky, 2020), have been widely used to perform structural annotation (Yandell and Ence, 2012).

Functional annotation consists of assigning biological information to genes, such as their involvement in biological processes, molecular functions, presence of functional protein domains, and subcellular localization, among others (Stein, 2001; Yandell and Ence, 2012). The assignment of biological functions to protein-coding genes is generally performed through similarity analysis with databases containing previously annotated protein sequences using sequence aligners such as BLAST (Camacho et al., 2009) or DIAMOND (Buchfink, Reuter, and Drost, 2021). The biological function of a predicted CDS is therefore assumed to be the same as the protein in the database that demonstrates the most significant similarity, leading to an annotation transfer (Hegyi and Gerstein, 2001). Thus, the accuracy of the annotated database is fundamental for genome annotation, allowing the quality of downstream analyses based on the transferred annotations. Especially with the use of high-throughput sequencing during the past years, several public genomic and proteomic databases from a variety of organisms are nowadays available. However, the exponential growth of datasets impairs the quality of a proper and detailed structural and functional annotation of genomes. For that, the use of curated databases such as SwissProt/UniProtKB (The UniProt Consortium, 2021) and Ensembl (Flicek et al., 2014), or even organism-specific databases, such as those contained in the VEuPathDB (Amos et al., 2022), is highly recommended to ensure high quality to the genome annotation.

Considering the growing datasets of genomic and proteomic databases, and the specific genomic features across taxa, combining different computational tools or pipelines to automatically assess gene structural and functional annotation has been widely used (Danchin et al., 2018). Composed of a set of data processing methods connecting inputs and outputs in series, automated pipelines can perform genome annotation by sequence similarity (Hyatt et al., 2010; Steinbiss et al., 2016) or functional annotation of proteins (Gotz et al., 2008; Vlasova et al., 2021; Törönen and Holm, 2022). Nevertheless, only a few genome annotation pipelines use expression experimental data (RNA-Seq or MS/MS) to validate the *in silico* annotation (Ghali et al., 2014; Sheynkman et al., 2014).

Large-scale genomic and transcriptomic studies based on high-throughput sequencing platforms in the past decade have provided increasing amounts of data (Kumar et al., 2016a), also providing extensive gene expression profiles based on transcribed RNAs (RNA-seq) sequencing. Moreover, extensive proteomic data acquired from sensitive mass spectrometry (MS) technologies are available from several databases (Vaudel et al., 2016), such as PRIDE (Perez-Riverol et al., 2022), MassIVE (Miao et al., 2012), and the ProteomeXchange Consortium (Vizcaíno et al., 2014). Thus, using transcription and expression evidence to annotate newly predicted CDS or reannotate formerly analyzed genomes would reveal novel biological aspects. The proteogenomic approach allows the cross-validation of genomic, transcriptomic, and proteomic data on both intra- and inter-specific analyzes (Nesvizhskii, 2014). However, this approach requires novel computational methods and pipelines. Thus, integrating the classic annotation analysis by sequence similarity with customizable parameters and databases, combined with functional prediction validated with RNA-seq and MS/MS data evidence, would enhance genome annotation as an essential step toward comprehending biological mechanisms.

In this study, we developed AnnotaPipeline, a proteogenomic computational tool for automatic annotation of eukaryotic genomes using support from high-throughput transcriptomic and proteomic data, allowing validation of gene function and expression.

Methods

AnnotaPipeline

Development and overview

The AnnotaPipeline overall scheme and processes are shown in Figure 1. This pipeline was developed using Python and runs on Unix-based systems, consisting of a series of tolls and in-house scripts for data preparation, processing, and analysis. Documentation related to installation instructions and scripts to run AnnotaPipeline are available at https://github.com/bioinformatics-ufsc/AnnotaPipeline.

Input and configuration files

AnnotaPipeline requires the input of at least one of the following different FASTA files: 1) a nucleotide sequence file, 2) a protein sequence file, 3) a protein sequence file, and structural annotation files in GFF3 format. If the first option is selected, AnnotaPipeline will perform gene prediction on the provided nucleotide sequence. Therefore, it is essential to use a trained AUGUSTUS model for the gene prediction process

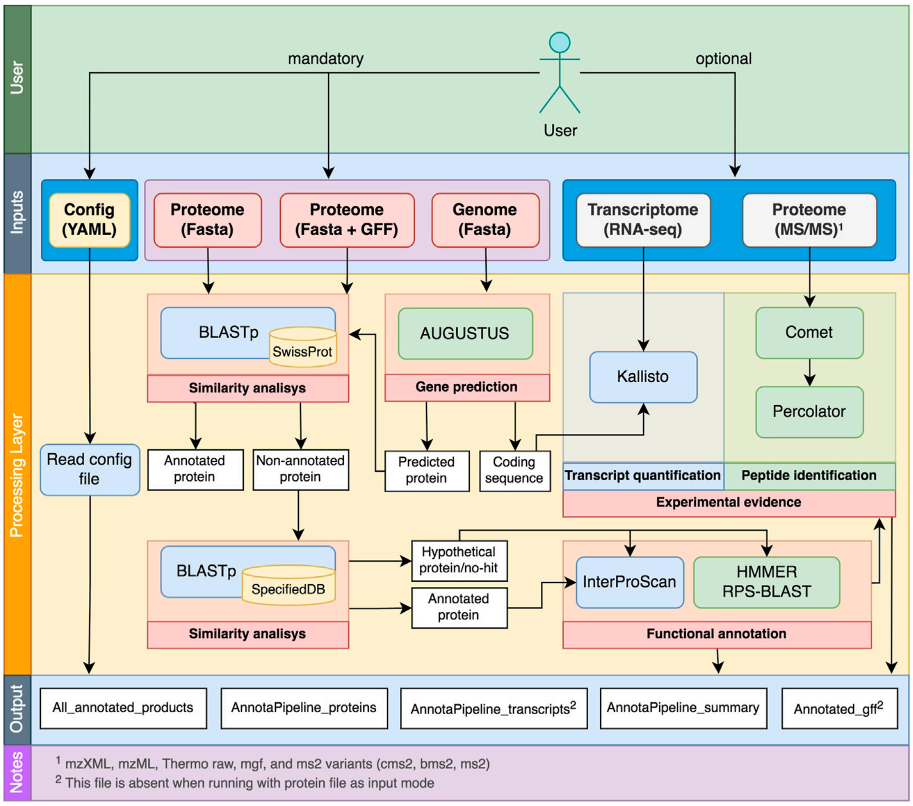


FIGURE 1
Overview of AnnotaPipeline workflow, indicating the optional and the required inputs from the user, the internal processes, and the output layers.

before executing the pipeline. This execution will produce an annotated GFF3, and CDS sequences will contain a complete header. For the second option, gene prediction will be skipped, and the final output file will contain only a simplified sequence header. The third option is executed equally to the second option, the pipeline will include annotations for each CDS from the provided GFF file. Also, it is recommended that the submitted GFF file is in GFF3 format, preferably from a previous AUGUSTUS gene prediction.

Aside from the molecular data input, it is also required from the user to access the YAML configuration file prior to running the pipeline, where locations of both software and databases required for the personalized analysis must be provided. Similarly, if analyses with experimental data will be carried out, it is also necessary to provide the locations of folders containing RNA-seq and MS/MS data.

Users can define the number of processing threads that will be used during the execution of the pipeline (default is set to 4 threads) and are required to define the cutoff parameters and specific keywords to classify hypothetical proteins during the similarity analysis process. This configuration step is facilitated if the user installs AnnotaPipeline using Conda from the environment file available at https://github.com/bioinformatics-ufsc/AnnotaPipeline.

Annotation process

The annotation process starting with a genomic file input is divided into three steps. Initially, gene prediction is performed by AUGUSTUS (Stanke and Waack, 2003). Although AnnotaPipeline is mainly focused on eukaryotic organisms, the pipeline accepts input of further gene prediction training models if absent in the AUGUSTUS standalone version. It is recommended to use the WebAUGUSTUS platform to generate custom training models (Hoff and Stanke, 2013).

Following gene prediction, the annotation process continues into similarity analysis performed by the BLASTp algorithm (Camacho et al., 2009) using (i) the SwissProt database, which contains about 570,000 manually curated protein sequences from a wide variety of organisms (The UniProt Consortium, 2021), and (ii) a user-specified database such as TrEMBL/UniProtKB, VEuPathDB and GenBank NR, or additional databases that must be specified in the AnnotaPipeline.yaml configuration file. Despite the used database, the pipeline contains parsing scripts that automatically will transfer the protein annotation for the predicted CDS on the output file. Proteins are then classified into three groups: annotated proteins (known function), hypothetical proteins, and no-hit proteins. Annotated proteins are those with attributed annotation either by the SwissProt or the user-specified database. In AnnotaPipeline, hypothetical proteins are considered those presenting similarities with proteins with no specific annotation in the databases (unknown function) and that contain filter keywords in their descriptions, such as “fragment”, “hypothetical”, “partial”, “uncharacterized”, “unknown”, and “unspecified”. These are the default keywords used by the pipeline, but users can change these in the AnnotaPipeline.yaml configuration file. Annotations in subject proteins will be disregarded if at least one description contains any of the provided keywords. No-hit proteins are proteins with no available match, and therefore no annotation, in either database used in the similarity analysis step. For downstream analysis steps, the no-hit and the hypothetical proteins are grouped by the pipeline. Furthermore, proteins revealing no matches with databases and presenting no supporting evidence from experimental data are considered true negative proteins.

The third step consists of the functional annotation of proteins, starting with analyzing both annotated and hypothetical protein groups by InterProScan software (Jones et al., 2014). Exclusively for the hypothetical protein/no-hit group, further analysis using the hmmscan algorithm of the HMMER suite (Finn, Clements, and Eddy, 2011) and the RPS-BLAST (Camacho et al., 2009) are performed. The resulting functional annotation is contained in a single output file where all predicted proteins will be annotated and can be used as input for the experimental validation analyses.

Experimental validation with proteogenomic data

The AnnotaPipeline accepts the input of RNA-seq and MS/MS data that will allow experimental validation of CDS prediction and annotation. Upon activation of the experimental analysis module, transcriptomic data will be processed by Kallisto (Bray et al., 2016), which performs a pseudo-alignment of RNA-seq reads to the annotated protein file. The result will be refined based on a quantification of aligned transcripts, which are accounted for transcripts per million (TPM). Users may concatenate their transcriptomic data into a single FASTQ file (for single-end RNA-seq) or two FASTQ files (R1 and R2, for paired-end RNA-seq) to run multiple experiments at once. For experimental validation using proteomic data (MS/MS), users can provide a single folder containing their MS/MS data files to run multiple experiments simultaneously. The search for MS/MS-derived peptides among the annotated proteins will be performed using Comet (Eng, Jahan, and Hoopmann, 2013), following the user-provided search parameters in comet.params configuration file, generating the input for the Percolator software (The et al., 2016). Then, the proteomic data will be searched among the annotated proteins dataset and parsed by the q-value threshold of the Percolator software.

Output files

The pipeline will create a log file and an output folder in the AnnotaPipeline directory. The log file contains details of script processing, software execution, and outputs of each computational tool. Also, this log may contain any possible warnings or errors relative to the software execution. Within the output folder, the pipeline will create (i) two FASTA files containing the annotated proteins and their respective annotated CDS, (ii) a GFF file including a transcript product field containing the final annotation for each CDS, (iii) a TXT file containing the all CDS product ID and annotated description, and (iv) a TSV file summarizing all annotated CDS and information regarding transcription (RNA-Seq) or expression (MS/MS) evidence. In addition to these main output files, within each of the folders created by AnnotaPipeline, other outputs can help the user manually curate the annotations suggested by the pipeline (Supplementary Table S1).

Comparative evaluation of AnnotaPipeline performance

Performance tests were carried out using a computational cluster equipped with 40 threads processor (3.2 GHz), 285 GB RAM memory (DDR4, 2,400 MHz), and 5 TB storage space (2.5 SATA HD, 7,200 RPM). Storage was mainly used for RNA-seq and MS/MS data of the testing organisms. Despite the availability of computing power, the number of processing threads used for testing was set to 12 in the AnnotaPipeline.yaml configuration file.

Molecular data from three different model organisms were used to test AnnotaPipeline: *Arabidopsis thaliana* (strain TAIR10), an essential model for plant biology and genetics; *Caenorhabditis elegans* (strain WBcel235), an important model for molecular and developmental biology; and *Candida albicans* (strain SC5314), a fungal pathogen model. Genomic data for each of these organisms were retrieved from GenBank under the following accession numbers: GCA_000001735.2, GCA_000002985.3, and GCA_000182965.3, respectively. RNA-seq data for each of these organisms were obtained from BioProject/NCBI under the following accession numbers: PRJNA779571, PRJNA809747, and PRJNA750749 for *A. thaliana*; PRJNA734346, PRJNA658149, and PRJNA755869 for *C. elegans*; PRJNA714869, PRJNA496318, PRJNA752883, and PRJNA744166 for *C. albicans*. MS/MS data for each of these organisms were obtained from ProteomeXchange, under the following accession numbers:

PXD012708 and PXD010730 for *A. thaliana*; PXD025128 for *C. elegans*; PXD005364 for *C. albicans*.

For the similarity analysis step, in addition to the SwissProt database, a specific database of protein sequences was used for each model organism: for *A. thaliana*, a subset of 370,680 protein sequences was obtained from the GenBank NR dataset; for *C. albicans*, the FungiDB v56 containing 2,331,868 protein sequences was obtained from VEuPathDB; and for *C. elegans*, a subset of 23,010 protein sequences was obtained from TrEMBL.

AnnotaPipeline was independently run with default parameters for every organism, using the genome FASTA file obtained for each organism as input. AUGUSTUS (version 3.4.0) prediction was performed with the gene model argument set to partial and using the prediction model dataset already provided by the software, as in: arabidopsis, for *A. thaliana*; *candida*_albicans, for *C. albicans*; and *caenorhabditis*, for *C. elegans*. Therefore, the gene prediction step was not optimized. BLASTp (version 2.12.0) execution was done assuming an e-value of 1e-5, the number of maximum target sequences set to 10. Also, a minimum threshold value of sequence coverage was set to 30, sequence identity 40, and sequence positivity 60 for the annotation transfer. The annotation was chosen based on the highest bit score between the analyzed sequences.

InterProScan (version 5.52–86.0) was run for the functional annotation step, allowing for the lookup of corresponding Gene Ontology annotation (--goterms). HMMscan (version 3.3.2) had the e-value of both sequences and domains set to 1e-5, and RPSblast (version 2.12.0) also had the minimum e-value of target sequences set to 1e-5. Kallisto (version 0.48.0) pseudo-alignment of RNA-seq dataset was run with 1,000 bootstraps, and the minimum threshold of TPM was selected as the mean. Comet (version 2021.01) was run for each MS/MS dataset with a scan range minimum and a maximum set to 200 and 4,000, respectively. After, Percolator (version 3.5) was run with Comet output files, and the results obtained were filtered by a q-value threshold of 0.05. As a complete example, all the output files from the *A. thaliana* dataset are available at https://github.com/bioinformatics-ufsc/AnnotaPipeline/blob/v1.0/Output%20Example/Annota_Athaliana.tar.xz.

The pipeline was further tested using two taxonomically close protozoa species of medical relevance containing over 50% of their CDS annotated as hypothetical proteins: *Trypanosoma cruzi* (strain Sylvio X10/1), the etiological agent of Chagas disease (Talavera-López et al., 2021) and *Trypanosoma rangeli* (strain SC58) an avirulent trypanosomatid of mammals (Stoco et al., 2014). Genomic data was retrieved from the TriTrypDB (version 57) under the following accession numbers: DS_107bdce9bb, and DS_9d0531db8e, respectively. For both organisms, the Augustus prediction model was trained online based on their respective available genome file and annotated transcripts files (tcruzi_sylviox10, for *T. cruzi*; and trypanosoma_rangeli, for *T. rangeli*). For the similarity analysis step, a database of 648,560 protein sequences obtained from the TriTrypDB was used, along with the mandatory SwissProt database. The AnnotaPipeline was run using default parameters for both trypanosomatid species, as previously mentioned.

Results

AnnotaPipeline workflow

The complete execution of AnnotaPipeline resulted in the expected output files that were named <basename>_AnnotaPipeline_<file>.<format>, allowing users to identify the results and perform multiple experiments in the same directory by swapping the <basename> of the experiments in the AnnotaPipeline.yaml configuration file.

The generated annotation files in FASTA format display for each sequence a header containing the following information separated by a pipe character "|": sequence identification; source organism; scaffold number; CDS start; CDS end; strand orientation; and sequence description, were functional annotations provided by GO and IPR are included. If no structural annotation GFF file is included in the analysis, information concerning strand orientation and scaffold location will be absent. Also, AnnotaPipeline changes the "transcript product" field of each CDS in the annotated GFF file to the corresponding sequence description present in the header of the FASTA file.

Comparative analysis of AnnotaPipeline results

AnnotaPipeline was comparatively tested using genomic data of different model organisms for which genome annotation is available. The pipeline enabled experimental evidence analyses and no gene prediction optimization. The summary of the obtained annotations, functional annotations, and experimental evidence results for the *A. thaliana*, *C. albicans*, and *C. elegans* datasets are presented in Table 1.

For *A. thaliana*, the pipeline annotated a total of 19,651 protein sequences in 29 h and 07 min; 5,377 protein sequences for *C. albicans* in 10 h and 06 min; and 14,278 protein sequences for *C. elegans* in 20 h and 58 min.

Among the genome analyzed, *C. albicans* had the highest percentage of annotated proteins with 99.48%, followed by *A. thaliana* with 98.90%. *C. elegans* had 22.62% of their protein sequences annotated as hypothetical proteins, and another 3.24% of proteins with no matches available in the analyzed databases. Comparatively to the current data from analyzed genomes available in public databases, AnnotaPipeline provided a

TABLE 1 Summary of AnnotaPipeline annotations, functional annotations, and experimental evidence results for different model organisms.

Parameter	*Arabidopsis thaliana* TAIR10		*Candida albicans* SC5314		*Caenorhabditis elegans* WBcel235	
	GenBank	AnnotaPipeline	GenBank	AnnotaPipeline	GenBank	AnnotaPipeline
Predicted proteins	27,562	19,651	6,043	5,377	19,984	14,278
Annotated proteins	25,151 (91.25%)	19,434 (98.90%)	3,735 (61.81%)	5,349 (99.48%)	13,186 (65.98%)	10,587 (74.15%)
Annotated by SwissProt	–	13,444 (69.18% of annotated)	–	2,914 (54.48% of annotated)	–	5,395 (50.96% of annotated)
Annotated by SpecificDB	–	5,990 (30.82% of annotated)	–	2,435 (45.52% of annotated)	–	5,192 (49.04% of annotated)
Hypothetical proteins	2,411	169	2,308	13	6,798	3,229
No hit proteins (true negative)*	–	48 (45)	–	15 (9)	–	462 (440)
Total hypothetical proteins	2,411 (8.75%)	217 (1.10%)	2,308 (38.19%)	28 (0.52%)	6,798 (34.02%)	3,691 (25.85%)
Proteins with at least 1 IPR term	–	17,974 (91.47%)	–	4,704 (87.48%)	–	11,050 (77.39%)
Proteins with at least 1 GO term	–	13,612 (69.27%)	–	3,705 (68.90%)	–	7,587 (53.14%)
Proteins with transcript evidence	–	3,228 (16.43%)	–	716 (13.32%)	–	1,714 (12.0%)
Proteins with peptide evidence	–	1,546 (7.87%)	–	809 (15.05%)	–	0 (%)

*True negative are proteins with no match on studied databases and no supporting evidence from experimental data, which could possibly be artifacts from gene prediction. Reference genome GenBank accession number: *Arabidopsis thaliana* (strain TAIR10) = GCA_000001735.2; *Caenorhabditis elegans* (strain WBcel235) = GCA_000002985.3; *Candida albicans* (strain SC5314) = GCA_000182965.3.

higher number of annotated proteins (known function) and fewer hypothetical proteins. Consequently, the number of hypothetical proteins in the *A. thaliana* dataset went down from 8.75% to 1.10% using the AnnotaPipeline, while for *C. elegans* and *C. albicans* datasets, the reduction was from 34.02% to 25.85% and 38.19%–0.52%, respectively.

Functional annotation of the *A. thaliana*, *C. albicans* and *C. elegans* genomes using the AnnotaPipeline revealed 69.27%, 68.90%, and 53.14% of their CDS associated with at least one GO term associated, respectively. When RNA-Seq and MS/MS data were included for the analysis of experimental evidence of transcription or expression, *A. thaliana*, *C. albicans* and *C. elegans* had 16.43%, 15.05%, and 12.00% of their annotated proteins validated with transcriptomic and proteomic data, respectively. Interestingly, no *C. elegans* annotated CDS were validated by the available MS/MS dataset.

Comparative analysis of the genome annotation for *T. cruzi* and *T. rangeli* retrieved from the TriTrypDB (version 57) and the annotation generated using AnnotaPipeline is shown in Supplementary Table S2. Although not including experimental data for validation (RNA-Seq or MS/MS), the pipeline was able to reduce the number of hypothetical proteins by 60.46% and 42.84% for *T. cruzi* and *T. rangeli*, respectively, while increasing the proportion of annotated CDS having at least one GO term assigned (Supplementary Table S2).

Considering the annotation provided by AnnotaPipeline, it is possible to classify the annotated protein sequences into eight different categories based on three different criteria: 1) available annotation based on sequence similarity with provided databases; 2) transcription evidence by quantifying RNA-seq reads; and 3) translation evidence supported by the identification of peptides matches from MS/MS information. As an example, result of the analysis of the *A. thaliana* dataset is shown in Table 2. From a total of 19,651 annotated CDS, the less represented categories are those who contains CDS having support from either RNA-Seq (12.65%) or MS/MS (4.09%) support, or both (3.78%).

Discussion

Whole genome annotation is one of the first and most essential steps in any genome study, consisting in a time-consuming and laborious work depending on the genome size, and no longer can be performed manually due to the amount of data generated by high-throughput sequencing (Ouzounis and Karp, 2002). AnnotaPipeline was designed to perform automatic annotation of genomes, having the unique feature to include experimental data derived from transcriptomic (RNA-Seq) or proteomic (MS/MS) approaches towards experimental validation of an annotated CDS. The pipeline is easy to install, runs on operating systems that support

TABLE 2 Classification table of annotated proteins by AnnotaPipeline for the *Arabidopsis thaliana* dataset.

Categories	Hypothetic Annotation	Transcript Evidence	Peptide Evidence	Number of sequences	Percentage (%)
1	Yes	No	No	203	1.03
2	No	No	No	15,417	78.45
3	Yes	Yes	No	5	0.03
4	Yes	No	Yes	7	0.04
5	No	Yes	No	2,480	12.62
6	No	No	Yes	796	4.05
7	Yes	Yes	Yes	2	0.01
8	No	Yes	Yes	741	3.77

command-line options, such as Unix-based systems, and does not require high computational demands, although the time-consuming tasks can be reduced while using more robust machines. It is also user-friendly and customizable to meet the user needs in terms of analysis stringency.

Although distinct genome annotation pipelines are available (Gotz et al., 2008; Hyatt et al., 2010; Ghali et al., 2014), AnnotaPipeline provides the possibility of using RNA-seq and MS/MS data to improve genome annotation simultaneously. Considering that proteomic data have become increasingly accessible (Nesvizhskii, 2014), and new RNA-seq technologies, such as single-cell or single-molecule sequencing, are improving significantly (Wang et al., 2019), the use of this pipeline would increase que quality and accuracy of the annotated genomes from a variety of organisms by providing several possible annotations for each protein sequence. On top of providing a more accurate automated analysis, the pipeline also offers information to support manual curation of the annotation by the user.

Comparison of the results obtained using AnnotaPipeline with the data available in public databases, it was possible to observe a reduction in the number of hypothetical proteins for *A. thaliana* (91.0%), *C. elegans* (45.70%), and *C. albicans* (98.79%), as shown in Table 1. This reduction can be due to the use of customizable databases and keywords but also to the use of combined proteogenomic data to complement gene annotation, increasing the reliability of gene prediction and automatic annotation.

In addition to these well-annotated genomes, AnnotaPipeline also showed good performance when used to annotate the repetitive genomes from two closely related species of *Trypanosoma* (*T. cruzi* and *T. rangeli*) retrieved from TriTrypDB, both lacking RNA-seq or MS/MS data for experimental validation. It was possible to observe a relative reduction of more than 60% in the number of proteins annotated as hypothetical (Supplementary Table S2).

The use of experimental data to validate CDS annotation raises a critical discussion, especially regarding hypothetical proteins. Categorizing hypothetical proteins according to their evidence of transcription or expression by AnnotaPipeline revealed interesting results. Although presenting experimental support from RNA-Seq, MS/MS or both, as observed for *A. thaliana* proteins belonging to Class 7 (Table 2), they remain annotated as hypothetical proteins in the studied databases. In this context, annotation pipelines using this multi-omics approach can provide fundamental insights into new and uncharacterized proteins and revise those whose functions are already annotated. Knowledge areas associated with medicine would benefit most since previously annotated hypothetical proteins could now be studied and thus allow for the re-evaluation of disease diagnosis or prognostic methods (Kumar et al., 2016b).

Furthermore, AnnotaPipeline can be used to guide the exploration of proteins because it adds functional annotation to protein annotation through the incorporation of GO and IPR terms. Especially for hypothetical or uncharacterized proteins, the classical description of annotations might not be biologically informative, so the lack of functional annotations (such as GO or IPR terms) increases this information gap (Lubec et al., 2005; Gotz et al., 2008). AnnotaPipeline provides descriptive and functional information for these proteins during the automated annotation process, which helps to identify potential prediction artifacts and streamline the process of manually curating the annotations. Lastly, the AnnotaPipeline summary file can provide to users the SUPERFAMILY protein information, adding yet another layer of detail to annotations. This information can provide new insights into the functionality of uncharacterized proteins, as they represent possibilities of new structures and functions to be explored (Lubec et al., 2005).

Conclusion

By integrating experimental data from RNA-seq and MS/MS analyses to validate prediction and annotations of protein-coding sequences, AnnotaPipeline, an integrated and modular genomic annotation pipeline, promoted the reduction of the number of hypothetical proteins for various organisms. The use of this original proteogenomic approach on reannotation of *A. thaliana*, *C. elegans*, *C. albicans*, *T. cruzi*, and *T. rangeli* datasets, have increased the proportion of annotated proteins, consequently reducing the number of hypothetical proteins if compared to the currently available annotation. AnnotaPipeline was developed as a generalist annotation pipeline, allowing the assessment of genomes from any eukaryotic organism with available molecular data.

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 in the article/Supplementary Material.

Author contributions

GM: participated in study design and manuscript writing. VF: participated in study design and manuscript writing; EK: participated in study design and manuscript writing. TS participated in study design and manuscript writing. RM: participated in study design and manuscript writing. EG: participated in manuscript writing. GW: participated in coordination, study design and manuscript writing.

Funding

This research was funded by CAPES (Coordination for the Improvement of Higher Education Personnel, Brazil)—Finance Code 001, CNPq (National Council for Scientific and Technological Development, Brazil), FAPESC (Santa Catarina Research Foundation) and UFSC (Federal University of Santa Catarina). GM and VF were recipients of CAPES scholarship. EK was recipient of CNPq scholarship. TS was recipient of FAPESC scholarship. The open access publication fee was funded by CAPES.

Acknowledgments

The development process and experimental steps were performed in the Laboratory of Bioinformatics, in the Department of Microbiology, Immunology, and Parasitology (MIP), in the Center for Biological Sciences (CCB) of the Federal University of Santa Catarina (UFSC), Brazil. The authors would like to thank Superintendencia de Governanca Eletronica e Tecnologia da Informacao e Comunicacao (SeTIC/UFSC) for the computational infrastructure support.

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.

Author disclaimer

The opinions expressed by authors contributing to this journal do not necessarily reflect the opinions of the Federal University of Santa Catarina, Brazil, or the institutions with which the authors are affiliated. The funders had no role in the study design, data analysis, or the decision to publish.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1020100/full#supplementary-material

References

Amos, B., Aurrecoechea, C., Barba, M., Barreto, A., Basenko, E. Y., Bazant, W., et al. (2022). VEuPathDB: The eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Res.* 50 (D1), D898–D911. doi:10.1093/nar/gkab929

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34 (5), 525–527. doi:10.1038/nbt.3519

Brůna, T., Lomsadze, A., and Borodovsky, M. (2020). GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. *Nar. Genom. Bioinform.* 2 (2), lqaa026. doi:10.1093/nargab/lqaa026

Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* 18 (4), 366–368. doi:10.1038/s41592-021-01101-x

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinforma.* 10 (1), 421. doi:10.1186/1471-2105-10-421

Danchin, A., Ouzounis, C., Tokuyasu, T., and Zucker, J. D. (2018). No wisdom in the crowd: Genome annotation in the era of big data - current status and future prospects. *Microb. Biotechnol.* 11 (4), 588–605. doi:10.1111/1751-7915.13284

Eng, J. K., Jahan, T. A., and Hoopmann, M. R. (2013). Comet: An open-source MS/MS sequence database search tool. *PROTEOMICS* 13 (1), 22–24. doi:10.1002/pmic.201200439

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi:10.1093/nar/gkr367

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., et al. (2014). Ensembl 2014. *Nucleic Acids Res.* 42 (D1), D749–D755. doi:10.1093/nar/gkt1196

Ghali, F., Krishna, R., Perkins, S., Collins, A., Xia, D., Wastling, J., et al. (2014). ProteoAnnotator - open source proteogenomics annotation software supporting PSI standards. *PROTEOMICS* 14 (23–24), 2731–2741. doi:10.1002/pmic.201400265

Gotz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36 (10), 3420–3435. doi:10.1093/nar/gkn176

Hegyi, H., and Gerstein, M. (2001). Annotation transfer for genomics: Measuring functional divergence in multi-domain proteins. *Genome Res.* 11 (10), 1632–1640. doi:10.1101/gr.183801

Hoff, K. J., and Stanke, M. (2013). WebAUGUSTUS — A web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic Acids Res.* 41 (W1), W123–W128. doi:10.1093/nar/gkt418

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* 11 (1), 119. doi:10.1186/1471-2105-11-119

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30 (9), 1236–1240. doi:10.1093/bioinformatics/btu031

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinforma.* 5, 59. doi:10.1186/1471-2105-5-59

Kumar, D., Bansal, G., Narang, A., Basak, T., Abbas, T., and Dash, D. (2016b). Integrating transcriptome and proteome profiling: Strategies and applications. *PROTEOMICS* 16 (19), 2533–2544. doi:10.1002/pmic.201600140

Kumar, D., Yadav, A. K., Jia, X., Mulvenna, J., and Dash, D. (2016a). Integrated transcriptomic-proteomic analysis using a proteogenomic workflow refines rat genome annotation. *Mol. Cell. Proteomics* 15 (1), 329–339. doi:10.1074/mcp.M114.047126

Lubec, G., Afjehi-Sadat, L., Yang, J. W., and John, J. P. P. (2005). Searching for hypothetical proteins: Theory and practice based upon original data and literature. *Prog. Neurobiol.* 77 (1–2), 90–127. doi:10.1016/j.pneurobio.2005.10.001

Miao, J. J., Chen, G. Y., Du, K., and Fang, Z. J. (2012). Towards big data to improve availability of massive database. *Appl. Mech. Mater.* 263–266, 3326–3329. doi:10.4028/www.scientific.net/AMM.263-266.3326

Nesvizhskii, A. I. (2014). Proteogenomics: Concepts, applications and computational strategies. *Nat. Methods* 11 (11), 1114–1125. doi:10.1038/nmeth.3144

Ouzounis, C. A., and Karp, P. D. (2002). The past, present and future of genome-wide re-annotation. *Genome Biol.* 3 (2), COMMENT2001. doi:10.1186/gb-2002-3-2-comment2001

Perez-Riverol, Y., Bai, J., Bandla, C., Garcia-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., et al. (2022). The PRIDE database resources in 2022: A hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* 50 (D1), D543–D552. doi:10.1093/nar/gkab1038

Sheynkman, G. M., Johnson, J. E., Jagtap, P. D., Shortreed, M. R., Onsongo, G., Frey, B. L., et al. (2014). Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC Genomics* 15 (1), 703. doi:10.1186/1471-2164-15-703

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19, ii215–ii225. doi:10.1093/bioinformatics/btg1080

Stein, L. (2001). Genome annotation: From sequence to biology. *Nat. Rev. Genet.* 2 (7), 493–503. doi:10.1038/35080529

Steinbiss, S., Silva-Franco, F., Brunk, B., Foth, B., Hertz-Fowler, C., Berriman, M., et al. (2016). *Companion*: A web server for annotation and analysis of parasite genomes. *Nucleic Acids Res.* 44 (W1), W29–W34. doi:10.1093/nar/gkw292

Stoco, P. H., Wagner, G., Talavera-Lopez, C., Gerber, A., Zaha, A., Thompson, C. E., et al. (2014). 'Genome of the avirulent human-infective trypanosome — *Trypanosoma rangeli*', *PLoS neglected tropical diseases*. *PLoS Negl. Trop. Dis.* 8 (9), e3176. doi:10.1371/journal.pntd.0003176

Talavera-López, C., Messenger, L. A., Lewis, M. D., Yeo, M., Reis-Cunha, J. L., Matos, G. M., et al. (2021). Repeat-driven generation of antigenic diversity in a major human pathogen, *Trypanosoma cruzi*. *Front. Cell. Infect. Microbiol.* 11, 614665. doi:10.3389/fcimb.2021.614665

The, M., MacCoss, M. J., Noble, W. S., and Kall, L. (2016). Fast and accurate protein false discovery rates on large-scale proteomics data sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* 27 (11), 1719–1727. doi:10.1007/s13361-016-1460-7

The UniProt Consortium (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49 (D1), D480–D489. doi:10.1093/nar/gkaa1100

Törönen, P., and Holm, L. (2022). Pannzer — a practical tool for protein function prediction. *Protein Sci.* 31 (1), 118–128. doi:10.1002/pro.4193

Vaudel, M., Verheggen, K., Csordas, A., Raeder, H., Berven, F. S., Martens, L., et al. (2016). Exploring the potential of public proteomics data. *PROTEOMICS* 16 (2), 214–225. doi:10.1002/pmic.201500295

Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Rios, D., et al. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 32 (3), 223–226. doi:10.1038/nbt.2839

Vlasova, A., Hermoso Pulido, T., Camara, F., Ponomarenko, J., and Guigo, R. (2021). FA-Nf: A functional annotation pipeline for proteins from non-model organisms implemented in nextflow. *Genes* 12 (10), 1645. doi:10.3390/genes12101645

Wang, B., Kumar, V., Olson, A., and Ware, D. (2019). Reviving the transcriptome studies: An insight into the emergence of single-molecule transcriptome sequencing. *Front. Genet.* 10, 384. doi:10.3389/fgene.2019.00384

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13 (5), 329–342. doi:10.1038/nrg3174

frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 24 November 2022
DOI 10.3389/fgene.2022.1017340

Applications of machine learning in metabolomics: Disease modeling and classification

Aya Galal[1,2†], Marwa Talal[1,3†] and Ahmed Moustafa[1,3,4*]

[1]Systems Genomics Laboratory, American University in Cairo, New Cairo, Egypt, [2]Institute of Global Health and Human Ecology, American University in Cairo, New Cairo, Egypt, [3]Biotechnology Graduate Program, American University in Cairo, New Cairo, Egypt, [4]Department of Biology, American University in Cairo, New Cairo, Egypt

Check for updates

OPEN ACCESS

EDITED BY
Mehdi Pirooznia,
Johnson & Johnson, United States

REVIEWED BY
Marco Vanoni,
University of Milano-Bicocca, Italy
Jagadheshwar Balan,
Mayo Clinic, United States

*CORRESPONDENCE
Ahmed Moustafa,
amoustafa@aucegypt.edu

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 11 August 2022
ACCEPTED 07 November 2022
PUBLISHED 24 November 2022

CITATION
Galal A, Talal M and Moustafa A (2022), Applications of machine learning in metabolomics: Disease modeling and classification.
Front. Genet. 13:1017340.
doi: 10.3389/fgene.2022.1017340

COPYRIGHT
© 2022 Galal, Talal and Moustafa. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Metabolomics research has recently gained popularity because it enables the study of biological traits at the biochemical level and, as a result, can directly reveal what occurs in a cell or a tissue based on health or disease status, complementing other omics such as genomics and transcriptomics. Like other high-throughput biological experiments, metabolomics produces vast volumes of complex data. The application of machine learning (ML) to analyze data, recognize patterns, and build models is expanding across multiple fields. In the same way, ML methods are utilized for the classification, regression, or clustering of highly complex metabolomic data. This review discusses how disease modeling and diagnosis can be enhanced via deep and comprehensive metabolomic profiling using ML. We discuss the general layout of a metabolic workflow and the fundamental ML techniques used to analyze metabolomic data, including support vector machines (SVM), decision trees, random forests (RF), neural networks (NN), and deep learning (DL). Finally, we present the advantages and disadvantages of various ML methods and provide suggestions for different metabolic data analysis scenarios.

KEYWORDS

metabolomics, machine learning, metabolic disorders, biomarkers, deep learning

Introduction

Metabolomics is the study of small metabolites or chemical processes involving small substrates in tissues or organisms. The metabolome is the representation of all metabolites in any biological cell, tissue, or organ and their subsequent cellular products. It provides a snapshot of the physiology of the cell under investigation and can be used to study biological information on the biochemical level. This provides an avenue of study that leads to understanding the biological phenotype, which can be used in the context of health and disease (Gowda et al., 2008). Roger Williams introduced the concept of a metabolic profile in the late 1940s (Gates and Sweeley 1978). He used paper chromatography to suggest that schizophrenia presents characteristic metabolic patterns in urine and saliva. Only with the technological advancements of the 1970s and with the introduction of gas chromatography and mass spectrometry was the term "metabolic profile" introduced (Griffiths and Wang 2009). The first comprehensive metabolomic tandem mass spectrometry database, Metabolite and Chemical Entity Database (METLIN), was developed in 2005 by the Scripps Research Institute (Smith et al., 2005; Guijas et al., 2018). In 2007, "The Human Metabolome Project," led by David S. Wishart, established the first draft of a database with ~2,500 metabolites, ~1,200 drugs, and ~3,500 food components (David S. Wishart et al., 2007; Wishart et al.,. 2009). Now, techniques such as mass spectrometry and gas chromatography have advanced so that they can detect thousands of independent features in a single specimen, making identifying metabolites associated with a disease or trait an increasingly difficult computational challenge. The field of metabolomics has enabled a comprehensive assessment of biological specimens and their associated compounds. This improved understanding of the biological system at the molecular level is crucial in aiding disease diagnosis and therapeutic development (Gowda et al., 2008). Within the omics field, metabolomics represents the underlying layer that reflects all information expressed and modulated by the upstream genetic regulation and processing layers. It is the closest link to the phenotype. It is at the forefront of personalized health, in terms of diagnosis and therapy, through its direct applicability to the area of biomarker discovery (Shah, Sureshkumar, and Shewade 2015; Aderemi et al., 2021). Biological systems are complex and often require the integration of several layers of omic data to decipher. Metabolomics is a potential solution for this, as it represents the product of the interaction between the various omic layers (Hasin, Seldin, and Lusis 2017; Misra et al., 2018).

Metabolic disorders are biochemical aberrations that can be detected through screening techniques or biomarker identification. However, biomarker identification requires extensive prior knowledge and numerous disease models for a single biomarker to be successfully linked to a disease. Metabolomics and other "omics" molecular profiling techniques provide essential tools for discovering new disease risk factors and biomarkers (Smith et al., 2005; Gowda et al., 2008) without the typical hurdles of time and money. The most studied metabolic disorders include diabetes mellitus (DM) (Friedrich 2012; Guasch-Ferré et al., 2016; Ahola-Olli et al., 2019; Sun et al., 2019; Hou, Wang, and Pan 2021), cardiovascular disease (CVD) (Müller et al., 2021; Iida, Harada, and Takebayashi 2019; Streese et al., 2021; McGranaghan et al., 2020; Cavus et al., 2019; Ruiz-Canela et al., 2017), and cancers (Gowda et al., 2008; Raffone et al., 2020; Yang et al., 2020; Schmidt et al., 2021).

For the purposes of this review, the main metabolomic experimental workflow can be divided into four main parts: 1) sample retrieval and preparation, 2) separation and detection of metabolites, 3) data processing, including data mining and extraction, and 4) data analysis (Figure 1, Middle Panel). Sample retrieval and preparation depend on the type of material to collect. Metabolites can be measured from a variety of different biological samples, e.g., tissue, biofluids, and cell culture. Depending on the disease or trait under investigation, the choice of specimen differs, as do the steps required to prepare the sample for the corresponding experiment. For example, tissue specimens should be immediately quenched with liquid nitrogen after harvesting to arrest the metabolism. Numerous sample preparation protocols entailing the details of metabolite extraction, enrichment, and depletion of proteins have been developed (Dettmer, Aronov, and Hammock 2007; D. S. Wishart 2005; Want et al., 2007). Separation and detection of metabolites can be achieved by two main protocols: nuclear magnetic resonance (NMR) and mass spectrometry and their assorted subtypes (Gowda et al., 2008). Both techniques are capable of high-throughput measurements of a large number of metabolites.

Metabolomics studies can be subclassified into three major approaches: targeted analysis (Shulaev 2006; Griffiths and Wang 2009; Mookherjee et al., 2020), metabolite profiling, i.e., untargeted analysis (Fiehn 2002; Halket et al., 2005), and metabolic fingerprinting, which is also known as exometabolomics and focuses on extracellular metabolites while utilizing analytical profiling approaches (Allen et al., 2003; Mapelli, Olsson, and Nielsen 2008; Silva and Northen 2015; Thomas et al., 2021). Targeted approaches are limited to a set of predetermined metabolites of interest for identifying and quantifying these specific metabolites. Untargeted approaches are conducted to identify a comprehensive metabolic profile in a specimen. The choice of metabolomics workflow and the associated downstream steps depends on the choice of experimental approach (Newgard 2017). Typically, untargeted metabolomics experiments generate substantial volumes of complex data requiring specialized computational processing and interpretation methods. Data interpretation software should ideally be capable of background noise elimination, peak identification and alignment, and peak normalization. While commercial and public domain software packages attempt to perform some of these tasks, there is no universal software for data extraction and analysis software. In metabolomics, hundreds of metabolites are detected and routinely analyzed. The complexity and magnitude of data produced from metabolomic studies necessitate the use of computational methods to analyze the data and elicit potential trends.

Artificial Intelligence (AI), both as a concept and research field, has gained attention across the twenty-first century. With its various applications in understanding the structures or trends in vast amounts of data collected or generated from modern high-throughput experiments, AI and machine learning (ML) offer countless possibilities. ML is used to develop models that can tackle large-scale data and, through learning, can solve complex problems. ML algorithms are fundamentally based on the ability to build mathematical models from a group of sample data (Dhall, Kaur, and Juneja

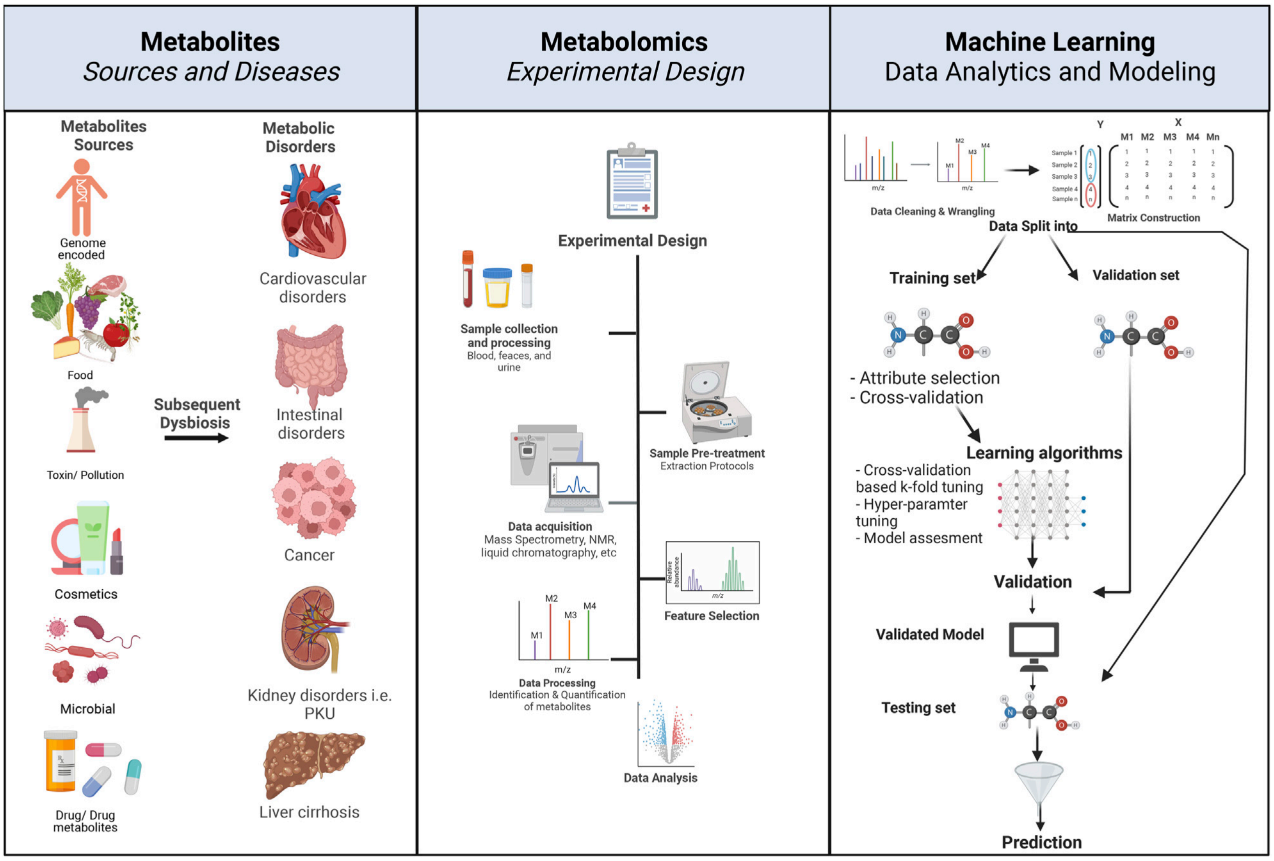


FIGURE 1
Principles of metabolomics experimental design and associated ML workflow. The left panel describes the various sources of metabolites. Metabolite exposure can be through endogenous and exogenous means, e.g., human-encoded, microbiome-encoded, food, drugs, and toxins. Metabolic dysbiosis can be associated with metabolic disorders, e.g., cancer, cardiovascular disease, intestinal disorders, and diabetes. The center panel describes the typical flow and design of a metabolomic experiment, starting with the 1) study design where disease and control groups are determined, 2) followed by sample selection, e.g., urine, stool, blood, and serum, 3) collected samples undergo pre-treatment and processing according to experimental design, 4) data acquisition, e.g., through mass spectrometry or NMR, 5) feature selection involves the identification of desired metabolite features that will undergo subsequent, 6) data processing through the quantification of metabolites, and finally, 7) data analysis depends on the study design. The right panel describes the concepts of ML workflow and prediction, starting with 1) data wrangling and cleaning, 2) matrix construction, where data from each metabolite is placed in a matrix in reference to the conditions, i.e., disease (marked in red), control (marked in blue), 3) data are then divided into testing, validation and training datasets, 4) ML algorithm is applied, and 5) cross-validation, and testing of the predictive power of the algorithm on a test dataset. Created with BioRender.com.

2020). Typically, a dataset used for developing a machine learning model is divided into a training subset, for example, comprised ~70% of the available data and used in the ML algorithm to build a model and make predictions, and a testing subset, for example, ~30% of the data used to provide an unbiased evaluation of the final model from the training step. Often, an intermediate validation step is added to assist in determining the most accurate model and obtaining optimal model hyper-parameters. In this instance, the data can be divided through a 60-20-20 split, where 20% of the data can act as an additional validation set. The initial learning process requires extensive data to allow the ML algorithm more opportunities to learn and improve the model. The ability of the algorithm to learn is formally through a mathematical function that maps specific inputs to certain outputs. The training dataset is used to guide the algorithms to make predictions without being explicitly programmed. This is achieved through a series of operations, where learning is made on the basis of weights and biases that will eventually make predictions in a finite number of steps (Cohen 2021). Having experienced the training dataset where the algorithm was able to learn and build a general model, the next step is testing the model's performance on an independent dataset that contains previously unseen data and producing sufficiently accurate predictions. Predictions are based on the algorithm's ability to assign each input to the chosen statistical representation defined by the user. The better the algorithm can learn from the input data provided, the more accurate the algorithm can produce predictions (Antonakoudis et al., 2020; Deepthi et al., 2020).

Constructing an ML model requires a series of steps: 1) Defining the training dataset: it involves identifying the type of data to be used as the training dataset; the input data would change depending on the problem that needs to be addressed. 2) Gathering the training dataset: a representation of the real-world use requires a set of inputs that will address the problem under investigation. 3) Input feature representation: The learned function's accuracy strongly depends on how the input object is represented. Input objects are transformed into feature vectors, which have several descriptive features. The number of features must be sufficient to contain enough information to predict the output accurately and not too large to affect the dimensionality. 4) Determining the type of algorithm to be used: this is the algorithm that will be used to fit the data during the testing/training phase into a model. The choice of the algorithm depends on several factors, including the question the analysis is trying to answer, the data, and the ML category used, i.e., supervised learning, unsupervised learning, reinforcement learning, and semi-supervised learning (it is expanded upon later). 5) Training the algorithm: running the algorithm on the gathered training dataset; this step might require additional user input depending on the choice of the algorithm. Cross-validation can be used to adjust hyper-parameters (variables that determine how the algorithm is trained, e.g., learning rate, number of branches, clusters, and epochs) and optimize performance on a subset of the training set. 6) Validation: the training phase is often followed by a validation step to fine-tune the hyper-parameters of the classifiers. This validation step is independent of the cross-validation performed on the training set and uses a separate validation dataset. Validation is typically necessary to compare the performance of the different candidate classifiers: it is used to obtain performance parameters, including accuracy, sensitivity, and specificity of the models, and to estimate the models' prediction error or bias. The model with the best performance on the validation set is then chosen to move forward to the testing phase. 7) Testing and evaluation: after hyper-parameter adjustments and learning, the accuracy of the learned function is assessed through the performance of the algorithm on an entirely new testing dataset, independent of the training and validation dataset (Figure 1: Principles of Metabolomics experimental design and associated ML workflow.).

Model performance assessment is an important step in properly evaluating the validity of a model's predictions and deciding which model is best for a given problem. Model assessment methods are varied, depend on the characteristics of the problem, and can include a process known as hyper-parameter tuning, where they can be used to control the learning process of the model. The most commonly used assessment methods for classification problems are accuracy (Gajda and Chlebus 2022), cross-entropy (Boubezoul, Paris, and Ouladsine 2008; Gajda and Chlebus 2022), area under the curve (AUC) (Airola et al., 2011; Yala et al., 2019; Gajda and Chlebus 2022), while for regression analysis, mean squared error (Bellet, Habrard, and Sebban 2013), mean absolute error (Airola et al., 2011; Bellet, Habrard, and Sebban 2013) and root mean squared error (Nguyen et al., 2019) are more commonly employed. However, other performance metrics are available, including variance and R^2 coefficient, to name a few. Determining model specificity (the ability of a model to identify true negatives correctly) and sensitivity (the ability of a model to correctly identify true positives) (Trevethan 2017) are additional methods that can inform researchers and apply some context to the data under investigation (Parikh et al., 2008; Trevethan 2017).

With every newly discovered metabolite, the field of metabolomics has grown, allowing for a comprehensive assessment of biological specimens and their associated compounds. This improved understanding of the human body at the molecular and biochemical levels is crucial in aiding disease diagnosis and therapeutic development (Gowda et al., 2008). Over the years, the significant contribution of AI and associated applications in various biomedical fields has grown, demonstrating the application of ML in disease prediction and diagnosis of multiple diseases, including cardiovascular disorders, cancer, and rare genetic diseases.

In 2019, an editorial published in *Nature* titled "Why the metabolism field risks missing out on the AI revolution" expressed concern with the lack of momentum in AI-assisted applications in the field of metabolic research as opposed to other areas, such as genetics, for example. The curation of high-quality datasets, as well as the collective efforts of various institutions and funding bodies over the past few years, has increased the number of AI-assisted metabolomics studies. The number of metabolomic publications with AI and ML-based methods has been consistently on the rise, with very few publications (~1/year) in the early 2000s, steadily rising to reach ~150 publications in 2021, and the number of ML-assisted publications in 2022 promising to surpass this. The most used ML methods in metabolomic studies in the past years are RF, SVM, logistic regression, and, more recently, DL (Figure 2).

The integration of metabolomics with analytical ML techniques can be used to answer questions that other omics approaches cannot answer alone (Gowda et al., 2008; Graham et al., 2018; Turi et al., 2018; Jendoubi 2021). Here, we discuss major ML approaches for analyzing metabolic profiles, focusing on biomarker discovery and disease diagnosis.

Types of machine learning algorithms

ML algorithms can be used to analyze ever-increasing amounts of generated and accumulated data. ML

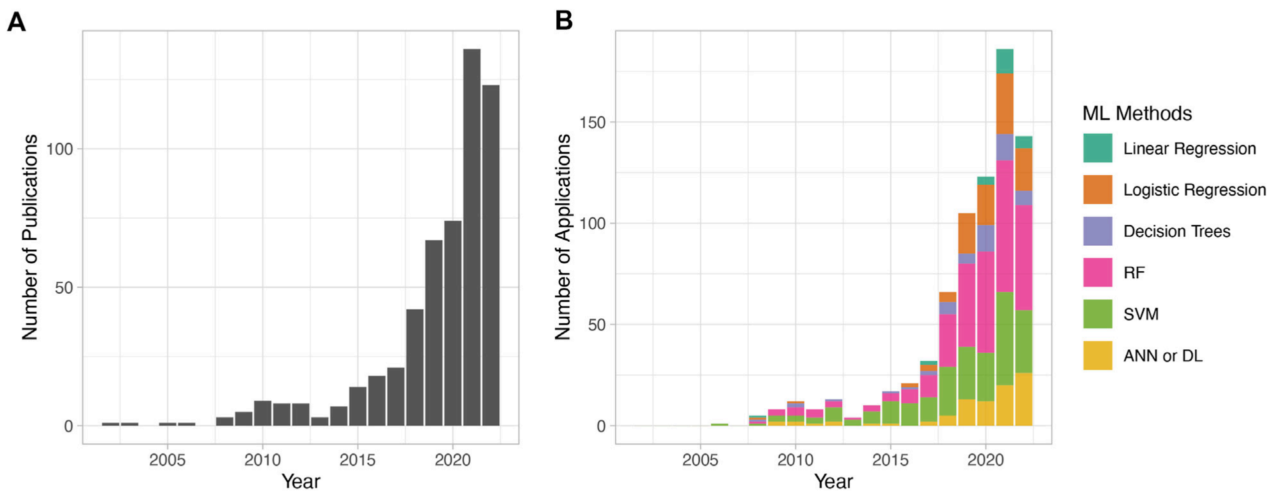


FIGURE 2
Metabolomic publications using machine learning in data analytics over the past 2 decades. PubMed was searched using the keywords "metabolomics" and "machine learning" from 2002 to 2022. Results were manually filtered to remove review articles and irrelevant publications. The counted publications include studies that use any of the mentioned ML algorithms in the context of metabolomic analysis, including classification problems, biomarker discovery, peak identification, metabolomic data analysis tools, and others. Only ML algorithms employed for disease model building are considered. **(A)** The total number of publications per year. **(B)** The number of publications using ML methods per year. The *y*-axis in **(A)** and **(B)** are different because in **(B)**, it indicates only the ML methods discussed in this review. The total number of publications across panels **(A)** and **(B)** varies because publications often utilize multiple ML algorithms.

algorithms are traditionally divided into supervised, unsupervised, semi-supervised, and reinforcement learning (Figure 3). For the purposes of this review, we focus on ML algorithms used in metabolomic studies, mainly supervised and unsupervised algorithms. The algorithms highlighted in the following sections do not exclusively belong to any of the mentioned ML categories; rather, the same algorithms can be used for multiple learning categories (e.g., *k*-Nearest Neighbor can be used in supervised and unsupervised learning).

Machine learning categories

Supervised learning

Supervised learning involves inferring a function from a labeled dataset input and a specific expected result (output), i.e., an input-output pair. With data containing continuous values, linear regression analysis is commonly used for objectives such as forecasting, prediction, and process optimization (Biswas, Saran, and Wilson 2021). Logistic regression is used with the classification into two categories. Classification for more than two categories can be performed using Support Vector Machines (SVM), decision trees, Random Forest (RF), and other methods (refer to Figure 3).

Unsupervised learning

In unsupervised machine learning, the algorithm learns patterns from unlabeled data. The algorithm takes a dataset with only inputs and attempts to find a structure in the data by grouping or clustering the data points (Dhall, Kaur, and Juneja 2020). Unlike supervised learning, where the algorithm learns from data that has been labeled, classified, or categorized, unsupervised algorithms identify trends or commonalities in the data and respond based on the presence or absence of these commonalities in the data. This analysis can have various goals, from identifying hidden data trends to reducing redundancy, i.e., dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville 2018) and t-distributed stochastic neighbor embedding (t-SNE) (Van Maaten and Hinton 2008), or grouping together similar data (Dhall, Kaur, and Juneja 2020), i.e., clustering. Examples of unsupervised algorithms include *k*-means clustering, hierarchical clustering, anomaly detection, neural networks (NN), principal component analysis (PCA), independent component analysis (ICA), and *apriori* algorithms.

Semi-supervised learning

Semi-supervised learning falls between unsupervised and supervised learning. It combines a small amount of labeled data with a large amount of unlabeled data during the

10.3389/fgene.2022.1017340

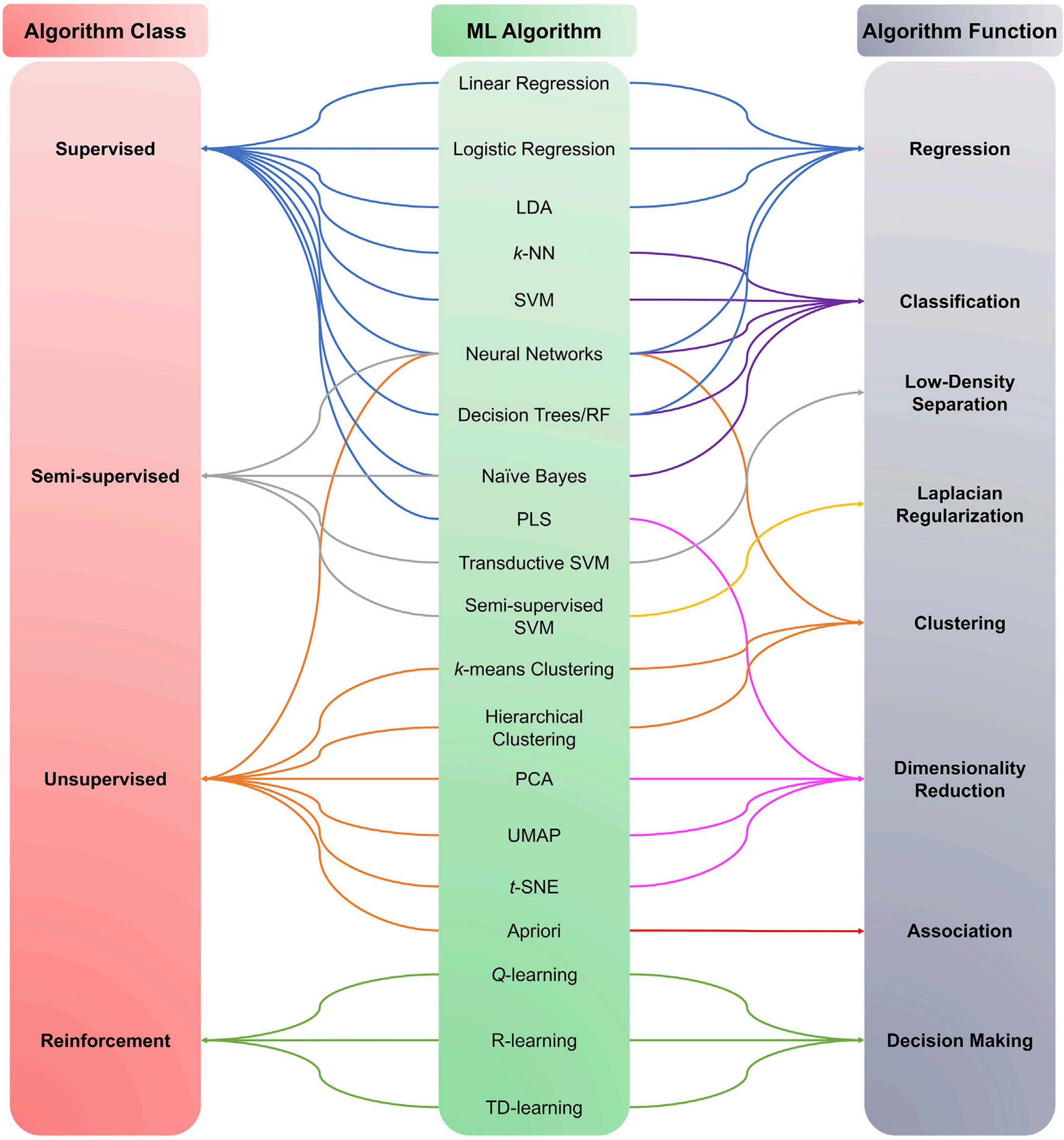
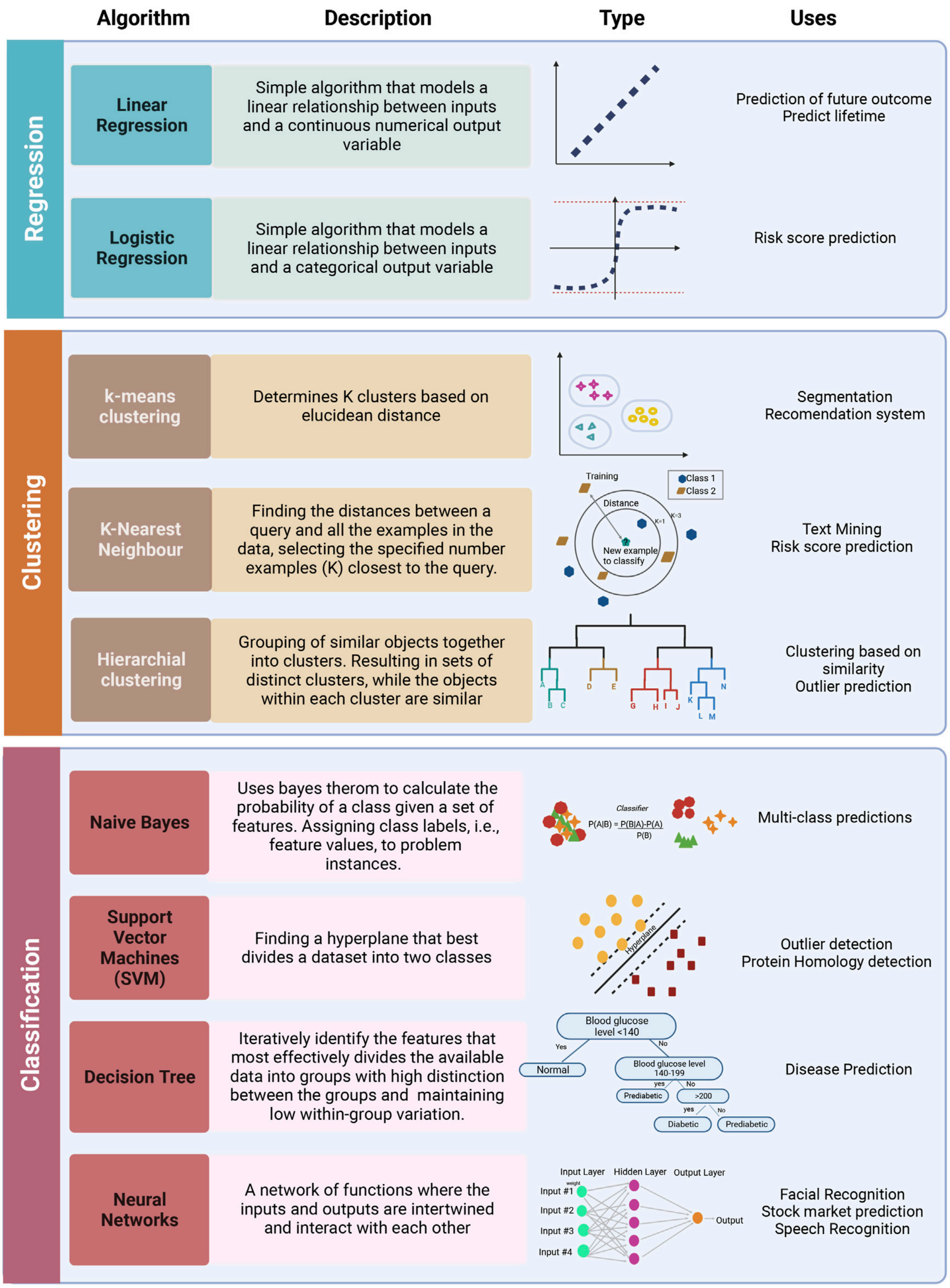


FIGURE 3
Machine learning algorithms categories. ML algorithms are divided into four main classes: Supervised, Unsupervised, Semi-supervised, and Reinforcement learning. The category choice depends on the type and nature of the data under investigation, i.e., labeled or unlabelled data.

training process and uses context to identify data patterns (Dhall, Kaur, and Juneja 2020). For example, this method can be used for classification problems that require a supervised learning algorithm to achieve the end goal; however, it would not require extensive labeling. It is faster than supervised learning because it involves a mixture of labeled and unlabeled data. Examples include generative models, low-density separation, Laplacian regularization, and heuristic approaches. This approach is not commonly used in the field of metabolomics, with few published applications (Libbrecht and Stafford Noble 2015; Migdadi et al., 2021; Abram and Douglas, 2022; Iqbal et al., 2022).

Reinforcement learning

This method was adopted to direct unsupervised ML by rewarding desired behavior and punishing undesired ones. Positive feedback strengthens the model's ability to connect target inputs and outputs (Dhall, Kaur, and Juneja 2020).



	Algorithm	Description	Type	Uses
Regression	Linear Regression	Simple algorithm that models a linear relationship between inputs and a continuous numerical output variable		Prediction of future outcome Predict lifetime
	Logistic Regression	Simple algorithm that models a linear relationship between inputs and a categorical output variable		Risk score prediction
Clustering	k-means clustering	Determines K clusters based on elucidean distance		Segmentation Recomendation system
	K-Nearest Neighbour	Finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query.		Text Mining Risk score prediction
	Hierarchial clustering	Grouping of similar objects together into clusters. Resulting in sets of distinct clusters, while the objects within each cluster are similar		Clustering based on similarity Outlier prediction
Classification	Naive Bayes	Uses bayes therom to calculate the probability of a class given a set of features. Assigning class labels, i.e., feature values, to problem instances.		Multi-class predictions
	Support Vector Machines (SVM)	Finding a hyperplane that best divides a dataset into two classes		Outlier detection Protein Homology detection
	Decision Tree	Iteratively identify the features that most effectively divides the available data into groups with high distinction between the groups and maintaining low within-group variation.		Disease Prediction
	Neural Networks	A network of functions where the inputs and outputs are intertwined and interact with each other		Facial Recognition Stock market prediction Speech Recognition

FIGURE 4
Representation of most commonly used ML algorithms with functional categorization accompanied by graphical representations of each algorithm and some potential applications. The most frequently used algorithms can be grouped into regression (linear and logistic), clustering (k-means, k-NN, hierarchical clustering, NN), and classification (Naive Bayes, SVM, Decision trees). Created with BioRender.com.

TABLE 1 Key applied machine learning algorithms.

Algorithm	Description
Linear regression	A linear approach to model a relationship between dependent and independent variables (Schneider, Hommel, and Blettner 2010)
Logistic Regression	Models the probability of an event ocuring out of two alternatives by defining the logarithmic odds of the event is a linear combination of independent variables (Stoltzfus 2011)
k-means clustering	Partitions data into groups of similar kinds of items by finding the similarity between the items using euclidean distance. (MacQueen 1967)
Partial Least Squares (PLS)	Reduces the dimensionality of correlated variables to a smaller set of variables that can then be used as predictors. Used when there is a high number of colinear variables. (Garthwaite 1994)
Linear Discriminant Analysis (LDA)	Finds a linear combination of features that can separate two or more object classes. Uses a generalization of Fischer's linear discriminant. (Riffenburgh 1957)
Boosting algorithms	Involves training a sequence of weak models, where each model compensates for the weakness of its predecessors. Thereby improving the overall predictive ability of the model. (Kearns and Valiant 1989)
Support Vector Machines (SVMs)	Based on finding a hyperplane that best divides a dataset into two classes (Boser, Guyon, and Vapnik 1992; Ben-Hur et al., 2002)
Naïve Bayes	Assigns class labels, i.e., feature values to problem instances (Bzdok, Altman, and Krzywinski 2018; Hastie et al., 2009).
k-Nearest Neighbors (*k*-NN)	Finding the distances between a query and all similar examples in the dataset, selecting the specified number of examples (K) closest to the query, when used for classification, the most frequent labels are counted and when used for regression, the labels are averaged (Altman 1992)
Decision Trees	Uses a tree-like model of decisions and consequences to predict the value of a target variable by learning simple decision rules from available data features (Shalev-Shwartz and Ben-David 2014)
Random Forest (RF)	Builds on the concept of multiple decision trees and takes the majority for classification and the average for regression (Hastie et al., 2009; Ho 1995)
Principal Component Analysis (PCA)	A dimensionality reduction technique that projects data onto a subspace of lower dimension that is able to retain the most variance among the data points. (Wold, Esbensen, and Geladi 1987; Jolliffe 2005)
Neural Networks (NN)	A network of functions where the inputs and outputs are intertwined and interact with each other (Hinton and Salakhutdinov 2006)

Examples include Monte Carlo, Q-learning, State–action–reward–state–action (SARSA), Q-learning Lambda, SARSA-Lambda, and Deep Q-Learning (DQN), to name a few. Reinforcement learning is often converged around fields such as game theory, operations research, and swarm intelligence, as they are highly dependent on using robotics.

On the functional level, different ML algorithms are mainly geared toward solving regression, clustering, or classification problems. A representation of different ML algorithms with functional categorization is depicted in Figure 4, and brief descriptions of the most commonly used ML algorithms are indicated in Table 1. Supervised ML algorithms are by far the most commonly used in the field of metabolomics. For this review, six algorithms centering around supervised learning are highlighted in the following section, and the application of these algorithms to metabolomic data will be expanded upon.

Machine learning algorithms

Regression analysis

Regression analysis is a group of statistical procedures used to estimate the relationship between a dependent variable (outcome or response) and one or more independent variables (predictors, covariates, or features). This method of statistical analysis progressed from the least-squares method to the regression. It can be used in a variety of fields. In order to interpret the output in real-world relationships, a number of assumptions are made, such as that the sample is representative of the entire population and that no errors occurred when measuring the independent variables (Vetter and Schober 2018). Regression analysis is used for two distinct purposes: inferring causal relations between the variables under investigation and prediction (Baumgartner, Böhm, and Baumgartner 2005).

Linear regression

Linear regression models the relationship between dependent and independent variables by fitting a straight line (linear equation) to the observed data (Schneider, Hommel, and Blettner 2010). Predictions based on linear regression are simple: data trend is observed, then a prediction is made on the basis of that trend (Casson and Farmer 2014; Vetter and Schober 2018). While not all data follow a linear trend, linear regression is often the first attempt used to understand data and for predictive analyses.

Logistic regression

A statistical model used to predict a binary outcome (one scenario out of two possible alternatives) based on a set of independent variables (those that influence the outcome) using a logarithmic odds scale (Stoltzfus 2011). Typically, logistic regression analysis is used for classification purposes and when dealing with binary outcomes i.e., two categories.

Decision trees

A statistical decision support tool that uses a tree-like model of decisions and possible consequences. Each tree is similar in structure to that of a flowchart. Each node represents a test, e.g., taking a vitamin, each subsequent branch represents the outcome of the test, i.e., "yes" or "no" for taking the vitamin, and each leaf node represents a class label (Shalev-Shwartz and Ben-David 2014; Kamiński, Jakubczyk, and Szufel 2018). Decision trees consist of three types of nodes: decision, chance, and end nodes (Kamiński, Jakubczyk, and Szufel 2018). Decision trees are constructed to iteratively identify the feature that most effectively divides the available data into groups with a high distinction between the groups in terms of outcome while maintaining a low within-group variation.

Random forest (RF)

A statistical classification method composed of an *assembly* of many decision trees constructed during the training phase. Generally outperforming decision trees as they correct the observed overfitting. New objects are classified based on the attributes of the data. Each tree is classified and gives a vote for the chosen attribute. When used for classification, the classification with the most votes is chosen, and when used for regression purposes, the average votes are used (Hastie et al., 2009; Dhall, Kaur, and Juneja 2020). RF models are among the most frequently used algorithms for prediction or classification purposes, with various omics applications from understanding the human gut microbiome, differentiating between healthy and disease metabolome, investigating the pregnancy metabolome, cancer diagnosis to the more recent COVID-19 diagnosis and classification of COVID-19 severity. Key studies using these ML algorithms for metabolomic understanding will be highlighted later.

Support vector machines (SVM)

Proposed in 1992 by Boser, Guyon, and Vapnik, SVMs (Boser, Guyon, and Vapnik 1992) has been popular classification tools in many fields, including bioinformatics and biological data analysis in general (Saeys, Inza, and Larrañaga 2007). SVMs split training observations into two classes by constructing a hyperplane, a decision boundary that separates the data points into two classes. The distance between the hyperplane and the nearest data points of each class is called the margin, and the points onto which this margin hits are called the "support vectors". The SVM is constructed so that the margin on either side of the hyperplane is maximized (Figure 5) (Vapnik 2006). In many cases, the data points cannot be fully segregated. Here, the SVM will try finding a "soft margin" that allows the misclassification of a few points while minimizing the cost of the training points that are on the wrong side of the classification boundary (Cortes and Vapnik 1995).

In the case of data that are not linearly separable, the data points are mapped into a higher dimensional feature space in which they become linearly separable (Cortes and Vapnik 1995) (Figure 6). This is known as the "kernel trick" and gives SVMs major advantage over other statistical multivariate methods, such as PCA, Partial Least Squares (PLS), and Orthogonal Projections to Latent Structures (OPLS), which cannot be applied to nonlinear cases. A variety of different kernel functions can be employed to transform the data, including the linear kernel, polynomial kernel, sigmoid kernel, and Gaussian radial basis function (RBF) kernel (Powell 1987), (Broomhead and Lowe 1988). A major drawback of SVM is that it is natively restricted to binary classification problems, i.e., it can only discriminate between two classes. However, it does not scale well with large datasets because of its computational complexity.

It is often beneficial to perform feature selection before training multivariate algorithms, such as SVMs, by only selecting subsets of features, in the case of metabolites, on which supervised learning is employed (Guyon 2003). Reducing the number of variables used for model building can simplify the interpretability of the data analysis results and prevent overfitting, which is often caused by non-informative input features (Liu and Motoda 2012). Feature selection methods have been reviewed elsewhere (Miao and Niu 2016). Popular feature selection methods used with SVM models in metabolomics studies include recursive feature elimination (RFE) (Guan et al., 2009; R. Shen et al., 2018), L1 norm SVM (Zhou et al., 2010) (Guan et al., 2009) (Zhou et al., 2010) and variable importance in projection (VIP) ((Zhang et al., 2018), (Cheng et al., 2019) (Z. Chen et al., 2021)).

Deep learning (DL)

Deep learning (LeCun, Bengio, and Hinton 2015) has risen to prominence as the most popular type of machine learning algorithm recently. It uses artificial neural networks (ANN) to construct complex relationships relating input variables to the outcome, advancing classifier performance beyond typical machine learning techniques, particularly in

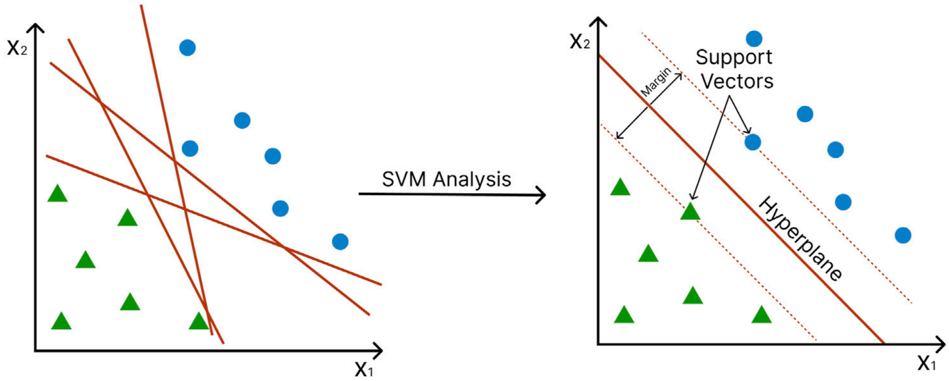


FIGURE 5
Support Vector Machines (SVM) construct a hyperplane to separate data into two classes. Axes represent different features. Green triangles and blue circles represent different conditions (e.g., disease vs. control). The margin (red dotted line) is the distance between the hyperplane and the support vectors (the nearest data point of each class).

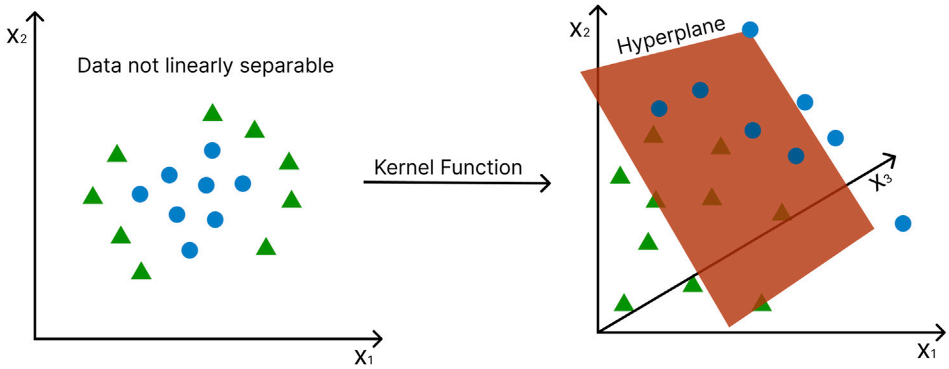


FIGURE 6
The "kernel trick" - non-linearly separable data points are mapped into a higher dimensional feature space in which they become linearly separable. Axes represent different features. Green triangles and blue circles represent different conditions (e.g., disease vs. control). The hyperplane, in this case, becomes a two-dimensional plane.

circumstances involving large-scale datasets with high dimensionality. The potential of deep learning is endless; however, it is an intensive process that requires considerable computational power, and its results are often difficult to interpret. In the case of metabolomics studies, it is difficult to evaluate from the model, which features affect classification the most. Deep learning's recent success has been fueled by an increase in computing power—particularly the introduction of graphics processing units, or GPUs —, as well as the availability of large-scale data sets to use for training the models.

Although there are applications of unsupervised deep learning, including autoencoders (Rumelhart, Hinton, and Williams 1985; Hinton and Salakhutdinov, 2006; Hinton and Salakhutdinov 2006) and generative adversarial networks (Goodfellow et al., 2014 (Goodfellow et al., 2014)), in this review, we focus on supervised deep learning.

An artificial NN is composed of units, termed neurons, that combine multiple inputs and produce a single output. The network approximates the functions that link inputs (e.g., gene expression levels, metabolite concentrations) to desired outputs (e.g., disease risk). Neurons are organized into several layers, with an input layer, an output layer, and intermediate layers, called "hidden layers" (LeCun, Bengio, and Hinton 2015). The variables from the input layer are multiplied by specific values called weights and fed into the neurons of the first hidden layer. Each neuron takes the input, and applies a nonlinear activation function to it, such

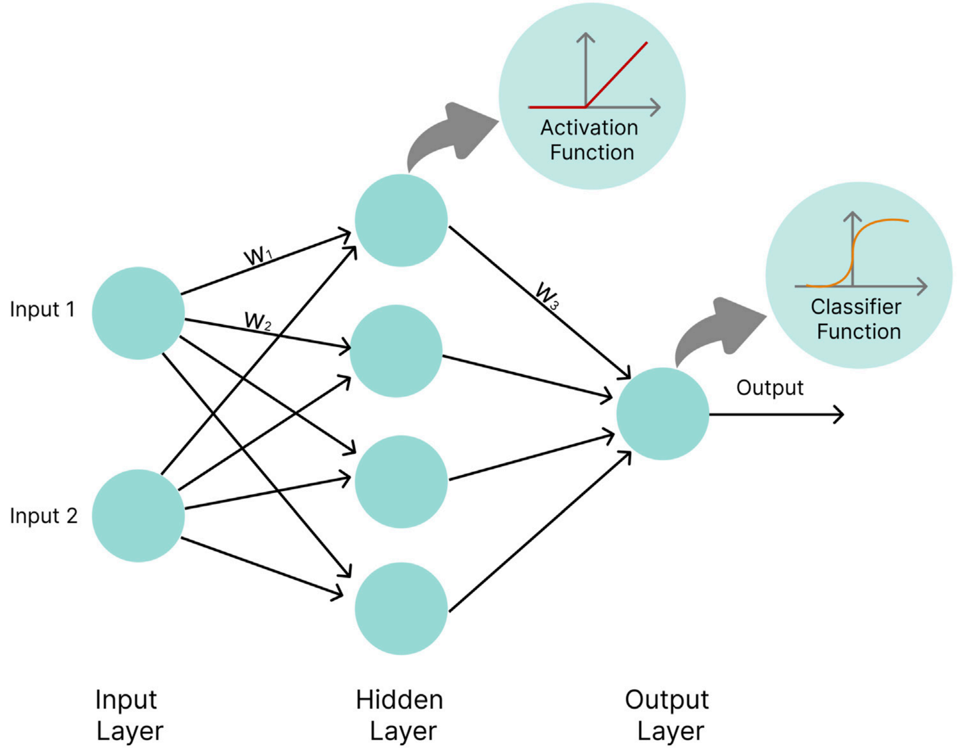


FIGURE 7
Basic neural network architecture. Circles represent neurons. w_1, w_2, and w_3 represent weights by which values calculated inside neurons are multiplied before being passed on to the next layer. In the hidden layer neurons, values are passed into an activation function (e.g., the ReLU function), while the output layer neuron applies a classifier function (e.g., the Softmax function) to input values.

as sigmoid (Narayan 1997) or rectified linear unit (ReLU) (Glorot et al., 2011), and modifies the outcome by adding a bias to it. The output is then passed on to the next hidden layer. Finally, the outputs of the hidden layers are linearly combined in the output layer and often passed through a classifier function, e.g., a Softmax function, to produce an output value. During supervised NN training, the tunable parameters of the network, i.e., the weights and biases, are optimized so that the distance between the network's computed outcome and the experimentally determined outcome is minimized (Figure 7).

Weights and biases are usually randomly initialized in an artificial neural network and then gradually optimized with the aid of a backpropagation algorithm. A cost function (e.g., the sum of squared errors, cross-entropy) computes the difference between the network's output and the desired output. The derivative of the cost function with respect to the weight can be used to evaluate how a slight change in a particular weight affects performance. The parameters of the network are adjusted in a direction that minimizes the cost. This process

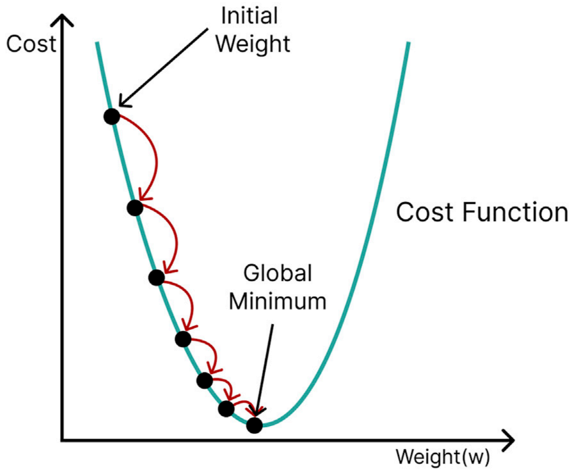


FIGURE 8
Gradient descent; initial network parameters (weights and biases) are adjusted in a direction that travels down the slope of the cost function (green curve) until the minimum is reached.

is termed gradient descent because it travels down the slope of the cost function in steps until, optimally, it reaches its global minimum (Figure 8). However, cost functions are often complicated in reality, with many local minima and saddle points to which gradient descent could converge. Since the slope in these regions is also zero, it is almost impossible to escape them. Stochastic gradient descent (Bottou 2010) offers a more efficient approach, in which only a subset (minibatch) of the training data is selected at random and used for cost minimization. Using different mini-batches for each calculation provides enough stochasticity to avoid getting stuck in local minima and saddle points, in addition to drastically reducing computation time and cost.

An artificial NN is considered 'deep' when it contains more than one hidden layer. It has been shown that a single hidden layer can approximate any function that maps input patterns to output patterns, given that sufficient neurons are employed (Cybenko 1989), (Hornik, Stinchcombe, and White 1989). However, using more hidden layers improves generalization and leads to more accurate modeling (LeCun, Bengio, and Hinton 2015). Some commonly used types of artificial NN include feed-forward NN, recurrent neural networks, convolutional neural networks (CNN), and deep Boltzmann machines. For an excellent review of NN types, refer to Min et al., 2016 (Min, Lee, and Yoon 2016); for potential applications, refer to (Mendez, Broadhurst, and Reinke 2019; Pomyen et al., 2020).

DL has only recently been used in the analysis of omics data, and the application of DL in metabolomics, in specific, is still emerging and comparatively low compared to other omics. Metabolomic studies that use DL algorithms are, therefore, much fewer than those that utuseilize other ML algorithms.

Use of ML approaches in metabolomic studies

Recently, ML techniques have been used for the analysis of metabolomics data from numerous diseases. For the purposes of this review, we are focusing on key studies that used the aforementioned ML approaches in metabolomic investigations, categorized according to the conditions being studied. For examples of metabolomic studies using ML approaches, refer to Table 2 and Supplementary Table S1.

Cancer

Ovarian cancer

In one of the earliest studies, Yu et al. developed an SVM classification model that achieved an average sensitivity of 97.38% and an average specificity of 93.30% for distinguishing cancer from healthy tissue, using a dataset provided by the National Cancer Institute containing serum metabolomic data from ovarian cancer and normal tissue (Yu et al., 2005).

The research group of Guan et al. also extensively studied ovarian cancer metabolites. In 2009, they constructed classifiers using linear and non-linear SVM to diagnose ovarian cancer from serum metabolites with over 90% accuracy, significantly better than a random classifier (Guan et al., 2009).

The same research team published in 2010 (Zhou et al., 2010) how they evaluated a customized fSVM algorithm (SVM for functional data classification (Rossi and Villa 2006)) coupled with ANOVA feature selection for detecting of ovarian cancer using serum metabolites. One of the tested models achieved 100% accuracy in split validation and 98.9% in leave-one-out cross-validation.

In a third study published in 2015 (Gaul et al., 2015), the authors were able to generate a further SVM model capable of identifying early-stage ovarian cancer with 100% accuracy, this time using a panel of sixteen serum metabolites selected by RFE. Eleven of the sixteen metabolites were identified, including phosphatidylinositol, as well as the lysophospholipids lysophosphatidylethanolamine and lysophosphatidylinositol.

Metabolomic analysis has also been found to predict ovarian cancer recurrence. An SVM prediction model was employed by Zhang et al. with ten significant plasma biomarkers, yielding area under the curve (AUC) values reaching 0.964 (Zhang et al., 2018). The results showed a clear clinical advantage over the most commonly used clinical biomarker, CA125, which by contrast, produced an AUC value of only 0.6126.

Breast cancer

An interesting metabolomics study on breast cancer by Henneges et al. focused on modified nucleosides (degradation products of cellular RNA metabolism) and ribosylated metabolites in urine samples (Henneges et al., 2009). From a set of 35 pruned metabolites, 44 pairwise combinations of metabolite features were employed for SVM-based analysis. The sensitivity and specificity of this model were 83.5% and 90.6%, respectively. S-adenosylhomocysteine (SAH) was the most commonly recurring compound in the metabolite pairs, underlining its importance for RNA methylation in cancer pathogenesis.

In another study conducted on breast cancer samples, Alakwaa et al. demonstrated that DL could reliably predict estrogen receptor status (Alakwaa, Chaudhary, and Garmire 2018). The authors used feed-forward networks with a sigmoid activation function and a softmax classifier on a dataset containing 162 metabolites. The predictions were compared to traditional ML methods like RF, SVM, prediction analysis for microarrays (PAMs), generalized boosted models, recursive partitioning and regression trees (RPART), and linear discriminant analysis, with DL models

TABLE 2 Examples of metabolomics studies utilizing ML algorithms.

#	Author	Journal	Publication year	Area of investigation	ML algorithm used	Brief description	Findings	Doi
1	Shen et al	Cell	2020	COVID-19	Random Forest	Identification of severe COVID-19 cases based on molecular signatures of proteins and metabolites	Severity identification was conducted on 18 non-severe and 13 severe patients. Identified 29 important variables (22 proteins, 7 metabolites) - > Incorrect classification of 1 patient Model was tested on an independent cohort of 10 patients - > all severe patients correctly identified except 1	doi: 10.1016/j.cell. 2020.05.032. Epub 2020 May 28. PMID: 32492406; PMCID: PMC7254001
2	Han et al	Nature	2021	Human gut microbiota	Random Forest	Identification of distinct metabolites to differentiate between different taxonomic groups	The model revealed subsets of chemical features that are highly conserved and predictive of taxonomic identification e.g., over-representation of amino acid metabolism	doi: 10.1038/s41586-021-03707-9. Epub 2021 Jul 14. PMID: 34262212; PMCID: PMC8939302
3	Liang et al	Cell	2020	Human pregnancy metabolome	Linear regression	Untargeted metabolomic profiling and identification of metabolic changes in human pregnancy	Detection of many of the previously reported pregnancy-associated metabolite profiles >95% of the pregnancy associated metabolites are previously unreported	doi: 10.1016/j.cell. 2020.05.002. PMID: 32589958; PMCID: PMC7327522
4	Hogan et al	EBioMedicine	2021	Influenza	Gradient boosted decision trees and random forest	Untargeted metabolomics approach for diagnosis of influenza infection	Untargeted metabolomics identified 3,318 ion features for further investigation Described LC/Q-TOF method in conjunction with machine learning model was able to differentiate between influenza samples (pos/neg) with sensitivity and specificity over 0.9	doi: 10.1016/j.ebiom. 2021.103546. Epub 2021 Aug 19. PMID: 34419924; PMCID: PMC8385175

(Continued on following page)

TABLE 2 (*Continued*) Examples of metabolomics studies utilizing ML algorithms.

#	Author	Journal	Publication year	Area of investigation	ML algorithm used	Brief description	Findings	Doi
5	Bifarin et al	J Proteome Res	2021	Renal Cell Carcinoma	Partial Least Squares Random Forest Recursive feature elimination K-NN	A 10-metabolite panel predicted Renal Cell Carcinoma within the test cohort with 88% accuracy	A total of 7,147 metabolites were narrowed down to a series of 10 and tested with 4 ML algorithms all of which were able to correctly identify RCC status with high accuracy in the test cohort	doi: 10.1021/acs.jproteome.1c00213. Epub 2021 Jun 23. PMID: 34161092
6	Tiedt et al	Ann Neurology	2020	Ischemic Stroke	Random Forest classification Linear discriminant analysis logistic regression K-NN naive Bayes SVM	Identified 4 metabolites showing high accuracy in differentiating between Ischemic stroke and Stroke Mimics	Levels of 41 metabolites showed significant association with Ischemic stroke compared to controls. Top 4 metabolites show high accuracy in differentiating between stroke and mimics	https://doi.org/10.1002/ana.25859
7	Liu et al	Mol Metabolite	2021	Diabetic kidney disease	Linear discriminant analysis SVM Random Forest Logistic regression	Serum integrative omics provide stable and accurate biomarkers for early warning and diagnosis of Diabetic Kidney Disease	combination of a2-macroglobulin, cathepsin D, and CD324 could serve as a surrogate protein biomarker using 4 different ML methods	doi: 10.1016/j.molmet.2021.101,367. Epub 2021 Nov 1. PMID: 34737094; PMCID: PMC8609166
8	Oh et al	Cell Metab	2020	Cirrhosis	Random Forest	Comparison of the dysregulation between gut microbiome in differentiating between advanced fibrosis and cirrhosis	Identified a core set of gut microbiome that could be used as universal non-invasive test for cirrhosis	doi: 10.1016/j.cmet.2020.06.005. PMID: 32610095; PMCID: PMC7822714
9	Delafiori et al	Anal Chem	2021	COVID-19	ADA tree boosting Gradient tree boosting Random forest partial least squares SVM	Combine ML with mass spectrometry to differentiate between COVID-19 in plasma samples within minutes	Diagnosis can be derived from raw data with diagnosis specificity 96%, sensitivity 83%	doi: 10.1021/acs.analchem.0c04497. Epub 2021 Jan 20. PMID: 33471512; PMCID: PMC8023531
10	Jung et al	Biomed Pharmacother	2021	Coronary artery disease	Logistic regression	10-year risk prediction model based on 5 selected serum metabolites	provided initial evidence that blood xanthine and uric acid levels play different roles in the development of machine learning models for primary/secondary prevention or diagnosis of CAD.	doi: 10.1016/j.biopha.2021.111,621. Epub 2021 May 10. PMID: 34243599

(Continued on following page)

TABLE 2 (*Continued*) Examples of metabolomics studies utilizing ML algorithms.

#	Author	Journal	Publication year	Area of investigation	ML algorithm used	Brief description	Findings	Doi
							Purine-related metabolites in blood are applicable to machine learning model development for CAD risk prediction and diagnosis	
11	Wallace et al	J Pathol	2020	Cancer	Linear discriminant analysis	Comparison between metabolic profile of tumor patients and the predictive ability of machine learning algorithm to interpret metabolite data	Application of machine learning algorithms to metabolite profiles improved predictive ability for hard-to-interpret cases of head and neck paragangliomas (99.2%)	doi: 10.1002/path.5472. Epub 2020 Jul 1. PMID: 32462735; PMCID: PMC7548960
12	Kouznetsova et al	Metabolomics	2019	Bladder cancer	Logistic regression	Elucidate the biomarkers including metabolites and corresponding genes for different stages of Bladder cancer, show their distinguishing and common features, and create a machine-learning model for classification of stages of Bladder cancer	The best performing model was able to predict metabolite class with an accuracy of 82.54%. The same model was applied to three separate sets of metabolites obtained from public sources, one set of the late-stage metabolites and two sets of the early-stage metabolites. The model was better at predicting early-stage metabolites with accuracies of 72% (18/25) and 95% (19/20) on the early sets, and an accuracy of 65.45% (36/55) on the late-stage metabolite set.	doi: 10.1007/s11306-019-1,555-9. PMID: 31222577
13	Murata et al	Breast Cancer Res Treat	2019	Breast Cancer	Multiple logistic regression	Combinations of salivary metabolomics and machine learning methods show potential for non-invasive screening of breast cancer	Polyamines were identified to be significantly elevated in saliva of breast cancer patients	doi: 10.1007/s10549-019-05330-9. Epub 2019 Jul 8. PMID: 31286302
14	Liu et al	BMC Genomics	2016	Major Depressive Disorder	SVM Random Forest	Identifying the metabolomics signature of major depressive disorder subtypes	~80% accuracy in classification of melancholic depression doi: 10.1186/s12864-016-2,953-2. PMID: 27549765; PMCID: PMC4994306	

displaying the highest accuracy (AUC 0.93). This DL method also identified eight unique metabolic pathways that seem to promote breast cancer. The study's findings suggest that DL may be used to deduce the topology of affected biochemical pathways from a network analysis of a metabolomics data set.

The predictive abilities of five potential urinary biomarkers for breast cancer were evaluated by Kim et al. (Kim et al., 2010). Multivariate methods (linear and Gaussian SVM algorithms, decision trees, and RF) were shown to outperform univariate methods by about 6.6–12.7%. It is noteworthy, however, that the linear SVM model scored the lowest in specificity.

Endometrial cancer

Cheng et al. (Cheng et al., 2019) applied 4 ML algorithms-SVM, Partial Least Square-Discriminant Analysis (PLS-DA), RF, and LR-to identify metabolomic biomarkers in cervicovaginal fluid for endometrial cancer detection. The SVM and RF techniques displayed the greatest accuracy of 78% (75% sensitivity and 80% specificity) in the testing dataset.

Hepatocellular carcinoma

Xue et al. (Xue et al., 2008) applied stepwise discriminant analysis (SDA) and SVM algorithms to identify a set of 13 serum metabolites to distinguish between patients with hepatocellular carcinoma and healthy controls with 75% accuracy. The metabolites included carbohydrates, amino acids, fatty acids, cholesterol, and low molecular weight organic acids.

Lung cancer

A more recent study used SVMs with untargeted lipidomics to identify features most important for early-stage lung cancer detection (Wang et al., 2022). Lung plasma lipidomic profiling was carried out on 311 participants using mass spectrometry. Using SVM feature selection, nine lipids were chosen for developing a liquid chromatography-mass spectrometry-based targeted assay. The authors validated the ability of these nine lipids to detect early-stage cancer across multiple independent cohorts, including a hospital-based lung cancer screening cohort of 1,036 participants and a prospective clinical cohort containing 109 participants, in which the assay reached more than 90% sensitivity and 92% specificity. The selected lipids were also shown to be differentially expressed in early-stage lung cancer tissues *in situ*. This assay, which the authors named "Lung Cancer Artificial Intelligence Detector," shows promise for the early detection of lung cancer and large-scale screening of high-risk populations for cancer prevention.

Squamous cell carcinoma

In their 2019 study, Hsu et al. uncovered potential metabolic biomarkers for oral cavity squamous cell carcinoma (Hsu et al., 2019). They constructed a three-marker panel consisting of putrescine, glycyl-leucine, and phenylalanine, using an SVM model that can discriminate cancerous from adjacent non-cancerous tissues with high sensitivity and specificity based on receiver operating characteristic (ROC) analysis.

RF and SVM also demonstrated favorable results in the identification of esophageal squamous cell carcinoma tissue based on differential metabolites (Z. Chen et al., 2021). Among the three models evaluated, RF had the highest predictive performance (100%), but required more computational time (8.99 s), compared to PLS and SVM models, which showed similar predictive performance (95%) and similar computational time (1.27 s and 1.11 s). It is of note, however, that the three models prioritized different features.

Non-Hodgkin's lymphoma

Bueno Duarte et al., 2020, identified a panel of 18 urine metabolites that can differentiate diffuse large B-cell lymphoma patients from healthy individuals with 99.8% accuracy using an SVM model (Duarte et al., 2020).

Renal cell carcinoma

In another cancer study, Bifarin et al. (Bifarin et al., 2021), identified candidate urine metabolic panels for renal cell carcinoma (RCC) as a noninvasive diagnostic assay. Information from patients and controls was gathered and divided into the model and test cohorts. Multiple ML algorithms were used to test the predictive ability. These include RF, KNN, linear kernel SVMs, and RBF kernel SVMs. A total of 7,147 metabolomic features were identified from the NMR and MS platforms. These were then merged and filtered to only those that showed a greater than 1-fold change between the RCC and control samples, and highly positively correlated features were removed. This hybrid model resulted in a selection of 10 metabolites for a panel. RCC status was tested across the used ML models, and all of them were able to predict RCC status accurately.

Osteosarcoma

An RF classifier demonstrated superiority over an SVM model, with an accuracy of 85% *versus* 81% for the classification of osteosarcoma and benign tumor patients using both X-ray image features and serum metabolomic data (R. Shen et al., 2018).

Non-cancer conditions

Coronavirus disease (COVID-19)

With the onset of the COVID-19 pandemic, research groups across the globe conducted numerous investigations trying to understand if there was any biological reasoning behind disease heterogeneity, in terms of disease severity, presentation, and even mortality rate. For example, Chen et al. (B. Shen et al., 2020) combined proteomic and metabolomic profiles of 31 COVID-19 patients (18 non-sever, 13 severe) to create an ML molecular classifier, which was eventually able to identify potential blood biomarkers for severe COVID-19. The devised RF model identified 29 variables of priority (22 proteins, seven metabolites); this model had a 0.957 AUC in the training set. Subsequent testing of the model against an independent cohort of 10 patients revealed accurate identification of severe COVID-19 patients for all but one of the cohort. The incorrectly identified patients had potential confounding factors, i.e., age, long period of administration of non-traditional medicine, and several comorbidities. The generated classifier was again tested against a model with 29 randomly selected molecules. The randomly generated model exhibited low accuracy when compared with the classifier.

Type 2 diabetes (T2D)

Shomorony et al. (Shomorony et al., 2020) identified a set of cardiometabolic biomarkers beyond the standard clinical biomarkers that can be used to stratify individuals into disease types and stages. Data features from 1,385 diverse modalities (microbiome, genetics, metabolome, advanced imaging) were collected from 1,253 self-assessed healthy individuals. A linear regression ML algorithm was used to identify whether there were any associated covariates. This was then validated through correlation analysis to identify any significant associations between features. Network analysis was performed to determine whether the identified modalities had biomarker signatures that corresponded to underlying biological systems. Finally, using the identified features, cluster analysis was performed to stratify participants into subsets consistent with their respective health status. The findings were validated in an independent cohort of 1,083 females. The authors highlighted several novel biomarkers in diabetes signature and gut microbiome health, i.e., 1-stearoyl-2 dihomo-linoleoyl-GPC and cinnamoyl glycine, respectively.

Nonalcoholic fatty liver disease (NAFLD)

universal gut-microbiome signatures can be used to predict various diseases. This is true for Oh et al. (Oh et al., 2020) who used stool microbiome from 163 nonalcoholic fatty liver (NAFLD) disease patients and applied an RF ML algorithm with a differential abundance analysis to identify microbial and metabolomic signatures to detect cirrhosis and the authors were able to test the generated model and its ability to differentiate between cirrhosis and fibrosis. The model was able to correctly distinguish between the various stages of fibrosis with high accuracy AUC 0.85. The incorporation of further information into the model, i.e., serum AST levels, showed marked improvement in model performance with AUC 0.94.

Perakakis et al. trained models for the non-invasive diagnosis of non-alcoholic steatohepatitis (NASH) and NAFLD (Perakakis et al., 2019) from serum samples. SVM models including 29 lipids or combining lipids with glycans and/or hormones were shown to classify the conditions with 90% accuracy, and a 10-lipid-model could detect liver fibrosis with 98% accuracy.

Acute myocardial ischemia (AMI)

A multilayer perceptron (MLP) neural network-based model achieved superior results in detecting acute myocardial ischemia (AMI) from serum metabolites in a rat model compared to several other classification algorithms, including SVM, RF, Gradient tree boosting (GTB), and LR (Cao et al., 2022). The model achieved accuracy of 96.67% in the rat model and 88.23% in predicting AMI type II in human autopsy cases of sudden cardiac death.

Chronic kidney disease (CKD)

In an attempt to classify chronic kidney disease patients from serum metabolites, Guo et al. (Guo et al., 2019) constructed two NN; a two-layered fully connected multi-layer NN with MLP with 128 neurons in the hidden layer, and a three-layered CNN with 16 and 32 neurons in the two hidden layers, respectively. The MLP achieved accuracy of 90.4%, while the CNN reached accuracy of 90.6%. Both NNs, as well as an SVM model, were outperformed by an RF classifier with 100% accuracy. A possible reason is the rigorous feature reduction steps performed prior to model application; DL methods specialized in the analysis of high-dimensional data and in this study, from thousands of measured metabolites, only five were retained for the final models.

Celiac disease

In one of the earliest and highly cited studies, metabolic signatures of celiac disease, detected by NMR, were used to construct an SVM model able to differentiate celiac disease patients from healthy controls with 83.4% accuracy using serum metabolites and 69.3% using urine metabolites. After a 12-month gluten-free diet, the same model correctly classified all but one of the patients as healthy (Bertini et al., 2009).

Multiple Sclerosis (MS)

Waddington et al. used ML models including SVM, RF, k-NN, decision tree, and least absolute shrinkage and selection operator (LASSO) logistic regression to predict the tendency of multiple sclerosis patients treated with beta interferons to develop anti-drug antibodies (Waddington et al., 2020). Among the five classification models tested for predicting future immunogenicity from serum metabolomics data, SVMs were one of the most successful at differentiating between cases with and without drug resistance.

Major depressive disorder

Metabolomic signatures associated with certain conditions may still persist after disease remission, as shown in a study by Hung et al., 2021. Eight plasma metabolites were identified as significantly differentially-expressed in patients with major depressive disorder (MDD) with full remission compared with healthy controls. These were then used to construct an SVM model capable of differentiating patients with MDD with full remission from healthy controls with predictive accuracy of nearly 85% (Hung et al., 2021).

Schizophrenia

Chen et al. uncovered metabolic biomarkers that can differentiate between schizophrenia patients with violence and those without violence (X. Chen et al., 2020). RF and SVM analyses unveiled ten and five plasma metabolites, respectively. The common metabolites formed a biomarker panel, including the ratio of L-asparagine to L-aspartic acid, vanillylmandelic acid, and glutaric acid, yielding an AUC of 0.808.

Autism spectrum disorders

In a study conducted by Chen et al., urine organic acids were detected in children with autism spectrum disorder (ASD) and combined with three algorithms, PLS-DA, SVM, and eXtreme Gradient Boosting (XGBoost), for the diagnosis of autism (Hung et al., 2021; Q. Chen et al., 2019). The work proved that autism spectrum disorders present with characteristic metabolic biomarkers that can be investigated for diagnosis of the condition as well as for future research on the pathogenesis of autism and possible interventions.

Gestational age

Another application of ML in metabolomics is the investigation of the human pregnancy metabolome conducted by Liang et al. (Liang et al.,. 2020), where the authors were able to identify a series of compounds (460) and associated metabolic pathways (34) that were significantly changed during pregnancy. The authors were able to construct a linear regression model that correlates certain plasma metabolites with time in gestational age; this model is in high accordance with the ultrasound. An additional two to three metabolites were able to identify the time of labor, e.g., prediction of 2, 4, 6, or 8 weeks to the time of delivery.

Methodological studies

The right choice of ML algorithm is a crucial factor for the success of a metabolomics study. Analysis results usually depend more on the data (type, quantity, quality) than the applied algorithm. Complex, multivariate approaches may be suitable for large, multidimensional datasets; however, in the case of simple, linearly separable data, conventional statistical approaches often outperform ML. Therefore, a large number of metabolomic studies make an effort to compare the predictive ability of different ML algorithms to each other, as well as to more traditional statistical methods.

One of the comprehensive comparative studies is the work by Mendez et al. (Mendez, Reinke, and Broadhurst 2019), in which the authors compared 8 ML algorithms, partial least squares regression (PLS), principal component regression (PCR), principal component logistic regression (PCLR), RF, linear kernel SVM, non-linear SVM with RBF, linear and non-linear ANN, for the binary classification of ten clinical metabolomic datasets. As for the ANNs, the linear network was composed of two layers, with a small number of linear neurons in the hidden layer and a single sigmoidal neuron in the output layer. For the non-linear NN, the activation function of the hidden layer neurons was changed to a sigmoidal function. Both networks were implemented using stochastic gradient descent with a binary cross-entropy loss function. The authors expected non-linear machine ML algorithms, especially the DL models, to outperform linear alternatives. Nevertheless, SVM and ANN only slightly

TABLE 3 Pros and cons of ML algorithms and applicability within the field of metabolomics.

Algorithm	Pros	Cons	Metabolomic application
Linear Regression	- Excellent for linearly separable data - Easy implementation	- Assumes linear relationship between dependent and independent variables - Outliers have significant impact - Prone to overfitting	- Unknown relationship between dependent and independent variables - Forecasting tasks
Logistic Regression	- Simple implementation - No Feature scaling needed - No hyper-parameter tuning needed	- Easily outperformed by other algorithms - Heavily reliant on proper identification of data	- Multiclass classification, i.e., when output class only has two possible outcomes e.g., cancer detection (yes or no) - Linear relationship between dependent and independent variables
Naive Bayes	- Fast predictions of dataset classes - Good for datasets with categorical variables	- Assumes all features are independent	- Dataset with highly independent features - For multi-class predictions
Support Vector Machines (SVMs)	- Works well for data that can be easily seperated with clear margin of separation - Effective for high dimension data	- Requires more training time for large datasets - Does not perform well when dataset has high level of noise i.e. overlapping target classes	- Medium sized dataset - Large number of features - Linear relationship between dependent and independent variables
k-Nearest Neighbors (*k*-NN)	- Easy implementation - Can solve multi-class problems - No data assumption needed	- Slow performance on large datasets - Data scaling required - Not for data with high dimensionality i.e. large number of features - Sensitive to missing values, outliers and imbalance data	- Small datasets with small number of features - Unknown relationship between dependant and independent variables - Useful for targeted metabolomics approaches
Decision Trees	- Scaling or normalization of data not needed - Able to handle missing values - Easy to visualize - Automatic feature selection	- Data sensitive - Might need more time to train trees - High chance of overfitting	- Known to suffer from a high chance of overfitting
Random Forest (RF)	- Good performance on imbalanced or missing data - Able to handle huge amounts of data - Feature importance extraction - Low chance of overfitting	- Predictions are uncorrelated - Influence of dependent variable on independent variable is unknown, i.e., Black box - Data sensitive	- Identification of variables with high importance - Useful for datasets with small sample population - Useful for metabolic fingerprinting approaches
Neural Networks (NN)	- Flexible network architecture i.e., can be used for regression and classification - Good with nonlinear data - Can handle large number of inputs - Fast predictions once trained	- Influence of dependent variable on the independent variable is unknown, i.e., Black box - Highly dependant on training data - Prone to overfitting and generalization - Extremely hardware dependant i.e., the larger the datasets, the more expensive and time-consuming the modeling process	- Data with a non-linear relationship between dependant and independent variables - Large datasets with a stipulation on time and cost - Can be applied to raw metabolomic data for feature extraction and multivariate classification combined into a single model - Integration of multi-omics data, i.e., collected over different times, multiple analytical platforms, biofluids, or omic platforms - Useful for metabolic profiling

surpassed PLS across all datasets, while RF performed poorly. In conclusion, no single DL or ML algorithm could be identified as superior.

In another 2019 study, Vu et al. evaluated the performance of five classification algorithms (PLS, OPLS, Principal component-Linear Discriminant Analysis (PC-LDA), RF, and SVM) using simulated and experimental 1D ^{1}H NMR spectral data sets (Vu et al., 2019). Datasets with clear group separation were classified equally well by all five models. However, when the data contained subtle differences between classes, OPLS produced the best results, as it was able to identify the useful discriminant features with good classification accuracy. It is noteworthy that although RF and PC-LDA classified the data more accurately than the other models, they achieved so using the wrong discriminant features.

The superiority of SVM and RF classifiers was demonstrated in an evaluation of seven classification techniques using both simulated and real metabolomics datasets (Trainor, DeFilippis, and Rai 2017). In the simulated datasets, the classifiers performed as follows (from least to greatest error): SVM, RF, Naïve Bayes, sparse PLS, ANN, PLS, and *k*-NN, while SVM and RF consistently outperformed the rest over the real datasets.

Expanding on the gut microbiome, Han et al. (Han et al., 2021) used RF models to identify sets of metabolites that are able to provide taxonomic distinction and classify the origin of microbial supernatants while also providing insights into highly conserved chemical features that are predictive of taxonomic identity. Han et al. were able to construct a chemical standard library-informed metabolomics pipeline that is both customizable and expandable. This method was used to construct an atlas of metabolic activity that can enable functional studies of the gut microbial communities and was validated using RF ML algorithms.

Concluding remarks

In this work, we provided a review of popular ML techniques as well as key studies that have applied them for the stratification of metabolites from various conditions.

RF and SVM have been among the most widely used algorithms in metabolomic studies. Although DL is a comparatively new player in the field, it is undoubtedly paving its way to metabolomics - and generally to the other omics and integrative multi-omics studies - as evident by the growing number of reports that use NNs in metabolomic analyses.

Cancer is by far the most studied condition, with ML algorithms having been applied to the supervised classification of cancer *versus* control sample sets from metabolic data obtained from various cancer types, including ovarian, breast, endometrial, lung and liver cancer, renal carcinoma, squamous cell carcinomas, osteosarcoma, and lymphomas.

Choosing the appropriate ML algorithm is crucial to the success of a metabolomics study. It is essential for researchers to be informed of the benefits of each ML approach and to choose one that best suits their needs to obtain reliable and interpretable outcomes. However, after reviewing a number of studies that compared different ML methods, no specific conclusion can be drawn regarding the choice of the algorithm. ML methods that produce good results in some investigations might perform poorly in others. The dimensionality, quality, and characteristics of input data and appropriate feature selection techniques play a significant role in the performance and behavior of the ML methods and their outcomes.

In addition, the choice of hyper-parameters and how they are tuned can influence the results remarkably. Accordingly, a detailed methodology for selecting the most suitable ML algorithm is a topic that needs further investigation. However, we can offer some insight into the pros and cons of each of the popular algorithms discussed in this review, as well as some suggested recommendations regarding their applications within the metabolomics field (Table 3) (Kell 2005; Kourou et al., 2015; Libbrecht and Stafford Noble, 2015; Soofi and Awan 2017; Malakar et al., 2018; Shinde and Shah 2018; Liebal et al., 2020).

Significantly altered metabolites generated by metabolomic experiments and unveiled by machine learning approaches can serve as a starting point for a number of investigations. Biomarker discovery is a definite main target. Nevertheless, their actual predictive ability needs to be further experimentally validated. Further investigations like enrichment studies and pathway analysis can provide new insights into the roles the identified metabolites play in the pathophysiology of various conditions. Additionally, the feasibility of targeting specific metabolites for disease treatment can be explored.

It is noteworthy that most of the reviewed work was published within the last 5 years, which aligns with the obvious rise in popularity ML has gained in recent years, enabled by an increase in computation power, efficiency and accessibility of ML tools, familiarity with the field and abundance of data. As more and large metabolomic data sets become available, it is expected that ML techniques, especially DL, will play a bigger role in building informative and predictive models that can be used to provide high-definition, personalized clinical diagnosis, and treatment.

Author contributions

AM planned and supervised the project. AG and MT wrote the manuscript draft and created the visualizations. AM, AG, and MT reviewed and edited the final manuscript.

Funding

AG was supported by the Mohamed Bin Abdulkarim Allehedan Ph.D. fellowship. MT was partially supported by the American University in Cairo Laboratory Fellowship. AM was partially supported by the American University in Cairo Centennial Grant.

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/fgene.2022.1017340/full#supplementary-material

References

Abram, K. J., and McCloskey, D. (2022). A comprehensive evaluation of metabolomics data preprocessing methods for deep learning. *Metabolites* 12 (3), 202. doi:10.3390/metabo12030202

Aderemi, A. V., Ayeleso, A. O., Oyedapo, O. O., and Mukwevho, E. (2021). Metabolomics: A scoping review of its role as a tool for disease biomarker discovery in selected non-communicable diseases. *Metabolites* 11 (7), 418. doi:10.3390/metabo11070418

Ahola-Olli, A. V., Mustelin, L., Kalimeri, M., Kettunen, J., Jokelainen, J., Auvinen, J., et al. (2019). Circulating metabolites and the risk of type 2 diabetes: A prospective study of 11, 896 young adults from four Finnish cohorts. *Diabetologia* 62 (12), 2298–2309. doi:10.1007/s00125-019-05001-w

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., and Salakoski, T. (2011). An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Comput. Statistics Data Analysis* 55 (4), 1828–1844. doi:10.1016/j.csda.2010.11.018

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 (1), 337–347. doi:10.1021/acs.jproteome.7b00595

Allen, J., Davey, H. M., Broadhurst, D., Heald, J. K., Rowland, J. J., Oliver, S. G., et al. (2003). High-throughput classification of yeast mutants for functional Genomics using metabolic footprinting. *Nat. Biotechnol.* 21 (6), 692–696. doi:10.1038/nbt823

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Statistician* 46 (3), 175–185. doi:10.2307/2685209

Antonakoudis, A., Barbosa, R., Kotidis, P., and Kontoravdi, C. (2020). The era of big data: Genome-scale modelling meets machine learning. *Comput. Struct. Biotechnol. J.* 18, 3287–3300. doi:10.1016/j.csbj.2020.10.011

Baumgartner, C., Böhm, C., and Baumgartner, D. (2005). Modelling of classification rules on metabolic patterns including machine learning and expert knowledge. *J. Biomed. Inf.* 38 (2), 89–98. doi:10.1016/j.jbi.2004.08.009

Bellet, A., Habrard, A., and Sebban, M. (2013). "A survey on metric learning for feature vectors and structured data." arXiv [cs.LG]. arXiv. Available at: http://arxiv.org/abs/1306.6709.

Ben-Hur, A., Horn, D., Siegelmann, H. T., and Vapnik, V. (2002). Support vector clustering. *J. Mach. Learn. Res. JMLR* 2, 125–137.

Bertini, I., Calabrò, A., De Carli, V., Luchinat, C., Nepi, S., Porfirio, B., et al. (2009). The metabonomic signature of celiac disease. *J. Proteome Res.* 8 (1), 170–177. doi:10.1021/pr800548z

Bifarin, O. O., Gaul, D. A., Sah, S., Arnold, R. S., Ogan, K., Master, V. A., et al. (2021). Machine learning-enabled renal cell carcinoma status prediction using multiplatform urine-based metabolomics. *J. Proteome Res.* 20 (7), 3629–3641. doi:10.1021/acs.jproteome.1c00213

Biswas, A., Saran, I., and Perry Wilson, F. (2021). Introduction to supervised machine learning. *Kidney360* 2 (5), 878–880. doi:10.34067/KID.0000182021

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers," in Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT. doi:10.1145/130385.13040192

Bottou, L. (2010). "Large-scale machine learning with stochastic gradient descent," in Proceedings of the COMPSTAT'2010. Physica-Verlag HD. Editors Y. Lechevallier and G. Saporta. doi:10.1007/978-3-7908-2604-3_16

Boubezoul, A., Paris, S., and Ouladsine, M. (2008). Application of the cross entropy method to the GLVQ algorithm. *Pattern Recognit.* 41 (10), 3173–3178. doi:10.1016/j.patcog.2008.03.016

Broomhead, D., and Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Syst.* 2, 321–355.

Bzdok, D., Altman, N., and Martin, K. (2018). Statistics versus machine learning. *Nat. Methods* 15 (4), 233–234. doi:10.1038/nmeth.4642

Cao, J., Li, J., Gu, Z., Niu, J-J., An, G-S., Jin, Q-Q., et al. (2022). Combined metabolomics and machine learning algorithms to explore metabolic biomarkers for diagnosis of acute myocardial ischemia. *Int. J. Leg. Med.* doi:10.1007/s00414-022-02816-y

Casson, R. J., and Farmer, L. D. M. (2014). Understanding and checking the assumptions of linear regression: A primer for medical researchers. *Clin. Exp. Ophthalmol.* 42 (6), 590–596. doi:10.1111/ceo.12358

Cavus, E., Karakas, M., Ojeda, F. M., Kontto, J., Veronesi, G., Ferrario, M. M., et al. (2019). Association of circulating metabolites with risk of coronary heart disease in a European population: Results from the biomarkers for cardiovascular risk assessment in europe (BiomarCaRE) consortium. *JAMA Cardiol.* 4 (12), 1270–1279. doi:10.1001/jamacardio.2019.4130

Chen, Q., Qiao, Y., Xu, X-J., You, X., and Tao, Y. (2019). Urine organic acids as potential biomarkers for autism-spectrum disorder in Chinese children. *Front. Cell. Neurosci.* 13, 150. doi:10.3389/fncel.2019.00150

Chen, X., Xu, J., Tang, J., Dai, X., Huang, H., Cao, R., et al. (2020). Dysregulation of amino acids and lipids metabolism in schizophrenia with violence. *BMC Psychiatry* 20 (1), 97. doi:10.1186/s12888-020-02499-y

Chen, Z., Gao, Y., Huang, X., Yao, Y., Chen, K., Su, Z., et al. (2021). Tissue-based metabolomics reveals metabolic biomarkers and potential therapeutic targets for esophageal squamous cell carcinoma. *J. Pharm. Biomed. Anal.* 197, 113937. doi:10.1016/j.jpba.2021.113937

Cheng, S-C., Chen, K., Chiu, C-Y., Lu, K-Y., Lu, H-Y., Chiang, M-H., et al. (2019). Metabolomic biomarkers in cervicovaginal fluid for detecting endometrial cancer through nuclear magnetic resonance spectroscopy. *Metabolomics* 15 (11), 146. doi:10.1007/s11306-019-1609-z

Cohen, S. (2021). "Chapter 1 - the evolution of machine learning: Past, present, and future," in *Artificial intelligence and deep learning in pathology. 1–12.* Editor S. Cohen (Elsevier).

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20 (3), 273–297. doi:10.1007/bf00994018

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control Signal. Syst.* 2 (4), 303–314. doi:10.1007/bf02551274

Deepthi, Y., Pavan Kalyan, K., Vyas, M., Radhika, K., Babu, D. K., and Krishna Rao, N. V. (2020). "Disease prediction based on symptoms using machine learning," in *Energy systems, drives and automations* (Singapore: Springer), 561–569.

Dettmer, K., Aronov, P. A., and Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrom. Rev.* 26 (1), 51–78. doi:10.1002/mas.20108

Dhall, D., Kaur, R., and Juneja, M. (2020). "Machine learning: A review of the algorithms and its applications," in Proceedings of ICRIC 2019 (Springer International Publishing).47–63

Duarte, G. H. B., Fernandes, A. A. D. P., Silva, A. A. R., Zamora-Obando, H. R., Amaral, A. G., Mesquita, A. D. S., et al. (2020). Gas chromatography-mass spectrometry untargeted profiling of non-hodgkin's lymphoma urinary metabolite markers. *Anal. Bioanal. Chem.* 412 (27), 7469–7480. doi:10.1007/s00216-020-02881-5

Fiehn, O. (2002). Metabolomics-the link between genotypes and phenotypes. *Plant Mol. Biol.* 48 (1-2), 155–171. doi:10.1023/a:1013713905833

Friedrich, N. (2012). Metabolomics in diabetes research. *J. Endocrinol.* 215 (1), 29–42. doi:10.1530/JOE-12-0120

Gajda, S., and Chlebus, M. (2022). A probability-based models ranking approach: An alternative method of machine-learning model performance assessment. *Sensors* 22 (17), 6361. doi:10.3390/s22176361

Garthwaite, P. H. (1994). An interpretation of partial least squares. *J. Am. Stat. Assoc.* 89 (425), 122–127. doi:10.1080/01621459.1994.10476452

Gates, S. C., and Sweeley, C. C. (1978). Quantitative metabolic profiling based on gas chromatography. *Clin. Chem.* 24 (10), 1663–1673. doi:10.1093/clinchem/24.10.1663

Gaul, D. A., Mezencev, R., Long, T. Q., Jones, C. M., Benigno, B. B., Gray, A., et al. (2015). Highly-accurate metabolomic detection of early-stage ovarian cancer. *Sci. Rep.* 5 (1), 16351–16357. doi:10.1038/srep16351

Glorot, X., Antoine, B., and Bengio, Y. (2011). "Deep sparse rectifier neural networks," in JMLR Workshop and Conference Proceedings. Available at: https://proceedings.mlr.press/v15/glorot11a.html.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 27. Available at: https://proceedings.neurips.cc/paper/5423-generative-adversarial-nets. doi:10.1145/3422622

Gowd, G. A. N., Zhang, S., Gu, H., Vincent, A., Shanaiah, N., and Raftery, D. (2008). Metabolomics-based methods for early disease diagnostics. *Expert Rev. Mol. diagn.* 8 (5), 617–633. doi:10.1586/14737159.8.5.617

Graham, E., Lee, J., Price, M., Tarailo-Graovac, M., Matthews, A., Engelke, U., et al. (2018). Integration of Genomics and metabolomics for prioritization of rare disease variants: A 2018 literature review. *J. Inherit. Metab. Dis.* 41 (3), 435–445. doi:10.1007/s10545-018-0139-6

Griffiths, W. J., and Wang., Y. (2009). Mass spectrometry: From proteomics to metabolomics and lipidomics. *Chem. Soc. Rev.* 38 (7), 1882–1896. doi:10.1039/b618553n

Guan, W., Zhou, M., Hampton, C. Y., Benigno, B. B., Deette Walker, L., Gray, A., et al. (2009). Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. *BMC Bioinforma.* 10 (1), 259–315. doi:10.1186/1471-2105-10-259

Guasch-Ferré, M., Hruby, A., Toledo, E., Clish, C. B., Martínez-González, M. A., Salas-Salvadó, J., et al. (2016). Metabolomics in prediabetes and diabetes: A systematic review and meta-analysis. *Diabetes Care* 39 (5), 833–846. doi:10.2337/dc15-2251

Guijas, C., Rafael Montenegro-Burke, J., Domingo-Almenara, X., Palermo, A., Warth, B., Hermann, G., et al. (2018). Metlin: A technology platform for identifying knowns and unknowns. *Anal. Chem.* 90 (5), 3156–3164. doi:10.1021/acs.analchem.7b04424

Guo, Y., Hui, Y., Chen, D., and Zhao, Y-Y. (2019). Machine learning distilled metabolite biomarkers for early stage renal injury. *Metabolomics* 16 (1), 4. doi:10.1007/s11306-019-1624-0

Guyon, I. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* Available at: http://citeseer.ist.psu.edu/viewdoc/summary?

Halket, J. M., Waterman, D., Przyborowska, A. M., Patel, R. K. P., Fraser, P. D., and Bramley, P. M. (2005). Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. *J. Exp. Bot.* 56 (410), 219–243. doi:10.1093/jxb/eri069

Han, S., Van Treuren, W., Fischer, C. R., Merrill, B. D., DeFelice, B. C., Sanchez, J. M., et al. (2021). A metabolomics pipeline for the mechanistic interrogation of the gut microbiome. *Nature* 595 (7867), 415–420. doi:10.1038/s41586-021-03707-9

Hasin, Y., Marcus, S., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18 (1), 83. doi:10.1186/s13059-017-1215-1

Hastie, T., Tibshirani, R., and Friedman, J. (2009). "The Elements of statistical learning" in *Springer Series in Statistics*. doi:10.1007/978-0-387-84858-7

Henneges, C., Bullinger, D., Fux, R., Friese, N., Seeger, H., Neubauer, H., et al. (2009). Prediction of breast cancer by profiling of urinary RNA metabolites using support vector machine-based feature selection. *BMC Cancer* 9, 104. doi:10.1186/1471-2407-9-104

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313 (5786), 504–507. doi:10.1126/science.1127647

Ho, T. K. (1995). Random decision Forests. *Proc. 3rd Int. Conf. Document Analysis Recognit.* 11, 278–282.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.* 2 (5), 359–366. doi:10.1016/0893-6080(89)90020-8

Hou, X-W., Wang, Y., and Pan, C-W. (2021). Metabolomics in diabetic retinopathy: A systematic review. *Invest. Ophthalmol. Vis. Sci.* 62 (10), 4. doi:10.1167/iovs.62.10.4

Hsu, C-W., Chen, Y-T., Hsieh, Y-J., Chang, K-P., Hsueh, P-C., Chen, T-W., et al. (2019). Integrated analyses utilizing metabolomics and transcriptomics reveal perturbation of the polyamine pathway in oral cavity squamous cell carcinoma. *Anal. Chim. Acta* 1050, 113–122. doi:10.1016/j.aca.2018.10.070

Hung, I., Lin, G., Chiang, M-H., and Chiu, C-Y. (2021). Metabolomics-based discrimination of patients with remitted depression from healthy controls using 1H-NMR spectroscopy. *Sci. Rep.* 11 (1), 15608. doi:10.1038/s41598-021-95221-1

Iida, M., Harada, S., and Takebayashi, T. (2019). Application of metabolomics to epidemiological studies of atherosclerosis and cardiovascular disease. *J. Atheroscler. Thromb.* 26 (9), 747–757. doi:10.5551/jat.RV17036

Iqbal, T., Elahi, A., Wijns, W., and Shahzad, A. (2022). Exploring unsupervised machine learning classification methods for physiological stress detection. *Front. Med. Technol.* 4, 782756. doi:10.3389/fmedt.2022.782756

Jendoubi, T. (2021). Approaches to integrating metabolomics and multi-omics data: A primer. *Metabolites* 11 (3), 184. doi:10.3390/metabo11030184

Jolliffe, I. (2005). "Principal component analysis," in *Encyclopedia of statistics in behavioral science* (Chichester, UK: John Wiley & Sons). doi:10.1002/0470013192.bsa501

Kamiński, B., Jakubczyk, M., and Szufel, P. (2018). A framework for sensitivity analysis of decision trees. *Cent. Eur. J. Oper. Res.* 26 (1), 135–159. doi:10.1007/s10100-017-0479-6

Kearns, M., and Valiant, L. G. (1989). "Crytographic limitations on learning boolean formulae and finite automata," in Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing (New York, NY, USA: Association for Computing Machinery). STOC '89.433–44

Kell, D. B. (2005). Metabolomics, machine learning and modelling: Towards an understanding of the language of cells. *Biochem. Soc. Trans.* 33 (3), 520–524. doi:10.1042/BST0330520

Kim, Y., Koo, I., Jung, B. H., Chung, B. C., and Lee, D. (2010). Multivariate classification of urine metabolome profiles for breast cancer diagnosis. *BMC Bioinforma.* 11, S4. doi:10.1186/1471-2105-11-S2-S4

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17. doi:10.1016/j.csbj.2014.11.005

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521 (7553), 436–444. doi:10.1038/nature14539

Liang, L., Rasmussen, M-L. H., Piening, B., Shen, X., Chen, S., Röst, H., et al. (2020). Metabolic dynamics and prediction of gestational age and time to delivery in pregnant women. *Cell* 181 (7), 1680–1692. e15. doi:10.1016/j.cell.2020.05.002

Libbrecht, M. W., and Stafford Noble, W. (2015). Machine learning applications in genetics and Genomics. *Nat. Rev. Genet.* 16 (6), 321–332. doi:10.1038/nrg3920

Liebal, U. W., Phan, A. N. T., Sudhakar, M., Raman, K., and Blank, L. M. (2020). Machine learning applications for mass spectrometry-based metabolomics. *Metabolites* 10 (6), E243. doi:10.3390/metabo10060243

Liu, H., and Motoda, H. (2012), Feature extraction, construction and selection: A data mining perspective. in *The springer international series in engineering and computer science* (New York, NY: Springer), 453.

Maaten, L. V. D., and Hinton, G. (2008). Visualizing data using T-SNE. *J. Mach. Learn. Res. JMLR*. Available at: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl.

MacQueen, J. (1967)., 5.1. Berkeley, CA: University of California Press, 281–298.Some methods for classification and analysis of multivariate observations*Proc. Fifth Berkeley Symposium Math. Statistics Probab.*

Malakar, P., Balaprakash, P., Vishwanath, V., Morozov, V., and Kumaran, K. (2018). "Benchmarking machine learning methods for performance modeling of scientific applications," in 2018 IEEE/ACM Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems, Dallas, TX, USA, 12-12 November 2018 (PMBS).33–44. doi:10.1109/PMBS.2018.8641686

Mapelli, V., Olsson, L., and Nielsen, J. (2008). Metabolic footprinting in microbiology: Methods and applications in functional Genomics and biotechnology. *Trends Biotechnol.* 26 (9), 490–497. doi:10.1016/j.tibtech.2008.05.008

McGranaghan, P., Saxena, A., Rubens, M., Radenkovic, J., Bach, D., Schleußner, L., et al. (2020). Predictive value of metabolomic biomarkers for cardiovascular disease risk: A systematic review and meta-analysis. *Biomarkers* 25 (2), 101–111. doi:10.1080/1354750X.2020.1716073

McInnes, Le, Healy, J., and James, M. (2018). "Umap: Uniform Manifold approximation and projection for dimension reduction." arXiv [stat.ML]. arXivAvailable at: http://arxiv.org/abs/1802.03426.

Mendez, K. M., Broadhurst, D. I., and Reinke, S. N. (2019). The application of artificial neural networks in metabolomics: A historical perspective. *Metabolomics: Official journal of the Metabolomic Society* 15 (11), 142. doi:10.1007/s11306-019-1608-0

Mendez, K. M., Reinke, S. N., and Broadhurst, D. 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 (12), 150. doi:10.1007/s11306-019-1612-4

Miao, J., and Niu, L. (2016). A survey on feature selection. *Procedia Comput. Sci.* 91 (1), 919–926. doi:10.1016/j.procs.2016.07.111

Migdadi, L., Lambert, J., Ahmad, T., Roland, H., and Wöhler, C. (2021). Automated metabolic assignment: Semi-supervised learning in metabolic analysis employing two dimensional nuclear magnetic resonance (NMR). *Comput. Struct. Biotechnol. J.* 19 (8), 5047–5058. doi:10.1016/j.csbj.2021.08.048

Min, S., Lee, B., and Yoon, S. (2016). Deep learning in bioinformatics. *Brief. Bioinform.* 18 (5), 851–869. doi:10.1093/bib/bbw068

Misra, B. B., Langefeld, C. D., Olivier, M., and Cox, L. A. (2018). Integrated omics: Tools, advances, and future approaches. *J. Mol. Endocrinol.* 62, R21–R45. doi:10.1530/JME-18-0055

Mookherjee, A., Mitra, M., Kutty, N. N., Mitra, A., and Maiti, M. K. (2020). Characterization of endo-metabolome exhibiting antimicrobial and antioxidant activities from endophytic fungus cercospora sp. PM018. *South Afr. J. Bot.* 134, 264–272. doi:10.1016/j.sajb.2020.01.040

Müller, J., Bertsch, T., Volke, J., Schmid, A., Klingbeil, R., Metodiev, Y., et al. (2021). Narrative review of metabolomics in cardiovascular disease. *J. Thorac. Dis.* 13 (4), 2532–2550. doi:10.21037/jtd-21-22

Narayan, S. (1997). The generalized sigmoid activation function: Competitive supervised learning. *Inf. Sci.* 99 (1), 69–82. doi:10.1016/s0020-0255(96)00200-9

Newgard, C. B. (2017). Metabolomics and metabolic diseases: Where do we stand? *Cell Metab.* 25 (1), 43–56. doi:10.1016/j.cmet.2016.09.018

Nguyen, H. V., Asif Naeem, M., Wichitaksorn, N., and Pears, R. (2019). A smart system for short-term price prediction using time series models. *Comput. Electr. Eng.* 76, 339–352. doi:10.1016/j.compeleceng.2019.04.013

Oh, T. G., Kim, S. M., Caussy, C., Fu, T., Guo, J., Bassirian, S., et al. (2020). A universal gut-microbiome-derived signature predicts cirrhosis. *Cell Metab.* 32 (5), 878–888. doi:10.1016/j.cmet.2020.06.005

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., and Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian J. Ophthalmol.* 56 (1), 45–50. doi:10.4103/0301-4738.37595

Perakakis, N., Polyzos, S. A., Yazdani, A., Sala-Vila, A., Kountouras, J., Anastasilakis, A. D., et al. (2019). Non-invasive diagnosis of non-alcoholic steatohepatitis and fibrosis with the use of omics and supervised learning: A proof of concept study. *Metabolism.* 101 , 154005. doi:10.1016/j.metabol.2019.154005

Pomyen, Y., Wanichthanarak, K., Poungsombat, P., Fahrmann, J., Grapov, D., and Khoomrung, S. (2020). Deep metabolome: Applications of deep learning in metabolomics. *Comput. Struct. Biotechnol. J.* 18, 2818–2825. doi:10.1016/j.csbj.2020.09.033

Powell, M. (1987). "Radial basis functions for multivariable interpolation: A review." Available at: https://www.semanticscholar.org/paper/c71ca26b183025b9f39f940f5e730f2c9a64e414.

Raffone, A., Troisi, J., Boccia, D., Travaglino, A., Capuano, G., Insabato, L., et al. (2020). Metabolomics in endometrial cancer diagnosis: A systematic review. *Acta Obstet. Gynecol. Scand.* 99 (9), 1135–1146. doi:10.1111/aogs.13847

Riffenburgh, R. H. (1957). "Linear discriminant analysis." Available at: https://vtechworks.lib.vt.edu/handle/10919/80187.

Rossi, F., and Villa, N. (2006). Support vector machine for functional data classification. *Neurocomputing* 69 (7), 730–742. doi:10.1016/j.neucom.2005.12.010

Ruiz-Canela, M., Hruby, A., Clish, C. B., Liang, L., Martínez-González, M. A., and Hu, F. B. (2017). Comprehensive metabolomic profiling and incident cardiovascular disease: A systematic review. *J. Am. Heart Assoc.* 6 (10), e005705. doi:10.1161/JAHA.117.005705

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). *Learning internal representations by error propagation*. California Univ San Diego La Jolla Inst for Cognitive Science. Available at: https://apps.dtic.mil/sti/citations/ADA164453.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23 (19), 2507–2517. doi:10.1093/bioinformatics/btm344

Schmidt, D. R., Patel, R., Kirsch, D. G., Lewis, C. A., Heiden, M. G. V., and Locasale, J. W. (2021). Metabolomics in cancer research and emerging applications in clinical oncology. *Ca. Cancer J. Clin.* 71 (4), 333–358. doi:10.3322/caac.21670

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

Shah, N. J., Sureshkumar, S., and Shewade, D. G. (2015). Metabolomics: A tool ahead for understanding molecular mechanisms of drugs and diseases. *Indian J. Clin. biochem.* 30 (3), 247–254. doi:10.1007/s12291-014-0455-z

Shalev-Shwartz, S., and Ben-David, S. (2014). "Decision trees," in *Understanding machine learning: From theory to algorithms* (Cambridge University Press), 212–18.

Shen, B., Xiao, Y., Sun, Y., Bi, X., Du, J., Zhang, C., et al. (2020). Proteomic and metabolomic characterization of COVID-19 patient sera. *Cell* 182 (1), 59–72. doi:10.1016/j.cell.2020.05.032

Shen, R., Li, Z., Zhang, L., Hua, Y., Mao, M., Li, Z., et al. (2018). "Osteosarcoma patients classification using plain X-rays and metabolomic data," in Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Engineering in Medicine and Biology Society), 690–693.

Shinde, P. P., and Shah., S. (2018). "A review of machine learning and deep learning applications," in Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).1–6. doi:10.1109/ICCUBEA.2018.8697857

Shomorony, I., Cirulli, E. T., Huang, L., Napier, L. A., Heister, R. R., Hicks, M., et al. (2020). An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. *Genome Med.* 12 (1), 7. doi:10.1186/s13073-019-0705-z

Shulaev, V. (2006). Metabolomics technology and bioinformatics. *Brief. Bioinform.* 7 (2), 128–139. doi:10.1093/bib/bbl012

Silva, L. P., and Northen, T. R. (2015). Exometabolomics and MSI: Deconstructing how cells interact to transform their small molecule environment. *Curr. Opin. Biotechnol.* 34, 209–216. doi:10.1016/j.copbio.2015.03.015

Smith, C. A., Grace O'Maille, E. J. W., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., et al. (2005). Metlin: A metabolite mass spectral database. *Ther. Drug Monit.* 27 (6), 747–751. doi:10.1097/01.ftd.0000179845.53213.39

Soofi, A. A., and Awan, Arshad (2017). Classification techniques in machine learning: Applications and issues. *J. Basic Appl. Sci.* 13, 459–465. doi:10.6000/1927-5129.2017.13.76

Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Acad. Emerg. Med.* 18 (10), 1099–1104. doi:10.1111/j.1553-2712.2011.01185.x

Streese, L., Springer, A. M., Deiseroth, A., Carrard, J., Infanger, D., Schmaderer, C., et al. (2021). Metabolic profiling links cardiovascular risk and vascular end organ damage. *Atherosclerosis* 331, 45–53. doi:10.1016/j.atherosclerosis.2021.07.005

Sun, Y., Gao, H-Y., Fan, Z-Y., Yan, H., and Yan, Y-X. (2019). Metabolomics signatures in type 2 diabetes: A systematic review and integrative analysis. *J. Clin. Endocrinol. Metab.* 105 (4), dgz240–1008. doi:10.1210/clinem/dgz240

Thomas, S. C., Payne, D., Tamadonfar, K. O., Seymour, C. O., Jiao, J-Y., Murugapiran, S. K., et al. (2021). Position-specific metabolic probing and metagenomics of microbial communities reveal conserved central carbon metabolic network activities at high temperatures. *Front. Microbiol.* 12, 1427. doi:10.3389/fmicb.2019.01427

Trainor, P. J., DeFilippis, A. P., and Rai, S. N. (2017). Evaluation of classifier performance for multiclass phenotype discrimination in untargeted metabolomics. *Metabolites* 7 (2), E30. doi:10.3390/metabo7020030

Trevethan, R. (2017). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Front. Public Health* 5, 307. doi:10.3389/fpubh.2017.00307

Turi, K. N., Romick-Rosendale, L., Ryckman, K. K., and Hartert, T. V. (2018). A review of metabolomics approaches and their application in identifying causal pathways of childhood asthma. *J. Allergy Clin. Immunol.* 141 (4), 1191–1201. doi:10.1016/j.jaci.2017.04.021

Vapnik, V. (2006). Estimation of dependences based on empirical data. *Inf. Sci. Statistics.* doi:10.1007/0-387-34239-7

Vetter, T. R., and Schober., P. (2018). Regression: The apple does not fall far from the tree. *Anesth. Analg.* 127 (1), 277–283. doi:10.1213/ANE.0000000000003424

Vu, T., Parker, S., Bhinderwala, F., Xu, Y., and Powers, R. (2019). Evaluation of multivariate classification models for analyzing NMR metabolomics data. *J. Proteome Res.* 18 (9), 3282–3294. doi:10.1021/acs.jproteome.9b00227

Waddington, K. E., Papadaki, A., Coelewij, L., Adriani, M., Nytrova, P., et al. (2020). Artemis Papadaki, Leda Coelewij, Marsilio Adriani, Petra Nytrova, Eva Kubala Havrdova, Anna Fogdell-Hahn, et alUsing Serum Metabolomics to Predict Development of Anti-Drug Antibodies in Multiple Sclerosis Patients Treated With IFNβ. *Front. Immunol.* 11, 1527. doi:10.3389/fimmu.2020.01527

Wang, G., Qiu, M., Xing, X., Zhou, J., Yao, H., Li, M., et al. (2022). Lung cancer scRNA-seq and lipidomics reveal aberrant lipid metabolism for early-stage diagnosis. *Sci. Transl. Med.* 14 (630), eabk2756. doi:10.1126/scitranslmed.abk2756

Want, E. J., Anders, N., Morita, H., and Gary, S. (2007). From exogenous to endogenous: The inevitable imprint of mass spectrometry in metabolomics. *J. Proteome Res.* 6 (2), 459–468. doi:10.1021/pr060505+

Wishart, D. S., Tzur, D., Craig, K., and Eisner, R. (2007). An chi Guo, nelson young, dean cheng, etHMDB: The human metabolome database. *Nucleic Acids Res.* 35, D521–D526. doi:10.1093/nar/gkl923

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., et al. (2009). Hmdb: A knowledgebase for the human metabolome. *Nucleic Acids Res.* 37, D603–D610. doi:10.1093/nar/gkn810

Wishart, D. S. (2005). Metabolomics: The Principles and potential applications to transplantation. *Am. J. Transpl.* 5 (12), 2814–2820. doi:10.1111/j.1600-6143.2005.01119.x

Wold, S., Kim, E., and Paul, G. (1987). Principal component analysis. *Chemom. Intelligent Laboratory Syst.* 2 (1), 37–52. doi:10.1016/0169-7439(87)80084-9

Xue, R., Lin, Z., Deng, C., Dong, L., Liu, T., Wang, J., et al. (2008). A serum metabolomic investigation on hepatocellular carcinoma patients by chemical derivatization followed by gas chromatography/mass spectrometry. *Rapid Commun. Mass Spectrom.* 22 (19), 3061–3068. doi:10.1002/rcm.3708

Yala, A., Lehman, C., Schuster, T., Portnoi, T., and Barzilay, R. (2019). A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* 292 (1), 60–66. doi:10.1148/radiol.2019182716

Yang, L., Wang, Y., Cai, H., Wang, S., Shen, Y., and Ke, C. (2020). Application of metabolomics in the diagnosis of breast cancer: A systematic review. *J. Cancer* 11 (9), 2540–2551. doi:10.7150/jca.37604

Yu, J. S., Ongarello, S., Fiedler, R., Chen, X. W., Toffolo, G., Cobelli, C., et al. (2005). Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. *Bioinformatics* 21 (10), 2200–2209. doi:10.1093/bioinformatics/bti370

Zhang, F., Zhang, Y., Ke, C., Li, A., Wang, W., Yang, K., et al. (2018). Predicting ovarian cancer recurrence by plasma metabolic profiles before and after surgery. *Metabolomics* 14 (5), 65. doi:10.1007/s11306-018-1354-8

Zhou, M., Guan, W., Walker, L. D., Mezencev, R., Benigno, B. B., Gray, A., et al. (2010)., 19. Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, 2262–2271.Rapid mass spectrometric metabolic profiling of blood sera detects ovarian cancer with high accuracy*Cosponsored by Am. Soc. Prev. Oncol.*9

Glossary

ML Machine Learning

AI Artificial Intelligence

DM Diabetes mellitus

CVD Cardiovascular Disease

NMR Nuclear Magnetic Resonance

MS Mass Spectrometry

AUC Area Under the Curve

SVM Support Vector Machine

***k*-NN** K-Nearest Neighbor

NN Neural Networks

RF Random Forests

UMAP Uniform Manifold Approximation and Projection

t-SNE t- stochastic neighbor embedding

PCA Principle Component Analysis

ICA Independent Component Analysis

PLS Partial Least Squares

OPLS Orthogonal Projections to Latent Structures

RBF Radial Basis Function

DL Deep Learning

ANN Artificial Neural Network

CNN Convolutional Neural Networks

PAM Prediction Analysis for Microarrays

RPART Recursive Partitioning And Regression Trees

PLS-DA Partial Least Square-Discriminant Analysis

SDA Stepwise discriminant Analysis

MLP Multilayer perceptron

GTB Gradient tree boosting

CKD Chronic Kidney Disease

MDD Major Depressive Disorders

XGBoost eXtreme Gradient Boosting

PC-LDA Principle component-Linear Discriminant Analysis

LASSO Least Absolute Shrinkage and Selection Operator

PCR Principal component regression

PCLR Principal component logistic regression

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 08 December 2022
DOI 10.3389/fgene.2022.1024577

Check for updates

OPEN ACCESS

EDITED BY
Marco Pellegrini,
National Research Council (CNR), Italy

REVIEWED BY
Ahmed Moustafa,
American University in Cairo, Egypt
Daljeet Singh Dhanjal,
Lovely Professional University, India

*CORRESPONDENCE
Noelle R. Noyes,
nnoyes@umn.edu

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 21 August 2022
ACCEPTED 25 October 2022
PUBLISHED 08 December 2022

CITATION
Nayar G, Terrizzano I, Seabolt E, Agarwal A, Boucher C, Ruiz J, Slizovskiy IB, Kaufman JH and Noyes NR (2022), ggMOB: Elucidation of genomic conjugative features and associated cargo genes across bacterial genera using genus-genus mobilization networks.
Front. Genet. 13:1024577.
doi: 10.3389/fgene.2022.1024577

COPYRIGHT
© 2022 Nayar, Terrizzano, Seabolt, Agarwal, Boucher, Ruiz, Slizovskiy, Kaufman and Noyes. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

ggMOB: Elucidation of genomic conjugative features and associated cargo genes across bacterial genera using genus-genus mobilization networks

Gowri Nayar[1], Ignacio Terrizzano[2], Ed Seabolt[2], Akshay Agarwal[2], Christina Boucher[3], Jaime Ruiz[3], Ilya B. Slizovskiy[4], James H. Kaufman[5] and Noelle R. Noyes[4]*

[1]Department of Biomedical Informatics, Stanford University, Stanford, CA, United States, [2]IBM Research Almaden, San Jose, CA, United States, [3]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, United States, [4]Department of Veterinary Population Medicine, University of Minnesota, Minneapolis, MN, United States, [5]Altos Labs, Redwood City, CA, United States

Horizontal gene transfer mediated by conjugation is considered an important evolutionary mechanism of bacteria. It allows organisms to quickly evolve new phenotypic properties including antimicrobial resistance (AMR) and virulence. The frequency of conjugation-mediated cargo gene exchange has not yet been comprehensively studied within and between bacterial taxa. We developed a frequency-based network of genus-genus conjugation features and candidate cargo genes from whole-genome sequence data of over 180,000 bacterial genomes, representing 1,345 genera. Using our method, which we refer to as ggMOB, we revealed that over half of the bacterial genomes contained one or more known conjugation features that matched exactly to at least one other genome. Moreover, the proportion of genomes containing these conjugation features varied substantially by genus and conjugation feature. These results and the genus-level network structure can be viewed interactively in the ggMOB interface, which allows for user-defined filtering of conjugation features and candidate cargo genes. Using the network data, we observed that the ratio of AMR gene representation in conjugative *versus* non-conjugative genomes exceeded 5:1, confirming that conjugation is a critical force for AMR spread across genera. Finally, we demonstrated that clustering genomes by conjugation profile sometimes correlated well with classical phylogenetic structuring; but that in some cases the clustering was highly discordant, suggesting that the importance of the accessory genome in driving bacterial evolution may be highly variable across both time and taxonomy. These results can advance scientific understanding of bacterial evolution, and can be used as a starting point for probing genus-genus gene exchange within complex microbial communities that include unculturable bacteria. ggMOB is publicly available under the GNU licence at https://ruiz-hci-lab.github.io/ggMOB/

KEYWORDS

antimicobial resistance, genomics, metagenomics, big data, mobile genetic elements (MGE)

1 Introduction

Several mechanisms of horizontal gene transfer (HGT) allow bacteria to exchange genetic material. One of these mechanisms, termed conjugation, occurs when bacterial cells form direct physical contacts that allow for passage of genetic material from one bacterium to another. The machinery required to form these contacts and initiate genetic exchange is often contained within integrative and conjugative elements (ICE), plasmids, and other mobile genetic elements (MGEs) (Frost et al., 2005; Wozniak and Waldor, 2010; Roberts and Mullany, 2011; Wiedenbeck and Cohan, 2011; Perry and Wright, 2013; Johnson and Grossman, 2015; Singer et al., 2016). The conditions that induce excision and conjugation are not fully elucidated, but DNA damage and subsequent SOS response seem to be an important trigger (Waldor et al., 2004; Koraimann and Wagner, 2014). The cost of acquiring and maintaining the new genetic material also influences the success of transfer events (Uhlemann et al., 2021).

Genes exchanged between bacteria during conjugation include functional domains associated with conjugative machinery (e.g., excisionases, integrases, conjugative transport proteins) as well as intervening "accessory" genes that are not necessary for conjugation, often termed "cargo genes" (Johnson and Grossman, 2015). By pairing conjugative machinery with an array of diverse cargo genes, bacterial communities can significantly expand their genetic repertoire, including between bacteria of diverse taxonomy (Guglielmini et al., 2011; Bellanger et al., 2014; Neil and Allard, 2021). Functions commonly associated with conjugative cargo include antimicrobial resistance (AMR) and virulence (Roberts and Mullany, 2011; Perry and Wright, 2013; Johnson and Grossman, 2015; Cury et al., 2017), which can pose a risk to human and animal health if transferred into pathogens (Partridge et al., 2018). Therefore, understanding the microbial ecology of conjugative elements and cargo genes (i.e., their distribution and behavior across bacterial taxa) is important in assessing the risk posed by various bacterial communities (Gaston et al., 2021). For example, how often do different bacterial taxa carry conjugative machinery and AMR genes; what resistance phenotypes are commonly associated with the presence of conjugative machinery within the genome; how often do different commensal bacterial taxa carry out conjugation to exchange cargo genes with pathogens; and what conditions foster conjugative exchange of specific cargo genes between pathogens and non-pathogens? These questions are fundamental to understanding how bacterial communities respond to external stimuli, and how these responses increase the overall risk posed by microbial communities of varying composition (Martínez et al., 2015; Oh et al., 2018).

However, the process of conjugative HGT is highly stochastic and therefore, difficult to predict (Lopatkin and Collin, 2020). One reason for this stochasticity is variability in the conjugation competency of donor and recipient bacterial cells for a given conjugative MGE; as well as variability in the capacity of a given type of conjugative MGE to also transfer unrelated cargo genes. Recent meta-analyses of conjugation rates for specific bacterial species and/or MGEs have highlighted these complexities (Alderliesten et al., 2020; Sheppard et al., 2020). Historically, the scientific process for estimating conjugative likelihood has stemmed from highly controlled *in vitro* experiments between pairs of bacterial isolates and specific MGEs. Results from such studies have been crucial for uncovering the behavior of MGEs and their importance for functions such as AMR. However, reductive experiments typically do not generalize well to the complex microbial communities found *in situ*, including host and environmental microbiomes. Furthermore, these experiments are necessarily restricted in their ability to characterize the full bacterial host range of a given MGE, as they typically involve only several distinct bacterial taxa. One major challenge that remains is to generate a conjugation likelihood for every host-donor-MGE combination observed across all bacterial taxa and MGE.

Insight into this challenge can be gained through the plethora of whole genome sequence (WGS) data which is now publicly available. As an example, the analysis of HGT-associated genes from just 336 genomes across 16 phyla was sufficient to significantly improve bacterial phylogenies as compared to those obtained from conserved marker genes (Abby et al., 2012). An analysis of 1,000 genomes demonstrated that ICE machinery is ubiquitous across diverse prokaryotes, and likely one of the most common mechanisms of bacterial evolution (Guglielmini et al., 2011). Currently, public datasets contain orders of magnitude more WGS data, which can be used to improve our understanding of the mechanisms by which critically important genes and pathogens emerge and persist (Botelho et al., 2020). However, despite the importance of HGT in bacterial evolution and pathogenicity, there has not yet been a comprehensive, systematic survey of the frequency of conjugation and cargo genes within or between bacterial genera. The objective of this work was to describe intra- and inter-genus conjugation-cargo dynamics by leveraging the comprehensive set of WGS data and conjugation sequences currently available within the Reference Sequence (RefSeq) and Short Read Archive (SRA) databases at the National Center for Biotechnology Information (NCBI). In particular, we analyzed 186,887 WGS datasets to identify putative conjugation events and corresponding candidate cargo genes, as well as to characterize the frequency of AMR genes with respect to the frequency of their occurrence with conjugative proteins. We were

TABLE 1 Conjugative features included in this study. Color label indicates the color used to represent the conjugative feature in Figure 2. Yellow indicates a conjugative feature defined by a single IPR code and red indicates a family of codes. Genome count indicates the number of genomes that contain the conjugative feature.

Label	Conjugative feature	Genome count	Description
🔴	ICE6013	93,267	Includes IS30-like DDE transposase. More closely related to ICEBs1 than Tn5801 Smyth and Robinson. (2009)
🔴	Tn916	59,718	TetM and other resistance genes Clewell et al. (1995)
🟡	IPR025955	49,841	Type-IV secretion system protein TraC Finn et al. (2016)
🔴	ICEEc2	49,543	set of three genes encoding DNA mobility enzymes and type IV pilus Roche et al. (2010)
🟡	IPR005094	42,760	Endonuclease relaxase, MobA, VirD2 Finn et al. (2016)
🔴	ICEhin1056	42,252	Antibiotic resistance island Mohd-Zain et al. (2004)
🟡	IPR011119	28,752	Unchar. domain, putative helicase, relaxase Finn et al. (2016)
🟡	IPR014862	26,295	TrwC relaxase Finn et al. (2016)
🟡	IPR014059	23,368	Conjugative relaxase, N-terminal Finn et al. (2016)
🔴	PAPI-1	22,855	Pathogenicity island PAPI-1 of strain PA14.115 gene cluster includes virulence phenotypes Qiu et al. (2006)
🔴	pKLC102	22,460	Hybrid of plasmid and phage origin includes replication, partitioning, conjugation, pili, & integrase genes Klockgether et al. (2004)
🟡	IPR021300	22,284	Integrating conjugative element protein Finn et al. (2016)
🟡	IPR022391	21,465	Integrating conjugative element relaxase, PFGI-1 class Finn et al. (2016)
🟡	IPR022303	19,664	Conjugative transfer ATPase Finn et al. (2016)
🔴	ICEPdaSpa1	19,424	An SXT-related ICE derived; causative agent of fish pasteurellosis Osorio et al. (2008)
🟡	IPR014129	18,029	Conjugative transfer relaxase protein TraI Finn et al. (2016)
🔴	SXT	17,525	Family of conjugative-transposon-like mobile elements encoding multiple AR genes Beaber et al. (2002); Burrus et al. (2006)
🔴	ICEEc1	10,170	High-pathogenicity island (HPI); evidence for Combinatorial Transfers Paauw et al. (2010)
🔴	R391	9,916	Archetype of IncJ; carries AR, DNA repair, & mercury resistance genes Böltner et al. (2002)
🔴	ICEKp1	9,117	Resembles functional ICEEc1 Paauw et al. (2010)
🔴	ICESde3396	9,088	Carries genes predicted to be involved in virulence and resistance to various metals Smyth et al. (2014)
🔴	ICEBs1	8,504	Plasmid mobilization and putative coupling protein Lee et al. (2012)
🔴	RD2	8,370	Encodes seven putative secreted extracellular proteins Sitkiewicz et al. (2011)
🟡	IPR011952	2,640	Conserved hypothetical protein CHP02256 Finn et al. (2016)
🟡	IPR014136	2,050	Ti-type conjugative transfer relaxase TraA Finn et al. (2016)
🔴	TnGBS2	1,630	See ICE6013 Everitt et al. (2014)
🔴	CTnBST	1,520	Tyrosine recombinase family Song et al. (2007)
🔴	ICEclc	1,465	Cargo for ortho-cleavage of chlorocatechols and aminophenol metabolism (amr genes) Obi et al. (2018)
🔴	GI3	1,340	Degradation of aromatic compounds and detoxification of heavy metals Lechner et al. (2009)

(Continued on following page)

TABLE 1 (*Continued*) Conjugative features included in this study. Color label indicates the color used to represent the conjugative feature in Figure 2. Yellow indicates a conjugative feature defined by a single IPR code and red indicates a family of codes. Genome count indicates the number of genomes that contain the conjugative feature.

Label	Conjugative feature	Genome count	Description
🔴	Tn1549	648	VanB-type resistance to glycopeptides with regions Garnier et al. (2000)
🔴	CTn341	389	Encodes tetracycline resistance and its transfer is induced by tetracycline Peed et al. (2010)
🟡	IPR020369	119	Mobilisation protein B Finn et al. (2016)
			(i) excision-integration process Garnier et al. (2000)
🔴	Tn4555	79	Includes cfxA gene encoding
			broad-spectrum beta-lactamase Smith and Parker (1993)
🔴	ICESt1	26	Integrative and putative transfer functions Burrus et al. (2002); Bellanger et al. (2009)
🔴	ICEMISymR7A	16	Rhizobial symbiosis genes Ramsay and Ronson (2015)
🔴	ICESt3	14	Integrative and putative transfer functions Bellanger et al. (2009)
			(ii) vanB2 operon replaces tet(M) Garnier et al. (2000)
			(iii) Conjugative transfer Garnier et al. (2000)
🔴	Tn4371	0	Biphenyl and 4-chlorobiphenyl degradation Springael et al. (1993)

able to identify over 95,000 genomes containing conjugative proteins, and more than 4 billion putative cargo genes between genomes. We summarize and disseminate this analysis through an open-source network that describes the genus-genus sharing of conjugation features and cargo genes, representing genomes from over 1,300 different genera. Our network, which we refer to as ggMOB, allows users to filter for both conjugation features and putative cargo genes. Using ggMOB we analyzed the ratio of AMR gene representation in conjugative *versus* non-conjugative genomes and found it to be greater than 5:1, confirming that conjugation is a critical force for AMR spread across genera. Finally, we demonstrated that clustering genomes by conjugation profile sometimes correlated well with taxonomic structuring, but in some cases was highly discordant, suggesting that the importance of the accessory genome in driving bacterial evolution may be highly variable across genera. These results demonstrates that ggMOB can be used to further probe potential genus-genus mobilization dynamics, and thus, provide insight into conjugative mobilization between unculturable bacteria and complex interactions involving multiple genera.

2 Results

2.1 Overview of ggMOB

The analyses conducted in this paper were derived from an existing resource (Seabolt et al., 2020), which was constructed by curating and annotating 186,887 genomes from NCBI (see MATERIALS AND METHODS). Using the sequence data and annotated features obtained from over 166,000 curated and high-quality WGS datasets, we identified genomes that contained conjugative features (Table 1), which we term conjugative genomes (see MATERIALS AND METHODS). By recording counts of shared features across these conjugative genomes, we constructed ggMOB (the "genus-genus mobilization" network), which contains information about features that are shared between conjugative genus-genus pairs.

2.2 Inter- and intra-genus conjugation profiles

Of the 106,443 genomes that contained at least one conjugative feature, 95,781 shared at least one conjugative feature with at least one other genome in the set, indicating a common evolutionary history. These 95,781 conjugative genomes represented close to 47% (631 of the 1,345) of the genera contained in the relational database (Seabolt et al., 2020). The lack of conjugative machinery in the other 714 genera may be a false negative finding (i.e., incomplete list of conjugative features, lack of representation in the utilized NCBI databases, or lack of inclusion in genome assemblies), or could indicate inherent differences in the conjugative ability of genera across the taxonomic tree. Similarly, one might reasonably expect the number of observed conjugative genomes to scale with the number of genomes available for each genus. However, genus

TABLE 2 Proportion of genomes that contained conjugative feature(s), by genus. All genera with over 100 representative genomes are listed, in descending order by the proportion of conjugative genomes in each genus.

Genus	Number of conjugative genomes	Number of total genomes	Proportion conjugative genomes
Legionella	1,672	1,686	0.99
Shigella	5,423	5,541	0.98
Klebsiella	4,682	5,304	0.88
Elizabethkingia	102	119	0.86
Escherichia	8,140	9,957	0.82
Stenotrophomonas	441	563	0.78
Enterobacter	894	1,210	0.74
Vibrio	2,902	4,017	0.72
Acinetobacter	2,621	3,770	0.70
Pseudomonas	3,222	4,750	0.68
Enterococcus	1,003	1,516	0.66
Citrobacter	131	203	0.65
Salmonella	24,123	38,808	0.62
Clostridioides	1,329	2,183	0.61
Streptococcus	8,244	13,766	0.60
Xanthomonas	201	357	0.56
Staphylococcus	18,034	32,661	0.55
Rhizobium	110	202	0.54
Yersinia	215	437	0.49
Lactococcus	57	117	0.49
Serratia	230	619	0.37
Sinorhizobium	44	121	0.36
Bifidobacterium	137	403	0.34
Moraxella	65	192	0.34
Bacillus	471	1,471	0.32
Campylobacter	5,340	19,501	0.27
Aeromonas	80	312	0.26
Brucella	230	970	0.24
Mesorhizobium	89	385	0.23
Helicobacter	118	529	0.22
Streptomyces	74	333	0.22
Corynebacterium	133	639	0.21
Neisseria	153	781	0.20
Burkholderia	341	2,053	0.17
Haemophilus	65	403	0.16
Lactobacillus	144	962	0.15
Listeria	848	7,716	0.11
Clostridium	40	454	0.09
Mycobacterium	1,120	13,129	0.09
Cutibacterium	10	118	0.08
Bordetella	60	733	0.08
Chlamydia	33	496	0.07
Bartonella	3	124	0.02
Mycoplasma	2	251	0.01
Francisella	0	120	0.00

TABLE 3 Count and proportion of inter- and intra-genus detection of conjugative features.

Conjugative Feature	No. Inter-genus Matches	Prop. Inter-genus Matches	No. of intra-genus Matches	Prop. Intra-genus Matches
IPR014059	122,066	0.02	6,098,286	0.98
IPR014862	123,175	0.02	6,246,342	0.98
IPR005094	415,744	0.11	3,224,556	0.89
IPR014136	43	0.00	83,448	1.00
Tn916	110,710,102	0.61	71,585,305	0.39
GI3	74,775	0.20	304,550	0.80
ICEPdaSpa1	8,356,908	0.37	14,101,581	0.63
ICEclc	75,322	0.20	304,722	0.80
ICEhin1056	128,834,838	0.69	58,590,390	0.31
IPR011119	454,906	0.03	16,981,338	0.97
IPR021300	69,832	0.00	20,685,606	1.00
IPR022303	44,341	0.01	7,859,462	0.99
IPR022391	52,638	0.00	14,265,248	1.00
IPR025955	55,3687	0.02	32,073,159	0.98
SXT	17,427,985	0.58	12,525,984	0.42
CTn341	7,760	0.72	3,036	0.28
ICEEc1	1,107,736	0.18	5,010,865	0.82
ICEEc2	28,902,519	0.33	58,830,977	0.67
ICEKp1	1,629,343	0.20	6,550,799	0.80
IPR011952	50	0.00	3,154,129	1.00
IPR014129	75,786	0.02	4,557,985	0.98
R391	796,084	0.10	7,476,937	0.90
Tn4555	972	0.72	382	0.28
pKLC102	10,135	0.00	5,277,724	1.00
IPR020369	577	0.12	4,406	0.88
Tn1549	866	0.04	23,336	0.96
ICESde3396	7,659	0.00	1,676,016	1.00
CTnBST	337,318	0.63	196,675	0.37
ICEBs1	0	0.00	1,220,869	1.00
ICE6013	46,112	0.00	406,783,868	1.00
ICESt1	31	0.12	237	0.88
ICEMISymR7A	0	0.00	37	1.00
PAPI-1	0	0.00	6,101,539	1.00
ICESt3	0	0.00	97	1.00
RD2	0	0.00	1,547,322	1.00
TnGBS2	0	0.00	56,243	1.00

representation in NCBI is not uniform across genera, leading to bias in available genomes per genus. To correct for this imbalance, we computed the conjugative genome proportion by normalizing the number of observed conjugative genomes to the total number of genomes per genus (Table 2). This analysis demonstrated that the genera with the largest fraction of conjugative genomes were *not* the genera with the most genomes in NCBI. For example, although *Salmonella* had by far the greatest number of high quality genomes (N = 39,574), it ranked fifth in terms of the proportion of genomes that contained a conjugative feature. The top 30 genera listed in Table 2 all had a conjugative genome frequency greater than 20%, with genomes from the genus *Legionella* containing conjugative proteins over 99% of the time. This high percentage may have been driven by sampling bias in the available NCBI WGS datasets (for example, the *Legionella pneumophila* WGS accessions appear to have been collected from a single site), or it may represent the propensity for conjugation-mediated processes to occur within individual genera.

Next, we identified a total of 5,956 conjugation proteins across genus pairs, i.e., triples of the form (genus1, genus2, protein-name), associated with a total of 23,353,196,048 cargo protein sequences. These results are visualizable as connected nodes in the ggMOB network. This count is non-distinct by protein name as, for instance, Tyrosine recombinase XerD was found both between Salmonella-Salmonella genomes, as well as between Oligella-Proteus genomes and other genera pairs. Of these 5,956 triples, 1,680 contained genome pairs belonging to the same genus (i.e., intra-genus). Some genera were much more likely to contain genomes with conjugative features that matched to genomes from other genera, i.e., inter-genus. For example, we observed over 30 million instances of matching conjugation proteins in genomes from the *Acinetobacter* genus. Of these instances, 84% were to genomes from other genera (i.e., inter-genus), *versus* 16% from genomes within *Acinetobacter* (i.e., intra-genus). Members of the *Acinetobacter* genus are well-known for genome plasticity (Chan et al., 2015), which contributes to important phenotypes such as AMR and biofilm formation.

Genera that shared the highest number of conjugation proteins with *Acinetobacter* genomes included *Escherichia*, *Shigella*, *Vibrio*, *Salmonella*, *Pseudomonas*, *Klebsiella*, *Enterobacter* and *Citrobacter*. While this result was not unexpected given the list of known pathogens contained within these genera, our database of intra- and inter-genus exact-match conjugative features also revealed many unreported and unexpected associations. For example, genomes within the genus Nitrosomonas contained 38,034 instances of conjugative proteins, nearly 100% of which were shared with genomes from the *Shigella* genus. The specific conjugative feature involved in the vast majority of these exact matches was ICEEc2, a relatively recently discovered ICE MGE that was previously shown to transfer competently between *Salmonella enterica* serovar Typhimurium strain and into a *Yersinia pseudotuberculosis* strain. Our results suggest that ICEEc2 has a very broad host-donor range, including many genera that may not yet be described in the literature in reference to this particular ICE.

We found that the likelihood of identifying conjugative features in pairs of genomes within *versus* across genera was highly variable. Of the 36 conjugative features analyzed, 13 were only identified in pairs of genomes belonging to the same genus, i.e., intra-genus (Table 3). However, six conjugative features were more likely to be observed across genera than within genera, i.e., $>50\%$ of the observations were inter-genus (Table 3). For example, the exact same CTn341 sequence was observed in pairs of genomes a total of 10,796 times; in 72% of these instances, the pairs of genomes belonged to different genera, indicating a history of inter-genus transfer of CTn341. This conjugative feature belongs to the ICE family of MGEs and plays an important role in tetracycline resistance, and is typically associated with the genus *Bacteroides*, including in most reports related to its functionality (Bacic et al., 2005). However, we observed that 17% of the genome pairs containing exact-match CTn341 sequences belonged to the *Bacteroides* and Alistipes genera, suggesting historical transfer of this important ICE between these genera. Alistipes is an emerging genus with potential health implications (Parker et al., 2020), and the epidemiology and ecology of CTn341 within this genus warrants further investigation. Furthermore, this finding provides additional insight into the potential importance of CTn341 in spreading tetracycline resistance genes across microbial taxa.

2.3 Cargo gene profiles

To characterize the set of genes that are most likely to have been associated with conjugative HGT events, we identified all proteins that were contained in at least two conjugative genomes with 100% sequence identity. Out of 51,362,178 total unique protein sequences in the source database, 28,042 were identified as conjugation-associated proteins (i.e., conjugation machinery), and 11,276,651 were identified as candidate cargo proteins. The full set of cargo proteins mapped to 20,550 distinct names (excluding "putative protein(s)" or "hypothetical protein(s)"), with a wide range of frequencies within and between genera. Annotation of the conjugative genomes demonstrated that the vast majority of conjugative and cargo proteins were adjacent within the genome (Figure 2). Moreover, within each genome, the conjugative features were more likely to be proximate to putative cargo protein *versus* non-cargo proteins, suggesting again a common evolutionary history.

2.4 Genotypic AMR and association with conjugation features

A subset of the observed cargo protein names were associated with a set of confirmed AMR protein names. We identified this subset by selecting only those names that Prokka assigned to sequences mapping to a name defined in MEGARes v1.0 (Lakin et al., 2016), a comprehensive AMR database. The entity relations in our database ensured a 1:1 mapping between gene and protein names and their respective sequences. Of the 3,824 distinct sequences contained in MEGARes, Prokka identified 3,674 as valid sequences coding for protein. These 3,674 distinct proteins were assigned 286 distinct names, excluding "putative protein" and "hypothetical protein". These unnamed proteins comprised just 1% of the MEGARes protein set. While this highly curated set is certainly not a comprehensive list of all proteins contributing to AMR, it is an initial set to estimate the fraction of AMR proteins within the larger set of conjugation and cargo proteins.

To further investigate the microbial genomics of AMR in relation to HGT events, we used plasmid sequences from NCBI

TABLE 4 Probability of observing AMR proteins in genomes that also contained conjugative features.

AMR protein name	Number of Genomes	Number of Conjugative Genomes	Proportion Conjugative Genomes
Rob DNA-binding transcriptional activator	4,559	4,559	1.00
Transposon Tn10 TetD protein	4,559	4,559	1.00
Tetracycline resistance gene Tet(M)	441	441	1.00
Outer membrane protein YedS	288	288	1.00
Regulator of RpoS	288	288	1.00
Beta-lactamase Toho-1	319	318	1.00
AcrAD-TolC permease subunit	1,584	1,576	0.99
Multidrug efflux pump subunit AcrB	1,584	1,576	0.99
DNA topoisomerase subunit A	1,091	1,085	0.99
MATE family multidrug efflux pump protein	1,607	1,596	0.99
Carbapenem-hydrolyzing beta-lactamase KPC	1,914	1,899	0.99
Beta-lactamase OXA-1	1,188	1,175	0.99
Inner membrane protein HsrA	350	346	0.99
Putative transport protein YdhC	342	338	0.99
Aclacinomycin methylesterase RdmC	419	414	0.99
Beta-lactamase OXA-2	199	196	0.98
Chloramphenicol efflux MFS transporter CmlA1	795	783	0.98
Beta-lactamase OXA-10	521	513	0.98
armA*	994	978	0.98
rRNA large subunit methyltransferase H	4,087	3,989	0.98
Multidrug resistance operon repressor	486	67	0.14
Outer membrane protein OprM	478	65	0.14
Methicillin-resistance regulatory protein MecR1	8,344	1,007	0.12
Phosphoethanolamine-lipid A transferase	8,344	1,007	0.12
HTH-type transcriptional repressor BepR	467	55	0.12
Bifunctional polymyxin resistance protein ArnA	507	53	0.10
Methicillin resistance regulatory protein MecI	6,583	395	0.06
Metallothiol transferase FosB	6,584	395	0.06
Multidrug efflux transporter MdtL	419	9	0.02
Multidrug efflux pump subunit AcrA	420	7	0.02
HTH-type transcriptional regulator SyrM 1	414	3	0.01
RND transporter permease subunit OqxB3	358	2	0.01
Aminoglycoside 2′-N-acetyltransferase	9,723	5	0.00
DNA-binding response regulator MtrA	9,854	2	0.00
Putative acetyltransferase	5,232	1	0.00
Quinolone resistance protein NorB	213	0	0.00

Only AMR proteins that appeared in more than 100 genomes were considered; and only AMR proteins that occurred in conjugative genomes with a probability ≥ 0.98 or ≤ 0.15 are listed in this table. Full data available in the Supplementary Table S3. *Full protein name: 16S rRNA (guanine(1405)-N(7))-methyltransferase.

to identify the set of AMR-specific cargo proteins found in a plasmid sequence. We then compared the distribution of AMR proteins in plasmids *versus* conjugative genomes (Figure 3). The distribution of AMR genes differed between plasmids and conjugative genomes, with a high probability that AMR genes were identified in conjugative (*versus* non-conjugative) genomes, and a much lower probability of being identified in plasmids (Figure 3). While plasmids are often critical to the microbial ecology of AMR, this analysis suggests that other conjugative processes may drive the vast majority of AMR gene exchanges between bacteria. This dynamic has been reported for specific AMR gene groups, including carbapenem AMR genes (Botelho et al., 2020), and also supports previous analyses of conjugative machinery across bacterial genera (Guglielmini et al., 2011).

To analyze the distribution of AMR proteins across genomes, we calculated the probability that each AMR gene was identified

TABLE 5 Associations between phenotypic AMR and representation of conjugative *versus* non-conjugative genomes.

Antibiotic	Number of Resistant Genomes	Number of Resistant Conjugative Genomes	Expected Number of Resistant Conjugative Genomes*	Observed Proportion Conjugative Genomes
doripenem	215	207	86 ± 7	0.96
cefepime	272	259	108 ± 8	0.95
ampicillin-sulbactam	433	402	173 ± 10	0.93
imipenem	429	396	171 ± 11	0.92
piperacillin-tazobactam	280	258	112 ± 9	0.92
meropenem	402	367	161 ± 11	0.91
trimethoprim-sulfamethoxazole	868	775	348 ± 14	0.89
ertapenem	222	197	89 ± 8	0.89
levofloxacin	741	644	297 ± 14	0.87
gentamicin	813	706	325 ± 13	0.87
ciprofloxacin	915	794	367 ± 14	0.87
amoxicillin-clavulanic acid	277	240	111 ± 8	0.87
ceftriaxone	984	849	394 ± 16	0.86
tetracycline	700	603	281 ± 12	0.86
ceftazidime	852	731	342 ± 14	0.86
cefotaxime	889	758	355 ± 15	0.85
tobramycin	674	574	270 ± 12	0.85
ampicillin	1042	852	418 ± 16	0.82
amikacin	383	313	154 ± 10	0.82
aztreonam	989	805	397 ± 15	0.81
cefazolin	1002	813	402 ± 16	0.81
cefoxitin	757	608	304 ± 15	0.8
nitrofurantoin	713	559	287 ± 12	0.78
cefotetan	124	95	49 ± 6	0.77

*Only drug compounds with 100 or more non-redundant resistant genome measurements and >76% representation in conjugative genomes are listed. The expected number of conjugative genomes was estimated based on a bootstrapped random selection process with 100 trials (null hypothesis), using the number of assays and the actual fraction of genomes with conjugative features (51%).

in conjugative *versus* non-conjugative genomes (Table 4). From the complete list of 286 AMR protein names, 220 were found more often in conjugative genomes; three were found with equal probability in conjugative and non-conjugative genomes; and 63 were found more often in genomes that did not contain any conjugation proteins. Given that the proportion of conjugative genomes in our database was 51%, these results suggest that AMR genes are disproportionately represented within conjugative genomes.

The above analysis does not distinguish between different types of conjugation features, and also treats AMR proteins as independent features. However, the data in Figure 2 demonstrate that many of the conjugation features defined in Table 1 often co-occur in the same genome, as do some of the AMR proteins. To gain insight into these correlations, and to identify groups of AMR proteins associated with different conjugation families, we performed a genomic co-occurrence analysis across all conjugation features for the 138 AMR proteins found in conjugative genomes with a frequency of at least 5 times compared to identification in non-conjugative genomes (i.e., these AMR proteins were highly represented in conjugative genomes, Figure 4). The results of this analysis demonstrated that some conjugation features frequently co-occurred within genomes with both other conjugation features as well as multiple AMR proteins. For example, IPR005094 co-occurred with the highest diversity of AMR protein names (N = 139, see Supplementary Table S2). Many co-occurrence patterns reflected known biological associations. For example, the Tn916 conjugation feature co-occurred most frequently with tetracycline ribosomal protection protein TetM, a genomic association discovered over 3 decades ago (Su et al., 1992). While TetM seemed to co-occur with a few select conjugation features (such as Tn916), other AMR protein names co-occurred with many conjugation features. For example, many of the AMR names associated with extended-spectrum beta-lactam and carbapenem resistance (e.g., beta-lactamases Toho-1, OXA-1, OXa-2, OXA-10, SHV-2, and KPC) co-occured with the majority of evaluated conjugation features, which may provide partial explanation for the observed rapid expansion of these important AMR genes within Enterobacteriaceae (Logan and Weinstein, 2017). Similarly, the recently widely-publicized mcr-1 protein co-occurred with multiple conjugation features, which both strengthens and expands upon recent findings that this AMR gene has been mobilized on numerous plasmid types (Wang et al., 2018). Co-occurrence data such as those provided in Figure 4 may represent a new and sustainable (i.e., easily updated) source of information regarding the potential for new and emerging AMR genes to expand within and across bacterial populations *via* HGT. This information, in turn, could help to prioritize and focus public health and human clinical decision-making regarding AMR.

Conversely, 29 AMR proteins occurred at least 5x more frequently in non-conjugative genomes compared to conjugative genomes (Supplementary Figure S1). Of note is the observation that no beta-lactam AMR protein names occur in this list of 29 AMR names, which contrasts starkly to the preponderance of beta-lactam-associated AMR names in Figure 4, again suggesting that beta-lactam resistance is tightly coupled with conjugative machinery, and that conjugation-mediated exchange is the primary evolutionary driver of beta-lactam resistance. By comparison, several mechanisms of multi-drug resistance (MDR) are contained within the list of 29 AMR proteins observed more frequently in non-conjugative *versus* conjugative genomes, i.e., AcrB, AcrE, OqxB7, mdtA, mdtE, mdtH, and MexB. These mechanisms of MDR tend to be multi-function, i.e., the proteins confer multiple functional benefits to bacteria, in addition to AMR. Together, the results of Figure 4 and Supplementary Figure S1 suggest that proteins with more specific AMR functions tend to be disproportionately represented amongst conjugative genomes, while more generalist proteins tend to be disproportionately represented within non-conjugative genomes. One hypothesis for this observation is that the fitness cost-benefit dynamics differ for generalist *versus* specialist genes, such that specialized genes are more likely to transiently yet rapidly spread within bacterial populations *via* the so-called ‘accessory genome’ (which includes conjugation-mediated exchange), whereas generalist genes are more likely to be maintained permanently within bacterial genomes, and thus are less likely to be identified as conjugation-associated cargo.

2.5 Phenotypic AMR and association with conjugative genomes

Given our hypothesis that conjugation-mediated spread of specialized AMR genes may be promoted by more specific evolutionary pressures such as antimicrobial drug exposures, we hypothesized that this signature of selective pressure may also manifest in the phenotypic properties of conjugative *versus* non-conjugative genomes. To evaluate this, we queried the NCBI BioSample assay metadata in our relational database, to identify isolates that had been subjected to phenotypic antibiotic susceptibility testing (AST) to known antibiotic compounds. For the 186,887 highest quality genomes, the NCBI assay metadata contained 15,286 phenotypically-confirmed resistant genome-compound AST results, representing 13,076 tests for conjugative genomes and 2,210 tests for non-conjugative genomes. Altogether, 1,242 genomes were used in these tests, of which 1,023 were conjugative genomes and 219 were non-conjugative genomes. For each antibiotic compound listed in the AST results, we computed the number of phenotypically resistant isolates with conjugative-vs. non-conjugative genomes.

The results of this analysis revealed that phenotypic resistance occurred in conjugative genomes with probability >80%, regardless of compound being tested (Table 5 and Supplementary Table S3). As with the disproportionate representation of AMR proteins within conjugative genomes, the phenotypic AMR data suggest that microbial AMR dynamics are driven largely by conjugation-mediated processes. However, it is also important to note that NCBI phenotypic assay data is likely biased due to the motivations for clinicians and researchers to submit isolates for phenotypic testing. Therefore, to test for SRA sampling bias with respect to these compounds, we also measured the phenotypically-resistant fraction expected for *randomly selected* genomes, based on the number of genomes tested per compound in Table 5 and the number of conjugative and non-conjugative genomes across the entire database. This null hypothesis was tested by running 100 bootstrapped trials for each compound (Table 3). The observed average probability that phenotypic AMR was expressed by an isolate with a conjugative genome was 0.85 ± 0.05 *independent of antibiotic compound*, i.e., weighted by total genomes tested per compound (Table 5). In a random process, the probability would be expected to be near 51% given the fraction of all genomes with conjugation features. These results further support the importance of conjugation in the microbial ecology and epidemiology of both genotypic and phenotypic AMR.

2.6 Genome clustering by conjugative feature profile

The specific proteins transferred between bacteria are known to vary by conjugation feature, for example as demonstrated by the co-occurrence patterns of conjugation features and AMR proteins within genomes (Figure 4 and Supplementary Figure S1). To demonstrate the structuring of bacterial populations by conjugation-cargo co-occurrence patterns, we generated the same co-occurrence matrix for all cargo proteins and all conjugative features within genomes that contained a high frequency of cargo proteins (Figure 5). We then calculated the genome-genome Euclidean distance for all genomes in the co-occurrence matrix, using a vector of normalized conjugation features (see *Methods*, Supplementary Figure S2). The results demonstrate clusters of genomes with similar conjugative features. While these clusters often reflect classical bacterial taxonomic structure, there are also instances of discordance between the clustering based on conjugation profile and traditional grouping based on taxonomy. These results reflect the evolutionary dynamics of bacterial populations, and suggest that the relative importance of vertical *versus* horizontal gene transfer events is highly variable. Such variability in the importance of HGT events has been previously reported, including differences in plasmid *versus* ICE-mediated exchange and interactions with bacterial host range (Cury et al., 2018). This finding has far-reaching and complex implications for applications that rely on a measure of phylogenetic relatedness, i.e., outbreak detection and source attribution. In some cases the use of the core genome may be sufficient to accurately reflect bacterial phylogenetic relatedness; while in others, the information in the core genome may obfuscate the true relationships between bacterial isolates. As WGS data become more widely used, these complexities must be considered, and in some cases, incorporated into phylogenetic analysis workflows.

The relational database underlying this work and the ggMOB tool is a necessary (yet not sufficient) component of improved interpretation and use of WGS data. We note that, in principle, one can use the co-occurrence matrix and vectorization procedure on genomic properties other than conjugative features. This provides a flexible and customizable approach for defining a different set of genomes of interest within the *mobilome*, which can then be used to (re)classify organisms not by name, but by genome-genome distance in a space of mobilization features. This capability could represent a powerful tool for improving our understanding of bacterial evolution while also informing the next generation of applied WGS computational and statistical pipelines.

3 Discussion

The public availability of large scale genomic data makes it possible to apply cloud computing technology and big data techniques to study important phenomena in molecular and microbiology. Curating these data in a relational database with biologically structured entity relations (i.e., linking genomes, genes, proteins, domains, and metadata) provides a powerful method with which to ask biological questions about the data. We leveraged this approach in the current study of cargo and conjugation, which is an essential mechanism by which bacteria acquire new phenotypes, transmit molecular functions, and adapt to stress. Furthermore, these events are critical for understanding bacterial evolution and phylogeny (Guglielmini et al., 2011; Abby et al., 2012; Bellanger et al., 2014). Our work not only sheds light on conjugation-mediated cargo transfers between and within genera, but also demonstrates the ability of mining and analyzing large datasets in improving our understanding of bacterial evolutionary dynamics. The network of putative genus-genus conjugation features and candidate cargo genes can be dynamically visualized using the ggMOB tool, which supports hypothesis generation and testing related to intra- and inter-genus conjugation dynamics.

In our analysis we identified sets of proteins with the strongest evidence as conjugation and cargo proteins. This was accomplished by selecting only those proteins that exhibited both 100% sequence identity and co-occurrence in pairs of genomes containing identical conjugation-associated sequences. With this strict selection process, the putative cargo proteins exhibited a high degree of spatial correlation within assembled contigs (i.e., they were highly adjacent to each other, as well as to the conjugation protein itself). Other proteins in these genomes may also have been transferred (or are transferable) by bacterial conjugation, but they did not meet our strict selection criteria. Considering only strictly-selected candidate cargo proteins, we were able to profile the frequency of conjugation-mediated protein exchange within and between genera.

Our results suggest that conjugation-mediated exchange is not uncommon, affirming prior studies (Guglielmini et al., 2011; Bellanger et al., 2014). Conjugation-related proteins were observable in 51% of bacterial genomes and in 631 of 1,345 genera (approximately 47%). Frequency of intra- and inter-genus conjugation-mediated exchange varied significantly depending on the taxa involved, suggesting that taxonomy greatly influences genetic exchange of, e.g., AMR or pathogenicity proteins (Delavat et al., 2017). By quantifying this across a large database of high-quality WGS data, we measured the "exchange likelihood" between different genera. These likelihoods can be visualized dynamically in the ggMOB tool, which reveals distinct clusters of genera that share conjugative features with exact sequence match. This suggests that the likelihood of protein transfer varies substantially by genus pair, and that the bacterial composition within a given environment is an important consideration when attempting to evaluate mobilization potential within a microbial community (Lopatkin and Collin, 2020; Neil and Allard, 2021).

While we have conducted this analysis for a specific set of conjugation features (Table 1), the analytic approach can be applied to any MGE(s) and cargo protein(s) of interest. As such, our overall approach represents a method for obtaining a long-range evolutionary view of transfer likelihood between diverse bacterial taxa, including pathogens and commensal bacteria (Guglielmini et al., 2011). These baseline exchange likelihoods are critical parameters for risk analysis at the microbial community level, including for applications such as personalized microbiome medicine, and microbiome-centric surveillance.

Bacterial taxon is not the only significant driver of exchange likelihood; we also observed that putative, successful transfer events were more likely to involve cargo proteins that infer fitness advantage to the involved bacterial populations, such as AMR. While any gene can, in principle, be transmitted as a cargo gene in conjugative exchange, only a subset of transferred proteins will increase the fitness of the receiving organism. The likelihood of observing successful transfer depends on a large number of factors including the environment, the existing proteins in the recipient chromosome, the cargo proteins themselves, and the survival probability of the organism (Cohen et al., 2010).

Conjugation-mediated protein transfer that improves fitness may increase survival probability. Therefore, chromosomal arrangements that group *fitness-conferring* cargo proteins near the conjugation machinery will be observed more frequently than those arrangements that involve neutral or disadvantageous proteins. Conversely, very common proteins that aid in stress response may be less likely to be transferred as cargo, since the relative fitness advantage is diminished for proteins that are already likely to be present within a bacterium (i.e., proteins that confer redundant function). The particular stressor—as well as the specific advantageous proteins of interest—depend on phenotype of interest. This view is exemplified by the data in Table 4 which shows that rare AMR proteins are more likely to be found as cargo in genomes that also contain conjugative proteins, as compared to genomes that do not. Conversely, common AMR proteins are less likely to be found in genomes that contain conjugative machinery. One might hypothesize that, with chromosomal rearrangement, nature effects a real world experiment to dynamically optimize cargo protein collections—thereby spreading rare (but useful) proteins and gene combinations over time.

The particular cargo proteins shared between chromosomes varied by conjugation feature, as demonstrated by the AMR proteins analyzed in Figures 1, 4. Considering all conjugation features used in this study, our results suggest that conjugation dynamics are important in structuring genomic content, and thus driving phylogenetic evolution. Based on Figure 5, it seems that sometimes these evolutionary conjugation dynamics can sometimes overpower other taxonomic drivers, such that genus-level genomes do not always cluster together. To demonstrate this conjugation-driven phylogeny, we used the data in Figure 5 to generate Figure 2, which represents the distance between all pairs of genomes based on Euclidean distance between their representations as normalized conjugation feature vectors. The resulting hierarchical clustering shows that the dominant conjugation features are represented in genomes across different genera and, conversely, that individual genera include genomes with differing conjugation profiles. This abrogation of genus-level taxonomy due to conjugation-related genomic content is an inevitable consequence of the inter-genus transfers visualized in ggMOB. Given the reality of conjugative exchange, there is no reason to expect that taxonomic classification by organism name will always predict the composition of conjugation-associated cargo proteins. However, by selecting genomes based on a particular phenotype of interest, it is possible to classify organisms and genome-genome distances based on a feature space defined by conjugation (or other mobilization) proteins, as in Figure 2. Given the ubiquity and diversity of conjugation and other types of HGT (Guglielmini et al., 2011), these types of genome clustering techniques may provide crucial information about bacterial evolution that is not contained within traditional phylogenies. In this regard, the ability to

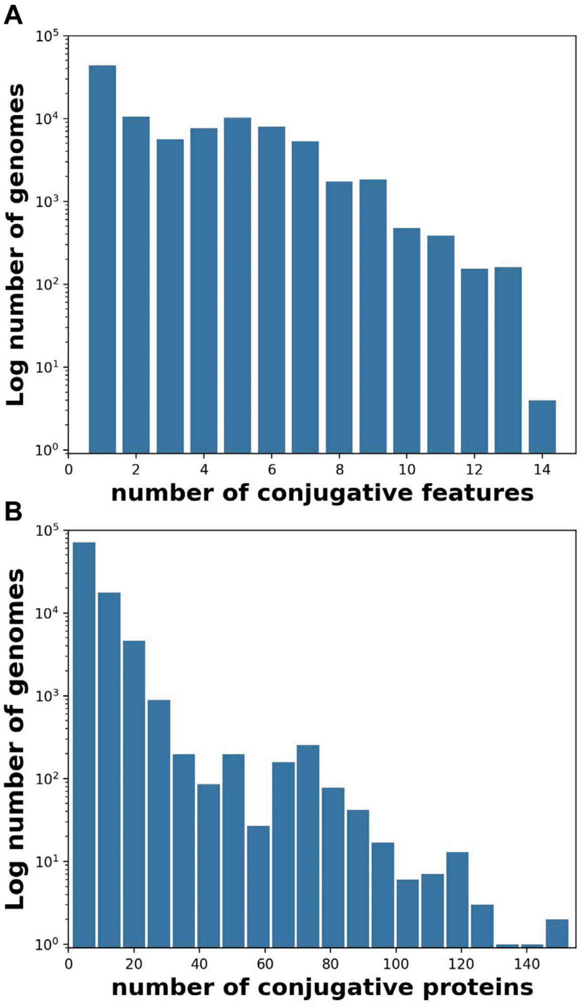


FIGURE 1
Individual genomes typically contained more than one conjugative feature (Table 1), and often contained more than one protein per feature. **(A)** Histogram (log frequency) of the number of conjugative features per genome, and **(B)** Histogram (log frequency) of the number of conjugative proteins per genome for all 106,433 genomes containing at least one conjugative feature.

filter the ggMOB data based on conjugation features of interest may particularly useful.

4 Materials and methods

4.1 Creation of relational database

Here we briefly describe the process used to create the relational database that underlies ggMOB; further details can be found at (Seabolt et al., 2020). First, we downloaded whole genome sequences from NCBI's Reference Sequence Database (RefSeq), which we then filtered to only obtain genomes that were identified as being of bacterial lineage, and as having an assembly level of "Complete Genome". We added to this set of genomes non-assembled sequence data from NCBI's Short Read Archive (SRA) by first downloading all datasets that had the following criteria: (1) the data consisted of WGS data generated from bacteria, as defined according to their taxonomic lineage; (2) the data were Illumina short-read sequence data from DNA; and (3) the sequence data were paired-end. Long-read and transcriptomic data were not considered. We note that we downloadeded all the SRA data in FASTQ format using the SRA toolkit (Sugawara and Shumway, 2010), and assembled them into contigs using SPAdes (Bankevich et al., 2012). We discarded any genome assemblies that contained more than 150 contigs (of size >500 bp) or had an N50 of less than 100 kbp. We note that only 48% of bacterial genomes met the aforementioned curation thresholds from the original corpus of SRA datasets. Next, we eliminated any assembled genomes in which a significant proportion of *k*-mers originated from multiple genomes across different genera. This removed another 13,044 genomes from consideration. This last step ensured all the genomes used for analysis were from a pure single bacterial isolate with valid genus-level classification, hence minimizing the probability of contamination. We obtained a total of 159,628 genome assemblies after filtering for all the above criteria, and a total of 186,887 genomes when including the genome assemblies from RefSeq.

Next, we annotated all the genomes using Prokka 1.12 (Prokka, 2014), resulting in the collation of the genome, gene, and protein data entities into CSV files. After annotation, we determined the protein domains using InterProScan 5.28–67.0 (Quevillon et al., 2005) with all available analyses provided by InterProScan (16). This resulted in 16 JSON files that were then parsed to create a set of CSV files. The annotation process yielded a total of 66,945,714 unique gene sequences; 51,362,178 unique protein sequences; and 138,327,556 unique protein domains along with related functional annotations.

Using the curated data, we created a relational database using IBM's DB2 system, which contained the following five different entity types: genomes, genes, proteins, protein domains, and functional annotations. These entities were determined by the above data curation, assembly and annotation steps; each entity in the database was stored using the MD5 hash of the sequence itself to create a unique identifier. Thus, we can quickly query for an entity within the database using the unique identifiers as a key. We stored the relations between entities as tables in the relational database, e.g., the genes corresponding to a particular genome in a table. We note this saves storage because unique sequences are stored only once in their respective tables and the tables point to all the parent entities in which they are found. Thus, while database construction required 1468 CPUs, 6TB RAM, and 160 TB of hard drive space, the final relational database scales efficiently with the addition of new sequences (Seabolt et al., 2020).

4.2 Creation of ggMOB

After curation and annotation of the data, we identified all candidate conjugative and cargo proteins in order to create ggMOB. In particular, we used both the primary literature and the InterProScan coding system to generate a list of conjugative features for ggMOB. This led us to consider the following 12 InterProScan codes:

1. IPR005094 describes relaxases and mobilisation proteins, as exemplified by MobA/VirD2 (Pansegrau et al., 1993; Byrd and Matson, 1997; Quevillon et al., 2005).
2. IPR011119 represents a domain found in Proteobacteria annotated as helicase, conjugative relaxase or nickase (Street et al., 2003; Quevillon et al., 2005).
3. IPR014059 codes for a domain in the N-terminal region of a relaxase-helicase (TrwC) that acts in plasmid R388 conjugation. It has been associated with both DNA cleavage and strand transfer activities, and members of this family are frequently found in genomic proximity to conjugative proteins thought to indicate the presence of integrated plasmids when identified in bacterial chromosomes (Quevillon et al., 2005; Boer et al., 2006).
4. IPR014129 represents proteins in the relaxosome complex, exemplified by TraI, which mediates the single-strand nicking and ATP-dependent unwinding of the plasmid molecule *via* two separate domains in the protein (Matson and Ragonese, 2005; Quevillon et al., 2005).
5. IPR014862 represents a conserved domain found in the relaxosome complex, as exemplified by TrwC (Quevillon et al., 2005; Boer et al., 2006).

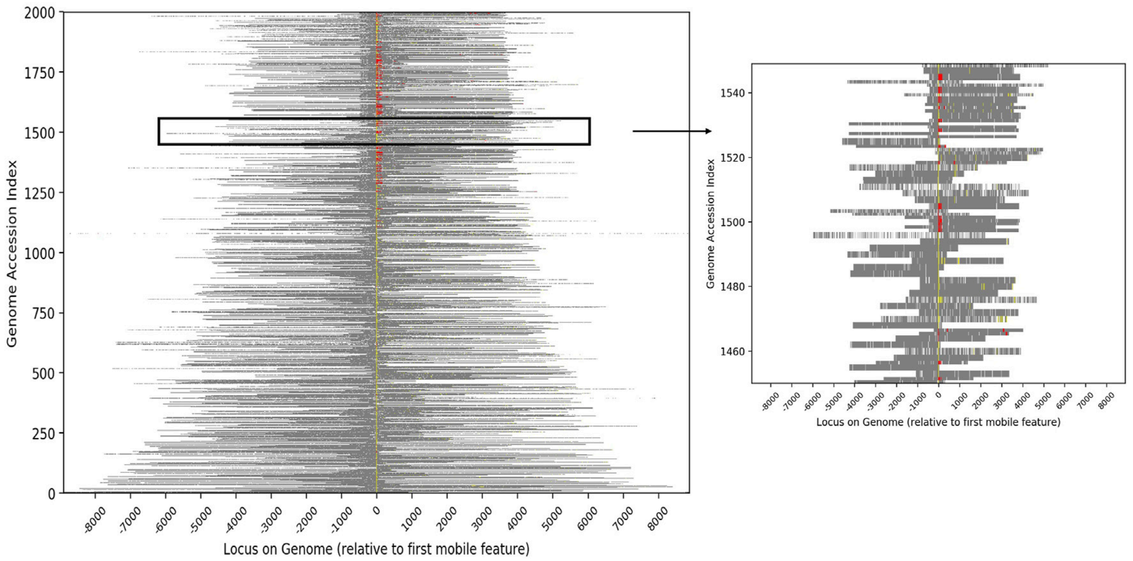


FIGURE 2
Heatmap showing the relative genomic positions of conjugative features and putative cargo proteins for the 2,000 genomes with the greatest number of cargo proteins. Conjugative features are represented as color pixels based on the colors shown in Table 1, with yellow representing proteins assigned specific IPR codes and red representing protein families from the literature. Cargo proteins are shown in grey and other chromosomal DNA in white. Each genome is bit shifted to the left until the first conjugative feature is centered in the figure. Most genomes contained more than one conjugative protein (all contain at least one). The inset highlights the genomes at indices 1450–1550 in order to expand a subset of the data.

6. IPR021300 represents a conserved domain observed in ICE elements in the protein family PFL_4695, and originally identified in *Pseudomonas fluorescens* Pf-5 (Quevillon et al., 2005; Mavrodi et al., 2009).
7. IPR022303 describes a family of conjugative transfer ATPases representing predicted ATP-binding proteins associated with DNA conjugative transfer. They are found both in plasmids and bacterial chromosomal regions that appear to derive from integrative elements such as conjugative transposons.
8. IPR025955 describes a family of TraC-related proteins observed in Proteobacteria. TraC is a cytoplasmic, membrane protein encoded by the F transfer region of conjugative plasmids, and is required for the assembly of the F pilus structure, which creates and maintains contact between the donor and recipient cells during conjugation. The family includes predicted ATPases associated with DNA conjugative transfer (Schandel et al., 1992; Quevillon et al., 2005).
9. IPR022391 represents the N-terminal domain of proteins associated with conjugative relaxases in the PFGI-1 class, which includes TraI putative relaxases required for ICE function. While these relaxases are similar in function to TraI relaxases of the F plasmid, they have no sequence homology (Quevillon et al., 2005).
10. IPR011952 represents CD-NTase-associated protein 3, a group of proteins that function as part of CBASS (cyclic oligonucleotide-based antiphage signaling system), which provides immunity against bacteriophages (Millman et al., 2020).
11. IPR014136 encompasses TraA, a Ti-type conjugative transfer relaxase that likely nicks the OriT site and unwinds the coiled plasmid prior to conjugative transfer (similar to TraI(F) in this respect) (Harris et al., 1999).
12. IPR020369 represents mobilization protein B (MobB), which is thought to play a role in conjugative exchange by presenting MobA and its covalently-linked plasmid DNA to the conjugative pore for subsequent export (Meyer, 2011).

We supplemented the IPR features with additional conjugative features that provide essential functions in the conjugation process, including conjugative relaxases, nickases, helicases, and other mobilisation proteins (Schandel et al., 1992; Pansegrau et al., 1993; Byrd and Matson, 1997; Boer et al., 2006) (Table 1). These conjugative features contain substantial sequence diversity, (Frost et al., 2005; Wozniak and Waldor, 2010; Perry and Wright, 2013; Johnson and Grossman, 2015; Singer et al., 2016), but also represent conserved domains involved in the machinery required for conjugative transfer. Using standard SQL queries, we obtained a list of the unique identifiers that have one or more of the features described above, resulting in a list of 28,042 candidate conjugation proteins. From this list, we removed those that appeared exactly in only a single genome, which further reduced the list of potential conjugation proteins to 15,398 across 95,781 genomes. We refer to this resulting set of genomes as *conjugative genomes* because they contain putative conjugation features (Supplementary Table S4). We note that these genomes represent 51% of all genomes in the relational database.

Next, we performed SQL queries in order to identify proteins most likely to be cargo based on evidence of conjugative transfer. To minimize misclassification of proteins as cargo, we applied two rules: (1) the proteins had to be present in at least two genomes with the same conjugative protein; and (2) they could not be present in any non-conjugative genome. To accomplish the query, we first queried the database for all proteins in the 95,781 conjugative genomes, which produced a list of 387,682,038 distinct < conjugative genome accession number, protein > tuples. In many cases a unique sequence was observed in more than one genome, and therefore, in total there were 21,207,794 distinct protein sequences in the set of conjugative genomes. Next, we filtered this list in order to identify the set of proteins that appeared in two or more conjugative genomes. To further reduce false positive identification of transfer by conjugation (vs. being vertically transferred), we discarded any protein that appeared in any of the 99,052 non-conjugative genomes. With this strict selection process, we identified 11,276,651 distinct sequences that we refer to as *cargo proteins*, i.e., proteins with the greatest evidence of conjugative transfer. Lastly, we tabulated these results to produce a list of triples of the form < cargo protein, conjugative genome A, conjugative genome B> which describe genome A and genome B contained at least one identical conjugative protein sequence, yielding a total of 4,938,737,476 putative transfers. We used this file as input to a custom python script that identifies all intra- and inter-genera cargo protein transfers for each protein by comparing all pairs of genomes in order to identify the intersection of conjugative proteins of each pair of genomes, and the cargo proteins (if any) in this intersection. The output of this script was used to create the ggMOB network, which contains a node for each genus, and an edge between any pair of nodes in which the value of co-occurrence of <conjugative protein, cargo protein> is non-empty (see https://github.com/Ruiz-HCI-Lab/ggMOB for source files and code).

4.3 Additional analyses

4.3.1 Conjugation and cargo gene proximity

To characterize the genomic proximity of conjugative and cargo genes within the conjugative genome pairs, we used our compiled list of genera, genomes, conjugative and cargo proteins, along with Prokka's accession index, which indicates the position of a gene or protein sequence within the assembled sequences. While this approach was limited by the fact that the order of assembled sequences is unknown, the Prokka index does indicate position of annotated sequences within each assembled sequence; this information was used in a visual display of genomic distance (Figure 2).

4.3.2 Characterizing AMR genes

We identified all AMR proteins by querying our relational database with all sequences in the MEGARes database (Lakin et al., 2016). To obtain consistent annotations, we annotated the sequences in MEGARes in the same format as was used to annotate the set of all proteins in the database. We note that we were able to maintain high confidence that these annotations represent AMR proteins because the annotations were derived by exact sequence matching. Next, we extracted the set of cargo proteins with (self-consistent) names that matched any MEGARes AMR protein name. These data were then used to compute the frequency of observing each AMR gene in genomes that contained and did not contain conjugative features.

4.3.3 NCBI antibiotic susceptibility testing BioSample data

To analyze associations between phenotypic AMR and genomic conjugation features, we retrieved metadata for each NCBI accession that contained antimicrobial susceptibility testing that contained AST data, which include genomic accession number for each isolate, the antibiotic compound(s) against which it was tested, the AST type, and the phenotypic outcome (resistant, susceptible, or intermediate). We only considered those isolates with a resistant phenotypic outcome to be resistant. By linking the BioSample accession with the SRA accessions in our relational database, we were able to identify genomes for which corresponding AST data were available. These genomes were used in our analysis of phenotyic AMR and conjugative features.

4.3.4 Classification of plasmid and conjugation proteins

We note that many of the InterPro codes listed in Table 1 contain conjugative machinery found in both plasmids and ICEs. Since these two groups of MGEs exhibit unique microbial ecological and epidemiological dynamics regarding AMR, we attempted to further annotate conjugative proteins identified within our set of genomes. To accomplish this, we first queried NCBI for bacterial plasmids. All of these assembled bacterial plasmids were

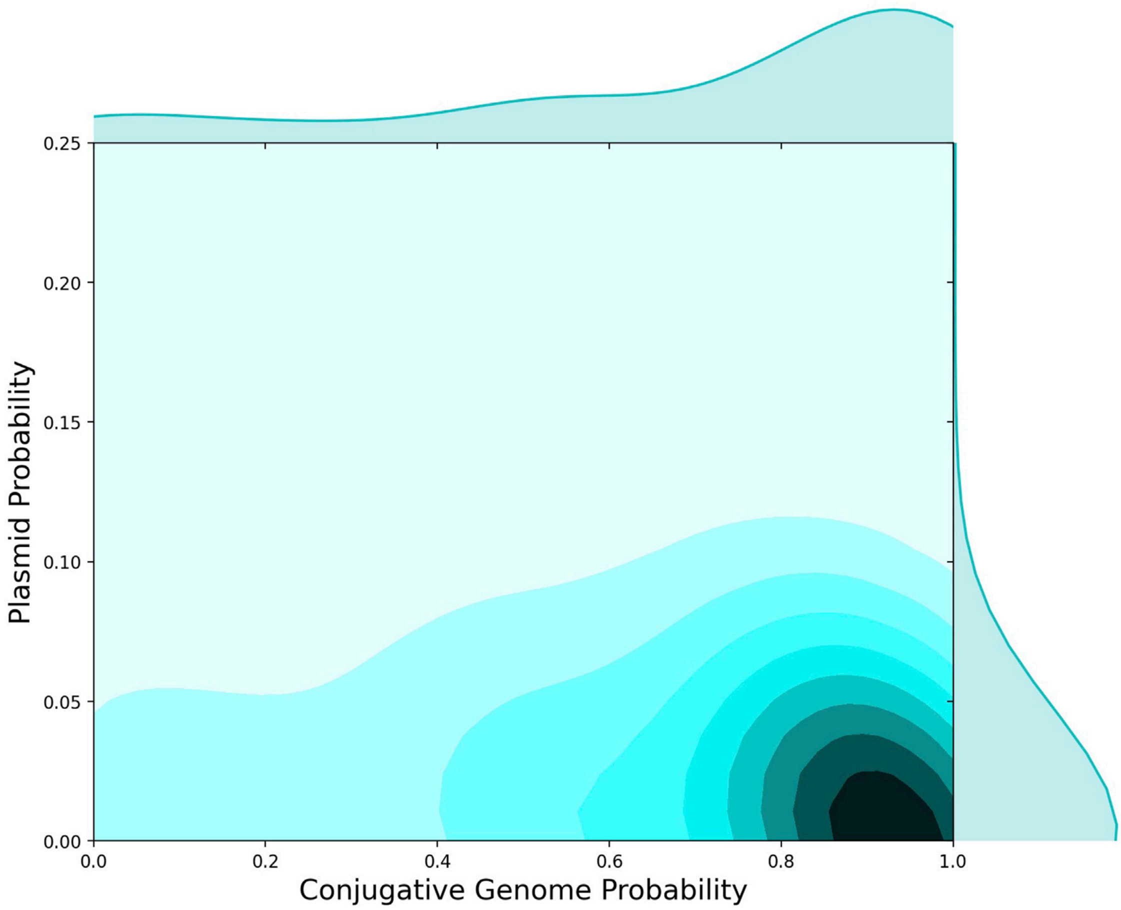


FIGURE 3
AMR Protein Detection in Plasmids versus Conjugative Genomes. 2D histogram showing the probability that AMR proteins were found in conjugative genomes (*x*-axis) *versus* the fraction of AMR proteins observed on plasmids (*y*-axis). Independent of presence on conjugative genomes, 5%–10% of AMR proteins were observed on plasmids, whereas the majority of AMR proteins were found in conjugative genomes.

downloaded from NCBI and annotated *via* our Prokka and InterProScan pipelines. The annotated plasmid associated proteins were then placed in a database table. The MD5 hash was used as the primary key for entries in this table, as was true for every sequence entity in the FGP database. Determining whether a protein had been observed on a plasmid in any of the reference genomes was then accomplished by querying for the primary (i.e., FGP) key of each protein within the table of plasmid-associated proteins. If the primary key existed in both tables, we considered that protein to be a plasmid-associated protein. These data were used in the analyses shown in Figure 3.

4.3.5 Tabulation of data

Conjugative genomes were tabulated by genus as shown in Table 2. Observed conjugative features from Table 1 were tabulated by genome for the analysis shown in Figures 1, 5. Similarly, proteins assigned AMR-associated annotations were tabulated by number of conjugative and non-conjugative genomes, and the fraction of unique sequences observed in plasmids was tabulated as well (See Figure 4).

4.3.6 Hierarchical clustering and co-occurrence

Hierarchical clustering was used to characterize the co-occurrence of proteins or genomes by conjugation feature (Figures 1, 4, 5. The co-occurrence matrix was generated using the Seaborn `clustermap` algorithm, which performs single linkage clustering to generate heatmaps and dendrograms (Waskom, 2015). The vector of features used to generate the heatmap shown in Figure 2 was the vector of conjugation features for each genome, which in order to compute the (Euclidean) distance between each genome. Again, Seaborn `clustermap` was used for hierarchical clustering and visualization.

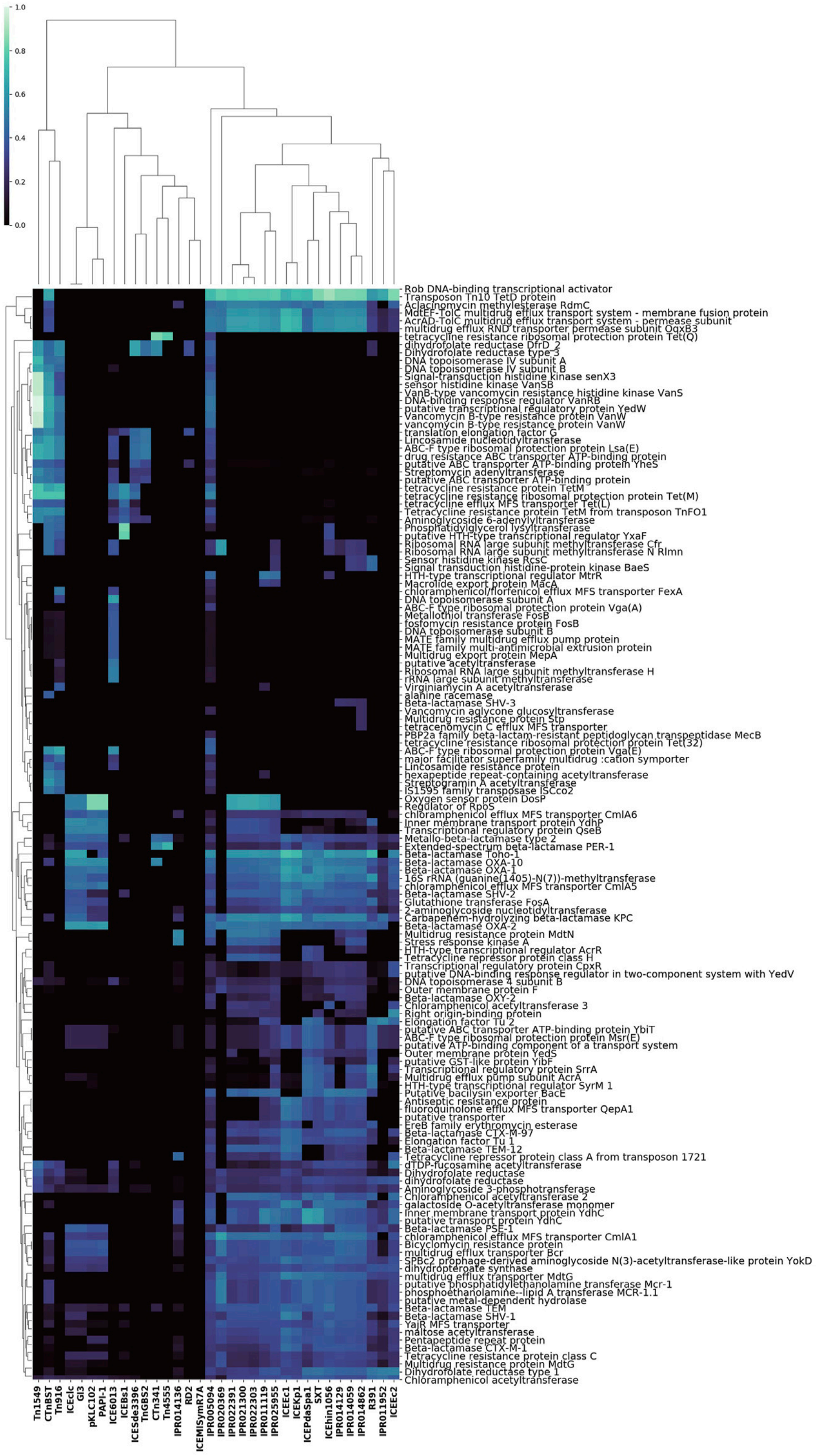


FIGURE 4
Co-occurrence of AMR proteins with conjugative features, for the 138 AMR proteins observed in conjugative genomes with a frequency greater than 5x the frequency of observation in non-conjugative genomes. Of these AMR proteins, 43 are *only* observed in conjugative genomes.

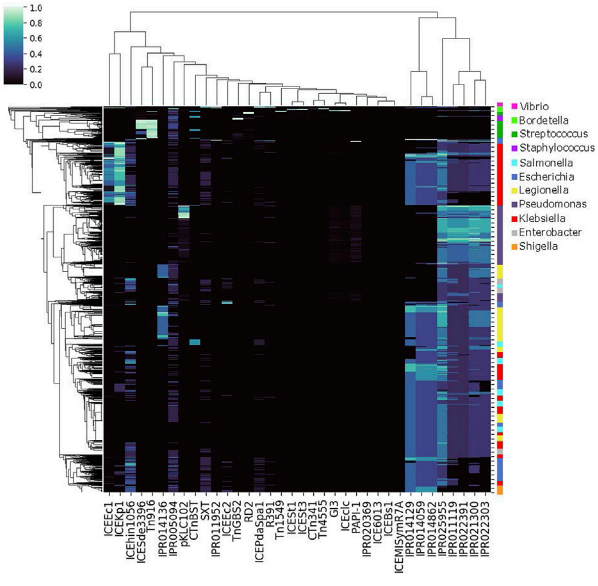


FIGURE 5
Genome-IPR Co-Occurence Map. Co-occurrence of conjugation features by genome, for the 6,000 genomes with the highest cargo protein fraction and the 500 genomes with the rarest conjugation features (see text). Co-occurrence of particular conjugation features does vary by taxonomic group.

Data availability statement

The original contributions presented in the study are included in the article Supplementary Material and at https://github.com/Ruiz-HCI-Lab/ggMOB. Further inquiries can be directed to the corresponding author.

Author contributions

GN, JK, IT, and IS performed and designed research, compiled supporting data, analyzed data, and wrote the paper. GN and AA developed big data visualizations, classified AMR genes and assay data, and contributed to the paper. ES built the database, pipeline, and contributed to the paper. CB provided input on research design, methods and visualizations, and contributed to the paper. NN proposed the study, provided input into the research design, and wrote the paper. JR built the online ggMOB dynamic network and contributed to the paper.

Funding

Financial support for this study was provided by the National Institute of Allergy and Infectious Diseases (NIAID) of the US. National Institutes of Health (NIH), Project No. 1R01AI141810-01 and the National Institute of Food and Agriculture of the United States Department of Agriculture (USDA NIFA), Project No. MINV62-058.

Acknowledgments

The authors would like to acknowledge contributions from M. Roth, H. Krishnareddy, K. Clarkson, and E. Kandogan.

Conflict of interest

ES and AA are currently employed by IBM, GN and JK were employed by IBM at the time of this work. JK is currently employed by Altos Labs.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1024577/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Co-occurrence of AMR proteins with conjugative features.Co-occurrence of AMR proteins with conjugative features for the 29 AMR proteins observed in non-conjugative genomes with a frequency at least 5x the frequency of observation in conjugative genomes..

SUPPLEMENTARY FIGURE S2
Heatmap of Genome-Genome Distances Based on Conjugation Feature Profiles. Genome-genome Euclidean distance between vectors of conjugation features for the 6500 genomes used in 5. For visualization purposes, 1 in 75 labels were rendered on each axis. The figure shows that different sets of genomes cluster based on different co-occurring conjugation features..

References

Abby, S. S., Tannier, E., Gouy, M., and Daubin, V. (2012). Lateral gene transfer as a support for the tree of life. *Proc. Natl. Acad. Sci. U. S. A.* 109 (13), 4962–4967. doi:10.1073/pnas.1116871109

Alderliesten, J. B., Duxbury, S. J. N., Zwart, M. P., de Visser, J. A. G. M., Stegeman, A., and Fischer, E. A. J. (2020). Effect of donor-recipient relatedness on the plasmid conjugation frequency: A meta-analysis. *BMC Microbiol.* 20, 135. doi:10.1186/s12866-020-01825-4

Bacic, M., Parker, A. C., Stagg, J., Whitley, H. P., Wells, W. G., Jacob, L. A., et al. (2005). Genetic and structural analysis of the Bacteroides conjugative transposon CTn341. *J. Bacteriol.* 187 (8), 2858–2869. doi:10.1128/JB.187.8.2858-2869.2005

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19 (5), 455–477. doi:10.1089/cmb.2012.0021

Beaber, J. W., Hochhut, B., and Waldor, M. K. (2002). Genomic and functional analyses of SXT, an integrating antibiotic resistance gene transfer element derived from *Vibrio cholerae*. *J. Bacteriol.* 184 (15), 4259–4269. doi:10.1128/jb.184.15.4259-4269.2002

Bellanger, X., Payot, S., Leblond-Bourget, N., and Guédon, G. (2014). Conjugative and mobilizable genomic islands in bacteria: Evolution and diversity. *FEMS Microbiol. Rev.* 38 (4), 720–760. doi:10.1111/1574-6976.12058

Bellanger, X., Roberts, A. P., Morel, C., Choulet, F., Pavlovic, G., Mullany, P., et al. (2009). Conjugative transfer of the integrative conjugative elements ICESt1 and ICESt3 from Streptococcus thermophilus. *J. Bacteriol.* 191 (8), 2764–2775. doi:10.1128/JB.01412-08

Boer, R., Russi, S., Guasch, A., Lucas, M., Blanco, A. G., Pérez-Luque, R., et al. (2006). Unveiling the molecular mechanism of a conjugative relaxase: The structure of TrwC complexed with a 27-mer DNA comprising the recognition hairpin and the cleavage site. *J. Mol. Biol.* 358 (3), 857–869. doi:10.1016/j.jmb.2006.02.018

Böltner, D., MacMahon, C., Pembroke, J. T., Strike, P., and Osborn, A. M. (2002). R391: A conjugative integrating mosaic comprised of phage, plasmid, and transposon elements. *J. Bacteriol.* 184 (18), 5158–5169. doi:10.1128/jb.184.18.5158-5169.2002

Botelho, J., Mourao, J., Roberts, A. P., and Peixe, L. (2020). Comprehensive genome data analysis establishes a triple whammy of carbapenemases, ICEs and multiple clinically relevant bacteria. *Microb. Genom.* 6. doi:10.1099/mgen.0.000424

Burrus, V., Marrero, J., and Waldor, M. K. (2006). The current ICE age: Biology and evolution of SXT-related integrating conjugative elements. *Plasmid* 55 (3), 173–183. doi:10.1016/j.plasmid.2006.01.001

Burrus, V., Pavlovic, G., Decaris, B., and Guédon, G. (2002). The ICESt1 element of Streptococcus thermophilus belongs to a large family of integrative and conjugative elements that exchange modules and change their specificity of integration. *Plasmid* 48 (2), 77–97. doi:10.1016/s0147-619x(02)00102-6

Byrd, D. R., and Matson, S. W. (1997). Nicking by transesterification: The reaction catalysed by a relaxase. *Mol. Microbiol.* 25 (6), 1011–1022. doi:10.1046/j.1365-2958.1997.5241885.x

Chan, A. P., Sutton, G., DePew, J., Krishnakumar, R., Choi, Y., Huang, X. Z., et al. (2015). A novel method of consensus pan-chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of Acinetobacter baumannii. *Genome Biol.* 16, 143. doi:10.1186/s13059-015-0701-6

Clewell, D. B., Flannagan, S. E., and Jaworski, D. D. (1995). Unconstrained bacterial promiscuity: The tn916–tn1545 family of conjugative transposons. *Trends Microbiol.* 3 (6), 229–236. doi:10.1016/s0966-842x(00)88930-1

Cohen, O., Gophna, U., and Pupko, T. (2010). The complexity hypothesis revisited: Connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol. Biol. Evol.* 28 (4), 1481–1489. doi:10.1093/molbev/msq333

Cury, J., Oliveira, P. H., de la Cruz, F., and Rocha, E. P. C. (2018). Host range and genetic plasticity explain the coexistence of integrative and extrachromosomal mobile genetic elements. *Mol. Biol. Evol.* 35 (9), 2850–2239. doi:10.1093/molbev/msy182

Cury, J., Touchon, M., and Rocha, E. P. C. (2017). Integrative and conjugative elements and their hosts: Composition, distribution and organization. *Nucleic Acids Res.* 45 (15), 8943–8956. doi:10.1093/nar/gkx607

Delavat, F., Miyazaki, R., Carraro, N., Pradervand, N., and van der Meer, J. R. (2017). The hidden life of integrative and conjugative elements. *FEMS Microbiol. Rev.* 41 (4), 512–537. doi:10.1093/femsre/fux008

Everitt, R. G., Didelot, X., Batty, E. M., Miller, R. R., Knox, K., Young, B. C., et al. (2014). Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus*. *Nat. Commun.* 5, 3956. doi:10.1038/ncomms4956

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., et al. (2016). Interpro in 2017 – beyond protein family and domain annotations. *Nucleic Acids Res.* 45 (D1), D190-D199. doi:10.1093/nar/gkw1107

Frost, L. S., Leplae, R., Summers, A. O., and Toussaint, A. (2005). Mobile genetic elements: The agents of open source evolution. *Nat. Rev. Microbiol.* 3 (9), 722–732. doi:10.1038/nrmicro1235

Garnier, F., Taourit, S., Glaser, P., Courvalin, P., and Galimand, M. (2000). Characterization of transposon Tn1549, conferring VanB-type resistance in Enterococcus spp. *Microbiology* 146 (6), 1481–1489. doi:10.1099/00221287-146-6-1481

Gaston, J. M., Zhao, S., Poyet, M., Groussin, M., et al. (2021). An omics-based framework for assessing the health risk of antimicrobial resistance genes. *Nat. Commun.* 12, 4765. doi:10.1038/s41467-021-25096-3

Guglielmini, J., Quintais, L., Garcillán-Barcia, M. P., de La Cruz, F., and Rocha, E. P. C. (2011). The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.* 7 (8), e1002222. doi:10.1371/journal.pgen.1002222

Harris, R. L., Sholl, K. A., Conrad, M. N., Dresser, M. E., and Silverman, P. M. (1999). Interaction between the F plasmid TraA (F-pilin) and TraQ proteins. *Mol. Microbiol.* 34 (4), 780–791. doi:10.1046/j.1365-2958.1999.01640.x

Johnson, C. M., and Grossman, A. D. (2015). Integrative and conjugative elements (ICEs): What they do and how they work. *Annu. Rev. Genet.* 49, 577–601. doi:10.1146/annurev-genet-112414-055018

Klockgether, J., Reva, O., Larbig, K., and Tümmler, B. (2004). Sequence analysis of the mobile genome island pKLC102 of *Pseudomonas aeruginosa* C. *J. Bacteriol.* 186 (2), 518–534. doi:10.1128/jb.186.2.518-534.2004

Koraimann, G., and Wagner, M. A. (2014). Social behavior and decision making in bacterial conjugation. *Front. Cell.. Infect. Microbiol.* 4, 54. doi:10.3389/fcimb.2014.00054

Lakin, S. M., Dean, C., Noyes, N. R., Dettenwanger, A., Ross, A. S., Doster, E., et al. (2016). MEGARes: An antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Res.* 45 (D1), D574-D580–D580. doi:10.1093/nar/gkw1009

Lechner, M., Schmitt, K., Bauer, S., Hot, D., Hubans, C., Levillain, E., et al. (2009). Genomic island excisions in Bordetella petrii. *BMC Microbiol.* 9 (1), 141. doi:10.1186/1471-2180-9-141

Lee, C. A., Thomas, J., and Grossman, A. D. (2012). The Bacillus subtilis conjugative transposon ICEBs1 mobilizes plasmids lacking dedicated mobilization functions. *J. Bacteriol.* 194 (12), 3165–3172. doi:10.1128/JB.00301-12

Logan, L. K., and Weinstein, R. A. (2017). The epidemiology of carbapenem-resistant Enterobacteriaceae: The impact and evolution of a global menace. *J. Infect. Dis.* 215 (1), S28–S36. S28–S36. doi:10.1093/infdis/jiw282

Lopatkin, A. J., and Collin, J. J. (2020). Predictive biology: Modelling, understanding and harnessing microbial complexity. *Nat. Rev. Microbiol.* 18, 507–520. doi:10.1038/s41579-020-0372-5

Martínez, J. L., Coque, T. M., and Baquero, F. (2015). What is a resistance gene? Ranking risk in resistomes. *Nat. Rev. Microbiol.* 13 (2), 116–123. doi:10.1038/nrmicro3399

Matson, S. W., and Ragonese, H. (2005). The F-plasmid TraI protein contains three functional domains required for conjugative DNA strand transfer. *J. Bacteriol.* 187 (2), 697–706. doi:10.1128/JB.187.2.697-706.2005

Mavrodi, D. V., Loper, J. E., Paulsen, I. T., and Thomashow, L. S. (2009). Mobile genetic elements in the genome of the beneficial rhizobacterium Pseudomonas fluorescens Pf-5. *BMC Microbiol.* 9 (1), 8. doi:10.1186/1471-2180-9-8

Meyer, R. (2011). Functional organization of MobB, a small protein required for efficient conjugal transfer of plasmid R1162. *J. Bacteriol.* 193 (15), 3904–3911. doi:10.1128/JB.05084-11

Millman, A., Melamed, S., Amitai, G., and Sorek, R. (2020). Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems. *Nat. Microbiol.* 5 (12), 1608–1615. doi:10.1038/s41564-020-0777-y

Mohd-Zain, Z., Turner, S. L., Cerdeno-Tárraga, A. M., Lilley, A. K., Inzana, T. J., Duncan, A. J., et al. (2004). Transferable antibiotic resistance elements in Haemophilus influenzae share a common evolutionary origin with a diverse family of syntenic genomic islands. *J. Bacteriol.* 186 (23), 8114–8122. doi:10.1128/JB.186.23.8114-8122.2004

Obi, C. C., Vayla, S., De Gannes, V., Berres, M. E., Walker, J., Pavelec, D., et al. (2018). The integrative conjugative element clc (ICEclc) of *Pseudomonas aeruginosa* JB2. *Front. Microbiol.* 9, 1532. doi:10.3389/fmicb.2018.01532

Oh, M., Pruden, A., Chen, C., Heath, L. S., Xia, K., and Zhang, L. (2018). MetaCompare: A computational pipeline for prioritizing environmental resistome risk. *FEMS Microbiol. Ecol.* 94 (6), fiy079. doi:10.1093/femsec/fiy079

Osorio, C. R., Marrero, J., Wozniak, R. A. F., Lemos, M. L., Burrus, V., and Waldor, M. K. (2008). Genomic and functional analysis of ICEPdaSpa1, a fish-pathogen-derived SXT-related integrating conjugative element that can mobilize a virulence plasmid. *J. Bacteriol.* 190 (9), 3353–3361. doi:10.1128/JB.00109-08

Paauw, A., Leverstein-van Hall, M. A., Verhoef, J., and Fluit, A. C. (2010). Evolution in quantum leaps: Multiple combinatorial transfers of HPI and other genetic modules in Enterobacteriaceae. *PLoS ONE* 5 (1), e8662. doi:10.1371/journal.pone.0008662

Pansegrau, W., Schoumacher, F., Hohn, B., and Lanka, E. (1993). Site-specific cleavage and joining of single-stranded DNA by VirD2 protein of agrobacterium tumefaciens Ti plasmids: Analogy to bacterial conjugation. *Proc. Natl. Acad. Sci. U. S. A.* 90 (24), 11538–11542. doi:10.1073/pnas.90.24.11538

Parker, B. J., Wearsch, P. A., Veloo, A. C. M., and Rodriguez-Palacios, A. (2020). The genus Alistipes: Gut bacteria with emerging implications to inflammation, cancer, and mental health. *Front. Immunol.* 11, 906. doi:10.3389/fimmu.2020.00906

Partridge, S. R., Kwong, S. M., Firth, N., and Jensen, S. O. (2018). Mobile genetic elements associated with antimicrobial resistance. *Clin. Microbiol. Rev.* 31. doi:10.1128/cmr.00088-17

Peed, L., Parker, A. C., and Smith, C. J. (2010). Genetic and functional analyses of the mob operon on conjugative transposon CTn341 from Bacteroides spp. *J. Bacteriol.* 192 (18), 4643–4650. doi:10.1128/JB.00317-10

Perry, J., and Wright, G. (2013). The antibiotic resistance "mobilome": Searching for the link between environment and clinic. *Front. Microbiol.* 4, 138. doi:10.3389/fmicb.2013.00138

Prokka, T. S. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30 (14), 2068–2069. doi:10.1093/bioinformatics/btu153

Qiu, X., Gurkar, A. U., and Lory, S. (2006). Interstrain transfer of the large pathogenicity island (PAPI-1) of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U. S. A.* 103 (52), 19830–19835. doi:10.1073/pnas.0606810104

Quevillon, E., SilVentoinen, V., Pillai, S., HarteN.MulderN.ApweileR, R., et al. (2005). InterProScan: Protein domains identifier. *Nucleic Acids Res.* 33 (2), W116–W120. –W120. doi:10.1093/nar/gki442

Ramsay, J. P., and Ronson, C. W. (2015). *Genetic regulation of symbiosis island transfer in Mesorhizobium loti*. New York, NY: John Wiley & Sons, 219.

Roberts, A. P., and Mullany, P. (2011). Tn916-like genetic elements: A diverse group of modular mobile elements conferring antibiotic resistance. *FEMS Microbiol. Rev.* 35 (5), 856–871. doi:10.1111/j.1574-6976.2011.00283.x

Roche, D., Flechard, M., Lallier, N., Reperant, M., Bree, A., Pascal, G., et al. (2010). ICEEc2, a new integrative and conjugative element belonging to the pKLC102/PAGI-2 family, identified in *Escherichia coli* strain BEN374. *J. Bacteriol.* 192 (19), 5026–5036. doi:10.1128/JB.00609-10

Neil, K., and Allard, N. (2021). Molecular mechanisms influencing bacterial conjugation in the intestinal microbiota. *Front. Microbiol.* 12, 673260. doi:10.3389/fmicb.2021.673260

Schandel, K. A., Muller, M. M., and Webster, R. E. (1992). Localization of TraC, a protein involved in assembly of the F conjugative pilus. *J. Bacteriol.* 174 (11), 3800–3806. doi:10.1128/jb.174.11.3800-3806.1992

Seabolt, E., Nayar, G., Krishnareddy, H., Agarwal, A., Beck, K. L., Terrizzano, I., et al. (2020). Functional genomics platform, A cloud-based platform for studying microbial life at scale. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 1 (–1), 940–952. doi:10.1109/TCBB.2020.3021231

Sheppard, R. J., Beddis, A. E., and Barraclough, T. G. (2020). The role of hosts, plasmids and environment in determining plasmid transfer rates: A meta-analysis. *Plasmid* 108, 102489. doi:10.1016/j.plasmid.2020.102489

Singer, A. C., Shaw, H., Rhodes, V., and Hart, A. (2016). Review of antimicrobial resistance in the environment and its relevance to environmental regulators. *Front. Microbiol.* 7, 1728. doi:10.3389/fmicb.2016.01728

Sitkiewicz, I., Green, N. M., Guo, N., Mereghetti, L., and Musser, J. M. (2011). Lateral gene transfer of streptococcal ICE element RD2 (region of difference 2) encoding secreted proteins. *BMC Microbiol.* 11 (1), 65. doi:10.1186/1471-2180-11-65

Smith, C. J., and Parker, A. C. (1993). Identification of a circular intermediate in the transfer and transposition of Tn4555, a mobilizable transposon from Bacteroides spp. *J. Bacteriol.* 175 (9), 2682–2691. doi:10.1128/jb.175.9.2682-2691.1993

Smyth, D. J., Shera, J., Bauer, M. J., Cameron, A., McNeilly, C. L., Sriprakash, K. S., et al. (2014). Conjugative transfer of ICE Sde 3396 between three β-hemolytic streptococcal species. *BMC Res. Notes* 7 (1), 521. doi:10.1186/1756-0500-7-521

Smyth, D. S., and Robinson, D. A. (2009). Integrative and sequence characteristics of a novel genetic element, ICE6013, in *Staphylococcus aureus*. *J. Bacteriol.* 191 (19), 5964–5975. doi:10.1128/JB.00352-09

Song, B., Shoemaker, N. B., Gardner, J. F., and Salyers, A. A. (2007). Integration site selection by the Bacteroides conjugative transposon CTnBST. *J. Bacteriol.* 189 (18), 6594–6601. doi:10.1128/JB.00668-07

Springael, D., Kreps, S., and Mergeay, M. (1993). Identification of a catabolic transposon, Tn4371, carrying biphenyl and 4-chlorobiphenyl degradation genes in Alcaligenes eutrophus A5. *J. Bacteriol.* 175 (6), 1674–1681. doi:10.1128/jb.175.6.1674-1681.1993

Street, L. M., Harley, M. J., Stern, J. C., Larkin, C., Williams, S. L., Miller, D. L., et al. (2003). Subdomain organization and catalytic residues of the F factor TraI relaxase domain. *Biochim. Biophys. Acta* 1646 (1-2), 86–99. doi:10.1016/s1570-9639(02)00553-8

Su, Y., He, P., and Clewell, D. B. (1992). Characterization of the tet (M) determinant of Tn916: Evidence for regulation by transcription attenuation. *Antimicrob. Agents Chemother.* 36 (4), 769–778. doi:10.1128/aac.36.4.769

Sugawara, H., and Shumway, M. (2010). The sequence read archive. *Nucleic Acids Res.* 39 (1), D19–D21. doi:10.1093/nar/gkq1019

Uhlemann, A. C. L., Aj Prensky, H., and Gomez-Simmonds, A. (2021). Conjugation dynamics depend on both the plasmid acquisition cost and the fitness cost. *Mol. Syst. Biol.* 17, e9913. doi:10.15252/msb.20209913

Waldor, M. K., Beaber, J. W., and Hochhut, B. (2004). SOS response promotes horizontal dissemination of antibiotic resistance genes. *Nature* 427, 72–74. doi:10.1038/nature02241

Wang, R., van Dorp, L., Shaw, L. P., Bradley, P., Wang, Q., Wang, X., et al. (2018). The global distribution and spread of the mobilized colistin resistance gene mcr-1. *Nat. Commun.* 9 (1), 1179–9. doi:10.1038/s41467-018-03205-z

Waskom, M., Botvinnik, O., Hobson, P., Warmenhoven, J., Cole, J. B., Halchenko, Y., et al. (2015). Seaborn: V0. 6.0 (June 2015). *Zenodo*. doi:10.5281/zenodo.19108

Wiedenbeck, J., and Cohan, F. M. (2011). Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol. Rev.* 35 (5), 957–976. doi:10.1111/j.1574-6976.2011.00292.x

Wozniak, R. A. F., and Waldor, M. K. (2010). Integrative and conjugative elements: Mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat. Rev. Microbiol.* 8 (8), 552–563. doi:10.1038/nrmicro2382

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 23 February 2023
DOI 10.3389/fgene.2023.1148470

Check for updates

OPEN ACCESS

EDITED BY
Quan Zou,
University of Electronic Science and Technology of China, China

REVIEWED BY
Ying Wang,
Xiamen University, China
Guangchuang Yu,
Southern Medical University, China

*CORRESPONDENCE
Dongmei Ai,
✉ aidongmei@ustb.edu.cn
Li C. Xia,
✉ lcxia@scut.edu.cn

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 20 January 2023
ACCEPTED 13 February 2023
PUBLISHED 23 February 2023

CITATION
Ai D, Wang M, Zhang Q, Cheng L, Wang Y, Liu X and Xia LC (2023), Regularized survival learning and cross-database analysis enabled identification of colorectal cancer prognosis-related immune genes.
Front. Genet. 14:1148470.
doi: 10.3389/fgene.2023.1148470

COPYRIGHT
© 2023 Ai, Wang, Zhang, Cheng, Wang, Liu and Xia. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Regularized survival learning and cross-database analysis enabled identification of colorectal cancer prognosis-related immune genes

Dongmei Ai[1]*, Mingmei Wang[1], Qingchuan Zhang[2], Longwei Cheng[1], Yishu Wang[1], Xiuqin Liu[1] and Li C. Xia[3]*

[1]School of Mathematics and Physics, University of Science and Technology Beijing, Beijing, China, [2]National Engineering Laboratory for Agri-Product Quality Traceability, Beijing Technology and Business University, Beijing, China, [3]School of Mathematics, South China University of Technology, Guangzhou, China

Colon adenocarcinoma is the most common type of colorectal cancer. The prognosis of advanced colorectal cancer patients who received treatment is still very poor. Therefore, identifying new biomarkers for prognosis prediction has important significance for improving treatment strategies. However, the power of biomarker analyses was limited by the used sample size of individual database. In this study, we combined Genotype-Tissue Expression (GTEx) and The Cancer Genome Atlas (TCGA) databases to expand the number of healthy tissue samples. We screened differentially expressed genes between the GTEx healthy samples and TCGA tumor samples. Subsequently, we applied least absolute shrinkage and selection operator (LASSO) regression and multivariate Cox analysis to identify nine prognosis-related immune genes: *ANGPTL4*, *IDO1*, *NOX1*, *CXCL3*, *LTB4R*, *IL1RL2*, *CD72*, *NOS2*, and *NUDT6*. We computed the risk scores of samples based on the expression levels of these genes and divided patients into high- and low-risk groups according to this risk score. Survival analysis results showed a significant difference in survival rate between the two risk groups. The high-risk group had a significantly lower overall survival rate and poorer prognosis. We found the receiver operating characteristic based on the risk score was showed to accurately predict patients' prognosis. These prognosis-related immune genes may be potential biomarkers for colorectal cancer diagnosis and treatment. Our open-source code is freely available from GitHub at https://github.com/gutmicrobes/Prognosis-model.git.

KEYWORDS

LASSO, multivariate cox analysis, prognosis, immune gene, colorectal cancer

1 Introduction

According to global cancer statistics 2020 data, colorectal cancer ranked third by cancer incidence and second by cancer mortality rate (Sung et al., 2021). According to predictions, the number of new colorectal cancers will reach 2.2 million and deaths will reach 1.1 million in 2030 (Arnold et al., 2017). Colorectal cancer usually occurs in the inner walls of the colon or rectum (Lao and Grady, 2011). When malignant cells are formed in the colon or rectum, it will lead to the occurrence of colorectal cancer (Wang et al., 2021). Based on histological classification, colon adenocarcinoma is the main type of colorectal cancer (Wei et al., 2018). The main causes of transformation of normal colonic epithelium to colon adenocarcinoma are genetic and epigenetic changes (Coppede, 2014). At present, the main method for treating colon adenocarcinoma is surgery combined with postoperative chemotherapy (Hashiguchi et al., 2020; Tarazona et al., 2020). Even with standard treatment, the outcomes of advanced colon adenocarcinoma patients are still very poor and varies widely (Andre et al., 2004; Nishihara et al., 2013; Sadanandam et al., 2013). Therefore, using simple conventional factors, such as clinicopathology stage, is insufficient for accurate prognostic prediction of colon adenocarcinoma patients, which calls for the discovery of new biomarkers to predict the prognosis of patients and improve treatment outcomes.

Biomarkers improve patients' prognosis by treating patients who may benefit from a given treatment (Blangero et al., 2020). In recent years, the rapid development of bioinformatics tools has enabled researchers to rapidly identify colorectal cancer biomarkers based on differentially expressed genes (DEGs). For examples, Dalerba et al. found that *CDX2* is a prognostic biomarker and that *CDX2* deletion is associated with poor prognosis in stage II or III colorectal cancer patients (Dalerba et al., 2016). Li et al. found that the immune gene *ULBP2* is a prognostic biomarker and that *TMEM37* and *GRP* may also be potential prognostic genes for colon cancer (Li et al., 2018). Wang et al. found that *MXRA5* is aberrantly expressed in colorectal cancer tissues and is a biomarker for the early detection of colorectal cancer (Wang et al., 2013). Den Uil et al. found that *KCNQ1* is a prognostic biomarker for predicting recurrence in stage II and III colon cancer patients (den Uil et al., 2016). Woischke et al. found that *CYB5R1* is intimately associated with poor prognosis in colorectal cancer (Woischke et al., 2016). Kandimalla et al. found that methylated *AXIN2* and *DKK1* are useful biomarkers for recurrence in stage II colon cancer patients (Kandimalla et al., 2017).

Compared with a single biomarker, combining multiple biomarkers in a model can predict patients' prognosis more accurately (Qu et al., 2018). For example, Lin et al. proposed a new prognosis risk score characteristic based on nine long non-coding RNAs (lncRNAs) associated with colon cancer prognosis (Lin et al., 2020). This characteristic has important clinical significance in improving the prediction results of colon cancer patients, and these lncRNAs as a whole may be biomarkers that affect prognosis. Zuo et al. carried out univariate and multivariate Cox analysis to identify six DEGs associated with colorectal cancer patients prognosis, including *EPHA6*, *TIMPI*, *IRX6*, *ART5*, *HIST3H2BB*, and *FOXD1* (Zuo et al., 2019). Their combined is an independent biomarker for predicting the survival rate.

Currently, immunotherapy has demonstrated huge potential in improving tumor prognosis, and studies have increasingly shown that expression of immune-related genes may be related to cancer patients' prognosis (Galon et al., 2013; Bedognetti et al., 2015). For example, Miao et al. identified 12 immune genes (*SLC10A2*, *CXCL3*, *NOX4*, *FABP4*, *ADIPOQ*, *IGKV1-33*, *IGLV6-57*, *INHBA*, *UCN*, *VIP*, *NGFR*, and *TRDC*) associated with the prognosis of colon adenocarcinoma patients (Miao et al., 2020). The associated risk score proved an independent prognostic factor. Therefore, the identification of colon adenocarcinoma-related immune genes is particularly useful to promote the development of tools to carry out colon adenocarcinoma immunotherapy.

However, the aforementioned studies only used healthy samples and tumor samples from The Cancer Genome Atlas (TCGA) database to identify DEGs between healthy samples and tumor samples. The differences in the number of samples in the TCGA database are very large. For example, several hundred tumor samples are available, but only a few dozen healthy samples (Mounir et al., 2019). This big difference will lead to inaccuracy in the identification of DEGs.

Therefore, in this study, we collected healthy tissue samples from the Genotype-Tissue Expression (GTEx) database and tumor tissue samples from the TCGA database when screening for DEGs. Large sample size enabled us to sensitively identify biomarkers based on DEGs. We employed least absolute shrinkage and selection operator (LASSO) regression and multivariate Cox analysis to construct a risk model based on multiple immune genes. This model can accurately predict patients' prognosis (AUC of training dataset >0.8), which has important clinical significance. The immune genes identified in the model could be used as potential biomarkers.

2 Materials and methods

2.1 Data sources

Healthy colon tissue RNA-seq data of 308 samples in the GTEx database were downloaded from the UCSC website (https://xenabrowser.net/, accessed on 25 March 2022), as fragments per kilobase of exon model per million mapped fragments (FPKM) values. Gene expression data were extracted from 308 healthy samples. We removed low-expressing genes that the mean expression level is less than 0.2. After removing low-expressing genes, the expression levels of 22,116 genes were retained.

The RNA-seq FPKM data of 391 colon adenocarcinoma samples were downloaded from the TCGA website (https://portal.gdc.cancer.gov/, accessed on 21 March 2022). Genes (mean expression level <0.2 in samples) were removed to obtain the expression levels of 14,791 genes. The clinical data of 391 colon adenocarcinoma patients were also downloaded from the TCGA website. The analysis flow chart is shown in Figure 1.

2.2 Screening of differentially expressed genes

The list of human immune genes was downloaded from the Immunology Database and Analysis Portal (IMMPORT) database (https://www.immport.org/home, accessed on 30 March 2022). Total 1793 immune genes were included. The GTEx dataset and TCGA dataset were combined to obtain 14,306 intersection genes. We used R package "limma" to screen DEGs between healthy samples and tumor samples through Wilcoxon test (Ritchie et al., 2015). False discovery rate (fdr) was computed to correct multiple testing. The screening criteria were $fdr < 0.05$ and $|log_2(fold\ change)| > 1$. After obtaining the list of DEGs, the intersection with immune genes was obtained as differentially expressed immune genes.

$$log_2(fold\ change) = log_2 \frac{mean\ value\ of\ gene\ in\ tumor\ group}{mean\ value\ of\ gene\ in\ healthy\ group} \tag{1}$$

2.3 Regularized survival analysis

Univariate Cox analysis is typically used to screen for prognosis-related genes in patients, and then a multivariate model is constructed to further confirm whether the association between gene and survival is independent. However, this method does not consider the multiple collinear effects between genes, and contradiction in hazard ratios (HR) obtained from univariate Cox regression and multivariate Cox regression occurs, causing model distortion. However, the multivariate analysis also suffers from the curse of dimensionality when the number of genes is greater than the sample size.

The modernized regularized survival analysis approach, such as LASSO, avoids the high-dimensionality issue by soft-selecting significant features. We thus employed LASSO Cox regression for gene screening before multivariate Cox regression model was used to establish prognostic characteristics. LASSO regularization, which was proposed by Tibshirani (Tibshirani, 1997), uses L1 norm for the shrinkage penalty in which the coefficients of not-so-important genes are compressed to 0, while the coefficients of important genes are retained at more than 0. This decreases the number of covariates in the Cox regression (i.e., genes). Genes with a coefficient >0 in LASSO-Cox regression were selected for further calculation of the risk score (Kidd et al., 2018). The formula of LASSO is as follows (Emmert-Streib and Dehmer, 2019):

$$\hat{\beta} = \arg\ min\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j x_{ij}\right)^2 + \lambda\|\beta\|_1\right\}$$
$$= \arg\ min\left\{\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right\} \tag{2}$$

The survival data of TCGA patients and the expression data of differentially expressed immune genes were combined. The 391 patient samples were randomized into a training dataset and a validation dataset. The training dataset accounted for 70% (273 samples) of the dataset, and the testing dataset accounted for 30% (118 samples) of the dataset. Data in the training dataset were used for LASSO regression. We used R package "glmnet" to conduct LASSO regression analysis. The objective was to minimize overfitting, i.e., removal of genes that will cause overfitting, and select differentially expressed immune genes significantly associated with survival.

2.4 Multivariate Cox analysis

The multivariate Cox regression model, also known as the proportional hazards model, is a semi-parametric regression model (Kleinbaum and Klein, 2012). In this model, survival outcome and survival time were used as dependent variables. The model can simultaneously analyze the effects of multiple variables (e.g., genes) on survival. Candidate immune genes related to prognosis were obtained through LASSO analysis, and then a risk model was constructed through multivariate Cox analysis. Multivariate Cox analysis will screen candidate immune genes by stepwise regression method. Multivariate Cox analysis was conducted using the R package "survival".

A multivariate Cox regression model was used to construct a prognostic characteristic of immune genes and calculate the risk score of each patient sample. The calculation formula is as follows:

$$Risk\ score = \sum_{i=1}^{n} exp_i * coef_i \tag{3}$$

where n is the number of characteristic genes included in the model, exp_i represents the expression level of gene i, and $coef_i$ represents the coefficient of gene i in the multivariate Cox regression analysis. We determined the optimal cut-off value of risk score according to the maximally selected log-rank statistics (Wright et al., 2017). Patients were divided into two groups based on the optimal cut-off value. Patients with risk scores greater than the cut-off value were included in the high-risk group, and patients whose risk scores did not exceed the cut-off value were included in the low-risk group.

2.5 Survival analysis and ROC curve computing

The Kaplan-Meier curve is also known as the survival curve and is a commonly used method in survival analysis. The Kaplan-Meier curve mainly analyzes the effect of a single factor on survival, and it is used to estimate the survival rate of patients. Survival time is the x-axis, survival rate is the y-axis, and a continuous stepped curve is computed to describe the relationship between survival time and survival rate. The log-rank test was used to evaluate survival differences between the two groups. We used the R package "survival" to conduct survival analysis. Receiver operating characteristic (ROC) curves were computed, and the area under the ROC curve (AUC) was calculated to assess the accuracy of the prognostic model. We used the R package "time ROC" package to calculate the AUC at different cutoff times.

2.6 Independence and model validation

Multivariate analysis was carried out for patient samples with clinical characteristics, and the prognostic value of the risk score was assessed. Based on multivariate analysis, the characteristics of $p < 0.05$ can be used as an independent prognostic factor. The entire TCGA dataset (391 samples) and testing dataset (118 samples) were used for model validation. The risk score of each sample was calculated based on the same formula [see Formula (4)], and samples were grouped into high- and low-risk groups based on the optimal cut-off value. Survival analysis was performed for these two groups to evaluate the survival differences between the two groups. A ROC curve was computed, and the AUC was calculated to assess model accuracy. Data analysis and visualization were performed using R software (version 4.1.3, https://www.rstudio.com/, accessed on 18 March 2022).

3 Results

3.1 Screening candidate immune biomarker

The Wilcoxon test was used to screen DEGs between GTEx healthy samples and TCGA tumor samples, and the screening criteria were $fdr < 0.05$ and $|log_2(fold\ change)| > 1$. By comparing with the healthy tissue group, 7670 DEGs were obtained. Among these, 6381 genes were downregulated, and 1289 genes were upregulated. A listing of 1793 immune genes was downloaded from the IMMPORT database, and the

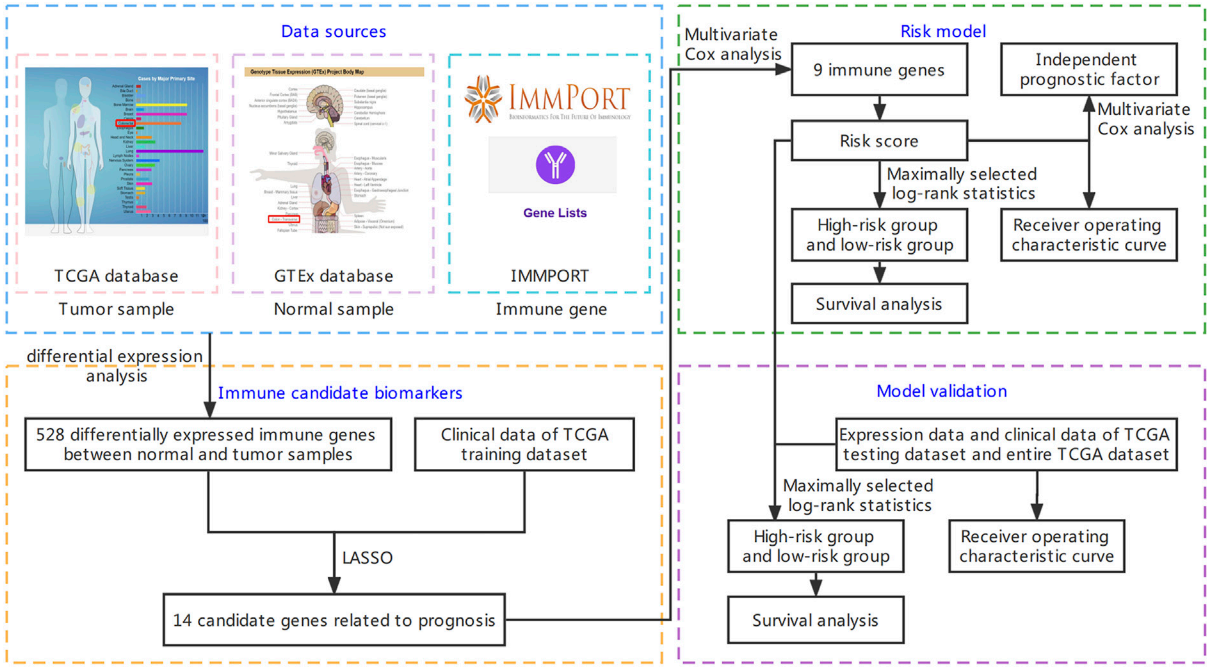

Data sources
TCGA database
Tumor sample
GTEx database
Normal sample
IMMPORT
Gene Lists
IMMPORT
Immune gene
differential expression analysis
Immune candidate biomarkers
528 differentially expressed immune genes between normal and tumor samples
Clinical data of TCGA training dataset
LASSO
14 candidate genes related to prognosis
Risk model
Multivariate Cox analysis
9 immune genes
Independent prognostic factor
Multivariate Cox analysis
Risk score
Maximally selected log-rank statistics
High-risk group and low-risk group
Receiver operating characteristic curve
Survival analysis
Model validation
Expression data and clinical data of TCGA testing dataset and entire TCGA dataset
Maximally selected log-rank statistics
High-risk group and low-risk group
Receiver operating characteristic curve
Survival analysis

FIGURE 1
Flow chart of this study. It is mainly divided into four parts: downloading data, screening immune candidate biomarkers, building risk model, and model validation. The detailed steps are shown in the figure.

TABLE 1 Summary of the clinical data of The Cancer Genome Atlas (TCGA) colon adenocarcinoma patients.

Clinical parameter	Variable	n (total = 341)	Percentage (%)
Age (years)	≤60	97	28.4
	>60	244	71.6
Gender	Female	155	45.5
	Male	186	54.5
Stage	Stage I	59	17.3
	Stage II	138	40.4
	Stage III	93	27.3
	Stage IV	51	15.0
Tumor	T1	8	2.4
	T2	57	16.7
	T3	236	69.2
	T4	40	11.7
Metastasis	M0	290	85.0
	M1	51	15.0
Lymph Node	N0	203	59.5
	N1	81	23.8
	N2	57	16.7
Survival status	Alive	282	82.7
	Dead	59	17.3

TABLE 2 Multivariate Cox analysis results of training dataset.

Gene symbol	Coef	Hazard ratios (HR)	95% CI of HR
ANGPTL4	0.109	1.115	1.069–1.163
IDO1	0.005	1.005	1.001–1.009
NOX1	−0.006	0.994	0.988–1.000
CXCL3	−0.016	0.984	0.962–1.007
LTB4R	0.076	1.078	1.010–1.152
IL1RL2	0.133	1.142	0.964–1.354
CD72	0.304	1.355	1.037–1.771
NOS2	−0.018	0.982	0.960–1.005
NUDT6	−1.689	0.185	0.031–1.082

intersection with DEGs, which contained 528 differentially expressed immune genes, was retained. Among these, 383 genes were downregulated, and 145 genes were upregulated.

Clinical data of 391 colon adenocarcinoma patients were downloaded from the TCGA database. The clinical information of 341 samples was retained by deleting some samples with unknown clinical characteristics. Table 1 shows the detailed clinical information. We divide the sample into two groups according to age, one group is no more than 60 years old, and the other group is over 60 years old (Lin et al., 2019).

TNM staging system is the most commonly used tumor staging system in the world. T is the first letter of “Tumor”, referring to the tumor size and local invasion range. T1 refers to the smaller primary part. T2 refers to the larger primary part. T3 refers to the larger primary part and/or the infiltration exceeds the edge of the primary organ. T4 refers to the very large primary part and/or the infiltration to adjacent organs. N is the first letter of “Node” in the lynch node, which refers to the involvement of regional lymph nodes. N0 refers to no lymph node metastasis. N1 refers to local lymph node metastasis. N2 refers to extensive lymph node metastasis. M is the first letter of “metastasis”, which refers to remote metastasis. M0 means no distal metastasis, and the tumor does not spread to other parts of the body. M1 refers to distal metastasis, and the tumor spreads to other parts of the body. Stage group determined from clinical information on the tumor (T), regional node (N) and metastases (M) and by grouping cases with similar prognosis for cancer. Stage includes stage Ⅰ, stage Ⅱ, stage Ⅲ and stage Ⅳ. Stage Ⅰ tumors are usually relatively early tumors with relatively good prognosis. The higher the stage, the higher the degree of tumor progression.

Expression and survival data of differentially expressed immune genes were combined to obtain the expression and survival data of differentially expressed immune genes of 391 samples. The 391 samples were randomized into the training dataset and testing dataset. The sample size of the training dataset accounted for 70% (273 samples) of the total sample size, and the sample size of the testing dataset accounted for 30% (118 samples) of the total sample size. To determine prognosis-related immune genes, training dataset samples were used for LASSO regression. Among the 528 differentially expressed immune genes between the healthy and tumor samples, 14 candidate genes were obtained (Supplementary Figure S1).

3.2 Predictive model construction through Multivariate Cox analysis

Multivariate Cox analysis was used for further screening of the 14 candidate biomarker genes, and nine biomarker genes were finally obtained (Table 2). The expression levels of these nine immune genes and their corresponding correlation coefficients were used to calculate risk scores. The calculation formula is as follows:

$$\begin{aligned} Risk\ score = {} & (0.109{*}expression\,level\,of\ ANGPTL4) \\ & + (0.005{*}expression\,level\,of\ IDO1) \\ & - (0.006{*}expression\,level\,of\ NOX1) \\ & - (0.016{*}expression\,level\,of\ CXCL3) \\ & + (0.076{*}expression\,level\,of\ LTB4R) \\ & + (0.133{*}expression\,level\,of\ IL1RL2) \\ & + (0.304{*}expression\,level\,of\ CD72) \\ & - (0.018{*}expression\,level\,of\ NOS2) \\ & - (1.689{*}expression\,level\,of\ NUDT6) \end{aligned} \tag{4}$$

The overall importance of the model was tested. The p values of the three tests were less than 0.05, which were likelihood ratio test ($p = 1e - 10$), wald test ($p = 1e - 10$) and score log rank test ($p < 2e - 16$). The optimal cut-off value of risk score is determined through the surv_cutpoint function of R. The optimal cut-off value of training dataset is 2.02 (Figure 2A). The 273 colon adenocarcinoma patients in the training dataset were divided into two groups based on the optimal cut-off value. Patients with risk scores greater than the cut-off were included in the high-risk group ($n = 73$), and patients with risk scores lower than the cut-off were included in the low-risk group ($n = 200$). Supplementary Figure S2 shows the survival distribution of the low- and high-risk groups. As risk score increased, the number of patient deaths increased, and the survival time decreased; that is, the number of deaths in the high-risk group was higher, and the survival rate was lower.

Supplementary Figure S3 shows the heatmap of nine immune genes included in the model. The $log_2\,(expression\ value)$ of genes in the healthy and tumor groups are also shown. *ANGPTL4*, *LTB4R*, *CD72*, and *NUDT6* were downregulated, as their expression levels were higher in the healthy group and lower in the tumor group. *IDO1*, *NOX1*, *CXCL3*, *IL1RL2* and *NOS2* were upregulated, as their expression levels were lower in the healthy group and higher in the tumor group.

3.3 Survival analysis and ROC characterization of training dataset

The genes were screened by LASSO regression, and the model was constructed by multifactor cox regression. The survival

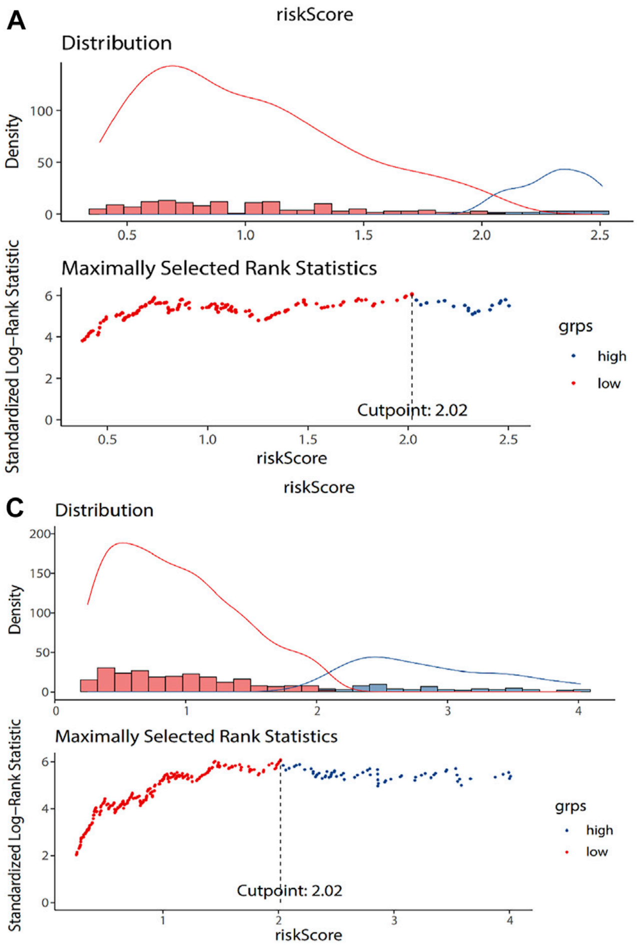


FIGURE 2
The grouping results of training dataset **(A)**, testing dataset **(B)**, and entire TCGA dataset **(C)**. The lower part of the figure is the optimal cut-off value calculated according to the maximum selection rank statistics. The risk score of the blue dot is lower than the cut-off value, which is a low-risk group. The risk score of the red dot is higher than the cut-off value, which is a high-risk group. The upper part of the figure is the data distribution histogram and density distribution curve of risk score. Blue represents low risk group, and red represents high risk group.

analysis results of the training set, the test set, and the entire data set are shown in Figures 3A,C,E. After screening the genes through univariate Cox analysis, the survival analysis results of the training set, test set and the entire data set are shown in Figures 3B,D,F. Comparing Figures 3C,D, we can see that the survival rate of high-risk group and low-risk group is significantly different without Univariate Cox analysis. Therefore, we choose not to add single factor cox analysis when building the model.

After patients were divided into high- and low-risk groups, Kaplan-Meier survival analysis was used to compare the survival differences between the two groups. Survival analysis results showed statistically significant difference in survival rate between the high- and low-risk groups ($p < 0.001$; Figure 3A). The high-risk group had lower overall survival rate and poorer prognosis. The median survival was more than 10 years and around 3 years in the low- and high-risk groups, respectively. The three- and 5-year survival rates of the low-risk group were 88% and 80%, respectively. The three- and 5-year survival rates of the high-risk group were 50% and 25%, respectively. The ROC curve was computed to assess the accuracy of the prognostic model. The AUC values of the 1-, 3-, and 5-year overall survival rates were 0.80, 0.81, and 0.82, respectively (Figure 4A), showing that the prognostic model had good accuracy.

3.4 Independent prognostic analysis of training dataset

Multivariate analysis was used to evaluate the independent prediction capacity of the model and the clinical characteristics. Clinical data of colon adenocarcinoma patients were downloaded from the TCGA database. Samples with missing clinical data were deleted to obtain 341 samples and their corresponding clinical data, including age, gender, stage, T, M, N, and risk score. Age is used as a numerical variable. Female in

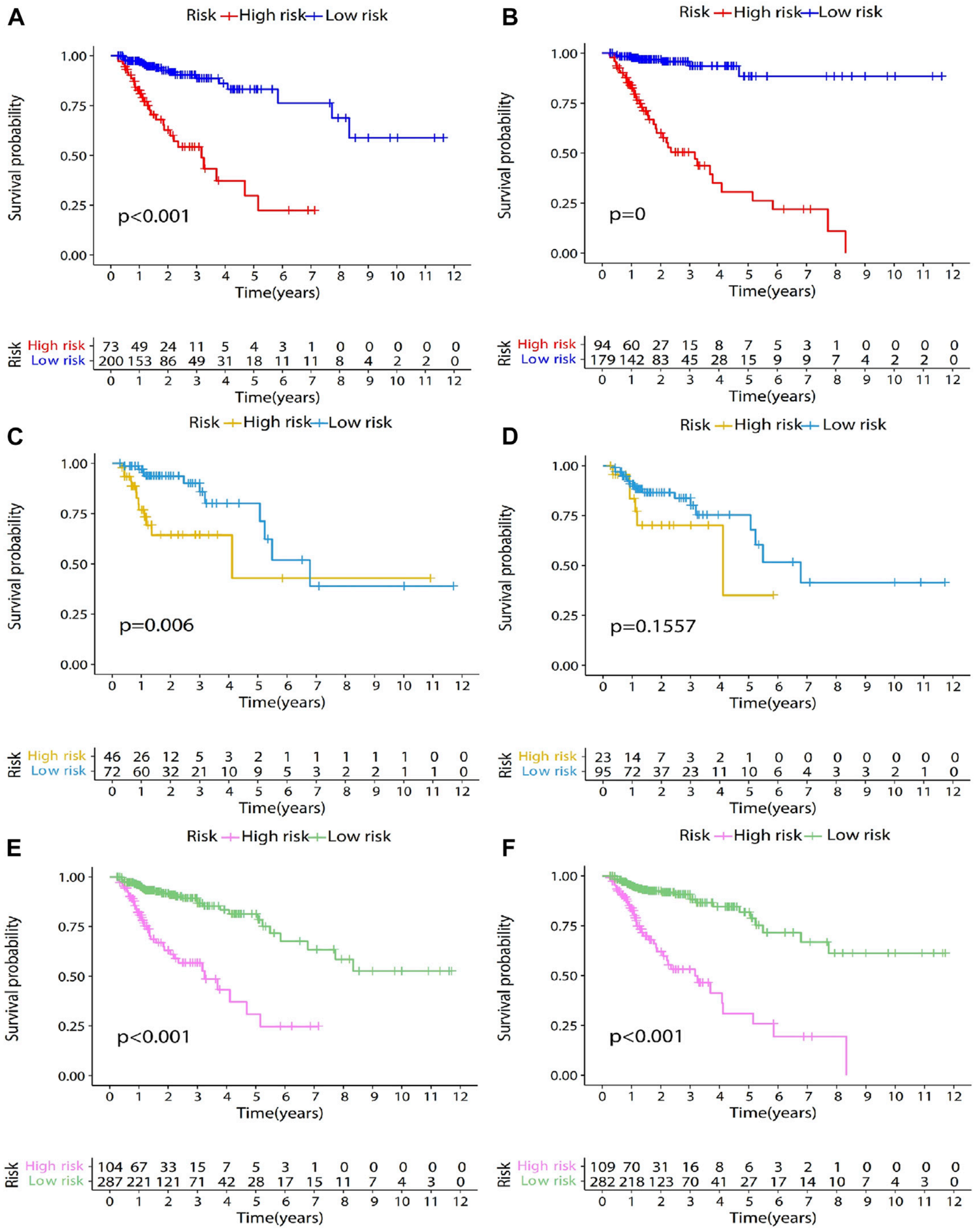


FIGURE 3
Survival analysis results of training dataset, testing dataset, and entire TCGA dataset. The genes were screened by LASSO regression, and the model was constructed by multifactor cox regression. The survival analysis results of the training dataset, the testing dataset, and the entire TCGA dataset are shown in Figures 3 **(A,C,E)**. After screening the genes through univariate Cox analysis, the survival analysis results of the training dataset, testing dataset and the entire TCGA dataset are shown in Figures 3 **(B,D,F)**. In the survival analysis chart, the x-coordinate represents the survival time, in years. The y-coordinate represents the survival probability. The patients were divided into two groups according to the optimal cut-off value. They are low-risk group and high-risk group. *p*-value represents the difference in survival between the two groups. At the bottom of the figure is a table. The abscissa is the survival time in years. The ordinate is the high-risk group and low-risk group, and the value represents the number of patients remaining at each time point.

gender is represented by 0 and male by 1. Each stage in the T, M, N and stage is represented by corresponding Arabic numerals. Multivariate analysis showed that the *p*-values of age, T, and risk score were all less than 0.05 and were independent prognostic factors (Table 3) that predicted patients' prognosis.

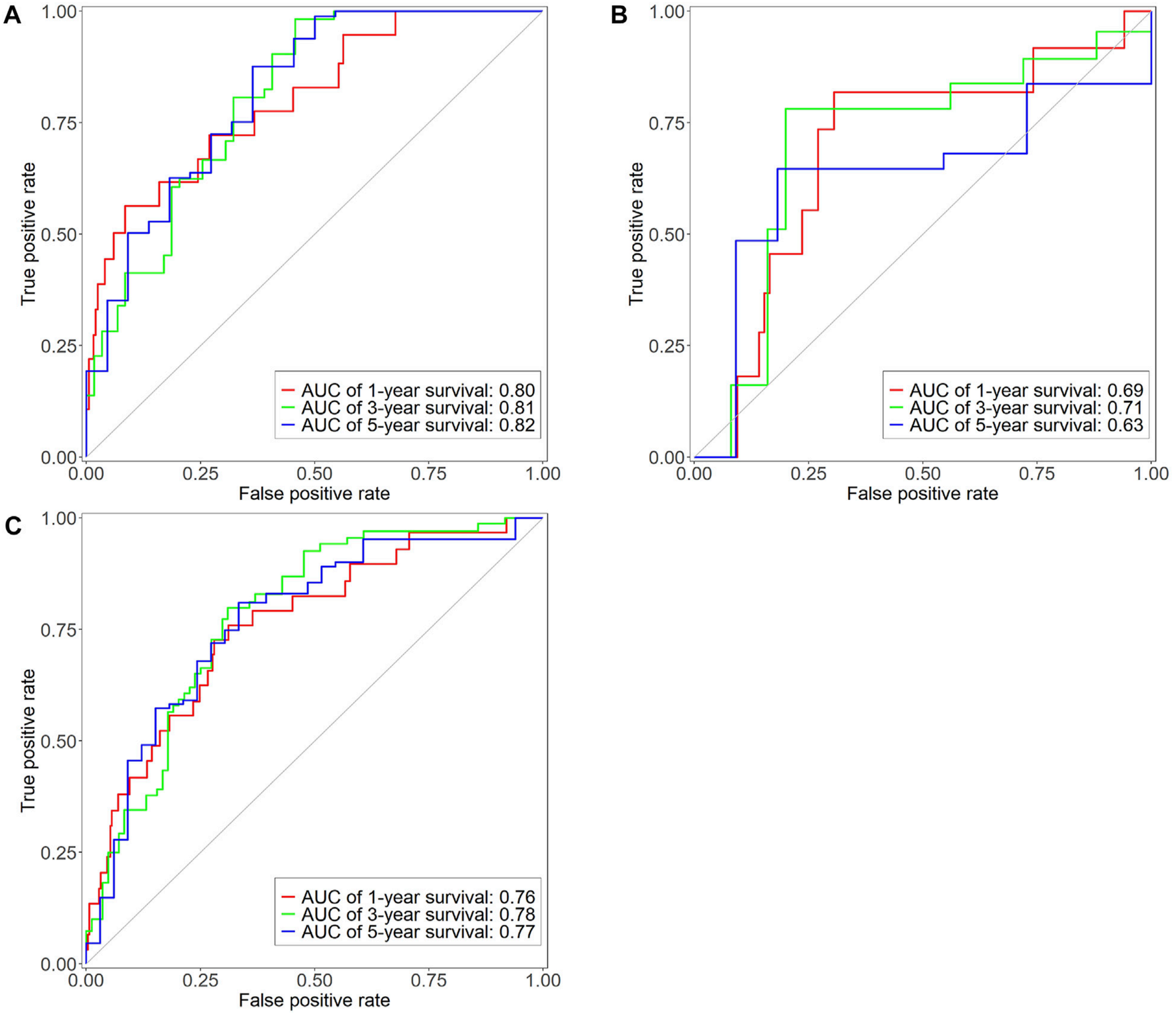


FIGURE 4
Time-dependent ROC curve of training dataset **(A)**, testing dataset **(B)**, and entire TCGA dataset **(C)**. *X*-axis represents false positive rate, *y*-axis represents true positive rate. Red, green and blue represent the curves of 1 year, 3 years and 5 years respectively.

3.5 Predictive model validation

The testing dataset (118 samples) and the entire TCGA dataset (391 samples) were used as validation sets for the prognostic model to evaluate model accuracy. The testing dataset included 118 colon adenocarcinoma patient samples. The risk score of each sample was calculated based on the same formula (Formula (4). The optimal cut-off value of risk score is determined through the surv_cutpoint function of R. The optimal cut-off value of testing dataset is 1.43 (Figure 2B). The optimal cut-off value was used to divide 118 patient samples into two groups, namely, the high- (n = 46) and low-risk groups (n = 72). Kaplan-Meier survival analysis was used to compare survival differences between the two groups. Survival analysis results showed differences in survival rate between the two groups ($p < 0.05$; Figure 3C). Overall survival of the high-risk group was lower, and the prognosis was worse. Median survival was more than 6 and 4 years in the low- and high-risk groups, respectively. The three- and 5-year survival rates of the low-risk group were 86% and 70%, respectively, while the three- and 5-year survival rates of the high-risk group were <65% and <40%, respectively. The reason for the intersection of survival curves at the end may have resulted from the low sample size. Figure 4B shows the ROC curve of the testing dataset. The AUC of the 3-year overall survival rate was 0.71. As the sample size was too small, fewer samples had overall survival rates of 1 and 5 years, so the AUC of 1-year and 5-year were low.

The entire TCGA set included 391 colon adenocarcinoma patient samples. The risk score of each sample was calculated based on Formula (4). The optimal cut-off value of risk score is determined through the surv_cutpoint function of R. The optimal cut-off value of entire TCGA set is 2.02 (Figure 2C). The optimal cut-off value was used to divide the 391 patient samples into two groups, namely, the high- (n = 104) and low-risk groups (n = 287). Kaplan-Meier survival analysis was used to compare the survival differences between the two groups. The survival

TABLE 3 Multivariate independent prognosis analysis results of training dataset.

Variable	HR	95% CI of HR	*p*-value
Age	1.043	1.012–1.074	**
Gender (Female vs. Male)	0.795	0.415–1.523	ns
Stage	1.062	0.370–3.047	ns
T	2.626	1.274–5.414	**
M	2.155	0.538–8.632	ns
N	1.409	0.717–2.769	ns
Risk score	1.004	1.001–1.007	**

**$p < 0.01$; ns, no significance.

analysis results showed differences in survival rate between the two groups ($p < 0.001$; Figure 3E). Overall survival of the high-risk group was lower, and the prognosis was worse. The median survival was more than 10 and 3 years in the low- and high-risk groups, respectively. The three- and 5-year survival rates of the low-risk group were 87% and 78%, respectively. The three- and 5-year survival rates of the high-risk group were 53% and 25%, respectively. Figure 4C shows ROC curves of the entire TCGA dataset. AUC values of the 1-, 3-, and 5-year overall survival rates were 0.76, 0.78, and 0.77, respectively, showing that the prognostic model had good accuracy.

4 Discussion

In this study, we found nine prognosis-related immune genes (*ANGPTL4*, *IDO1*, *NOX1*, *CXCL3*, *LTB4R*, *IL1RL2*, *CD72*, *NOS2*, and *NUDT6*), and we calculated the risk score according to their gene expression and correlation coefficient. Previous experiments have shed light on aberration in these immune genes can lead to tumorigenesis and tumour progression.

Nakayama et al. studied the expression of *ANGPTL4* in colorectal cancer and showed that its expression is associated with venous and lymphatic invasion and that it promotes distal metastasis, i.e., *ANGPTL4* is one critical factor of colorectal cancer progression (Nakayama et al., 2011). Huang et al. showed that *ANGPTL4* expression was more frequent in colorectal cancer tissues than in healthy tissues and that it mediates metastasis through the cytoskeleton signalling pathway to promote colorectal cancer invasion and metastasis (Huang et al., 2012).

Bishnupuri et al. found that *IDO1* activity in epithelial cells and kynurenine pathway metabolites activate tumour epithelial PI3K-Akt signalling, which promotes cell proliferation and anti-apoptosis, thus promoting colon tumorigenesis (Bishnupuri et al., 2019). Thaker et al. found that *IDO1* directly promotes tumour growth and tumour epithelial proliferation in a cell-independent manner through the synthesis of uric acid metabolites and activation of β-catenin signalling, showing that *IDO1* can be a potential therapeutic target (Thaker et al., 2013).

Wang et al. found that *NOX1* regulates colorectal cancer cell proliferation and invasion through the *ADAM17-EGFR-PI3K-Akt* axis to promote colorectal cancer metastasis, showing that *NOX1* can also be a potential target in colorectal cancer treatment (Wang et al., 2016). Ohata et al. studied the biological pathways of cancer stem cell proliferation and demonstrated that *NOX1* induces mTORC1 activation through lysosomal S100A9 oxidation and promotes colon cancer proliferation (Ohata et al., 2019).

According to Farquharson et al., insulin and adiponectin can regulate the expression level of *CXCL3* and thereby participate in colorectal cancer tumorigenesis (Farquharson et al., 2012). Liao et al. showed that *CXCL3* can bind to *CXCLR2* on myeloid-derived suppressor cells to promote its migration to the tumour microenvironment (Liao et al., 2019).

LTB4R is a receptor of leukotriene B4 and exists in two forms. One is the high-affinity LTB4 receptor *BLT1*, which is expressed in different leukocyte subsets and is responsible for LTB4-dependent leukocyte migration. The other is the low-affinity LTB4 receptor *BLT2*, which is expressed in epidermal keratinocytes and epithelial cells and has wound healing and epidermal barrier functions (Yokomizo et al., 2018). Sharma et al. showed that *BLT1* expression in CD8+T cells plays an important role in tumour metastasis (Sharma et al., 2013). Chheda et al. found that *BLT1* plays a critical role in regulating the migration of cytotoxic T lymphocytes to tumours and anti-tumour immunity (Chheda et al., 2016).

Tomuschat et al. studied the expression of *IL1RL2* in patients with congenital Hirschsprung's disease (Tomuschat et al., 2017). Their results showed that *IL1RL2* is an important mediator of inflammatory responses and that a significant reduction in its expression can increase inflammatory responses and cause changes in mucosal healing, thereby resulting in susceptibility to Hirschsprung-associated enterocolitis. In addition, Penha et al. showed that *IL1RL2* is expressed in intestinal T lymphocytes and can induce CD4$^+$ lymphocyte proliferation, relating to human intestinal diseases (Penha et al., 2016). *CD72* is expressed by various immune, inflammatory and epithelial cells. CD100-CD72 interaction can regulate the intensity of B cell receptor signal pathway, enhance cell activation and maintain immune homeostasis (Wu et al., 2016).

5 Conclusion

We downloaded transcriptome data of colorectal cancer healthy tissues from GTEx and then downloaded transcriptome data and clinical data of colorectal adenocarcinoma patients from TCGA. LASSO regression was carried out on DEGs between healthy samples and tumor samples to identify prognosis-related immune genes. Multivariate Cox regression and prognosis-related immune genes (*ANGPTL4*, *IDO1*, *NOX1*, *CXCL3*, *LTB4R*, *IL1RL2*, *CD72*, *NOS2* and *NUDT6*) were used to construct an immune-related prognosis risk score model for colon adenocarcinoma patients. This score was used to divide colon adenocarcinoma patients into high- and low-risk groups. Survival analysis found that the high-risk group had lower overall survival rate and poorer prognosis.

To validate the prognostic value of the model, we computed ROC curves. The model AUC values of the 1-, 3-, and 5-year overall survival rates were 0.76, 0.78, and 0.77, respectively, showing good prediction results for patients' prognosis. Further multivariate analysis demonstrated that the risk score was an independent prognostic factor. A validation dataset was used to further demonstrate the accuracy of this score. The model also identified immune genes as potential prognostic biomarkers and therapeutic targets in colorectal cancer, however, further validation in clinical trials is required, the mechanism by which immune genes affect cancer progress should be further studied.

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.

Author contributions

DA, LCX, and MW conceived and designed the study. MW performed the analyses and summarized the data. LCX and DA supervised the study. DA, MW, and LCX wrote the manuscript with inputs from QZ, LC, YW, and XL. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by grants from the National Natural Science Foundation of China (grant number 61873027) and Open Project of the National Engineering Laboratory for Agri-product Quality Traceability (No.AQT-2020-YB6).

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/fgene.2023.1148470/full#supplementary-material

References

Andre, T., Boni, C., Mounedji-Boudiaf, L., Navarro, M., Tabernero, J., Hickish, T., et al. (2004). Oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment for colon cancer. *N. Engl. J. Med.* 350 (23), 2343–2351. doi:10.1056/NEJMoa032709

Arnold, M., Sierra, M. S., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2017). Global patterns and trends in colorectal cancer incidence and mortality. *Gut* 66 (4), 683–691. doi:10.1136/gutjnl-2015-310912

Bedognetti, D., Hendrickx, W., Marincola, F. M., and Miller, L. D. (2015). Prognostic and predictive immune gene signatures in breast cancer. *Curr. Opin. Oncol.* 27 (6), 433–444. doi:10.1097/CCO.0000000000000234

Bishnupuri, K. S., Alvarado, D. M., Khouri, A. N., Shabsovich, M., Chen, B., Dieckgraefe, B. K., et al. (2019). Ido1 and kynurenine pathway metabolites activate PI3K-akt signaling in the neoplastic colon epithelium to promote cancer cell proliferation and inhibit apoptosis. *Cancer Res.* 79 (6), 1138–1150. doi:10.1158/0008-5472.CAN-18-0668

Blangero, Y., Rabilloud, M., Ecochard, R., and Subtil, F. (2020). A Bayesian method to estimate the optimal threshold of a marker used to select patients' treatment. *Stat. Methods Med. Res.* 29 (1), 29–43. doi:10.1177/0962280218821394

Chheda, Z. S., Sharma, R. K., Jala, V. R., Luster, A. D., and Haribabu, B. (2016). Chemoattractant receptors BLT1 and CXCR3 regulate antitumor immunity by facilitating CD8+ T cell migration into tumors. *J. Immunol.* 197 (5), 2016–2026. doi:10.4049/jimmunol.1502376

Coppede, F. (2014). The role of epigenetics in colorectal cancer. *Expert Rev. Gastroenterol. Hepatol.* 8 (8), 935–948. doi:10.1586/17474124.2014.924397

Dalerba, P., Sahoo, D., Paik, S., Guo, X., Yothers, G., Song, N., et al. (2016). CDX2 as a prognostic biomarker in stage II and stage III colon cancer. *N. Engl. J. Med.* 374 (3), 211–222. doi:10.1056/NEJMoa1506597

den Uil, S. H., Coupe, V. M., Linnekamp, J. F., van den Broek, E., Goos, J. A., Delis-van Diemen, P. M., et al. (2016). Loss of KCNQ1 expression in stage II and stage III colon cancer is a strong prognostic factor for disease recurrence. *Br. J. Cancer* 115 (12), 1565–1574. doi:10.1038/bjc.2016.376

Emmert-Streib, F., and Dehmer, M. (2019). High-dimensional LASSO-based computational regression models: Regularization, shrinkage, and selection. *Mach. Learn. Knowl. Extr.* 1 (1), 359–383. doi:10.3390/make1010021

Farquharson, A. J., Steele, R. J., Carey, F. A., and Drew, J. E. (2012). Novel multiplex method to assess insulin, leptin and adiponectin regulation of inflammatory cytokines associated with colon cancer. *Mol. Biol. Rep.* 39 (5), 5727–5736. doi:10.1007/s11033-011-1382-1

Galon, J., Angell, H. K., Bedognetti, D., and Marincola, F. M. (2013). The continuum of cancer immunosurveillance: Prognostic, predictive, and mechanistic signatures. *Immunity* 39 (1), 11–26. doi:10.1016/j.immuni.2013.07.008

Hashiguchi, Y., Muro, K., Saito, Y., Ito, Y., Ajioka, Y., Hamaguchi, T., et al. (2020). Japanese Society for Cancer of the Colon and Rectum (JSCCR) guidelines 2019 for the treatment of colorectal cancer. *Int. J. Clin. Oncol.* 25 (1), 1–42. doi:10.1007/s10147-019-01485-z

Huang, X. F., Han, J., Hu, X. T., and He, C. (2012). Mechanisms involved in biological behavior changes associated with Angptl4 expression in colon cancer cell lines. *Oncol. Rep.* 27 (5), 1541–1547. doi:10.3892/or.2012.1672

Kandimalla, R., Linnekamp, J. F., van Hooff, S., Castells, A., Llor, X., Andreu, M., et al. (2017). Methylation of WNT target genes AXIN2 and DKK1 as robust biomarkers for recurrence prediction in stage II colon cancer. *Oncogenesis* 6 (4), e308. doi:10.1038/oncsis.2017.9

Kidd, A. C., McGettrick, M., Tsim, S., Halligan, D. L., Bylesjo, M., and Blyth, K. G. (2018). Survival prediction in mesothelioma using a scalable lasso regression model: Instructions for use and initial performance using clinical predictors. *BMJ Open Respir. Res.* 5 (1), e000240. doi:10.1136/bmjresp-2017-000240

Kleinbaum, D. G., and Klein, M. (2012). "The cox proportional hazards model and its characteristics," in *Survival analysis*, 97–159.

Lao, V. V., and Grady, W. M. (2011). Epigenetics and colorectal cancer. *Nat. Rev. Gastroenterol. Hepatol.* 8 (12), 686–700. doi:10.1038/nrgastro.2011.173

Li, C., Shen, Z., Zhou, Y., and Yu, W. (2018). Independent prognostic genes and mechanism investigation for colon cancer. *Biol. Res.* 51 (1), 10. doi:10.1186/s40659-018-0158-7

Liao, W., Overman, M. J., Boutin, A. T., Shang, X., Zhao, D., Dey, P., et al. (2019). KRAS-IRF2 Axis drives immune suppression and immune therapy resistance in colorectal cancer. *Cancer Cell* 35 (4), 559–572.e7. doi:10.1016/j.ccell.2019.02.008

Lin, P., Guo, Y. N., Shi, L., Li, X. J., Yang, H., He, Y., et al. (2019). Development of a prognostic index based on an immunogenomic landscape analysis of papillary thyroid cancer. *Aging (Albany NY)* 11 (2), 480–500. doi:10.18632/aging.101754

Lin, Y., Pan, X., Chen, Z., Lin, S., and Chen, S. (2020). Identification of an immune-related nine-lncRNA signature predictive of overall survival in colon cancer. *Front. Genet.* 11, 318. doi:10.3389/fgene.2020.00318

Miao, Y., Wang, J., Ma, X., Yang, Y., and Mi, D. (2020). Identification prognosis-associated immune genes in colon adenocarcinoma. *Biosci. Rep.* 40 (11). doi:10.1042/BSR20201734

Mounir, M., Lucchetta, M., Silva, T. C., Olsen, C., Bontempi, G., Chen, X., et al. (2019). New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Comput. Biol.* 15 (3), e1006701. doi:10.1371/journal.pcbi.1006701

Nakayama, T., Hirakawa, H., Shibata, K., Nazneen, A., Abe, K., Nagayasu, T., et al. (2011). Expression of angiopoietin-like 4 (ANGPTL4) in human colorectal cancer: ANGPTL4 promotes venous invasion and distant metastasis. *Oncol. Rep.* 25 (4), 929–935. doi:10.3892/or.2011.1176

Nishihara, R., Wu, K., Lochhead, P., Morikawa, T., Liao, X., Qian, Z. R., et al. (2013). Long-term colorectal-cancer incidence and mortality after lower endoscopy. *N. Engl. J. Med.* 369 (12), 1095–1105. doi:10.1056/NEJMoa1301969

Ohata, H., Shiokawa, D., Obata, Y., Sato, A., Sakai, H., Fukami, M., et al. (2019). NOX1-Dependent mTORC1 activation via S100A9 oxidation in cancer stem-like cells leads to colon cancer progression. *Cell Rep.* 28 (5), 1282–1295.e8. doi:10.1016/j.celrep.2019.06.085

Penha, R., Higgins, J., Mutamba, S., Barrow, P., Mahida, Y., and Foster, N. (2016). IL-36 receptor is expressed by human blood and intestinal T lymphocytes and is dose-dependently activated via IL-36β and induces CD4+ lymphocyte proliferation. *Cytokine* 85, 18–25. doi:10.1016/j.cyto.2016.05.023

Qu, L., Wang, Z. L., Chen, Q., Li, Y. M., He, H. W., Hsieh, J. J., et al. (2018). Prognostic value of a long non-coding RNA signature in localized clear cell renal cell carcinoma. *Eur. Urol.* 74 (6), 756–763. doi:10.1016/j.eururo.2018.07.032

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 (7), e47. doi:10.1093/nar/gkv007

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., et al. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.* 19 (5), 619–625. doi:10.1038/nm.3175

Sharma, R. K., Chheda, Z., Jala, V. R., and Haribabu, B. (2013). Expression of leukotriene B_4 receptor-1 on $CD8^+$ T cells is required for their migration into tumors to elicit effective antitumor immunity. *J. Immunol.* 191 (6), 3462–3470. doi:10.4049/jimmunol.1300967

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Tarazona, N., Gimeno-Valiente, F., Gambardella, V., Huerta, M., Rosello, S., Zuniga, S., et al. (2020). Detection of postoperative plasma circulating tumour DNA and lack of CDX2 expression as markers of recurrence in patients with localised colon cancer. *ESMO Open* 5 (5), e000847. doi:10.1136/esmoopen-2020-000847

Thaker, A. I., Rao, M. S., Bishnupuri, K. S., Kerr, T. A., Foster, L., Marinshaw, J. M., et al. (2013). Ido1 metabolites activate beta-catenin signaling to promote cancer cell proliferation and colon tumorigenesis in mice. *Gastroenterology* 145 (2), 416–425.e1-4. doi:10.1053/j.gastro.2013.05.002

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* 16 (4), 385–395. doi:10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

Tomuschat, C., O'Donnell, A. M., Coyle, D., and Puri, P. (2017). Altered expression of IL36γ and IL36 receptor (IL1RL2) in the colon of patients with Hirschsprung's disease. *Pediatr. Surg. Int.* 33 (2), 181–186. doi:10.1007/s00383-016-4011-1

Wang, B., Li, J., and Wang, X. (2021). Change point detection in Cox proportional hazards mixture cure model. *Stat. Methods Med. Res.* 30 (2), 440–457. doi:10.1177/0962280220959118

Wang, G. H., Yao, L., Xu, H. W., Tang, W. T., Fu, J. H., Hu, X. F., et al. (2013). Identification of MXRA5 as a novel biomarker in colorectal cancer. *Oncol. Lett.* 5 (2), 544–548. doi:10.3892/ol.2012.1038

Wang, H. P., Wang, X., Gong, L. F., Chen, W. J., Hao, Z., Feng, S. W., et al. (2016). Nox1 promotes colon cancer cell metastasis via activation of the ADAM17 pathway. *Eur. Rev. Med. Pharmacol. Sci.* 20 (21), 4474–4481.

Wei, H. T., Guo, E. N., Liao, X. W., Chen, L. S., Wang, J. L., Ni, M., et al. (2018). Genome-scale analysis to identify potential prognostic microRNA biomarkers for predicting overall survival in patients with colon adenocarcinoma. *Oncol. Rep.* 40 (4), 1947–1958. doi:10.3892/or.2018.6607

Woischke, C., Blaj, C., Schmidt, E. M., Lamprecht, S., Engel, J., Hermeking, H., et al. (2016). CYB5R1 links epithelial-mesenchymal transition and poor prognosis in colorectal cancer. *Oncotarget* 7 (21), 31350–31360. doi:10.18632/oncotarget.8912

Wright, M. N., Dankowski, T., and Ziegler, A. (2017). Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Stat. Med.* 36 (8), 1272–1284. doi:10.1002/sim.7212

Wu, M., Li, J., Gao, Q., and Ye, F. (2016). The role of Sema4D/CD100 as a therapeutic target for tumor microenvironments and for autoimmune, neuroimmune and bone diseases. *Expert Opin. Ther. Targets* 20 (7), 885–901. doi:10.1517/14728222.2016.1139083

Yokomizo, T., Nakamura, M., and Shimizu, T. (2018). Leukotriene receptors as potential therapeutic targets. *J. Clin. Invest* 128 (7), 2691–2701. doi:10.1172/JCI97946

Zuo, S., Dai, G., and Ren, X. (2019). Identification of a 6-gene signature predicting prognosis for colorectal cancer. *Cancer Cell Int.* 19, 6. doi:10.1186/s12935-018-0724-7

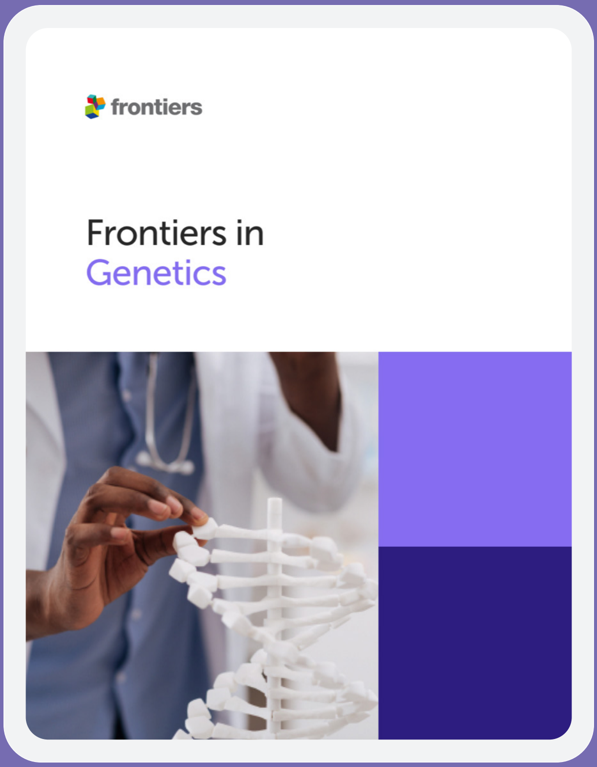
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Genetics

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact