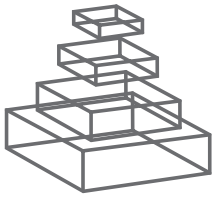
frontiers

RESEARCH TOPICS

COMPUTATIONAL APPROACHES IN AID OF ADVANCING UNDERSTANDING IN PLANT PHYSIOLOGY

Hosted by
Alisdair Fernie

frontiers in
PLANT SCIENCE



frontiers

Cover image provided by Ibbl sarl, Lausanne CH

ISSN 1664-8714
ISBN 978-2-88919-017-1
DOI 10.3389/978-2-88919-017-1

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

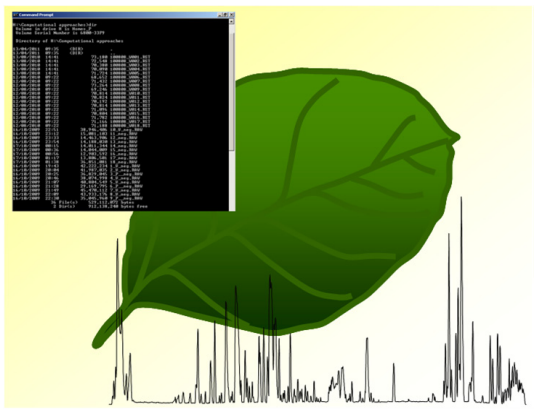
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - 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: researchtopics@frontiersin.org

COMPUTATIONAL APPROACHES IN AID OF ADVANCING UNDERSTANDING IN PLANT PHYSIOLOGY

Hosted By
Alisdair Fernie, Max Planck Institut for Plant Physiology, Germany



The recent data flood has required greater and greater reliance on computational usage in plant biology. This special issue will focus on the utility of computational approaches across the breadth of modern plant biology with particular focus on the following areas:

(i) Comparative genomics- gene family size in the green lineage
(ii) Adaptive evolution - specifics of development
(iii) Adaptive evolution - specifics of secondary metabolism
(iv) Translational biology- co-response analysis from arabidopsis outwards
(v) Conserved(and differential) transcriptional response to stress
(vi) Transcriptomics databases
(vii) Translatomics
(ix) Proteomics- abundance
(x) Proteomics- location,
(xi) Proteomics- interactions
(xii) Proteomics databases
(xiii) The activome
(xiv) Metabolite-abundance
(xv) Metabolite- location
(xvi) Experimental flux calculations,
(xvii) Advanced metabolomic technologies
(xviii) Metabolite databases
(xix) Genome wide metabolic modelling

Table of Contents

frontiers in
PLANT SCIENCE

EDITORIAL
published: 25 November 2011
doi: 10.3389/fpls.2011.00078

Editorial overview – computational approaches in aid of advancing understanding in plant physiology

Alisdair R. Fernie*

Max-Planck-Institute of Molecular Plant Physiology, Potsdam, Germany
**Correspondence: fernie@mpimp-golm.mpg.de*

The exact impact of computers on any branch of science is impossible to predict, however, as highlighted in a recent article in Science the internet has already had a profound effect on the way academics use their brains as information retrieval units (Sparrow et al., 2011). Whilst Coleridge is often cited as the last man who read (or would have been capable of reading) every article in print, the advent of the internet dwarfs even the rapid expansion of the printing presses. It's impact on science, though vast, is clearly incalculable. From a data, as opposed to a text, perspective computation has undoubtedly greatly enabled genomics – a discipline that would certainly not exist in its current form without recent advances in computational power. For example the recently sequenced potato genome (Xu et al., 2011) could easily be stored on a standard laptop computer (Usadel, personal communication).

Within this special issue of Frontiers are collected both reviews and primary research papers in which computational support played a major role. Whilst the call for papers was open to submissions from any plant biological discipline the collected papers are focused on next generation RNA sequencing, co-expression analysis, protein sub-cellular compartment prediction, sub-cellular metabolite analyses, the expansion of capacities for metabolite profiling, translational metabolomics, and metabolic flux analyses. There is thus a clear bias toward studies focused on metabolites, however, it is likely that this reflects their relative complexity both in terms of chemical structure and difficulty of analyses (Stitt and Fernie, 2003; Matsuda and Saito, 2010) as much as the interest in these problems from a biological standpoint.

The article by Jimenez-Gomez provides a detailed perspective of how computational analysis has begun, and will continue, to revolutionize the analysis of continuous phenotypic trait variation. It highlights the current state of the art in using next generation sequencing methods for the analysis of expression quantitative trait loci (eQTL) detailing recent technical, computational, and technical innovations which have facilitated the detection of molecular markers at higher resolution than previously achievable (Jimenez-Gomez, 2011). In addition to providing examples of how this works within the context of species for which a reference genome sequence is present Jimenez-Gomez also describes the utility of next generation sequencing in cases where it is absent – a technique which will prove highly useful in addressing the grand challenge of translational biology (Huber, 2011). In addition he describes complexities of next generation sequencing with respect to expression profiling and the identification of allele specific expression. Two further studies in this collection, those or Ruprecht et al. (2011) and Tohge et al. (2011), also address aspects of translational biology by means of co-expression analysis of transcript data and by addressing experimental and computational caveats of the translational application of metabolite profiling protocols, respectively. In the article by Ruprecht et al. (2011) the recently described PlaNet platform (Mutwil et al., 2011) was utilized to perform large scale condition-dependent comparisons of primary and secondary cell wall related cellulose synthase A co-expression networks. The authors used this approach to select genes from gene families that were conserved across seven species to correlate with cellulose synthase A and analyzed cell wall properties of *Arabidopsis* mutant lines of these gene families. One of these lines was demonstrated to be lignin deficient thus demonstrating the utility of this approach and suggesting that it will likely be a highly useful strategy for gene functional annotation. Also taking a cross-species approach, Tohge et al. (2011) analyzed the difficulties of using standard metabolite profiling approaches in cross-species experiments. Presented results support arguments for the need for the adoption of closely controlled empirical adaptations each time a new species or tissue is analyzed (Fernie et al., 2011). Such experiments are required because reliability of protocols for harvesting, handling, and analysis depends on biological features and chemical composition of the plant tissue. Tohge et al. (2011) provide cases studies of two different liquid chromatography mass spectrometry (LC–MS) based metabolomics platforms and four species in order to illustrate how measurement errors can be detected and circumvented.

The article of Voll et al. (2011) investigated the transcriptomic and metabolomics response of three diverse pathosystems, the barley powdery mildew fungus (*Blumeria graminis* f. sp. *hordei*), the corn smut fungus *Ustilago maydis,* and the maize anthracnose fungus *Colletotrichum graminicola.* Intriguingly, analysis of 42 water-soluble metabolites, allowed the separation of early biotrophic from late biotrophic interactions by hierarchical cluster analysis and principal component analysis, irrespective of the plant host. Both metabolome and transcript data were employed to generate models of leaf primary metabolism during early biotrophy for the three investigated interactions and these models will likely prove highly important for future studies of these pathosystems.

Sticking with metabolomics, Matsuda and co-workers present a novel framework for automated elucidation of metabolite structures in LC–MS and a co-responding metabolite ontology system. As a proof of concept the metabolome of 20 *Arabidopsis* accessions was evaluated and 704 metabolites were analyzed (Matsuda et al., 2011). Exact chemical structure determination remains one of the grand challenges of metabolomics and this strategy allowed structural estimates for an impressive 30% of these signals. In a similar vein, the paper of Hummel et al. (2011) describes a novel Ultra Performance Liquid Chromatography-based method as an alternative to widely used direct infusion based shotgun-lipidomics approaches coupled to a database search software which allows both

targeted and non-targeted lipidomic and metabolomics analysis of all kinds of mass spectral data. Widespread adoption of either the Matsuda et al. (2011) or the Hummel et al. (2011), strategy will likely greatly enhance knowledge retrieval from data acquired by similar methods.

The identification of biomarkers and complex metabolic signatures is receiving increasing attention. One of the first such searches defined the metabolite signature associated with high growth rates in *Arabidopsis* (Meyer et al., 2007) whilst recent studies have identified starch and protein levels to be key integrators of metabolism and growth (Sulpice et al., 2009, 2010). In this issue Kusano et al. (2011) describe the use of gas chromatography mass spectrometry (GC–MS) aligned with multivariate projection methods to define a metabolite signature associated with short-day induced growth cessation in aspen. In this paper the authors use this casestudy to highlight the power of statistical data analyses including principal component analysis (PCA) and orthogonal projection to latent structures (OPLS) in data interpretation.

One of the greatest challenges we currently face in plant biology is that of understanding spatial compartmentation of metabolic pathways and indeed of any other biological function. This problem is particularly acute in plants due to the myriad of cell types and sub-cellular compartments they contain (Fernie, 2007; Lunn, 2007). Whilst considerable advances have been made in determining protein location using a combination of reporter gene constructs, subcellular proteomics, and *in silico* sequence analysis (see for example Millar et al., 2009), our inventories for protein content remain incomplete and in some instances inaccurate. In the paper by Ryngajllo et al. (2011), the authors explored whether gene expression data could be harnessed to enhance bioinformatics location prediction performance. In this paper they show that utilizing their approach they could greatly enhance plastid localization prediction with notable improvements for the mitochondrion, Golgi apparatus, and plasma membrane. On the basis of these results they created the SLocX sub-cellular location predictor engine that even works in cases where only partial gene sequences are available suggesting that it may additionally have great utility for non-sequenced or poorly annotated genomes. The sub-cellular localization of metabolites is additionally currently seeing somewhat of a renaissance. Early plant studies were initiated in the 80s (Gerhardt and Heldt, 1984), however, these techniques were not commonly adopted prior to the advent of metabolomics (Farre et al., 2001). In their article Klie et al. (2011) discusses computational aspects associated with the non-aqueous fractionation method. In addition they provide a new version of the BestFit command line tool for calculation and evaluation of sub-cellular metabolite distributions and also discuss caveats and benefits of the approach.

The final two articles of Schwender (2011) and Sweetlove and Ratcliffe (2011) describe two different approaches for assessing metabolic fluxes. In the first, the ^{13}C-metabolic flux profiling approach is reviewed and Schwender describes the principle of the approach before outlining how the model boundaries are defined and the need for reaction stoichiometries for this approach. He also details computational aspects of ^{13}C-metabolic flux profiling as defined by the modeling framework of Wiechert et al. (2001), providing recent examples of network definition and validation in plants before ending with a perspective for future developments of this approach. The article of Sweetlove and Ratcliffe (2011) reviews the complementary technique of flux balance modeling. They define flux balance modeling as a constraints-based approach in which steady-state fluxes are predicted using optimization algorithms within an experimentally bounded solution space. Sweetlove and Ratcliffe argue that despite the undoubted power of the approach described by Schwender it has several limitations and postulate that these have driven to the adoption of alternate flux balance based approaches. They provide a comprehensive review of the field from its early beginnings in microbial systems to the several plant models which have been published in the last 2 years, covering modeling of specific cell types, accounting for cell maintenance energy costs, and the evaluation of metabolic efficiency via this approach. In concluding their article Sweetlove and Ratcliffe make convincing arguments for the adoption of flux balance modeling as an important complement to ^{13}C-metabolic flux profiling both for understanding metabolic regulation and ultimately as a means to determine targets for rational crop improvement.

When taken as a whole these articles cover many, although by no means all, of the ways in which computational approaches are rapidly advancing our understanding of plant function. As well as providing informative overviews of the fields defined in the opening paragraphs several of the articles also describe and provide software which should allow a relatively simple adoption of the described techniques by researchers from other laboratories. I thank all the authors for their support in putting together this special issue and hope people enjoy reading it as much as I enjoyed editing it.

REFERENCES

Farre, E. M., Tiessen, A., Roessner, U., Geigenberger, P., Trethewey, R. N., and Willmitzer, L. (2001). Analysis of the compartmentation of glycolytic intermediates, nucleotides, sugars, organic acids, amino acids and sugar alcohols in potato tubers using a nonaqueous fractionation method. *Plant Physiol.* 127, 685–700.

Fernie, A. (2007). The future of metabolic phytochemistry: larger numbers of metabolites, higher resolution, greater understanding. *Phytochemistry* 68, 2861–2880.

Fernie, A. R., Aharoni, A., Willmitzer, L., Stitt, M., Tohge, T., Kopka, J., Carroll, A. J., Saito, K., Fraser, P. D., and DeLuca, V. (2011). Recommendations for reporting metabolite data. *Plant Cell* 23, 2477–2482.

Gerhardt, R., and Heldt, H. W. (1984). Measurement of subcellular metabolite levels in leaves by fractionation of freeze stopped material in nonaqueous media. *Plant Physiol.* 75, 542–547.

Huber, S. (2011). Grand challenges in plant physiology: the underpinning of translational research. *Front. Plant Sci.* 2:48. doi: 10.3389/fpls.2011.00048

Hummel, J., Segu, S., Irgang, S., Jueppner, J., and Giavalisco, P. (2011). Ultra performance liquid chromatography and high resolution mass spectrometry fort he analysis of plant lipids. *Front. Plant Sci.* 2:54. doi: 10.3389/fpls.2011.00054

Jimenez-Gomez, J. M. (2011). Next generation quantitative genetics in plants. *Front. Plant Sci.* 2:77. doi: 10.3389/fpls.2011.00077

Klie, S. Krueger, S., Krall, L., Giavalisco, P., Flügge, U. I., Willmitzer, L., and Steinhauser, D. (2011). Analysis of the compartmentalized metabolome – a validation oft he non-aqueous fractionation technique. *Front. Plant Sci.* 2:55. doi: 10.3389/fpls.2011.00055

Kusano, M., Jonsson, P., Fukushima, A., Gulberg, J., Sjöstrom, M., Trygg, J., and Moritz, T. (2011). Metabolite signature during short-day induced growth cessation in *Populus*. *Front. Plant Sci.* 2:29. doi: 10.3389/fpls.2011.00029

Lunn, J. E. (2007). Compartmentation in plant metabolism. *J. Exp. Bot.* 58, 35–47.

Matsuda, F., Nakabayashi, R., Sawada, Y., Suzuki, M., Hirai, M. Y., Kanaya, S., and Saito, K. (2011). Mass spectra-based framework for automated structural elucidation of metabolome data to explore phytochemical diversity.

Front. Plant Sci. 2:40. doi: 10.3389/fpls.2011.00040

Matsuda, F., and Saito, K. (2010). Metabolomics for functional genomics, systems biology and biotechnology. *Annu. Rev. Plant Biol.* 61, 463–489.

Meyer, R. C., Steinfath, M., Lisec, J., Becher, M., Witucka-Wall, H., Törjek, O., Fiehn, O., Eckardt, A., Willmitzer, L., Selbig, J., and Altmann, T. (2007). The metabolic signature related to high plant growth rate in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 13, 4759–4764.

Millar, A. H., Carrie, C., Pogson, B., and Whelan, J. (2009). Exploring the function-location nexus: using multiple lines of evidence in defining the subcellular location of plant proteins. *Plant Cell* 21, 1625–1631.

Mutwil, M., Klie, S., Tohge, T., Giorgi, F. M., Wilkins, O., and Campbell, M. M. (2011). PlaNet: Combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* 23, 895–910.

Ruprecht, C., Mutwil, M., Saxe, F., Edler, M., Nikoloski, Z., and Persson, S. (2011). Large-scale co-expression approach to dissect secondary cell wall formation across plant species. *Front. Plant Sci.* 2:23. doi: 10.3389/fpls.2011.00023

Ryngajllo, M., Childs, L., Lohse, M., Giorgi, F. M., Lude, A., Selbig, J., and Usadel, B. (2011). SLocX: predicting subcellular localization of *Arabidopsis* proteins leveraging gene expression data. *Front. Plant Sci.* 2:43. doi: 10.3389/fpls.2011.00043

Schwender, J. (2011). Experimental flux measurements on a network scale. *Front. Plant Sci.* 2:63. doi: 10.3389/fpls.2011.00063

Sparrow, B., Liu, J., and Wegner, D. M. (2011). Google effects on memory: cognitive consequences of having information at our fingertips. *Science* 333, 776–778.

Stitt, M., and Fernie, A. R. (2003). From measurements of metabolites to metabolomics: an "on the fly" perspective illustrated by recent studies of carbon-nitrogen interactions. *Curr. Opin. Biotechnol.* 14, 136–144.

Sulpice, R., Pyl, E. T., Ishihara, H., Trenkamp, S., Steinfath, M., Witucka-Wall, H., Gibon, Y., Usadel, B., Poree, F., Piques, M. C., Von Korff, M., Steinhauser, M. C., Keurentjes, J. J., Guenter, M., Hoehne, M., Selbig, J., Fernie, A. R., Altmann, T., and Stitt, M. (2009). Starch as a major integrator in the regulation of plant growth. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10348–10353.

Sulpice, R., Trenkamp, S, Steinfath, M., Usadel, B., Gibon, Y., Witucka-Wall, H., Pyl, E. T., Tschoep, H., Steinhauser, M. C., Guenther, M., Hoehne, M., Rohwer, J. M., Altmann, T., Fernie, A. R., and Stitt, M. (2010). Network analysis of enzyme activities and metabolite levels and their relationship to biomass in a large panel of *Arabidopsis* accessions. *Plant Cell* 22, 2872–2893.

Sweetlove, L. J., and Ratcliffe, R. G. (2011). Flux-balance modeling of plant metabolism. *Front. Plant Sci.* 2:38. 2:38. doi: 10.3389/fpls.2011.00038

Tohge, T., Mettler, T., Arrivault, S., Carroll, A. J., Stitt, M., and Fernie, A. R. (2011). From models to crop species: caveats and solutions for translational metabolomics. *Front. Plant Sci.* 2:61. doi: 10.3389/fpls.2011.00061

Voll, L. M., Horst, R. J., Voitsik, A.-M., Zajic, D., Samans, B., Pons-Kühnemann, J., Doehlemann, G., Münch, S., Wahl, R., Molitor, A., Hofmann, J., Schmiedl, A., Waller, F., Deising, H. B., Kahmann, R., Kämper, J., Kogel, K.-H., and Sonnewald, U. (2011). Common motifs in the response of cereal primary metabolism to fungal pathogens are not based on similar transcriptional reprogramming. *Front. Plant Sci.* 2:39. doi: 10.3389/fpls.2011.00039

Wiechert, W., Möllney, M., Petersen, S., and De Graaf, A. A. (2001). 13C metabolic flux analysis. *Metab. Eng.* 3, 265–283.

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Orieda, G., Guzman, F., Torres, M., Lozano, R., Ponce, O., Martinez, D., De la Cruz, G., Chakrabartu, S. K., Patil, V. U., Skrvabin, K. G., Kuznetsov, B. B., Ravin, N. V., Kolganova, T. V., Beletsky, A. V., Mardanov, A. V., Di Genova, A., Bolser, D. M., Martin, D. M., Li, G., Yang, Y., Kuang, H., Hu, Q, Xiong, X., Bishop, G. J., Sagredo, B., Majia, N., Zagorski, W., Gromadka, R., Gawor, J, Szczesny, P., Huang, S., Zhang, Z., Liang, C., He, J., Li, Y., He, Y., Xu, J., Zhang, Y., Xie, B., Du, Y., Qu, D., Bonierbale, M., Ghislain, M., Herrera Mdel, R., Giuliano, G., Pietrella, M., Perrotta, G., Facello, P., O'Brien, K., Feingold, S. E., Barreiro, L. E., Massa, G. A., Diambra, L., Whitty, B. R., Vaillancourt, B., Lin, H., Massa, A. N., Geoffroy, M., Lundback, S., Dellapenna, D., Buell, C. R., Sharma, S. K., Marshall, D. F., Waugh, R., Bryan, G. J., Destefanis, M., Nagy, I., Milbourne, D., Thomson, S. J., Fiers, M., Jacobs, J. M., Nielsen, K. L., Sonderkaer, M., Iovene, M., Torres, G. A., Jiang, J., Veilleux, R. E., Bachem, C. W., de Boer, J., Borm, T., Kloosterman, B., van Eck, H., Datema, E., Hekkert, B. L., Goverse, A., van Ham, R. C., and Visser, R. G. (2011). Genome sequence and analysis of the tuber crop potato. *Nature* 475, 189–195.

Received: 14 September 2011; accepted: 26 October 2011; published online: 25 November 2011.

Citation: Fernie AR (2011) Editorial overview – computational approaches in aid of advancing understanding in plant physiology. Front. Plant Sci. ***2****:78. doi: 10.3389/fpls.2011.00078*

This article was submitted to Frontiers in Plant Physiology, a specialty of Frontiers in Plant Science.

frontiers in
PLANT SCIENCE

REVIEW ARTICLE
published: 15 November 2011
doi: 10.3389/fpls.2011.00077

Next generation quantitative genetics in plants

José M. Jiménez-Gómez*

Department of Plant Breeding and Genetics, Max Planck Institute for Plant Breeding Research, Köln, Germany

Edited by:
Alisdair Fernie, Max Planck Institute for Plant Physiology, Germany

Reviewed by:
Alisdair Fernie, Max Planck Institute for Plant Physiology, Germany
Mathilde Causse, National Institute of Agricultural Research, France

***Correspondence:**
José M. Jiménez-Gómez, Department of Plant Breeding and Genetics, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Köln, Germany.
e-mail: jmjimenez@mpipz.mpg.de

Most characteristics in living organisms show continuous variation, which suggests that they are controlled by multiple genes. Quantitative trait loci (QTL) analysis can identify the genes underlying continuous traits by establishing associations between genetic markers and observed phenotypic variation in a segregating population. The new high-throughput sequencing (HTS) technologies greatly facilitate QTL analysis by providing genetic markers at genome-wide resolution in any species without previous knowledge of its genome. In addition HTS serves to quantify molecular phenotypes, which aids to identify the loci responsible for QTLs and to understand the mechanisms underlying diversity. The constant improvements in price, experimental protocols, computational pipelines, and statistical frameworks are making feasible the use of HTS for any research group interested in quantitative genetics. In this review I discuss the application of HTS for molecular marker discovery, population genotyping, and expression profiling in QTL analysis.

Keywords: QTL analysis, plant genetics, next generation sequencing, genomics, eQTL analysis, RNA-seq

INTRODUCTION

For almost one century scientists have dissected the genetic architecture of quantitative traits in plants using Quantitative trait loci (QTL) analysis (Fisher, 1918). These analyses establish associations between genetic markers and the phenotypic variation of a quantitative trait in a segregating population. The techniques used to obtain markers and physiological phenotypes have been constantly improved through history (Schlotterer, 2004; Montes et al., 2007). Recently, the price drop of high-throughput technologies have allowed plant researchers to quantify the general abundance of transcripts, proteins, or metabolites in segregating populations (Kirst et al., 2005; Vuylsteke et al., 2005, 2006; Decook et al., 2006; Keurentjes et al., 2007; West et al., 2007; Lisec et al., 2008; Potokina et al., 2008; Drost et al., 2010). These studies show that there are multiple benefits in using "omic" technologies for QTL analyses, even when the goal is to characterize physiological phenotypic diversity. First, molecular phenotypes are the initial step toward the production of physiological phenotypes and its regulation underlies much of phenotypic diversity (Hoekstra and Coyne, 2007; Stern and Orgogozo, 2008). Second, the availability of genome-wide information significantly increases the ability to identify candidate genes for QTLs (Jimenez-Gomez et al., 2010). Third, molecular traits measured at system scale allow estimation of the effect of QTLs in the genetic pathways of interest, or identification of additional gene networks altered by the loci responsible for the variation (Kliebenstein et al., 2006). Finally, molecular traits offer researchers a better understanding of how mutation drives physiological variation and what are the evolutionary forces acting at primary levels.

High-throughput sequencing, or HTS, allows the rapid and cost–effective generation of massive amounts of short sequences or reads (Metzker, 2010). The potential of this technology for mapping loci responsible for phenotypic differences in plants has already been demonstrated by identifying genes containing EMS-induced mutations in samples of pooled F2 individuals (Schneeberger et al., 2009; Austin et al., 2011). HTS technologies have been in the market for a few years, and new methods are being developed that will be cheaper, require less sample processing, and will produce more and longer reads (Munroe and Harris, 2010; Glenn, 2011; Niedringhaus et al., 2011). It is therefore clear that very soon HTS will be the tool of choice for QTL analyses. One important limiting factor remains to be eliminated: Data analysis. It requires long and computationally intensive pipelines that need to be customized for each particular experimental set up. An increasing number of new algorithms are constantly released to the community, and the debate on which pipelines return the most accurate results is still ongoing. Comparing, combining, and customizing these pipelines requires simple Unix or Linux commands and greatly benefits from knowledge in powerful statistical software such as R, and in scripting languages, such as Perl or Python (R Development Core Team, 2009). For non-bioinformaticians, integrated solutions with convenient interfaces are becoming popular both from collaborative open projects and companies (Blankenberg et al., 2010; Goecks et al., 2010). A popular website that keeps an actualized list of the available software tools is www.seqanswers.com, where users and developers also discuss new technological advances and pipelines. In terms of the computational equipment required for HTS data analysis, the majority of tools are developed for Linux or Unix based systems. Although parts of the analysis can be performed in any modern computer, machines with dozens of gigabytes of RAM are recommended in cases where reference sequences form the species considered are available, or with hundreds if no reference exists. An alternative option that is likely to become popular is to rent storage and computing power in specialized centers, or "the cloud" (Stein, 2010).

Due to the fast improvement of HTS, this review intends only to capture a snapshot in time of the possibilities that it offers for

molecular marker discovery, genotyping, and molecular phenotyping in segregating populations of plants. This review has the purpose of helping researchers who have not incorporated this technology to their work to think about the requirements and possibilities of HTS. By no means this review refers to all available experimental designs or analysis tools, and the solutions proposed here are mere suggestions that will certainly soon be substituted by new and better ones. A guide map of the methods proposed in this review is depicted in **Figure 1**.

LIBRARY PREPARATION

Sample preparation protocols are continuously improved to use fewer amounts of biological material, be completed faster, and reduce the bias in their output. As an example, most current protocols allow multiplexing samples by adding a short sequence tag to all reads in a library, a convenient feature given the increasing numbers of reads produced per HTS run. The same companies that developed the HTS sequencers commercialize library preparation protocols optimized for the most common experimental designs. There are also kits from other companies that give comparable results and may be more cost efficient. Finally, many researchers are developing custom protocols to obtain specific information such as the transcribed strand in RNA-seq experiments, the rate of RNA degradation, or the positions occupied by RNA polymerases, just to name a few (Addo-Quaye et al., 2008; Core et al., 2008; German et al., 2008; Parkhomchuk et al., 2009).

QUALITY CONTROL AND PRE-PROCESSING

Assessing the quality of HTS reads includes detection of biases on base composition, base quality, and sample complexity. The quality of the sequences has an impact on the reliability of the biological interpretations resulting from the analysis (Dohm et al., 2008). Part of these biases are introduced by the sample preparation protocols (Schwartz et al., 2011), particularly during cDNA synthesis in RNA-seq experiments (Hansen et al., 2010; Li et al., 2010b) and PCR amplification (Aird et al., 2011). Additional biases are particular to each HTS technology (Smith et al., 2008; Quince et al., 2011) or specific to each run of the sequencers (Auer and Doerge, 2010).

After quality control it is usually necessary to pre-process the reads by trimming low quality nucleotides and adapter sequences. At this stage, foreign sequences such as vectors or DNA from organisms contaminating the samples can also be removed. Depending of the type of libraries sequenced further pre-processing may be needed, such as trimming poly A or poly T tails and terminal transferase tails in RNA-seq libraries. In cases where several libraries have been multiplexed, reads should be separated by their barcode.

Both quality control and pre-processing can be easily performed with basic scripts written in Perl (Bioperl), R (Bioconductor), or Python (Biopython; Stajich et al., 2002; Gentleman et al., 2004; Cock et al., 2009; R Development Core Team, 2009). For non-programmers, there are some convenient tools that can carry out all or some of these tasks (FastQC, 2008; FASTX-Toolkit, 2009; Blankenberg et al., 2010; Falgueras et al., 2010; Goecks et al., 2010; Cutadapt, 2010; Schmieder et al., 2010; Schmieder and Edwards, 2011).

MOLECULAR MARKER DISCOVERY

Depending on the availability of a reference sequence short reads will be aligned or *de novo* assembled using one of the multiple tools available. There are a number of recent articles that compare the most popular algorithms and software available for these purposes (Bao et al., 2011; Lin et al., 2011; Ruffalo et al., 2011). Please note that the methods proposed below are directed to developing molecular markers for QTL analysis and not to

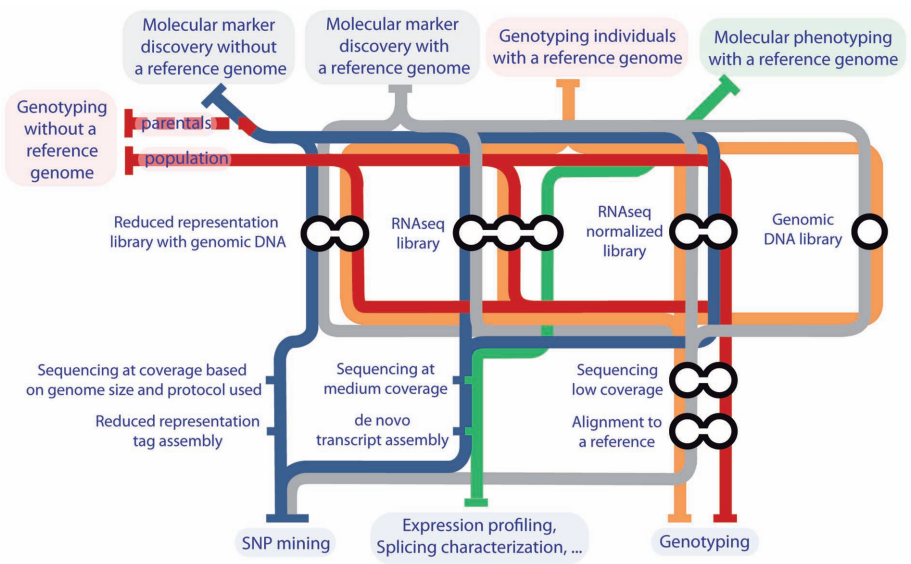


FIGURE 1 | Guide map to the proposed pipelines for SNP identification, genotyping, and molecular phenotyping for QTL analysis in plants. Medium coverage is considered from 20× to 100× the genome or transcriptome size under study. Low coverage is considered under 15× the genome or transcriptome size under study.

identify the mutation underlying the QTL, which requires much deeper sequencing.

WITH A REFERENCE SEQUENCE

A cost efficient solution to obtain molecular markers is to sequence DNA or RNA from the parental genotypes and mine polymorphisms from the resulting reads. These polymorphisms can be used later to design PCR markers or a high-throughput genotyping assay for the full population. This approach works remarkably well in diploid and polyploidy species using as low an amount of sequence as 5× coverage, meaning five times the size of the genome under study (Ossowski et al., 2008; Gore et al., 2009; Trick et al., 2009; Lai et al., 2010; Lam et al., 2010; Arai-Kichise et al., 2011; Geraldes et al., 2011). A recent article reviews the methods and tools available for single nucleotide polymorphism (SNP) identification and genotyping (Nielsen et al., 2011). To align the reads to the reference, mapping softwares based in "seed methods" are preferred despite their slower nature because their robustness to polymorphisms. Before SNP calling users may consider removal of the reads that map to multiple locations in the reference, and of duplicated reads that may have been generated from PCR artifacts. A recent pipeline also recalibrates the quality of the nucleotides in the reads to correct for the high error rates in HTS, and realigns reads in complex genomic positions where the fast processing alignment algorithms may have failed (Depristo et al., 2011). Commonly used indicators of the veracity of polymorphisms are based in the amount and quality of reads showing the polymorphism, frequency of the observed alleles, quality of the alignment, and/or proximity to other polymorphisms. There are some basic and popular options for calling polymorphisms from aligned reads (Li et al., 2009a,b; Depristo et al., 2011), tools specialized in the analysis of reads from particular sequencing platforms (Souaiaia et al., 2011), that have the ability to detect structural variation (Chen et al., 2009; Hormozdiari et al., 2009, 2010), or that have into account the quality of the reference in addition to the quality of the reads (Frohler and Dieterich, 2010). An essential method to control for the quality of the data analysis process is visual inspection through genome viewers specialized in HTS datasets (Huang and Marth, 2008; Bao et al., 2009; Milne et al., 2010; Robinson et al., 2011).

WITHOUT A REFERENCE SEQUENCE

High-throughput sequencing sequences can serve to construct the necessary reference to identify molecular markers if it is not already available. Although assembling *de novo* a complete genome sequence is possible with HTS, it requires very deep sequencing and extensive bioinformatic analysis, even more given the relatively large size of most plant genomes. A more efficient option is sequencing mRNA, which greatly reduces sample complexity in comparison with genome sequencing and has the advantage of offering functional information such as coding polymorphisms or expression levels (Graham et al., 2010; Mizrachi et al., 2010; Bancroft et al., 2011; Everett et al., 2011; Garg et al., 2011; Guo et al., 2011; Ibarra-Laclette et al., 2011; Ness et al., 2011; Su et al., 2011; Wei et al., 2011). A comprehensive compilation of the methods and tools available for transcriptome assembly has been recently published (Martin and Wang, 2011). *De novo* assembly algorithms greatly benefit from long and paired-end reads, but are extremely sensitive to errors and polymorphisms and will not perform well during assembly of datasets from mixed genotypes or highly heterozygous individuals. The amount of new genomic positions detected in RNA-seq experiments decrease exponentially as the number of reads increases (**Figure 2**). The majority of medium and highly expressed transcripts in a sample are detected at low coverage, and increasing coverage will mainly add non-coding RNAs and low expressed transcripts at a very high cost (Tarazona et al., 2011). If the objective is to assemble complete transcriptomes, obtaining samples from diverse tissues, time points, and conditions is preferred to depth of sequencing. Even in the best possible conditions assemblies from RNA-seq reads will return only a subset of the existing transcripts, many of which will be fragmented. This is expected due to low expression of particular transcripts, the non-uniform read coverage, and the presence of different isoforms per gene. To help assembly of low expressed transcripts researchers can use normalization protocols that deplete the most abundant transcripts from the samples (Christodoulou et al., 2011). In any case, contigs resulting from *de novo* assembly can be effectively used as a reference for molecular marker detection and characterization of transcripts in un-sequenced genomes (Parchman et al., 2010; Wang et al., 2010e; Angeloni et al., 2011; Hiremath et al., 2011; Kaur et al., 2011).

When highly similar genotypes are compared, RNA-seq may not be the best option since it mostly targets coding regions, which are less diverse than non-coding regions. In these cases researchers can construct reduced representation libraries by shearing DNA using restriction endonucleases and size-selecting the fragments that will be sequenced. Reads from these libraries can be clustered by similarity and mined for polymorphisms close to the restriction sites; or used to detect the presence–absence of particular tags, indicating a polymorphism in the restriction site itself (Kerstens et al., 2009; Sanchez et al., 2009; Etter et al., 2011). Obtaining polymorphisms from reduced representation libraries is more efficient when a reference sequence is available (Van Tassell et al., 2008; Wu et al., 2010). However, researchers have already developed tools to genotype samples from these tags using a low number

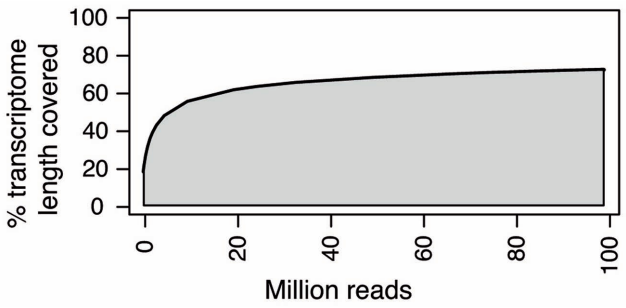


FIGURE 2 | Percentage of transcriptome covered versus number of RNA-seq reads used. Eighty-one base pair paired-end RNA-seq reads from *S. lycopersicum* were randomly sampled in different subset sizes and aligned to the *S. lycopersicum* genome reference. The percentage of the length of the transcriptome covered by at least one read is represented at different coverages.

of reads from organisms without a reference (Ratan et al., 2010), or to reconstruct part of the targeted genome using paired-end sequencing (Willing et al., 2011). Additional protocols to obtain markers from reduced representation libraries exist in which different combination of restriction enzymes are used for each of the genotypes involved (Hyten et al., 2010), or that do not shear the DNA but filter the reads for single copy sequences (You et al., 2011). The amount of reads necessary to perform this type of analysis depends on the size of the genome, the restriction enzymes used, and the availability of a reference.

GENOTYPING POPULATIONS

With the price drop of the HTS technologies and the possibility of multiplexing samples, genotyping an entire population has become realistic (Schneeberger and Weigel, 2011). In the case of a sequenced system such as rice, generating reads from the individuals of a population at 0.02–0.055× coverage allowed high-density genotyping by comparisons with the parental genotypes (Huang et al., 2009), or by inferring the parental genotypes from the polymorphisms found in the population (Xie et al., 2010). Since erroneous polymorphism calls are expected at low coverage, more or less complex algorithms need to be defined to correctly genotype each polymorphism in each individual (Huang et al., 2009; Xie et al., 2010; Li et al., 2011). In addition, a reference sequence can serve researchers to design enrichment essays that will target their preferred genomic locations, although at high cost (Blow, 2009; Mamanova et al., 2010; Nijman et al., 2010; Kenny et al., 2011). For species where a genome sequence is not available, a very practical approach is to sequence reduced representation libraries as mentioned above (Baird et al., 2008; Emerson et al., 2010b; Hohenlohe et al., 2010, 2011).

MOLECULAR PHENOTYPING

The list of molecular phenotypes that can be quantified with HTS is extensive and is rapidly increasing (Hawkins et al., 2010). Examples of these phenotypes are protein–RNA interactions (Licatalosi et al., 2008; Hafner et al., 2010), translation rates (Ingolia et al., 2009; Ingolia, 2010), transcription rates (Core et al., 2008; Churchman and Weissman, 2011), protein–DNA interactions (Albert et al., 2007; Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007; Robertson et al., 2007; Chen et al., 2008; Hesselberth et al., 2009), RNA degradation rates (Addo-Quaye et al., 2008; German et al., 2008), RNA secondary structure (Kertesz et al., 2010; Underwood et al., 2010), transcription start positions (Plessy et al., 2010), chromatin accessibility (Boyle et al., 2008), methylation states (Cokus et al., 2008; Down et al., 2008; Lister et al., 2008; Meissner et al., 2008), natural antisense transcription (Cloonan et al., 2008; Core et al., 2008; He et al., 2008; Armour et al., 2009; Parkhomchuk et al., 2009) or small RNA profiles (Lu et al., 2005). QTL analysis using these phenotypes as traits is an exciting field that remains un-explored. Therefore, the computational frameworks to quantitatively compare these phenotypes between individuals will need to be established.

EXPRESSION PROFILING WITH HTS

Although many cases of phenotypic variation caused by coding polymorphisms have been documented, variation in gene expression has been shown to underlie much of phenotypic diversity (Reviewed in Hoekstra and Coyne, 2007; Wray, 2007; Stern and Orgogozo, 2008). One method to detect differences in expression between individuals using HTS is to sequence 26–27 nucleotide-long tags from expressed transcripts (Matsumura et al., 2010; Hong et al., 2011). A recent study shows that this method reaches saturation in mice with 6–8 million reads per sample (Hong et al., 2011). Its advantages over sequencing full transcripts are the lower cost, higher sensitivity, reduced bias during amplification due to the fixed fragment lengths, and use of simplified statistical models to calculate differential expression. On the other hand, methods based in tags will not detect the majority of coding polymorphisms and isoforms, and require a close enough reference sequence to extract biologically meaningful results.

RNA-seq is rapidly becoming a standard in expression profiling because of its simple protocol of preparation, digital nature, large dynamic range, and high sensitivity in comparison with previous technologies (Marioni et al., 2008; Bradford et al., 2010; Liu et al., 2010). In addition, it can serve to genotype individuals, identify novel transcripts, characterize alternative splicing, and quantify allele specific expression (Reviewed in Wang et al., 2009; Costa et al., 2010; Marguerat and Bahler, 2010). Due to the novelty of the technique there is no consensus on which sample preparation protocols present fewer biases (Raz et al., 2011). However, strand-specific methods could become a standard because of their increased precision due to their ability to distinguish between sense and antisense transcripts (He et al., 2008; Levin et al., 2010). In terms of experimental designs, it is necessary to randomize and replicate biological samples, as with any other type of genome-wide analysis (Auer and Doerge, 2010; Fang and Cui, 2011; Hansen et al., 2011). There is little consensus about the depth of sequence needed for expression profiling with RNA-seq. Recent estimates range between 30 million reads to compare the expression profiles of two samples, to 100 million reads to detect most transcribed genes and quantify isoforms, to 500 million to obtain accurate profiles, including low expressed transcripts (Zhang et al., 2010; ENCODE, 2011; Toung et al., 2011). In any case, it is advisable to balance the number of reads between samples in the same experiment in order to perform accurate expression comparisons (Tarazona et al., 2011).

Expression profiling from HTS datasets is necessarily based on counting the reads mapped to each transcript in a reference sequence. When a reference genome or transcriptome is not available, it can be reconstructed using *de novo* assembly of the reads for at least one of the genotypes as described above. The simpler and less computational intensive protocol for expression profiling is to map the RNA-seq reads to known (or *de novo* assembled) transcripts and a set of possible exon–exon junctions (when available) to detect alternative splicing. However, in organisms with sequenced genomes this protocol will not allow detection of novel exons, transcripts, and isoforms. The preferred pipeline involves aligning the reads to the genomic reference using an alignment tool that splices the reads to detect intron–exon junctions (For example Trapnell et al., 2009; Ameur et al., 2010; Au et al., 2010; Guttman et al., 2010; Wang et al., 2010b; Lou et al., 2011).

A challenge for expression analyses in samples from two unrelated individuals is the need to perform robust quantification of reads generated from two or more alleles. This implies that reads

with the closer genotype to the reference will align better than reads from a more distant genotype, in which more polymorphisms may interfere with their ability to map (Fontanillas et al., 2010). In these cases, aligners based in seed methods will perform better than those based in the Burrows–Wheeler Transform algorithm (For a review see Garber et al., 2011). Although most studies ignore this problem, there are solutions that go from identifying and removing the polymorphisms that cause these biases (Degner et al., 2009), aligning the reads to all references from the genotypes involved (Bullard et al., 2010a) or including the polymorphisms found in the references (Gan et al., 2011). When two references are used, a potential problem may arise from motifs that are more abundant in one reference with respect to the other if only uniquely mapped reads are counted. The use of longer reads and/or paired-end reads greatly decreases the number of ambiguously mapped reads. In addition, there are robust methods to assign these multi-mapped reads to a single location (Faulkner et al., 2008; Mortazavi et al., 2008; Hashimoto et al., 2009; Li et al., 2010a; Wang et al., 2010a; Ji et al., 2011).

There are a number of tools to count the number of reads aligned to each transcriptional unit to calculate expression, most of which require knowledge of Perl, Phyton, Linux/Unix, or R (Carlson et al., 2009; Bio::DB::Sam, 2009; Anders, 2010; Morgan and Pagès, 2010; Quinlan and Hall, 2010). Some alignment tools can directly calculate the number of reads per transcript and/or a measure of expression based in the reads (or fragments) per gene size in kilobases per million reads mapped, called RPKM (or FPKM; Mortazavi et al., 2008; Trapnell et al., 2010). However, these expression units show biases depending on the length, number, abundance of the transcripts present in the samples, or because of technical replication (Oshlack and Wakefield, 2009; Bullard et al., 2010b; Mcintyre et al., 2011). For this reason researchers have developed dedicated R/Bioconductor packages to calculate differential expression between samples based on raw read counts per transcript (Anders and Huber, 2010; Bullard et al., 2010b; Hardcastle and Kelly, 2010; Robinson et al., 2010; Wang et al., 2010c). In addition, there are software packages that take into consideration the biases inherent to RNA-seq when calculating expression or performing downstream analyses such as gene ontology over-representation studies (Young et al., 2010; Zheng et al., 2011).

High-throughput sequencing datasets allow quantification of expression for each isoform separately, resulting in significantly more accurate estimates than calculating expression at the gene level (Wang et al., 2010d). For this, users must first identify splicing events from the reads that align to exon–exon junctions. Quantifying isoform expression is complicated since most reads in an alternatively spliced transcript cannot be assigned to a single isoform. The most promising methods to address this complex problem take advantage from the information offered by paired-end and/or unambiguously mapped reads (Guttman et al., 2010; Katz et al., 2010; Li et al., 2010a; Trapnell et al., 2010; Nicolae et al., 2011). One advantage of going through the intricate process of identification of alternative splicing is that it can also be used as a trait for QTL analysis (Li et al., 2010c; Montgomery et al., 2010; Pickrell et al., 2010; Lalonde et al., 2011).

ALLELE SPECIFIC EXPRESSION IN HYBRIDS

An alternative to sequencing a full segregating population to perform eQTL analyses is to sequence F1 hybrid individuals, where allele specific expression can be calculated for loci with coding polymorphisms (Babak et al., 2008, 2010; Bullard et al., 2010a; Emerson et al., 2010a; Mcmanus et al., 2010; Pickrell et al., 2010). For any gene, both alleles in the hybrid share the same cellular environment and, as a result, changes in expression between alleles must necessarily be due to *cis*-acting regulators (Cowles et al., 2002). Trans-acting eQTLs can be inferred by performing RNA-seq in the parentals and comparing the differences in expression levels between alleles in the hybrid with the differences between the parentals (Wittkopp et al., 2004). Despite the considerable reduction in price and simplicity of experimental design, this method has several drawbacks. Allele specific expression can only be calculated in transcripts with coding polymorphisms that are highly covered, and it is very dependent on read and transcript length (Degner et al., 2009; Fontanillas et al., 2010). New statistical approaches are being developed that will overcome these limitations, starting by being able to estimate false discovery rates and allele specific alternative splicing (Skelly et al., 2011).

In summary, HTS is changing the way we perform QTL analysis by allowing high-throughput genotyping of populations and phenotyping of traits with a precision not achievable before. It is clear that HTS has not reached its peak of development, and that tools and algorithms will have to be modified according to the new technological improvements. Nevertheless, the first experiments using this technology have already identified exciting possibilities for the characterization of natural variation in plants.

REFERENCES

Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., and Axtell, M. J. (2008). Endogenous siRNA and miRNA targets identified by sequencing of the *Arabidopsis* degradome. *Curr. Biol.* 18, 758–762.

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 12, R18.

Albert, I., Mavrich, T. N., Tomsho, L. P., Qi, J., Zanton, S. J., Schuster, S. C., and Pugh, B. F. (2007). Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* 446, 572–576.

Ameur, A., Wetterbom, A., Feuk, L., and Gyllensten, U. (2010). Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol.* 11, R34.

Anders, S. (2010). *HTSeq: Analysing High-Throughput Sequencing Data With Python*. Available at: http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html#author

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11, R106.

Angeloni, F., Wagemaker, C. A., Jetten, M. S., Op Den Camp, H. J., Janssen-Megens, E. M., Francoijs, K. J., Stunnenberg, H. G., and Ouborg, N. J. (2011). De novo transcriptome characterization and development of genomic tools for *Scabiosa columbaria* L. using next-generation sequencing techniques. *Mol. Ecol. Resour.* 11, 662–674.

Arai-Kichise, Y., Shiwa, Y., Nagasaki, H., Ebana, K., Yoshikawa, H., Yano, M., and Wakasa, K. (2011). Discovery of genome-wide DNA polymorphisms in a landrace cultivar of Japonica rice by whole-genome sequencing. *Plant Cell Physiol.* 52, 274–282.

Armour, C. D., Castle, J. C., Chen, R., Babak, T., Loerch, P., Jackson, S., Shah, J. K., Dey, J., Rohl, C. A.,

Johnson, J. M., and Raymond, C. K. (2009). Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat. Methods* 6, 647–649.

Au, K. F., Jiang, H., Lin, L., Xing, Y., and Wong, W. H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.* 38, 4570–4578.

Auer, P. L., and Doerge, R. W. (2010). Statistical design and analysis of RNA sequencing data. *Genetics* 185, 405–416.

Austin, R. S., Vidaurre, D., Stamatiou, G., Breit, R., Provart, N. J., Bonetta, D., Zhang, J., Fung, P., Gong, Y., Wang, P. W., Mccourt, P., and Guttman, D. S. (2011). Next-generation mapping of *Arabidopsis* genes. *Plant J.* 67, 715–725.

Babak, T., Deveale, B., Armour, C., Raymond, C., Cleary, M. A., Van Der Kooy, D., Johnson, J. M., and Lim, L. P. (2008). Global survey of genomic imprinting by transcriptome sequencing. *Curr. Biol.* 18, 1735–1741.

Babak, T., Garrett-Engele, P., Armour, C. D., Raymond, C. K., Keller, M. P., Chen, R., Rohl, C. A., Johnson, J. M., Attie, A. D., Fraser, H. B., and Schadt, E. E. (2010). Genetic validation of whole-transcriptome sequencing for mapping expression affected by cis-regulatory variation. *BMC Genomics* 11, 473. doi:10.1186/1471-2164-11-473

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., and Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3, e3376. doi:10.1371/journal.pone.0003376

Bancroft, I., Morgan, C., Fraser, F., Higgins, J., Wells, R., Clissold, L., Baker, D., Long, Y., Meng, J., Wang, X., Liu, S., and Trick, M. (2011). Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. *Nat. Biotechnol.* 29, 762–766.

Bao, H., Guo, H., Wang, J., Zhou, R., Lu, X., and Shi, S. (2009). MapView: visualization of short reads alignment on a desktop computer. *Bioinformatics* 25, 1554–1555.

Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X., and Song, Y. Q. (2011). Evaluation of next-generation sequencing software in mapping and assembly. *J. Hum. Genet.* 56, 406–414.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837.

Bio::DB::Sam. (2009). Available at: http://search.cpan.org/∼lds/Bio-SamTools/lib/Bio/DB/Bam/Alignment.pm

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., and Nekrutenko, A. (2010). Manipulation of FASTQ data with galaxy. *Bioinformatics* 26, 1783–1785.

Blow, N. (2009). Genomics: catch me if you can. *Nat. Methods* 6, 539–544.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132, 311–322.

Bradford, J. R., Hey, Y., Yates, T., Li, Y., Pepper, S. D., and Miller, C. J. (2010). A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC Genomics* 11, 282. doi:10.1186/1471-2164-11-282

Bullard, J. H., Mostovoy, Y., Dudoit, S., and Brem, R. B. (2010a). Polygenic and directional regulatory evolution across pathways in *Saccharomyces*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5058–5063.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010b). Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics* 11, 94. doi:10.1186/1471-2105-11-94

Carlson, M., Pages, H., Aboyoun, P., Falcon, S., Morgan, M., Sarkar, D., and Lawrence, M. (2009). *Genomic Features: Tools for Making and Manipulating Transcript Centric Annotations.* Available at: http://www.bioconductor.org/packages/2.6/bioc/html/GenomicFeatures.html

Chen, K., Wallis, J. W., Mclellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., Mcgrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., Shi, X., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L., and Mardis, E. R. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* 6, 677–681.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y. H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W. K., Clarke, N. D., Wei, C. L., and Ng, H. H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133, 1106–1117.

Christodoulou, D. C., Gorham, J. M., Herman, D. S., and Seidman, J. (2011). Construction of normalized RNA-seq libraries for next-generation sequencing using the crab duplex-specific nuclease. *Curr. Protoc. Mol. Biol.* 94, 4.12.1–4.12.11.

Churchman, L. S., and Weissman, J. S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* 469, 368–373.

Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., Mckernan, K. J., and Grimmond, S. M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5, 613–619.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and De Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* 452, 215–219.

Core, L. J., Waterfall, J. J., and Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 322, 1845–1848.

Costa, V., Angelini, C., De Feis, I., and Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-seq. *J. Biomed. Biotechnol.* 2010, 853916.

Cowles, C. R., Hirschhorn, J. N., Altshuler, D., and Lander, E. S. (2002). Detection of regulatory variation in mouse genes. *Nat. Genet.* 32, 432–437.

Cutadapt. (2010). *A Tool That Removes Adapter Sequences From DNA Sequencing Reads.* Available at: http://code.google.com/p/cutadapt/

Decook, R., Lall, S., Nettleton, D., and Howell, S. H. (2006). Genetic regulation of gene expression during shoot development in *Arabidopsis*. *Genetics* 172, 1155–1164.

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., and Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* 25, 3207–3212.

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., Mckenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498.

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* 36, e105.

Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E., Graf, S., Johnson, N., Herrero, J., Tomazou, E. M., Thorne, N. P., Backdahl, L., Herberth, M., Howe, K. L., Jackson, D. K., Miretti, M. M., Marioni, J. C., Birney, E., Hubbard, T. J., Durbin, R., Tavare, S., and Beck, S. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.* 26, 779–785.

Drost, D. R., Benedict, C. I., Berg, A., Novaes, E., Novaes, C. R., Yu, Q., Dervinis, C., Maia, J. M., Yap, J., Miles, B., and Kirst, M. (2010). Diversification in the genetic architecture of gene expression and transcriptional networks in organ differentiation of *Populus*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 8492–8497.

Emerson, J. J., Hsieh, L. C., Sung, H. M., Wang, T. Y., Huang, C. J., Lu, H. H., Lu, M. Y., Wu, S. H., and Li, W. H. (2010a). Natural selection on cis and trans regulation in yeasts. *Genome Res.* 20, 826–836.

Emerson, K. J., Merz, C. R., Catchen, J. M., Hohenlohe, P. A., Cresko, W. A., Bradshaw, W. E., and Holzapfel, C. M. (2010b). Resolving postglacial phylogeography using high-throughput sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 107, 16196–16200.

ENCODE. (2011). Standards, guidelines and best practices for RNA-seq. V1.0.

Etter, P. D., Preston, J. L., Bassham, S., Cresko, W. A., and Johnson, E. A. (2011). Local de novo assembly of RAD paired-end contigs using short sequencing reads. *PLoS ONE* 6, e18561. doi:10.1371/journal.pone.0018561

Everett, M. V., Grau, E. D., and Seeb, J. E. (2011). Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Mol. Ecol. Resour.* 11(Suppl. 1), 93–108.

Falgueras, J., Lara, A. J., Fernandez-Pozo, N., Canton, F. R., Perez-Trabado, G., and Claros, M. G. (2010). SeqTrim: a high-throughput pipeline for preprocessing any type of sequence read. *BMC Bioinformatics* 11, 38. doi:10.1186/1471-2105-11-38

Fang, Z., and Cui, X. (2011). Design and validation issues in RNA-seq experiments. *Brief. Bioinform.* 12, 280–287.

FastQC. (2008). Available at: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

FASTX-Toolkit. (2009). Available at: http://hannonlab.cshl.edu/fastx_toolkit/index.html

Faulkner, G. J., Forrest, A. R., Chalk, A. M., Schroder, K., Hayashizaki, Y., Carninci, P., Hume, D. A., and Grimmond, S. M. (2008). A rescue strategy for multi mapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics* 91, 281–288.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Philos. Trans. R. Soc. Edinb.* 52, 399–433.

Fontanillas, P., Landry, C. R., Wittkopp, P. J., Russ, C., Gruber, J. D., Nusbaum, C., and Hartl, D. L. (2010). Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing. *Mol. Ecol.* 19(Suppl. 1), 212–227.

Frohler, S., and Dieterich, C. (2010). ACCUSA – accurate SNP calling on draft genomes. *Bioinformatics* 26, 1364–1365.

Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E. J., Harberd, N. P., Kemen, E., Toomajian, C., Kover, P. X., Clark, R. M., Ratsch, G., and Mott, R. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 477, 419–423.

Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* 8, 469–477.

Garg, R., Patel, R. K., Jhanwar, S., Priya, P., Bhattacharjee, A., Yadav, G., Bhatia, S., Chattopadhyay, D., Tyagi, A. K., and Jain, M. (2011). Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. *Plant Physiol.* 156, 1661–1678.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5, R80.

Geraldes, A., Pang, J., Thiessen, N., Cezard, T., Moore, R., Zhao, Y., Tam, A., Wang, S., Friedmann, M., Birol, I., Jones, S. J., Cronk, Q. C., and Douglas, C. J. (2011). SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. *Mol. Ecol. Resour.* 11(Suppl. 1), 81–92.

German, M. A., Pillay, M., Jeong, D. H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L. A., Nobuta, K., German, R., De Paoli, E., Lu, C., Schroth, G., Meyers, B. C., and Green, P. J. (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat. Biotechnol.* 26, 941–946.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* 11, 759–769.

Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11, R86.

Gore, M. A., Chia, J. M., Elshire, R. J., Sun, Q., Ersoz, E. S., Hurwitz, B. L., Peiffer, J. A., Mcmullen, M. D., Grills, G. S., Ross-Ibarra, J., Ware, D. H., and Buckler, E. S. (2009). A first-generation haplotype map of maize. *Science* 326, 1115–1117.

Graham, I. A., Besser, K., Blumer, S., Branigan, C. A., Czechowski, T., Elias, L., Guterman, I., Harvey, D., Isaac, P. G., Khan, A. M., Larson, T. R., Li, Y., Pawson, T., Penfield, T., Rae, A. M., Rathbone, D. A., Reid, S., Ross, J., Smallwood, M. F., Segura, V., Townsend, T., Vyas, D., Winzer, T., and Bowles, D. (2010). The genetic map of *Artemisia annua* L. identifies loci affecting yield of the antimalarial drug artemisinin. *Science* 327, 328–331.

Guo, S., Liu, J., Zheng, Y., Huang, M., Zhang, H., Gong, G., He, H., Ren, Y., Zhong, S., Fei, Z., and Xu, Y. (2011). Characterization of transcriptome dynamics during watermelon fruit development: sequencing, assembly, annotation and gene expression profiles. *BMC Genomics* 12, 454. doi:10.1186/1471-2164-12-454

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., and Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* 28, 503–510.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M. Jr., Jungkamp, A. C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141, 129–141.

Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* 38, e131.

Hansen, K. D., Zhijin, W., Irizarry, R. A., and Leek, J. T. (2011). Sequencing technology does not eliminate biological variability. *Nat. Biotechnol.* 29, 575–573.

Hardcastle, T. J., and Kelly, K. A. (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 11, 422. doi:10.1186/1471-2105-11-422

Hashimoto, T., De Hoon, M. J., Grimmond, S. M., Daub, C. O., Hayashizaki, Y., and Faulkner, G. J. (2009). Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite. *Bioinformatics* 25, 2613–2614.

Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nat. Rev. Genet.* 11, 476–486.

He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N., and Kinzler, K. W. (2008). The antisense transcriptomes of human cells. *Science* 322, 1855–1857.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S., and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* 6, 283–289.

Hiremath, P. J., Farmer, A., Cannon, S. B., Woodward, J., Kudapa, H., Tuteja, R., Kumar, A., Bhanuprakash, A., Mulaosmanovic, B., Gujaria, N., Krishnamurthy, L., Gaur, P. M., Kavikishor, P. B., Shah, T., Srinivasan, R., Lohse, M., Xiao, Y., Town, C. D., Cook, D. R., May, G. D., and Varshney, R. K. (2011). Large-scale transcriptome analysis in chickpea (*Cicer arietinum* L.), an orphan legume crop of the semi-arid tropics of Asia and Africa. *Plant Biotechnol. J.* 9, 922–931.

Hoekstra, H. E., and Coyne, J. A. (2007). The locus of evolution: evo devo and the genetics of adaptation. *Evolution* 61, 995–1016.

Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W., and Luikart, G. (2011). Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and west slope cutthroat trout. *Mol. Ecol. Resour.* 11(Suppl. 1), 117–122.

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., and Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.* 6, e1000862. doi:10.1371/journal.pgen.1000862

Hong, L. Z., Li, J., Schmidt-Kuntzel, A., Warren, W. C., and Barsh, G. S. (2011). Digital gene expression for non-model organisms. *Genome Res.* doi: 10.1101/gr.122135.111. [Epub ahead of print].

Hormozdiari, F., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* 19, 1270–1278.

Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2010). Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 26, i350–i357.

Huang, W., and Marth, G. (2008). EagleView: a genome assembly viewer for next-generation sequencing technologies. *Genome Res.* 18, 1538–1543.

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T., Dong, G., Sang, T., and Han, B. (2009). High-throughput genotyping by whole-genome resequencing. *Genome Res.* 19, 1068–1076.

Hyten, D. L., Song, Q., Fickus, E. W., Quigley, C. V., Lim, J. S., Choi, I. Y., Hwang, E. Y., Pastor-Corrales, M., and Cregan, P. B. (2010). High-throughput SNP discovery and assay development in common bean. *BMC Genomics* 11, 475. doi: 10.1186/1471-2164-11-475

Ibarra-Laclette, E., Albert, V. A., Perez-Torres, C. A., Zamudio-Hernandez, F., Ortega-Estrada Mde, J., Herrera-Estrella, A., and Herrera-Estrella, L. (2011). Transcriptomics and molecular evolutionary rate analysis of the bladderwort (*Utricularia*), a carnivorous plant with a minimal genome. *BMC Plant Biol.* 11, 101. doi:10.1186/1471-2229-11-101

Ingolia, N. T. (2010). Genome-wide translational profiling by ribosome footprinting. *Meth. Enzymol.* 470, 119–142.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218–223.

Ji, Y., Xu, Y., Zhang, Q., Tsui, K.-W., Yuan, Y., Norris Jr., C., Liang, S., and Liang, H. (2011). BM-map: Bayesian mapping of multireads for next-generation sequencing data. *Biometrics*. doi: 10.1111/j.1541-0420.2011.01605.x. [Epub ahead of print].

Jimenez-Gomez, J. M., Wallace, A. D., and Maloof, J. N. (2010). Network analysis identifies ELF3 as a QTL for the shade avoidance response in *Arabidopsis*. *PLoS Genet.* 6, e1001100. doi:10.1371/journal.pgen.1001100

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316, 1497–1502.

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* 7, 1009–1015.

Kaur, S., Cogan, N. O., Pembleton, L. W., Shinozuka, M., Savin, K. W., Materne, M., and Forster, J. W. (2011). Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. *BMC Genomics* 12, 265. doi:10.1186/1471-2164-12-265

Kenny, E. M., Cormican, P., Gilks, W. P., Gates, A. S., O'dushlaine, C. T., Pinto, C., Corvin, A. P., Gill, M., and Morris, D. W. (2011). Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection. *DNA Res.* 18, 31–38.

Kerstens, H., Crooijmans, R., Veenendaal, A., Dibbits, B., Chin-a-Woeng, T., Den Dunnen, J., and Groenen, M. (2009). Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. *BMC Genomics* 10, 479. doi:10.1186/1471-2164-10-479

Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. *Nature* 467, 103–107.

Keurentjes, J. J. B., Fu, J., Terpstra, I. R., Garcia, J. M., Van Den Ackerveken, G., Snoek, L. B., Peeters, A. J. M., Vreugdenhil, D., Koornneef, M., and Jansen, R. C. (2007). Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1708–1713.

Kirst, M., Basten, C. J., Myburg, A. A., Zeng, Z. B., and Sederoff, R. R. (2005). Genetic architecture of transcript-level variation in differentiating xylem of a eucalyptus hybrid. *Genetics* 169, 2295–2303.

Kliebenstein, D. J., West, M. A., Van Leeuwen, H., Loudet, O., Doerge, R. W., and St Clair, D. A. (2006). Identification of QTLs controlling gene expression networks defined a priori. *BMC Bioinformatics* 7, 308. doi:10.1186/1471-2105-7-308

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., Song, W., Ying, K., Zhang, M., Jiao, Y., Ni, P., Zhang, J., Li, D., Guo, X., Ye, K., Jian, M., Wang, B., Zheng, H., Liang, H., Zhang, X., Wang, S., Chen, S., Li, J., Fu, Y., Springer, N. M., Yang, H., Wang, J., Dai, J., and Schnable, P. S. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.* 42, 1027–1030.

Lalonde, E., Ha, K. C., Wang, Z., Bemmo, A., Kleinman, C. L., Kwan, T., Pastinen, T., and Majewski, J. (2011). RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res.* 21, 545–554.

Lam, H. M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F. L., Li, M. W., He, W., Qin, N., Wang, B., Li, J., Jian, M., Wang, J., Shao, G., Sun, S. S., and Zhang, G. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* 42, 1053–1059.

Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* 7, 709–715.

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. (2010a). RNA-seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26, 493–500.

Li, J., Jiang, H., and Wong, W. H. (2010b). Modeling non-uniformity in short-read rates in RNA-seq data. *Genome Biol.* 11, R50.

Li, Y., Breitling, R., Snoek, L. B., Van Der Velde, K. J., Swertz, M. A., Riksen, J., Jansen, R. C., and Kammenga, J. E. (2010c). Global genetic robustness of the alternative splicing machinery in *Caenorhabditis elegans*. *Genetics* 186, 405–410.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009a). The sequence alignment/map format and SAM tools. *Bioinformatics* 25, 2078–2079.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., and Kristiansen, K. (2009b). SNP detection for massively parallel whole-genome resequencing. *Genome Res.* 19, 1124–1132.

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., and Abecasis, G. R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* 21, 940–951.

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C., and Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456, 464–469.

Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C. J., and Deng, H. W. (2011). Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics* 27, 2031–2037.

Lisec, J., Meyer, R. C., Steinfath, M., Redestig, H., Becher, M., Witucka-Wall, H., Fiehn, O., Torjek, O., Selbig, J., Altmann, T., and Willmitzer, L. (2008). Identification of metabolic and biomass QTL in *Arabidopsis thaliana* in a parallel analysis of RIL and IL populations. *Plant J.* 53, 960–972.

Lister, R., O'malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133, 523–536.

Liu, S., Lin, L., Jiang, P., Wang, D., and Xing, Y. (2010). A comparison of RNA-seq and high-density exon array for detecting differential gene expression between closely related species. *Nucleic Acids Res.* 39, 578–588.

Lou, S. K., Ni, B., Lo, L. Y., Tsui, S. K., Chan, T. F., and Leung, K. S. (2011). ABMapper: a suffix array-based tool for multi-location searching and splice-junction mapping. *Bioinformatics* 27, 421–422.

Lu, C., Tej, S. S., Luo, S., Haudenschild, C. D., Meyers, B. C., and Green, P. J. (2005). Elucidation of the small RNA component of the transcriptome. *Science* 309, 1567–1569.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., and Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nat. Methods* 7, 111–118.

Marguerat, S., and Bahler, J. (2010). RNA-seq: from technology to biology. *Cell. Mol. Life Sci.* 67, 569–579.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18, 1509–1517.

Martin, J. A., and Wang, Z. (2011). Next-generation transcriptome assembly. *Nat. Rev. Genet.* 12, 671–682.

Matsumura, H., Yoshida, K., Luo, S., Kimura, E., Fujibe, T., Albertyn, Z., Barrero, R. A., Kruger, D. H., Kahl, G., Schroth, G. P., and Terauchi, R. (2010). High-throughput SuperSAGE for digital gene expression analysis of multiple samples using next generation sequencing. *PLoS ONE* 5, e12010. doi:10.1371/journal.pone.0012010

Mcintyre, L. M., Lopiano, K. K., Morse, A. M., Amin, V., Oberg, A. L., Young, L. J., and Nuzhdin, S. V. (2011). RNA-seq: technical variability and sampling. *BMC Genomics* 12, 293. doi:10.1186/1471-2164-12-293

Mcmanus, C. J., Coolon, J. D., Duff, M. O., Eipper-Mains, J., Graveley, B. R., and Wittkopp, P. J. (2010). Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Res.* 20, 816–825.

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., Gnirke, A., Jaenisch, R., and Lander, E. S. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454, 766–770.

Metzker, M. L. (2010). Sequencing technologies – the next generation. *Nat. Rev. Genet.* 11, 31–46.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R. P., Lee, W., Mendenhall, E., O'donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., and Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553–560.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., and Marshall, D. (2010). Tablet – next generation sequence assembly visualization. *Bioinformatics* 26, 401–402.

Mizrachi, E., Hefer, C. A., Ranik, M., Joubert, F., and Myburg, A. A. (2010). De novo assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-seq. *BMC Genomics* 11, 681. doi:10.1186/1471-2164-11-681

Montes, J. M., Melchinger, A. E., and Reif, J. C. (2007). Novel throughput phenotyping platforms in plant genetic studies. *Trends Plant Sci.* 12, 433–436.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, 773–777.

Morgan, M., and Pagès, H. (2010). *Rsamtools: Import Aligned BAM File Format Sequences Into R/Bioconductor*. Available at: http://bioconductor.org/packages/release/bioc/html/Rsamtools.html

Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* 5, 621–628.

Munroe, D. J., and Harris, T. J. (2010). Third-generation sequencing fireworks at Marco Island. *Nat. Biotechnol.* 28, 426–428.

Ness, R. W., Siol, M., and Barrett, S. C. (2011). De novo sequence assembly and characterization of the floral transcriptome in cross- and self-fertilizing plants. *BMC Genomics* 12, 298. doi:10.1186/1471-2164-12-298

Nicolae, M., Mangul, S., Mandoiu, I. I., and Zelikovsky, A. (2011). Estimation of alternative splicing isoform frequencies from RNA-seq data. *Algorithms Mol. Biol.* 6, 9.

Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P., and Barron, A. E. (2011). Landscape of next-generation sequencing technologies. *Anal. Chem.* 83, 4327–4341.

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451.

Nijman, I. J., Mokry, M., Van Boxtel, R., Toonen, P., De Bruijn, E., and Cuppen, E. (2010). Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. *Nat. Methods* 7, 913–915.

Oshlack, A., and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* 4, 14.

Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* 18, 2024–2033.

Parchman, T. L., Geist, K. S., Grahnen, J. A., Benkman, C. W., and Buerkle, C. A. (2010). Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* 11, 180. doi:10.1186/1471-2164-11-180

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* 37, e123.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J. B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–772.

Plessy, C., Bertin, N., Takahashi, H., Simone, R., Salimullah, M., Lassmann, T., Vitezic, M., Severin, J., Olivarius, S., Lazarevic, D., Hornig, N., Orlando, V., Bell, I., Gao, H., Dumais, J., Kapranov, P., Wang, H., Davis, C. A., Gingeras, T. R., Kawai, J., Daub, C. O., Hayashizaki, Y., Gustincich, S., and Carninci, P. (2010). Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nat. Methods* 7, 528–534.

Potokina, E., Druka, A., Luo, Z., Wise, R., Waugh, R., and Kearsey, M. (2008). Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. *Plant J.* 53, 90–101.

Quince, C., Lanzen, A., Davenport, R. J., and Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12, 38. doi:10.1186/1471-2105-12-38

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842.

R Development Core Team. (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Ratan, A., Zhang, Y., Hayes, V., Schuster, S., and Miller, W. (2010). Calling SNPs without a reference sequence. *BMC Bioinformatics* 11, 130. doi:10.1186/1471-2105-11-130

Raz, T., Kapranov, P., Lipson, D., Letovsky, S., Milos, P. M., and Thompson, J. F. (2011). Protocol dependence of sequencing-based gene expression measurements. *PLoS ONE* 6, e19287. doi:10.1371/journal.pone.0019287

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* 4, 651–657.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26.

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.

Ruffalo, M., Laframboise, T., and Koyuturk, M. (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* 27, 2790–2796.

Sanchez, C. C., Smith, T. P., Wiedmann, R. T., Vallejo, R. L., Salem, M., Yao, J., and Rexroad, C. E. III. (2009). Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics* 10, 559. doi:10.1186/1471-2164-10-559

Schlotterer, C. (2004). The evolution of molecular markers – just a matter of fashion? *Nat. Rev. Genet.* 5, 63–69.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864.

Schmieder, R., Lim, Y. W., Rohwer, F., and Edwards, R. (2010). TagCleaner: identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* 11, 341. doi:10.1186/1471-2105-11-341

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A. H., Nielsen, K. L., Jorgensen, J. E., Weigel, D., and Andersen, S. U. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods* 6, 550–551.

Schneeberger, K., and Weigel, D. (2011). Fast-forward genetics enabled by new sequencing technologies. *Trends Plant Sci.* 16, 282–288.

Schwartz, S., Oren, R., and Ast, G. (2011). Detection and removal of biases in the analysis of next-generation sequencing reads. *PLoS ONE* 6, e16685. doi:10.1371/journal.pone.0016685

Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J., and Akey, J. M. (2011). A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.* doi: 10.1101/gr.119784.110. [Epub ahead of print].

Smith, D. R., Quinlan, A. R., Peckham, H. E., Makowsky, K., Tao, W., Woolf, B., Shen, L., Donahue, W. F., Tusneem, N., Stromberg, M. P., Stewart, D. A., Zhang, L., Ranade, S. S., Warner, J. B., Lee, C. C., Coleman, B. E., Zhang, Z., Mclaughlin, S. F., Malek, J. A., Sorenson, J. M., Blanchard, A. P., Chapman, J., Hillman, D., Chen, F., Rokhsar, D. S., Mckernan, K. J., Jeffries, T. W., Marth, G. T., and Richardson, P. M. (2008). Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.* 18, 1638–1642.

Souaiaia, T., Frazier, Z., and Chen, T. (2011). ComB: SNP calling and mapping analysis for color and nucleotide space platforms. *J. Comput. Biol.* 18, 795–807.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12, 1611–1618.

Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome Biol.* 11, 207.

Stern, D. L., and Orgogozo, V. (2008). The loci of evolution: how predictable is genetic evolution? *Evolution* 62, 2155–2177.

Su, C. L., Chao, Y. T., Alex Chang, Y. C., Chen, W. C., Chen, C. Y., Lee, A. Y., Hwa, K. T., and Shih, M. C. (2011). De novo assembly of expressed transcripts and global analysis of the *Phalaenopsis aphrodite* transcriptome. *Plant Cell Physiol.* 52, 1501–1514.

Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Res.* doi: 10.1101/gr.124321.111. [Epub ahead of print].

Toung, J. M., Morley, M., Li, M., and Cheung, V. G. (2011). RNA-sequence analysis of human B-cells. *Genome Res.* 21, 991–998.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* 25, 1105–1111.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515.

Trick, M., Long, Y., Meng, J., and Bancroft, I. (2009). Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnol. J.* 7, 334–346.

Underwood, J. G., Uzilov, A. V., Katzman, S., Onodera, C. S., Mainzer, J. E., Mathews, D. H., Lowe, T. M., Salama, S. R., and Haussler, D. (2010). FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods* 7, 995–1001.

Van Tassell, C. P., Smith, T. P., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., Moore, S. S., Warren, W. C., and Sonstegard, T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* 5, 247–252.

Vuylsteke, M., Daele, H., Vercauteren, A., Zabeau, M., and Kuiper, M. (2006). Genetic dissection of transcriptional regulation by cDNA-AFLP. *Plant J.* 45, 439–446.

Vuylsteke, M., Van Eeuwijk, F., Van Hummelen, P., Kuiper, M., and Zabeau, M. (2005). Genetic analysis of variation in gene expression in *Arabidopsis thaliana*. *Genetics* 171, 1267–1275.

Wang, J., Huda, A., Lunyak, V. V., and Jordan, I. K. (2010a). A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. *Bioinformatics* 26, 2501–2508.

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., Macleod, J. N., Chiang, D. Y., Prins, J. F., and Liu, J. (2010b). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* 38, e178.

Wang, L., Feng, Z., Wang, X., and Zhang, X. (2010c). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26, 136–138.

Wang, X., Wu, Z., and Zhang, X. (2010d). Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq. *J. Bioinform. Comput. Biol.* 8(Suppl. 1), 177–192.

Wang, Z., Fang, B., Chen, J., Zhang, X., Luo, Z., Huang, L., Chen, X., and Li, Y. (2010e). De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (*Ipomoea batatas*). *BMC Genomics* 11, 726. doi:10.1186/1471-2164-11-726

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63.

Wei, W., Qi, X., Wang, L., Zhang, Y., Hua, W., Li, D., Lv, H., and Zhang, X. (2011). Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. *BMC Genomics* 12, 451. doi:10.1186/1471-2164-12-451

West, M. A. L., Kim, K., Kliebenstein, D. J., Van Leeuwen, H., Michelmore, R. W., Doerge, R. W., and St Clair, D. A. (2007). Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* 175, 1441–1450.

Willing, E. M., Hoffmann, M., Klein, J. D., Weigel, D., and Dreyer, C. (2011). Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics* 27, 2187–2193.

Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature* 430, 85–88.

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8, 206–216.

Wu, X., Ren, C., Joshi, T., Vuong, T., Xu, D., and Nguyen, H. T. (2010). SNP discovery by high-throughput sequencing in soybean. *BMC Genomics* 11, 469. doi:10.1186/1471-2164-11-469

Xie, W., Feng, Q., Yu, H., Huang, X., Zhao, Q., Xing, Y., Yu, S., Han, B., and Zhang, Q. (2010). Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10578–10583.

You, F. M., Huo, N., Deal, K. R., Gu, Y. Q., Luo, M. C., Mcguire, P. E., Dvorak, J., and Anderson, O. D. (2011). Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. *BMC Genomics* 12, 59. doi:10.1186/1471-2164-12-59

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11, R14.

Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., Zhuang, R., Lu, Z., He, Z., Fang, X., Chen, L., Tian, W., Tao, Y., Kristiansen, K., Zhang, X., Li, S., Yang, H., and Wang, J. (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res.* 20, 646–654.

Zheng, W., Chung, L. M., and Zhao, H. (2011). Bias detection and correction in RNA-sequencing data. *BMC Bioinformatics* 12, 290. doi:10.1186/1471-2105-12-290

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

Received: 29 April 2011; accepted: 23 October 2011; published online: 15 November 2011.

Citation: Jiménez-Gómez JM (2011) Next generation quantitative genetics in plants. Front. Plant Sci. ***2****:77. doi: 10.3389/fpls.2011.00077*

This article was submitted to Frontiers in Plant Physiology, a specialty of Frontiers in Plant Science.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 01 July 2011
doi: 10.3389/fpls.2011.00023

Large-scale co-expression approach to dissect secondary cell wall formation across plant species

Colin Ruprecht[1], Marek Mutwil[1], Friederike Saxe[2], Michaela Eder[2], Zoran Nikoloski[1,3] and Staffan Persson[1]*

[1] *Independent Research Group, Max-Planck-Institute of Molecular Plant Physiology, Potsdam, Germany*
[2] *Department of Biomaterials, Max-Planck-Institute of Colloids and Interfaces, Potsdam, Germany*
[3] *Institute of Biochemistry and Biology, University of Potsdam, Potsdam, Germany*

Edited by:
Alisdair Fernie, Max Planck Institut for Plant Physiology, Germany

Reviewed by:
Nicholas Provart, University of Toronto, Canada
Jesper Harholt, University of Copenhagen, Denmark

****Correspondence:***
Staffan Persson, Max-Planck-Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476 Potsdam, Germany.
e-mail: persson@mpimp-golm.mpg.de

Plant cell walls are complex composites largely consisting of carbohydrate-based polymers, and are generally divided into primary and secondary walls based on content and characteristics. Cellulose microfibrils constitute a major component of both primary and secondary cell walls and are synthesized at the plasma membrane by cellulose synthase (CESA) complexes. Several studies in *Arabidopsis* have demonstrated the power of co-expression analyses to identify new genes associated with secondary wall cellulose biosynthesis. However, across-species comparative co-expression analyses remain largely unexplored. Here, we compared co-expressed gene vicinity networks of primary and secondary wall *CESA*sin *Arabidopsis*, barley, rice, poplar, soybean, Medicago, and wheat, and identified gene families that are consistently co-regulated with cellulose biosynthesis. In addition to the expected polysaccharide acting enzymes, we also found many gene families associated with cytoskeleton, signaling, transcriptional regulation, oxidation, and protein degradation. Based on these analyses, we selected and biochemically analyzed T-DNA insertion lines corresponding to approximately twenty genes from gene families that re-occur in the co-expressed gene vicinity networks of secondary wall *CESAs* across the seven species. We developed a statistical pipeline using principal component analysis and optimal clustering based on silhouette width to analyze sugar profiles. One of the mutants, corresponding to a pinoresinol reductase gene, displayed disturbed xylem morphology and held lower levels of lignin molecules. We propose that this type of large-scale co-expression approach, coupled with statistical analysis of the cell wall contents, will be useful to facilitate rapid knowledge transfer across plant species.

Keywords: secondary cell wall, comparative co-expression analysis, *Arabidopsis*, cellulose

INTRODUCTION

Plant cell walls constitute a cellular exoskeleton that molds the cell shape and protects the cell against environmental threats (Somerville et al., 2004; Liepman et al., 2010). The cell wall mainly holds carbohydrate-based polymers, such as cellulose, hemicelluloses, and pectins, but also polyphenolic macromolecules, or lignins, and various highly glycosylated proteins. Historically, cell walls have been divided into primary and secondary walls, largely depending on the wall function and on the structural contents (Carpita and McCann, 2000). While the primary wall in most higher plants holds cellulose, hemicelluloses, and pectins, the secondary wall is mainly composed of cellulose, xylans, and lignin.

The carbohydrate-based cell wall components are, with the exception of cellulose, synthesized as oligomeric structures in the Golgi, and are subsequently transported to the cell surface where they are incorporated into the growing cell wall matrix (Geisler et al., 2008). In essence, these oligomers are assembled by different glycosyltransferases, perhaps working as larger protein complexes during synthesis (Lerouxel et al., 2006; Scheller and Ulvskov, 2010). Cellulose, on the other hand, is synthesized at the plasma membrane by large cellulose synthase (CESA) complexes (Somerville, 2006; Mutwil et al., 2008a; Taylor, 2008). These complexes consist of three different, yet structurally related, CESA proteins. Consequently, the CESA-complex that is active during secondary wall formation consists of the three CESA proteins CESA 4, 7, and 8 (Turner and Somerville, 1997; Taylor et al., 2000), and the primary wall complex of CESA 1, 3, and 6-related CESA proteins (Arioli et al., 1998; Desprez et al., 2007; Persson et al., 2007a). At least the primary wall complexes are assumed to be guided by microtubules at the cell cortex (Paredez et al., 2006); however, the mechanism of this process still remains unclear. The absolute need for the three CESA proteins for a functional CESA-complex suggests that the corresponding genes may exhibit similar spatiotemporal expression. Indeed, large-scale co-expression analyses have confirmed such behavior (Brown et al., 2005; Persson et al., 2005). In addition, it was also shown that the three *CESA* genes, either the primary or secondary wall *CESA*s, could readily be used as baits to find other co-expressed genes associated with cell wall production. These studies revealed that several crucial genes for xylan and lignin synthesis were transcriptionally coordinated with the secondary wall *CESAs* (Brown et al., 2005; Persson et al., 2005). More recently, similar approaches have also been utilized for genes involved in the synthesis of the primary wall hemicellulose xyloglucan (Cocuron et al., 2007). This study showed that the *CSLC4* gene in *Arabidopsis*, which is presumed to make the glucan backbone for the xyloglucan, was co-expressed with other genes that

are associated with xyloglucan synthesis. Furthermore, a broader analysis of transcriptional coordination of cell wall-related genes in *Arabidopsis* revealed that members of some gene families tend to be co-expressed, e.g., different *GH19* family members tend to be co-expressed with different *CESA* members (Mutwil et al., 2009).

To our knowledge, the possibilities of comparative co-expression analysis across species remain largely unexplored, with the exception of a recent study that explored similarities in co-expression networks between *Arabidopsis* and rice for xylan synthesis-related genes (Oikawa et al., 2010). By using PlaNet (Mutwil et al., 2011), we performed large-scale condition-independent comparisons (Mutwil et al., 2008b; Usadel et al., 2009) of primary and secondary cell wall-related *CESA* co-expression networks from seven different plant species to discover gene families that are consistently transcriptionally coordinated with cellulose synthesis across species. To identify new genes involved in secondary cell wall formation in *Arabidopsis*, we selected genes from gene families that are conserved in the co-expression networks of the secondary CESAs across the seven species and analyzed their mutant lines. We established a statistical pipeline based on biochemical characteristics of the cell wall and show that at least one of the analyzed mutants is deficient in the secondary wall-related polymer lignin.

MATERIALS AND METHODS

COMPARATIVE CO-EXPRESSION ANALYSIS

The respective primary and secondary *CESA* genes for *Arabidopsis* (253428_at, at4g32410, AtCESA1, and 246425_at, at5g17420, AtCESA7), poplar (PtpAffx.23691.1.S1_at, PtCESA1, and Ptp.3087.1.S1_at, PtCESA7), rice (Os.10183.1.S2_at, Os05g08370, OsCESA1, and Os.10206.1.S1_at, Os09g25490, OsCESA9), barley (Contig3478_at, aaf89964.1, HvCESA1, and Contig15116_at, bab67900.1, HvCESA5/7), medicago (Mtr.14653.1.S1_s_at, Medtr3g136720/Medtr7g099810, and Mtr.10615.1.S1_at, Medtr8g145000), soybean (Gma.10862.2.S1_x_at, Glyma04g07220, and GmaAffx.3712.1.S1_a, Glyma06g30860.1), and wheat (Ta.28561.1.S1_a, UniRef90_A2Y0X2, and Ta.4321.1.A1_at, UniRef90_A2WV32) were analyzed using the Network Comparer tool from PlaNet (http://aranet.mpimp-golm.mpg.de/aranet/NetworkComparer), which is based on the AraGenNet co-expression analysis platform (Mutwil et al., 2010). The tool classifies genes according to their PFAM (Protein family, Finn et al., 2010) annotation and compares gene vicinity networks two steps away ($N = 2$; Mutwil et al., 2010) from the query genes for re-occurring PFAMs.

PLANT MATERIAL AND GROWTH CONDITIONS

Seeds for all plant-lines used in this study were obtained from the Nottingham *Arabidopsis* Stock Centre (NASC, http://arabidopsis.info). Mutants used for the neutral sugar analysis were all in Col-0 background. Homozygous mutants were obtained by genotyping using the T-DNA line specific primers and the respective left border primer of the T-DNA listed in supplementary **Table S3** in Supplementary Material. Seedlings were first grown on MS medium containing 1% sucrose for 2 weeks. Then, plants were transferred to standard soil (Einheitserde GS90; Gebrüder Patzer, Sinntal-Jossa, Germany) and grown in a greenhouse under a 16 h light/8 h dark regime at temperatures 21°C (day) and 17°C (night).

BIOCHEMICAL CELL WALL ANALYSES

For neutral sugar analysis, stems of more than ten different individual 9-week-old plants were pooled per sample and then ground in liquid nitrogen. The three replicates obtained from this plant material were then consecutively washed with 10 ml 70% ethanol, 10 ml methanol:chloroform (1:1, v:v) and 10 ml acetone. The resulting crude cell wall material was air-dried for 2 days. To extract the different cell wall components the material was fractionated. First, pectins were extracted by adding 1.5 ml CDTA (1,2-Diaminocyclohexane tetraacetic acid) and shaking the samples for 12 h at 4°C. After centrifugation for 5 min at 13000 rpm, the supernatant was transferred into a fresh 15 ml Falcon tube. This extraction was repeated twice and the pooled supernatants were dialysed using Spectra/Por dialysis tubes (MWCO: 3.5 kDa, Spectrum Laboratories, Rancho Dominguez, CA, USA) for 3 days at 4°C in double distilled water, which was exchanged every 12 h. With the resulting pellet, this whole procedure was repeated with Na_2CO_3 and then 4 M KOH. The remaining material after these three extractions was the insoluble fraction. All four fractions were dried in an Alpha 2–4 lyophilisator (Christ, Osterode, Germany). For the analysis of the neutral sugar composition, 1 mg cell wall material was transferred to screw-capped eppendorf tubes and 30 μg inositol was added as internal standard. After hydrolysis with 2 M trifluoroacetic acid (TFA), alditol acetates were analyzed as described in Neumetzler (2010), which is a modified version of the original protocol from Albersheim et al. (1967). Detection was performed with an Agilent 6890N GC System coupled with an Agilent 5973N Mass Selective Detector (Waldbronn, Germany). For analysis of cellulose in the crude cell wall material, Seaman hydrolysis (Selvendran et al., 1979) was performed of the pellet after trifluoroacetic acid hydrolysis. After this, the hexose content was determined with the anthrone assay described in Dische (1962).

MICROSCOPIC ANALYSES OF XYLEM VESSELS

To determine the thickness of the cell wall in xylem cells, 0.5 cm long segments from the base of the main stem were fixed in a mixture of 2% paraformaldehyde and 2.5% glutaraldehyde on cacodylate buffer, pH 7.4 for 4 h at room temperature. The samples were then fixed with 2% OsO_4 on the same buffer for 2 h, dehydrated in series of ethanol and propylene oxide and finally embedded in Spurr's low viscosity epoxy resin (Spurr, 1969). The embedded stem segments were cut perpendicular to their longitudinal axes. Afterward, the surfaces of the created cross sections were diamond-polished down to 1 μm. The samples were then coated with a 5 nm gold–palladium layer and observed in a Jeol JSM-7500F field emission scanning electron microscope with an acceleration voltage of 5 kV using a secondary electron in-lens detector. The obtained images were analyzed using ImageJ (Rasband, 1997) by measuring the thickness of the cell wall in the middle of the edge of adjacent cells. For analysis of the disturbed xylem phenotype, 0.5 cm long pieces from the basal part of the main stem were embedded in paraffin as previously described (Weigel and Glazebrook, 2002) using an ASP300S embedding automat (Leica, Wetzlar, Germany). Then, 10 μm thin sections were prepared with a RM2265 rotary microtome (Leica, Wetzlar, Germany). Phoroglucinol-HCl staining was performed directly on the slides. Observations of the xylem cells were made with a BX61 (Olympus, Hamburg, Germany) microscope using a

20× objective. Imaging was carried out with a ColorView III digital camera (Olympus, Hamburg, Germany) controlled with the cell^P software from Olympus. Images were processed for publication using Adobe Photoshop CS2 (Adobe, Dublin, Ireland).

LIGNIN MEASUREMENTS

The amount of lignins in selected mutants was analyzed with the thioglycolic-acid (TGA) assay as previously described (Campbell and Ellis, 1992). However, here, 2 mg of dry crude cell wall material was used and directly incubated with 750 µl water, 250 µl concentrated HCl, and 100 µl TGA.

DATA PREPROCESSING

Seven data sets, each with three replicates, from 18 plant-lines were considered in the analysis. The first five data sets correspond to the mol percentage values for the crude cell wall material and the four different fractions for each of the following sugars: Rhamnose (Rha), Fucose (Fuc), Arabinose (Ara), Xylose (Xyl), Mannose (Man), Galactose (Gal), and Glucose (Glc). The sixth dataset comprises the weight percentage values for each of the following four fractions: CDTA, Na_2CO_3, KOH, and insoluble. Since different amounts of sugars could be hydrolyzed from each fraction, the dry weight percentages of the material extracted by the fractionation were normalized according to the total amount of sugars that could be measured on the GC–MS for that fraction. Finally, the last data set integrates the mol percentage values for the seven sugars of the four fractions (CDTA, Na_2CO_3, KOH, and insoluble) by normalizing each of them to the weight percentage values from the sixth dataset. Note that the values for Fuc and Ara in the insoluble fraction were not considered in the analysis (as they were below detection limit across all considered variables, i.e., plants). Each data set was represented by a matrix with rows corresponding to the mutants and the columns corresponding to the fractions/sugars, such that a row defines a profile of a mutant. Clustering and principal component analysis (PCA) were then performed on the mean percentage values from the three replicates, which were row-wise normalized (centered) to zero mean and unit variance.

CLUSTERING

We applied the k-medoid clustering method, which is a more robust version of k-means (Theodoridis and Koutroumbas, 2006). We used the Euclidean distance as a similarity measure for mutant profiles. To determine the number of clusters k, we employed the silhouette validation method (Rousseeuw, 1987). For each mutant i, the silhouette width, $s(i)$, is defined as follows: Let $a(i)$ denote the average dissimilarity between i and all other mutants placed in the same cluster as mutant i. Let $b(i)$ denote the smallest average dissimilarity of mutant i compared to the mutants in another cluster. Then

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

If $s(i)$ is close to 1, it means that the mutant has been assigned an appropriate cluster. A value of 0 for $s(i)$ implies that the mutant lies between clusters, while a value of −1 signifies misclassification. Using this method, each cluster could be represented by the average silhouette width of the mutants in the cluster. For a clustering with k clusters, one can then calculate the overall average of the silhouette widths of the k clusters. Larger overall average silhouette width indicates better clustering; therefore, the number of clusters with maximum overall average silhouette width was taken as the optimal number of clusters. The R programming package *cluster* was used to determine the optimal number of clusters with which k-medoid clustering was subsequently conducted.

PCA AND BIPLOTS

Principal component analysis is a standard technique for data reduction, from which useful summary biplots can be created. Each biplot allows a two-dimensional representation of the mutants based on their scores from the first two principle components (PCs). We combined the biplot with the clustering results from the k-medoids, by including ellipses around each cluster. The loadings for the variables (fractions/sugars) are represented in blue in the supplementary biplots. PCA was conducted by the R function *princomp*.

RESULTS AND DISCUSSION

CONSERVATION OF CERTAIN CO-EXPRESSED PROTEIN FAMILIES ACROSS SPECIES USING PRIMARY AND SECONDARY WALL CESA GENE VICINITY NETWORKS

Genes that are transcriptionally coordinated tend to be functionally related (Usadel et al., 2009). For example, many genes that are co-expressed with the secondary wall *CESA* genes in *Arabidopsis* are involved in secondary cell wall formation (Brown et al., 2005; Persson et al., 2005; Zhong et al., 2008). While these relationships now are obvious in *Arabidopsis*, no large-scale comparative studies have been performed to analyze such relationships in other species. To carry out such an analysis, we first created a co-expressed gene vicinity network for *AtCESA7* using the AraGenNet platform (Mutwil et al., 2010), which includes most of the essential genes for secondary cell wall biosynthesis (**Figure 1**). This co-expression network contains the other two *CESA*s responsible for secondary wall cellulose *AtCESA4* and *AtCESA8* and many other genes that are involved in xylan production, including *IRX8* (*IRREGULAR XYLEM* 8), *IRX9*, *GUX1* (*GLUCURONIC ACID SUBSTITUTION OF XYLAN 1*) and the recently identified *IRX15* (Peña et al., 2007; Persson et al., 2007b; Brown et al., 2009, 2011; Jensen et al., 2011; Mortimer et al., 2010), and tentatively in lignin synthesis, such as *IRX12* (Brown et al., 2005). In addition, several transcription factors, such as *SND1 (SECONDARY WALL-ASSOCIATED NAC DOMAIN 1)*, and *SND2*, *MYB46*, *85*, and *103*, and *IRX11* (Brown et al., 2005; Zhong et al., 2007, 2008), which regulate different aspects of secondary cell wall formation are also transcriptionally coordinated with the secondary *CESAs*.

The *AtCESA7* co-expression network also displayed the corresponding PFAM (Finn et al., 2010) for each gene (**Figure 1**). Based on these PFAM associations, we have compared primary and secondary cellulose synthesis-related co-expression networks of seven species to investigate if those networks consistently include genes from certain PFAMs across species. We identified primary and secondary cell wall specific *CESAs* for barley, rice, poplar, Medicago, soybean, and wheat to create similar co-expression networks as for *Arabidopsis*. The secondary wall *CESAs* are normally expressed

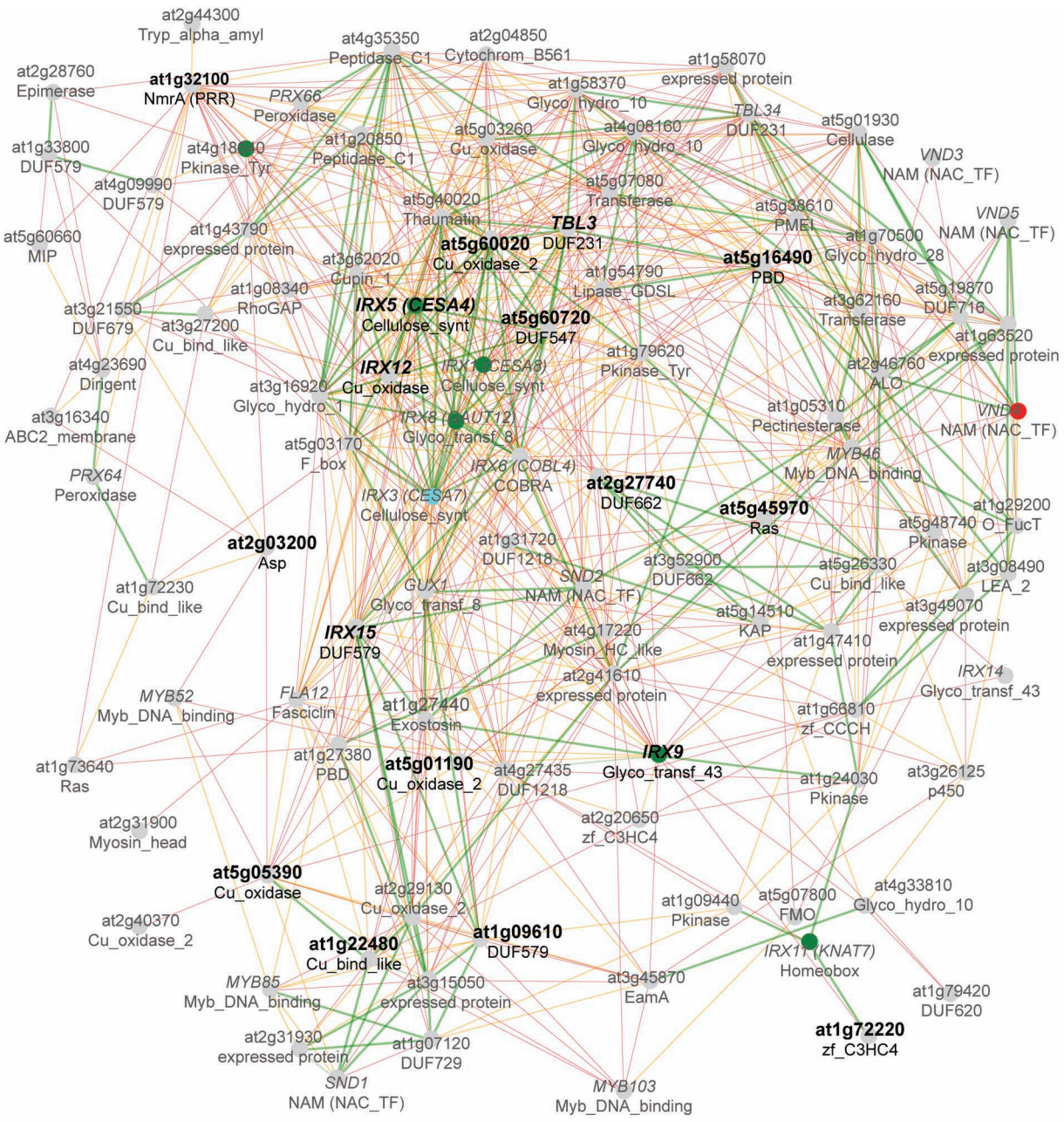


FIGURE 1 | Co-expression gene vicinity network for *AtCESA7* (turqois). Nodes indicate individual genes, and edges indicate whether two genes are co-expressed above a certain mutual rank. Red, green, and gray nodes indicate whether mutations in the gene cause embryophytic lethality (red), any biological phenotype (green), or if no mutant phenotype currently is available (gray) according to TAIR. Green, orange, and red edges indicate a mutual rank relationship 10 (green), 10 but 20 (orange), and 20 but 30 (red), respectively, for each connected gene pair. Most genes with acronyms have been associated with secondary cell wall production. The genes were associated to gene families according to PFAM (protein family) classification, and the respective PFAM is indicated below the gene name/AGI-code. Bold genes have been included in this study. The network was generated, and modified from AraGenNet (http://aranet.mpimp-golm.mpg.de/aranet; Mutwil et al., 2010).

in roots and stems where large amounts of secondary walls are produced, while the primary wall *CESAs* tend to be more uniformly expressed across plant tissues and organs. Based on these assumptions we selected the closest homologs of the *Arabidopsis* primary wall-related *CESA1* and the secondary wall-related *CESA7* gene from the six other species, evaluated the expression profiles (**Figure S1** in Supplementary Material), and validated the genes with previously published results (Tanaka et al., 2003; Burton et al., 2004; Kumar et al., 2009). First, we compared the co-expressed gene vicinity networks of the secondary *CESA*s from the seven

species using the Network Comparer Tool from PlaNet (Mutwil et al., 2011). This analysis revealed that many gene families are conserved across species, because we found genes from the respective PFAM annotation in at least five (green nodes) or four (orange nodes) of the seven networks (**Figure 2**).

For a more comprehensive representation of conserved components in cellulose synthesis across species, we have extended the analysis of the secondary *CESA* genes and included also the co-expression networks of the primary cell wall-related *CESAs* from the seven species. **Figure 3** shows only the highly conserved gene families that appear in at least eight of the 14 analyzed networks (for a complete list see supplementary **Table S1** in Supplementary Material). The gene families were grouped into functional categories based on their PFAM description (PFAM version 24.0, http://pfam.sanger.ac.uk). Core components and gene families that were enriched in primary or secondary cell wall and monocots or dicots were defined based on the difference of occurrence in primary or secondary wall and monocot and dicot specific networks, respectively (**Figure 3**). We hypothesize that the enrichments of certain gene families might reflect the differences of cellulose biosynthesis between primary and secondary cell wall as well as between monocots and dicots. Interestingly, many genes known to be involved in cell wall synthesis in *Arabidopsis* had homologs in other species that are also transcriptionally coordinated with the respective *CESA* genes. For example, most co-expression networks contained Glycosyltransferase family 8 genes, for which *IRX8/*

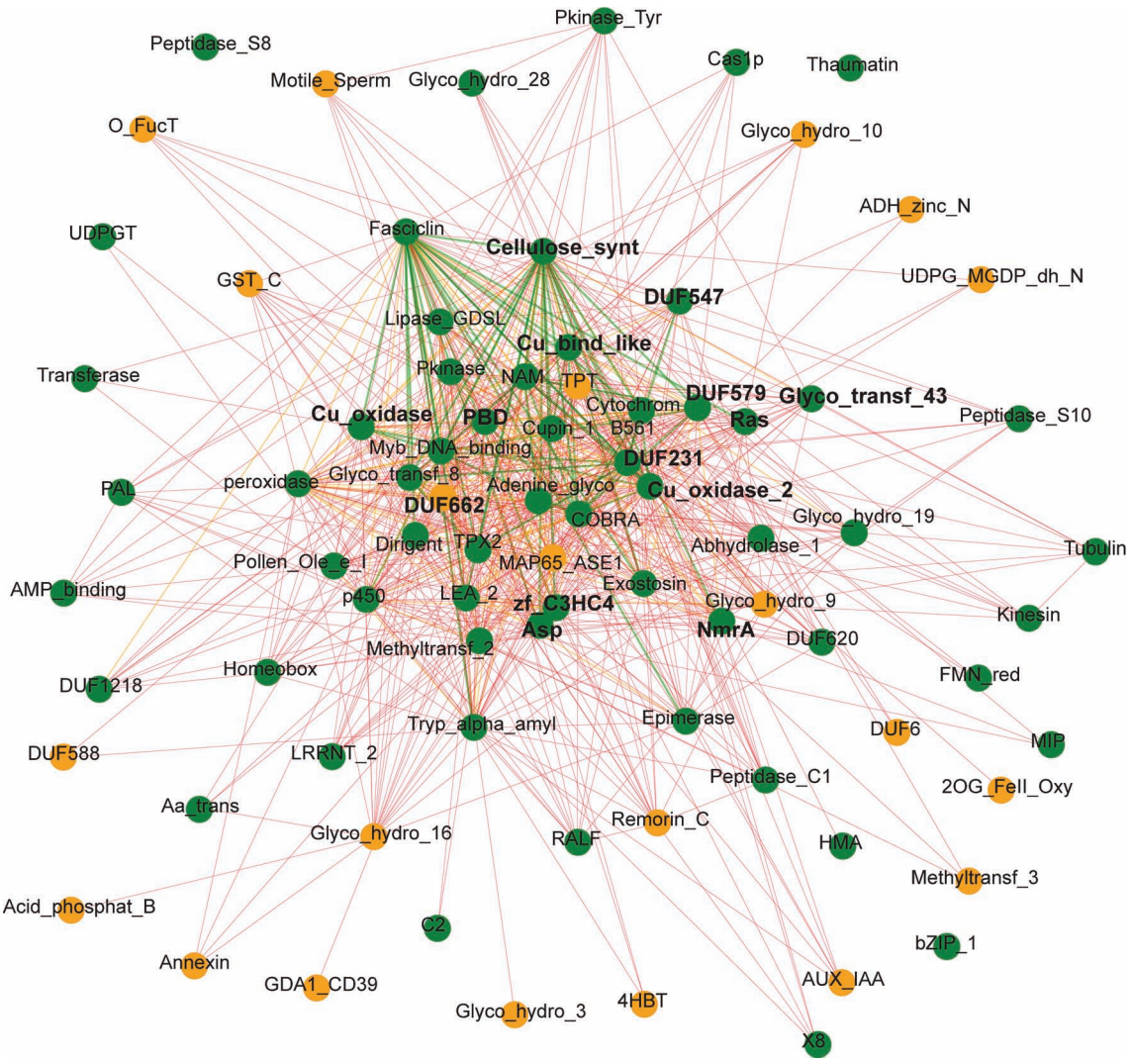


FIGURE 2 | Consensus co-expression network of secondary cell wall genes from seven plant species. Gene vicinity co-expression networks of the *Arabidopsis CESA7* homologs from barley, rice, poplar, medicago, soybean, and wheat were compared for re-occurring gene families (PFAMs). Nodes with their respective PFAM description indicate gene families that are enriched across species. Nodes and edges colored green, orange, and red represent gene families and co-expression relationships between them, that are present in at least five, four or two of the selected networks, respectively. For example, in at least five of the seven co-expression networks from the different species there is at least one gene belonging to the gene family DUF579 (green node in the upper right center). Bold PFAMs indicate gene families from which we selected genes for further analysis. This network was generated using the Network Comparer Tool from PlaNet (http://aranet.mpimp-golm.mpg.de/aranet/NetworkComparer; Mutwil et al., 2011), and red nodes were removed for clarity.

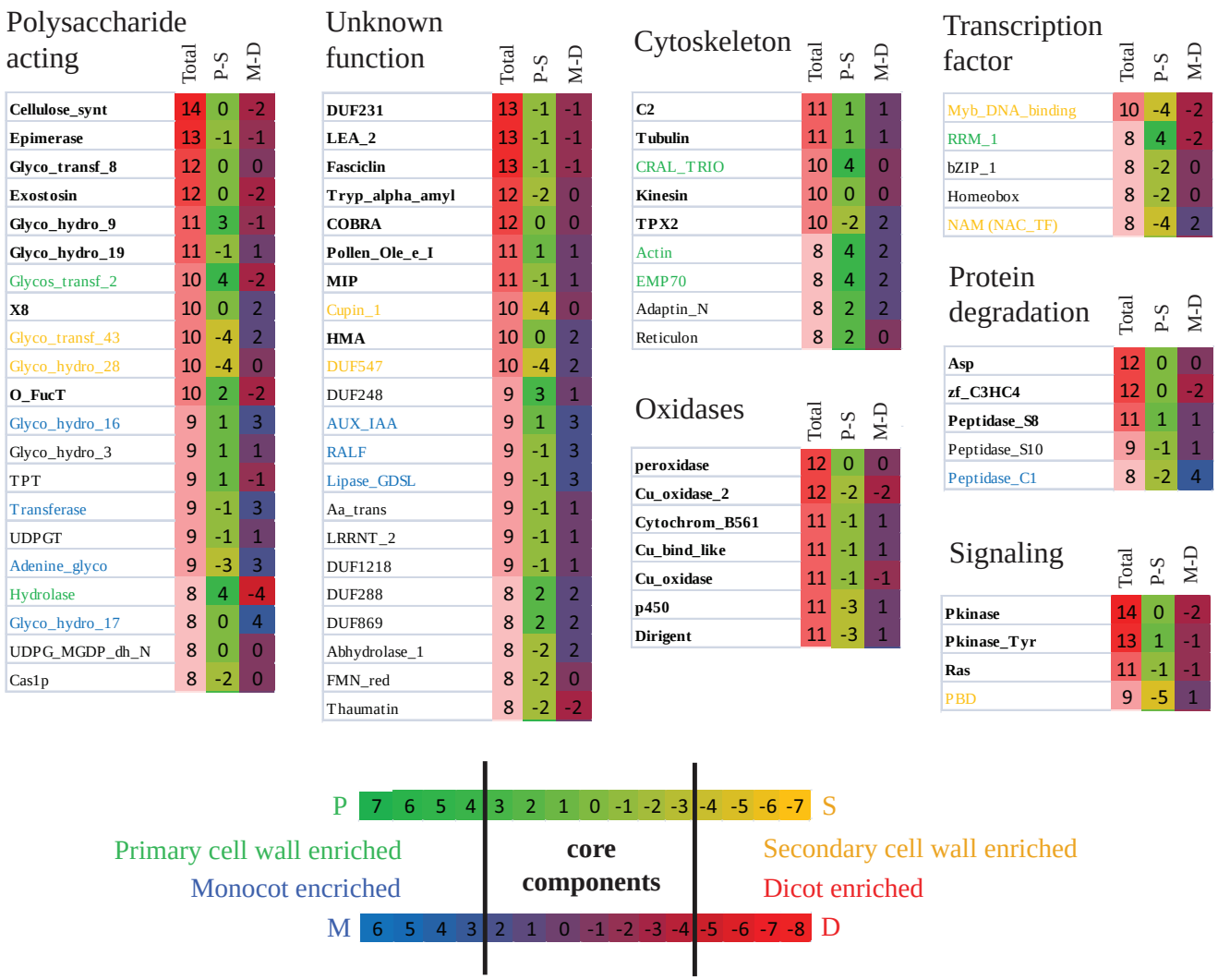


FIGURE 3 | Analysis of primary and secondary cellulose biosynthesis networks in the seven species using Network Comparer. Seven primary cell wall and seven secondary *CESA* gene vicinity co-expression networks have been analyzed for enrichment of PFAMs. Total presence of at least one gene of the respective PFAM in the fourteen co-expression networks is indicated in the first column. The heatmaps in the second and third column represent enrichment of a gene family in primary (P) or secondary cell walls (S) and monocots (M) or dicots (D) based on the difference of occurrence of genes of the respective PFAM in the 14 co-expression networks. For example, the genes from the NAM family (Transcription factor section) are present in two primary (P) and six secondary (S) co-expression networks (**Table S1** in Supplementary Material), which sums up to eight in total (first column), and the difference of P – S is –4 (second column). At the same time, genes from this NAM family are present in five monocot (M) and three dicot (D) co-expression networks resulting in a difference M – D of two (third column). Any PFAM was defined as enriched in primary or secondary cell walls and marked green or orange, if P – S ³ 4 or P – S £ –4, respectively. Enrichment for monocots or dicots, with corresponding PFAMs marked in blue and red, was defined as M – D ³ 3 or M – D £ –5, respectively, because eight dicot and only six monocot networks (corresponding to four dicot and three monocot species) were analyzed. Core components (marked in bold letters) were defined as gene families that were present in at least ten networks, without being enriched in monocots or dicots, nor primary or secondary cell wall biosynthesis.

GAUT12 (*GALACTURONOSYL TRANSFERASE 12)* and *GUX1* (Peña et al., 2007; Mortimer et al., 2010) have been identified in *Arabidopsis*, though their exact function is still unknown. Also, *COBRA*, a GPI-anchored protein important for anisotropic growth (Roudier et al., 2005), *KORRIGAN*, encoding for a endo-1,4-beta-D-glucanase (Glycosyl hydrolase family 9, Nicol et al., 1998), and *CTL1*, a chitinase-like protein (Glycosyl hydrolase family 19, Zhong et al., 2002), had homologs in most of the primary and secondary CESA co-expression networks of the seven species, underlining their importance in cellulose related cell wall synthesis. In addition, DUF231 family members were consistently co-expressed with the *CESA* genes across different species, and at least two DUF231-related gene products have recently been implicated in cellulose production (Bischoff et al., 2010).

Intriguingly, also transcription factors, oxidases, as well as tentative cytoskeletal components, protein degradation, and signaling related genes seem to be consistently co-expressed with the *CESA* genes. In particular, several MYB and NAC transcription factors,

which are secondary cell wall specific according to our analysis, have been shown to be involved in secondary cell wall formation (e.g., MYB46, MYB83, SND1, SND2, Zhong et al., 2008). Oxidases, such as laccases (PFAM annotation: Cu_oxidase) and peroxidases were only slightly enriched in secondary cell wall co-expression networks, suggesting that in addition to their role in lignification of the secondary cell wall they might also have a function in primary cell wall biosynthesis. Furthermore, the recently identified cellulose synthase interacting protein (CSI) 1 (Gu et al., 2010) belongs to the C2 domain-containing family, which was highly enriched in the *CESA* vicinity networks. Although the exact function of CSI1 is still unclear, the conserved co-expression across species implies an important role in cellulose synthesis of this protein. Interestingly, actin and two other cytoskeleton related genes appear more primary cell wall specific, suggesting a more prominent role of these components during primary cell wall biosynthesis. To our surprise also several protein degradation and signaling components appeared in our analysis. For example, the highly conserved gene family Pkinase_Tyr comprises the *THESEUS1* homolog *FERONIA* in *Arabidopsis*. Both receptor-like kinases are involved in control of growth regulation (Hématy and Höfte, 2008, Kessler et al., 2010), suggesting that their homologs might play a similar role in other species. The function of protein degradation in cell wall biosynthesis is unclear. However, given the importance of trafficking and recycling of the CESA-complex (Wightman and Turner, 2010), we hypothesize that these components might be involved in removing inhibited or defect CESA-complex subunits.

We conclude that many genes in the co-expression networks of primary and secondary *CESA* genes are conserved across species indicating that similar genetic modules for cellulose biosynthesis are present in higher plants. This demonstrates that cellulose synthesis-related knowledge obtained in the model species *Arabidopsis* is likely to be transferable to other species. However, using this comparative analysis we can also attempt to infer gene functions in *Arabidopsis*. For example, several glycosyl hydrolase families (GH3, GH16, GH17) are highly conserved in *CESA* co-regulated clusters across species, but do not have any corresponding homologs in the *Arabidopsis* co-expression networks (**Table S1** in Supplementary Material). This might be due to the fact that only about 63% of the *Arabidopsis* genes are represented on the Affymetrix ATH1 chip and the probesets for the respective genes might be missing on the chip (e.g., At3g47040 for GH3, At4g13090 and At5g65730 for GH16, and At1g11820 for GH17 are not represented on the ATH1 chip). We hypothesize that members of these families, whose expression is not determined by microarrays, might actually be co-expressed with the primary or secondary *CESA* in *Arabidopsis* and could constitute functional homologs of the respective genes from the other species.

IDENTIFICATION OF PUTATIVE SECONDARY WALL-RELATED GENES

Several studies have successfully employed the co-expression approach to identify new genes, which are associated with secondary wall cellulose synthesis (for example Brown et al., 2005; Persson et al., 2005). However, there are a large number of genes that are also closely co-expressed with the secondary *CESAs*, which have not been characterized in previous analyses (**Figure 1**). To identify new genes that are important for secondary wall biosynthesis we made use of the comparative analysis of secondary wall co-expression networks and selected genes from gene families that are conserved across species (highlighted PFAMs in **Figure 2**). We obtained homozygous T-DNA mutant lines corresponding to 17 genes that were used for further analyses (**Table 1; Figure 1**). To gain a broader overview on the cellulose synthesis-related gene families, these genes covered our previously defined categories "Unknown function" (i.e., DUF231, 547, 579, 662), "Protein degradation" (i.e. zf_C3HC4, Asp), "Signaling" (i.e., PBD and Ras) and "Oxidases" (i.e., Cu_oxidase, Cu_bind_like, **Figure 3**). Gene homology searches revealed that many of the genes also are part of larger gene families, perhaps suggesting functional redundancies in the absence of one homolog. It is also important to point out that several studies have been undertaken to identify irregular xylem (*irx*) mutants, and we therefore reasoned that while it is unlikely that any of the new mutant lines would exhibit strong defects in xylem morphology it appeared plausible that more subtle changes associated with the secondary cell wall, such as the sugar compositions, may be evident.

CELL WALL ANALYSES

To provide a statistical pipeline to assess similarities and differences of sugar profiles in mutant lines we developed a combined PCA with k-medoid clustering of sugar profiles. To elucidate whether this approach may reveal differences of known and unknown secondary wall mutants we harvested the lower part of stems and analyzed the sugar alditols using GC–MS, and the cellulose content using the anthrone assay. **Figure 4** shows a subset of the sugar alditol estimates from crude cell wall material for wild-type, *irx5*, *irx9*, *irx12*, and three T-DNA lines affecting the genes At5g05390 (mutant number 14), At5g60020 (mutant number 16), and At1g32100 (mutant number 18; **Table 1**). The complete set of sugar alditol estimates is available in **Table S2** in Supplementary Material, and the cellulose measurements in **Figure S2** in Supplementary Material. As previously shown the *irx9* had substantial reduction in xylose (Brown et al., 2005; Bauer et al., 2006), and the *CESA4* deficient *irx5* displayed about 60% reduction in cellulose as estimated from the crude cell wall material. On the other hand, the *irx* mutant *irx12* displayed very minor changes in its sugar composition and cellulose content, similar to what Brown et al. (2005) reported.

To get a more conclusive picture of how the sugar profiles for the different mutants relate to each other we used the corresponding values of the profiles for the pipeline outlined above. The PCA explain the highest variation among the samples (**Figure 4B**), and should detect similar patterns in the sugar profiles for certain mutants. To manage the latter, we assessed whether the sugar profiles for certain mutants clustered together, i.e., we tried to find mutant sugar profiles that were similar to each other but dissimilar to the other mutant sugar profiles. We did this by using a k-means clustering derivative, referred to as k-medoid (Theodoridis and Koutroumbas, 2006). Choosing the right number of clusters is very important for the result of these clustering algorithms. To obtain a statistically reliable number of clusters we analyzed the sugar profiles using the silhouette validation method (Rousseeuw, 1987). This method estimates whether a certain mutant sugar profile should be classified as belonging to a distinct cluster. This is performed by first quantifying the average dissimilarity between one mutant sugar profile to other mutant sugar profiles in the same cluster, and then comparing this difference against the smallest dissimilarity

Table 1 | T-DNA insertion information for selected genes, which are enriched in secondary cell wall co-expression networks across species.

Mutant number	AGI-code	T-DNA line(s)[a]	Annotation	PFAM	Other species[b]
1			**Wild-type Col-0**		
2	**At5g44030**	**SALK_084627 (exon)**	**CESA4 (IRX5)**	Cellulose_synt	All 6 species
3	**At2g37090**	**SALK_057033 (exon)**	**IRX9**	Glyco_transf_43	All 6 species
4	At5g01360	SALK_103316 (exon)	Protein of unknown function	DUF231	All 6 species
5	At5g60720	SALK_055553 (exon)	Protein of unknown function	DUF547	All 6 species
6	At1g09610	SALK_050883 (exon)	Protein of unknown function	DUF579	Os, Hv, Pt, Gm, Ta
7	At3g50220	GABI_735E12 (exon)	IRX15	DUF579	Os, Hv, Pt, Gm, Ta
8	At2g27740	SALK_013255 (exon)	Protein of unknown function	DUF662	Os, Ta, Hv
9	At2g03200	SALK_148906 (5′UTR)	Aspartyl protease family protein; similar to CDR1	Asp	Os, Hv, Pt, Mt, Ta
10	At1g72220	SALK_104510 (5′UTR)	Zinc finger family protein	zf_C3HC4	Os, Hv, Pt, Gm, Mt
11	At5g16490	SALK_015799 (exon)	p21-rho-binding domain-containing protein	PBD	All 6 species
12	At5g45970	GABI_212D04 (exon)	*Arabidopsis thaliana* RAC 2	Ras	Os, Hv, Pt, Gm, Mt
13	**At2g38080**	**SAIL_196_A02 (exon)**	**Laccase (IRX12)**	Cu_oxidase	Os, Hv, Pt, Gm, Ta
14	At5g05390	SALK_004019 (exon)	Laccase 12	Cu_oxidase	Os, Hv, Pt, Gm, Ta
15	At5g01190	SAIL_77_A02 (exon)	Laccase 10	Cu_oxidase_2	All 6 species
16	At5g60020	SALK_016748 (exon)	Laccase 17	Cu_oxidase_2	All 6 species
17	At1g22480	SAIL_381_C11 (intron)	Plastocyanin-like domain-containing protein	Cu_bind_like	Os, Hv, Pt, Gm, Ta
18	At1g32100	SALK_087014 (intron) SALK_058467 (exon) SALK_090999 (intron)	Pinoresinol-lariciresinol reductase (PRR1)	NmrA	Hv, Pt, Gm, Ta

[a]All lines are in Col-0 background. [b]The respective co-expression networks of the Arabidopsis CESA7 homologs of rice (Os), poplar (Pt), barley (Hv), soybean (Gm), medicago (Mt), and wheat (Ta) contained at least one homolog of the respective gene. Bold indicates known genotypes, and corresponding secondary wall phenotypes.

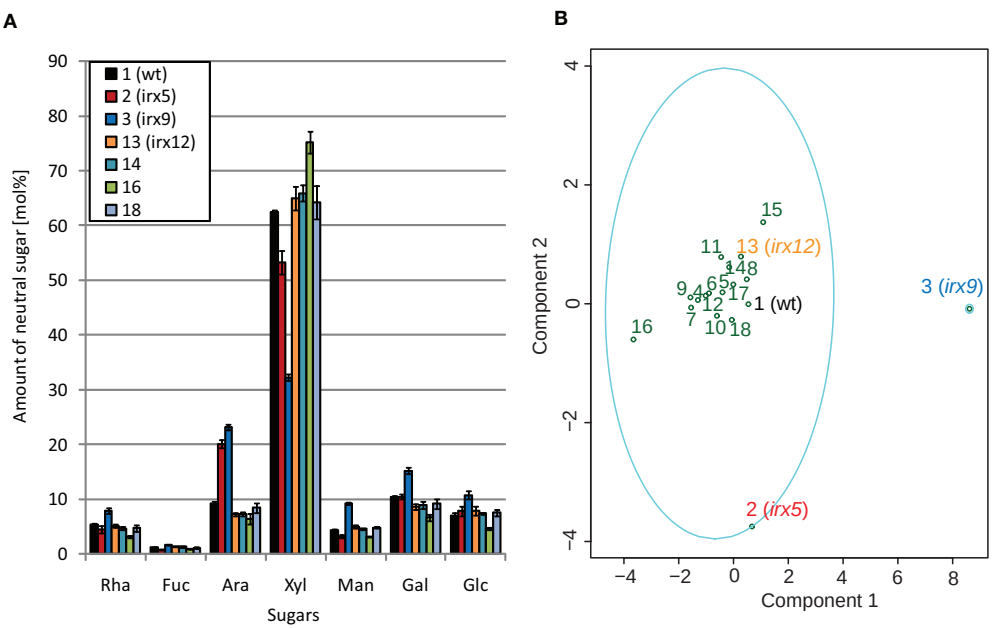


FIGURE 4 | Principal component analysis and clustering of crude cell wall sugar estimates for the selected mutants. (A) GC–MS analysis of sugar alditols of crude cell wall material for indicated mutants. **(B)** Principal component analysis (PCA) of the mol% sugar values from the crude cell wall material. The first two components explain 93.12% of the variation. The loadings for the variables are displayed in supplementary **Figure S3** in Supplementary Material. The ellipses indicate clustering of mutant sugar profiles based on the k-mediod algorithm. The number of clusters was estimated based on the silhouette width, i.e., the dissimilarity in sugar profiles as measured by the Euclidean distances between the profiles. Wild-type (black), and mutants affecting cellulose (*irx5*; red), xylan (*irx9*; blue), and potentially lignin (*irx12*; orange) are indicated in different colors.

between the one mutant sugar profile and mutant sugar profiles that are assigned to other clusters. The scores range between −1 and 1, where a value close to −1 means that the mutant should be assigned to another cluster, and a value close to 1 means that the mutant is correctly classified. The clustering with the highest average silhouette width for the sugar profiles from the crude cell wall material resulted in an optimal number of two clusters. As seen in **Figure 4B**, most of the mutant profiles were classified as belonging to one major cluster. However, at least one of the mutant profiles, corresponding to *irx9*, was retained in its own cluster, and the other severe *irx* mutant, *irx5*, deviated quite dramatically from the other sugar profiles in the larger cluster. These data showed that changes in the cell wall composition can be captured by the two methods, i.e., PCA and the clustering evaluation.

PCA AND CLUSTERING ANALYSES OF SUGAR PROFILES REVEAL SEVERAL SECONDARY WALL MUTANT CLASSES

While the crude cell wall sugar measurements are informative for mutants with dramatic alterations in certain monosaccharides, it is relatively difficult to detect smaller changes associated with distinct polymers. To enrich for such small putative changes, we fractionated the crude cell walls into four fractions using CDTA, which mainly releases Ca^{2+}-chelated polymers such as pectins, Na_2CO_3, which releases pectic polymers that are associated to other polymers by weak hydrogen-bonds, and 4 M KOH, which releases hemicellulosic polymers associated by stronger hydrogen-bonds to the remaining matrix. In addition, we analyzed the remaining pellet, which largely contains cellulosic polymers. Consistent with the supposed fractionation pattern we obtained relatively more pectin related monosaccharides, e.g., rhamnose, galactose, and arabinose, in the first two fractions, and mainly xylose in the third fraction (**Table S2** in Supplementary Material). Similar to the analysis undertaken for the crude cell wall sugar profiles we performed PCA and silhouette width based clustering of the mutant sugar profiles for the different fractions (**Figure 5**). In the CDTA fraction we obtained two clusters, one with most of the mutants and one containing only *irx5*. Interestingly, *irx12* mutant, and also mutant number 14, 16, and 18 were separated on the PCA plot. These three mutants correspond to two laccase genes and to a pinoresinol reductase (*PRR1*) gene, respectively (**Table 1**). It is important to note that both the *IRX12* and the two laccase gene products are proposed to be associated with lignin synthesis, and that the gene product from the *PRR1* has been shown to be involved in the synthesis of lignan (Nakatsubo et al., 2008). The main cause for these mutants to separate from the other mutants in the PCA plot appeared to be a relative changes in xylose and galactose (**Figure S3** in Supplementary Material). A similar, but less clear, pattern was also seen in the Na_2CO_3 fraction where at least two of the laccases and the *prr1* mutants are contained in a separate cluster from the wild-type mainly because of less arabinose in the mutants (**Figure 5; Figure S3** in Supplementary Material). However, these changes were relatively small. It is important to note that the results obtained here are based on one T-DNA per mutant due to the extensive work load involved in generating the profiles from the fractionated material. We can therefore not rule out that the observed changes emanate from additional mutations in the T-DNA line backgrounds.

To merge the different sugar profiles from the four fractions into one estimate, we normalized the mol percentages of the mutants based on their amount of extractable material of the individual fractions and performed silhouette width-driven clustering on all the values for the four fractions. Hence, these estimates reflect the composition of the cell wall in a more detailed way than by only analyzing the neutral sugar composition in the crude cell wall material. The result in **Figure 6** shows that three clusters were apparent, where *irx9* solely occupied one cluster, and the other two clusters held the rest of the mutants.

In summary, we propose that the combined sugar profiling and clustering analyses may be useful to classify mutants, a task that may be relatively difficult using the raw sugar alditol estimates.

A PINORESINOL REDUCTASE IS ASSOCIATED WITH SECONDARY WALL INTEGRITY

To investigate whether mutations in some of the genes resulted in weaker secondary cell walls, we generated hand-cut stem sections and stained these with Toluidine blue. As expected, none of the mutants showed any severe *irx* phenotype. This may be due to extensive genetic redundancy for some of the gene families. For example, At5g01360 (assigned to DUF231 pfam) is part of a gene family of over 40 genes (Bischoff et al., 2010), of which many have over-lapping expression pattern with At5g01360. One of these genes is At2g38320, which is co-expressed with the secondary wall *CESA* genes. However, some of the mutants appeared to have more disturbed xylem vessel shapes as compared to the wild-type. We selected one of these mutants, mutant number 18 or *prr1*, to analyze more in detail, and embedded basal mutant stem parts in paraffin and cut sections (10 μm) using a microtome. We subsequently stained these sections with either Toluidine blue or Phloroglucinol–HCl. The mutant stem displayed what appeared to be weakened secondary cell walls, with disturbed xylem vessel morphology (**Figures 7A,B**). Since deformed xylem vessels were observed in wild-type stems occasionally, the number of deformed xylem vessels was counted in three different mutant lines corresponding to the *PRR1* gene and in the wild-type. **Table 2** clearly shows that all of the three *prr1* mutant lines contained about twice as many xylem vessels with disturbed shapes as wild-type sections. Several of these sections also indicated that the secondary walls were thinner in the mutants compared to wild-type. Since it is difficult to estimate cell wall thickness by using light microscopy we embedded basal stem parts in Spurr's resin, created a plane surface perpendicular to the stem axis for detailed analysis by using a scanning electron microscope (**Figure 7**). The thickness of the secondary cell walls of the xylem related cells in one of the mutants and wild-type was measured. The results indicate that impairment of the PRR1 function results in thinner cell walls, and that this most likely affects the integrity of the wall.

The PRR1 can reduce pinoresinols to lariciresinols (Nakatsubo et al., 2008), and the latter can subsequently be converted into secoisolariciresinols. These structures are part of a larger family of molecules generally referred to as lignans, and may work as antioxidants and phytoestrogens (Pan et al., 2009). In addition, some of these structures have also been found in lignin through two-dimensional-NMR studies (Zhang et al., 2003). To investigate whether the *prr1* mutants, and some of the laccase mutants, caused alterations in lignin related

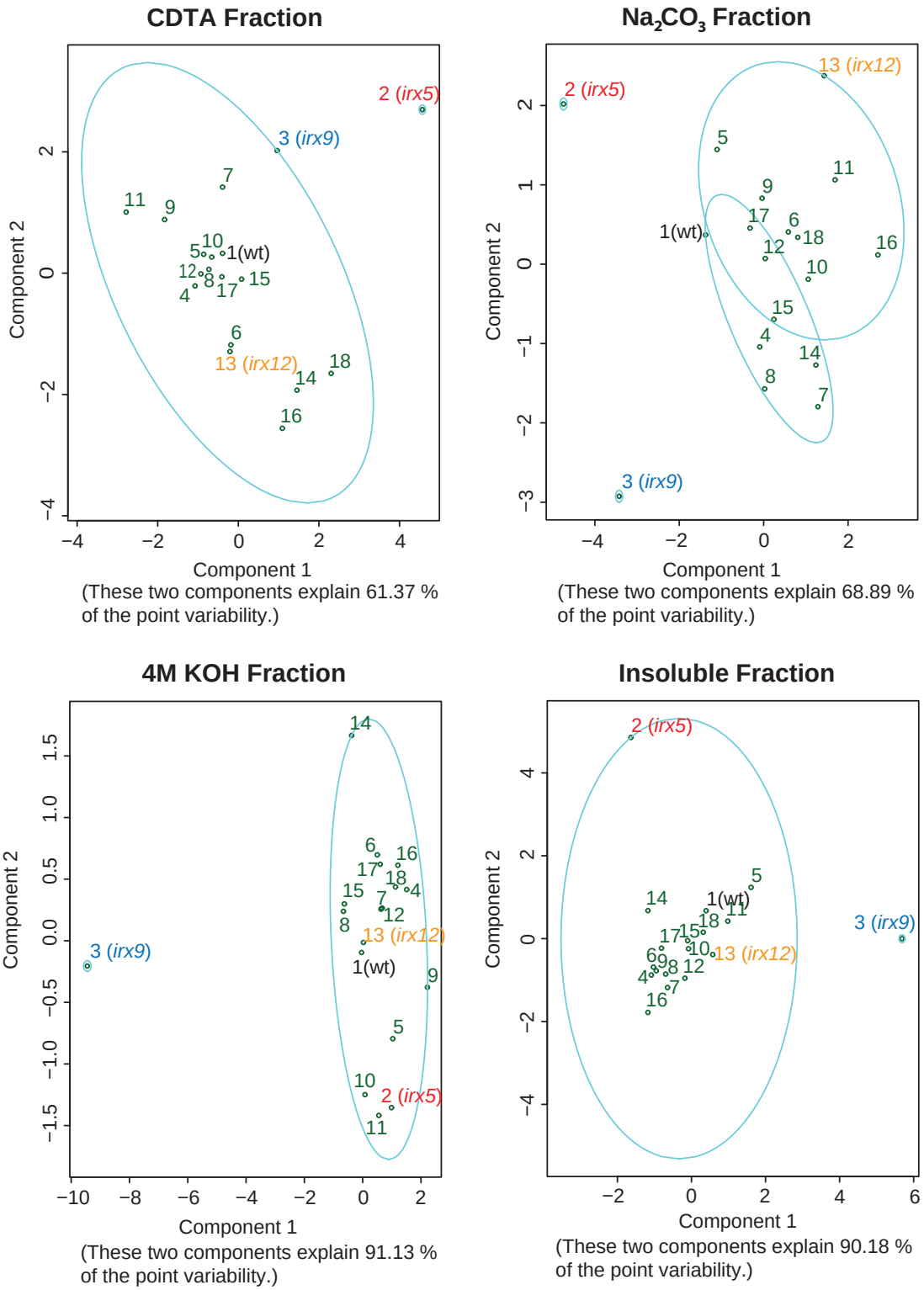


FIGURE 5 | Principal component analysis and cluster analyses of sugar contents from fractionated cell wall material. Principal component analyses of the mol% sugar values in the different fractions for each mutant. The separation based on the scores from the first two principle components is displayed, and the explained variation of these components is indicated below each graph. The loadings for the variables are included in supplementary **Figure S3** in Supplementary Material. The ellipses indicate clustering of mutants based on the k-mediod algorithm. The number of clusters was estimated based on the silhouette width, i.e., the dissimilarity in sugar profiles as measured by the Euclidean distances between the profiles. Wild-type (black), and mutants affecting cellulose (*irx5*; red), xylan (*irx9*; blue), and potentially lignin (*irx12*; orange) are indicated in different colors.

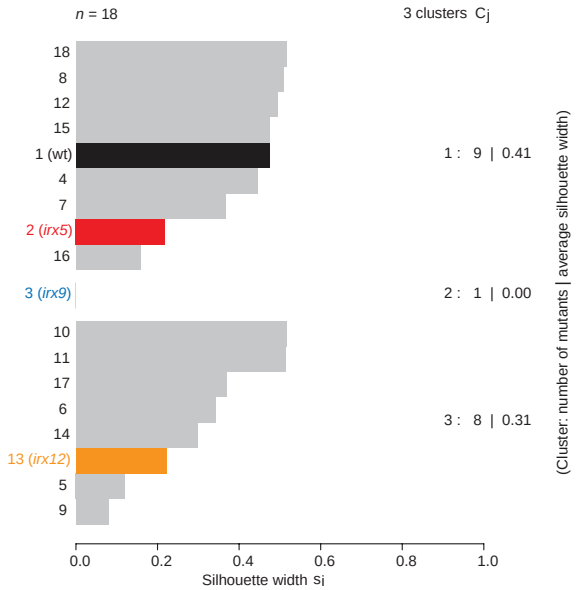


FIGURE 6 | Clustering of combined sugar profiles from the four different cell wall fractions. Clustering summary of mutants based on the k-mediod algorithm. The number of clusters (3) was estimated based on the silhouette width, i.e., the dissimilarity in sugar profiles as measured by the Euclidean distances between the combined profiles. Wild-type (black), and mutants affecting cellulose (*irx5*; red), xylan (*irx9*; blue), and potentially lignin (*irx12*; orange) are indicated in different colors. The average silhoutte width was calculated to 0.34.

structures we measured the lignin content using the Thioglycolic-acid (TGA) assay (Campbell and Ellis, 1992). Using this method we found that the *prr1*, *irx12*, and the laccase mutant number 16 indeed held lower levels of lignin related structures, whereas the laccase mutant number 14 had similar levels as the wild-type control (**Figure 7**). It is important to note that these analyses only estimate the levels of lignin related structures, and do not reflect differences in the structure of the structures. Given that these mutants displayed similar trends for the sugar profiles in the CDTA fraction it is possible that defects in the lignin polymers affect pectin levels, or extractability.

CONCLUSION

The remarkable transcriptional coordination of the genes associated with secondary cell wall formation in *Arabidopsis* suggested that similar relationships would also be present in other plant species. Indeed, by comparing the co-expression networks of primary and secondary *CESA* genes from seven different plant species we find that many components in these networks are conserved across

Table 2 | Quantification of xylem vessels with disturbed shapes for *prr1* (At1g32100).

Genotype	Distorted xylem vessel (%)	Counted cells
WT Col-0	7.8	*N* = 1364
SALK_058467	17.7	*N* = 458
SALK_090999	16.7	*N* = 436
SALK_087014	20.1	*N* = 523

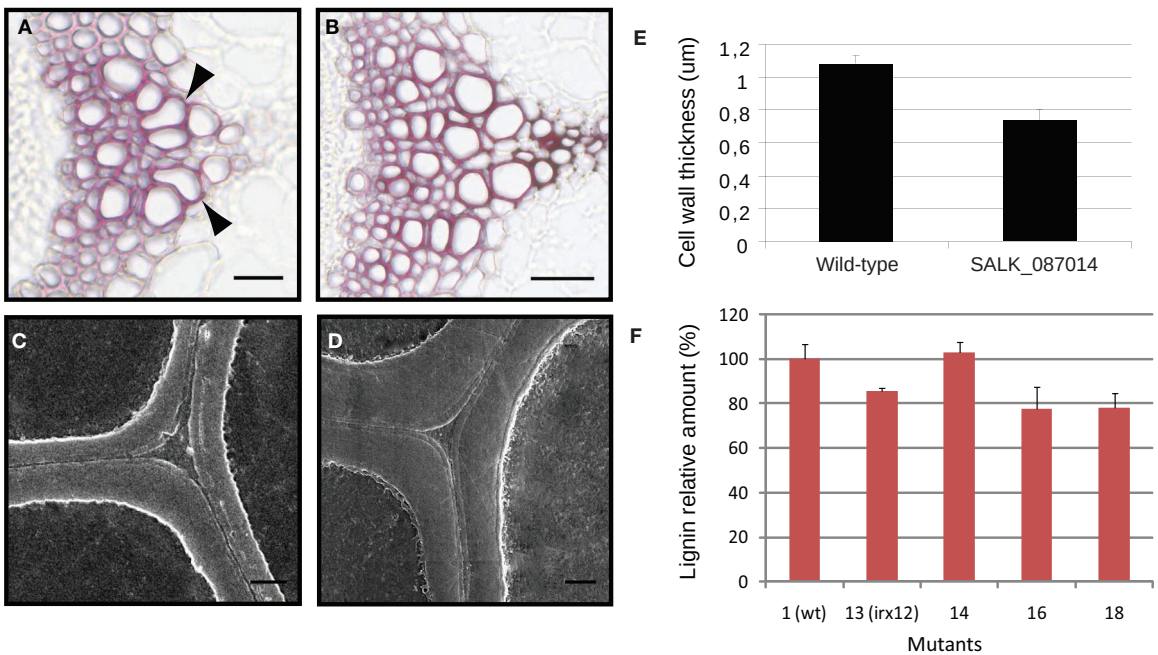


FIGURE 7 | Mutations in *PRR1* cause disturbed xylem shapes, thinner secondary cell walls and a decrease in lignin related structures. (A,B) Example of stem sections stained with phloroglucinol–HCl for SALK_087014 **(A)**; affecting *PRR1*) and wild-type **(B)**. Arrowheads indicate xylem vessels with disturbed morphology. Scale bars = 25 mm. **(C,D)** Secondary cell walls from stem sections using scanning electron microscopy from SALK_087014 **(C)** and wild-type **(D)**. Scale bars = 1 mm. **(E)** Secondary cell wall width from wild-type (*N* = 687) and SALK_087014 (*N* = 585). Displayed are the means of three different stems per line. **(F)** Thioglycolic-acid analysis of lignin contents in different mutants, presumably affecting stem lignification.

species. Based on sequence similarity and co-expression vicinities, we argue that their gene products are likely to perform similar functions in the different species, and thus that these comparative analyses may constitute an excellent tool to transfer knowledge obtained in *Arabidopsis*. We also developed a statistical pipeline of PCA and clustering analyses of cell wall sugar profiles that may be used to classify cell wall mutants. We propose that comparative co-expression analysis is a powerful approach to select and identify new genes involved in cell wall formation.

ACKNOWLEDGMENTS

We would like to thank Federico Giorgifor help with the association of probe sets to pfams, and Susann Weichold and Annemarie Martins for assistance with electron microscopy, and Malgorzata Janczarska for technical help with the biochemical cell wall analyses. This work was supported by a fellowship to Colin Ruprecht from the IMPRS program at the MPI-MP. Marek Mutwil and Staffan Persson were supported by the MPG, Friederike Saxe was financed by the IMPRS program at the MPI-KGF, Michaela Eder was funded by CASPIC, and Zoran Nikoloski was supported by the *Go*FORSYS project funded by the German Federal Ministry of Education and Research, Grant No. 0313924.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/plant_physiology/10.3389/fpls.2011.00023/abstract

REFERENCES

Albersheim, P., Nevins, D. J., English, P. D., and Karr, A. (1967). A method for the analysis of sugars in plant cell-wall polysaccharides by gas-liquid chromatography (*Acer pseudoplatanus* tissue culture cells). *Carbohydr. Res.* 5, 340–345.

Arioli, T., Peng, L., Betzner, A. S., Burn, J., Wittke, W., Herth, W., Camilleri, C., Höfte, H., Plazinski, J., Birch, R., Cork, A., Glover, J., Redmond, J., and Williamson, R. E. (1998). Molecular analysis of cellulose biosynthesis in *Arabidopsis*. *Science* 279, 717–720.

Bauer, S., Vasu, P., Persson, S., Mort, A. J., and Somerville, C. R. (2006). Development and application of a suite of polysaccharide-degrading enzymes for analyzing plant cell walls. *Proc. Natl. Acad. Sci. U.S.A.* 103, 11417–11422.

Bischoff, V., Nita, S., Neumetzler, L., Schindelasch, D., Urbain, A., Eshed, R., Persson, S., Delmer, D., and Scheible, W. (2010). Trichome birefringence and its homolog AT5G01360 encode plant-specific DUF231 proteins required for cellulose biosynthesis in *Arabidopsis*. *Plant Physiol.* 153, 590–602.

Brown, D., Wightman, R., Zhang, Z., Gomez, L. D., Atanassov, I., Bukowski, J. P., Tryfona, T., McQueen-Mason, S. J., Dupree, P., and Turner, S. (2011). Arabidopsis genes IRREGULAR XYLEM (IRX15) and IRX15L encode DUF579-containing proteins that are essential for normal xylan deposition in the secondary cell wall. *Plant J.* 66, 401–413.

Brown, D. M., Zeef, L. A., Ellis, J., Goodacre, R., and Turner, S. R. (2005). Identification of novel genes in *Arabidopsis* involved in secondary cell wall formation using expression profiling and reverse genetics. *Plant Cell* 17, 2281–2295.

Brown, D. M., Zhang, Z., Stephens, E., Dupree, P., and Turner, S. R. (2009). Characterization of IRX10 and IRX10-like reveals an essential role in glucuronoxylan biosynthesis in *Arabidopsis*. *Plant J.* 57, 732–746.

Burton, R. A., Shirley, N. J., King, B. J., Harvey, A. J., and Fincher, G. B. (2004). The CesA gene family of barley. Quantitative analysis of transcripts reveals two groups of co-expressed genes. *Plant Physiol.* 134, 224–236.

Campbell, M. M., and Ellis, B. E. (1992). Fungal elicitor-mediated responses in pine cell-cultures. 1. Induction of phenylpropanoid metabolism. *Planta* 186, 409–417.

Carpita, N., and McCann, M. (2000). "The plant cell wall," *Biochemistry and Molecular Biology of Plants*, eds B. Buchanan, W. Gruissem, and R. Jones (Rockville, MD: American Society of Plant Biologists), 52–108.

Cocuron, J. C., Lerouxel, O., Drakakaki, G., Alonso, A. P., Liepman, A. H., Keegstra, K., Raikhel, N., and Wilkerson, C. G. (2007). A gene from the cellulose synthase-like C family encodes a beta-1,4glucan synthase. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8550–8555.

Desprez, T., Juraniec, M., Crowell, E. F., Jouy, H., Pochylova, Z., Parcy, F., Höfte, H., Gonneau, M., and Vernhettes, S. (2007). Organization of cellulose synthase complexes involved in primary cell wall synthesis in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 104, 15572–15577.

Dische, Z. (1962). "Colour reactions of carbohydrates," *Methods in Carbohydrate Chemistry*, Vol. 1, eds R. L. Whistler, and M. L. Wolfrom (New York, NY: Academic Press Inc.), 478–548.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunesekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L., Eddy, S. R., and Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Res.* 38, D211–D222.

Geisler, D. A., Sampathkumar, A., Mutwil, M., and Persson, S. (2008). Laying down the bricks: logistic aspects of cell wall biosynthesis. *Curr. Opin. Plant Biol.* 11, 647–652.

Gu, Y., Kaplinsky, N., Bringmann, M., Cobb, A., Carroll, A., Sampathkumar, A., Baskin, T. I., Persson, S., and Somerville, C. R. (2010). Identification of a cellulose synthase-associated protein required for cellulose biosynthesis. *Proc. Natl. Acad. Sci. U S A.* 107, 12866–12871.

Hématy, K., and Höfte, H. (2008). Novel receptor kinases involved in growth regulation. *Curr. Opin. Plant Biol.* 11, 321–328.

Jensen, J. K., Kim, H., Cocuron, J. C., Orler, R., Ralph, J., and Wilkerson, C. G. (2011). The DUF579 domain containing proteins IRX15 and IRX15-L affect xylan synthesis in Arabidopsis. *Plant J.* 66, 387–400.

Kessler, S. A., Shimosato-Asano, H., Keinath, N. F., Wuest, S. E., Ingram, G., Panstruga, R., and Grossniklaus, U. (2010). Conserved molecular components for pollen tube reception and fungal invasion. *Science* 330, 968–971.

Kumar, M., Thammannagowda, S., Bulone, V., Chiang, V., Han, K. H., Joshi, C. P., Mansfield, S. D., Mellerowicz, E., Sundberg, B., Teeri, T., and Ellis, B. E. (2009). An update on the nomenclature for the cellulose synthase genes in *Populus*. *Trends Plant Sci.* 14, 248–254.

Lerouxel, O., Cavalier, D. M., Liepman, A. H., and Keegstra, K. (2006). Biosynthesis of plant cell wall polysaccharides – a complex process. *Curr. Opin. Plant Biol.* 9, 621–630.

Liepman, A. H., Wightman, R., Geshi, N., Turner, S. R., and Scheller, H. V. (2010). *Arabidopsis* – a powerful model system for plant cell wall research. *Plant J.* 61, 1107–1121.

Mortimer, J. C., Miles, G. P., Brown, D. M., Zhang, Z., Segura, M. P., Weimar, T., Yu, X., Seffen, K. A., Stephens, E., Turner, S. R., and Dupree, P. (2010). Absence of branches from xylan in *Arabidopsis* gux mutants reveals potential for simplification of lignocellulosic biomass. *Proc. Natl. Acad. Sci. U.S.A.* 107, 17409–17414.

Mutwil, M., Debolt, S., and Persson, S. (2008a). Cellulose synthesis: a complex complex. *Curr. Opin. Plant Biol.* 11, 252–257.

Mutwil, M., Obro, J., Willats, W. G., and Persson, S. (2008b). GeneCAT-novel webtools that combine BLAST and co-expression analyses. *Nucleic Acids Res.* 36, W320–W326.

Mutwil, M., Klie, S., Tohge, T., Giorgi, F. M., Wilkins, O., Campbell, M. M., Fernie, A. R., Usadel, B., Nikoloski, Z., and Persson, S. (2011). PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell.* 23, 895–910.

Mutwil, M., Ruprecht, C., Giorgi, F. M., Bringmann, M., Usadel, B., and Persson, S. (2009). Transcriptional wiring of cell wall-related genes in *Arabidopsis*. *Mol. Plant* 2, 1015–1024.

Mutwil, M., Usadel, B., Schütte, M., Loraine, A., Ebenhöh, O., and Persson, S. (2010). Assembly of an interactive correlation network for the *Arabidopsis* genome using a novel heuristic clustering algorithm. *Plant Physiol.* 152, 29–43.

Nakatsubo, T., Mizutani, M., Suzuki, S., Hattori, T., and Umezawa, T. (2008). Characterization of *Arabidopsis thaliana* pinoresinol reductase, a new type of enzyme involved in lignan biosynthesis. *J. Biol. Chem.* 283, 15550–15557.

Neumetzler, L. (2010). *Identification and Characterization of Arabidopsis Mutants Associated with Xyloglucan Metabolism*. Berlin: Rhombos-Verlag, 33–35.

Nicol, F., His, I., Jauneau, A., Vernhettes, S., Canut, H., and Höfte, H. (1998). A plasma membrane-bound putative

endo-1,4-beta-D-glucanase is required for normal wall assembly and cell elongation in *Arabidopsis*. *EMBO J.* 17, 5563–5576.

Oikawa, A., Joshi, H. J., Rennie, E. A., Ebert, B., Manisseri, C., Heazlewood, J. L., and Scheller, H. V. (2010). An integrative approach to the identification of *Arabidopsis* and rice genes involved in xylan and secondary wall development. *PLoS ONE* 5, e15481. doi: 10.1371/journal.pone.0015481

Pan, J. Y., Chen, S. L., Yang, M. H., Wu, J., Sinkkonen, J., and Zou, K. (2009). An update on lignans: natural products and synthesis. *Nat. Prod. Rep.* 26, 1251–1292.

Paredez, A. R., Somerville, C. R., and Ehrhardt, D. W. (2006). Visualization of cellulose synthase demonstrates functional association with microtubules. *Science* 312, 1491–1495.

Peña, M. J., Zhong, R., Zhou, G. K., Richardson, E. A., O'Neill, M. A., Darvill, A. G., York, W. S., and Ye, Z. H. (2007). *Arabidopsis* irregular xylem8 and irregular xylem9: implications for the complexity of glucuronoxylan biosynthesis. *Plant Cell* 19, 549–563.

Persson, S., Paredez, A., Carroll, A., Palsdottir, H., Doblin, M., Poindexter, P., Khitrov, N., Auer, M., and Somerville, C. R. (2007a). Genetic evidence for three unique components in primary cell-wall cellulose synthase complexes in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 104, 15566–15571.

Persson, S., Caffall, K. H., Freshour, G., Hilley, M. T., Bauer, S., Poindexter, P., Hahn, M. G., Mohnen, D., and Somerville, C. (2007b). The *Arabidopsis* irregular xylem8 mutant is deficient in glucuronoxylan and homogalacturonan, which are essential for secondary cell wall integrity. *Plant Cell* 19, 237–255.

Persson, S., Wei, H., Milne, J., Page, G. P., and Somerville, C. R. (2005). Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc. Natl. Acad. Sci. U.S.A.* 102, 8633–8638.

Rasband, W. S. (1997). *ImageJ*. Bethesda, MD: U. S. National Institutes of Health. Available at: http://rsb.info.nih.gov/ij/

Roudier, F., Fernandez, A. G., Fujita, M., Himmelspach, R., Borner, G. H., Schindelman, G., Song, S., Baskin, T. I., Dupree, P., Wasteneys, G. O., and Benfey, P. N. (2005). COBRA, an *Arabidopsis* extracellular glycosyl-phosphatidyl inositol-anchored protein, specifically controls highly anisotropic expansion through its involvement in cellulose microfibril orientation. *Plant Cell* 17, 1749–1763.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65.

Scheller, H. V., and Ulvskov, P. (2010). Hemicelluloses. *Annu. Rev. Plant Biol.* 61, 263–289.

Selvendran, R. R., Ring, S. G., and Dupont, M. S. (1979). Assessment of procedures used for analyzing dietary fiber and some recent developments. *Chem. Ind.* 225–230.

Somerville, C. (2006). Cellulose synthesis in higher plants. *Annu. Rev. Cell Dev. Biol.* 22, 53–78.

Somerville, C., Bauer, S., Brininstool, G., Facette, M., Hamann, T., Milne, J., Osborne, E., Paredez, A., Persson, S., Raab, T., Vorwerk, S., and Youngs, H. (2004). Toward a systems approach to understanding plant cell walls. *Science* 306, 2206–2211.

Spurr, A. R. (1969). A low viscosity epoxy resin embedding medium for electron microscopy. *J. Ultrastruct. Res.* 26, 31–43.

Tanaka, K., Murata, K., Yamazaki, M., Onosato, K., Miyao, A., and Hirochika, H. (2003). Three distinct rice cellulose synthase catalytic subunit genes required for cellulose synthesis in the secondary wall. *Plant Physiol.* 133, 73–83.

Taylor, N. G. (2008). Cellulose biosynthesis and deposition in higher plants. *New Phytol.* 178, 239–252.

Taylor, N. G., Laurie, S., and Turner, S. R. (2000). Multiple cellulose synthase catalytic subunits are required for cellulose synthesis in *Arabidopsis*. *Plant Cell* 12, 2529–2540.

Theodoridis, S., and Koutroumbas, K. (2006). *Pattern Recognition*, *3rd Edn*. San Diego, CA: Academic Press, 635.

Turner, S. R., and Somerville, C. R. (1997). Collapsed xylem phenotype of *Arabidopsis* identifies mutants deficient in cellulose deposition in the secondary cell wall. *Plant Cell* 9, 689–701.

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F. M., Bassel, G. W., Tanimoto, M., Chow, A., Steinhauser, D., Persson, S., and Provart, N. J. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ.* 32, 1633–1651.

Weigel, D., and Glazebrook, J. (2002). *Arabidopsis: A Laboratory Manual*. New York: Cold Spring Habour Laboratory Press, 195–203.

Wightman, R., and Turner, S. (2010). Trafficking of the cellulose synthase complex in developing xylem vessels. *Biochem. Soc. Trans.* 38, 755–760.

Zhang, L., Henriksson, G., and Gellerstedt, G. (2003). The formation of beta-beta structures in lignin biosynthesis–are there two different pathways? *Org. Biomol. Chem.* 1, 3621–3624.

Zhong, R., Kays, S. J, Schroeder, B. P., and Ye, Z. (2002). Mutation of a chitinase-like gene causes ectopic deposition of lignin, aberrant cell shapes, and overproduction of ethylene. *Plant Cell* 14, 165–179.

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.

Zhong, R., Richardson, E. A., and Ye, Z. H. (2007). The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in *Arabidopsis*. *Plant Cell* 19, 2776–2792.

Conflict of Interest Statement: 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.

Received: 31 March 2011; accepted: 14 June 2011; published online: 01 July 2011.
*Citation: Ruprecht C, Mutwil M, Saxe F, Eder M, Nikoloski Z and Persson S (2011) Large-scale co-expression approach to dissect secondary cell wall formation across plant species. Front. Plant Sci. **2**:23. doi: 10.3389/fpls.2011.00023*
This article was submitted to Frontiers in Plant Physiology, a specialty of Frontiers in Plant Science.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 03 October 2011
doi: 10.3389/fpls.2011.00061

From models to crop species: caveats and solutions for translational metabolomics

Takayuki Tohge[1]*, Tabea Mettler[1], Stéphanie Arrivault[1], Adam James Carroll[2], Mark Stitt[1] and Alisdair R. Fernie[1]

[1] *Max-Planck-Institute for Molecular Plant Physiology, Potsdam-Golm, Germany*
[2] *Australian Research Council Centre of Excellence in Plant Energy Biology, The Australian National University, Canberra, ACT, Australia*

Edited by:
Wolf B. Frommer, Carnegie Institution for Science, USA

Reviewed by:
Jin Chen, Michigan State University, USA
Woei-Jiun Guo, National Cheng Kung University, Taiwan

****Correspondence:***
Takayuki Tohge, Max-Planck-Institute for Molecular Plant Physiology, Am Muehlenberg 1, 14476 Potsdam-Golm, Germany.
e-mail: tohge@mpimp-golm.mpg.de

Although plant metabolomics is largely carried out on *Arabidopsis* it is essentially genome-independent, and thus potentially applicable to a wide range of species. However, transfer between species, or even between different tissues of the same species, is not facile. This is because the reliability of protocols for harvesting, handling and analysis depends on the biological features and chemical composition of the plant tissue. In parallel with the diversification of model species it is important to establish good handling and analytic practice, in order to augment computational comparisons between tissues and species. Liquid chromatography–mass spectrometry (LC–MS)-based metabolomics is one of the powerful approaches for metabolite profiling. By using a combination of different extraction methods, separation columns, and ion detection, a very wide range of metabolites can be analyzed. However, its application requires careful attention to exclude potential pitfalls, including artifactual changes in metabolite levels during sample preparation under variations of light or temperature and analytic errors due to ion suppression. Here we provide case studies with two different LC–MS-based metabolomics platforms and four species (*Arabidopsis thaliana*, *Chlamydomonas reinhardtii, Solanum lycopersicum,* and *Oryza sativa*) that illustrate how such dangers can be detected and circumvented.

Keywords: plant metabolomics, sample preparation, ion suppression, chemical diversity, translational biology, LC–MS

INTRODUCTION

Plant metabolomics is a relatively new analytic strategy which provides complementary information to transcriptomic and proteomic studies as well as important information in its own right concerning the regulation of metabolic networks (Hall et al., 2002; Bino et al., 2004). Initial applications of metabolic profiling were largely focused on the model plant *Arabidopsis thaliana* (von Roepenack-Lahaye et al., 2004; Tohge et al., 2005; Gibon et al., 2006; Trenkamp et al., 2009; Araujo et al., 2010; Kerwin et al., 2011), however, several studies have been carried out on the green algae *Chlamydomonas reinhardtii* (Giroud et al., 1988; Bolling and Fiehn, 2005; May et al., 2008; Boyle and Morgan, 2009; Renberg et al., 2010) with other successful applications being reported for *Catharanthus roseus* (Rischer et al., 2006), *Fragaria x ananassa* (Aharoni et al., 2000, 2002; Hanhineva et al., 2008), *Hordeum vulgare* (Widodo Patterson et al., 2009), *Medicago truncatula* (Achnine et al., 2005), *Nicotiana tabacum* (Goossens et al., 2003), *Oryza sativa* (Albinsky et al., 2010), *Perilla frutescens* (Yamazaki et al., 2008), *Pisum sativum* (Jom et al., 2010), and *Solanum lycopersicum* (Schauer et al., 2005, 2006; Moco et al., 2006; Fraser et al., 2007) as well as the unicellular prokaryotes *Synechocystis* sp. (Krall et al., 2009) and the diatom *Phaeodactylum tricornutum* (Allen et al., 2008).

Initially, the use of metabolic profiling in plants, as indeed in all species, was restricted to diagnostic approaches in which the obtained profiles were used as markers for a range of biological conditions (Sauter et al., 1988; Meyer et al., 2007; Semel et al., 2007; Carmo-Silva et al., 2009; Scherling et al., 2009; Widodo Patterson et al., 2009). Although such studies remain highly important, particularly in medical research (Nicholson and Wilson, 2003; Griffin and Nicholls, 2006), more sophisticated uses of metabolic profiling have recently been developed, including identifying regulated enzymes and exploring the regulatory structure of pathways (Tiessen et al., 2002; Arrivault et al., 2009), searching for unexpected effects of genetic manipulation (Catchpole et al., 2005), screening wild species for beneficial chemical composition (Zhu and Wang, 2000; El-Lithy et al., 2005), gaining a more comprehensive view of metabolic regulation and as part of integrative analyses for the systemic response of environmental genetic perturbations (Hirai et al., 2004, 2005; Fukushima et al., 2009; Sulpice et al., 2009; Trenkamp et al., 2009). In addition to these uses, metabolomics is proving to be a powerful tool for gene functional annotation in plants. There are now several examples of *Arabidopsis* genes that have been identified with the help of metabolomic approaches including MYB transcription factors (Hirai et al., 2007; Stracke et al., 2007), *O*-methyltransferase (Tohge et al., 2007), glycosyltransferases (Tohge et al., 2005; Yonekura-Sakakibara et al., 2007, 2008), acyltransferases (Luo et al., 2007), UDP-rhamnose synthase (Yonekura-Sakakibara et al., 2008), and pyrophosphorylase (Okazaki et al., 2009) with the approach being equally effective in other species (Aharoni et al., 2000; Goossens et al., 2003; Achnine et al., 2005; Yamazaki et al., 2008).

One advantage that metabolomics has over transcriptomics [with the exception of next-generation sequencing tools, see Detlef Weigels recent review (Schneeberger and Weigel, 2011)] and proteomics is that it is essentially genome-independent (Stitt and Fernie, 2003) and as such can be applied to a species whose genome has not been sequenced as easily as those whose has. This "democratization" of biology allows in depth functional analyses of many species for which a complete and fully annotated genome is not yet available (Schneeberger and Weigel, 2011). Despite this fact caution needs to be taken when adopting a method set up for one tissue of one species to analyze another tissue of that species or even another species. This is especially so for metabolite profiling. The plant kingdom contains an incredibly rich chemical diversity (St-Pierre and De Luca, 2000). It is obvious that this chemical diversity poses a large challenge and stimulates research in developing new and increasingly powerful approaches to separate, detect, and identify metabolites. However, it also raises important challenges for experimental design, sample handling, and validation of analytic procedures. This is because tissue composition affects the reliability with which a particular metabolite can be reliably extracted and analyzed. This problem is particularly acute when using liquid chromatography–mass spectrometry (LC–MS) due to the so called ion suppression effects wherein the composition of the extract affects the efficiency of ionization of some of its constituent analytes (Fernie et al., 2004). That said, a number of relatively simple control tests, in combination with the growing number of chemoinformatic tools for metabolomics (Tohge and Fernie, 2009; Bais et al., 2010; Carroll et al., 2010; Cottret et al., 2010; Xia and Wishart, 2010), should at least ameliorate this phenomenon and hence facilitate high-quality translational metabolomics.

Driven by the increasing diversification of plant research away from the principle model species *A. thaliana* we present here case studies in which methods developed for this species are assessed for use in determining metabolite levels either in the unicellular algae *C. reinhardtii* or in the crop species rice and tomato. For the former we assessed the analysis of primary metabolism using an LC–MS/MS method developed to deliver validated measurements of the levels of Calvin–Benson cycle intermediates, organic acids, nucleotide-sugars, and nucleotides in *Arabidopsis* rosettes (Arrivault et al., 2009). Given that information documenting the transfer of gas chromatography–mass spectrometry (GC–MS)-based methods of analysis of primary metabolites has already been extensively supplied for potato and tomato (Roessner et al., 2001; Roessner-Tunali et al., 2003), we chose crop species to focus our studies on secondary metabolism. The two LC–MS-based methods applied in this study complement standard and well-established GC–MS methods by greatly increasing the range of metabolites that can be analyzed.

Here some examples of how can be performed using two different LC–MS-based metabolomics platforms, on one algal and two crop species and *A. thaliana* are shown. The combined results illustrate important experimental controls which should be implemented alongside computation algorithms in order to successfully adapt protocols that have been established for another biological system. This also applies to other LC–MS-based methods (Okazaki et al., 2009; Kanno et al., 2010). We additionally discuss how such studies could be used in conjuncture with novel tools for combined sequence comparison and co-expression analysis (Mutwil et al., 2011, and Ruprecht et al., this issue) in order to improve gene functional predictions from *Arabidopsis* to crop species.

MATERIALS AND METHODS

CELL CULTURE AND EXTRACTION PROCEDURES

Chlamydomonas reinhardtii strain CC-1690 wild type mt+ was acquired from the *Chlamydomonas* Genetics Center (Duke University, Durham, NC, USA). Single colonies were used to inoculate the growth media containing 5 mM Hepes, 1 mM K-phosphates, Beijerinck salts (final concentrations of 7.5 mM NH_4Cl, 0.34 mM $CaCl_2$, 0.41 mM $MgSO_4$) and trace salt solution (final concentrations of 184 μM H_3BO_3, 77 μM $ZnSO_4$, 26 μM $MnCl_2$, 18 μM $FeSO_4$, 7 μM $CoCl_2$, 6 μM $CuSO_4$, 1 μM $(NH_4)_6Mo_7O_{24}$; Harris, 1989; May et al., 2008; Kempa et al., 2009) at 25°C under constant illumination with 400 μmol photon m^{-2} s^{-1} and continuous shaking. The amount of NH_4Cl was reduced to 4 mM for the experiment shown in **Figure 1** to reduce the impact of ion suppression. Before harvesting, cells were grown to a density of 3×10^6 cells ml^{-1} and dark-adapted for a minimum of 20 min before transferring 1 ml of cells to a cuvette and exposed them to 660 μmol photon m^{-2} s^{-1} under continuous stirring. Before illumination and at different time points after illumination the suspension was quenched by vigorously adding 2 ml of −70°C methanol (70%). The entire mix was then lyophilized to dryness at −80°C and extracted at 4°C by a chloroform:methanol:water (1:2:5 [v/v]) mixture. Water fractions of three subsequent washes were collected, concentrated by lyophilization at −80°C and filtered before metabolite measurement (80 μl of extracted culture in 100 μl sample measured) by ion pair (reverse-phase) chromatography triple quadrupole MS (IPC–MS/MS) detection.

PLANT MATERIALS AND EXTRACTION PROCEDURES

Arabidopsis thaliana ecotype Col-0 and *S. lycopersicum* (M82) were grown in soil in a controlled environmental chamber (16 h light/8 h dark photoperiod; 21°C at 145 μmol photon m^{-2} s^{-1} and 25°C at 500 μmol photon m^{-2} s^{-1}, respectively). *A. thaliana* used for recovery test was grown in general 1/2 MS agar plate in controlled plant growth chamber (16 h light/8 h dark photoperiod; 21°C). Rice (*O. sativa*, Nipponbare) seeds were pre-germinated in tap water at 28°C for 10 days. Plantlets were transferred to a controlled growth chamber with 12 h day length at 700 μmol photon m^{-2} s^{-1}. Plant material was frozen in liquid nitrogen, ground into powder and stored at −80°C until use.

For assessment of ion suppression and recovery tests, extraction for secondary metabolite profiling was conducted as described in Tohge and Fernie (2010). Extraction buffer was added to reach 0.2 mg FW μl^{-1}. To evaluate secondary metabolite degradation due to enzymatic activities, four different extraction procedures were carried out using aliquots of frozen powders from a pool of plant materials from at least three plants: (a) extraction was conducted as described in Tohge and Fernie, 2010; extraction buffer was added to frozen material kept at liquid nitrogen temperature); (b) extracts obtained by method (a) were incubated at 37°C for 1 h, (c) extraction buffer was immediately added to frozen material on ice; (d) plant material was incubated at 37°C for 1 h prior addition of extraction buffer. To assess ion suppression in different plant

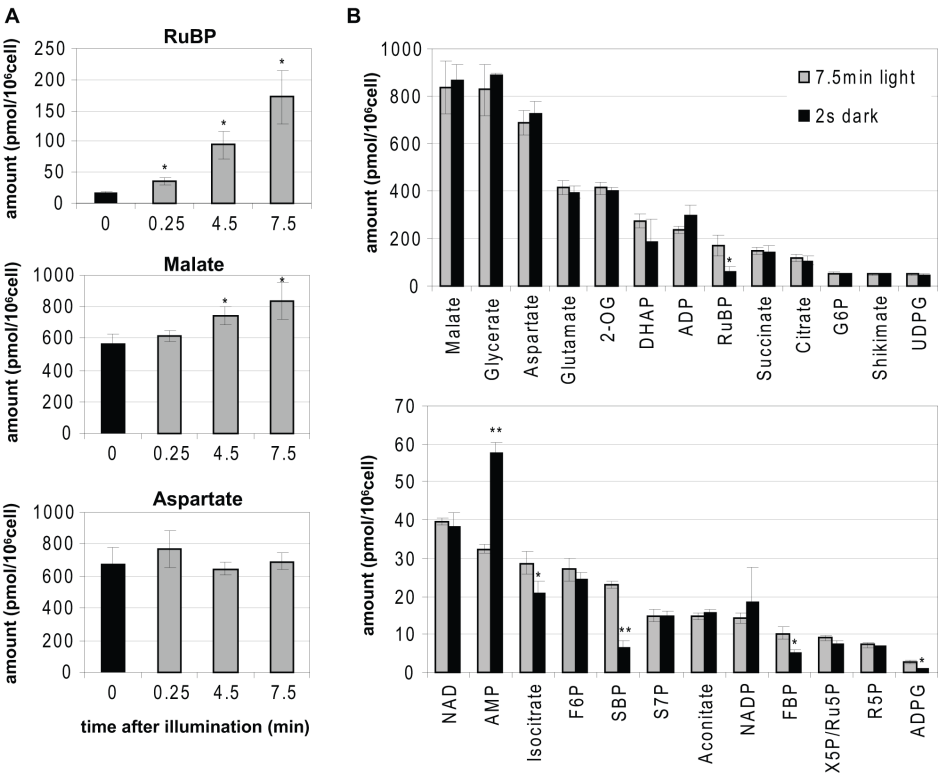


FIGURE 1 | Example of rapid metabolic response by light and darkness. Metabolites in *Chlamydomonas reinhardtii* CC-1690 were measured after quenching in an excess of cold (−70°C) methanol, lyophilization, and extraction in chloroform-methanol using IPC–MS/MS. **(A)** Ribulose-1,5-bisphosphate (RuBP), malate, and aspartate levels in *Chlamydomonas* CC-1690 cells are shown after dark-adaption for 20 min (black bars) and exposure to 660 μmol photon m^{-2} s^{-1} for 0.25, 4.5, and 7.5 min (gray bars). *Y*-axis indicates amount (pmol/10^6 cell). **(B)** *Chlamydomonas* cells were harvested in cuvettes after 20 min dark-adaption and exposure to 660 μmol photon m^{-2} s^{-1} for 7.5 min without (gray bars) or with an additional 2 s of darkness (black bars).. Levels of metabolites are presented as absolute values ($n = 3$, ±SD, two asterisks: Student's *t*-test $p < 0.01$, one asterisk: Student's *t*-test $p < 0.05$). 2-OG, 2-oxoglutarate; DHAP, dihydroxyacetone-phosphate; ADP, adenosine diphosphate; G6P, glucose-6-phosphate; UDPG, UDP-glucose; NAD, nicotinamide adenine dinucleotide; AMP, adenosine monophosphate; F6P, fructose-6-phosphate; SBP, sedoheptulose-1,7-bisphosphate; S7P, sedoheptulose-7-phosphate; NADP, nicotinamide adenine bisnucleotide phosphate; FBP, fructose-1,6-phosphate; X5P, xylulose-5-phosphate; Ru5P, ribulose-5-phosphate; R5P, ribose-5-phosphate; ADPG, ADP-glucose.

species, an internal standard (IS) mixture containing three standard compounds (isovitexin, CAS: 29702-25-8; saponarin, CAS: 20310-89-3; sinigrin, CAS: 3952-98-5) was prepared at four different concentrations (20, 10, 5, 1 μg ml^{-1}). Identical volume of standard mixture and plant extracts were added, resulting in a final sample containing 0.1 mg FW μl^{-1} of plant extracts and 10, 5, 2.5, or 0.5 μg ml^{-1} of standard compounds. Recovery test was carried out with *Arabidopsis* extracts (0.2 mg FW μl^{-1}) of leaves and roots grown on agar plates for 3 weeks, and flowers harvested from the plants grown on soil for 4 weeks. Extracts from leaves "A" and roots "B" (or flowers) were mixed at different ratios [(A:B), 90:10, 80:20, 50:50, 20:80, 10:90], respectively. The percentage recovery was estimated for evaluation using theoretical concentration of extracts mixture, [(level in leaves × A%) + (level in roots (or flowers) × B%)]/100.

ION PAIR (REVERSE-PHASE) CHROMATOGRAPHY TRIPLE QUADRUPOLE MS (IPC–MS/MS)

Primary metabolite analysis by IPC–MS/MS was carried out on a Dionex HPLC system (Sunnyvale, CA, USA) coupled to a Finnigan TSQ Quantum Discovery MS-Q3 (Thermo Fisher Scientific, Waltham, USA) equipped with an electrospray ionization (ESI) interface. It was operated as described in Arrivault et al. (2009). Chromatographic separation was obtained at 35°C by a multi-step gradient with online-degassed eluent A (10 mM tributylamine aqueous solution, adjusted to pH 4.95 with 15 mM acetic acid) and eluent B (methanol) applied to a Gemini (C18) 150 mm × 2.00 mm inner diameter 5 μm, 110 Å particle column (Phenomenex, Aschaffenburg, Germany). The MS-Q3 device was operated in the negative ion scanning mode with selected reaction monitoring (SRM). The MS-parameters for each compound are documented in the Supplementary Data of Arrivault et al. (2009). Calibration curves using authentic standards were used to calculate absolute amounts of metabolites in algal samples. Data were processed using LC-quan 2.5.6 SP1 software.

REVERSE-PHASE HPLC–MS ANALYSIS

Secondary metabolite analysis by LC–MS was performed on HPLC system Surveyor (Thermo Finnigan, USA) coupled to Finnigan LTQ-XP system (Thermo Finnigan, USA) as described by Tohge

and Fernie (2010). All data were processed using Xcalibur 2.1 software (Thermo Fisher Scientific, Waltham, USA). Metabolite identification and annotation were performed using standard compounds (Nakabayashi et al., 2009) and reference metabolomics databases (Moco et al., 2006; Shinbo et al., 2006; Iijima et al., 2008; Tohge and Fernie, 2009).

RESULTS AND DISCUSSION

HARVESTING – OBTAINING REPRESENTATIVE MATERIAL AND AVOIDING HANDLING-INDUCED CHANGES

Expression of genes and activity of enzymes associated with photosynthesis, respiration, and energy metabolism are rapidly affected by changes in environmental conditions. Transcriptional and metabolic regulation by the circadian clock has been defined (Harmer et al., 2000; Gibon et al., 2006; Fukushima et al., 2009; Kerwin et al., 2011). Many metabolites showed marked diurnal changes. Problems related to variation in clock and diurnal rhythms can be circumvented by harvesting plants at the same time in the 24-h cycle. They can also be affected by shorter term fluctuations. The exact timing of harvesting and avoidance of perturbation of metabolism during harvesting, by for instance shading of leaves or changes in the oxygen tension are therefore critical (see Geigenberger et al., 2000). Additionally, rapid and complete quenching of metabolic activity is crucial to ensure faithful measurement of the intracellular metabolite content (discussed in details below).

To highlight the rapid metabolic response following light treatment, metabolite profiling was performed on the model organism *C. reinhardtii* (**Figure 1**). Quenching was performed by rapid mixing of the algal suspension in the light with an excess of very cold (−70°C) methanol to instantaneously freeze the cells. Metabolite profiling by IPC–MS/MS analysis of short term light treatment was performed in dark-adapted material and after 0.25, 4.5, and 7.5 min illumination. Ribulose-1,5-bisphosphate (RuBP), which is a major metabolite in the Calvin–Benson cycle, was already significantly elevated by a light treatment of 15 s (**Figure 1A**). The levels of malate, the late step in TCA cycle, also displayed significant increases upon illumination, whilst the level of the amino acid aspartate was not altered.

Figure 1B illustrates why avoiding perturbations of the conditions by the mean of rapid harvesting is critical for metabolite profiling. Darkening significantly and almost instantaneously influences operation of the photosystems and the delivery of ATP and NADPH. The levels of RuBP, sedoheptulose-1,7-bisphosphate (SBP), fructose-1,6-bisphosphate (FBP), ADP-glucose (ADPG), and isocitrate were significantly decreased, while adenosine 5′-diphosphate (ADP) and adenosine monophosphate (AMP) increased within 2 s of darkening. Thus, harvesting protocols that lead to even very brief decrease or increase in the light intensity preceding quenching or during the quenching process will lead to erroneous estimates for the levels of metabolites. An identical problem arises in higher plants. This is due to the simple fact that the fluxes in the Calvin–Benson cycle are so high that many of the metabolites in the cycle as well as ATP and NADPH have short turnover times of 1 s or less (Arrivault et al., 2009).

This problem is of course especially critical for processes like photosynthesis, where fluxes are very fast and metabolite pools are small and turn over very quickly. However, it illustrates the more general points that (i) all available information about the turnover times of the metabolites-of-interest should be collected, evaluated, and used to design an appropriate harvesting and quenching protocol and (ii) that this protocol should be validated by checking if slowing down or speeding up the harvesting process modifies the levels of metabolites that are found in the harvested material.

QUENCHING OF ENZYMATIC ACTIVITIES AND DIFFERENCES BETWEEN PLANT SPECIES AND CHEMICAL PROPERTIES

Quenching of metabolic activity is not only essential to stop metabolic turnover in the running pathways, but also to inhibit other enzymatic activities that can destroy the metabolite after tissue disruption. An old but still instructive example of this is the precautions needed to determine pyrophosphate levels in plants (Weiner et al., 1987). Pyrophosphatase activity in leaves is so high that it can hydrolyse all the pyrophosphate in a leaf extract in <0.05 s. In an intact tissue, the vast majority of the pyrophosphatase activity is in the plastids whilst the pyrophosphate is in the cytosol. As soon as the tissue is disrupted the pyrophosphatase comes into contact with and destroys the pyrophosphate. To measure pyrophosphate it is therefore essential that enzymatic activity is completely stopped by rapid quenching and remains totally inactive during all subsequent stages in sample handlings as a fraction of a percent would be enough to destroy all the pyrophosphate within few seconds. Such problems can be routinely identified by recovery experiments in which representative amounts of authentic standards are added to the plant material before extraction, and it is checked that the added standard can be quantitatively detected in the final extract (Fernie et al., 2011).

Whilst secondary metabolites do not display such rapid responses to changes in the environment as those observed for primary metabolism (see an example in Kusano et al., 2011), they, like primary metabolites, are highly susceptible to degradation by enzymes that come in contact with them after tissue disruption. For example glucosinolates are converted into isothiocyanates by myrosinase in *Arabidopsis* (Tierens et al., 2001; Barth and Jander, 2006). Given that degradative enzymes typically remain potent subsequent to freezing in liquid nitrogen when the extract is thawed, particular care must be taken during the extraction procedure. To illustrate this point, three extraction procedures were conducted in addition of our original extraction procedure (extraction a), using frozen powder of plant materials. To test if breakdown enzymes were definitively inactivated during this extraction procedure, extracts were incubated at 37°C for 1 h (extraction b). A pre-incubation of sample material at 37°C for 1 h prior extraction was conducted to test the extent of metabolite degradation upon thawing (extraction d). Secondary metabolite extraction is routinely performed at liquid nitrogen temperature, so to test if another temperature would affect metabolic composition, addition of buffer on frozen sample was performed at ice temperature (extraction c). All extractions were performed using pre-cooled extraction buffer (10°C). Metabolite breakdown was assessed in *A. thaliana* leaves, *O. sativa* leaves, and *S. lycopersicum* fruits by mean of LC–MS. Total ion chromatograms and relative peak areas of selected metabolites are presented in **Figures 2** and **3**, respectively.

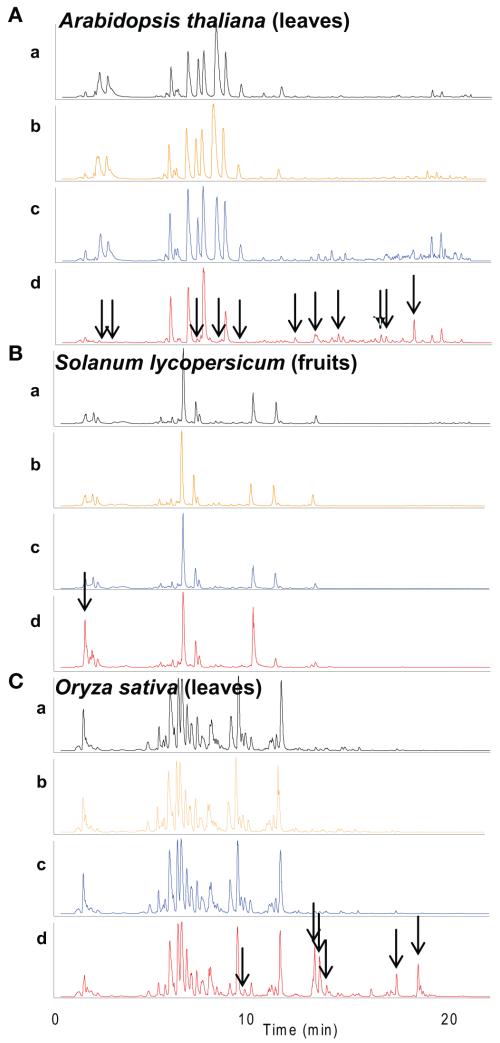


FIGURE 2 | Effect of different extraction methods on secondary metabolite breakdown in *Arabidopsis* leaves, tomato fruits, and rice leaves. Total ion chromatograms (TIC) monitored by negative ion detection mode of extracts of **(A)** *Arabidopsis* leaves, **(B)** tomato fruits, and **(C)** rice leaves, are shown. **(a–d)** indicate different extraction methods. **(a)** extraction method as described in Tohge and Fernie (2010), **(b)** extracts obtained by **(a)** method were incubated at 37˚C for 1 h, **(c)** extraction buffer was immediately added to frozen material on ice, **(d)** frozen sample was incubated at 37˚C for 1 h before extraction. All extractions were performed using pre-cooled extraction buffer (10˚C). Arrows show peaks which were newly or not detected in treated samples.

For almost all plant species, an incubation of the material after addition of extraction buffer at 37˚C for 1 h had no observable consequences on the total ion chromatograms and on the relative peak areas of selected metabolites [**Figures 2** and **3**, respectively. Comparison between (a) and (b)]. This result indicates that the major secondary metabolites are not broken down at 37˚C, provided the tissue has been taken up in the extraction buffer. However, metabolite profiling of samples extracted after 1 h (pre-incubation of the disrupted tissue at 37˚C) revealed that samples were significantly changed in some compound species (pointed by arrows in **Figure 2**). A more detailed analysis revealed that within a plant species and between various plant species the metabolite classes were differently affected [**Figure 3**, comparison between (a) and (d)]. In general non-pigmented flavonoids such as flavonol glycoside in *Arabidopsis* leaves and *S. lycopersicum* fruits, and glycoflavone in *O. sativa* leaves were stable. By contrast, red-pigmented flavonoids namely anthocyanin derivatives in *Arabidopsis* were significantly decreased following the 37˚C pre-incubation. Furthermore, phenylpropanoids in *Arabidopsis* such as sinapoyl-derivatives were broken down by pre-incubation. The Brassica species specific secondary metabolites glucosinolates are well-known compounds which can be broken down by myrosinase (Tierens et al., 2001; Barth and Jander, 2006). The glucosinolate levels in extracts with 37˚C incubation before extraction were not detected [**Figure 3A**, comparison between (a) and (d)]. Despite the breakdown of phenylpropanoids by enzyme activity in *Arabidopsis*, phenylpropanoids in tomato fruit such as chlorogenic acid related compounds were generally unaffected by enzymes. With the exception of esculeoside related compounds, levels of glycoalkaloids which are the major alkaloid in tomato fruits were not significantly changed (**Figure 3B**). These data show that, whatever the plant species, a pre-incubation of the plant material at 37˚C prior to extraction leads generally to various levels of secondary metabolite breakdown due to the presence of active enzymes when plant material is thawed out. Addition of extraction buffer to frozen material not at liquid nitrogen temperature (extraction c) led to a significant decrease in the levels of glucosinolates and anthocyanin derivatives (**Figure 3A**). This shows that the temperature during addition of the extraction buffer is also an important factor.

These results taken together illustrate that the tissue extraction should be carried out in the proper way with attention being taken to empirically optimize the extraction method for each and every new tissue measured. The effect of 20 min sonication was also evaluated in the same manner, but no differences were observed (data not shown). That said it is important to note that sonication should only be performed if this can be managed without an increase in temperature.

ION SUPPRESSION EFFECTS CAUSED BY GROWTH MEDIA

Although LC–MS analysis is a highly sensitive technique, ion suppression is a general problem of LC–MS analytical platforms due to altered ESI of a target ion by a contamination (Ikonomou et al., 1990; Kebarle and Tang, 1993; Buhrman et al., 1996; Matuszewski et al., 1998, 2003; King et al., 2000; Fernie et al., 2004). It is actually not a single event but a range of response–reducing phenomena which should be avoided as much as possible. While there is, however, no universal solution to this problem, understanding difference between samples and assessing the effects of ion suppression affords greater confidence in the accuracy of the results.

An example of ion suppression caused by growth conditions of *Chlamydomonas* is shown in **Figure 4**; **Table 1**. This example illustrates how the composition of the growth media can dramatically affect the reliability of metabolite analyses in *Chlamydomonas*. Following quenching of the algal suspension by mixing with an excess of cold methanol (see above), we took the entire suspension for analysis. This was necessary because some metabolites leak

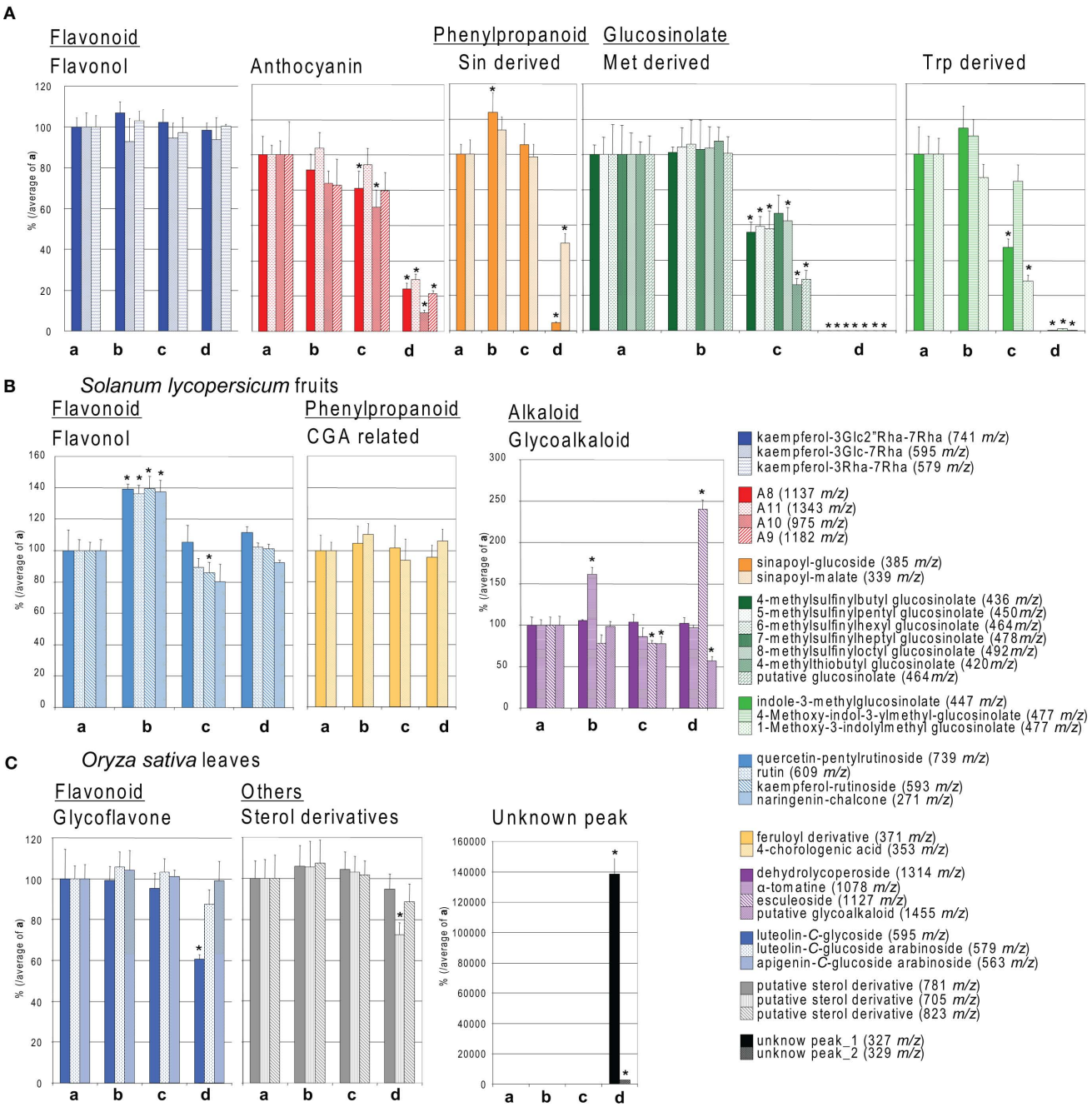


FIGURE 3 | Comparison of secondary metabolite levels between different extraction methods. Relative peak area of the result described in **Figure 2** of **(A)** *Arabidopsis* leaves, **(B)** tomato fruits, and **(C)** rice leaves, are shown. *Y*-axis indicates percentage calculated by average of peak area of treatment **(a)**. **(a–d)** indicate different extraction methods as described in **Figure 2** and Section "Materials and Methods." Blue related colors indicate flavonoid related compounds (flavonol and glycoflavone). Red color indicates red-pigmented flavonoid related compounds (anthocyanin). Yellow, phenylpropanoid related compounds (sinapoyl-derivatives and chlorogenate-related compound); green, glucosinolate related compounds in *Arabidopsis* (methionine-derived aliphatic glucosinolate and tryptophan-related aromatic glucosinolate); purple, indicates glycoalkaloid; gray, putative sterol derivative; black, unknown compounds. The *m/z* value detected in negative ion detection is indicated between parentheses. Average of three experimental replicates ±SD. Significant differences by *t*-test ($p < 0.05$) are marked by asterisks.

out of the cells into the methanol–water mix and are therefore lost when the quenched cells are harvested by centrifugation (data not shown, see also Krall et al., 2009). This means that metabolites from the cells must be analyzed in a matrix that contains methanol and all the components of the suspension medium. Because the *Chlamydomonas* cells are quite diluted, components of the growth

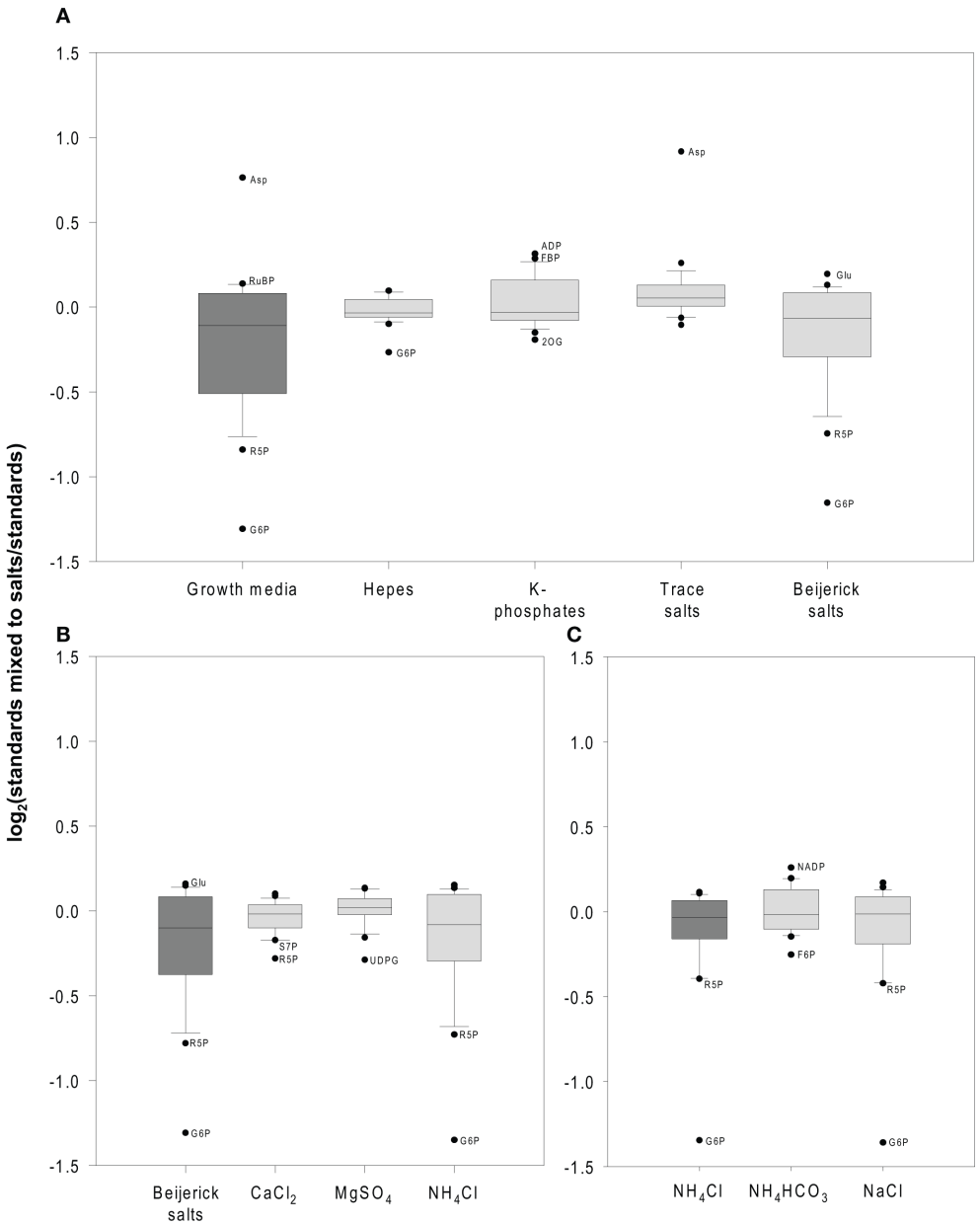


FIGURE 4 | Example of ion suppression caused by growth media. (A–C) A standard mix containing all measured metabolites was mixed with individual components of the growth media to identify which component(s) of the growth media most severely influence ionization during electrospray ionization (ESI). **(A)** The standard mix was mixed with the whole growth medium (dark gray) and independently with the medium components Hepes, K-phosphates, trace salts, and Beijerinck salts (light gray). **(B)** In a second experiment, the individual components of the Beijerinck salts were tested for ion suppression and enhancement. Standards were mixed to all Beijerinck salts (dark gray) and independently to its components $CaCl_2$, $MgSO_4$, and NH_4Cl (light gray). **(C)** A third experiment investigated if the anion or the cation was responsible for ion suppression caused by ammonium chloride. Standards were mixed to NH_4Cl (dark gray), NH_4HCO_3, or NaCl (light gray). The data is shown as box plots of the average values (calculated from three technical replicates) for 24 metabolites. Significant outliers ($p < 0.05$) are identified in the figure panels. For a complete list of percentage of ion suppression and enhancement for all metabolites see **Table A1** in Appendix.

medium are present in rather large amounts compared to metabolites in the cells. Unfortunately, some components of the growth medium lead to ion suppression.

We were alerted to ion suppression by three routine checks. First, comparison of the spectrum of metabolites with those expected from earlier studies of metabolites during photosynthesis in *Arabidopsis* showed low levels of several metabolites. Second, we checked whether the signal for each metabolite shows a strictly linear relationship to the amount of extract applied. In the case of ion suppression, the estimated levels of many metabolites decreased

Table 1 | Ion suppression mainly caused by growth media.

	I	II	III	IV	V
Methanol	−	+	−	+	+
Growth media	−	−	+	+	+
Cells	−	−	−	−	+
RuBP	100	96	110	106	165
Citrate	100	107	104	104	128
FBP	100	102	96	98	115
SBP	100	102	101	98	111
AMP	100	105	94	94	110
ADPG	100	97	101	100	109
ADP	100	137	89	116	107
2OG	100	117	86	84	101
NADP	100	98	99	97	98
Malate	100	95	94	96	94
Isocitrate	100	98	96	101	93
Aspartate	100	104	134	131	92
Glutamate	100	100	88	90	90
DHAP	100	125	86	86	85
NAD	100	115	77	76	84
Aconitate	100	103	92	93	83
UDPG	100	106	65	64	83
X5P/Ru5P	100	119	77	85	83
S7P	100	108	88	90	74
G1P	91	97	70	75	nd
F6P	100	108	78	79	73
Glycerate	100	119	72	71	66
R5P	100	115	67	70	55
G6P	100	110	35	35	39

Standards were mixed to methanol and growth media independently or together (column II–IV). In addition, standards were mixed to Chlamydomonas reinhardtii CC1690 cells grown in growth media and quenched by methanol (column V). Values represent recovery ratios for all standards (n = 3, SD of raw data <18%).

when more samples were applied. Third, we checked the recovery of authentic standards added to the extract and found it was very low.

Attempts to analyze high concentrations of sample resulted in an almost complete suppression of all metabolite signals, including those of spiked standards (**Figure A1** in Appendix). A five-time dilution allowed an average of 87–93% recovery of the spiked standards (**Table 1**; **Figure 4**). However, there was still a residual ion suppression, and this varied from metabolite to metabolite (**Table 1**). This obviously still prevents reliable and comparable analysis of metabolite levels. We therefore carried out a further series of experiments to identify the major sources of ion suppression, in order to modify the growth medium and circumvent this problem. To show that ion suppression was caused by components of the growth medium and not the biological sample itself or the methanol, we first mixed known amounts of standards with either methanol and/or the growth media compared to the same known amounts of standards mixed with a sample containing *Chlamydomonas* cells (**Table 1**). For more than half of the metabolites, <15% of the signal was suppressed by the media, but for many other metabolites including dihydroxyacetone-phosphate (DHAP), fructose-6-phosphate (F6P), glycerate, NAD^+, ribose-5-phosphate (R5P), sedoheptulose-7-phosphate (S7P), UDP-glucose (UDPG), and xylulose-5-phosphate/ribulose-5-phosphate (X5P/Ru5P) the signals were reduced by 15–50%. For G6P ion suppression caused a decrease of intensities by >50%. Therefore, for all these metabolites the absolute values have to be treated with extreme caution. Further, small changes in the extent of ion suppression can lead to changes in the relative signals for the various metabolites. It almost goes without saying that mixing ISs of these metabolites to each sample would allow a much more precise determination of their absolute amounts.

To minimize such errors, we systematically investigated which of the salts in the medium could contribute to the loss of signal due to ion suppression or ion enhancement during ionization by ESI. The growth media used in this study consisted of Hepes, K-phosphates, Beijerinck salts, and trace salts (for details see Materials and Methods). A sequence of experiments was performed to unravel the effects of the individual salts from this growth media (**Figure 4**, for details see **Table A1** in Appendix). Hepes, K-phosphate, and the trace salt solution had only minor ion suppression effects (**Figure 4A**, for details see **Table A1** in Appendix). Hepes caused significant ion suppression of G6P but less than the Beijerinck salts. K-phosphates caused weak but significant ion induction of ADP, and FBP, RuBP, SBP, aconitate, and isocitrate. The trace salt solution caused overestimation of aspartate due to ion enhancement. However, most of the ion suppression observed due to the growth media in the sample could be attributed to the Beijerinck salts present in the media (**Figure 4A**). In a second experiment, ammonium chloride, one component of the Beijerinck salts, was shown to be responsible for the major part of the residual ion suppression whereas $MgSO_4$ and $CaCl_2$ had minor effects (**Figure 4B**). In a third experiment, the chloride anion was found to be the main reason why ammonium chloride causes ion suppression (**Figure 4C**). From the 24 metabolites routinely measured with this method, the signal of 10 was significantly suppressed in the presence of the Beijerinck salt (**Figure A1** in Appendix). With the exception of UDPG which was found to be suppressed by $MgSO_4$, the chloride anion in the growth media was found to be responsible for ion suppression. Thus, simply replacing the chloride anion with bicarbonate greatly decreased ion suppression (**Figure 4C**).

For subsequent measurements a growth medium with lower ammonium chloride concentration was used (**Figure 1**). Alternatively, a medium in which ammonium chloride is replaced with ammonium bicarbonate could be used or, as mentioned earlier ISs for each metabolite could be added to the sample to assure accurate metabolite measurements. This example shows that different degrees of ion suppression, both with respect of individual metabolites and with respect to different samples, can be generated by differences in growth conditions and culture components. These results imply that erroneous results will also be obtained if such changes occur as a result of growing algae in different conditions, or are generated in time as a result of the algae using nutrients.

The specific issue with growth medium components does not arise with higher plants. Nevertheless, this example serves as a warning that differential ion accumulation in plant tissues might affect ion suppression. In attempt to circumvent this problem the total ion chromatogram should be carefully checked in areas which appear to be strongly affected. If strong ion suppression is observed, both dilution and recovery tests in which standard compounds are added to the extracts should be performed (see Fernie and Keurentjes, 2011). More generally, such problems can be identified by routine checks that the signal is linear with the amount of applied extract and (where available) that authentic samples can be quantitatively recovered after addition to the extract.

ION SUPPRESSION EFFECTS CAUSED BY DIFFERENT TISSUE TYPES

Metabolite composition varies between plant species and also between different tissues of the same plant. It is therefore expected to observe different levels of ion suppression within these samples. To evaluate this, an IS mixture (sinigrin, isovitexin, and saponarin) was added to the same volume (ratio 1/1) of extracts from *Arabidopsis* leaves, tomato fruits, and rice leaves, respectively. This was performed with four different known concentrations of the IS mixture with three experimental replicates of the step of mixing solutions. As control, the 50% diluted original standard mixture was also analyzed without being mixed with plant extract. Peak areas for each IS were determined and are presented in **Figure 5**. For all IS compounds, the strongest ion suppression was observed when they were added to rice leaves, followed by *Arabidopsis* leaves, and the lowest was seen for tomato fruit extracts. As expected, the ion suppression in a mixture of *Arabidopsis* leaf and tomato fruit extracts (ratio 1/1) was intermediate to what was observed in the corresponding independent extracts.

Plant material used for metabolic determination is often a mixture of tissues. For example, plant seedling is a mixture of hypocotyl and root, or fruit samples are a mixture of pericarp, seed, and peel. Ion suppression caused by differences between tissue types, in comparative analysis between mutants, transgenic, time course, and stress treatment is relatively minor. But in case of comparison between different plant species (as shown above with *Arabidopsis* and rice leaves), wild accessions which have important phenotypical differences, or mutants with strong phenotypic differences, such problems due to differential ion suppression could easily arise. For example, seedling samples that differ in the relative amount of shoot/root or hypocotyl/root might be particularly susceptible to differential ion suppression. The same problem is raised in the case of harvesting flower samples with a varying ratio of flower/sepal and/or (flower/pedicel).

To evaluate this problem, recombination analyses (Fernie and Keurentjes, 2011) were evaluated using *Arabidopsis* samples containing various ratio of leaves and roots or flower extracts, focusing on the general secondary metabolites which were detected in both tissues. The percentage recovery was simply estimated for evaluation using theoretical concentration of extracts mixture, [(level in leaves × A%) + (level in roots (or flowers) × B%)]/100] (**Table 2**). Mixture of different tissue types results in >100% recovery (i.e., less ion suppression) for some peaks (e.g., IS), presumably because the area which is strongly suppressed differs between leaf and root extracts. Increase of recovery was observed in some peaks (<117%) at an equivalent mixture of leaf and root extracts. On the other hand, recovery ratio in mixture of leaves and flowers showed more significant variance (43 ∼ 122%). This experiment highlights the value of preliminary analyses in order to check for ion suppression. It is furthermore useful to understand the range of variance in instances in which the value of IS is unstable between samples.

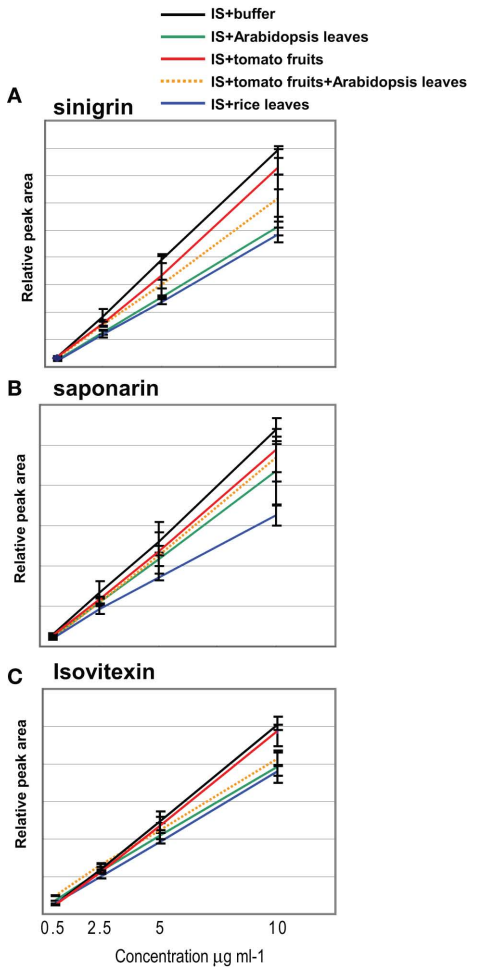


FIGURE 5 | Example of ion suppression caused by plant extracts. Standard solutions (20, 10, 5, 1 $\mu g\,ml^{-1}$) of three standard compounds, **(A)** sinigrin, **(B)** saponarin, and **(C)** isovitexin, were mixed to the same volume of plant extracts (0.2 mg FW μl^{-1}) of *Arabidopsis* leaves, tomato fruits, rice leaves and mixture of *Arabidopsis* leaves and tomato fruits (ratio 1/1). Average and ±SD were calculated from three experimental replicates.

CHEMICAL DIVERSITY AND PEAK ANNOTATION USING CROSS SPECIES COMPARISON

Many metabolite databases for LC–MS are available, such as MASSBANK (Horai et al., 2010), METLIN (Smith et al., 2005), MS2T (Matsuda et al., 2009), KNApSAcK (Shinbo et al., 2006), and Flavonoid Viewer (Arita and Suwa, 2008). These greatly aid in the prediction and annotation of detected peaks (Tohge and Fernie, 2010). That said, technical improvement of peak identification and annotation still represents a major hurdle for LC–MS-based

Table 2 | Recovery test with mixture of extract of leaves, roots, and flowers.

Tissue type								
	Leaves (%)	100	90	80	50	20	10	100
	Roots (%)	0	10	20	50	80	90	0
Compound	*m/z*				Recovery (%)			
Saponarin (IS)	593	100	103	101	100	100	104	100
Isovitexin (IS)	431	100	104	107	113	113	111	100
Kaempferol-3Glc2″Rha-7Rha	741	100	104	102	105	106	108	100
Kaempferol-3Glc-7Rha	595	100	103	100	99	100	103	100
Kaempferol-3Rha-7Rha	577	100	104	104	111	111	106	100
Quercetin-3Glc-7Rha	609	100	103	101	96	94	94	100
Sinapoyl glucoside	385	100	114	111	116	115	111	100
7-methylsulfinylheptyl glucosinolate	478	100	106	109	115	117	111	100
8-methylsulfinyloctyl GLS	492	100	101	97	100	102	100	100
	Leaves (%)	100	90	80	50	20	10	100
	Flowers (%)	0	10	20	50	80	90	0
Compound	*m/z*				Recovery (%)			
Saponarin (IS)	593	100	93	87	68	52	55	100
Isovitexin (IS)	431	100	109	111	122	116	116	100
Kaempferol-3Glc2″Rha-7Rha	741	100	98	92	77	64	70	100
Kaempferol-3Glc-7Rha	595	100	93	87	68	52	55	100
Kaempferol-3Rha-7Rha	577	100	100	99	95	100	108	100
Quercetin-3Glc-7Rha	609	100	98	88	66	46	43	100
Sinapoyl glucoside	385	100	105	108	109	111	113	100
7-methylsulfinylheptyl glucosinolate	478	100	100	95	90	91	102	100
8-methylsulfinyloctyl GLS	492	100	99	94	81	70	80	100

The peaks which were detected in leaves, roots, and flowers, were used for recovery test. The percentage recovery was estimated for evaluation using theoretical concentration of extracts mixture, [(level in leaves × A%) + (level in roots (or flowers) × B%)]/100], respectively. Three internal standard compounds, saponarin and isovitexin were used. Analysis was evaluated by three experimental replicates. (n = 3, SD of raw data <25.7%).

metabolite profiling platforms. The identification of secondary metabolites is obstructed by the insufficient availability of standard compounds. It is impossible to comprehensively purchase standard substances since the diversity of their chemical structure is far too large. Moreover, complex compounds are largely unavailable commercially and those that are available are often prohibitively expensive. Furthermore, LC–MS studies are complicated by the fact that the levels of secondary metabolites are highly divergent between different organs, growth conditions, and species (Petersen, 2007; Hanhineva et al., 2008; Matsuda et al., 2010). For these reasons, peak identification is generally performed by the use of combinatorial strategies whereby the literature information is taken alongside available compounds in an attempt to identify specific peaks (see for example Tohge et al., 2005; Giavalisco et al., 2009).

Given the recent explosion of genome information afforded firstly by microarray analyses and more recently by next-generation sequencing (review of (Schneeberger and Weigel, 2011), further tools for translational biology are becoming available. One such example, PlaNet, was described recently by Mutwil et al. (2011). Following this approach gene sequences can be connected between plant species on the basis of BLAST homology searches and then the positions in co-expression networks can be ascertained and finally it is possible to link unknown genes to annotated metabolic genes. As such this approach holds great promise both for gene functional annotation and via use of mutant plants in the annotation of unknown metabolites (Tohge and Fernie, 2010; Mutwil et al., 2011). It demonstrated the utility of this approach by identifying candidate genes of the general and species specific flavonoid pathways. It is likely that integrating metabolomics data on all the species currently in PlaNet will greatly aid this process and is certainly a research avenue that should be pursued in the near future.

CONCLUSION

Whilst applying a method established for another species is likely not to be overly problematic for screening purposes and for a first insight into the metabolome of an organism, the examples presented here demonstrate that when more precise information is required considerable effort should be put into establishing both the qualitative and quantitative reliability of any LC–MS-based metabolic profiling method. As evidenced by the ion suppression (and ion enhancement) examples particular care must be taken with this issue as well as in ensuring that the extraction procedure is appropriate for the tissue under study. Once these important controls have been adhered to a wide array of computational resources are available (Tohge and Fernie, 2009), which will greatly aid in translational research. Given that the trend

in plant science research is to move away from the single model species of *A. thaliana*, such tools will become increasingly important. However, it is prudent to note that uncritical use of such tools without adequate controls of the type demonstrated here may well result in inaccurate representations of the metabolome. The best way to approach a new tissue, species, or even a dramatic mutant/transgenic line is to adopt both experimental and computational approaches to ensure the highest possible data quality.

ACKNOWLEDGMENTS

We thank Dr. Yozo Okazaki in RIKEN PSC for useful discussion, and Dr. Mark-Aurel Schöttler at the Max-Planck-Institute of Molecular Plant Physiology (MPIMP) for kindly providing instrumentation for *Chlamydomonas* illumination and harvesting. We thank Prof. Dr. Martin Steup in University of Potsdam and Dr. Wagner L. Araújo at MPIMP for expert comments. This work is partially funded by the German Federal Ministry of Education and Research by the FORSYS BMBF grant (*Go*FORSYS) and the Max-Planck Society (MPG). Research activity of Takayuki Tohge was supported by the Alexander von Humboldt Foundation. Tabea Mettler was supported by the International Max-Planck Research School (IMPRS). Adam James Carroll was supported through a grant to the Australian Research Council Centre of Excellence in Plant Energy Biology.

REFERENCES

Achnine, L., Huhman, D. V., Farag, M. A., Sumner, L. W., Blount, J. W., and Dixon, R. A. (2005). Genomics-based selection and functional characterization of triterpene glycosyltransferases from the model legume *Medicago truncatula*. *Plant J.* 41, 875–887.

Aharoni, A., Keizer, L. C. P., Bouwmeester, H. J., Sun, Z. K., Alvarez-Huerta, M., Verhoeven, H. A., Blaas, J., van Houwelingen, A., De Vos, R. C. H., van der Voet, H., Jansen, R. C., Guis, M., Mol, J., Davis, R. W., Schena, M., van Tunen, A. J., and O'Connell, A. P. (2000). Identification of the SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays. *Plant Cell* 12, 647–661.

Aharoni, A., Ric de Vos, C. H., Verhoeven, H. A., Maliepaard, C. A., Kruppa, G., Bino, R. J., and Goodenowe, D. B. (2002). Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. *OMICS* 6, 217–234.

Albinsky, D., Kusano, M., Higuchi, M., Hayashi, N., Kobayashi, M., Fukushima, A., Mori, M., Ichikawa, T., Matsui, K., Kuroda, H., Horii, Y., Tsumoto, Y., Sakakibara, H., Hirochika, H., Matsui, M., and Saito, K. (2010). Metabolomic screening applied to rice FOX *Arabidopsis* lines leads to the identification of a gene-changing nitrogen metabolism. *Mol. Plant* 3, 125–142.

Allen, A. E., LaRoche, J., Maheswari, U., Lommer, M., Schauer, N., Lopez, P. J., Finazzi, G., Fernie, A. R., and Bowler, C. (2008). Whole-cell response of the pennate diatom *Phaeodactylum tricornutum* to iron starvation. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10438–10443.

Araujo, W. L., Ishizaki, K., Nunes-Nesi, A., Larson, T. R., Tohge, T., Krahnert, I., Witt, S., Obata, T., Schauer, N., Graham, I. A., Leaver, C. J., and Fernie, A. R. (2010). Identification of the 2-hydroxyglutarate and isovaleryl-CoA dehydrogenases as alternative electron donors linking lysine catabolism to the electron transport chain of *Arabidopsis* mitochondria. *Plant Cell* 22, 1549–1563.

Arita, M., and Suwa, K. (2008). Search extension transforms Wiki into a relational system: a case for flavonoid metabolite database. *BioData Min.* 7, 8.

Arrivault, S., Guenther, M., Ivakov, A., Feil, R., Vosloh, D., van Dongen, J. T., Sulpice, R., and Stitt, M. (2009). Use of reverse-phase liquid chromatography, linked to tandem mass spectrometry, to profile the Calvin cycle and other metabolic intermediates in *Arabidopsis* rosettes at different carbon dioxide concentrations. *Plant J.* 59, 824–839.

Bais, P., Moon, S. M., He, K., Leitao, R., Dreher, K., Walk, T., Sucaet, Y., Barkan, L., Wohlgemuth, G., Roth, M. R., Wurtele, E. S., Dixon, P., Fiehn, O., Lange, B. M., Shulaev, V., Sumner, L. W., Welti, R., Nikolau, B. J., Rhee, S. Y., and Dickerson, J. A. (2010). Plant metabolomics. org: a web portal for plant metabolomics experiments. *Plant Physiol.* 152, 1807–1816.

Barth, C., and Jander, G. (2006). *Arabidopsis* myrosinases TGG1 and TGG2 have redundant function in glucosinolate breakdown and insect defence. *Plant J.* 46, 549–562.

Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B. J., Mendes, P., Roessner-Tunali, U., Beale, M. H., Trethewey, R. N., Lange, B. M., Wurtele, E. S., and Sumner, L. W. (2004). Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.* 9, 418–425.

Bolling, C., and Fiehn, O. (2005). Metabolite profiling of *Chlamydomonas reinhardtii* under nutrient deprivation. *Plant Physiol.* 139, 1995–2005.

Boyle, N. R., and Morgan, J. A. (2009). Flux balance analysis of primary metabolism in *Chlamydomonas reinhardtii*. *BMC Syst. Biol.* 3, 4. doi:10.1186/1752-0509-3-4

Buhrman, D. L., Price, P. I., and Rudewicz, P. J. (1996). Quantitation of SR 27417 in human plasma using electrospray liquid chromatography tandem mass spectrometry: a study of ion suppression. *J. Am. Soc. Mass Spectrom.* 7, 1099–1105.

Carmo-Silva, A. E., Keys, A. J., Beale, M. H., Ward, J. L., Baker, J. M., Hawkins, N. D., Arrabaca, M. C., and Parry, M. A. J. (2009). Drought stress increases the production of 5-hydroxynorvaline in two C-4 grasses. *Phytochemistry* 70, 664–671.

Carroll, A. J., Badger, M. R., and Harvey Millar, A. (2010). The MetabolomeExpress Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. *BMC Bioinformatics* 11, 376. doi:10.1186/1471-2105-11-376

Catchpole, G. S., Beckmann, M., Enot, D. P., Mondhe, M., Zywicki, B., Taylor, J., Hardy, N., Smith, A., King, R. D., Kell, D. B., Fiehn, O., and Draper, J. (2005). Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proc. Natl. Acad. Sci. U.S.A.* 102, 14458–14462.

Cottret, L., Wildridge, D., Vinson, F., Barrett, M. P., Charles, H., Sagot, M. F., and Jourdan, F. (2010). MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Res.* 38, W132–W137.

El-Lithy, M. E., Rodrigues, G. C., van Rensen, J. J. S., Snel, J. F. H., Dassen, H., Koornneef, M., Jansen, M. A. K., Aarts, M. G. M., and Vreugdenhil, D. (2005). Altered photosynthetic performance of a natural *Arabidopsis* accession is associated with atrazine resistance. *J. Exp. Bot.* 56, 1625–1634.

Fernie, A. R., Aharoni, A., Willmitzer, L., Stitt, M., Tohge, T., Kopka, J., Carroll, A. J., Saito, K., Fraser, P. D., and DeLuca, V. (2011). Recommendations for reporting metabolite data. *Plant Cell* 23, 2477–2482.

Fernie, A. R., and Keurentjes, J. J. B. (2011). Genetics, genomics and metabolomics. *Annu. Plant Rev.* 43, 219–246.

Fernie, A. R., Trethewey, R. N., Krotzky, A. J., and Willmitzer, L. (2004). Innovation – metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.* 5, 763–769.

Fraser, P. D., Enfissi, E. M. A., Goodfellow, M., Eguchi, T., and Bramley, P. M. (2007). Metabolite profiling of plant carotenoids using the matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Plant J.* 49, 552–564.

Fukushima, A., Kusano, M., Nakamichi, N., Kobayashi, M., Hayashi, N., Sakakibara, H., Mizuno, T., and Saito, K. (2009). Impact of clock-associated *Arabidopsis* pseudo-response regulators in metabolic coordination. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7251–7256.

Geigenberger, P., Fernie, A. R., Gibon, Y., Christ, M., and Stitt, M. (2000). Metabolic activity decreases as an adaptive response to low internal oxygen in growing potato tubers. *Biol. Chem.* 381, 723–740.

Giavalisco, P., Kohl, K., Hummel, J., Seiwert, B., and Willmitzer, L. (2009). C-13 isotope-labeled metabolomes allowing for improved compound annotation and relative quantification in liquid chromatography-mass spectrometry-based metabolomic research. *Anal. Chem.* 81, 6546–6551.

Gibon, Y., Usadel, B., Blaesing, O. E., Kamlage, B., Hoehne, M., Trethewey, R., and Stitt, M. (2006). Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in *Arabidopsis* rosettes. *Genome Biol.* 7, R76.

Giroud, C., Gerber, A., and Eichenberger, W. (1988). Lipids of *Chlamydomonas reinhardtii* – analysis of molecular-species and intracellular site(s) of biosynthesis. *Plant Cell Physiol.* 29, 587–595.

Goossens, A., Hakkinen, S. T., Laakso, I., Seppanen-Laakso, T., Biondi, S., De Sutter, V., Lammertyn, F., Nuutila, A. M., Soderlund, H., Zabeau, M., Inze, D., and Oksman-Caldentey, K. M. (2003). A functional genomics approach toward the understanding of secondary metabolism in plant cells. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8595–8600.

Griffin, J. L., and Nicholls, A. W. (2006). Metabolomics as a functional genomic tool for understanding lipid dysfunction in diabetes, obesity and related disorders. *Pharmacogenomics* 7, 1095–1107.

Hall, R., Beale, M., Fiehn, O., Hardy, N., Sumner, L., and Bino, R. (2002). Plant metabolomics: the missing link in functional genomics strategies. *Plant Cell* 14, 1437–1440.

Hanhineva, K., Rogachev, I., Kokko, H., Mintz-Oron, S., Venger, I., Karenlampi, S., and Aharoni, A. (2008). Non-targeted analysis of spatial metabolite composition in strawberry (*Fragaria x ananassa*) flowers. *Phytochemistry* 69, 2463–2481.

Harmer, S. L., Hogenesch, L. B., Straume, M., Chang, H. S., Han, B., Zhu, T., Wang, X., Kreps, J. A., and Kay, S. A. (2000). Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock. *Science* 290, 2110–2113.

Harris, E. H. (1989). *The Chlamydomonas Sourcebook*. New York, NY: Academic Press.

Hirai, M. Y., Klein, M., Fujikawa, Y., Yano, M., Goodenowe, D. B., Yamazaki, Y., Kanaya, S., Nakamura, Y., Kitayama, M., Suzuki, H., Sakurai, N., Shibata, D., Tokuhisa, J., Reichelt, M., Gershenzon, J., Papenbrock, J., and Saito, K. (2005). Elucidation of gene-to-gene and metabolite-to-gene networks in *Arabidopsis* by integration of metabolomics and transcriptomics. *J. Biol. Chem.* 280, 25590–25595.

Hirai, M. Y., Sugiyama, K., Sawada, Y., Tohge, T., Obayashi, T., Suzuki, A., Araki, R., Sakurai, N., Suzuki, H., Aoki, K., Goda, H., Nishizawa, O. I., Shibata, D., and Saito, K. (2007). Omics-based identification of *Arabidopsis* Myb transcription factors regulating aliphatic glucosinolate biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6478–6483.

Hirai, M. Y., Yano, M., Goodenowe, D. B., Kanaya, S., Kimura, T., Awazuhara, M., Arita, M., Fujiwara, T., and Saito, K. (2004). Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 101, 10205–10210.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M. Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., and Nishioka, T. (2010). MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* 45, 703–714.

Iijima, Y., Nakamura, Y., Ogata, Y., Tanaka, K., Sakurai, N., Suda, K., Suzuki, T., Suzuki, H., Okazaki, K., Kitayama, M., Kanaya, S., Aoki, K., and Shibata, D. (2008). Metabolite annotations based on the integration of mass spectral information. *Plant J.* 54, 949–962.

Ikonomou, M. G., Blades, A. T., and Kebarle, P. (1990). Investigations of the electrospray interface for liquid-chromatography mass-spectrometry. *Anal. Chem.* 62, 957–967.

Jom, K. N., Frank, T., and Engel, K. H. (2010). A metabolite profiling approach to follow the sprouting process of mung beans (*Vigna radiata*). *Metabolomics* 7, 102–117.

Kanno, Y., Jikumaru, Y., Hanada, A., Nambara, E., Abrams, S. R., Kamiya, Y., and Seo, M. (2010). Comprehensive hormone profiling in developing *Arabidopsis* seeds: examination of the site of ABA biosynthesis, ABA transport and hormone interactions. *Plant Cell Physiol.* 51, 1988–2001.

Kebarle, P., and Tang, L. (1993). From ions in solution to ions in the gas-phase – the mechanism of electrospray mass-spectrometry. *Anal. Chem.* 65, A972–A986.

Kempa, S., Hummel, J., Schwemmer, T., Pietzke, M., Strehmel, N., Wienkoop, S., Kopka, J., and Weckwerth, W. (2009). An automated GCxGC-TOF-MS protocol for batch-wise extraction and alignment of mass isotopomer matrixes from differential C-13-labelling experiments: a case study for photoautotrophic-mixotrophic grown *Chlamydomonas reinhardtii* cells. *J. Basic Microbiol.* 49, 82–91.

Kerwin, R. E., Jimenez-Gomez, J. M., Fulop, D., Harmer, S. L., Maloof, J. N., and Kliebenstein, D. J. (2011). Network quantitative trait loci mapping of circadian clock outputs identifies metabolic pathway-to-clock linkages in *Arabidopsis*. *Plant Cell* 23, 471–485.

King, R., Bonfiglio, R., Fernandez-Metzler, C., Miller-Stein, C., and Olah, T. (2000). Mechanistic investigation of ionization suppression in electrospray ionization. *J. Am. Soc. Mass Spectrom.* 11, 942–950.

Krall, L., Huege, J., Catchpole, G., Steinhauser, D., and Willmitzer, L. (2009). Assessment of sampling strategies for gas chromatography-mass spectrometry (GC-MS) based metabolomics of cyanobacteria. *J. Chromatogr.* 877, 2952–2960.

Kusano, M., Tohge, T., Fukushima, A., Kobayashi, M., Hayashi, N., Otsuki, H., Kondou, Y., Goto, H., Kawashima, M., Matsuda, F., Niida, R., Matsui, M., Saito, K., and Fernie, A. R. (2011). Metabolomics reveals comprehensive reprogramming involving two independent metabolic responses of *Arabidopsis* to ultraviolet-B light. *Plant J.* 67, 354–369.

Luo, J., Nishiyama, Y., Fuell, C., Taguchi, G., Elliott, K., Hill, L., Tanaka, Y., Kitayama, M., Yamazaki, M., Bailey, P., Parr, A., Michael, A. J., Saito, K., and Martin, C. (2007). Convergent evolution in the BAHD family of acyltransferases: identification and characterization of anthocyanin acyltransferases from *Arabidopsis thaliana*. *Plant J.* 50, 678–695.

Matsuda, F., Hirai, M. Y., Sasaki, E., Akiyama, K., Yonekura-Sakakibara, K., Provart, N. J., Sakurai, T., Shimada, Y., and Saito, K. (2010). AtMetExpress development: a phytochemical atlas of *Arabidopsis* development. *Plant Physiol.* 152, 566–578.

Matsuda, F., Yonekura-Sakakibara, K., Niida, R., Kuromori, T., Shinozaki, K., and Saito, K. (2009). MS/MS spectral tag-based annotation of non-targeted profile of plant secondary metabolites. *Plant J.* 57, 555–577.

Matuszewski, B. K., Constanzer, M. L., and Chavez-Eng, C. M. (1998). Matrix effect in quantitative LC/MS/MS analyses of biological fluids: a method for determination of finasteride in human plasma at picogram per milliliter concentrations. *Anal. Chem.* 70, 882–889.

Matuszewski, B. K., Constanzer, M. L., and Chavez-Eng, C. M. (2003). Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC-MS/MS. *Anal. Chem.* 75, 3019–3030.

May, P., Wienkoop, S., Kempa, S., Usadel, B., Christian, N., Rupprecht, J., Weiss, J., Recuenco-Munoz, L., Ebenhoh, O., Weckwerth, W., and Walther, D. (2008). Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of *Chlamydomonas reinhardtii*. *Genetics* 179, 157–166.

Meyer, R. C., Steinfath, M., Lisec, J., Becher, M., Witucka-Wall, H., Torjek, O., Fiehn, O., Eckardt, A., Willmitzer, L., Selbig, J., and Altmann, T. (2007). The metabolic signature related to high plant growth rate in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 104, 4759–4764.

Moco, S., Bino, R. J., Vorst, O., Verhoeven, H. A., de Groot, J., van Beek, T. A., Vervoort, J., and de Vos, C. H. R. (2006). A liquid chromatography-mass spectrometry-based metabolome database for tomato. *Plant Physiol.* 141, 1205–1218.

Mutwil, M., Klie, S., Tohge, T., Giorgi, F. M., Wilkins, O., Campbell, M. M., Fernie, A. R., Usadel, B., Nikoloski, Z., and Persson, S. (2011). PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* 23, 895–910.

Nakabayashi, R., Kusano, M., Kobayashi, M., Tohge, T., Yonekura-Sakakibara, K., Kogure, N., Yamazaki, M., Kitajima, M., Saito, K., and Takayama, H. (2009). Metabolomics-oriented isolation and structure elucidation of 37 compounds including two anthocyanins from *Arabidopsis thaliana*. *Phytochemistry* 70, 1017–1029.

Nicholson, J. K., and Wilson, I. D. (2003). Understanding "global" systems biology: metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.* 2, 668–676.

Okazaki, Y., Shimojima, M., Sawada, Y., Toyooka, K., Narisawa, T., Mochida, K., Tanaka, H., Matsuda, F., Hirai, A., Hirai, M. Y., Ohta, H., and Saito, K. (2009). A chloroplastic UDP-glucose pyrophosphorylase from *Arabidopsis* is the committed enzyme for the first step of sulfolipid biosynthesis. *Plant Cell* 21, 892–909.

Petersen, M. (2007). Current status of metabolic phytochemistry. *Phytochemistry* 68, 2847–2860.

Renberg, L., Johansson, A. I., Shutova, T., Stenlund, H., Aksmann, A., Raven, J. A., Gardestrom, P., Moritz, T., and Samuelsson, G. (2010). A metabolomic approach to study major metabolite changes during acclimation to limiting CO2 in *Chlamydomonas reinhardtii*. *Plant Physiol.* 154, 187–196.

Rischer, H., Oresic, M., Seppanen-Laakso, T., Katajamaa, M., Lammertyn, F., Ardiles-Diaz, W., Van Montagu, M. C. E., Inze, D., Oksman-Caldentey, K. M., and Goossens, A. (2006). Gene-to-metabolite networks for terpenoid indole alkaloid biosynthesis in *Catharanthus roseus* cells. *Proc. Natl. Acad. Sci. U.S.A.* 103, 5614–5619.

Roessner, U., Willmitzer, L., and Fernie, A. R. (2001). High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.* 127, 749–764.

Roessner-Tunali, U., Hegemann, B., Lytovchenko, A., Carrari, F., Bruedigam, C., Granot, D., and Fernie, A. R. (2003). Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. *Plant Physiol.* 133, 84–99.

Sauter, H., Lauer, M., and Fritsch, H. (1988). Metabolite profiling of plants – a new diagnostic technique. *Abstr. Pap. Am. Chem. Soc.* 195, 129.

Schauer, N., Semel, Y., Roessner, U., Gur, A., Balbo, I., Carrari, F., Pleban, T., Perez-Melis, A., Bruedigam, C., Kopka, J., Willmitzer, L., Zamir, D., and Fernie, A. R. (2006). Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nat. Biotechnol.* 24, 447–454.

Schauer, N., Zamir, D., and Fernie, A. R. (2005). Metabolic profiling of leaves and fruit of wild species tomato: a survey of the *Solanum lycopersicum* complex. *J. Exp. Bot.* 56, 297–307.

Scherling, C., Ulrich, K., Ewald, D., and Weckwerth, W. (2009). A metabolic signature of the beneficial interaction of the endophyte paenibacillus sp isolate and in vitro-grown poplar plants revealed by metabolomics. *Mol. Plant Microbe Interact.* 22, 1032–1037.

Schneeberger, K., and Weigel, D. (2011). Fast-forward genetics enabled by new sequencing technologies. *Trends Plant Sci.* 16, 282–288.

Semel, Y., Schauer, N., Roessner, U., Zamir, D., and Fernie, A. (2007). Metabolite analysis for the comparison of irrigated and non-irrigated field grown tomato of varying genotype. *Metabolomics* 3, 289–295.

Shinbo, Y., Nakamura, Y., Altaf-Ul-Amin, M., Asahi, H., Kurokawa, K., Arita, M., Saito, K., Ohta, D., Shibata, D., and Kanaya, S. (2006). KNApSAcK: a comprehensive species-metabolite relationship database. *Plant Metabolomics* 165–181.

Smith, C. A., O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R., and Siuzdak, G. (2005). METLIN – a metabolite mass spectral database. *Ther. Drug Monit.* 27, 747–751.

Stitt, M., and Fernie, A. R. (2003). From measurements of metabolites to metabolomics: an "on the fly" perspective illustrated by recent studies of carbon-nitrogen interactions. *Curr. Opin. Biotechnol.* 14, 136–144.

St-Pierre, B., and De Luca, V. (2000). "Evolution of acyltransferase genes: origin and diversification of the BAHD superfamily of acyltransferases involved in secondary metabolism," in *Evolution of Metabolic Pathways,* ed J. T. Romeo (Amsterdam: Elsevier Science), 285–315.

Stracke, R., Ishihara, H., Barsch, G. H. A., Mehrtens, F., Niehaus, K., and Weisshaar, B. (2007). Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the *Arabidopsis thaliana* seedling. *Plant J.* 50, 660–677.

Sulpice, R., Pyl, E. T., Ishihara, H., Trenkamp, S., Steinfath, M., Witucka-Wall, H., Gibon, Y., Usadel, B., Poree, F., Piques, M. C., Von Korff, M., Steinhauser, M. C., Keurentjes, J. J. B., Guenther, M., Hoehne, M., Selbig, J., Fernie, A. R., Altmann, T., and Stitt, M. (2009). Starch as a major integrator in the regulation of plant growth. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10348–10353.

Tierens, K., Thomma, B. P. H., Brouwer, M., Schmidt, J., Kistner, K., Porzel, A., Mauch-Mani, B., Cammue, B. P. A., and Broekaert, W. F. (2001). Study of the role of antimicrobial glucosinolate-derived isothiocyanates in resistance of *Arabidopsis* to microbial pathogens. *Plant Physiol.* 125, 1688–1699.

Tiessen, A., Hendriks, J. H. M., Stitt, M., Branscheid, A., Gibon, Y., Farre, E. M., and Geigenberger, P. (2002). Starch synthesis in potato tubers is regulated by post-translational redox modification of ADP-glucose pyrophosphorylase: a novel regulatory mechanism linking starch synthesis to the sucrose supply. *Plant Cell* 14, 2191–2213.

Tohge, T., and Fernie, A. R. (2009). Web-based resources for mass-spectrometry-based metabolomics: a user's guide. *Phytochemistry* 70, 450–456.

Tohge, T., and Fernie, A. R. (2010). Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function. *Nat. Protoc.* 5, 1210–1227.

Tohge, T., Nishiyama, Y., Hirai, M. Y., Yano, M., Nakajima, J., Awazuhara, M., Inoue, E., Takahashi, H., Goodenowe, D. B., Kitayama, M., Noji, M., Yamazaki, M., and Saito, K. (2005). Functional genomics by integrated analysis of metabolome and transcriptome of *Arabidopsis* plants over-expressing an MYB transcription factor. *Plant J.* 42, 218–235.

Tohge, T., Yonekura-Sakakibara, K., Niida, R., Watanabe-Takahashi, A., and Saito, K. (2007). Phytochemical genomics in *Arabidopsis thaliana*: a case study for functional identification of flavonoid biosynthesis genes. *Pure Appl. Chem.* 79, 811–823.

Trenkamp, S., Eckes, P., Busch, M., and Fernie, A. R. (2009). Temporally resolved GC-MS-based metabolic profiling of herbicide treated plants treated reveals that changes in polar primary metabolites alone can distinguish herbicides of differing mode of action. *Metabolomics* 5, 277–291.

von Roepenack-Lahaye, E., Degenkolb, T., Zerjeski, M., Franz, M., Roth, U., Wessjohann, L., Schmidt, J., Scheel, D., and Clemens, S. (2004). Profiling of *Arabidopsis* secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiol.* 134, 548–559.

Weiner, H., Stitt, M., and Heldt, H. W. (1987). Subcellular compartmentation of pyrophosphate and alkaline pyrophosphatase in leaves. *Biochim. Biophys. Acta* 893, 13–21.

Widodo Patterson, J. H., Newbigin, E., Tester, M., Bacic, A., and Roessner, U. (2009). Metabolic responses to salt stress of barley (*Hordeum vulgare* L.) cultivars, Sahara and Clipper, which differ in salinity tolerance. *J. Exp. Bot.* 60, 4089–4103.

Xia, J., and Wishart, D. S. (2010). MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.* 38, W71–W77.

Yamazaki, M., Shibata, M., Nishiyama, Y., Springob, K., Kitayama, M., Shimada, N., Aoki, T., Ayabe, S. I., and Saito, K. (2008). Differential gene expression profiles of red and green forms of *Perilla frutescens* leading to comprehensive identification of anthocyanin biosynthetic genes. *FEBS J.* 275, 3494–3502.

Yonekura-Sakakibara, K., Tohge, T., Matsuda, F., Nakabayashi, R., Takayama, H., Niida, R., Watanabe-Takahashi, A., Inoue, E., and Saito, K. (2008). Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene-metabolite correlations in *Arabidopsis*. *Plant Cell* 20, 2160–2176.

Yonekura-Sakakibara, K., Tohge, T., Niida, R., and Saito, K. (2007). Identification of a flavonol 7-O-rhamnosyltransferase gene determining flavonoid pattern in *Arabidopsis* by transcriptome coexpression analysis and reverse genetics. *J. Biol. Chem.* 282, 14932–14941.

Zhu, T., and Wang, X. (2000). Large-scale profiling of the *Arabidopsis* transcriptome. *Plant Physiol.* 124, 1472–1476.

Conflict of Interest Statement: 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.

Received: 04 May 2011; accepted: 13 September 2011; published online: 03 October 2011.

*Citation: Tohge T, Mettler T, Arrivault S, Carroll AJ, Stitt M and Fernie AR (2011) From models to crop species: caveats and solutions for translational metabolomics. Front. Plant Sci. **2**:61. doi: 10.3389/fpls.2011.00061*

This article was submitted to Frontiers in Plant Physiology, a specialty of Frontiers in Plant Science.

APPENDIX

Table A1 | Ion suppression of growth media mainly caused by ammonium chloride.

	Components of growth media					Beijerinck salts			Kation and anion of NH_4Cl	
	Growth media	HEPES	K-phosphates	Trace salts	Beijerinck salts	$CaCl_2$	$MgSO_4$	NH_4Cl	NH_4HCO_3	NaCl
2-OG	−19.2**	−4.4	−12.6*	−3.8	−19.0*	−2.8	−10.3	−9.6*	−6.8	−10.9*
Aconitate	−2.7	5.4	13.2*	0.9	2.4	2.9	1.3	1.2	11.4	5.4
ADP	9.1	5.7	24.2*	3.0	9.5	7.3	9.7	8.2	19.6	12.4
ADPG	3.7	−0.7	3.8	0.0	6.1	1.3	3.0	1.2	2.5	1.3
AMP	−6.6	−3.2	−7.1	−3.6	−5.6	−0.6	4.5	2.5	2.6	0.0
Aspartate	69.5**	−3.5	−1.8	88.5**	7.5	−2.0	−0.7	2.6	−8.9	6.1
Citrate	−8.0	−2.5	−2.7	1.9	−5.2	0.5	−1.8	−6.9	−9.7	−6.9
DHAP	−37.8*	−5.1	−5.9	12.1	−36.5**	−11.2*	0.2	−23.5**	−3.3	−24.9**
F6P	−35.1*	−4.9	−4.1	−4.4	−23.9**	−6.4	5.4	−24.0**	−8.4*	−25.4**
FBP	0.5	6.8	21.8*	8.9	6.0	6.2	7.3	5.4	12.5	6.5
G1P	−30.8*	−3.4	−3.4	12.4	−15.1	−8.8	0.8	−4.8	−4.0	0.2
G6P	−59.7**	−17.0*	−5.6	10.0	−59.6**	−10.4*	1.3	−60.7**	−4.9	−61.0**
Glutamate	6.4	−2.5	−1.8	6.5*	11.6**	−1.9	1.7	7.8	−6.9	8.3
Glycerate	−20.6	−1.6	−10.0	4.2	−21.2*	−5.8	−7.8	−22.9**	−16.2	−20.1**
Isocitrate	6.8**	1.0	6.8**	3.5	−6.0	−0.7	−4.0	−2.1	3.9	−4.6
Malate	−3.3	−2.0	0.9	1.7	−7.5	−6.8	−5.2	−3.6	−1.8	−3.1
NAD	−27.9**	−2.3	−3.6	−7.1	−24.5**	−4.5	2.7	−10.8**	2.2	−12.9**
NADP	9.5	3.9	16.7	8.3	6.0	3.6	3.5	6.7	14.6*	7.2
R5P	−44.2**	−6.8	−5.7	9.6	−41.9**	−17.7*	0.9	−36.1**	−5.1	−35.7**
RuBP	9.9*	6.8	18.8*	7.8	10.7	3.1	9.1	6.1	13.3	8.5
S7P	−27.5*	−3.4	−4.6	7.9	−17.8**	−11.4**	1.1	−14.8*	−1.0	−14.9*
SBP	−5.8	4.6	16.2*	3.1	5.6	4.5	8.2	6.7	14.3	10.4
UDPG	−12.7**	−5.1	−3.1	−1.9	−18.5*	1.5	−18.2**	−2.8	1.8	−1.2
X5P/Ru5P	−26.1	6.9	13.8	27.2	−23.5**	0.6	9.6	−24.7**	−3.3	−18.5**

*Raw data of **Figure 4**. Values are expressed as % ion suppression ($x < 0$) or % ion enhancement ($x > 0$) respectively, according to the formula by Buhrman et al. (1996): $x = 100$ *(observed concentration-expected concentration)/ expected concentration Blue, significant ion suppression; red, significant ion enhancement; two asterisks: Student's t-test: $p < 0.01$; one asterisk, Student's t-test $p < 0.05$ ($n = 3$). Abbreviations of metabolites according to **Figure 1**.*

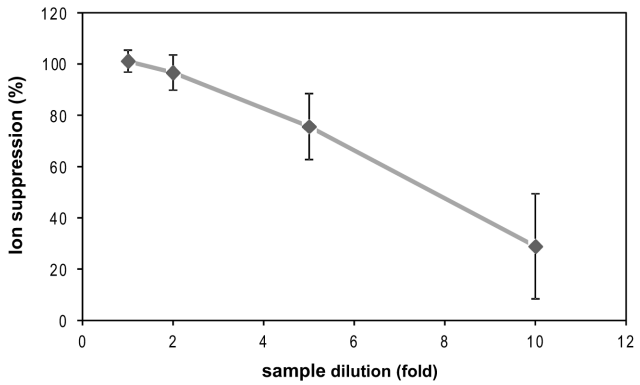


FIGURE A1 | Sample dilution led to decreased ion suppression. Illuminated *Chlamydomonas reinhardtii* CC-1690 suspension was quenched, harvested, and extracted as in Figure 1. Four different dilutions of the extract (diluted with water) were then spiked with a reference standard mix and compared to the standard mix on its own to assess the extent of ion suppression. The fivefold dilution was selected for experiments shown in **Figures 1** and **4**; **Table 1** as for several metabolites the 10-fold dilution was below detection limit. Note that the average of ion suppression was 77% for the fivefold dilution in this experiment. For experiment shown in **Figures 1** and **4**; **Table 1**, the average ion suppression for the fivefold dilution was <13%. This improvement was vey likely achieved by using a quadrupole (Finning TSQ Quantum Ultra) with a larger ion transfer tube diameter. The plot shows the mean ion suppression for 24 metabolites, each measured with two technical replicates, ±SD.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 22 August 2011
doi: 10.3389/fpls.2011.00039

Common motifs in the response of cereal primary metabolism to fungal pathogens are not based on similar transcriptional reprogramming

Lars Matthias Voll[1]*, Robin Jonathan Horst[1†], Anna-Maria Voitsik[1], Doreen Zajic[1], Birgit Samans[2,3], Jörn Pons-Kühnemann[2,3], Gunther Doehlemann[4], Steffen Münch[5], Ramon Wahl[6], Alexandra Molitor[3,7†], Jörg Hofmann[1], Alfred Schmiedl[1], Frank Waller[3,7†], Holger Bruno Deising[5], Regine Kahmann[4], Jörg Kämper[6], Karl-Heinz Kogel[3,7] and Uwe Sonnewald[1]

[1] *Division of Biochemistry, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany*
[2] *Institute of Biometry and Population Genetics, Justus Liebig University, Giessen, Germany*
[3] *Research Center for BioSystems, Land Use and Nutrition, Justus Liebig University, Giessen, Germany*
[4] *Max Planck Institute for Terrestrial Microbiology, Marburg, Germany*
[5] *Faculty of Agricultural and Nutritional Sciences, Phytopathology and Plant Protection, Martin-Luther-University Halle-Wittenberg, Halle, Germany*
[6] *Department of Genetics, Institute of Applied Biosciences, University of Karlsruhe, Karlsruhe, Germany*
[7] *Institute of Phytopathology and Applied Zoology, Justus Liebig University, Giessen, Germany*

Edited by:
Alisdair Fernie, Max Planck Institute for Plant Physiology, Germany

Reviewed by:
Andreas P. M. Weber, University of Duesseldorf, Germany
Veronica Graciela Maurino, Heinrich-Heine-Universität Düsseldorf, Germany

****Correspondence:***
Lars Matthias Voll, Division of Biochemistry, Friedrich-Alexander-University Erlangen-Nuremberg, Staudtstrasse 5, D-91058 Erlangen, Germany.
e-mail: lvoll@biologie.uni-erlangen.de

†Present address:
Robin Jonathan Horst, Department of Biology, University of Washington, Seattle, WA, USA;
Alexandra Molitor, KWS Saat AG, Einbeck, Germany;
Frank Waller, Julius-von-Sachs Institute, Julius-Maximilians-Universität Würzburg, Würzburg, Germany.

During compatible interactions with their host plants, biotrophic plant–pathogens subvert host metabolism to ensure the sustained provision of nutrient assimilates by the colonized host cells. To investigate, whether common motifs can be revealed in the response of primary carbon and nitrogen metabolism toward colonization with biotrophic fungi in cereal leaves, we have conducted a combined metabolome and transcriptome study of three quite divergent pathosystems, the barley powdery mildew fungus (*Blumeria graminis* f.sp. *hordei*), the corn smut fungus *Ustilago maydis*, and the maize anthracnose fungus *Colletotrichum graminicola*, the latter being a hemibiotroph that only exhibits an initial biotrophic phase during its establishment. Based on the analysis of 42 water-soluble metabolites, we were able to separate early biotrophic from late biotrophic interactions by hierarchical cluster analysis and principal component analysis, irrespective of the plant host. Interestingly, the corresponding transcriptome dataset could not discriminate between these stages of biotrophy, irrespective, of whether transcript data for genes of central metabolism or the entire transcriptome dataset was used. Strong differences in the transcriptional regulation of photosynthesis, glycolysis, the TCA cycle, lipid biosynthesis, and cell wall metabolism were observed between the pathosystems. However, increased contents of Gln, Asn, and glucose as well as diminished contents of PEP and 3-PGA were common to early post-penetration stages of all interactions. On the transcriptional level, genes of the TCA cycle, nucleotide energy metabolism and amino acid biosynthesis exhibited consistent trends among the compared biotrophic interactions, identifying the requirement for metabolic energy and the rearrangement of amino acid pools as common transcriptional motifs during early biotrophy. Both metabolome and transcript data were employed to generate models of leaf primary metabolism during early biotrophy for the three investigated interactions.

Keywords: maize, barley, *Ustilago maydis*, *Blumeria graminis* f.sp. *hordei*, *Colletotrichum graminicola*, compatible interaction, metabolite analysis, transcriptome analysis

INTRODUCTION

Substantial effort is being devoted to gain insight into plant–pathogen interactions to improve crop plants for sustainable agriculture. Phytopathogenic bacteria and fungi drive their own cellular metabolism with substrates being diverted from the colonized and/or surrounding host cells. Nutrient acquisition from the host cells is crucial for the successful establishment of bacterial and fungal pathogens (reviewed by Divon and Fluhr, 2007). Plant–pathogens have evolved different strategies to divert nutrients from their plant hosts. While necrotrophic pathogens rapidly kill plant tissue usually by the secretion of highly efficient toxins and cell wall degrading enzymes (van Kan, 2006) – and thrive on the dead plant material, biotrophic pathogens strictly rely on living tissue to survive and complete their life cycle (Divon and Fluhr, 2007). In contrast, hemibiotrophs establish themselves during an initial biotrophic phase before necrotrophic growth is initiated (Mendgen and Hahn, 2002; Münch et al., 2008). In general, infection sites of biotrophic fungi represent strong local metabolic sinks that drain nutrients from the host environment. Evidence obtained for the rust fungus *Uromyces fabae* suggest that nutrients are mainly taken up as hexoses (generated by secreted fungal invertase) and amino acids (Hahn et al., 1997; Voegele et al., 2001; Struck et al., 2002, 2004). Recently,

a novel high-affinity *U. maydis* sucrose transporter Srt1 has been characterized, which is required for full virulence (Wahl et al., 2010). Effective nutrient provision by host cells is necessary to establish a compatible interaction with biotrophs, as indicated by increased resistance of the variegated barley *albostrians* mutant toward powdery mildew fungus or by increased resistance of *Arabidopsis* overexpressing invertase inhibitors toward clubroot disease (Jain et al., 2004; Siemens et al., 2011). In addition, it was recently found that the induction of sugar efflux carriers in infected tissue by TAL-effectors of the bacterial rice pathogen *Xanthomonas oryzae pv. oryzae* (*Xoo*) is required for pathogenicity (Chen et al., 2010).

Vice versa, a vast array of fungal genes coding for metabolic enzymes was found to be induced upon host colonization, providing evidence that pathogen metabolism adapts to the host environment and nutrient availability (as reviewed by Divon and Fluhr, 2007). Despite its importance for hexose provision to the invaders, the induction of invertases, and the concomitant increase in free hexoses can serve as a signal for the repression of photosynthetic gene expression (as reviewed in Biemelt and Sonnewald, 2006). Furthermore, elevated hexose contents constitute an important cue in defense signaling (as reviewed by Bolton, 2009). Similarly, the support of the host defense response by the provision of reducing equivalents in the cytosol via glucose-6-phosphate dehydrogenase (G6PDH) seems to be an essential metabolic process that heightens defense effectiveness (Scharte et al., 2009). In *Arabidopsis*, strong evidence has been gathered that lipid metabolism in the chloroplast is involved in regulating the balance between SA- and JA-mediated defense responses and the induction of the hypersensitive response, HR (Kachroo et al., 2003; Chanda et al., 2008; Chaturvedi et al., 2008; Raffaele et al., 2008).

Although metabolic processes are important determinants of compatibility during plant–pathogen interactions, our knowledge on metabolic compatibility factors is scarce. Nevertheless, an increase in the sucrose/hexose ratio (Chou et al., 2000; Swarbrick et al., 2006) and elevated contents of nitrogen storage amino acids Gln and Asn (Olea et al., 2004; Tavernier et al., 2007; Horst et al., 2010a) have frequently been observed during biotrophic interactions, nourishing the hypothesis that a direct or indirect metabolic reprogramming of host metabolism occurs during the establishment of fungal biotrophs on their hosts. Employing comparative metabolome analysis, our study aims at identifying metabolic processes that are commonly altered during compatible interactions of biotrophic fungal leaf pathogens with agriculturally relevant cereal hosts. Pathosystems were selected to maximize biological diversity in the analyzed interactions and to minimize the chance of identifying effects specific to certain subclasses of pathogens. First, we have chosen to compare the response of barley, a C_3-plant, with that of maize, a C_4 plant, and second, the biotrophic lifestyle of the three fungal pathogens is quite diverse.

Ustilago maydis (*Um*), the causal agent of corn smut disease, is a biotrophic basidiomycete parasitizing maize and its natural ancestor teosinte. It can induce the formation of tumors on all aerial organs (Banuett, 1995) and exhibits a dimorphic lifestyle (Kahmann and Kämper, 2004): While haploid sporidia are not infectious and grow saprophytically in a yeast-like manner, filamentous growth is initiated upon mating of two compatible sporidia on the plant surface. Filamentous hyphae quickly form appressoria that penetrate host cells. Immediately upon host entry at around 24 h post inoculation, the invading biotrophic hyphae grow both inter- and intra-cellular without disrupting the host plasma membrane. About 4 days after penetration, the formation of hypertrophic host cells and concomitant tumor development are induced, while the fungal hyphae start proliferating in the apoplastic spaces that develop as a consequence of cell wall degradation and induced host cell enlargement (Doehlemann et al., 2008a,b).

Blumeria graminis f.sp. *hordei* (*Bgh*) is an obligate biotroph that causes powdery mildew disease on barley. Germination of wind-dispersed *Bgh* conidia on the barley leaf surface first produces a short primary germ tube prior to the formation of the infectious secondary germ tube, at the tip of which a hooked appressorium is formed. From the appressorium, a penetration peg is ejected within 15 h post inoculation (Hückelhoven et al., 1999; Both et al., 2005) that penetrates cuticle and wall of the host epidermis cell beneath and subsequently, a haustorium is established in the periplasmic space of the colonized host cell that serves as a strongly invaginated feeding organ. Unlike *U. maydis* hyphae that grow filamentously through the colonized maize tissue, only the haustoria of *Bgh* reside inside the infected leaf, while the predominant portion of fungal hyphae are growing epiphytically, occasionally forming secondary haustoria in adjacent epidermal cells. Eventually at 5 days post inoculation, conidiophores emerge from the epiphytic mycelium that shed series of conidiospores from their tips.

In contrast to *U. maydis* and *Bgh*, the maize pathogen *Colletotrichum graminicola* leads a hemibiotrophic lifestyle (as reviewed by Bergstrom and Nicholson, 1999; Mendgen and Hahn, 2002; Münch et al., 2008). Rain-dispersed conidia land on the leaf surface, produce germ tubes, which then differentiate sophisticated appressoria. During maturation, appressoria form rigid cell walls which melanize and synthesis of high concentrations of compatible solutes results in generation of enormous appressorial turgor pressure by diffusion of water into the appressorium. At the appressorial base, turgor pressure is translated into mechanical force that breaches the host cell wall. In the penetrated host epidermis cells, *C. graminicola* establishes itself as a biotroph within 36 h post inoculation by forming an infection vesicle that produces lobed biotrophic primary hyphae. During the subsequent colonization of neighboring cells at around 72 h post infection, the formation of narrow-bore secondary hyphae is initiated, which grow rapidly, are highly destructive and represent the necrotrophic lifestyle of the pathogen.

Thus, our set of fungal pathogens extends (i) an obligate biotroph that nourishes via epidermis-localized haustoria, *Bgh*, (ii) a biotroph that colonizes the entire leaf tissue by intra- and intercellularly growing hyphae, *Um*, and (iii) a hemibiotroph, *Cg*, that switches from biotrophic colonization of epidermis cells to vast proliferation by necrotrophic hyphae throughout the entire leaf.

MATERIALS AND METHODS

PLANT AND FUNGAL CULTIVATION AND INFECTION CONDITIONS

For combined metabolite and transcript profiling experiments, *Zea mays* cv. Early Golden Bantam was cultivated as described in (Doehlemann et al., 2008a) and infected with *U. maydis* strain SG200 as described by Doehlemann et al. (2008a) or with *C. graminicola* strain CgM2 as described in Münch et al. (2011).

Combined metabolite and transcript profiling experiments with barley (*cv. Golden Promise*) after challenge with *Bgh* isolate B6 were conducted as described in Molitor et al. (2011).

TRANSCRIPTOME ANALYSIS BY DNA MICROARRAY

Transcriptome data from *U. maydis*-infected maize leaf tissue was obtained from the same set of material described in Doehlemann et al. (2008a), which is deposited in the Gene Expression Omnibus[1] under the accession number GSE10023. The transcriptome dataset of *Bgh* infected barley leaves represents the same dataset as in Molitor et al. (2011). Transcriptome data for *C. graminicola* infected maize leaves (infection procedure as in Münch et al., 2011) were obtained as described in Doehlemann et al. (2008a) and are deposited in the Gene Expression Omnibus (see text footnote 1) under the accession number GSE31188. If not stated otherwise, a low stringent threshold of >1.5-fold change with no *p*-value filter was used for comparative analyses of transcriptome data.

MATCHING OF BARLEY AND MAIZE MICROARRAY DATA

To connect the transcripts from different microarray platforms, we used the microarray platform translator on the PlexDB homepage[2]. The transformation was performed with the default settings.

CALCULATION OF MAPMAN BIN ENRICHMENT AND SUBSEQUENT HCA

The tool MapMan (Thimm et al., 2004) adapted for maize and barley Affymetrix microarrays was used to visualize the transcriptome data that was obtained as described above. For the analysis, the mean values from all three replicate experiments were employed. To calculate the percentage of regulated genes per MapMan BIN of primary carbon and nitrogen metabolism, the number of regulated features with fold change >2.0 was expressed as percentage of total number of features in the respective BINs, to enable a comparison of maize and barley data that do not share the same number of accessions per BIN. Percentage up-regulated and percentage down-regulated features were scored separately and used for hierarchical cluster analysis (HCA) analysis after log transformation, median centering, and normalization as described for metabolite data below.

METABOLITE QUANTIFICATION AND ANALYSIS

For all three interactions analyzed in this report, metabolite contents were determined in three independent experiments from subsets of the leaf material pools that were employed for transcriptome analysis, such that material of four independent samples for metabolite analysis were pooled to generate one sample pool for transcript analysis per time point. All metabolite assays were conducted as described by Horst et al. (2010a).

MULTIVARIATE DATA ANALYSIS

Mean values of the four biological replicates taken per time point and experiment were calculated for all individual metabolites prior to calculating the metabolite ratio between infected vs. non-infected tissue, which was employed for HCA. After log transformation of the data, median centered ratios were normalized and HCA was performed using the complete linkage algorithm of the program Cluster V2.11 (Eisen et al., 1998) and the results were visualized using Maple Tree[3].

[1]http://www.ncbi.nlm.nih.gov/geo/

[2]http://www.plexdb.org/modules/MPT/mpt_help.php#overview

[3]http://mapletree.sourceforge.net/

Principle component analysis of log-transformed metabolite ratios was performed with the MarkerView software (Version 1.1.0.7, Applied Biosystems, Foster City, CA, USA) using the autoscale algorithm for scaling.

RESULTS

EXPERIMENTAL DESIGN AND SAMPLING STRATEGY

The sampling time points in all three pathosystems were carefully adjusted to the *on planta* development of the respective pathogen and to the diurnal light/dark cycles of the growth regimen (**Figure 1**). For every interaction, infected leaves were harvested at two crucial stages: (i) shortly after the establishment of biotrophy and (ii) at time points late in the biotrophic interaction, with a corresponding sampling time point during necrotrophic colonization by *C. graminicola* at 96hpi serving as a reference for non-biotrophic colonization. To minimize artifacts by diurnal oscillations of metabolite contents, leaf material harvested at the end of the subjective light phase was prioritized for comparative analysis described below.

To produce leaf infections of barley (*cv. Golden Promise*) and maize (*cv. Early Golden Bantam*) with *Bgh* and *Cg*, respectively, expanding leaves of young plants were inoculated with conidia of *Bgh* and *Cg*. In contrast, the infection of *Early Golden Bantam* with *Um* was performed by injecting sporidia suspension into the leaf canal with a syringe, giving rise to infections on meristematic tissue of developing leaves. For all three interactions studied, targeted analysis of 42 metabolites of central carbon and nitrogen metabolism as well as major low-molecular antioxidants was conducted in four biological replicates per time point and treatment in three independent experiments. For each of the independent experiments, material from all four biological replicates used for metabolite determination was pooled for subsequent transcriptome analysis (as described by Doehlemann et al., 2008).

COMPARATIVE METABOLOME ANALYSIS

Since the aim of our work was to assess, whether common metabolic signatures of biotrophy can be identified in cereal leaves during compatible interactions with fungal leaf pathogens, we first tried to identify similarities between the patterns of the 42 determined metabolites by HCA. As our goal was comparing the dynamics of host metabolism, we employed metabolite ratios between infected and mock control leaves for the HCA analysis, in order to avoid complications by species and experiment specific variation in steady state contents of metabolites. **Table S1** in Supplementary Material contains a compilation of the individual metabolite contents ± SE and the calculated metabolite ratios infected/mock ± SE for all three replicate experiments for all time points and pathosystems analyzed. For the sake of clarity, only two of the three replicate datasets were used for subsequent multivariate data analysis, with the results remaining comparable.

In the HCA, three major clusters could be distinguished that correspond to three different types of interaction (**Figure 2**). The most prominent cluster contained samples derived from *U. maydis*-induced tumors, irrespective, whether the samples were taken at the beginning (*Um* 108hpi) or at the end of the subjective light phase (*Um* 96hpi and *Um* 192hpi), and independent of the developmental state of the tumors. This indicates that tumor

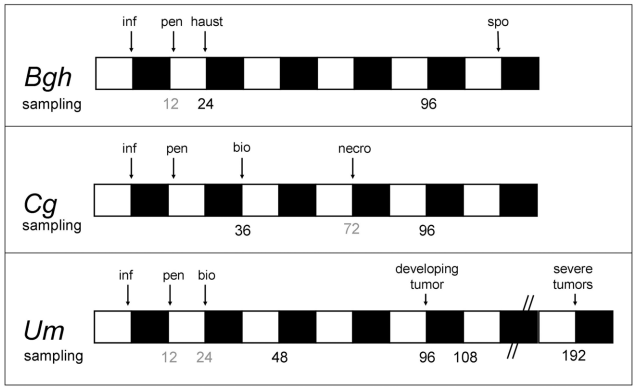


FIGURE 1 | Mapping of sampling time points to respective infection stages in the three studied interactions. *Bgh* – barley (*cv. Golden Promise*) infected with *Blumeria graminis f.sp. hordei* isolate A6 (Wiberg, 1974); *Cg* – maize (*cv. Early Golden Bantam*) infected with *Colletotrichum graminicola* strain CgM2; *Um* – maize (*cv. Early Golden Bantam*) infected with the solopathogenic *Ustilago maydis* strain SG200 (Kämper et al., 2006). White bars indicate light period, and dark bars indicate dark period. Sampling time points analyzed in this study are indicated in black below the bars, while sampling time points that were disregarded are printed in gray. inf, Inoculation; pen, penetration, bio, initiation of biotrophic growth, haus, haustoria establishment; necro, commencement of necrotrophic growth; spo, formation of conidiospores begins.

development determines very profound changes in infected maize leaves (as already observed by Horst et al., 2010a) that even superimpose diurnal variations in metabolite contents. Consequently, these samples were not within the focus of our further analysis, as many metabolic changes specific to tumor formation occur at late stages of the *U. maydis* – maize interaction. However, the metabolite changes in maize leaves during the initial colonization phase at 48hpi, when no tumors had yet been formed, was most similar to that of barley leaves with strong powdery mildew colonization (*Bgh* 96hpi), suggesting that this cluster represents established biotrophic interactions. The third cluster is comprised of samples taken immediately after penetration (*Cg* 36hpi and *Bgh* 24hpi). For the two latter clusters, it is remarkable that the physiological situation of the samples, i.e., immediate post-penetration (*Cg* 36hpi and *Bgh* 24hpi) and established biotrophic interaction (*Um* 48hpi and *Bgh* 96hpi), respectively, appears to be more important for sample parsing than host or pathogen involved. All three clusters mentioned so far were separated from the samples obtained from the necrotrophic phase of *C. graminicola* infection (*Cg* 96hpi). Interestingly, replicate samples of pre-penetration stages (*Bgh* 12hpi, *Um* 12hpi) or from developing leaf tissue (*Um* 12hpi and *Um* 24hpi) did not cluster together when included in the HCA (not shown), indicating that despite strong transcriptional changes for genes involved in central metabolism during basal defense reaction (see corresponding publications by Horst et al., 2010a and Molitor et al., 2011), central metabolism itself was not strongly altered at post-penetration stages. This indicates that changes in central leaf metabolism only occur upon physical interaction with pathogens inside the host tissue, when the drainage of nutrients to the pathogen and the suppression of host defense is being established.

IDENTIFICATION OF METABOLITE DETERMINANTS SPECIFIC FOR INTERACTION STAGES

The obtained results indicate that there must be certain metabolites, which can be used to discriminate the three major clusters produced in the HCA. Therefore, we conducted a principal component analysis (PCA) to identify those metabolite changes that contribute most to the distinction between early post-penetration (*Cg* 36hpi and *Bgh* 24hpi), established biotrophic interaction (*Um* 48hpi and *Bgh* 96hpi) and *U. maydis*-induced tumors. As already suggested by the HCA, principal component 1 (PC1), explaining 41% of the variation, distinguished *U. maydis* tumor samples from the rest (**Figures 3A,B**). Including the time point 108hpi sampled at dawn did not affect the clustering (not shown). As inferred from the metabolite loading scores, Glc, Asn, Ser, Tyr, Gln, and Arg showed the strongest positive distinction, while 3-PGA, PEP, and pyrophosphate exhibited the strongest negative loading in the *U. maydis* tumor samples. PC2, explaining 17% of the variance, separated the necrotrophic interaction (*Cg* 96hpi), indicating substantial differences in the metabolite pattern to all other samples (**Figure 3B**), which was reflected by a strong positive loading of the phosphorylated intermediate F16bP and negative loading of the major amino acids Asp, Ala, and Glu. Finally, PC3, corresponding to 13% of the overall variance, was able to subdivide the early biotrophic interaction time points (above the abscissa) from leaves with established biotrophy (below the abscissa, **Figure 3B**). Branched-chain amino acids, Gly, and His as well as phosphorylated intermediates of carbohydrate metabolism, G16bP, RubP, G1P, UDPglc, pyrophosphate, and the end product sucrose were the most important metabolites to separate these interaction stages from one another.

In general, we have observed numerous metabolite changes at most interaction stages (**Tables 1 and 2**), and therefore we analyzed not only the differences between the early interaction phase (*Cg* 36hpi and *Bgh* 24hpi) and established biotrophy (*Um* 48hpi and *Bgh* 96hpi), but also assessed common metabolite dynamics among these stages. Looking only at those metabolites that changed in average more than 1.4-fold in all four situations of interest (*Cg* 36hpi, *Bgh* 24hpi, *Um* 48hpi, and *Bgh* 96hpi), we could identify glucose and the nitrogen storage amino acids Glutamine and Asparagine being consistently increased, while the glycolytic intermediate PEP and the Calvin cycle intermediate 3-PGA were commonly decreased (**Table 1**). This might indicate that the balance between carbon and nitrogen metabolism and respiration is already readjusted early during compatible interactions. As indicated by the low number of metabolites that were consistently altered more than 1.4-fold in infected leaves in all three pathosystems, the stringency of the interspecies comparison needs to be low in order to identify common metabolic changes. For the vast majority of the regarded biotrophic interaction stages, changes in the abovementioned five metabolites were statistically significant in a Welch–Satterthwaite *t*-test, but not after Benjamini–Hochberg FDR correction.

DISTINCT METABOLITE RESPONSES ARE NOT CAUSED BY CONFINED TRANSCRIPTIONAL PROGRAMS

By multivariate data analysis, we were able to identify common and distinct metabolite changes associated with different phases of compatible biotrophic interactions, which could represent potential metabolic compatibility factors. To identify potential host targets of

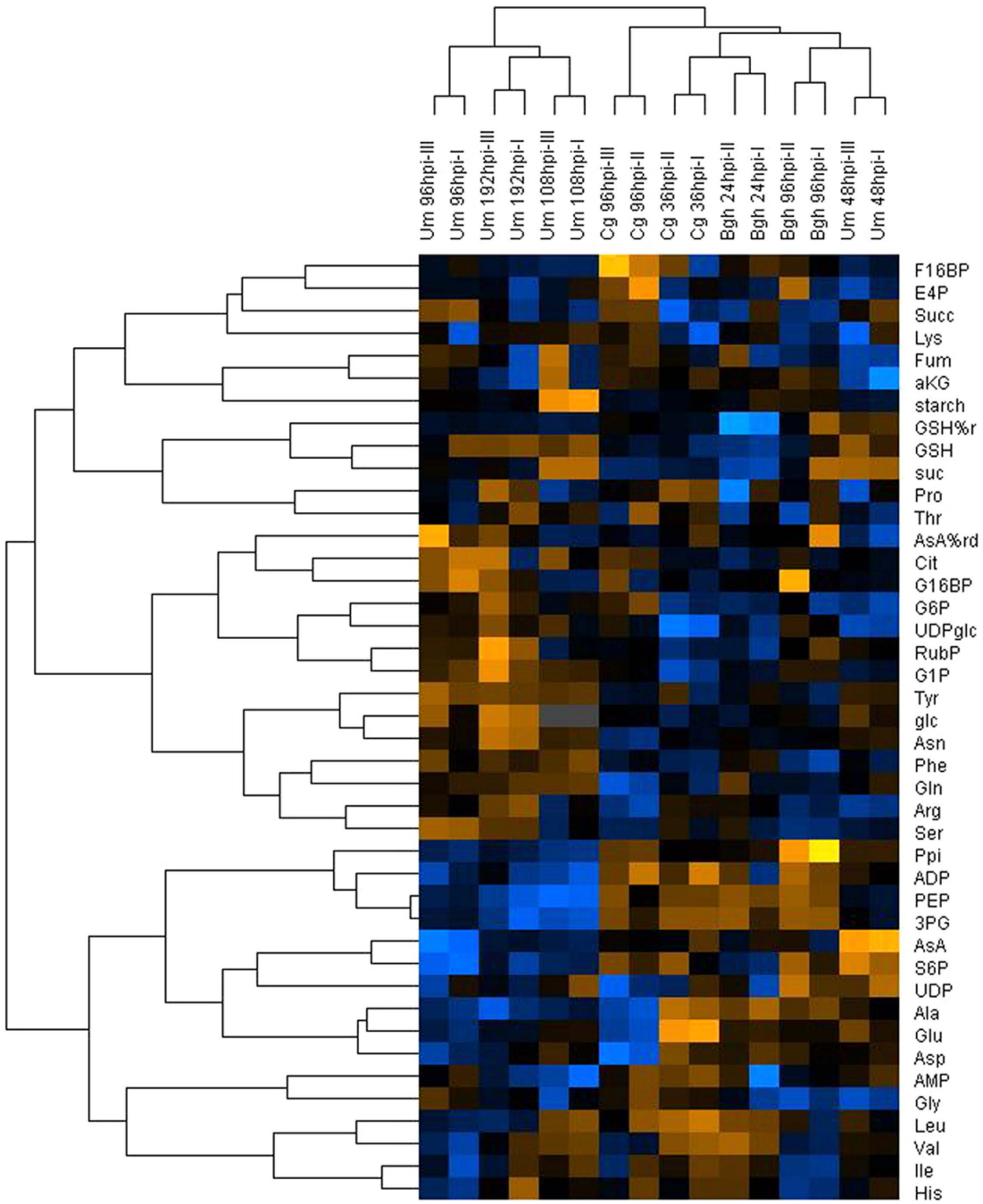


FIGURE 2 | Hierarchical cluster analysis of metabolome data from infected leaves. Mean values of metabolite contents from four biological replicates of infected and mock control leaves harvested at the indicated time points after infection with the respective pathogens (*Bgh*, *Blumeria graminis f.sp. hordei*; *Cg*, *Colletotrichum graminicola*; *Um*, *Ustilago maydis*) were used to calculate the metabolite ratio infected/mock for the indicated experimental replicates. After log transformation of the data, median centered ratios were normalized and hierarchical clustering analysis (HCA) was performed using the complete linkage algorithm of the program Cluster v2.11 (Eisen et al., 1998 – www.eisenlab.org) and the results were visualized using MapleTree (http://mapletree.sourceforge.net/). Metabolite ratios from two independent experiments (indicated by Roman numbers) of every pathosystem were used for HCA. Color intensity correlates with degree of increase (yellow) and decrease (blue) relative to the mean metabolite ratio. hpi, hours post infection. Amino acids and nucleotides are abbreviated according to three letter code, aKG, (α-ketoglutarate) Asc (ascorbate); %AsA red, (% reduced ascorbate); Cit, (citrate); E4P, (erythrose-4-phosphate); F16BP, (fructose-1,6-bisphosphate); F6P, (fructose-6-phosphate); frc, (fructose); Fum, (fumarate); G16BP, (glucose-1,6-bisphosphate); G1P, (glucose-1-phosphate); G6P, (glucose-6-phosphate); glc, (glucose); GSH, (glutathione); %GSH red, (% glutathione reduced); Icit, (isocitrate); PEP, (phospho*enol* pyruvate); 3PG, (3-phosphoglycerate); 6PG, (6-phosphogluconate); Ppi, (pyrophosphate); Pyr, (pyruvate); RubP, (ribulose-1,5-bisphosphate); S6P, (sucrose-6-phosphate); suc, (sucrose); Succ, (succinate); T6P, (trehalose-6-phosphate); UDPglc, (UDP-glucose).

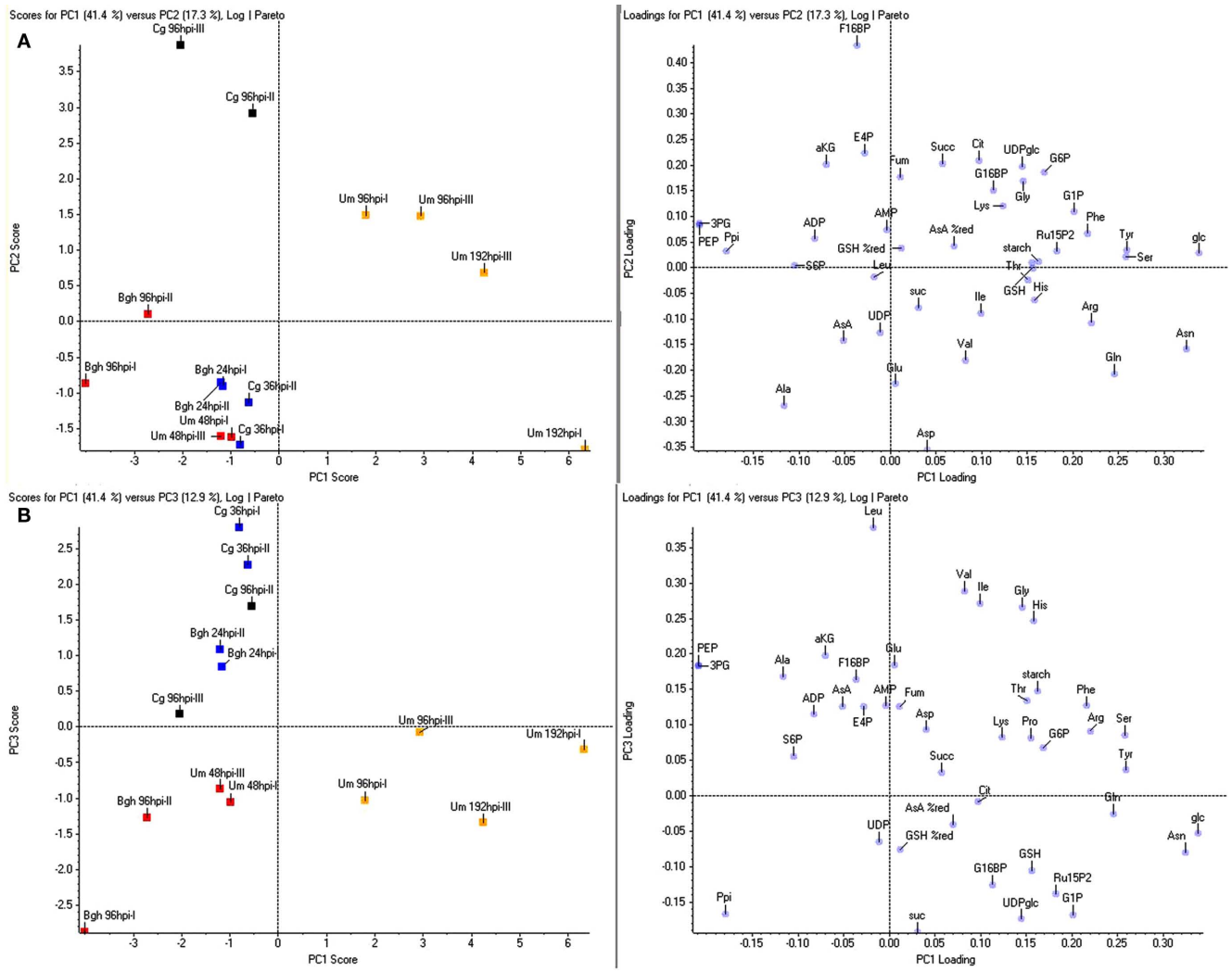


FIGURE 3 | Principle component analysis of metabolite data. For principle component analysis (PCA), the same metabolite ratios as in **Figure 1** were used, representing the ratio infected/mock calculated from the mean values of four sample replicates each. Per pathosystem, data from two independent experiments (designated by Roman numbers) were employed for PCA. **(A)** PC1 vs. PC2. **(B)** PC1 vs. PC3. *Bgh*, *Blumeria graminis f.sp. hordei*; *Cg*, *Colletotrichum graminicola*; *Um*, *Ustilago maydis*; yellow, tumor, red, established biotrophy, blue, early biotrophy, black, necrotrophy. Amino acids and nucleotides are abbreviated according to three letter code, aKG, (α-ketoglutarate) AsA (ascorbate); %AsA red, (% reduced ascorbate); Cit, (citrate); E4P, (erythrose-4-phosphate); F16BP, (fructose-1,6-bisphosphate); F6P, (fructose-6-phosphate); frc, (fructose); Fum, (fumarate); G16BP, (glucose-1,6-bisphosphate); G1P, (glucose-1-phosphate); G6P, (glucose-6-phosphate); glc, (glucose); GSH, (glutathione); %GSH red, (% glutathione reduced); Icit, (isocitrate); PEP, (phospho*enol* pyruvate); 3PG, (3-phosphoglycerate); 6PG, (6-phosphogluconate); Ppi, (pyrophosphate); Pyr, (pyruvate); RubP, (ribulose-1,5-bisphosphate); S6P, (sucrose-6-phosphate); suc, (sucrose); Succ, (succinate); T6P, (trehalose-6-phosphate); UDPglc (UDP-glucose).

metabolic reprogramming by biotrophic fungi, we aimed at refining the underlying transcriptional changes governing the observed metabolic redirections.

We analyzed the corresponding transcriptome data obtained from the same pooled material that was used for metabolite analysis for transcriptional changes that could account for the observed dynamics in the metabolome. As not all genes in a pathway are subject to transcriptional regulation, it appeared instrumental to analyze the enrichment of transcriptional regulation within entire metabolic pathways. To avoid complications by annotation artifacts in the pairwise assignment of the maize and barley microarray features, we preferred to calculate the enrichment of regulated genes within MapMan BINs (Thimm et al., 2004) associated with central primary carbon and nitrogen metabolism, of which the functional annotations are quite robust (**Table 3**). If transcriptional reprogramming of metabolic pathways would account for the observed differences in the metabolome between early post-penetration and at established biotrophy, we would expect a similar clustering result of the transcript data as for the metabolite data. Surprisingly, an HCA comparing the fraction of regulated genes in MapMan BINs assigned to central carbon and nitrogen metabolism gave a completely different picture compared to the metabolome analysis (**Figure 4**). Only the samples reflecting established biotrophy (*Um* 48hpi and *Bgh* 96hpi) still clustered together. On the pathway level,

Table 1 | Metabolites consistently altered in all biotrophic interactions.

Metabolite	*Bgh* 24hpi	*Bgh* 96hpi	*Cg* 36hpi	*Um* 48hpi	Average
Glutamine	3.16 ± 0.52	2.19 ± 0.93	2.42 ± 0.34	2.48 ± 0.18	2.56
Glucose	1.37 ± 0.19	1.35 ± 0.16	1.43 ± 0.10	1.64 ± 0.26	1.44
Asparagine	1.81 ± 0.17	1.67 ± 0.52	2.76 ± 0.53	2.07 ± 0.10	2.07
3-PGA	–1.28 ± 0.11	–1.41 ± 0.08	–1.12 ± 0.10	–2.04 ± 0.07	–1.46
PEP	–1.25 ± 0.05	–1.37 ± 0.10	–1.25 ± 0.10	–2.08 ± 0.05	–1.48
NO. OF METABOLITE WITH F.C. > 1.5 IN INFECTED LEAVES					
Metabolites increased	18 (12 – 3)	11 (3 – 0)	21 (11 – 0)	15 (12 – 11)	
Metabolites decreased	3 (0 – 0)	6 (1 – 0)	2 (0 – 0)	5 (4 – 2)	

Mean values of metabolite contents from four biological replicates of infected and mock control leaves harvested at the indicated time points after infection with the respective pathogens (Bgh, Blumeria graminis f.sp. hordei; Cg, Colletotrichum graminicola; Um, Ustilago maydis) were used to calculate the metabolite ratio infected/mock. Only metabolites with an average f.c. > 1.4 in both early and established biotrophic interactions are displayed (see right column). The total number of metabolites that were increased or decreased at the indicated time point of infection is indicated in the lower part of the table, with the first number in brackets giving significant changes in a t-test with p < 0.05, and the second number giving significant changes with p < 0.05 after Benjamini–Hochberg FDR correction. For individual p-values and Benjamini–Hochberg-corrected p-values, please see ***Table 2****.*

deregulation of major carbon metabolism was consistent enough to parse the MapMan BINs Calvin cycle, sucrose, and starch biosynthesis into the same cluster.

During early post-penetration biotrophy (*Bgh* 24hpi and *Cg* 36hpi), the most pronounced changes in metabolite contents had occurred in the accumulation of most free amino acids (see **Figure 3B** and **Table 2**) as well as by decreased contents of phosphorylated intermediates of starch and sucrose biosynthesis (see **Figure 3B**). While between 8 and 19% of genes annotated to central carbon metabolism and between 12 and 53% of genes annotated to amino acid biosynthesis are up-regulated at 24hpi after *Bgh* infection, most of these MapMan BINs are not regulated at all at 36hpi after *Cg* infection (**Table 3**), demonstrating that although both early post-penetration situations exhibit similar metabolite changes, transcriptional regulation of the corresponding metabolic pathways is utterly different. Likewise, the samples attributed to established biotrophy (*Um* 48hpi and *Bgh* 96hpi) were refined by PCA based on concomitant changes in phosphorylated intermediates of central carbon metabolism (see **Figure 3B**). In addition, the contents of the glycolytic intermediate PEP, and the Calvin cycle intermediates 3-PGA and F16bP (which are predominantly localized in the stroma in illuminated leaves, see Gerhardt et al., 1987; Heineke et al., 1994 and Leidreiter et al., 1995) were consistently diminished at 48hpi after *Um* infection and at 96hpi after *Bgh* infection (**Table 2**). MapMan BINs for sucrose and starch biosynthesis, the Calvin cycle, glycolysis, and major amino acid biosynthesis were much stronger deregulated in *Um* 48hpi than in *Bgh* 96hpi (highlighted in **Table 3**), again indicating a sincere difference on the transcriptional level despite similar metabolite changes as revealed by PCA.

To evaluate whether a more global transcriptome analysis would result in a similar outcome compared to the focused analysis of transcripts involved in central metabolism, we matched all features on the Barley1 and the maize Affymetrix arrays via the corresponding gene annotations deposited at PlexDB[4]. An HCA employing features with fold change >2 from the whole transcriptome dataset resulted in a different sample parsing than in the previous analysis employing data categorized into MapMan BINs. Mostly, clustering of samples occurred predominantly according to pathosystems, irrespective, whether only the previously analyzed samples were clustered or whether all available samples were used for the computation (**Figure 5**). Similar results were obtained when barley and maize genes were matched based on their closest homolog in rice (not shown). Although we still cannot rule out that part of the clustering is influenced by artifacts arising from matching the array annotations, the fact that *Cg* and *Bgh* samples form one cluster in the full transcriptome HCAs (**Figure 5A**) argues against a strong influence by such misinterpretations. At the bottom line, no infection stage specific clustering could be observed when transcriptome data were analyzed.

IDENTIFICATION OF METABOLIC GENES REGULATED IN RESPONSE TO FUNGAL INFECTION

Although we were unable to identify common motifs in the transcriptional response of metabolic pathways at the early post-penetration stage and during established biotrophy, we were surveying the transcriptome data for metabolic genes that were found to be regulated during biotrophic interactions in more than one pathosystem. To address this question, we had to decrease the fold change threshold down to 1.5-fold, as the standard twofold threshold appeared to be too stringent for such a cross-species comparison. **Table 4** shows that genes involved in the TCA cycle and carboxylate metabolism as well as genes regulating the energy status of the nucleotide pool are consistently induced in more than one pathosystem. Similarly, remodeling of amino acid metabolism appears to be a common theme during compatible biotrophic interactions. Surprisingly, the number of targets in central carbohydrate metabolism is quite scarce. While there seems to be different ways of transcriptional regulation of fructose-2,6-bisphosphate homeostasis in all pathosystems, only few more genes in central carbohydrate metabolism were found, but not as consistent as genes involved in carboxylate, nucleotide, and amino acid metabolism, indicating that there is no strong regulation of carbohydrate flux on the transcriptional level.

Thus, a conserved transcriptional program that is activated to redirect primary metabolism during biotrophic interactions does not exist, indicating that the manipulation of host metabolism

[4]www.plexdb.org

Table 2 | Compilation of all substantial metabolite changes at biotrophic interaction time points.

Metabolite	*Bgh*24	*p*-Value	adj *p*-Value	*Bgh*96	*p*-Value	adj *p*-Value	*Cg*36	*p*-Value	adj *p*-Value	*Um*48	*p*-Value	adj *p*-Value
INCREASED IN INFECTED LEAVES												
aKG							1.25 ± 0.24	0.304	0.361			
Ala	1.54 ± 0.20	0.039	0.124									
AMP							1.98 ± 0.12	0.022	0.119			
Arg	1.60 ± 0.11	0.016	0.099				2.03 ± 0.18	0.004	0.050			
Asn	1.81 ± 0.17	0.071	0.136	1.67 ± 0.52	0.094	0.396	2.76 ± 0.53	0.046	0.146	2.07 ± 0.10	0.021	0.056
Asp	1.58 ± 0.18	0.023	0.107	1.28 ± 0.26	0.238	0.452	2.47 ± 0.61	0.054	0.145			
F16BP							1.68 ± 0.63	0.350	0.758			
G1P	1.34 ± 0.12	0.052	0.115	1.29 ± 0.09	0.032	0.242						
G6P							1.25 ± 0.09	0.054	0.136			
Glc	1.37 ± 0.19	0.269	0.310	1.35 ± 0.16	0.299	0.517	1.43 ± 0.20	0.038	0.144	1.64 ± 0.26	0.028	0.067
Gln	3.16 ± 0.62	0.056	0.118	2.19 ± 0.93	0.182	0.385	2.42 ± 0.34	0.043	0.150	2.48 ± 0.18	0.007	0.037
Glu							2.07 ± 0.25	0.004	0.070			
Gly	1.32 ± 0.16	0.097	0.154				2.50 ± 0.19	0.056	0.133			
His	1.94 ± 0.33	0.001	0.016				2.36 ± 0.43	0.018	0.111	1.88 ± 0.06	0.005	0.044
Ile	1.74 ± 0.10	0.013	0.099				2.17 ± 0.17	0.022	0.105	1.44 ± 0.08	0.010	0.034
Leu	1.45 ± 0.08	0.001	0.020	1.26 ± 0.42	0.931	0.982	2.07 ± 0.29	0.291	0.357			
Lys	1.54 ± 0.07	0.020	0.109							1.50 ± 0.43	0.224	0.517
Phe	2.47 ± 0.13	0.001	0.034	1.35 ± 0.39	0.312	0.492	1.99 ± 0.04	0.006	0.059	1.88 ± 0.13	0.008	0.030
Ppi				3.88 ± 0.92	0.064	0.302						
Pro							2.13 ± 0.51	0.088	0.177			
S6P							1.25 ± 0.11	0.065	0.146	1.25 ± 0.12	0.092	0.160
Ser	1.59 ± 0.18	0.050	0.135				2.22 ± 0.30	0.066	0.139	1.25 ± 0.02	0.004	0.048
Suc										1.59 ± 0.09	0.006	0.044
Succ										1.43 ± 0.05	0.001	0.024
Thr	1.77 ± 0.24	0.041	0.119	1.45 ± 0.33	0.166	0.394	2.38 ± 0.41	0.034	0.144	1.79 ± 0.15	0.014	0.044
Tyr	1.52 ± 0.21	0.031	0.118				1.97 ± 0.39	0.051	0.150	1.81 ± 0.04	0.000	0.003
UDP				1.29 ± 0.09	0.025	0.318						
UDPglc	1.50 ± 0.13	0.035	0.120	1.65 ± 0.19	0.015	0.294				1.28 ± 0.12	0.094	0.155
Val	1.81 ± 0.12	0.024	0.103				2.24 ± 0.14	0.007	0.051	1.30 ± 0.05	0.007	0.031
REDUCED IN INFECTED LEAVES												
3-PGA	1.28 ± 0.11	0.075	0.130	1.41 ± 0.08	0.026	0.243				2.04 ± 0.07	0.033	0.074
aKG										1.89 ± 0.11	0.063	0.113
F16BP				1.21 ± 0.17	0.137	0.435				1.45 ± 0.08	0.047	0.099
Gly				1.42 ± 0.12	0.142	0.416						
Lys				1.25 ± 0.30	0.363	0.511	1.25 ± 0.27	0.291	0.357			
PEP	1.25 ± 0.05	0.051	0.128	1.37 ± 0.10	0.052	0.333	1.25 ± 0.10	0.095	0.181	2.08 ± 0.05	0.017	0.050
Ser												
Suc	1.25 ± 0.06	0.058	0.116									
Starch				1.63 ± 0.11	0.094	0.357				2.27 ± 0.04	0.006	0.041

The metabolite ratio infected/mock is given as the mean value of three independent experiments ± SE. In each experimental replicate, four biological replicates were analyzed. Leaves of infected and mock control leaves were harvested at the indicated time points after infection with the respective pathogens (Bgh, Blumeria graminis f.sp. hordei; Cg, Colletotrichum graminicola; Um, Ustilago maydis). Metabolites with an average f.c. > 1.25 at any biotrophic interaction time point are displayed. p-Values were calculated employing a Welch–Satterthwaite t-test and for multiple testing correction of p-values, Benjamini–Hochberg false discovery rate (FDR) was determined (adj p-value). For abbreviations, see legend of ***Figure 1****.*

depends on the individual pathogen and the effector proteins it produces. Nevertheless, some metabolic pathways seem to be consistently addressed on the transcriptional level in all investigated pathosystems. Nitrogen metabolism and energy status appeared to be regulated more consistently on the transcriptional level than carbohydrate metabolism, which might be rather controlled on the post-translational level or by interaction-specific modulations.

IDENTIFICATION OF TRANSCRIPTIONAL SIGNATURES IN THE INVESTIGATED PATHOSYSTEMS

Our analysis has only revealed a few genes of central primary metabolism that were regulated in all investigated pathosystems. As stated above, a conserved transcriptional program that is activated to redirect primary metabolism during biotrophic interactions does not exist. Therefore, we set out to identify particular

Table 3 | Percentage of up- and down-regulated genes in MapMan BINs of central primary metabolism.

MapMan BIN	*Bgh* 24hpi %	*Bgh* 96hpi %	*Um* 48hpi %	*Um* 96hpi %	*Um* 108hpi %	*Cg* 36hpi%	*Cg* 96hpi %
PERCENTAGE OF UP-REGULATED							
Starch BS	8.1	0.0	3.3	23.3	23.3**	0.0	10.0
Starch Deg	14.8	**0.0**	**21.7****	30.4	34.8	0.0	0.0
Sucrose BS	0.0	**7.1**	**16.7**	16.7**	0.0	0.0	16.7
Sucrose Deg	19.2	11.5	18.5	18.5	37.0	0.0	18.5**
OPPP	20.5	**23.1**	**6.9**	20.7	44.8	6.9	10.3**
Glycolysis	10.1	**7.1**	**25.8**	29.0*	38.7	9.7*	17.7*
Fermentation	27.6	13.8	26.3	42.1	52.6	10.5	21.1**
TCA cycle	17.7	11.4	16.7	18.8	35.4**	2.1	22.9***
Calvin cycle	8.2	13.7	10.0***	13.3***	16.7***	3.3**	10.0
Photorespiration	10.0	16.0	11.9**	9.5***	11.9	0.0	4.8
aa Deg	11.8	9.2	13.3**	16.3	26.7*	2.2	9.6
Glu aa BS	53.3***	**46.7*****	**11.8**	17.6	41.2	11.8	5.9**
Asp aa BS	17.3***	13.5***	22.0	26.8*	24.4	4.9	12.2
bc-aa BS	23.1*	**0.0**	**12.5**	37.5	43.8	0.0	12.5
Ser BS	23.1	15.4**	25.0	25.0**	41.7***	12.5**	25.0***
aro-aa BS	37.7***	43.4***	36.4**	38.6	36.4	0.0	31.8**
His BS	12.5	18.8***	12.5	6.3	6.3	0.0	6.3**
Nucleotides	16.4**	15.1***	13.2	27.2	41.2**	3.5**	14.0**
PERCENTAGE OF DOWN-REGULATED							
Starch BS	10.8**	**5.4****	**21.7***	43.5**	60.9**	0.0	8.7
Starch Deg	3.7*	**3.7**	**16.7****	16.7	30.0**	3.3**	6.7
Sucrose BS	0.0	**14.3****	**33.3****	66.7**	66.7**	0.0	16.7**
Sucrose Deg	1.9	7.7**	7.4	18.5	22.2	3.7	3.7**
OPPP	10.3**	2.6	3.4	20.7	24.1	0.0	3.4
Glycolysis	5.1***	4.0**	8.1**	11.3**	38.7**	0.0	3.2*
Fermentation	3.4	**3.4**	**15.8***	31.6**	26.3**	0.0	5.3
TCA cycle	1.3**	1.3***	2.1**	2.1**	12.5**	0.0	2.1**
Calvin cycle	11.0***	**6.8****	**16.7***	56.7**	66.7***	3.3**	6.7**
photorespiration	6.0***	**0.0**	**14.3***	31.0**	45.2**	2.4**	31.0
aa Deg	6.5***	**2.6****	**10.4*****	21.5**	22.2**	0.7	6.7
Glu aa BS	0.0	**0.0**	**17.6**	17.6	11.8	0.0	5.9*
Asp aa BS	0.0	**1.9**	**9.8**	12.2	14.6	0.0	4.9
bc-aa BS	7.7	7.7	12.5	18.8	56.3	0.0	12.5
Ser BS	0.0	**0.0**	**12.5****	16.7*	33.3**	0.0	0.0
aro-aa BS	3.8	1.9	2.3	13.6	15.9	6.8	11.4**
His BS	0.0	0.0	6.3	6.3	12.5	0.0	6.3
Nucleotides	4.8**	**2.7****	**8.8****	14.0	15.8**	0.0	2.6*

*The percentage of up- (upper half) and down-regulated genes (lower half) in the indicated MapMan BINs was calculated based on the mean fold changes from all three replicate experiments. Strong differences between the two time points of established biotrophy, Bgh 96hpi and Um 48hpi, are indicated in bold. Significant enrichment of MapMan BINs was calculated with a Wilcoxon rank sum test: *$p < 0.1$ and **$p < 0.05$. ***indicates $p < 0.05$ in a Wilcoxon rank sum test after Benjamini–Hochberg FDR correction. BS, biosynthesis; Deg, degradation; aro-aa, aromatic amino acids; bc-aa, branched-chain amino acids; OPPP, oxidative pentose phosphate pathway.*

differences in the transcriptional responses of metabolic pathways between the pathosystems. In **Figure 6**, MapMan representations of those pathways were compiled that exhibit most pronounced differences between early biotrophy in the barley powdery mildew interaction (*Bgh* 24hpi), biotrophy in the *U. maydis*-maize interaction (*Um* 48hpi), necrotrophy (*Cg* 96hpi), and *U. maydis*-induced tumors (*Um* > 96hpi), each representing one cluster in the HCA of the metabolite data depicted in **Figure 2**. Please note that the transcriptional changes upon *Bgh* infection were very similar at 24 and 96hpi (also see **Figure 5A**).

As already published (Doehlemann et al., 2008a; Horst et al., 2010a,b), *U. maydis*-induced tumors exhibit substantial transcriptional changes in almost all displayed metabolic pathways compared to mock control leaves (**Figure 6**; **Table 3**). While the majority of the genes involved in the light reaction, the Calvin cycle, and the photorespiratory C_2 cycle were transcriptionally repressed in tumors, genes of lipid biosynthesis and remodeling, cell wall biosynthesis were significantly induced in comparison to mock control leaves >4 dpi. More subtle transcriptional differences at early stages of the three interactions could be identified. In *Bgh* infected leaves,

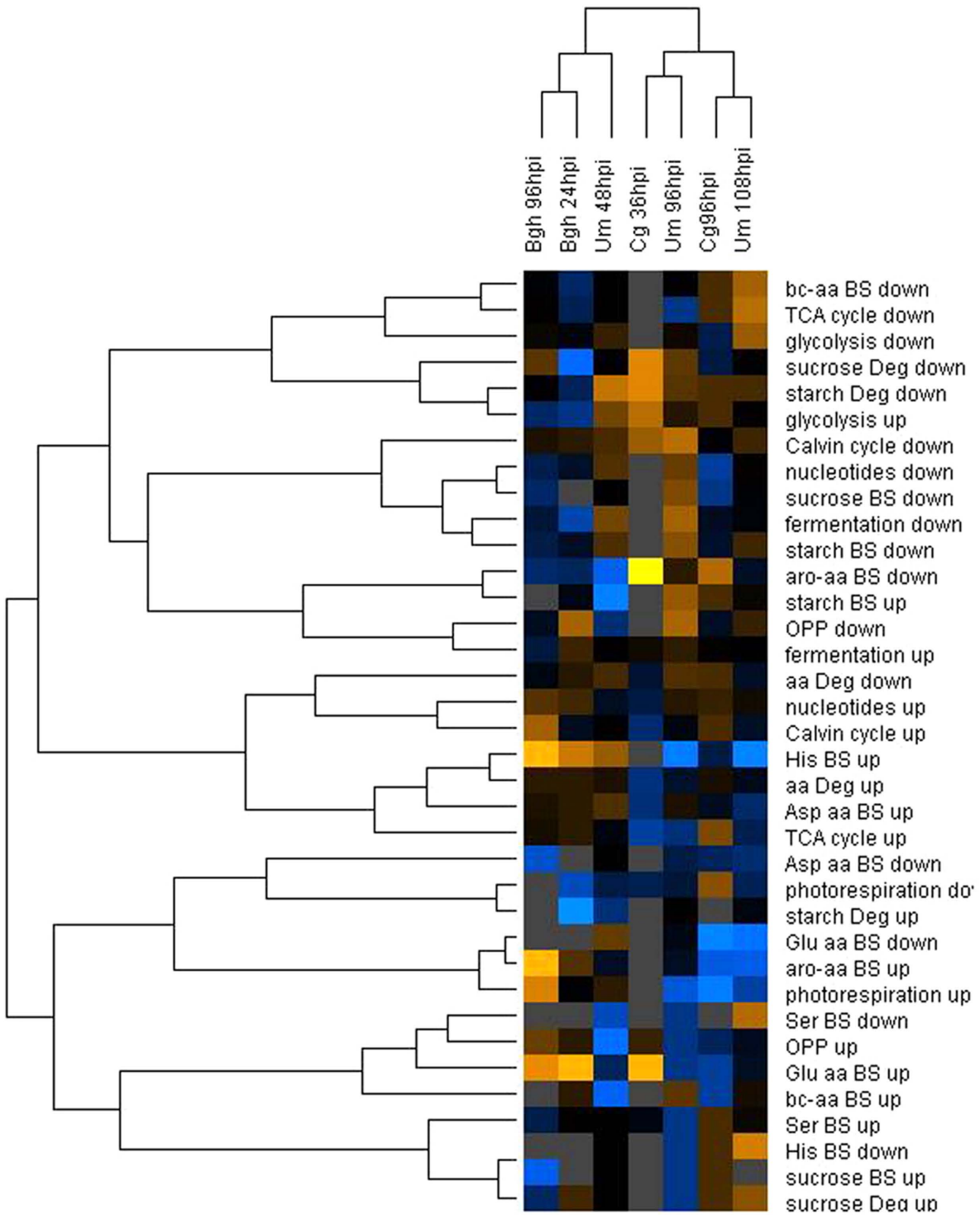


FIGURE 4 | Hierarchical cluster analysis of MapMan BIN enrichment. The percentage of up- and down-regulated genes in each MapMan BIN was calculated based on the mean fold changes from all three replicate experiments. After log transformation of the data, median centered percentages were normalized and hierarchical clustering analysis (HCA) was performed using the complete linkage algorithm of the program Cluster v2.11 (Eisen et al., 1998 – www.eisenlab.org) and the results were visualized using MapleTree (http://mapletree.sourceforge.net/). Color intensity correlates with degree of increase (yellow) and decrease (blue) relative to the BIN mean of all samples, while gray corresponds to 0% regulated genes in the MapMan BINs. *Bgh, Blumeria graminis f.sp. hordei*; *Cg, Colletotrichum graminicola*; *Um, Ustilago maydis;* hpi, hours post infection. Amino acids are abbreviated by three letter code; aro-aa, aromatic amino acids; bc-aa, branched-chain amino acids; BS, biosynthesis; Deg, degradation; OPPP, oxidative pentose phosphate pathway.

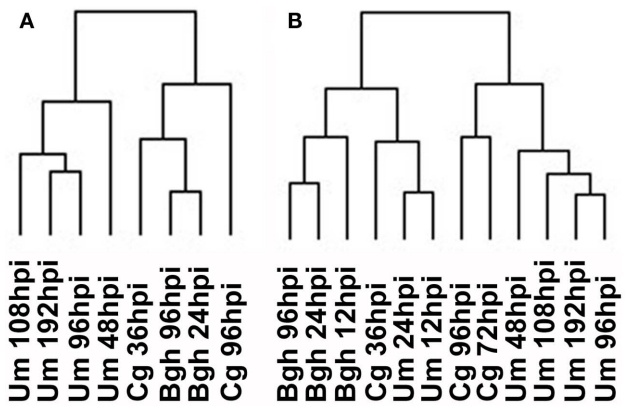


FIGURE 5 | Hierarchical cluster analysis of whole transcriptome data. **(A)** HCA for the timepoints analyzed in this report. **(B)** HCA for all timepoints sampled. Barley and maize Affymetrix data were matched via PlexDB (www.plexDB.org) as described in the methods section and the mean transcript fold changes of infected vs. mock samples were calculated based on all three replicate experiments and were subsequently used for HCA. After log transformation of the data, median centered data were normalized and hierarchical clustering analysis (HCA) was performed using the complete linkage algorithm of the program Cluster v2.11 (Eisen et al., 1998 – www.eisenlab.org) and the results were visualized using Maple Tree (http://mapletree.sourceforge.net/). *Bgh, Blumeria graminis f.sp. hordei*; *Cg, Colletotrichum graminicola*; *Um, Ustilago maydis;* hpi, hours post infection.

transcriptional suppression of the light reaction is substantially more pronounced, while transcriptional regulation of other metabolic pathways is much weaker compared to *Um* biotrophy or in *Cg* necrotrophy (also see **Table 3**). This corroborates that changes in metabolic flux during *Bgh* infection are not predominantly caused by transcriptional regulation, but rather by post-translational fine-tuning. During the biotrophic (pre-tumor) colonization of maize by *U. maydis*, lipid biosynthesis, and cell wall biosynthesis are much stronger induced than in the other two interactions, already reflecting initial hypertrophic growth. Finally, necrotrophic colonization of maize leaves by *C. graminicola* results in a significant induction of glycolysis, TCA cycle, and fermentation, indicating that an increase in respiratory flux might occur during the challenge with the necrotroph.

DISCUSSION

CONSISTENT MOTIFS IN METABOLIC REPROGRAMMING DURING PLANT–PATHOGEN INTERACTIONS

Despite the extensive use of metabolomics for the analysis of plant metabolism (Bino et al., 2004), metabolomic studies of plant–pathogen interactions are rare, most of which rely on data acquisition by FIE-MS and NMR-based metabolite profiling and fingerprinting techniques and subsequent deconvolution by supervised or non-supervised multivariate data analysis (Widarto et al., 2006; Parker et al., 2009; Sana et al., 2010; Ward et al., 2010).

In these approaches, defense-associated metabolites like glucosinolates (Ward et al., 2010), indoles (Ward et al., 2010), and phenylpropanoids (Widarto et al., 2006; Parker et al., 2009; Sana et al., 2010; Ward et al., 2010) were commonly the most prominent metabolite changes that could be identified in infected leaf tissue. The pool sizes of branched-chain amino acids and aromatic amino acids fueling glucosinolate and phenylpropanoid biosynthesis with building blocks increased concomitantly (Parker et al., 2009; Sana et al., 2010; Ward et al., 2010; this study). Nevertheless, elevated Gln and Asn contents in infected leaves represented a major consistent change in primary metabolism at any stage during the biotrophic interactions in the cereal pathosystems investigated in our study (**Table 1**). Taken together with the reported results from the above cited and other studies (reviewed by Bolton, 2009), this indicates that a substantial reprogramming of central amino acid metabolism takes place already early during infection. It has to be stressed that early after the establishment of the three investigated interactions, only five metabolites exhibited consistent changes, most of which were only moderately significant. Likewise, only few genes coding for enzymes of primary metabolism were consistently altered on the transcriptional level during the early stages of infection. Most of these genes were only deregulated in two out of the three pathosystems. This indicates that the congruence of the metabolic response is rather low in the three examined cereal pathosystems.

Nevertheless, malate dehydrogenase (MDH) was found to be consistently induced early in all interactions we investigated and, with the exception of *U. maydis*-infected leaves, various isoforms of malic enzyme were also induced swiftly after inoculation (**Table 4**). Likewise, rice leaves challenged with *Magnaporthe grisea* (Parker et al., 2009) and *Arabidopsis* leaves in defense of the hemibiotroph *Colletotrichum higginsianum* (Voll et al., unpublished) exhibited an induction of malic enzyme activity that was shown to support the global defense response by providing reducing equivalents (Parker et al., 2009), identifying malic enzyme as a conserved player in early, i.e., basal plant defense. Like malic enzyme, MDH, would also produce reducing equivalents from the oxidation of malate in the cytosol, yet producting oxaloacetate instead of pyruvate, thereby competing with ME for the substrate malate.

Cell wall bound invertase is known to be involved in the defense response of several plant species (e.g., Bonfig et al., 2006; Swarbrick et al., 2006; Voegele et al., 2006; Essmann et al., 2008; Horst et al., 2008; Kocal et al., 2008; Siemens et al., 2011). Interestingly, we could only observe an induction of cell wall invertase (cw-INV) at late interaction stages of the two maize pathosystems, indicating that its induction might be slower in maize than in other species. As both malic enzyme (Parker et al., 2009) and invertase (see citations above) have been shown to be induced much stronger and faster in incompatible than in compatible interactions, we can rule out that their transcriptional induction represents a susceptibility factor.

An increase in TCA cycle intermediates Citrate, Malate, Succinate, and Fumarate had been observed during the necrotrophic phase of *Magnaporthe grisea* infection (Parker et al., 2009). Similarly, we have observed an accumulation of these carboxylates with isocitrate exhibiting the most pronounced increase in maize leaves during necrotrophic colonization with *C. graminicola* at 96hpi (**Table S1** in Supplementary Material). In addition, TCA cycle, glycolysis, and respiration displayed the strongest induction on the transcriptional level in *C. graminicola* infected leaves at that time point (**Figure 6**), suggesting that necrotrophic growth in grass species might commonly provoke a strong induction of

Table 4 | Consistent transcriptional changes among the pathosystems.

Gene	Probe set Barley1	*Bgh* 24hpi	*Bgh* 96hpi	Probe set Maize Affy	*Cg 96hpi*	*Um* 48hpi	Classification
TCA CYCLE/CARBOXYLATE METABOLISM							
Citrate lyase	Contig3815_at	2.1		Zm.1942.1.A1_at	4.4		Early
Aconitase	Contig3351_s_at	1.5		Zm.12697.1.S1_at	2.1	2.6	Global
Oxoglutarate dehydrogenase	Contig4963_at		1.6	Zm.6807.1.A1_at	5.0	2.0	Global
Cytosolic malate dehydrogenase	Contig3610_s_at	1.9	3.5	Zm.2061.1.A1_at	3.0	4.0	Global
NAD malic enzyme	HV_CEb0015P21f_S_at	1.8	1.5	Zm.3666.1.A1_at	1.9		Global
Pyruvate decarboxylase	Contig5532_s_at		1.7	Zm.3994.1.S1_at		10.2	Late
NUCLEOTIDE METABOLISM							
Adenosine/uridine kinase	Contig2829_at	1.7	1.9	Zm.247.2.A1_at	5.2	32	Global
UMP synthase	Contig16393_at	1.9	2.3	Zm.908.1.A1_at	3.0		Global
Nucleoside diphosphosphate kinases*	Contig2124_at	2.3		Zm.19303.1. S1_at (Zm.17247.1. A1_at)	27	4.1	Global
AMINO ACID METABOLISM							
Serine hydroxymethyltransferase	Contig2168_s_at	1.8	2.6	Zm.3136.1.A1_at		2.4	Global
Methylene tetrahydrofolate reductase	Contig3235_s_at	1.8		Zm.475.1.S1_at	1.9		Early
Amino acid transporters	Contig26356_at	1.5	1.5	Zm.1788.1.A1_at	–	2.6	Global
Aminotransferases**	Contig1672_s_at	–2.5	–1.8	Zm.13511.1.A1_at Zm.2321.1.A1_at	1.5	–4.2	Global biotr
Glutamine synthetase	Contig1646_at	–1.7		Zm.3455.3.A1_at	–1.6		Early
PRPP synthetase	Contig8025_at	–5.3	2.3	Zm.1727.1.A1_at	–2.0	1.6	Early < to > late
SUCROSE METABOLISM							
PFK2			1.5	Zm.711.1.S1_at		3.0	Late
H^+/PPase	Contig385_s_at	–2.0		Zm.6095.1.A1_at	2.1	4.3	
STARCH METABOLISM/CALVIN CYCLE							
AGPase	Contig5267_at	1.6		Zm.312.1.A1_at	1.9		Early
Phosphoribulokinase	rbaal2124_s_at	–1.8		Zm.2248.1.A1_at		–2.5	
Aldolase	Contig4817_at		1.8	Zm.4778.1.A1_at		2.5	Late

Transcriptional changes during biotrophic Bgh colonization (early stage – Bgh 24hpi; late stage Bgh 96hpi), necrotrophic Cg colonization (Cg 96hpi), and biotrophic Um colonization (Um 48hpi) is compiled. Fold changes were calculated based on mean values from all three replicate experiments, respectively. Only genes with f.c. > 1.5 are displayed that are regulated in more than one pathosystem.

**For Cg 96hpi, data for adenylate kinase (Zm.17247.1.A1_at) are given.*

***For Bgh, and Cg (Zm.13511.1.A1_at), the induced aminotransferases are annotated as aspartate glutamate aminotransferases, while the Um induced aminotransferase (Zm.2321.1.A1_at) is supposed to be an alanine oxoglutarate aminotransferase. Both of these aminotransferase activities are associated with central nitrogen metabolism.*

carboxylate metabolism via the TCA cycle. Increased metabolic flux into carboxylate production via the TCA cycle could either provide ample supply of reducing equivalents and ATP to the host cells or it could indicate increased respiratory flux as a result of impaired photoautotrophy during necrotrophic colonization. We also observed a strong induction of the TCA cycle in *U. maydis*-induced tumors (**Figure 6**). However, a thorough inspection disclosed that in leaf tumors, the TCA cycle will most probably provide nitrogen assimilation with carbon skeletons (Horst et al., 2010a,b).

Apart from only few very conserved responses of primary metabolism on the transcriptional and metabolic level between all three pathosystems, we could also identify changes that were quite specific to only one of the pathosystems investigated here. For instance, the induction of the TCA cycle and the accumulation of carboxylates was most pronounced during necrotrophic colonization of maize by *Cg* (as just discussed), whereas lipid and cell wall biosynthesis were most severely affected in the early interaction of maize leaves with *Um*, reflecting the initiation of hypertrophic growth. Genes involved in photosynthesis were most quickly suppressed in *Bgh* infected barley leaves. Given the connection of the *Cg* and the *Um* response to particular physiological situation described above, we can assume that these specific effects on the transcriptome and the metabolome of primary metabolism arise from the divergent strategies of the fungal pathogens to manipulate host metabolism. It appears likely, that besides specific targets in defense signaling (as reviewed by de Wit et al., 2009), also different enzymes and metabolic pathways are targeted by the fungi to match the metabolic requirements of the individual pathogens.

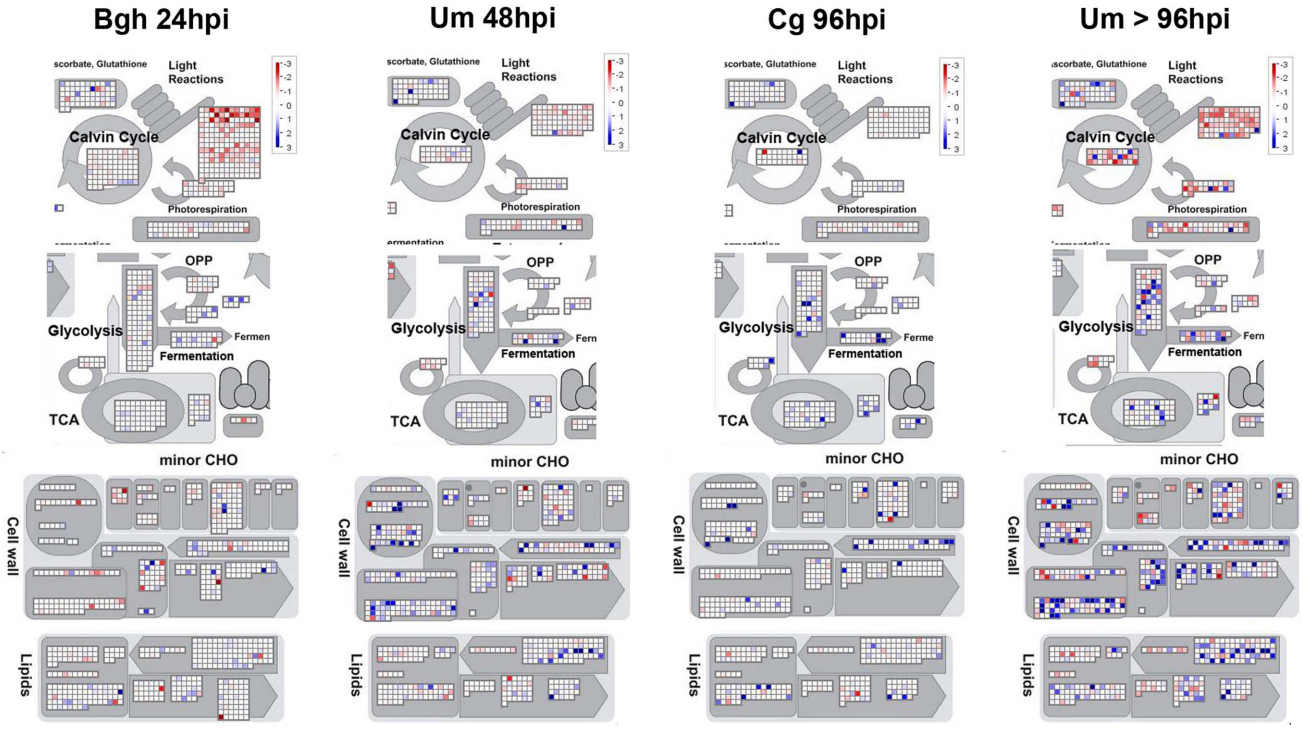


FIGURE 6 | Visualization of specific transcriptional changes between biotrophy, necrotrophy, and tumors by MapMan analysis. Mean fold change of the transcripts in infected vs. mock treated samples was calculated based on all three replicate experiments per time point and log-scaled data were visualized using MapMan with all filters disabled. Data for established biotrophy (*Bgh* 96hpi and *Um* 48hpi), necrotrophy (*Cg* 96hpi), and *Um* tumors (*Um* > 96hpi) are displayed from left to right. The scale bar represents fold change and reaches from –3 (red) to +3 (blue) on the log scale, which corresponds to –8-fold (red) to +8-fold (blue) in the linear scale. Photosynthesis (Light reactions), Calvin cycle, and photorespiration are depicted in the upper panel, glycolysis, TCA cycle, and fermentation are shown in the second tier from the top, cell wall metabolism, and minor carbohydrate metabolism (minor CHO) are represented in the second panel from the bottom, while lipid metabolism is displayed in the bottom panel.

METABOLIC CHANGES ARE NOT CAUSED BY A CONSERVED TRANSCRIPTIONAL REPROGRAMMING

Based on the comparison between pathosystems from our study and published data, we could resolve some recurring metabolic motifs in response to pathogen infection.

In the few references available to date, it remains controversial, however, whether the observed changes in metabolism fit to the corresponding transcriptome dynamics in infected leaf tissue. Sana et al. (2010) only reported a weak accordance of metabolome dynamics and the corresponding transcriptional changes in the assessed compatible and incompatible *X. oryzae* pv. *oryzae* (*Xoo*)-rice interactions. In contrast, Ward et al. (2010) stated a quite substantial congruence when aligning their metabolome data with the transcriptome analysis of publically available data for *Pst* infections on *Arabidopsis* (Truman et al., 2006). A specific re-assessment of the data by Ward et al. (2010) and Truman et al. (2006) did, however, not reveal a substantial number of regulated genes involved in central carbon and nitrogen metabolism, while the highest agreement of metabolome and transcriptome data was achieved for glucosinolate and phenylpropanoid metabolism (Ward et al., 2010). Similarly, we were also unable to identify a strong congruence between the observed changes in primary metabolism and the corresponding transcript data. Nevertheless, both Sana et al. (2010) and Ward et al. (2010) observed few, but consistent changes in some pathways of primary metabolism on the transcript and the metabolite level, which is in accordance with our results.

Although our metabolome dataset was restricted to quantitative data for only 42 metabolites of central primary metabolism obtained via targeted LC and LC-MS-based methods, we could separate discrete infection stages of the three interactions by HCA, i.e., early biotrophy, established biotrophy, necrotrophy, and tumors (**Figure 3**), indicating that the information in our dataset provided sufficient divergence. However, we cannot rule out that metabolic flux through certain pathways even differs between those interaction stages that clustered together in the HCA, because our metabolome data comprises of steady state contents that indicate individual metabolite accumulation, but do not reflect metabolic flux.

In addition, multivariate analysis of the metabolome data did not yield comparable results to any transcriptome based analysis in our study. Even if only transcripts coding for proteins involved in the corresponding pathways were regarded, no similarity to the metabolome data could be attained. This could be due to several reasons. First, secreted effectors of *Bgh*, *Cg*, and *Um* are very likely to target different molecular processes in their respective hosts, leading to interaction-specific variation in the observed transcriptional response that could mask common motifs in the defense response. From the complementary point of view, the defense reactions that are not suppressed by

the pathogens during the investigated compatible interactions will diverge on the molecular level between the pathosystems. Second, transcript amounts and steady state contents of metabolites, which have been assessed in this study, are not directly correlated with metabolic flux. Primary carbon metabolism is strongly regulated on both, the post-transcriptional and the post-translational level throughout the diurnal cycle (e.g., Gibon et al., 2004), which could lead to a discrepancy between the assessed transcript amounts and actual *in vivo* activity of most enzymes in central carbon metabolism – which we did not determine. In addition, we have only measured steady state contents of the metabolites included in our metabolome dataset. As outlined above, despite similar steady state pools of most metabolites, flux could be utterly different between two specimen. Nevertheless, we have obtained evidence that allosteric regulation of key steps in central carbon and nitrogen metabolism is likely to account for some of the regulation of metabolic flux during fungal biotrophy, as indicated in the models shown below.

MODELS FOR THE REDIRECTION OF PRIMARY METABOLISM DURING EARLY BIOTROPHIC INTERACTIONS

By analyzing steady state contents of 42 metabolites in primary carbon and nitrogen metabolism, we were able to reveal similarities and differences in the response of host metabolism toward *Bgh* infection in barley leaves, *Cg* infection in maize leaves, and *Um* infection in maize leaves. Together with the transcriptome data obtained from the same samples, we integrated all the information into individual models of host metabolism at early time points in the investigated biotrophic interactions. We assumed that individual changes on the transcriptional and the metabolic level would not necessarily have to be comparable in strength. Therefore, we used a low stringent evaluation of our data for the generation of the presented models of primary metabolism, in order to better allow for comparisons between the pathosystems. When taken together, the integrated transcriptional and metabolite data were highly consistent for most of the depicted pathways in all three analyzed interactions.

Our survey for consistently regulated genes had already revealed that the TCA cycle, nucleotide energy status and amino acid metabolism represented strongly regulated pathways at early stages of all three interactions (**Table 4**). A concomitant induction of the TCA cycle and nucleotide diphosphate kinases apparently reflects an increase requirement for building blocks, reducing power, and energy in host leaves during the early interaction stage.

Consequently, a comparison of the models for *Bgh*, *Cg,* and *Um* infected leaves during early biotrophic colonization reveals quite similar gross tendencies between two or more pathosystems, despite all the singular differences discussed earlier (**Figure 7**): (i) the biosynthesis of the major amino acids Gln and Asn as well as of the defense-associated branched-chain and aromatic amino acids are commonly induced, (ii) the Calvin cycle and/or starch biosynthesis are reduced while (iii) glycolysis and the TCA cycle are more frequented. Mostly, (iv) photorespiration is elevated, while sucrose biosynthesis is hampered. Because these changes in primary metabolism are not specific to one particular pathosystem, it appears likely that they are part of a common response of cereal primary metabolism during the early infection phase rather than being associated with particular responses of the hosts toward targeted manipulation by individual pathogens.

As suggested by our comparative analyses, there are important differences between the interactions. In the early powdery mildew (*Bgh*) infection (**Figure 7A**), the Calvin cycle genes RubisCO and phosphoribulokinase become transcriptionally repressed, which is also reflected by diminished 3-PGA contents that commonly correlate with activity of the Calvin cycle. Based on steady state metabolite contents and the transcriptome data, metabolic flux appears to be directed toward the biosynthesis of free amino acids, with the major amino acids Gln and Asn representing transient stores for organic N. In contrast, the TCA cycle is induced on the transcriptional level, but the involved carboxylates did not accumulate. Therefore, the depletion of the PEP pool (including the related transcriptional changes) can rather be interpreted due to its anaplerotic function of plastid-localized biosynthesis of branched-chain and aromatic amino acids (Schulze-Siebert et al., 1984; Herrmann and Weaver, 1999). As both plastidic and cytosolic protein biosynthesis are significantly induced processes in the barley–*Bgh* interaction, it appears likely that elevated production of amino acids will serve as building blocks for both PR proteins and low-molecular weight compounds like glucosinolates and phenylpropanoids. Sucrose breakdown by cw-INV and sucrose synthase (SuSy) could lead to an increase in the hexose/sucrose ratio. Due to the transcriptional repression of the Calvin cycle and the photosynthetic electron transport chain (**Figure 6**) diminished triose phosphate export and increased flux toward PEP could limit sucrose biosynthesis in turn. An accumulation of UDPglc further indicates diminished formation of sucrose in *Bgh* infected leaves.

During the biotrophic phase of the *C. graminicola* infection at 36hpi (**Figure 7B**), almost no transcriptional changes were observed, in contrast to numerous changes in steady state metabolite contents. The observed changes in metabolite contents are either the result of endogenous post-transcriptional, perhaps of allosteric regulation, or are due to altered flux through the respective pathways determined by substrate availability or substrate compartmentation, or it might be effectuated by the action of fungal effectors. As indicated by the high similarity to *Bgh* 24hpi in the HCA analysis of metabolite data (**Figure 2**), the changes in the metabolome of *Cg* 36hpi are almost congruent to those of *Bgh* infected barley leaves at 24hpi, rendering it unlikely that secreted effectors of *Bgh* and *Cg* exert identical effects on host metabolism. Furthermore, the sampled maize leaves perform C_3–C_4 intermediate photosynthesis, while barley is a C_3-plant. The only substantial difference between the *Bgh* infected barley and the *Cg* infected maize leaves is an increased accumulation of the TCA cycle intermediates and amino acid building blocks α-ketoglutarate and isocitrate. At 96hpi, the majority of the genes involved in the TCA cycle are strongly induced on the transcriptional level in *Cg* infected leaves. Besides, an increased Gly/Ser ratio indicates increased photorespiration during biotrophic colonization of maize leaves with *C. graminicola*.

Maize leaves infected with *U. maydis* differ in two important aspects from the two previously regarded pathosystems (**Figure 7C**). First, the accumulation of free amino acids resembles the previously described situation for *Bgh* and *Cg*, except for the fact that anaplerotic provision of carbon skeletons by the TCA cycle does not appear to be substantially induced. Second, the balance of sucrose biosynthesis and sucrose degradation seems to be strongly regulated

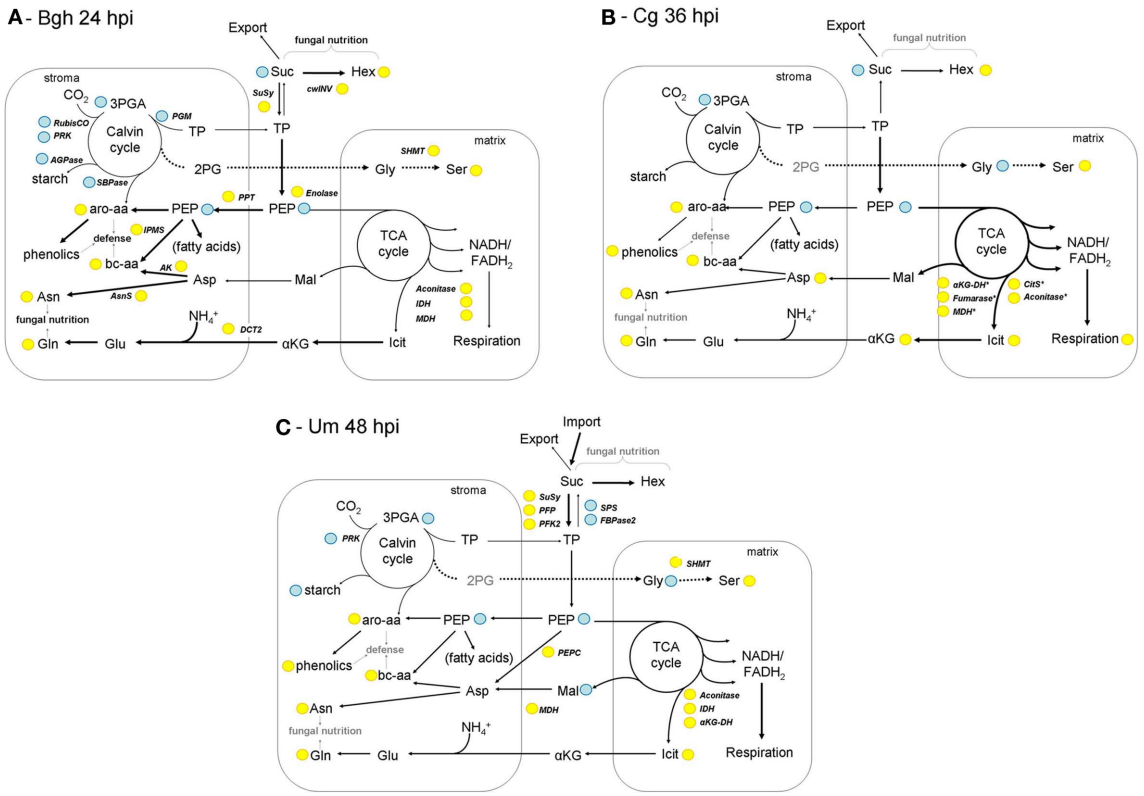


FIGURE 7 | Models of leaf metabolism during early interaction stages. Based on the results of the combined metabolome and transcriptome analysis, models illustrating the reprogramming of host metabolism during early biotrophic interactions are depicted for *Bgh* infected barley leaves at 24hpi **(A)**, *Cg* infected maize leaves at 36hpi **(B)** and *Um* infected maize leaves at 48hpi **(C)**. Please note that for simplicity, C_4 metabolism has been omitted from the maize models. Yellow – up compared to mock control; blue – down compared to mock control. Arrow thickness correlates with the proposed metabolic flux relative to the other depicted metabolic pathways. For explanations, please see the discussion text. Amino acids are abbreviated according to three letter code, 2PG, (2-phosphoglycolate); aKG, (α-ketoglutarate) Hex (hexoses); Icit, (isocitrate); PEP, (phospho*enol* pyruvate); 3-PGA, (3-phosphoglycerate); Suc, (sucrose); TP (triose phosphates); αKG-DH, (α-ketoglutarate dehydrogenase); AK, (aspartate kinase); AsnS, (asparagine synthetase); CitS, (citrate synthase); cw-INV, (cell wall invertase); DCT2, (dicarboxylate translocator); FBPase2, (fructose-2,6-bisphosphatase); IDH, (isocitrate dehydrogenase); IPMS, (isopropylmalate synthase); MDH, (malate dehydrogenase); PEPC, (PEP carboxylase); PFK2, (phosphofructokinase 2); PFP, (pyrophosphate-dependent phosphofructokinase); PPT, (phosphoenolpyruvate/phosphate translocator); SHMT, (serine hydroxymethyl transferase); SPS, (sucrose phosphate synthase); SuSy, (sucrose synthase).

in favor of glycolytic utilization of sucrose on the transcriptional level. As discussed by Horst et al. (2008) for *U. maydis*-induced tumors, the altered regulation of sucrose metabolism indicates that the developing leaves at 48hpi might represent *sink* characteristics and cover part of their carbohydrate budget by import of sugar (as indicated in **Figure 7C**). In this light, increased amino acid contents without clearly elevated supply of carbon moieties by the TCA cycle might also indicate that part of the free amino acid pool is replenished by import from systemic leaves (as discussed for tumors in Horst et al., 2010b).

ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG) in the framework of the priority program FOR 666.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/plant_physiology/10.3389/fpls.2011.00039/abstract/

Table S1 | Complete metabolome data and calculated metabolite ratios.

REFERENCES

Banuett, F. (1995). Genetics of *Ustilago maydis*, a fungal pathogen that induces tumors in maize. *Annu. Rev. Genet.* 29, 179–208.

Bergstrom, G. C., and Nicholson, R. L. (1999). The biology of corn anthracnose - knowledge for improved management. *Plant Dis.* 83, 596–608.

Biemelt, S., and Sonnewald, U. (2006). Plant–microbe interactions to probe the regulation of plant carbon metabolism. *J. Plant Physiol.* 163, 307–318.

Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B. J., Mendes, P., Roessner-Tunali, U., Beale, M. H., Trethewey, R. N., Lange, B. M., Wurtele, E. S., and Sumner, L. W. (2004). Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.* 9, 418–425.

Bolton, M. D. (2009). Primary metabolism and plant defense – fuel for the fire. *Mol. Plant Microbe Interact.* 22, 487–497.

Bonfig, K. B., Schreiber, U., Gabler, A., Roitsch, T., and Berger, S. (2006). Infection with virulent and avirulent *P. syringae* strains differentially affects photosynthesis and sink metabolism in *Arabidopsis* leaves. *Planta* 225, 1–12.

Both, M., Csukai, M., Stumpf, M. P. H., and Spanu, P. D. (2005). Gene expression profiles of *Blumeria graminis* indicate dynamic changes to primary

metabolism during development of an obligate biotrophic pathogen. *Plant Cell* 17, 2107–2122.

Chanda, B., Venugopal, S. C., Kulshrestha, S., Navarre, D. A., Downie, B., Vaillancourt, L., Kachroo, A., and Kachroo, P. (2008). Glycerol-3-phosphate levels are associates with basal resistance to the hemibiotrophic fungus *Colletotrichum higginsianum* in *Arabidopsis*. *Plant Physiol.* 147, 2017–2029.

Chaturvedi, R., Krothapalli, K., Makandar, R., Nandi, A., Sparks, A. A., Roth, M. R., Welti, R., and Shah, J. (2008). Plastid omega 3-fatty acid desaturase-dependent accumulation of a systemic acquired resistance inducing activity in petiole exudates of *Arabidopsis thaliana* is independent of jasmonic acid. *Plant J.* 54, 106–117.

Chen, L.-Q., Hou, B.-H., Lalonde, S., Takanaga, H., Hartung, M. L., Qu, X.-Q., Guo, W.-J., Kim, J.-G., Underwood, W., Chaudhuri, B., Chermak, D., Antony, G., White, F. F., Sommerville, S. C., Mudgett, M. B., and Frommer, W. B. (2010). Sugar transporters for intercellular exchange and nutrition of pathogens. *Nature* 468, 527–532.

Chou, H. M., Bundock, N., Rolfe, S. A., and Scholes, J. D. (2000). Infection of *Arabidopsis thaliana* leaves with *Albugo candida* (white blister rust) causes a reprogramming of host metabolism. *Mol. Plant Pathol.* 1, 99–113.

de Wit, P. J. G. M., Mehrabi, R., van den Burg, H. A., and Stergiopoulos, I. (2009). Fungal effector proteins: past, present and future. *Mol. Plant Pathol.* 10, 735–747.

Divon, H. H., and Fluhr, R. (2007). Nutrition acquisition strategies during fungal infection of plants. *FEMS Microbiol. Lett.* 266, 65–74.

Doehlemann, G., Wahl, R., Horst, R. J., Voll, L., Usadel, B., Poree, F., Stitt, M., Pons-Kuehnemann, J., Sonnewald, U., Kahmann, R., and Kämper, J. (2008a). Reprogramming a maize plant: transcriptional and metabolic changes induced by the fungal biotroph *Ustilago maydis*. *Plant J.* 56, 181–195.

Doehlemann, G., Wahl, R., Vranes, M., de Vries, R. P., Kämper, J., and Kahmann, R. (2008b). Establishment of compatibility in the *Ustilago maydis*/maize pathosystem. *J. Plant Physiol.* 165, 29–40.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS* 95, 14863–14868.

Essmann, J., Schmitz-Thom, I., Schon, H., Sonnewald, S., Weis, E., and Scharte, J. (2008). RNA interference-mediated repression of cell wall invertase impairs defense in source leaves of tobacco. *Plant Physiol.* 147,1288–1299.

Gerhardt, R., Stitt, M., and Heldt, H. W. (1987). Subcellular metabolite levels in spinach leaves: regulation of sucrose synthesis during diurnal alterations in photosynthetic partitioning. *Plant Physiol.* 83, 399–407.

Gibon, Y., Blaesing, O. E., Hannemann, J., Carillo, P., Höhne, M., Hendriks, J. H. M., Palacios, N., Cross, J., Selbig, J., and Stitt, M. (2004). A robot-based platform to measure multiple enzyme activities in *Arabidopsis* using a set of cycling assays: comparison of changes of enzyme activities and transcript levels during diurnal cycles and in prolonged darkness. *Plant Cell* 16, 3304–3325.

Hahn, M., Neef, U., Struck, C., Gottfert, M., and Mendgen, K. (1997). A putative amino acid transporter is specifically expressed in haustoria of the rust fungus *Uromyces fabae*. *Mol. Plant Microbe Interact.* 10, 438–445.

Heineke, D., Wildenberger, K., Sonnewald, U., Willmitzer, L., and Heldt, H.-W. (1994). Accumulation of hexoses in leaf vacuoles – studies with transgenic tobacco plants expressing yeast-derived invertase in the cytosol, vacuole or appoplasm. *Planta* 194, 29–33.

Herrmann, K. M., and Weaver, L. M. (1999). The shikimate pathway. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 50, 473–503.

Horst, R. J., Doehlemann, G., Wahl, R., Kahmann, R., Kämper, J., Sonnewald, U., and Voll, L. M. (2010a). *Ustilago maydis* infection strongly alters organic nitrogen allocation in maize and stimulates productivity of systemic source leaves. *Plant Physiol.* 152, 293–308.

Horst, R. J., Doehlemann, G., Wahl, R., Hofmann, J., Schmiedl, A., Kahmann, R., Kämper, J., and Voll, L. M. (2010b). A model of *Ustilago maydis* leaf tumor metabolism. *Plant Signal. Behav.* 5, 1446–1449.

Horst, R. J., Engelsdorf, T., Sonnewald, U., and Voll, L. M. (2008). Infection of maize leaves with *Ustilago maydis* prevents establishment of C-4 photosynthesis. *J. Plant Physiol.* 165, 19–28.

Hückelhoven, R., Fodor, J., Preis, C., and Kogel, K. H. (1999). Hypersensitive cell death and papilla formation in barley attacked by the powdery mildew fungus are associated with h2o2 accumulation but are not accompanied by enhanced concentrations of salicylic acid. *Plant Physiol.* 119, 1251–1260.

Jain, S. K., Langen, G., Hess, W., Borner, T., Huckelhoven, R., and Kogel, K.-H. (2004). The white barley mutant *albostrians* shows enhanced resistance to the biotroph *Blumeria graminis* f. sp. hordei. *Mol. Plant Microbe Interact.* 17, 374–382.

Kachroo, A., Lapchyk, L., Fukushige, H., Hildebrand, D., Klessig, D., and Kachroo, P. (2003). Plastidial fatty acid signaling modulates salicylic acid- and jasmonic acid-mediated defense pathways in the *Arabidopsis* ssi2 mutant. *Plant Cell* 15, 2952–2965.

Kahmann, R., and Kämper, J. (2004). *Ustilago maydis*: how its biology relates to pathogenic development. *New Phytol.* 164, 31–42.

Kämper, J., Kahmann, R., Bolker, M., Ma, L.-J., Brefort, T., Saville, B. J., Banuett, F., Kronstad, J. W., Gold, S. E., Muller, O., Perlin, M. H., Wosten, H. A. B., de Vries, R., Ruiz-Herrera, J., Reynaga-Pena, C. G., Snetselaar, K., McCann, M., Perez-Martin, J., Feldbrugge, M., Basse, C. W., Steinberg, G., Ibeas, J. I., Holloman, W., Guzman, P., Farman, M., Stajich, J. E., Sentandreu, R., Gonzalez-Prieto, J. M., Kennell, J. C., Molina, L., Schirawski, J., Mendoza-Mendoza, A., Greilinger, D., Munch, K., Rossel, N., Scherer, M., Vranes, M., Ladendorf, O., Vincon, V., Fuchs, U., Sandrock, B., Meng, S., Ho, E. C. H., Cahill, M. J., Boyce, K. J., Klose, J., Klosterman, S. J., Deelstra, H. J., Ortiz-Castellanos, L., Li, W., Sanchez-Alonso, P., Schreier, P. H., Hauser-Hahn, I., Vaupel, M., Koopmann, E., Friedrich, G., Voss, H., Schluter, T., Margolis, J., Platt, D., Swimmer, C., Gnirke, A., Chen, F., Vysotskaia, V., Mannhaupt, G., Guldener, U., Munsterkotter, M., Haase, D., Oesterheld, M., Mewes, H.-W., Mauceli, E. W., DeCaprio, D., Wade, C. M., Butler, J., Young, S., Jaffe, D. B., Calvo, S., Nusbaum, C., Galagan, J., and Birren, B. W. (2006). Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature* 444, 97–101.

Kocal, N., Sonnewald, U., and Sonnewald, S. (2008). Cell wall-bound invertase limits sucrose export and is involved in symptom development and inhibition of photosynthesis during compatible interaction between tomato and *Xanthomonas* campestris pv vesicatoria. *Plant Physiol.* 148, 1523–1536.

Leidreiter, K., Kruse, A., Heineke, D., Robinson, D. G., and Heldt, H.-W. (1995). Subcellular volumes and metabolite concentrations in potato (*Solanum tuberosum* cv. Désirée) leaves. *Bot. Acta* 108, 439–444.

Mendgen, K., and Hahn, M. (2002). Plant infection and the establishment of fungal biotrophy. *Trends Plant Sci.* 7, 352–356.

Molitor, A., Zajic, D., Voll, L. M., Pons-Kühnemann, J., Kogel, K. H., and Waller, F. (2011). Transcriptome and metabolite analysis of the barley-powdery mildew interaction reveals priming as a mechanism of *P. indica*-induced systemic resistance. *Mol. Plant Microbe Interact.* [Epub ahead of print].

Münch, S., Lingner, U., Floss, D. S., Ludwig, N., Sauer, N., and Deising, H. (2008). The hemibiotrophic lifestyle of *Colletotrichum* species. *J. Plant Physiol.* 165, 41–51.

Münch, S., Ludwig, N., Floß, D. S., Sugui, J. A., Koszucka, A. M., Voll, L. M., Sonnewald, U., and Deising, H. B. (2011). Identification of virulence genes in the corn pathogen *Colletotrichum graminicola* by *Agrobacterium tumefaciens*-mediated transformation. *Mol. Plant Pathol.* 12, 43–55.

Olea, F., Perez-Garcia, A., Canton, F. R., Rivera, M. E., Canas, R., Avila, C., Cazorla, F. M., Canovas, F. M., and de Vicente, A. (2004). Up-regulation and localization of asparagine synthetase in tomato leaves infected by the bacterial pathogen *Pseudomonas syringae*. *Plant Cell Physiol.* 45, 770–780.

Parker, D., Beckmann, M., Zubair, H., Enot, D. P., Caracuel-Rios, Z., Overy, D. P., Snowdon, S., Talbot, N. J., and Draper, J. (2009). Metabolomic analysis reveals a common pattern of metabolic re-programming during invasion of three host plant species by *Magnaporthe grisea*. *Plant J.* 59, 723–737.

Raffaele, S., Vailleau, F., Léger, A., Joubès, J., Miersch, O., Huard, C., Blée, E., Mongrand, S., Domergue, F., and Roby, D. (2008). A MYB transcription factor regulates very-long-chain fatty acid biosynthesis for activation of the hypersensitive cell death response in *Arabidopsis*. *Plant Cell* 20, 752–67.

Sana, T. R., Fischer, S., Wohlgemuth, G., Katrekar, A., Jung, K. H., Ronald, P. C., and Fiehn, O. (2010). Metabolomic and transcriptomic analysis of the rice response to the bacterial blight pathogen *Xanthomonas oryzae* pv. *oryzae*. *Metabolomics* 6,451–465.

Scharte, J., Schön, H., Tjaden, Z., Weis, E., and von Schaewen, A. (2009). Isoenzyme replacement of glucose-6-phosphate dehydrogenase in the cytosol improves stress tolerance in plants. *Proc. Natl. Acad. Sci. U.S.A.* 106, 8061–8066.

Schulze-Siebert, D., Heineke, D., Scharf, H., and Schulz, G. (1984). Pyruvate-derived amino acids in spinach chloroplasts: synthesis and regulation during photosynthetic carbon metabolism. *Plant Physiol.* 76, 465–71.

Siemens, J., González, M. C., Wolf, S., Hofmann, C., Greiner, S., Du, Y., Rausch, T., Roitsch, T., and Ludwig-Müller, J. (2011). Extracellular invertase is involved in the regulation

of clubroot disease in *Arabidopsis thaliana*. *Mol. Plant Pathol.* 12, 247–262.

Struck, C., Ernst, M., and Hahn, M. (2002). Characterization of a developmentally regulated amino acid transporter (AAT1p) of the rust fungus *Uromyces fabae*. *Mol. Plant Pathol.* 3, 23–30.

Struck, C., Mueller, E., Martin, H., and Lohaus, G. (2004). The *Uromyces fabae* UfAAT3 gene encodes a general amino acid permease that prefers uptake of in planta scarce amino acids. *Mol. Plant Pathol.* 5, 183–189.

Swarbrick, P. J., Schulze-Lefert, P., and Scholes, J. D. (2006). Metabolic consequences of susceptibility and resistance (race-specific and broad-spectrum) in barley leaves challenged with powdery mildew. *Plant Cell Environ.* 29, 1061–1076.

Tavernier, V., Cadiou, S., Pageau, K., Lauge, R., Reisdorf-Cren, M., Langin, T., and Masclaux-Daubresse, C. (2007). The plant nitrogen mobilization promoted by *Colletotrichum lindemuthianum* in *Phaseolus* leaves depends on fungus pathogenicity. *J. Exp. Bot.* 58, 3351–3360.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L. A., Rhee, S. Y., and Stitt, M. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37, 914–939.

Truman, W., de Zabala, M. T., and Grant, M. (2006). Type III effectors orchestrate a complex interplay between transcriptional networks to modify basal defence responses during pathogenesis and resistance. *Plant J.* 46, 14–33.

van Kan, J. A. L. (2006). Licensed to kill: the lifestyle of a necrotrophic plant pathogen. *Trends Plant Sci.* 11, 247–253.

Voegele, R. T., Struck, C., Hahn, M., and Mendgen, K. (2001). The role of haustoria in sugar supply during infection of broad bean by the rust fungus *Uromyces fabae*. *Proc. Natl. Acad. Sci. U.S.A.* 98, 8133–8138.

Voegele, R. T., Wirsel, S., Möll, U., Lechner, M., and Mendgen, K. (2006). Cloning and characterization of a novel invertase from the obligate biotroph *Uromyces fabae* and analysis of expression patterns of host and pathogen invertases in the course of infection. *Mol. Plant Microbe Interact.* 19, 625–634.

Wahl, R., Wippel, K., Goos, S., and Kämper, J. (2010). Norbert Sauer a novel high-affinity sucrose transporter is required for virulence of the plant pathogen *Ustilago maydis*. *PLoS Biol.* 8, e1000303. doi: 10.1371/journal.pbio.1000303

Ward, J. L., Forcat, S., Beckmann, M., Bennett, M., Miller, S. J., Baker, J. M., Hawkins, N. D., Vermeer, C. P., Lu, C., Lin, W., Truman, W. M., Beale, M. H., Draper, J., Mansfield, J. W., and Grant, M(2010). The metabolic transition during disease following infection of *Arabidopsis thaliana* by *Pseudomonas syringae* pv. tomato. *Plant J.* 63, 443–457.

Wiberg, A. (1974). Genetical studies of spontaneous sources of resistance to powdery mildew in barley. *Hereditas* 77, 89–148.

Widarto, H. T., Van Der Meijden, E., Lefeber, A. W., Erkelens, C., Kim, H. K., Choi, Y. H., and Verpoorte, R. (2006). Metabolomic differentiation of *Brassica rapa* following herbivory by different insect instars using two-dimensional nuclear magnetic resonance spectroscopy. *J. Chem. Ecol.* 32, 2417–2428.

Conflict of Interest Statement: 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.

Received: 19 May 2011; accepted: 01 August 2011; published online: 22 August 2011.
*Citation: Voll LM, Horst RJ, Voitsik A-M, Zajic D, Samans B, Pons-Kühnemann J, Doehlemann G, Münch S, Wahl R, Molitor A, Hofmann J, Schmiedl A, Waller F, Deising HB, Kahmann R, Kämper J, Kogel K-H and Sonnewald U (2011) Common motifs in the response of cereal primary metabolism to fungal pathogens are not based on similar transcriptional reprogramming. Front. Plant Sci. **2**:39. doi: 10.3389/fpls.2011.00039*
This article was submitted to Frontiers in Plant Physiology, a specialty of Frontiers in Plant Science.

frontiers in
PLANT SCIENCE

METHODS ARTICLE
published: 22 August 2011
doi: 10.3389/fpls.2011.00040

Mass spectra-based framework for automated structural elucidation of metabolome data to explore phytochemical diversity

Fumio Matsuda[1,2], Ryo Nakabayashi[1], Yuji Sawada[1,3], Makoto Suzuki[1], Masami Y. Hirai[1,3], Shigehiko Kanaya[1,4] and Kazuki Saito[1,5]*

[1] *RIKEN Plant Science Center, Yokohama, Japan*
[2] *Organization of Advanced Science and Technology, Kobe University, Kobe, Japan*
[3] *Japan Science and Technology Agency, CREST, Kawaguchi, Japan*
[4] *Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan*
[5] *Graduate School of Pharmaceutical Sciences, Chiba University, Chiba, Japan*

Edited by:
Alisdair Fernie, Max Planck Institut for Plant Physiology, Germany

Reviewed by:
Asaph Aharoni, Weizmann Institute of Science, Israel
Takayuki Tohge, Max Planck Institute of Molecular Plant Physiology, Germany

****Correspondence:***
Kazuki Saito, Metabolome Research Group, RIKEN Plant Science Center, Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan.
e-mail: ksaito@psc.riken.jp

A novel framework for automated elucidation of metabolite structures in liquid chromatography–mass spectrometer metabolome data was constructed by integrating databases. High-resolution tandem mass spectra data automatically acquired from each metabolite signal were used for database searches. Three distinct databases, KNApSAcK, ReSpect, and the PRIMe standard compound database, were employed for the structural elucidation. The outputs were retrieved using the CAS metabolite identifier for identification and putative annotation. A simple metabolite ontology system was also introduced to attain putative characterization of the metabolite signals. The automated method was applied for the metabolome data sets obtained from the rosette leaves of 20 *Arabidopsis* accessions. Phenotypic variations in novel *Arabidopsis* metabolites among these accessions could be investigated using this method.

Keywords: metabolome analysis, liquid chromatography-mass spectrometry, structural elucidation, database searching, natural variations in secondary metabolite

INTRODUCTION

The ability to produce various secondary metabolites has evolved in plants for the purpose of self-defense, environmental adaptation, and interaction with other organisms. Because humans utilize phytochemicals as a rich resource for various purposes such as the production of pharmaceuticals, further understanding of the genetic background behind the diversity of secondary metabolites produced by plants will facilitate more intensive application of these compounds (Saito and Matsuda, 2010). Recent progress in gene sequencing has enabled generation of a large volume of data on genetic polymorphisms that is related to natural variations in phytochemicals (Clark et al., 2007; Ossowski et al., 2008; Zeller et al., 2008). Accordingly, it is expected that novel genes and functions of plant secondary metabolism as well as those involved in evolution could be investigated based on the association between genotypes and metabolic phenotypes (metabolotypes; Plantegenet et al., 2009; Weigel and Mott, 2009). Since the metabolotype data required for such analyses is both qualitative (structure of secondary metabolites) and quantitative (amount of metabolite), metabolic profiling analysis using liquid chromatography–tandem mass spectrometry (LC–MS) has been used to obtain comprehensive profiles of plant secondary metabolites (De Vos et al., 2007). While qualitative data describing hundreds of metabolite signals have routinely been acquired during analysis (Keurentjes et al., 2006), structural elucidation of the observed signals using LC–MS is still difficult (Moco et al., 2006; Bottcher et al., 2007; Iijima et al., 2008; Matsuda et al., 2010a).

The structure-related information available from LC–MS analysis includes the retention time, exact mass number, and tandem mass spectrum (MS/MS spectrum). The structure associated with each metabolite signal has been estimated by searching databases containing reference data using the information obtained from LC–MS analysis (Moco et al., 2007; Kind and Fiehn, 2010; Neumann and Bocker, 2010). The amount of information obtained from database searches varies among metabolite peaks; therefore, four levels of structural elucidation have been standardized by the metabolome standard initiative (MSI) as follows (Fiehn et al., 2007; Sumner et al., 2007): (1) Identified: a minimum of two independent data points relative to an authentic compound analyzed under identical experimental conditions. (2) Putatively annotated: without chemical reference standards, based on physicochemical properties and/or spectral similarity with public/commercial spectral libraries. (3) Putatively characterized: based on characteristic physicochemical properties of a chemical class of compounds, or spectral similarity to known compounds of a chemical class. (4) Unknown. Based on the standardized format, a framework for automated structural elucidation is required to explore the structural diversity of phytochemicals. However, several technical issues must be solved before database-assisted elucidation of metabolite structures (Kind and Fiehn, 2010; Neumann and Bocker, 2010). One bottleneck is represented by a shortage of standard compounds and their associated MS/MS spectra data. Owing to the poor availability of plant secondary metabolites, only a very low percentage of the observed metabolite signals can be assigned by comparison of the chromatographic behavior

with chemical reference standards (Matsuda et al., 2010a). Although great effort has been put into construction of the MS/MS spectral databases (Moco et al., 2006; Wishart et al., 2007; Horai et al., 2010), further enrichment is required for structural elucidation of the wider range of metabolites. Another difficulty is the low reproducibility of the structure-related information. For instance, the fragment patterns in MS/MS spectra depend on the mass spectrometers and their operating conditions. The error derived from the analysis also exists in the high-resolution mass spectral data (Mihaleva et al., 2008; Matsuda et al., 2009b). Owing to these technical problems, elucidation of the structure associated with signals corresponding to metabolomes is time consuming, which has hampered the investigation of phytochemical diversity across plant species or ecotypes.

In this study, a novel framework for the automated elucidation of metabolite structures in LC–MS metabolome data was constructed by integrating three different databases. To overcome the aforementioned problems, the MS/MS spectra databases were enriched using literature reported information. Additionally, the high-resolution MS/MS spectra data were redundantly acquired from each metabolite signal to improve the quality of structure-related information that was used to search the databases. The outputs were retrieved using the CAS metabolite identifier for identification and putative annotation. A simple metabolite ontology system was also introduced to enable putative characterization of the metabolite signals. The automated method developed here was applied for metabolome data sets obtained from the rosette leaves of 20 *Arabidopsis* accessions, from which phenotypic variations in novel *Arabidopsis* metabolites among these accessions could be investigated.

MATERIALS AND METHODS

PLANT MATERIALS

Seeds of 20 accessions of *Arabidopsis thaliana*, CS22676 Bay-0, CS22677 Bor-4, CS22678 Br-0, CS22679 Bur-0, CS22680 C24, CS22681 Col-0, CS22682 Cvi-0, CS22683 Est-1, CS22684 Fei-0, CS22685 Goettingen-7, CS22686 Ler-1, CS22687 NFA-8, CS22688 RRS-7, CS22689 RRS-10, CS22690 Sha, CS22691 Tamm-2, CS22692 Ts-1, CS22693 Tsu-1, CS22694 Van-0, and CS22695 Lov-5, were obtained from the ABRC. The seeds were soaked on MS agar plates and then incubated at 22°C under 16 h day and 8 h night conditions. At 18 days after germination, the aerial parts of the seedlings were harvested.

METABOLOME ANALYSIS USING LC-ESI-Q-Tof/MS

The collected sample tissues were weighed and stored at −80°C until analysis. The frozen tissues of independent plants were homogenized in five volumes of 80% aqueous methanol containing 0.1% acetic acid, 0.5 mg/l of lidocaine, and *d*-camphor sulfonic acid (Tokyo Kasei, Tokyo, Japan) using a mixer mill (MM 300, Retsch) with a zirconia bead for 6 min at 20 Hz. Next, the samples were centrifuged at 15,000 g for 10 min and filtered (Ultrafree-MC filter, 0.2 μm; Millipore, Bedford, MA, USA). The sample extracts were then applied to an HLB μElution plate (Waters, Milford, MA, USA) that had been equilibrated with 80% aqueous methanol containing 0.1% acetic acid. The eluates (3 μl) were subsequently subjected to metabolome analysis by LC coupled with electrospray quadrupole time-of-flight tandem MS using an Acquity BEH ODS column (LC-ESI-Q-Tof/MS, HPLC: Waters Acquity UPLC system; MS: Waters Q-Tof Premier). The metabolome analysis and data processing were conducted according to a previously described method (Matsuda et al., 2009c, 2010a). Briefly, the metabolome data were obtained in the negative ion mode (*m*/*z* 100–2,000; dwell time: 0.45 s; interscan delay: 0.05 s, centroid), from which a data matrix was generated with the aid of MetAlign (De Vos et al., 2007; Lommen, 2009). In order to reduce a redundancy of the data matrix, fragment ions were removed by a following procedure. A metabolite signal was removed from the matrix when there is another intense peak eluted at similar retention times [within the retention time threshold (<0.5 s)] with the highest correlation coefficient above the threshold value (>0.8). The analysis was conducted using five biological replicates of 20 accessions, from which a data matrix composed of 703 signals (peaks) was obtained (**Table S1** in Supplementary Material). The number of signals would not reflect an exact number of detected metabolites due to the complex nature of the metabolome data.

To construct MS2T libraries, the extracts of five ecotypes were mixed and utilized for the MS2T data acquisition. The analyses were repeatedly conducted for four mixtures by previously described methods (Matsuda et al., 2009c). Each MS2T entry was assigned a unique accession code, such as ATH10n03690, in which ATH10n is the name of the library and 03690 is the entry number. All data obtained in this study are available at the PRIMe website[1] (Akiyama et al., 2008).

DATABASES AND SOFTWARE

The ReSpect (RIKEN MS/MS spectra database for phytochemicals; 2011 January version), KNApSAcK (2010.12.24 version; Shinbo et al., 2006; Takahashi et al., 2008), and PRIMe standard compound database (2009 November version) were used in this study. The genetic polymorphism data from 20 *Arabidopsis* accessions were downloaded from the TAIR web site (Clark et al., 2007; Poole, 2007). All data processing procedures were conducted using the in-house script written with Perl. Structural elucidation work was performed in-batch search for all metabolite signals.

In the automated structural elucidation procedure, several thresholds were required to conduct the database searches. The thresholds used in this study are described in **Figures 2 and 3**. To search the MS/MS spectra, the similarity scores were determined by employing dot product method with mass tolerance at 0.5 Da (Stein and Scott, 1994). The two spectra were considered to be the similar when the similarity score was greater than 0.6. For hierarchical clustering analysis, log2-transformed *Z*-scored signal intensity data were processed using MEV version 4.4 (Saeed et al., 2003, 2006).

RESULTS

ACQUISITION OF METABOLOME DATA FROM 20 *ARABIDOPSIS* ACCESSIONS

To investigate variations in the composition of secondary metabolites among *Arabidopsis* strains (accessions), metabolic profile data were obtained from the rosette leaves of 20 accessions of *Arabidopsis* by LC-ESI-Q-Tof/MS analysis (Matsuda et al., 2009c, 2010a). The 20 diverse accessions evaluated herein were previously selected by Clark et al. (2007) to investigate the genetic variations within the popula-

[1]http://prime.psc.riken.jp/

tion of *Arabidopsis*. The analysis was conducted using five biological replicates of 20 accessions, from which a data matrix composed of 703 metabolite signals (peaks) was obtained (**Table S1** in Supplementary Material). Here, the dataset was designed as AtMetExpress 20 Ecotypes and each metabolite signal was addressed by a unique ID, such as aen00884. Hierarchical clustering analysis of the dataset revealed that there were large variations in the metabolic profiles across 20 accessions, which should be derived from those genetic polymorphisms (**Figure 1**). To acquire information for structural elucidation of those metabolite signals, MS/MS spectra data were obtained from identical extracts by using the automated data acquisition methods described in Section "Materials and Methods." Since the analyses were conducted repeatedly, multiple MS/MS spectra data were recorded for each metabolite signal (Matsuda et al., 2009c). Consequently, MS/MS spectral tag (MS2T) libraries containing 126,889 accessions were constructed (**Table 1**). Each MS2T entry was assigned a unique ID such as ATH67n06391. Based on the MS2T data, the structure of each metabolite signal was elucidated by searching the databases.

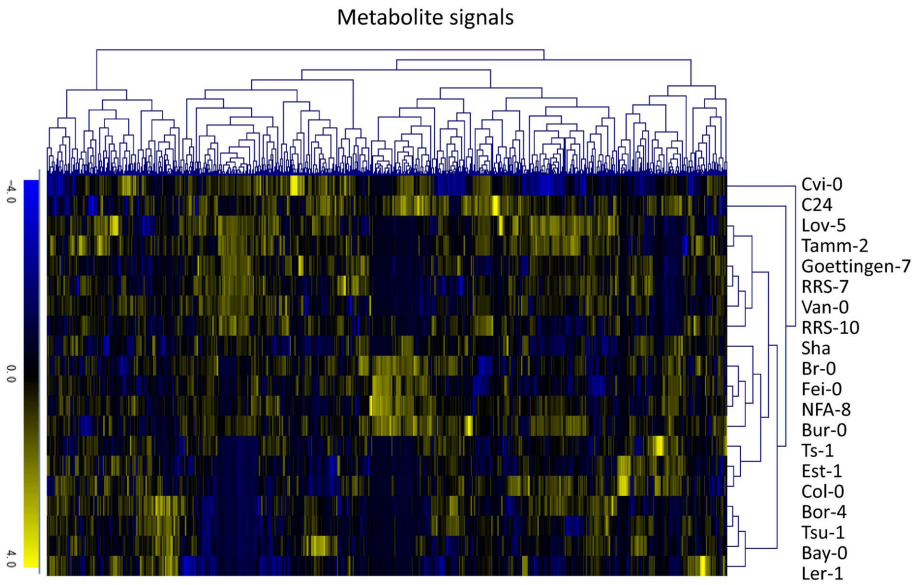


FIGURE 1 | Hierarchical clustering analysis of metabolic profile data of the AtMetExpress 20 ecotype dataset. Log2-transformed and *Z*-scored signal intensity data were hierarchically classified using the average linkage clustering methods.

Table 1 | List of databases and datasets used in this study.

Databases	Description	Number of accessions	Data source
AtMetExpress 20 ecotype	Metabolic profile data obtained from 20 accessions of *Arabidopsis* strains	100 metabolic profile data (20 accessions by five biological replicates) containing 703 metabolite signals	http://prime.psc.riken.jp/?action = drop_index
MS2T library	Library of high-resolution MS/MS spectra data obtained from the actual *Arabidopsis* extracts	Subset of MS2T library containing 126,889 accessions obtained from the *Arabidopsis* ecotypes were used in this study	http://prime.psc.riken.jp/lcms/ms2tview/ms2tview.html
ReSpect for phytochemicals	MS/MS spectra database of standard and literature reported phytochemicals	Literature data: 3,136 records corresponding to 2,741 metabolites Q-TOF/MS data 1,050 records/575 standard compounds QqQ/MS data: 4,258 records/861 standards. Total 8,444 records/3,595 metabolites	http://spectra.psc.riken.jp/
RIKEN Standard compound database	List of standard compounds and physicochemical data	LC–MS/MS retention time and *m*/*z* data of 600 compounds	http://prime.psc.riken.jp/lcms/data/StandardCompound/
KNApSAcK	Comprehensive species–metabolite relationship database	Collection of 50,048 unique metabolites and 101,500 metabolite–species pairs	http://kanaya.naist.jp/KNApSAcK_Family/
Metabolite ontology	Simple classification of phytochemicals	322 ontology terms are assigned for the ReSpect database	In preparation

PREPARATION OF STANDARD COMPOUND DATABASES AND THE COMPOUND ONTOLOGY SYSTEM

Three distinct databases, KNApSAcK, ReSpect, and the PRIMe standard compound database, were employed for the structural elucidation (**Table 1**). ReSpect is a new web data resource that incorporates records from existing literature as well as the MS/MS data from our standard compounds. This database contains 8,444 records corresponding to 3,595 metabolites. ReSpect is the first tool for annotation of phytochemicals that is based on downloadable MS/MS data resources and databases (Sawada et al., in preparation). KNApSAcK is a comprehensive species–metabolite relationship database developed by the Kanaya lab in NAIST (Shinbo et al., 2006; Takahashi et al., 2008). KNApSAcK contains the structural data of 50,048 metabolites and 101,500 metabolite–species pairs. In this study, KNApSAcK was used to elucidate molecular formulas of candidate metabolites from the high-resolution mass spectra data. The PRIMe standard compound database contains a retention time and *m*/*z* data of 600 authentic compounds acquired using an identical analytical method (Matsuda et al., 2009c). For the automated metabolite annotations, accessions in these databases were assigned with corresponding CAS identifiers.

Since CAS identifiers basically address a structurally confirmed metabolites (Matsuda et al., 2009a), the metabolite annotation procedure based on the identifier cannot deal with information describing partially characterized metabolites. For example, the metabolite structures were often estimated to be from a compound class such as "kaempferol glycoside" and "amino acid derivative" (Bottcher et al., 2007; Iijima et al., 2008; Matsuda et al., 2010a). In the case of gene annotation, each gene was tentatively annotated by gene ontology terms that were manually assigned or automatically estimated from the sequence similarities. Although detailed compound ontology systems and vocabularies have been developed using several databases such as CheBi and KEGG (Degtyarenko et al., 2008; Kanehisa et al., 2008; Matsuda et al., 2009a), a simple compound ontology system was newly introduced in this study to cover the wide range of phytochemicals. Here entries in the PRIMe databases were classified within three levels, ranging from basic (Class 1) to detailed (Class 3) with considering the basic skeleton and modified parts of metabolites (**Table S2** in Supplementary Material). The ontology terms prepared in this study is not comprehensive, since the classification system was arbitrary prepared by manually curating the entries of ReSpect MS/MS spectra database for an assistance of structural elucidation of metabolome data. For instance, partially characterized metabolites could be classified as follows: kaempferol-3,7-dirhamnoside is a member of Class 1: flavonoid, Class 2: flavonol, and Class 3: kaempferol glycoside; tryptophan is a member of Class 1: amino acid and Class 2: tryptophan; and pinoresinol-dihexoside is a member of Class 1: phenylpropanoid, Class 2: lignan, and Class 3: pinoresinol glycoside.

These metabolite classifications have been assigned to all accessions in the ReSpect and PRIMe standard compound databases. A detailed classification study is currently in progress for KNApSAcK, and 60% of the accessions in this database have been assigned to Class 1 or 2.

IDENTIFICATION AND PUTATIVE ANNOTATION USING CAS IDENTIFIERS

Based on the MS2T libraries and reference databases, the metabolite signals in the AtMetExpress 20 ecotypes dataset were identified or putatively annotated using the following automated procedure. For the case of a metabolite signal assigned as aen00884 (Rt 3.82 and *m*/*z* 422), the MS2T library contains 19 MS2T accessions acquired from the identical metabolite signal with various spectral quality (**Figure 2A**). In other words, the metabolite signal was tagged with 19 accessions of corresponding MS2Ts. Each MS2T accession consists of the exact mass number of the precursor ion and MS/MS spectra data. Thus, MS/MS spectra data were submitted to the ReSpect database to identify metabolites producing similar MS/MS spectra. In the case of the MS2T accession, ATH67n06391, the MS/MS spectrum was similar to that of 13 compounds whose CAS numbers are obtained as search results (**Figure 2B**). Additionally, the exact mass number of the precursor ion was used to search the KNApSAcK database to find metabolites possessing a highly similar mass number, by which the CAS number of 1 metabolite was obtained. A common CAS number (499-30-9, 2-phenylglucolsinolate) observed in the outputs of both the ReSpect and KNApSAcK searches indicated that it is a candidate structure of the metabolite signal deduced from the MS2T data. To improve the search quality, the procedure was repeated for all 19 MS2T accessions, and the same results were observed for 11 MS2Ts. Since identical metabolites were elucidated using two distinct search methods with high reproducibility (>50%), it is likely that the metabolite signal was derived from 2-phenylethylglucosinolate or its structural isomers. Based on the MSI standard, the metabolite signal could be putatively annotated using the automated structure elucidation procedure (**Figure 2D**).

Furthermore, an automated search of the PRIMe standard compound database revealed that the authentic compound of 2-phenylethylglucosinolate, CAS 499-30-9, was also detected at a similar retention time and mass number as the queried metabolite signal. Since three distinct pieces of information, including the MS/MS spectra, exact mass number, and chromatographic behavior, were matched to the identical metabolite, the metabolite signal was identified as 2-phenylethylglucosinolate (**Figure 2F**).

Among the 703 metabolite signals in the AtMetExpress 20 ecotype dataset, 25 and 106 peaks could be identified and putatively annotated, respectively, using the procedure described above (**Table S1** in Supplementary Material). Additionally, comparison with the manually curated results produced in our previous study (Matsuda et al., 2010a) revealed no significant error among the 32 commonly annotated metabolite signals.

PROCESSING OF PUTATIVELY CHARACTERIZED METABOLITES

In addition to the identification and putative annotation using the CAS metabolite identifiers, putative characterization of the metabolite signals was conducted by introducing the metabolite ontology system. The procedure is explained using the metabolite signal described above as an example (peak ID: aen008844). For each MS2T accession tagged to the metabolite signal, MS/MS spectra data and the exact mass number were used for ReSpect (**Figure 3A**) and KNApSAcK (**Figure 3B**) searches. The compound ontology information instead of CAS identifiers was obtained as outputs in these procedures. The outputs of KNAsSAcK and ReSpect searches were compared to identify a common result, which is a compound ontology estimated from the MS2T accession. Repeated searching for 19 MS2T accessions of aen008844 resulted in 11 MS2Ts being identified as glucosinolate based on

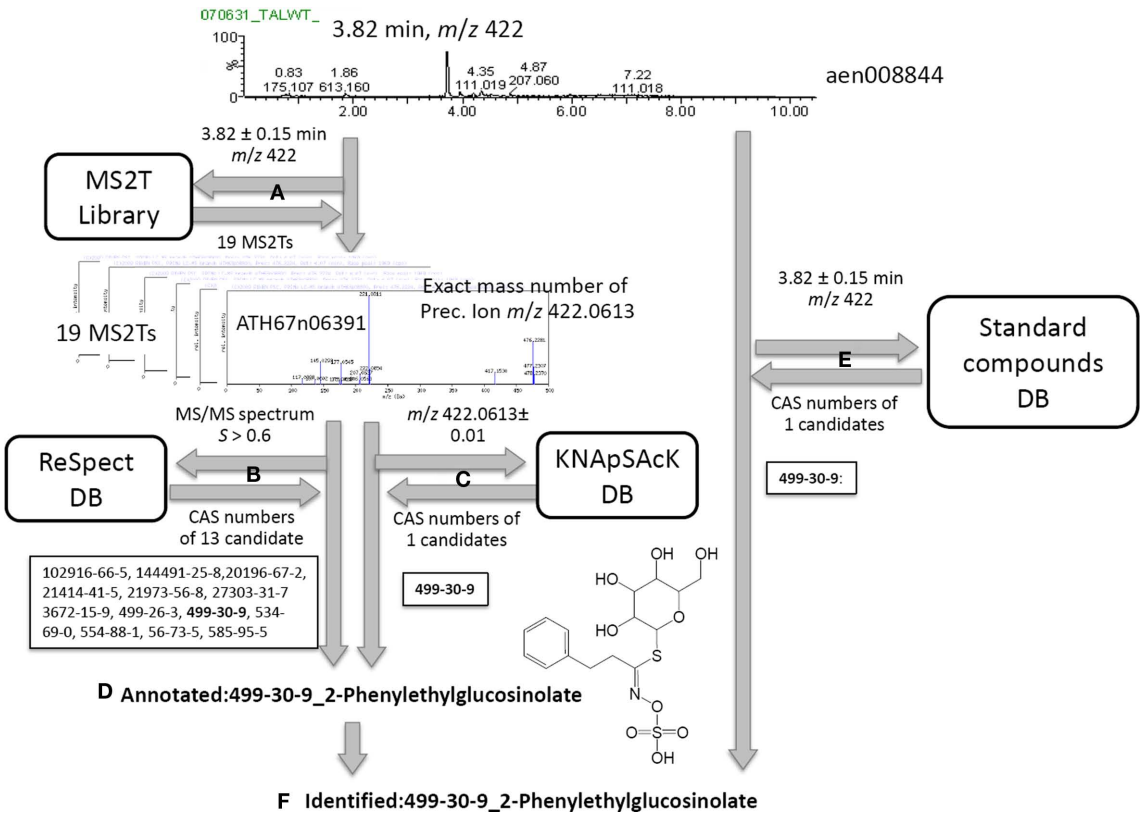


FIGURE 2 | Procedure for peak identification and putative annotation using CAS identifiers. For the case of a metabolite signal, aen00884, the MS2T library contains 19 MS2T accessions acquired from the identical metabolite signal **(A)**. MS/MS spectra data in MS2T were submitted to the ReSpect database **(B)**. The exact mass number data of the precursor ion was used to search the KNApSAcK database **(C)**. Since a common CAS number (499-30-9, 2-phenylglucolsinolate) in the outputs of both searches were observed for 11 MS2Ts, the metabolite signal was putatively annotated as 2-phenylethylglucosinolate **(D)**. Since the authentic compound of 2-phenylethylglucosinolate,CAS 499-30-9, was also detected at a similar retention time and mass number **(E)**, the metabolite signal was identified as 2-phenylethylglucosinolate **(F)**.

the Class 1 ontology. The Class 2 ontology benzylglucosinolate was not accepted, because the result was estimated from only 2 MS2T accessions. Using the procedure, the metabolite signal was successfully characterized as glucosinolate based on the Class 1 ontology (**Figure 3C**).

This procedure was conducted for all metabolite signals of the AtMetExpress 20 Ecotype dataset, and 188 among 703 metabolite signals were automatically characterized. In the case of Class 1 ontology, 1 alkaloids, 7 amino acids, 33 flavonoids, 68 glucosinolates, 47 phenylpropanoids, 4 terpenoids, and 28 other characterizations were assigned to the metabolome data (**Table S2** in Supplementary Material).

STRUCTURAL ELUCIDATION OF METABOLITE SIGNALS USING THE DATABASE SEARCH RESULTS

Based on the results obtained using the automated methods, the structures of the novel *Arabidopsis* metabolites were manually elucidated. Among the putatively characterized metabolite signals, the metabolite signal aen006966 (Rt 4.051 min and *m/z* 369) was putatively characterized as being in Class 1: phenylpropanoid. The MS/MS spectral data for ATH67n05643 (**Figure 4A**) indicated that the metabolite would be a coumarin hexoside based on the fragment pattern. An additional KNApSAcK search suggested that a plausible candidate of the metabolite is fraxin (CAS 524-32-1), although the position of glycosylation is unclear. Using a similar procedure, a metabolites putatively characterized as Class 1: phenylpropanoid (aen012096: Rt 3.848 min, *m/z* 501) were found to be malonyl-hexosyl-sinapate (**Figure 4B**).

Structural elucidation of the putative phenylpropanoid aen006925 (Rt 5.762, *m/z* 367) indicated that this metabolite is a hexoside of an unknown aglycon (**Figure 4C**). Because the molecular formula of the aglycone was deduced to be $C_{11}H_9O_4$ (*m/z* 205.0502 *obsd*, *m/z* 205.0500 *theor*), the aglycone should be a methylated hydroxy-coumarin (according to the presence of four oxygen atoms, aglycone should contain at least two hydroxy-groups on the coumarin moiety), or dimethoxycoumarin. Thus, the compound aen006925 can be a glycoside (or *C*-glycoside) of these two aglycones, both of which are novel *Arabidopsis* metabolites. While strict structural elucidation must be conducted following the protocols accepted for natural product chemistry (Nakabayashi et al., 2009; Matsuda et al., 2010b), the results presented here demonstrate that a portion of the phytochemical diversity in *Arabidopsis* could be elucidated from MS/MS spectra via automated structural elucidation.

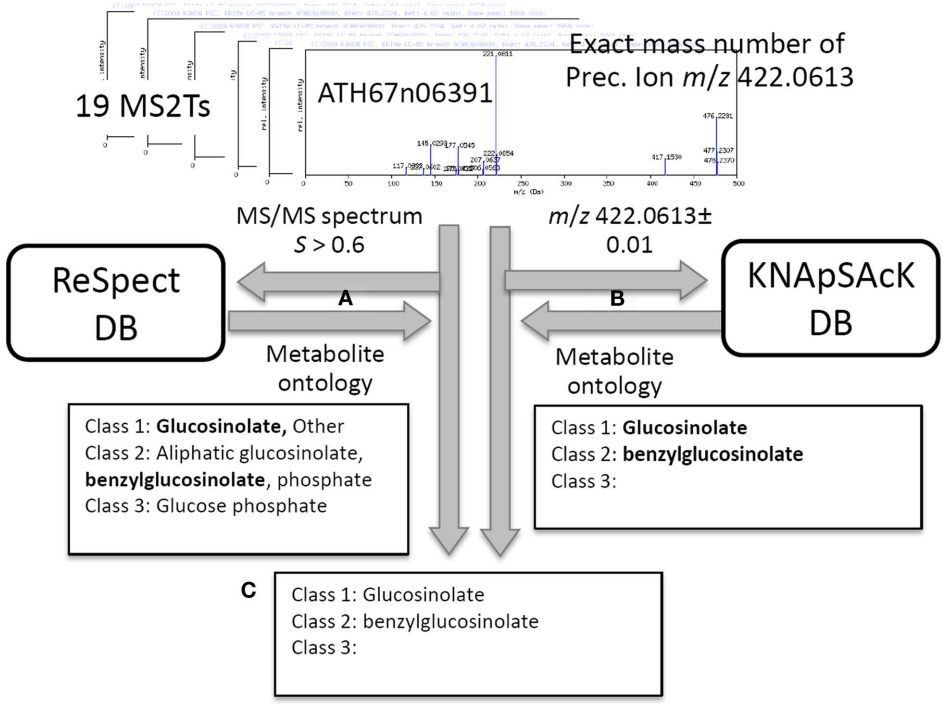


FIGURE 3 | Procedure for putative characterization of metabolite signal using metabolite ontology. For 19 MS2T accessions tagged to the metabolite signal, aen008844, MS/MS spectra data and the exact mass number were used for ReSpect **(A)** and KNApSAcK **(B)** searches. The compound ontology information in KNAsSAcK and ReSpect searches were compared to identify a common result. Repeated searching for 19 MS2T accessions resulted in 11 MS2Ts being identified as glucosinolate based on the Class 1 ontology **(C)**.

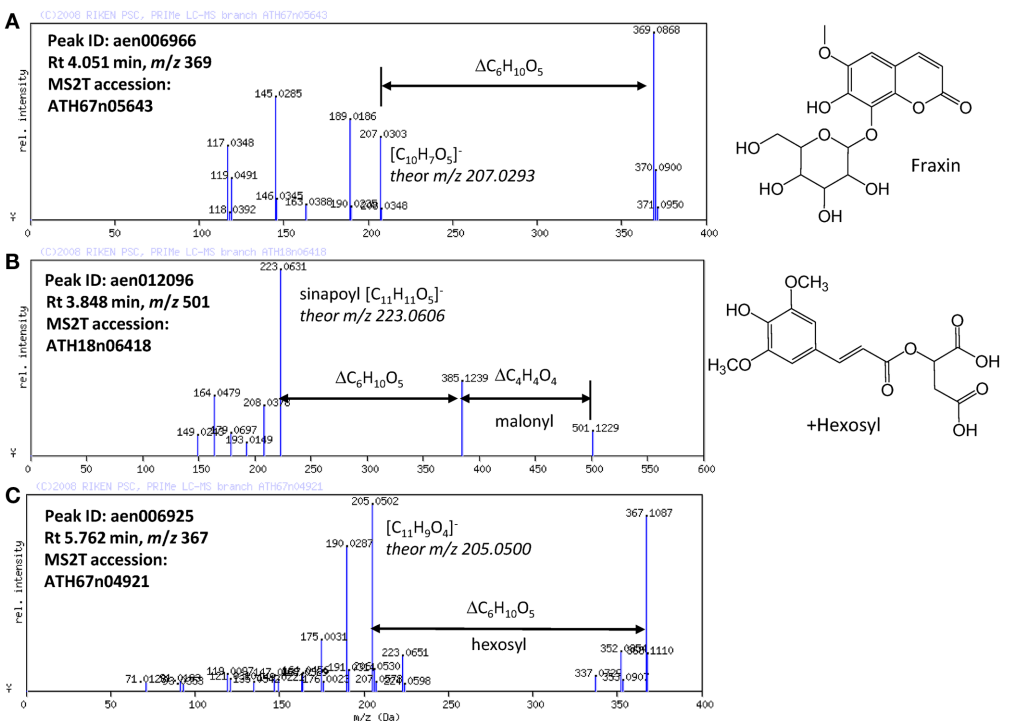


FIGURE 4 | MS/MS spectra of putatively characterized metabolites. The predicted molecular formulas of key fragments, neutral losses, and elucidated structures of **(A)** putative fraxin, **(B)** hexosylsinapoylmalate, and **(C)** hexosyl-coumarin are shown in the figure.

PHENOTYPIC VARIATIONS ACROSS *ARABIDOPSIS* ACCESSIONS

The structural elucidation based on the compound ontology information enabled us to deal with putatively characterized metabolite signals such as glucosinolates and flavonoids without strict metabolite identification or annotation. Here, the natural variations in accumulation levels among 20 *Arabidopsis* accessions were compared for metabolites belonging to lignan, amino acids, flavonoids, and glucosinolate (**Figure 5**). The metabolites assigned by lignan (**Figure 5A**) and amino acid (**Figure 5B**) were constitutively accumulated with small natural variations, suggesting that the production of those metabolites is essential for *Arabidopsis* (Matsuda et al., 2010a). Indeed, more than 10 genes encoding the dirigent protein for lignan biosynthesis are present in the *Arabidopsis* genome (Burlat et al., 2001; Davin and Lewis, 2005; Nakatsubo et al., 2008). This redundancy would contribute to the constitutive production of lignans, although the details regarding their physiological role in the growth of *Arabidopsis* remain unknown. In contrast, the metabolites identified as flavonoids (**Figure 5C**) and glucosinolates (**Figure 5D**) tended to show larger natural variations among the 20 accessions. These results suggest that the levels of flavonoids and glucosinolates in rosette leaves are controlled by genetic polymorphisms, which would contribute to the adaptation of each accession to local environments (Li et al., 2008; Bednarek and Osbourn, 2009; Janowitz et al., 2009; Sawada et al., 2009; Manzaneda et al., 2010; De Kraker and Gershenzon, 2011). To investigate the association between large variations in metabolic phenotypes and genetic polymorphisms, we considered the levels of 3-hydroxy-*n*-propylglucosinolate (aen007244) among 20 accessions. Despite significant production of Bor-4, Tsu-1, Bay-0, and Ler-1, the glucosinolate was not detected from other accessions, including Col-0 (**Figure 6A**). Single nucleotide polymorphisms (SNPs) that commonly occurred in Bor-4, Tsu-1, Bay-0, and Ler-1, as well as did not occurred in other accessions, were searched against the re-sequence data produced by Clark et al. (2007). The results revealed that 80 SNPs of 96 corresponding SNPs formed a linkage disequilibrium (LD) block along the long arm of chromosome 4 (**Figure 6B**). Among the 28 ORFs in the 11-kb region (from At4g02870 to At4g03090), there is an enzyme gene responsible

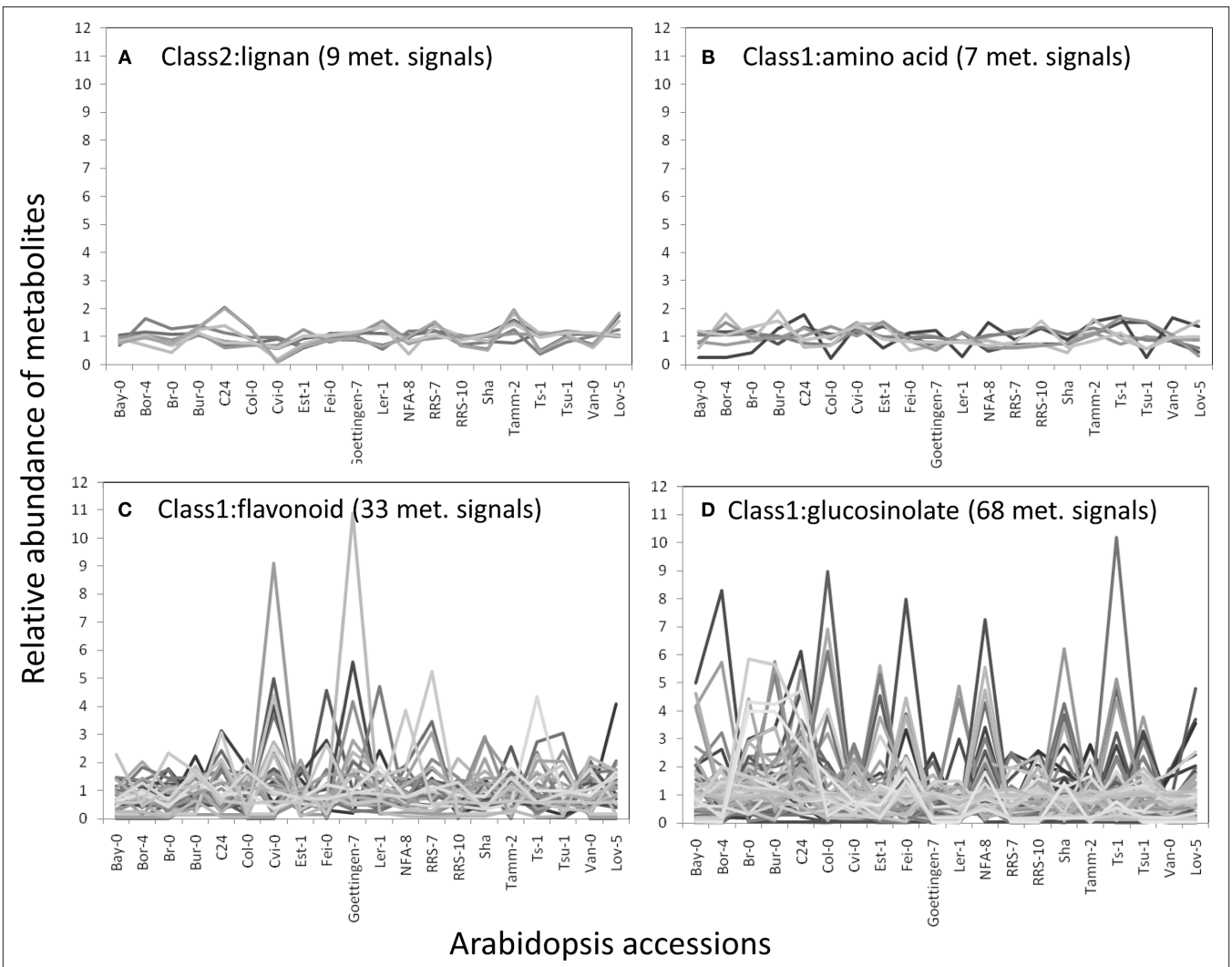


FIGURE 5 | Natural variation in metabolite levels among 20 *Arabidopsis* accessions belonging to lignan (A), amino acid (B), flavonoid (C), and glucosinolate (D). The relative abundances of metabolites were determined by dividing each metabolite level by the average level.

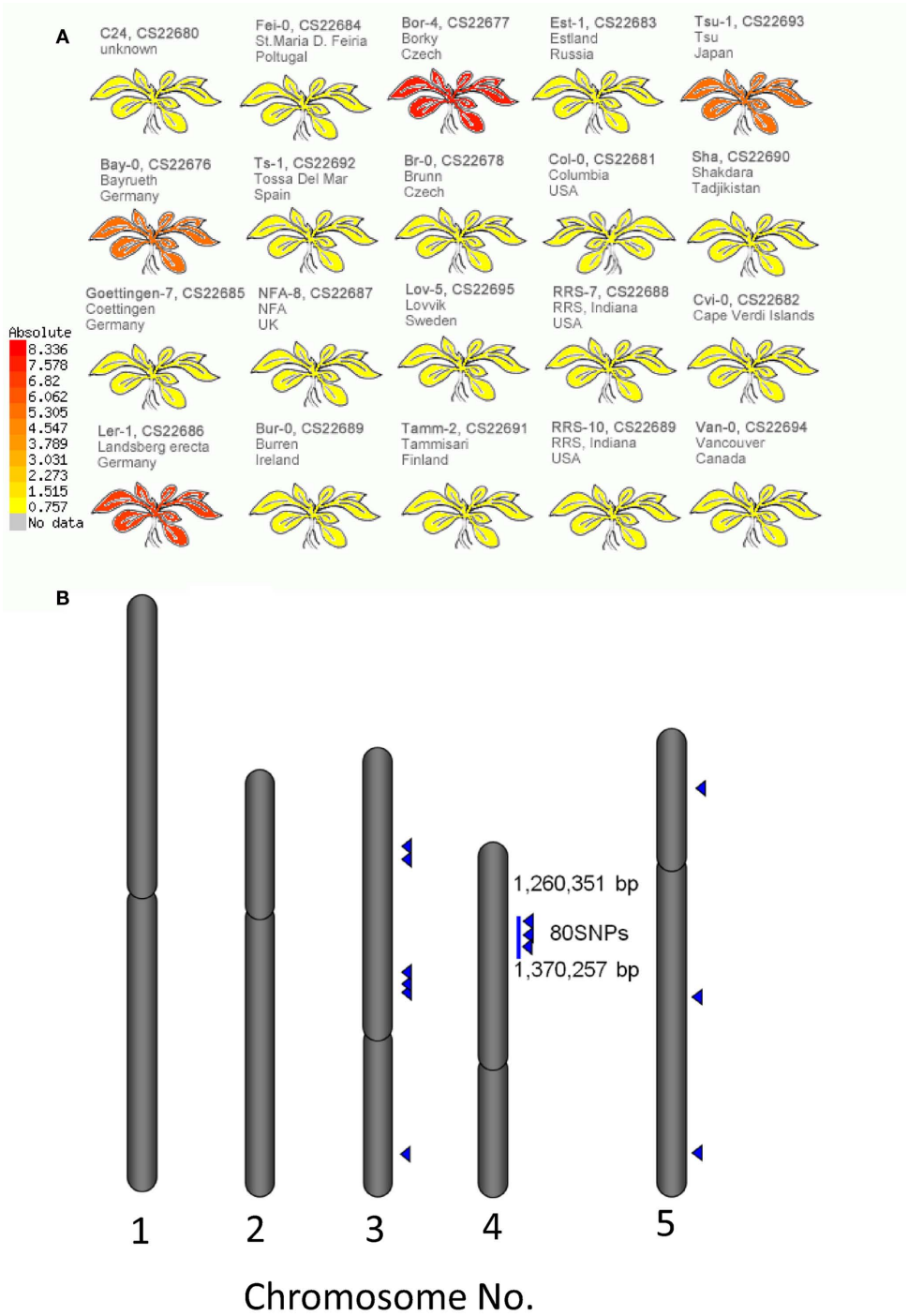


FIGURE 6 | Association between levels of 3-hydroxy-n-propylglucosinolate and single nucleotide polymorphisms (SNPs) across 20 accessions of *Arabidopsis*. (A) Heat-map representation of 3-hydroxy-n-propylglucosinolate levels in each accession. **(B)** Positions of SNPs associated with 3-hydroxy-n-propylglucosinolate levels on the *Arabidopsis* genome. Blue triangles indicate positions of the SNPs.

for the last step of hydroxyalkylglucosinolate biosynthesis (AOP3, At4g03050). Although the biological meaning of the LD are unclear, the association between the natural variations in the 3-hydroxy-*n*-propylglucosinolate levels and genetic polymorphisms in the AOP3 gene has been reported (Kliebenstein et al., 2001; Wentzell et al., 2007).

DISCUSSION

A framework for the automated structural elucidation of LC–MS metabolome data was developed to investigate the structural diversity of phytochemicals. Although the framework requires a large amount of structure-related information (MS2T library) and

intensive searches of large databases (**Figures 2 and 3**), the processing of the AtMetExpress 20 ecotype dataset (**Figure 1**) demonstrated that the method is able to reasonably estimate metabolite structures. By referring to the automatically assigned information, the effort required for the manual curation of metabolome data could be drastically reduced (**Figure 4**), which accelerated the investigation of natural variations in the *Arabidopsis* secondary metabolites (**Figures 5 and 6**). These results demonstrated that the framework is effective for the structural elucidation of LC–MS metabolome data, although several technical improvements are required for more comprehensive annotation of the metabolites.

Since the MS/MS spectra database is one of the most important kernels in the framework (**Figures 3 and 4**), the search results are highly dependent on the database quality. For example, processing of the AtMetExpress 20 ecotype dataset failed to identify metabolites belonging to alkaloids and terpenoids, probably because the current version of ReSpect contains poor entries of those metabolites in contrast to the rich flavonoids and glucosinolates data[2]. This bias is derived from the available standard compounds and published MS/MS spectra data. However, the data dependency indicated that further enrichment of the MS/MS spectra database by the addition of alkaloids, terpenoids, and other phytochemicals could directly improve the results of the structural elucidation. To promote the integration and sharing of spectral data, all ReSpect contents were opened to the public from the PRIME Web site (**Table 1**).

Structures elucidated by an automated method should contain incorrect hits derived from errors in mass analyses, indicating that the false discovery rate (FDR) of large-scale search results must be evaluated (Matsuda et al., 2009b; Saito and Matsuda, 2010). In the case of the homology searches of gene sequences, the levels of FDR could be controlled using a probability-based searching algorithm such as BLAST (Altschul and Erickson, 1985). In this study, the cosine product (dot product) method was employed to search MS/MS spectra because it is robust enough to identify identical spectra (Stein and Scott, 1994). A drawback of this method is a FDR control since the similarity score is not based on probability. To reduce false positives, the output obtained from the ReSpect search was compared with that derived from KNApSAcK to identify common results (**Figures 2 and 3**). The cross-check strategy should reduce false-positive hits, but many metabolite signals were assigned with no structural information. In the case of the AtMetExpress 20 Ecotype dataset, 94% of 703 metabolite signals were tagged by at least one MS2T, and the metabolite structures could be somehow estimated for approximately 30% of the signals (**Table S1** in Supplementary Material). Further development of a probability-based algorithm to determine the similarity between MS/MS spectra is required to increase the numbers of structurally elucidated metabolite signals while controlling FDR (Mylonas et al., 2009).

In the framework developed herein, putative characterization of the metabolite signal could be attained by introducing a new simple ontology system to cover the wider range of plant metabolites. Additionally, the performance of the ontology system was demonstrated for the AtMetExpress 20 ecotype datasets, which revealed the diversity of secondary metabolites in *Arabidopsis* based on structural elucidation using the putatively characterized information. The comparison of levels of putatively characterized metabolites revealed the genetic background of metabolotype variations, which would facilitate the analysis of these associations with genetic polymorphism and evolution.

[2]http://spectra.psc.riken.jp/menta.cgi/analyses/explains?chapter=1

ACKNOWLEDGMENT

We would like to thank Drs. K. Hanada, K. Akiyama, T. Sakurai, R. Niida, and A. Takahashi (RIKEN PSC) for their useful comments regarding this manuscript and their technical support. This work was partly supported by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Genomics for Agricultural Innovation, NVR-0005).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Plant_Physiology/10.3389/fpls.2011.00040/abstract/

REFERENCES

Akiyama, K., Chikayama, E., Yuasa, H., Shimada, Y., Tohge, T., Shinozaki, K., Hirai, M. Y., Sakurai, T., Kikuchi, J., and Saito, K. (2008). PRIMe: a web site that assembles tools for metabolomics and transcriptomics. *In Silico Biol.* 8, 339–345.

Altschul, S. F., and Erickson, B. W. (1985). Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol.* 2, 526–538.

Bednarek, P., and Osbourn, A. (2009). Plant-microbe interactions: chemical diversity in plant defense. *Science* 324, 746–748.

Bottcher, C., Roepenack-Lahaye, E. V., Willscher, E., Scheel, D., and Clemens, S. (2007). Evaluation of matrix effects in metabolite profiling based on capillary liquid chromatography electrospray ionization quadrupole time-of-flight mass spectrometry. *Anal. Chem.* 79, 1507–1513.

Burlat, V., Kwon, M., Davin, L. B., and Lewis, N. G. (2001). Dirigent proteins and dirigent sites in lignifying tissues. *Phytochemistry* 57, 883–897.

Clark, R. M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T. T., Fu, G., Hinds, D. A., Chen, H., Frazer, K. A., Huson, D. H., Scholkopf, B., Nordborg, M., Ratsch, G., Ecker, J. R., and Weigel, D. (2007). Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science* 317, 338–342.

Davin, L. B., and Lewis, N. G. (2005). Lignin primary structures and dirigent sites. *Curr. Opin. Biotechnol.* 16, 407–415.

De Kraker, J. W., and Gershenzon, J. (2011). From amino acid to glucosinolate biosynthesis: protein sequence changes in the evolution of methylthioalkylmalate synthase in Arabidopsis. *Plant Cell* 23, 38–53.

De Vos, R. C., Moco, S., Lommen, A., Keurentjes, J. J., Bino, R. J., and Hall, R. D. (2007). Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* 2, 778–791.

Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., Mcnaught, A., Alcantara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* 36, D344–350.

Fiehn, O., Robertson, D., Griffin, J., Werf, M. V. D., Nikolau, B., Morrison, N., Sumner, L. W., Goodacre, R., Hardy, N. W., Taylor, C., Fostel, J., Kristal, B., Kaddurah-Daouk, R., Mendes, P., Ommen, B. V., Lindon, J. C., and Sansone, S.-A. (2007). The metabolomics standards initiative (MSI). *Metabolomics* 3, 175–178.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M. Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., and Nishioka, T. (2010). MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* 45, 703–714.

Iijima, Y., Nakamura, Y., Ogata, Y., Tanaka, K., Sakurai, N., Suda, K., Suzuki, T., Suzuki, H., Okazaki, K., Kitayama, M., Kanaya, S., Aoki, K., and Shibata, D. (2008). Metabolite annotations based on the integration of mass spectral information. *Plant J.* 54, 949–962.

Janowitz, T., Trompetter, I., and Piotrowski, M. (2009). Evolution of nitrilases in glucosinolate-containing plants. *Phytochemistry* 70, 1680–1686.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480–D484.

Keurentjes, J. J., Fu, J., De Vos, C. H., Lommen, A., Hall, R. D., Bino, R. J., Van Der Plas, L. H., Jansen, R. C., Vreugdenhil, D., and Koornneef, M. (2006). The genetics of plant metabolism. *Nat. Genet.* 38, 842–849.

Kind, T., and Fiehn, O. (2010). Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal. Rev.* 2, 23–60.

Kliebenstein, D. J., Lambrix, V. M., Reichelt, M., Gershenzon, J., and Mitchell-Olds, T. (2001). Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in Arabidopsis. *Plant Cell* 13, 681–693.

Li, J., Hansen, B. G., Ober, J. A., Kliebenstein, D. J., and Halkier, B. A. (2008). Subclade of flavin-monooxygenases involved in aliphatic glucosinolate biosynthesis. *Plant Physiol.* 148, 1721–1733.

Lommen, A. (2009). MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal. Chem.* 81, 3079–3086.

Manzaneda, A. J., Prasad, K. V., and Mitchell-Olds, T. (2010). Variation and fitness costs for tolerance to different types of herbivore damage in Boechera stricta genotypes with contrasting glucosinolate structures. *New Phytol.* 188, 464–477.

Matsuda, F., Hirai, M. Y., Sasaki, E., Akiyama, K., Yonekura-Sakakibara, K., Provart, N. J., Sakurai, T., Shimada, Y., and Saito, K. (2010a). AtMetExpress development: a phytochemical atlas of Arabidopsis development. *Plant Physiol.* 152, 566–578.

Matsuda, F., Ishihara, A., Takanashi, K., Morino, K., Miyazawa, H., Wakasa, K., and Miyagawa, H. (2010b). Metabolic profiling analysis of genetically modified rice seedlings that overproduce tryptophan reveals the occurrence of its inter-tissue translocation. *Plant Biotechnol.* 27, 17–27.

Matsuda, F., Redestig, H., Sawada, Y., Shinbo, Y., Hirai, M. Y., Kanaya, S., and Saito, K. (2009a). Visualization of metabolite identifier information. *Plant Biotechnol.* 26, 479–483.

Matsuda, F., Shinbo, Y., Oikawa, A., Hira, M. Y., Fiehn, O., Kanaya, S., and Saito, K. (2009b). Assessment of metabolome annotation quality: a method for evaluating the false discovery rate of elemental composition searches. *PLoS ONE* 4, e7490. doi: 10.1371/journal.pone.0007490

Matsuda, F., Yonekura-Sakakibara, K., Niida, R., Kuromori, T., Shinozaki, K., and Saito, K. (2009c). MS/MS spectral tag (MS2T)-based annotation of non-targeted profile of plant secondary metabolites. *Plant J.* 57, 555–577.

Mihaleva, V. V., Vorst, O., Maliepaard, C., Verhoeven, H. A., Vos, R. C. H. D., Hall, R. D., and Ham, R. C. H. J. V. (2008). Accurate mass error correction in liquid chromatography time-of-flight mass spectrometry based metabolomics. *Metabolomics* 4, 171–182.

Moco, S., Bino, R. J., Vorst, O., Verhoeven, H. A., De Groot, J., Van Beek, T. A., Vervoort, J., and De Vos, C. H. (2006). A liquid chromatography-mass spectrometry-based metabolome database for tomato. *Plant Physiol.* 141, 1205–1218.

Moco, S., Bino, R. J., Vos, R. C. H. D., and Vervoort, J. (2007). Metabolomics technologies and metabolite identification. *Trends Analyt. Chem.* 26, 855–866.

Mylonas, R., Mauron, Y., Masselot, A., Binz, P. A., Budin, N., Fathi, M., Viette, V., Hochstrasser, D. F., and Lisacek, F. (2009). X-Rank: a robust algorithm for small molecule identification using tandem mass spectrometry. *Anal. Chem.* 81, 7604–7610.

Nakabayashi, R., Kusano, M., Kobayashi, M., Tohge, T., Yonekura-Sakakibara, K., Kogure, N., Yamazaki, M., Kitajima, M., Saito, K., and Takayama, H. (2009). Metabolomics-oriented isolation and structure elucidation of 37 compounds including two anthocyanins from Arabidopsis thaliana. *Phytochemistry* 70, 1017–1029.

Nakatsubo, T., Mizutani, M., Suzuki, S., Hattori, T., and Umezawa, T. (2008). Characterization of Arabidopsis thaliana pinoresinol reductase, a new type of enzyme involved in lignan biosynthesis. *J. Biol. Chem.* 283, 15550–15557.

Neumann, S., and Bocker, S. (2010). Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. *Anal. Bioanal. Chem.* 398, 2779–2788.

Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res.* 18, 2024–2033.

Plantegenet, S., Weber, J., Goldstein, D. R., Zeller, G., Nussbaumer, C., Thomas, J., Weigel, D., Harshman, K., and Hardtke, C. S. (2009). Comprehensive analysis of Arabidopsis expression level polymorphisms with simple inheritance. *Mol. Syst. Biol.* 5, 242.

Poole, R. L. (2007). The TAIR database. *Methods Mol. Biol.* 406, 179–212.

Saeed, A. I., Bhagabati, N. K., Braisted, J. C., Liang, W., Sharov, V., Howe, E. A., Li, J., Thiagarajan, M., White, J. A., and Quackenbush, J. (2006). TM4 microarray software suite. *Meth. Enzymol.* 411, 134–193.

Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., and Quackenbush, J. (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34, 374–378.

Saito, K., and Matsuda, F. (2010). Metabolomics for functional genomics, systems biology, and biotechnology. *Annu. Rev. Plant Biol.* 61, 463–489.

Sawada, Y., Kuwahara, A., Nagano, M., Narisawa, T., Sakata, A., Saito, K., and Hirai, M. Y. (2009). Omics-based approaches to methionine side chain elongation in Arabidopsis: characterization of the genes encoding methylthioalkylmalate isomerase and methylthioalkylmalate dehydrogenase. *Plant Cell Physiol.* 50, 1181–1190.

Shinbo, Y., Nakamura, Y., Altaf-Ul-Amin, M., Asahi, H., Kurokawa, K., Arita, M., Saito, K., Ohta, D., Shibata, D., and Kanaya, S. (2006). "KNApSAcK: A comprehensive species-metabolite relationship database," in *Biotechnology in Agriculture and Forestry 57 Plant Metabolomics*, eds K. Saito, R. A. Dixon, and L. Willmitzer (Berlin: Springer), 165–181.

Stein, S. E., and Scott, D. R. (1994). Optimization and testing of mass-spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* 5, 859–866.

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W.-M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., Reily, M. D., Thaden, J. J., and Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics* 3, 211–221.

Takahashi, H., Kai, K., Shinbo, Y., Tanaka, K., Ohta, D., Oshima, T., Altaf-Ul-Amin, M., Kurokawa, K., Ogasawara, N., and Kanaya, S. (2008). Metabolomics approach for determining growth-specific metabolites based on Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Bioanal. Chem.* 391, 2769–2782.

Weigel, D., and Mott, R. (2009). The 1001 genomes project for Arabidopsis thaliana. *Genome Biol.* 10, 107.

Wentzell, A. M., Rowe, H. C., Hansen, B. G., Ticconi, C., Halkier, B. A., and Kliebenstein, D. J. (2007). Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. *PLoS Genet.* 3, e162. doi: 10.1371/journal.pgen.0030162

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M. A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D. D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G. E., Macinnis, G. D., Weljie, A. M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B. D., Vogel, H. J., and Querengesser, L. (2007). HMDB: the human metabolome database. *Nucleic Acids Res.* 35, D521–D526.

Zeller, G., Clark, R. M., Schneeberger, K., Bohlen, A., Weigel, D., and Ratsch, G. (2008). Detecting polymorphic regions in Arabidopsis thaliana with resequencing microarrays. *Genome Res.* 18, 918–929.

Conflict of Interest Statement: 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.

Received: 29 April 2011; accepted: 01 August 2011; published online: 22 August 2011.
Citation: Matsuda F, Nakabayashi R, Sawada Y, Suzuki M, Hirai MY, Kanaya S and Saito K (2011) Mass spectra-based framework for automated structural elucidation of metabolome data to explore phytochemical diversity. Front. Plant Sci. 2:40. doi: 10.3389/fpls.2011.00040
This article was submitted to Frontiers in Plant Physiology, a specialty of Frontiers in Plant Science.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 12 October 2011
doi: 10.3389/fpls.2011.00054

Ultra performance liquid chromatography and high resolution mass spectrometry for the analysis of plant lipids

Jan Hummel, Shruthi Segu, Yan Li, Susann Irgang, Jessica Jueppner and Patrick Giavalisco*

Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany

Edited by:
Alisdair Fernie, Max Planck Institute for Plant Physiology, Germany

Reviewed by:
Kazuki Saito, Chiba University, Japan
Asaph Aharoni, Weizmann Institute of Science, Israel

***Correspondence:**
Patrick Giavalisco, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, Golm, 14476 Potsdam, Germany.
e-mail: giavalisco@mpimp-golm.mpg.de

Holistic analysis of lipids is becoming increasingly popular in the life sciences. Recently, several interesting, mass spectrometry-based studies have been conducted, especially in plant biology. However, while great advancements have been made we are still far from detecting all the lipids species in an organism. In this study we developed an ultra performance liquid chromatography-based method using a high resolution, accurate mass, mass spectrometer for the comprehensive profiling of more than 260 polar and non-polar *Arabidopsis thaliana* leaf lipids. The method is fully compatible to the commonly used lipid extraction protocols and provides a viable alternative to the commonly used direct infusion-based shotgun lipidomics approaches. The whole process is described in detail and compared to alternative lipidomic approaches. Next to the developed method we also introduce an in-house developed database search software (GoBioSpace), which allows one to perform targeted or un-targeted lipidomic and metabolomic analysis on mass spectrometric data of every kind.

Keywords: lipidomics, ultra performance liquid chromatography, high resolution mass spectrometry, accurate mass, database, all-ion fragmentation, *Arabidopsis thaliana*, metabolomics

INTRODUCTION

Holistic analysis of a cellular metabolome, the complement of all small molecules within a cell (Oliver et al., 1998), is still quite complicated due to the huge complexity and the large chemical heterogeneity of all the contained molecules. Besides the polar compounds, like sugars and amino- and organic-acids, there are also a large number of non-polar (water insoluble) compounds which need to be analyzed. The high complexity and chemical diversity, but also the huge difference in the molar abundance of these compounds explains why up to now no single analytical platform has been developed that is able to detect and quantify all of these compounds in a single analysis (Oldiges et al., 2007). As a consequence, different sample extraction and fractionation methods have been developed which allow a rough separation of the metabolites into less complex and more homogeneous fractions before their analysis (Vuckovic et al., 2010). One functionally and chemically distinct metabolic fraction that can be efficiently separated from crude extracts contains the water insoluble, generally hydrophobic lipids.

Lipids have essential functions for all living cells, not only because they are the building blocks of the membranes, which enclose the cell and the internal organelles (Van Meer et al., 2008), but also by functioning as energy storage or signaling molecules (Downes and Currie, 1998; Spiegel and Milstien, 2003; Wenk, 2005; Wymann and Schneiter, 2008). For this purpose it is not surprising that a complete new branch in the metabolomics area, namely the field of lipidomics, emerged, and has made great advancement within the last few years (Dennis, 2009; Blanksby and Mitchell, 2010; Wenk, 2010; Harkewicz and Dennis, 2011). Lipids, which are often defined by their inability to dissolve in water, do still cover a broad spectrum of diverse substances ranging from slightly polar [e.g., glycosylated sphingolipids (Merrill et al., 2009) to highly non-polar lipids like, e.g., triacylglycerol (Kuksis, 2007)]. Estimations on lipid numbers within eukaryotic cells range from a few 100 to several 1,000 lipid species (Dennis, 2009), indicating the expected high complexity. To structure this complexity and to generate a uniform nomenclature for the known lipids a general classification and nomenclature system was required. The publicly funded LIPID MAPS Consortium (Fahy et al., 2005, 2009) provided a new definition system, which is mostly based on the biosynthetic origin of the different lipids and not only on the solubility of the compound. Therefore the lipids are now defined as hydrophobic or amphiphatic small molecules, which originate from carbanion-based condensation of thioesters or by carbocation-based condensation of isoprene units (Fahy et al., 2005). This new definition is not only more precise then the old water insolubility-based definition, but it also allows to classify the commonly known lipids into homogenous functional subclasses: namely the fatty acids, glycerolipids, glycerophospholipids, sphingolipids, sterols, prenols, saccharolipids, and polyketides (Fahy et al., 2005).

The fact that no single analytical technology has allowed the identification and quantification of all metabolite species in a single experiment is also true for the analysis of all the different lipids from a cell (Wenk, 2010). Historically, lipids have been analyzed by diverse chromatography-based separation methods (Bausch, 1993). Commonly used technologies comprised methods like one or two dimensional thin layer chromatography in combination

with different visualization strategies (Touchstone, 1995), but also high performance liquid chromatography (HPLC) methods in combination with various detection systems (Picchioni et al., 1996). Even though these methods have proven useful for many purposes, it seems that their limitations for large scale quantitative lipid analysis are more evident (Blanksby and Mitchell, 2010). As a consequence, mass spectrometry (MS)-based methods, with or without chromatographic separation techniques, have evolved to fill this technological gap (Welti et al., 2007b; Griffiths and Wang, 2009; Blanksby and Mitchell, 2010; Wenk, 2010; Harkewicz and Dennis, 2011).

There are many different MS instruments available which can be combined with an even larger number of separation systems (Griffiths and Wang, 2009; Wenk, 2010). Still, only two main strategies for the analysis of lipids have been used in most of the described reports: on one hand there is the most successfully used method, namely shotgun lipidomics, which relies on a separation free (direct infusion) analysis of a crude lipid extract on triple quadrupole (QqQ) or quadrupole time-of-flight (qTOF) mass spectrometers (Welti and Wang, 2004; Han and Gross, 2005; Ejsing et al., 2009; Yang et al., 2009), on the other hand there is chromatography-based separation prior to the mass spectrometric measurement for the lipid analysis, which has been used only in a small number of studies thus far (Markham and Jaworski, 2007; Rainville et al., 2007; Glauser et al., 2008b; Nakanishi et al., 2009; Nygren et al., 2011). Both methods have their advantages and disadvantages: for example, the shotgun approach is prone to strong ion suppression effects, which can in part be compensated for by large sample dilutions or by the use of internal reference compounds (Moore et al., 2007). While the chromatography-based methods are less sensitive to these suppression effects, due to the chromatographic separation (Muller et al., 2002; Annesley, 2003), these approaches were thus far unsuitable for absolute lipid quantification (Stahlman et al., 2009).

In the field of plant metabolomics both technologies have found their applications, while the polar glycerolipids have been widely analyzed by the shotgun lipidomic approach (Devaiah et al., 2007; Welti et al., 2007b; Zhang et al., 2009; Kilaru et al., 2010), sphingolipids have been most successfully analyzed by targeted LC–MS-based approaches (Markham et al., 2006; Markham and Jaworski, 2007; Chen et al., 2008). Still, since most of these studies made use of highly sensitive, but low resolution mass spectrometers, they were mostly performed in a targeted way, by simply profiling a limited number of known lipid species (Lu et al., 2008).

In this report we describe a versatile and reproducible ultra performance liquid chromatography (UPLC)-based separation system, coupled to a high resolution mass spectrometer operating in MS as well as all-ion fragmentation mode. The developed system allows for the accurate qualitative and semi-quantitative targeted analysis of several hundred different lipid species extracted from a single plant sample. Additionally, due to the combination of chromatography and high resolution MS and all-ion MS/MS, the method allows to revisit the data long after the actual measurement and therefore extract and possibly elucidate novel structures (Harkewicz and Dennis, 2011). For the actual data mining we introduce a novel database search (GoBioSpace), which allows one to perform either targeted or un-targeted database searches with the acquired lipid data.

MATERIALS AND METHODS

PLANT GROWTH

The *Arabidopsis thaliana* Col-0 plants used for the metabolite extraction were grown in a light and temperature controlled phytotron under constant CO_2 conditions using a BioBox growth chamber (GMS Gaswechsel-Messsysteme GmbH, Berlin, Germany). The plant material preparation and the experimental settings for the BioBox were as previously described (Huege et al., 2007). Plant growth in the BioBox was performed for 42 days. The aerial parts of the plants were separated from the roots by cutting, and immediately snap frozen in liquid nitrogen.

LIPID EXTRACTION PROTOCOL

Lipids were extracted from three independent biological replicates of *Arabidopsis thaliana* leaves. In brief: 50 mg of frozen leaf tissue was homogenized in a 2 ml Eppendorf tube (Eppendorf, Hamburg, Germany) for two times 1 min at maximum speed within a Retsch mill (MM 301, Retsch, Düsseldorf, Germany). The lipids were extracted from each aliquot using 1 ml of a pre-cooled (−20˚C) homogenous methanol:methyl-tert-butyl-ether (1:3) mixture, spiked with 0.1 μg/ml PE 34:0 (17:0, 17:0), and PC 34:0 (17:0, 17:0) as internal standards. For the extraction, the samples were incubated for 10 min in a shaker at 4˚C (Thermostat Plus, Eppendorf), followed by another 10 min incubation in an ultrasonication bath at RT. After adding 500 μl of UPLC grade water:methanol (3:1), the homogenate was vortexed and centrifuged for 5 min at 4˚C in a table top centrifuge (Eppendorf). The addition of water:methanol leads to a phase separation producing an upper organic phase, containing the lipids, and a lower phase containing the polar and semi-polar metabolites. The upper organic phase was removed, dried in a speed-vac concentrator, and stored at −80˚C until used.

UPLC–FT–MS MEASUREMENT OF LIPIDS

The dried lipid extracts were re-suspended in 500 μl buffer B (see below) and transferred to a glass vial. Two microliters of this sample were injected on a C_8 reversed phase column (100 mm × 2.1 mm × 1.7 μm particles waters), using a Waters Acquity UPLC system. The two mobile phases were water (UPLC MS grade, BioSolve) with 1% 1 M NH_4Ac, 0.1% acetic acid (Buffer A,), and acetonitrile:isopropanol (7:3, UPLC grade BioSolve) containing 1% 1 M NH_4Ac, 0.1% acetic acid (Buffer B). The gradient separation, which was performed at a flow rate of 400 μl/min, was: 1 min 45% A, 3 min linear gradient from 45% A to 35% A, 8 min linear gradient from 25 to 11% A, 3 min linear gradient from 11% A to 1% A. After washing the column for 3 min with 1% A the buffer was set back to 45% A and the column was re-equilibrated for 4 min (22 min total run time).

The mass spectra were acquired using an Exactive mass spectrometer (Thermo-Fisher, Bremen, Germany). The spectra were recorded using altering full scan and all-ion fragmentation scan mode, covering a mass range from 100–1500 m/z. The resolution was set to 10,000 with 10 scans per second, restricting the Orbitrap loading time to a maximum of 100 ms with a target value of 1E6

ions. The capillary voltage was set to 3 kV with a sheath gas flow value of 60 and an auxiliary gas flow of 35. The capillary temperature was set to 150°C, while the drying gas in the heated electro spray source was set to 350°C. The skimmer voltage was held at 25 V while the tube lens was set to a value of 130 V. The spectra were recorded from min 1 to min 20 of the UPLC gradients.

MANUAL AND AUTOMATED PEAK EXTRACTION AND ALIGNMENT

Chromatograms from the UPLC–FT–MS runs were analyzed and processed either by using Xcalibur (Version 2.10, Thermo-Fisher, Bremen, Germany), ToxID (Version 2.1.1, Thermo-Fisher), or automatically with the Refiner MS® software (Version 6.0, GeneData, Basel, Switzerland). In the automated approach the molecular masses, retention time, and associated peak intensities for the three replicates of each sample were extracted from the raw files, which contained the full scan MS and the all-ion fragmentation MS data. The processing of the MS data included the separate processing of the full scan spectra and the all-ion fragmentation spectra. Chemical noise was automatically removed from the spectra before the chromatograms were aligned using a pair wise-based alignment tree algorithm (Refiner MS 6.0).

Further peak filtering on the manually extracted spectra or the aligned data matrices was performed in Excel or Access (Microsoft, Seattle, WA, USA).

GOBIOSPACE DATABASE

Based on the fact that the masses measured in the mass spectrometer are almost directly connected to the elemental composition of a measured analyte, considering either an addition or loss of a sub structure – so called adducts (i.e., $[M+H]^+$ protonation, $[M-H]^-$ de-protonation, $M+NH_4]^+$ Ammonium-, $[M+Na]^+$ Sodium-, $[M+Ca]^+$ Calcium-adduct), GoBioSpace (Golm Biochemical Space) was conceptualized as a repository of elemental compositions with source tagged annotations for properties such as InChI strings, CAS numbers, IUPAC names, synonyms, cross references or KEGG Pathway names, among others.

The source of an annotation – the so called depositor – facilitates as a filter for the biological relevance of elemental compositions. The meaningful interpretation of search results in a biological context is accomplished by a targeted search limiting the formula to biology related depositors such as KEGG and BioCyc, among others. In contrast, relaxed searches in regard to the formula's depositor (i.e., including those elemental compositions only reported from vendors of potentially synthesized chemicals) result in search hits with lower biological interpretability.

To date, we collected more than 366 million meta information for 2.1 million unique elemental compositions from more than 150 public available databases (143 included in PubChem), such as the chemical focused databases PubChem Substance[1] and ChemSpider[2] or biological focused databases such as the Human Metabolome Database[3] and Metabolome.JP[4] into the GoBioSpace repository. Our approach also facilitates the search against potentially putative elemental compositions such as described for lipids in the chapter "Targeting Specific Lipids within the Total Ion Chromatogram: Pick What You Know."

For high resolution mass search queries, the accurate isotopic masses for either ambient ^{12}C or fully isotopic labeled ^{13}C, ^{15}N, and ^{34}S formula were calculated according to Böhlke et al. (2005). An indexed view in the database allows the single step matching of measured masses to elemental compositions, tolerating a given mass error and considering user defined sets of expected analytical adducts and depositors to correct the measured masses. In addition, the client side search application supports the restriction of elemental composition hits based on atom number constraints.

To make the mass search functionality accessible to the community, we implemented a Web Service within the Golm Metabolome Database (GMD[5]; Kopka et al., 2005; Hummel et al., 2010) and integrated this web service into a graphical user interface which is also made available http://gmd.mpimp-golm.mpg.de/GoBioSpace.aspx. Here, elemental compositions and individual or batched (tabulator formatted text files) masses can easily be configured and searched against databases of interest. The matching results are returned as browse- and sort-able tables which can be exported for further analysis as tabular formatted text files. However, the web services can be integrated for non-commercial use into any data processing pipeline. All software is implemented using the Microsoft .NET 4.0 framework, the C# language, and Microsoft Visual Studio® 2010. The data back end is based on a Microsoft® SQL Server® 2005.

RESULTS AND DISCUSSION

UPLC–FT/MS-BASED SEPARATION AND MEASUREMENT OF CRUDE *ARABIDOPSIS* LIPID EXTRACTS

Arabidopsis thaliana lipids were extracted using a buffer system containing methyl-tert-butyl-ether instead of chloroform as the organic solvent (Matyash et al., 2008). This extraction protocol enabled us not only to extract the lipids with a higher efficiency, but also to extract lipids, polar and semi-polar metabolites, starch, and proteins from a single sample (Giavalisco et al., 2011). The extracted lipids were analyzed on a C_8 reversed phase UPLC column, using 1.7 µm particles (Rainville et al., 2007), in a 22 min method. Both steps, the extraction as well as the chromatographic separation are simple and high-throughput compatible methods, and are applicable for several different plants but also other, non-photosynthetic organisms like, e.g., yeast, *Drosophila*, *C. elegans*, or mammalian tissue (data not shown).

All mass spectrometric measurements were performed on a standalone high resolution Orbitrap (Exactive) mass spectrometer (Lu et al., 2010), coupled to an ultra performance liquid chromatography system. This "smaller" version of an Orbitrap (lacking a the linear ion trap in front of the Orbitrap analyzer), which actually does not cost more than a QqQ mass spectrometer, still matches all the demands of an high resolution mass spectrometer [fast scanning (up to 10 Hz), high resolution (up to 100,000 R), and accurate mass (<2 ppm)]. The combination of these attributes therefore allows one not only to distinguish compounds with

[1]http://www.ncbi.nlm.nih.gov/pcsubstance
[2]http://www.chemspider.com/
[3]http://www.hmdb.ca/
[4]http://www.metabolome.jp/
[5]http://gmd.mpimp-golm.mpg.de/

very similar masses, but also to directly annotate elemental compositions, without a need for a reference compound, based on the measured accurate masses (Giavalisco et al., 2008; Xu et al., 2010).

Each lipid extract was separated and measured twice, once in positive ionization (**Figure 1A**) and once in negative ionization mode (**Figure 1B**). The reason for this duplicated measurement can be easily seen by looking at the two chromatograms, as they appear quite different. The explanation for this difference comes from the chemical nature of the detected lipid species (Han and Gross, 2005; Devaiah et al., 2006). Even though all of these lipids are constructed from a small number of building blocks (a glycerol backbone linked to a number of fatty acids), their general mass spectrometric behavior is controlled by the chemical property of their class-specific head group (Yang et al., 2009). Accordingly, even though most of these lipids ionize in both ionization modes, they do have a clear bias for a specific adduct and, as a consequence, a specific polarity (**Table 1**).

For example, monogalactosyldiacylglycerol (MGDG) 34:6 can be detected with three different adducts in the positive ionization mode ($[M+H]$, $[M+NH_4]$, and $[M+Na]$) and another two adducts can be detected in the negative ionization mode ($[M-H]$, $[M+\text{Acetate}-H]$). The appearance of these multiple adducts proves to be an extremely useful feature, even if it increases the spectral complexity, since it improves the analysis and the correct annotation of the measured lipid classes. As can be seen in **Figure 1A**, peak pairs with precise distances can be identified. A difference of m/z 21.98 (± 5 ppm) indicates a $[M+H]$ and a $[M+Na]$ ion pair, while distances of m/z 17.02 (± 5 ppm) indicate a $[M+H]$ and a $[M+NH_4]$ ion pair (**Figure 1A**).

The correct adduct annotation is of particular importance, especially if looking at lipids where the different adducts might have very similar (or even identical) masses. One example for such a case is given in **Figure 2** for a phosphatidylserine (PS) and phosphatidylglycerol (PG) lipid. As the protonated PS 34:2 is only 0.02 ppm different from the ammonium adduct of PG 34:4, which means that for the mass of 760.51385 ± 5 ppm we will get two lipid peaks from our positive mode spectrum. Looking at the adduct patterns of the spectra (including also the negative ion mode spectra), helps to solve the above mentioned annotation dilemma for these two compounds, since only the peak with a retention time of 7.17 min pairs to a sister peak with a distance of 17.02, which indicates that this peak is the ammonium adduct of PG 34:4, while the peak at RT 7.97 min can be annotated as the PS 34:2.

TARGETING SPECIFIC LIPIDS WITHIN THE TOTAL ION CHROMATOGRAM: PICK WHAT YOU KNOW

In almost all cases lipidomics studies performed in the plant field were conducted in a targeted way, meaning that a number (a few dozen to several 100) expected lipids species were profiled (Devaiah et al., 2006; Markham and Jaworski, 2007). To validate our system, we decided to profile the lipids from these previously conducted studies by selectively extracting the expected masses from our chromatograms. In total we prepared a target list containing 332 different lipid species types [168 sphingolipids (Markham and Jaworski, 2007), 147 phosphoglycero- or galacto-lipids (Devaiah et al., 2006), and 17 oxylipin species (Buseman et al., 2006)], which were detected in three independent studies, using three different extraction protocols, and three different types of mass spectrometers. As illustrated in **Figure 3A** we conducted the peak extraction by simply extracting each single mass associated to a specific lipid and relatively quantified the intensity of the different adducts from each chromatogram (Table S1 in Supplementary Material). In the same way it is also possible to extract several masses, belonging to different lipids, within a specific lipids class or from different classes, and quantitatively compare them to each other in parallel (e.g., whole PC 36: 1–6 and PC 34: 1–6 series is displayed in **Figure 3B**).

By manually extracting the masses from the chromatograms we matched 187 of the 332 different lipids, including 127/147 of the previously described phospho-, lysophospho-, and galactolipids (Devaiah et al., 2006), all 17 of the 17 previously described oxylipins (Buseman et al., 2006), and 43 of the 168 possible sphingolipids (Markham and Jaworski, 2007). Compared to the excellent coverage of lipid species from the phosphoglycero and galacto lipids the result achieved for the sphingolipids were less comprehensive, only covering the most abundant lipid species from the Markham and Jaworski (2007) study. This indicated that we were not having a general loss of sphingolipids in our method, but rather a sensitivity problem, which can often be observed if ion trap-like mass spectrometers are compared to QqQ-type mass spectrometers (Mcluckey and Wells, 2001). Additionally, we noticed that the sample preparation method used in the sphingolipid study was highly sophisticated and specifically tailored to this lipid class, including a depletion step of the highly abundant phospholipids, which will lead to a higher detection sensitivity due to strongly decreased ion suppression effects (Markham and Jaworski, 2007).

Taken together we can conclude that we do see most of the expected lipid species in our samples and most of them with several different ion species (different adducts). The data of these initially extracted and validated lipid species is collected in Table S1 in Supplementary Material.

SYSTEMATIC DISTRIBUTION OF RETENTION TIME AND MASS AIDS TO VALIDATE THE ANNOTATION OF THE MEASURED LIPIDS

Confidence in the annotation of a measured compound can be increased with the number of parameters this compound shares with related compounds. Since lipids are constructed as modular molecules (Fahy et al., 2009; Yang et al., 2009), which usually vary only slightly between the different species within a lipid class (extension of the fatty acid chain length or the degree of saturation), they have a very systematic mass and retention time behavior (Hermansson et al., 2005). Therefore, both these parameters allows the validation of lipids within a specific class by simply plotting the m/z and RT values of the measured species of the most abundant adduct in a scatter plot. As can be seen for **Figure 4** (scatter plot for the measured PCs from Table S1 in Supplementary Material), the lipids with longer fatty acid chains lead to a higher mass and increased retention time, while fatty acids with higher degrees of un-saturation result in lipids with lower masses and decreased retention times. As a consequence, a diagonal series appears within the plots. These contain lipid species with the same number of carbons atoms in the fatty acid chains but show decreasing number of double bonds from left to right (**Figure 4**). Wrongly annotated

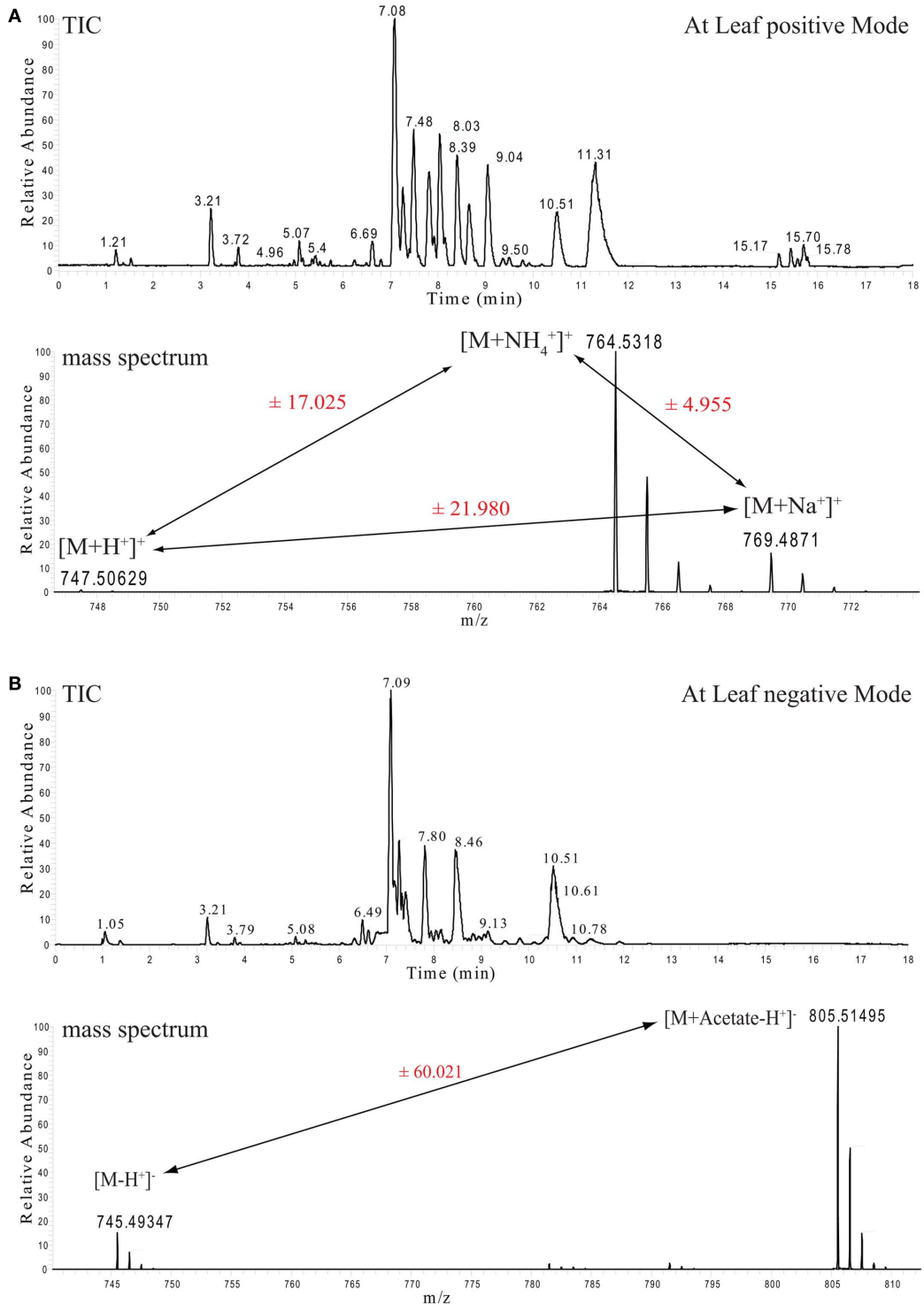


FIGURE 1 | Ultra performance liquid chromatography chromatograms and selected mass spectra from *Arabidopsis thaliana* leaf lipid extracts. (A) Total ion chromatogram (TIC, upper part) of mass spectra recorded in positive ion mode. The lower part shows the mass spectrum from the apex of the MGDG 34:6 peak with the retention time of 7.08 min and its associated ionization adducts. **(B)** As above, but here the TIC and the spectrum of the negative ion mode measurements are shown.

or unusually distributed lipids can be easily detected within these patterns since they appear as dots outside the systematic scatter pattern. A curious and unexplained example is given for the PCs with 40 carbons in the two fatty acid chains (**Figure 4**). Even

Table 1 | Ionization adducts of the detected lipid classes within the UPLC chromatograms.

Lipid class	Detected ions	Most abundant ion
PC	$[M+H]^+$, $[M+Na]^+$,$[M+Ac-H]^-$	$[M+H]^+$
PE	$[M+H]^+$, $[M+Na]^+$, $[M-H]^-$	$[M+H]^+$
PG	$[M+H]^+$, $[M+NH4]^+$, $[M+Na]^+$, $[M-H]^-$	$[M-H]^-$
PI	$[M+H]^+$, $[M+NH4]^+$, $[M-H]^-$	$[M-H]^-$
PS	$[M+H]^+$, $[M-H]^-$	$[M+H]^+$
MGDG	$[M+NH4]^+$, $[M+Na]^+$, $[M-H]^-$	$[M+Na]^+$
DGDG	$[M+NH4]^+$, $[M+Na]^+$, $[M-H]^-$	$[M+Na]^+$
SQDG	$[M+NH4]^+$, $[M+Na]^+$, $[M-H]^-$	$[M-H]^-$
Cer	$[M+H]^+$, $[M+NH4]^+$, $[M+Na]^+$, $[M-H]^-$	$[M-H]^-$, $[M+H]^+$
GlcCer	$[M+H]^+$, $[M+NH4]^+$, $[M+Na]^+$, $[M-H]^-$	$[M-H]^-$, $[M+H]^+$
GIPC	$[M+H]^+$, $[M+NH4]^+$, $[M+Na]^+$, $[M-H]^-$	$[M+H]^+$
Oxylipins	$[M+NH4]^+$, $[M+Na]^+$, $[M-H]^-$	$[M+Na]^+$
TAG	$[M+NH4]^+$, $[M+Na]^+$	$[M+NH4]^+$
DAG	$[M+NH4]^+$, $[M+Na]^+$	$[M+Na]^+$
FA	$[M-H]^-$	$[M-H]^-$

PC, phosphatidylcholine; PE, phosphatidylethanolamine; PG, phosphatidylglycerol; PI, phosphatidylinositol; PS, phosphatidylserine; MGDG, monogalactosyldiacylglycerol; DGDG, digalactosyldiacylglycerol; SQDG, sulfoquinovosyldiacylglycerol; Cer, ceramide; GlcCer, glucosylceramides; GIPC, glycosylinositolphosphoceramides; TAG, triacylglycerols; DAG, diacylglycerols; FA, fatty acids.

though these lipids are systematically distributed by themselves, it is evident from the plot that they are not matching the distribution of the other, shorter fatty acid chain lipids in this lipid class. The PC 40:2 for example, which would be predicted to have a later elution time than the PC 38:2, does actually elute almost a minute earlier than the shorter chain classmate (**Figure 4**). This could indicate that the PC 40:X lipids have been either annotated wrongly or there is a systematic shift in these longer fatty acid chain lipids.

Next to the exclusion of possibly wrongly annotated lipids, the scatter plot representation allows one to also quickly detect missing lipid species within a systematic series. In this case one or several dots would be missing within the diagonal line. In **Figure 4** we can see for example that we could not detect PC 38:1. Even rechecking the spectra at the expected retention time did not allow us to detect the expected peak.

ALL-ION FRAGMENTATION DATA FOR THE LIPID ANNOTATION VALIDATION

Using high resolution accurate mass data is in many cases sufficient to predict an elemental composition of a measured peak (Giavalisco et al., 2009). Still the accuracy and probability for a correct annotation is increased if along with the accurate mass of the intact molecule (precursor) an additional mass of a compound-specific fragment can be detected. The measurement of the mass of the intact precursor and one or several fragments are the essential values for the peak identification in shotgun lipidomic analysis (Han and Gross, 2005). The occurrence of these specific fragment ions results from either a specific loss of a charged molecule (e.g., choline head group from PC lipids) or from the loss of an

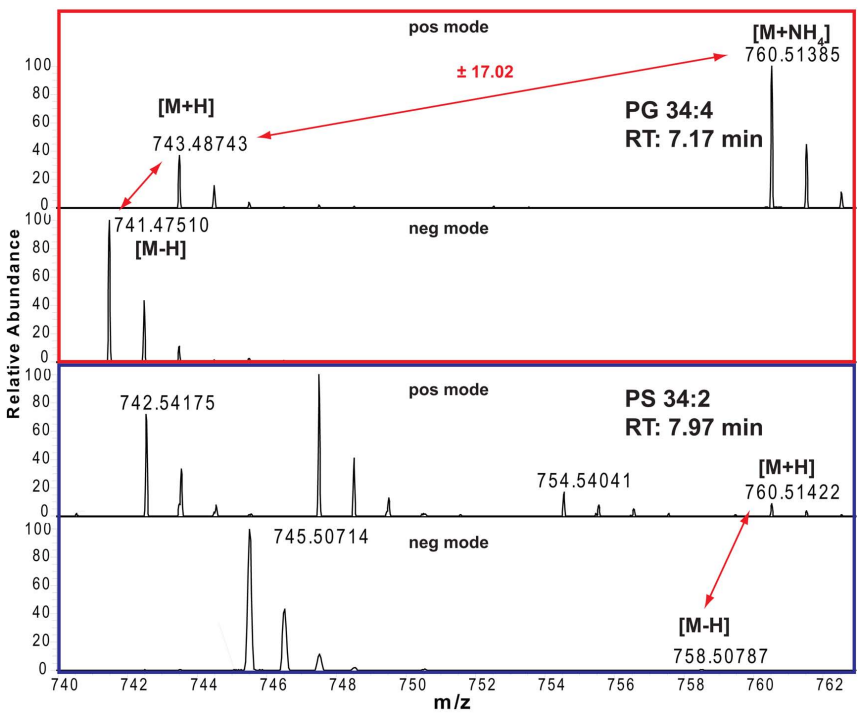


FIGURE 2 | Positive and negative ion mode spectra and adduct annotations of PG 34:4 (red boxed) and PS 34:2 (blue boxed).

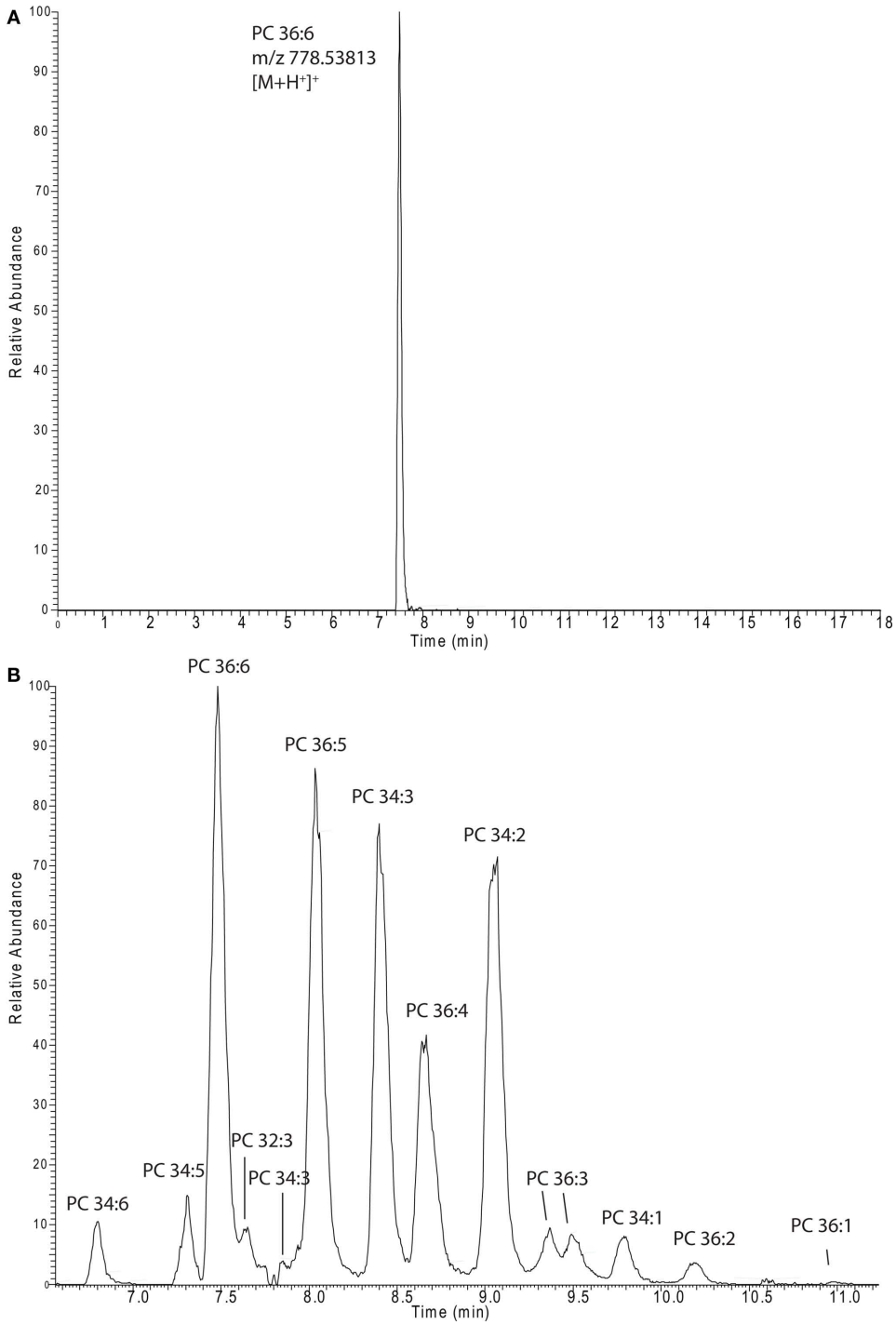


FIGURE 3 | Extracted ion chromatograms of a single lipid species [PC 36:6 (A)] or a whole series of lipids [PCs (B)]. The spectra were recorded in and extracted from positive ion mode lipid chromatograms (**Figure 1A**).

uncharged fragment (neutral loss). This technique can also be used on LC–MS-based systems in non-shotgun lipidomic studies, but only if fragmentation mass spectra are recorded.

The main advantage of high pressure sub 2 μm particle UPLC systems, compared to conventional, lower pressure, larger particle HPLC systems, is its fast, sensitive, and highly reproducible chromatography (Plumb et al., 2004). The faster chromatography and the smaller peak width, which is a consequence of the higher plate number achieved in the UPLC system, turns into a disadvantage when the number of scans/time of the mass spectrometer

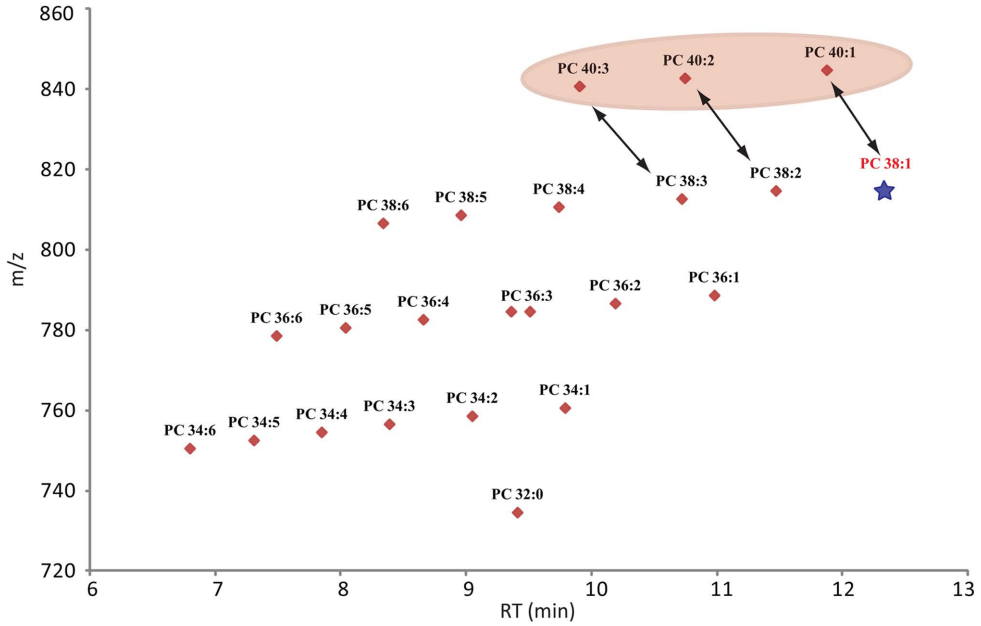


FIGURE 4 | Scatter plot of all lipids annotated as phosphocholines (Table S2 in Supplementary Material). The plot contains the measured retention time in minutes on the *x*-axes and the recorded mass of the $[M+H]^+$ adduct on the *y*-axes. Due to the modular building block structure of lipids within a homogenous class, systematic patterns of parallel lines should be observed. From bottom to top these lines should contain lipids with increasing fatty acid chain length, while the number of double bond should decrease from left to right. The star under PC 38:1 indicates that this lipid was not detected in the analyzed samples, but it would have been expected at this retention time. The PC 40:X series is highlighted since these compounds seem to elute too early and therefore do not match the expected elution pattern given by the whole class.

are too low to perform the survey full scans and data-dependent MS/MS measurements of the most abundant peaks (Schmitt-Kopplin et al., 2008). The FT–MS instrument used in this study, which has a scan speed of up to 10 Hz at a resolution of 10,000 can circumvent this problem partially, but still, even 10 scans/s are not enough time to perform classical data-dependent MS/MS analysis of several eluting masses while recording sufficient information for good peak integration, especially if the eluting peaks are only 3–6 s long (**Figure 3A**). The solution for this problem, which has originally been developed and implemented under the name MS^e as a scan method for qTOF mass spectrometers (Bateman et al., 2007), and simply relies on the fragmentation of all precursor ions measured in the full scan instead of selecting individual masses. This approach has successfully been used in a proteomic study in the Exactive MS and was called all-ion fragmentation (Geiger et al., 2010). In **Figure 5A** an illustration of the measurement method used for our lipidomic analysis is given, showing that we constantly alter between low energy full scans and high energy all-ion fragmentation scans throughout the whole chromatographic separation. The advantage of this procedure is that two independent MS data-sets are generated, one contains the intact mass information for all the compounds eluting during the chromatographic separation, while the second contains the fragmentation data for the selfsame compounds. To integrate this data and to validate a predicted lipid it is only necessary to connect the elution profile of a full scan (low energy) mass to the similarly eluting masses from the all-ion MS/MS (high energy) spectra. In **Figure 5B** this procedure is illustrated for PC 36:6. As can be seen, three fragment masses (m/z 184.07381, m/z 500.31598, and m/z 518.32513) within the mass spectra between 7.2 and 7.8 min are exactly co-eluting to the phosphocholine lipid (m/z 778.53894) and should therefore be associated. Another two masses (m/z 728.52446 and m/z 573.48822), which are closely co-eluting, show clearly differential elution profiles and can therefore excluded to be associated to PC36:6, indicating that they should represent different lipids.

The systematic analysis of these all-ion MS/MS spectra therefore allows us to uncover a number of lipid specific fragments, which can be used to validate a specific lipid species, e.g., the masses m/z 500.31598 and m/z 518.32513, which are specific fragments of PG 36:6 (**Figure 5B**). As well, we can also find class-specific fragments, like the m/z 184.07381, which is the positively charged choline fragment that can be detected for all phosphocholine lipids.

AUTOMATED LIPID ANNOTATION STRATEGIES

The strategy presented for the analysis of lipids thus far still requires a high manual input, especially for the validation of the lipid annotation. Of course this is only true if a novel sample (a new organism or a new tissue) is analyzed. Once a sample is annotated and no major changes in the extraction procedure or the chromatographic separation are introduced, the following lipid profiles can be simply matched to the results of the initially performed peak annotation.

The chromatographic and the spectral compatibility between different samples, namely the retention time and the spectral

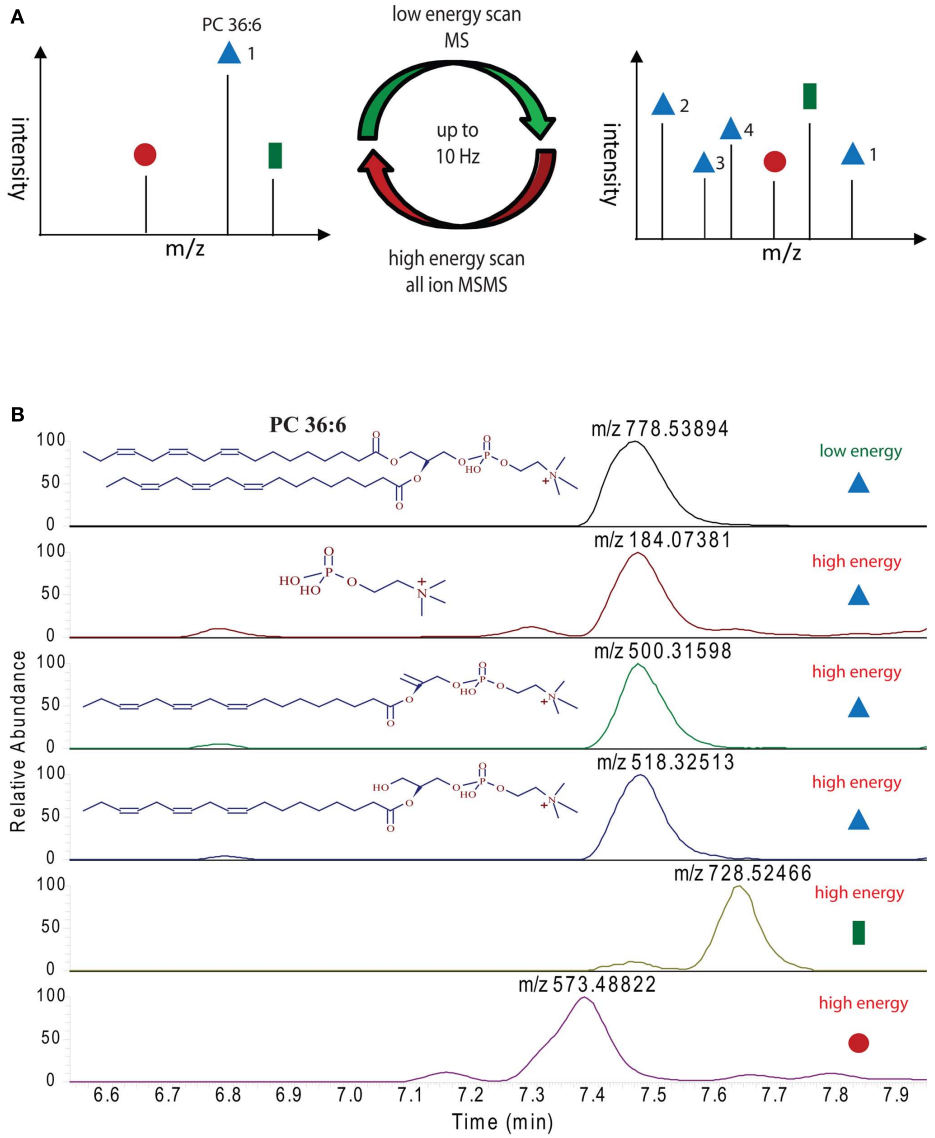


FIGURE 5 | Ultra performance liquid chromatography–MS measurement strategy employed for the lipid analysis in this study: (A) illustration of high and low energy alteration for the acquisition of full scan and all-ion MS/MS spectra. (B) Extracted ion chromatograms of the indicated masses (derived either from the high or low energy mass spectra) from a representative positive ion modes UPLC chromatogram. Peaks with the same elution profile can be regarded as co-eluting masses, which are derived from the same precursor molecule. Differentially eluting peaks have to be regarded as different compounds, requiring different annotations.

intensities, are achieved by using the two internal standards (PE 34:0 and PC 34:0), which we have spiked into the extraction buffer. Increasing the number of internal standards might be useful in the long run if the retention time system needs to be converted into a retention index system, which would possibly allow one to not only match lipids within a single experiment, but also between different experiments.

After having annotated the initial expected lipids from a novel matrix the data analysis can be automated by using one of the two different strategies depicted in **Figure 6**. The main distinction between the two approaches lies in the fact that one strategy directly targets only the peaks of interest by selectively extracting the masses of lipids of interest at specific retention times from the generated chromatograms (left part of **Figure 6**), while the second strategy relies on a slightly different approach. Here, all the peaks from the chromatograms are extracted and aligned into a data matrix before matching these peaks to the m/z and RT values of an annotated peak list (right part of **Figure 6**). The result in both cases should be almost identical. The major difference between the two approaches lies in the fact that in the first approach only annotated peaks can be used for the analysis, while the second approach allows for the further use of an un-annotated matrix, derived from the peak picking software, providing the basis for fully un-targeted lipidomics.

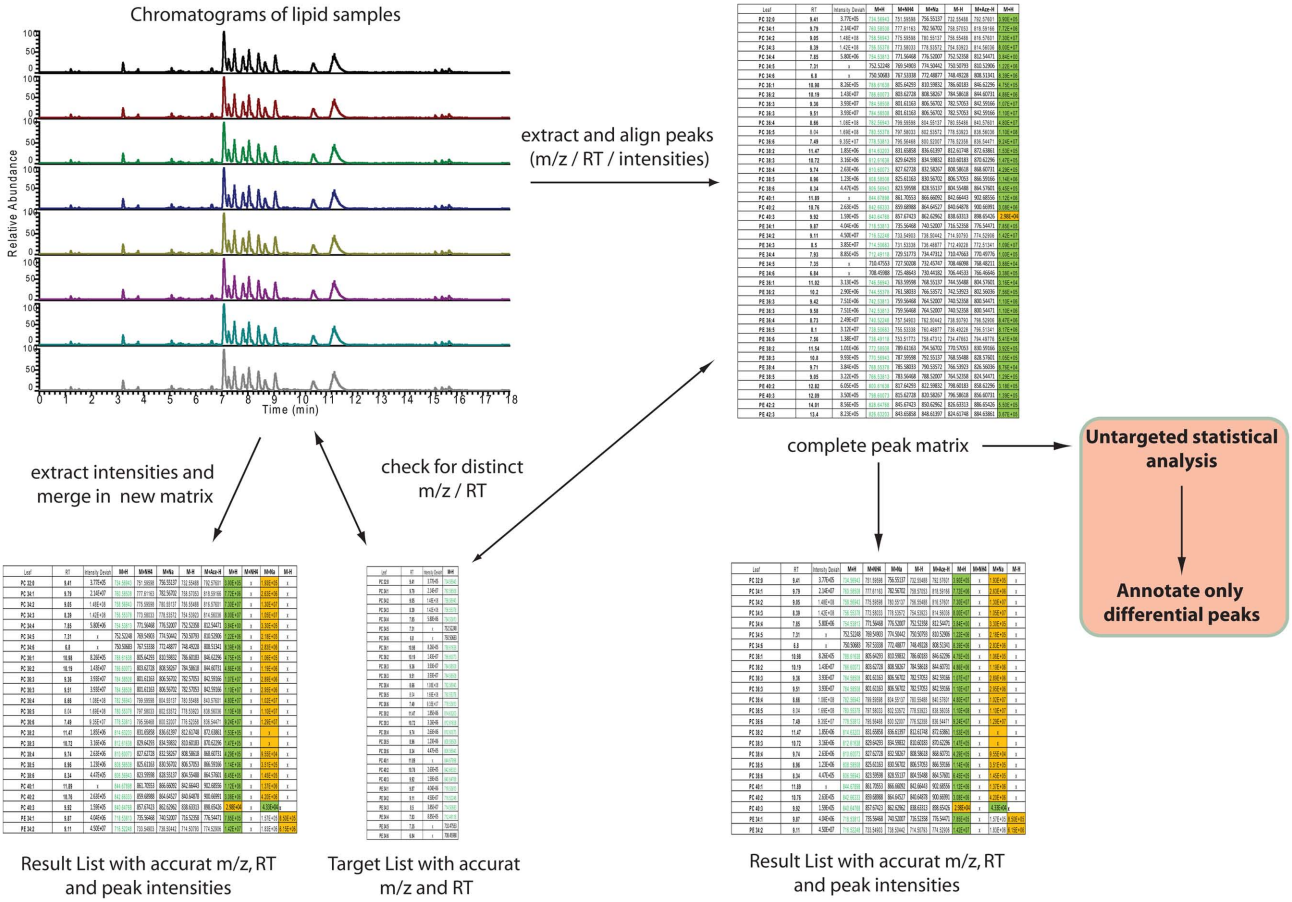


FIGURE 6 | Automated, software-assisted, strategies for targeted but also un-targeted lipid profiling. On the left hand side a purely targeted strategy is depicted, where based on a target list several chromatograms are searched for the occurrence of specific m/z and retention times. If a peak is found, within certain tolerance boundaries, the intensity is loaded to a result table. The strategy on the right hand side indicates a diverse strategy. Here all peaks are extracted first and these are written into an un-annotated data matrix. This matrix can then be compared against a target list (same as for the first strategy) or used for statistical analysis of significantly differential peaks, which then have to be annotated.

For the purpose of targeted peak picking (left part of **Figure 6**), software is usually provided by the vendor of the mass spectrometer. This software can be used by uploading a target list containing the name, the m/z, and the RT of the peaks of interest. This target list is then used to query the chromatograms generated during the analysis. The output of such a search is a list where every peak of interest is associated to the compound name, the measured m/z and RT, and an intensity value, which is equivalent to the relative amount of the compound within the sample. For the analysis of Exactive or other Thermo-Fisher MS data two software packages are available: either a processing method [which has to be entered compound by compound within Xcalibur (Thermo-Fisher, Bremen, Germany)] can be generated, or if the ToxID software package (Thermo-Fisher) is used, a comma separated text file can be employed for the targeted analysis of the lipidomic data.

For the purpose of targeted, but also un-targeted data analysis (right part of **Figure 6**), peak picking and matrix alignment of all peaks is necessary first. Here several commercial, but also open source software packages are available (Katajamaa et al., 2006; Smith et al., 2006; Katajamaa and Oresic, 2007; Benton et al., 2008; Lommen, 2009; Pluskal et al., 2010). Once the initial, un-annotated matrix is generated from a suitable software package, this matrix can be further filtered and compared to the previously generated reference lists.

Usually a matrix from *Arabidopsis* leaf tissue contains 30,000 or more reproducible peaks which are above a minimal threshold of 10,000 counts (data not shown). The difference in dimensions between the target list and the global matrix already indicates that even though we are mining a significant portion of lipids from these samples (200–300 lipid species, Tables S1 and S2 in Supplementary Material), the majority of the detectable peaks remains un-annotated.

GOBIOSPACE: A DATABASE SEARCH INTERFACE FOR MASS SPECTROMETRIC DATA

As shown in **Figure 6** the un-targeted global matrix, which contains all the extractable peaks from the recorded mass spectra, can be compared against a reference list of annotated compounds. The size and the content of these lists can vary significantly: therefore one can use the reference list generated in this study (Table S1 in

Supplementary Material) or other more comprehensive customer made lists. Furthermore public and commercial databases like, e.g., the Lipid Maps (Fahy et al., 2005, 2009), the KNApSAcK (Shinbo et al., 2006), KEGG (Kanehisa et al., 2008), PubChem (Wang et al., 2009), or ChemSpider (Williams, 2008) can be employed for even more comprehensive or specific database searches. The problem with these comparisons is that first of all not all these databases are easily accessible, but also even if they are, it still requires experience and personal effort with appropriate tools to compile these databases into a suitable resource. For this purpose we decided to develop a distributed client-server application utilizing a graphical user interface which supports the matching of measured masses to elemental compositions deposited in a relational database and make this tool publicly available.

We named this software GoBioSpace (for Golm Biochemical Space), which can be installed on Microsoft Windows XP Service Pack 3 and later desktop computers using the ClickOnce deployment[6]. The database server is accessed in-house directly using ADO.NET[7], while internet users fall back to WSDL-based [W3C (2001) Web Services Description Language (WSDL)[8]] web services[9].

The main functionality of GoBioSpace is to compare measured masses from mass spectrometric measurements, now including all kind of mass spectrometric data (high accurate mass but also lower mass accuracy), against a single or several databases (see Materials and Methods). As illustrated in **Figure 7**, the workflow for the data analysis is simple: a single mass or an elemental composition, but also a list of masses or formulas (tab-delimited text file) can be loaded into the software and searched against a single or several databases (at the moment more than 150 public databases are hosted, including the whole PubChem collection). Prior to the database search a number of parameters have to be specified, including the possible adducts of the measured mass (e.g., $[M+H]^+$, $[M+Na]^+$, $[M+NH_4]^+$, $[M-2H]^{2-}$, $[M-\text{Acetate}+H]^-$), the mass accuracy of the entered data, and finally a selection of elements expected to be contained in the matching compounds. The database search by itself (the in-house version) is quite fast and can process easily 2,000 searches per second, meaning that even a large list containing 30,000 peaks is processed within 15 s. However, reasoned by the increased complexity of protocol layers utilizing *xml* (eXtensible Markup Language)[10] and *http* (Hypertext Transport Protocol)[11] for data encapsulation and transport over the internet, we expect the performance of the internet version to fall below this value, also depending on the final capacity of the web and database servers. The output format of the result list, which is again a tab-delimited text file, contains all the information contained in the input table (measured m/z, RT and intensity of the measured peaks) added by the possible elemental composition of the measured mass, the adduct used to match measured and calculated mass, the database this hit was derived from, one or several compound name(s) if specified within the selected databases, and the mass error between the measured mass and the matched hit.

To re-validate our chromatographic data we searched the 30,000 peaks against an in-house assembled lipid database containing approximately 1,000 entries. This table contained the previously described lipids profiled in *Arabidopsis thaliana* lipids samples (Buseman et al., 2006; Devaiah et al., 2006; Markham et al., 2006; Markham and Jaworski, 2007; Glauser et al., 2008a,b), but also a large set of other lipid species including sterols (Benveniste, 2004; Hemmerlin et al., 2004), several di- and tri-acylglycerols, fatty acids, chlorophylls (Tanaka and Tanaka, 2006), and other plant pigments (Grotewold, 2006).

This database search resulted initially in a list of more than 4,000 hits for the positive mode spectra and 1,500 hits for the negative mode spectra. After correcting for the accurate adducts (**Table 1**) but also the expected retention times of the expected lipids within their lipid classes (Table S1 in Supplementary Material) we annotated, still very conservatively, 577 distinct peaks which were annotated to 265 unique elemental compositions (Tables S1 and S2 in Supplementary Material). Still, the number of hits within the already highly targeted database search seems to promise that this data-set contains many more compounds awaiting a proper annotation.

For overview purposes and to visualize the annotated data we mapped all the annotated lipids from Table S2 in Supplementary Material into a scatter plot (**Figure A1** in Appendix) and the different lipid classes and their distribution within the positive mode UPLC chromatogram (**Figure 8**).

PROS AND CONS OF DIFFERENT LIDOMIC STRATEGIES

The most common approach for systematic lipid profiling is still the well-established shotgun lipidomic approach (Han et al., 2005; Welti et al., 2007b; Yang et al., 2009), which was conceptually developed more than 15 years ago (Han and Gross, 1994). Due to this fact, there are several publications available (including comprehensive plant studies), which either made directly use of the QqQ approach (Welti and Wang, 2004; Devaiah et al., 2006; Welti et al., 2007b) or modified it for the use on different mass spectrometers like qTOF (Ekroos et al., 2002; Ejsing et al., 2006; Esch et al., 2007) or the Orbitrap (Yang et al., 2007. As a consequence different commercial and open source software packages were developed to make use of this kind of data (Ejsing et al., 2006 #127; Graessler et al., 2009; Yang et al., 2009; Herzog et al., 2011).

The developments and the application of LC–MS lipidomics, especially in the plant field, seems to be less popular, even though a number of groups developed different open source software packages for these applications (Haimi et al., 2006, 2009; Taguchi and Ishikawa, 2010; Nygren et al., 2011). The lack of absolute quantification, or better the lack of control of ion suppression in LC–MS-based lipidomic studies and the increased analytical complexity seem to be the main reasons for this discrepancy (Stahlman et al., 2009).

Ion suppression in shotgun lipidomic studies cannot be eliminated, even if lipid class-specific internal standards are used. The function of these internal standards is basically to corrected for the differential suppression effects on each measured lipid molecule

[6]http://msdn.microsoft.com/en-us/library/t71a733d.aspx
[7]http://msdn.microsoft.com/en-us/library/h43ks021%28v=VS.100%29.aspx
[8]http://www.w3.org/TR/wsdl
[9]http://gmd.mpimp-golm.mpg.de/webservices/wsGoBioSpace.asmx
[10]http://www.w3.org/XML/
[11]http://www.w3.org/Protocols/

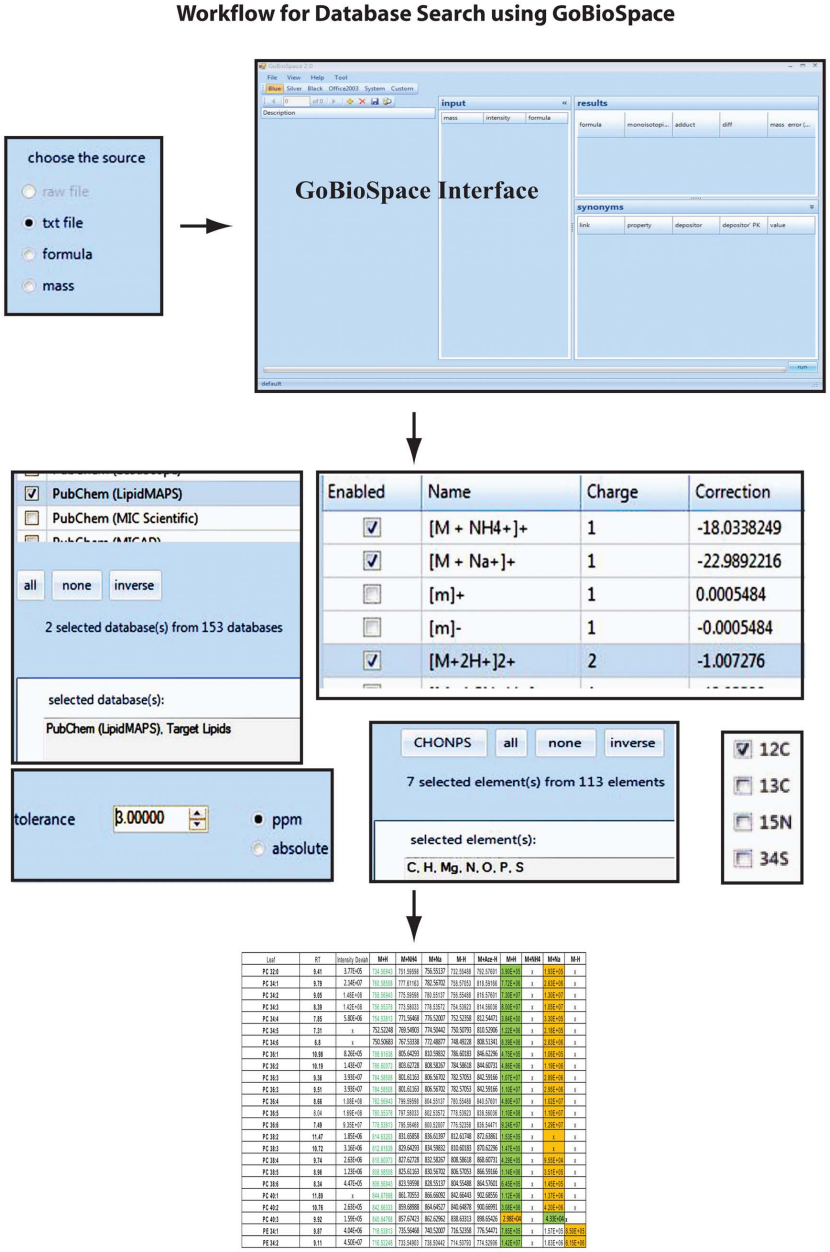


FIGURE 7 | Overview with screenshots of the GoBioSpace-assisted database search procedure. The workflow is separated in three steps: data input (single mass, mass list, or formula), specification of search criteria (databases, expected mass adducts, mass error tolerance, expected elements, and isotope label), and data output.

(Stahlman et al., 2009; Yang et al., 2009). Making use of mixtures of internal standards (in best case using one or two standard lipids per lipid class (Welti and Wang, 2004)) for LC–MS-based lipidomic studies could be possible if these mixtures are spiked in the eluting sample post-column, using a second pump and a t-connection. Such an on-line LC–MS approach using continuously infused internal standards at low concentrations, which has not been demonstrated yet, would definitely be an excellent compromise between complicated and time consuming off-line sample pre-fractionation (Stahlman et al., 2009), and the use of strongly ion-suppressed shotgun lipidomics. Our developed system could therefore provide an excellent test case for such an approach.

Alternatively, the use of fully labeled metabolomes or lipidomes (Ekroos et al., 2002; Hegeman et al., 2007; Giavalisco et al., 2008, 2009) could be an alternative way to quantify and annotate lipids in LC–MS-based studies. For this purpose analytical samples will be spiked with the same amount of the isotope-labeled matrix (Giavalisco et al., 2009). This approach, which has been tested by us (data not shown), is of course more complicated and expensive than the post-column spiking with a handful of reference

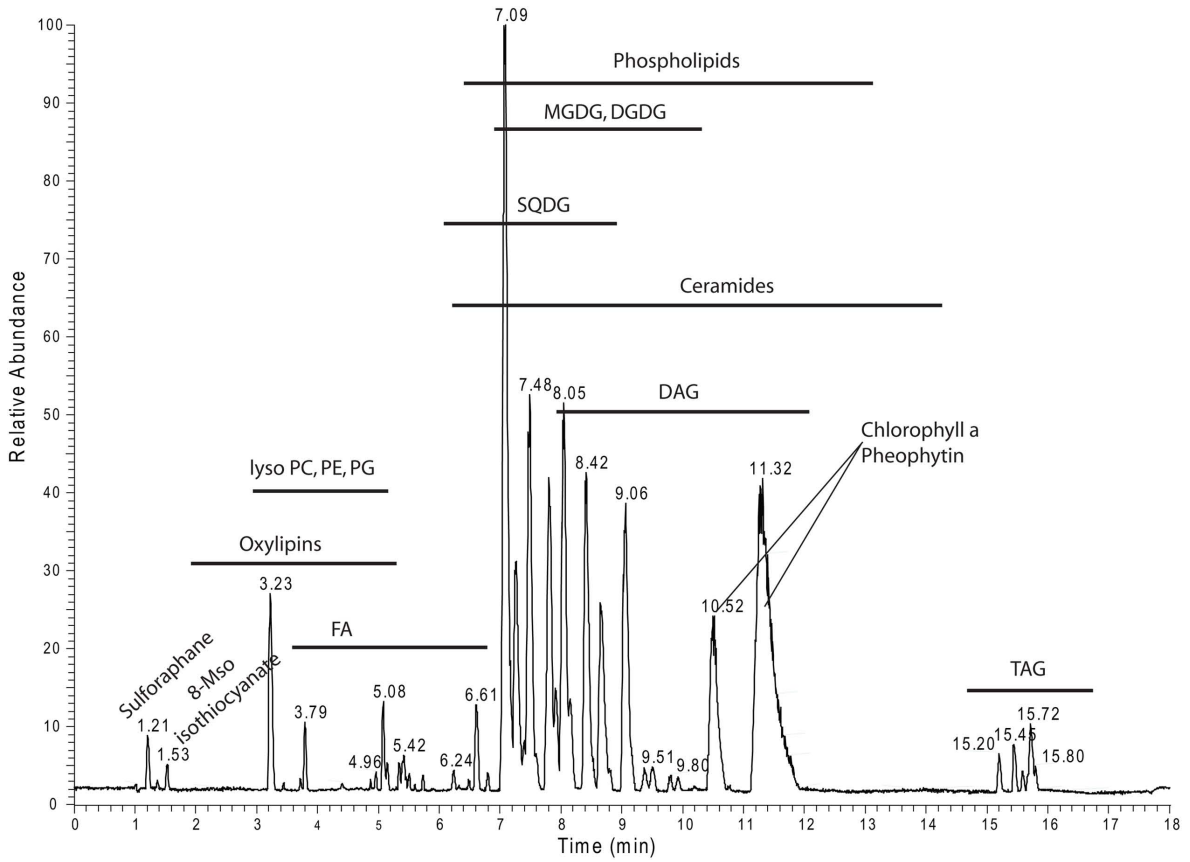


FIGURE 8 | Positive ion mode chromatogram from Figure 1A delineating the retention time areas of the different detected lipids from Table S2 in Supplementary Material.

compounds, but next to the relative quantification, it will also allow the reliable annotation of previously unknown compounds (Giavalisco et al., 2008, 2009).

ANNOTATING LIPIDS WITH DIFFERENT STRATEGIES: HOW MANY LIPIDS REMAIN UN-ANNOTATED?

One of the biggest differences between targeted and un-targeted lipid analysis lies in the fact that even though a number of 150 profiled and quantified lipids enables a meaningful analysis of an organism (Ejsing et al., 2009), there still remain many unidentified peaks to be annotated before we can really call it a lipidomic analysis. Looking at the data from our study already shows that of the 30,000 extractable peaks "only" 577 were annotated to a compound by using a targeted approach (Table S2 in Supplementary Material). Increasing the size of the employed databases would therefore directly provide a larger number of possible annotations, but this comes, in dependence of the database size used for the annotation, at the price of also annotating more false positives (Matsuda et al., 2009). Here the use of additional, orthogonal, physico-chemical properties can increase the validation of the recorded data. While the use of fragmentation data will greatly help to exclude false positives, also the use of the retention time information will improve the predictability of an annotation, which strongly argues in favor of LC–MS-based lipidomics (**Figures 4** and **8**).

Another advantage of LC–MS-based lipidomics in combination with global, un-targeted peak extraction lies in the statistically analyzed whole data-set consisting of 30,000 peaks prior to peak annotation. As a consequence, only the differential peaks would be regarded as potentially interesting and therefore subjected to more sophisticated peak annotation strategies. The annotation strategy could include isotope-labeling (see above) or analytical preparation techniques, including peak collection from the chromatographic run and subsequent analysis using higher order MS/MS, analysis on a high resolution mass spectrometer (Schwudke et al., 2007), or other orthogonal analytical techniques such as NMR.

COME BACK LATER: REVISITING OLD SPECTRA WITH NEW KNOWLEDGE

High resolution full scan and all-ion fragmentation spectra containing thousands of peaks are not only a rich source of biological information for a "one-pass" analysis but could serve as a repository of information, which can be reused with new knowledge repeatedly.

We demonstrated in our study that the use of targeted data, derived from a limited number of plant lipidomic studies (Buseman et al., 2006; Devaiah et al., 2006; Esch et al., 2007; Markham and Jaworski, 2007; Welti et al., 2007a,b; Glauser

et al., 2008a,b), allowed us to profile and annotate more than 260 lipid species. Increasing the list of targets by annotating novel lipid species, or simply checking literature for previously un-targeted lipids like *N*-acyl phosphatidylethanolamines (NAPE) and more complex sphingolipids (Welti and Wang, 2004), or tetra galactolipids (Moreau et al., 2008), will increase the length of the list of lipids which can be profiled. This includes the repercussive profiling of old data. Therefore, in the future more knowledge about thus far unidentified lipid moieties will allow us to annotate and profile more and more lipid species; we will not have to rerun all of our old experiments, since we can simply revisit our old high resolution chromatograms and reexamine them. This cannot be done using shotgun lipidomics with highly sensitive, but low resolution mass spectrometers.

ACKNOWLEDGMENTS

The authors wish to thank Prof. Dr. Lothar Willmitzer for discussions, support, and the opportunity to develop the project. Furthermore we would like to thank Dr. Leonard Krall for proof reading and commenting on the manuscript. Antony Williams is acknowledged for providing data from ChemSpider, while Drs. Dirk Walther and Joachim Kopka are acknowledged for hosting the GoBioSpace DB in the frame of the Golm Metabolome Database (GMD), which is currently funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under the LIS-program call Information Infrastructures for Research Projects, grant WA 1285/2-1 "Development of the Golm Metabolome Database as a central plant metabolomics information resource." Last but not least Änne Eckardt and Gudrun Wolter are most kindly acknowledged for the outstanding technical support.

REFERENCES

Annesley, T. M. (2003). Ion suppression in mass spectrometry. *Clin. Chem.* 49, 1041–1044.

Bateman, K. P., Castro-Perez, J., Wrona, M., Shockcor, J. P., Yu, K., Oballa, R., and Nicoll-Griffith, D. A. (2007). MSE with mass defect filtering for in vitro and in vivo metabolite identification. *Rapid Commun. Mass Spectrom.* 21, 1485–1496.

Bausch, J. N. (1993). Lipid analysis. *Curr. Opin. Biotechnol.* 4, 57–62.

Benton, H. P., Wong, D. M., Trauger, S. A., and Siuzdak, G. (2008). XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization. *Anal. Chem.* 80, 6382–6389.

Benveniste, P. (2004). Biosynthesis and accumulation of sterols. *Annu. Rev. Plant Biol.* 55, 429–457.

Blanksby, S. J., and Mitchell, T. W. (2010). Advances in mass spectrometry for lipidomics. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)* 3, 433–465.

Böhlke, J. K., De Laeter, J. R., De Bièvre, P., Hidaka, H., Peiser, H. S., Rosman, K. J. R., and Taylor, P. D. P. (2005). Isotopic compositions of the elements, 2001. *J. Phys. Chem. Ref. Data* 34, 57–67.

Buseman, C. M., Tamura, P., Sparks, A. A., Baughman, E. J., Maatta, S., Zhao, J., Roth, M. R., Esch, S. W., Shah, J., Williams, T. D., and Welti, R. (2006). Wounding stimulates the accumulation of glycerolipids containing oxo phytodienoic acid and dinor-oxophytodienoic acid in *Arabidopsis* leaves. *Plant Physiol.* 142, 28–39.

Chen, M., Markham, J. E., Dietrich, C. R., Jaworski, J. G., and Cahoon, E. B. (2008). Sphingolipid long-chain base hydroxylation is important for growth and regulation of sphingolipid content and composition in *Arabidopsis*. *Plant Cell* 20, 1862–1878.

Dennis, E. A. (2009). Lipidomics joins the omics evolution. *Proc. Natl. Acad. Sci. U.S.A.* 106, 2089–2090.

Devaiah, S. P., Pan, X., Hong, Y., Roth, M., Welti, R., and Wang, X. (2007). Enhancing seed quality and viability by suppressing phospholipase D in *Arabidopsis*. *Plant J.* 50, 950–957.

Devaiah, S. P., Roth, M. R., Baughman, E., Li, M., Tamura, P., Jeannotte, R., Welti, R., and Wang, X. (2006). Quantitative profiling of polar glycerolipid species from organs of wild-type *Arabidopsis* and a phospholipase Dalpha1 knockout mutant. *Phytochemistry* 67, 1907–1924.

Downes, C. P., and Currie, R. A. (1998). Lipid signalling. *Curr. Biol.* 8, R865–R867.

Ejsing, C. S., Duchoslav, E., Sampaio, J., Simons, K., Bonner, R., Thiele, C., Ekroos, K., and Shevchenko, A. (2006). Automated identification and quantification of glycerophospholipid molecular species by multiple precursor ion scanning. *Anal. Chem.* 78, 6202–6214.

Ejsing, C. S., Sampaio, J. L., Surendranath, V., Duchoslav, E., Ekroos, K., Klemm, R. W., Simons, K., and Shevchenko, A. (2009). Global analysis of the yeast lipidome by quantitative shotgun mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* 106, 2136–2141.

Ekroos, K., Chernushevich, I. V., Simons, K., and Shevchenko, A. (2002). Quantitative profiling of phospholipids by multiple precursor ion scanning on a hybrid quadrupole time-of-flight mass spectrometer. *Anal. Chem.* 74, 941–949.

Esch, S. W., Tamura, P., Sparks, A. A., Roth, M. R., Devaiah, S. P., Heinz, E., Wang, X., Williams, T. D., and Welti, R. (2007). Rapid characterization of the fatty acyl composition of complex lipids by collision-induced dissociation time-of-flight mass spectrometry. *J. Lipid Res.* 48, 235–241.

Fahy, E., Subramaniam, S., Brown, H. A., Glass, C. K., Merrill, A. H. Jr., Murphy, R. C., Raetz, C. R., Russell, D. W., Seyama, Y., Shaw, W., Shimizu, T., Spener, F., Van Meer, G., Vannieuwenhze, M. S., White, S. H., Witztum, J. L., and Dennis, E. A. (2005). A comprehensive classification system for lipids. *J. Lipid Res.* 46, 839–861.

Fahy, E., Subramaniam, S., Murphy, R. C., Nishijima, M., Raetz, C. R., Shimizu, T., Spener, F., Van Meer, G., Wakelam, M. J., and Dennis, E. A. (2009). Update of the LIPID MAPS comprehensive classification system for lipids. *J. Lipid Res.* 50, S9–S14.

Geiger, T., Cox, J., and Mann, M. (2010). Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol. Cell Proteomics* 9, 2252–2261.

Giavalisco, P., Hummel, J., Lisec, J., Inostroza, A. C., Catchpole, G., and Willmitzer, L. (2008). High-resolution direct infusion-based mass spectrometry in combination with whole 13C metabolome isotope labeling allows unambiguous assignment of chemical sum formulas. *Anal. Chem.* 80, 9417–9425.

Giavalisco, P., Kohl, K., Hummel, J., Seiwert, B., and Willmitzer, L. (2009). 13C isotope-labeled metabolomes allowing for improved compound annotation and relative quantification in liquid chromatography-mass spectrometry-based metabolomic research. *Anal. Chem.* 81, 6546–6551.

Giavalisco, P., Li, Y., Matthes, A., Eckhardt, A., Hubberten, H. M., Hesse, H., Segu, S., Hummel, J., Köhl, K., and Willmitzer, L. (2011). Elemental formula annotation of polar and lipophilic metabolites using (13) C, (15) N and (34) S isotope labelling, in combination with high-resolution mass spectrometry. *Plant J.* doi: 10.1111/j.1365-313X.2011.04682.x. [Epub ahead of print].

Glauser, G., Grata, E., Dubugnon, L., Rudaz, S., Farmer, E. E., and Wolfender, J. L. (2008a). Spatial and temporal dynamics of jasmonate synthesis and accumulation in Arabidopsis in response to wounding. *J. Biol. Chem.* 283, 16400–16407.

Glauser, G., Grata, E., Rudaz, S., and Wolfender, J. L. (2008b). High-resolution profiling of oxylipin-containing galactolipids in *Arabidopsis* extracts by ultra-performance liquid chromatography/time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* 22, 3154–3160.

Graessler, J., Schwudke, D., Schwarz, P. E., Herzog, R., Shevchenko, A., and Bornstein, S. R. (2009). Top-down lipidomics reveals ether lipid deficiency in blood plasma of hypertensive patients. *PLoS ONE* 4, e6261. doi: 10.1371/journal.pone.0006261

Griffiths, W. J., and Wang, Y. (2009). Mass spectrometry: from proteomics to metabolomics and lipidomics. *Chem. Soc. Rev.* 38, 1882–1896.

Grotewold, E. (2006). The genetics and biochemistry of floral pigments. *Annu. Rev. Plant Biol.* 57, 761–780.

Haimi, P., Chaithanya, K., Kainu, V., Hermansson, M., and Somerharju, P. (2009). Instrument-independent software tools for the analysis of MS-MS and LC-MS lipidomics data. *Methods Mol. Biol.* 580, 285–294.

Haimi, P., Uphoff, A., Hermansson, M., and Somerharju, P. (2006). Software tools for analysis of mass spectrometric lipidome data. *Anal. Chem.* 78, 8324–8331.

Han, X., and Gross, R. W. (1994). Electrospray ionization mass spectroscopic analysis of human erythrocyte plasma membrane phospholipids. *Proc. Natl. Acad. Sci. U.S.A.* 91, 10635–10639.

Han, X., and Gross, R. W. (2005). Shotgun lipidomics: electrospray ionization mass spectrometric analysis and quantitation of cellular lipidomes directly from crude extracts of biological samples. *Mass Spectrom. Rev.* 24, 367–412.

Han, X., Yang, K., Cheng, H., Fikes, K. N., and Gross, R. W. (2005). Shotgun lipidomics of phosphoethanolamine-containing lipids in biological samples after one-step in situ derivatization. *J. Lipid Res.* 46, 1548–1560.

Harkewicz, R., and Dennis, E. A. (2011). Applications of mass spectrometry to lipids and membranes. *Annu Rev Biochem.* 80, 301–325.

Hegeman, A. D., Schulte, C. F., Cui, Q., Lewis, I. A., Huttlin, E. L., Eghbalnia, H., Harms, A. C., Ulrich, E. L., Markley, J. L., and Sussman, M. R. (2007). Stable isotope assisted assignment of elemental compositions for metabolomics. *Anal. Chem.* 79, 6912–6921.

Hemmerlin, A., Gerber, E., Feldtrauer, J. F., Wentzinger, L., Hartmann, M. A., Tritsch, D., Hoeffler, J. F., Rohmer, M., and Bach, T. J. (2004). A review of tobacco BY-2 cells as an excellent system to study the synthesis and function of sterols and other isoprenoids. *Lipids* 39, 723–735.

Hermansson, M., Uphoff, A., Kakela, R., and Somerharju, P. (2005). Automated quantitative analysis of complex lipidomes by liquid chromatography/mass spectrometry. *Anal. Chem.* 77, 2166–2175.

Herzog, R., Schwudke, D., Schuhmann, K., Sampaio, J. L., Bornstein, S. R., Schroeder, M., and Shevchenko, A. (2011). A novel informatics concept for high-throughput shotgun lipidomics based on the molecular fragmentation query language. *Genome Biol.* 12, R8.

Huege, J., Sulpice, R., Gibon, Y., Lisec, J., Koehl, K., and Kopka, J. (2007). GC-EI-TOF-MS analysis of in vivo carbon-partitioning into soluble metabolite pools of higher plants by monitoring isotope dilution after (13)CO(2) labelling. *Phytochemistry* 68, 2258–2272.

Hummel, J., Strehmel, N., Selbig, J., Walther, D., and Kopka, J. (2010). Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics* 6, 322–333.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480–D484.

Katajamaa, M., Miettinen, J., and Oresic, M. (2006). MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 22, 634–636.

Katajamaa, M., and Oresic, M. (2007). Data processing for mass spectrometry-based metabolomics. *J. Chromatogr. A* 1158, 318–328.

Kilaru, A., Isaac, G., Tamura, P., Baxter, D., Duncan, S. R., Venables, B. J., Welti, R., Koulen, P., and Chapman, K. D. (2010). Lipid profiling reveals tissue-specific differences for ethanolamide lipids in mice lacking fatty acid amide hydrolase. *Lipids* 45, 863–875.

Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmüller, E., Dörmann, P., Weckwerth, W., Gibon, Y., Stitt, M., Willmitzer, L., Fernie, A. R., Steinhauser, D. (2005). GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* 21, 1635–1638.

Kuksis, A. (2007). Lipidomics in triacylglycerol and cholesteryl ester oxidation. *Front. Biosci.* 12, 3203–3246.

Lommen, A. (2009). MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal. Chem.* 81, 3079–3086.

Lu, W., Bennett, B. D., and Rabinowitz, J. D. (2008). Analytical strategies for LC-MS-based targeted metabolomics. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 871, 236–242.

Lu, W., Clasquin, M. F., Melamud, E., Amador-Noguez, D., Caudy, A. A., and Rabinowitz, J. D. (2010). Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. *Anal. Chem.* 82, 3212–3221.

Markham, J. E., and Jaworski, J. G. (2007). Rapid measurement of sphingolipids from *Arabidopsis thaliana* by reversed-phase high-performance liquid chromatography coupled to electrospray ionization tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 21, 1304–1314.

Markham, J. E., Li, J., Cahoon, E. B., and Jaworski, J. G. (2006). Separation and identification of major plant sphingolipid classes from leaves. *J. Biol. Chem.* 281, 22684–22694.

Matsuda, F., Shinbo, Y., Oikawa, A., Hirai, M. Y., Fiehn, O., Kanaya, S., and Saito, K. (2009). Assessment of metabolome annotation quality: a method for evaluating the false discovery rate of elemental composition searches. *PLoS ONE* 4, e7490. doi: 10.1371/journal.pone.0007490

Matyash, V., Liebisch, G., Kurzchalia, T. V., Shevchenko, A., and Schwudke, D. (2008). Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. *J. Lipid Res.* 49, 1137–1146.

Mcluckey, S. A., and Wells, J. M. (2001). Mass analysis at the advent of the 21st century. *Chem. Rev.* 101, 571–606.

Merrill, A. H. Jr., Stokes, T. H., Momin, A., Park, H., Portz, B. J., Kelly, S., Wang, E., Sullards, M. C., and Wang, M. D. (2009). Sphingolipidomics: a valuable tool for understanding the roles of sphingolipids in biology and disease. *J. Lipid Res.* 50, S97–S102.

Moore, J. D., Caufield, W. V., and Shaw, W. A. (2007). Quantitation and standardization of lipid internal standards for mass spectroscopy. *Methods Enzymol.* 432, 351–367.

Moreau, R. A., Doehlert, D. C., Welti, R., Isaac, G., Roth, M., Tamura, P., and Nunez, A. (2008). The identification of mono-, di-, tri-, and tetragalactosyl-diacylglycerols and their natural estolides in oat kernels. *Lipids* 43, 533–548.

Muller, C., Schafer, P., Stortzel, M., Vogt, S., and Weinmann, W. (2002). Ion suppression effects in liquid chromatography-electrospray-ionisation transport-region collision induced dissociation mass spectrometry with different serum extraction methods for systematic toxicological analysis with mass spectra libraries. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 773, 47–52.

Nakanishi, H., Ogiso, H., and Taguchi, R. (2009). Qualitative and quantitative analyses of phospholipids by LC-MS for lipidomics. *Methods Mol. Biol.* 579, 287–313.

Nygren, H., Seppanen-Laakso, T., Castillo, S., Hyotylainen, T., and Oresic, M. (2011). Liquid chromatography-mass spectrometry (LC-MS)-based lipidomics for studies of body fluids and tissues. *Methods Mol. Biol.* 708, 247–257.

Oldiges, M., Lutz, S., Pflug, S., Schroer, K., Stein, N., and Wiendahl, C. (2007). Metabolomics: current state and evolving methodologies and tools. *Appl. Microbiol. Biotechnol.* 76, 495–511.

Oliver, S. G., Winson, M. K., Kell, D. B., and Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends Biotechnol.* 16, 373–378.

Picchioni, G. A., Watada, A. E., and Whitaker, B. D. (1996). Quantitative high-performance liquid chromatography analysis of plant phospholipids and glycolipids using light-scattering detection. *Lipids* 31, 217–221.

Plumb, R., Castro-Perez, J., Granger, J., Beattie, I., Joncour, K., and Wright, A. (2004). Ultra-performance liquid chromatography coupled to quadrupole-orthogonal time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* 18, 2331–2337.

Pluskal, T., Castillo, S., Villar-Briones, A., and Oresic, M. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 11, 395. doi: 10.1186/1471-2105-11-395

Rainville, P. D., Stumpf, C. L., Shockcor, J. P., Plumb, R. S., and Nicholson, J. K. (2007). Novel application of reversed-phase UPLC-oaTOF-MS for lipid analysis in complex biological mixtures: a new tool for lipidomics. *J. Proteome Res.* 6, 552–558.

Schmitt-Kopplin, P., Englmann, M., Rossello-Mora, R., Schiewek, R., Brockmann, K. J., Benter, T., and Schmitz, O. J. (2008). Combining chip-ESI with APLI (cESILI) as a multimode source for analysis of complex mixtures with ultrahigh-resolution mass spectrometry. *Anal. Bioanal. Chem.* 391, 2803–2809.

Schwudke, D., Hannich, J. T., Surendranath, V., Grimard, V., Moehring, T., Burton, L., Kurzchalia, T., and Shevchenko, A. (2007). Top-down lipidomic screens by multivariate analysis of high-resolution survey mass spectra. *Anal. Chem.* 79, 4083–4093.

Shinbo, Y., Nakamura, Y., Altaf-Ul-Amin, M., Asah, H., Kurokawa, K., Arita, M., Saito, K., Ohta, D., Shibata, D., and Kanaya, S. (2006). KNApSAcK: a comprehensive species-metabolite relationship database. *Biotechnol. Agr. Forest.* 57, 165–181.

Smith, C. A., Want, E. J., O'maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 78, 779–787.

Spiegel, S., and Milstien, S. (2003). Sphingosine-1-phosphate: an enigmatic signalling lipid. *Nat. Rev. Mol. Cell Biol.* 4, 397–407.

Stahlman, M., Ejsing, C. S., Tarasov, K., Perman, J., Boren, J., and Ekroos, K. (2009). High-throughput shotgun lipidomics by quadrupole time-of-flight mass spectrometry. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 877, 2664–2672.

Taguchi, R., and Ishikawa, M. (2010). Precise and global identification of phospholipid molecular species by an Orbitrap mass spectrometer and automated search engine lipid search. *J. Chromatogr. A* 1217, 4229–4239.

Tanaka, A., and Tanaka, R. (2006). Chlorophyll metabolism. *Curr. Opin. Plant Biol.* 9, 248–255.

Touchstone, J. C. (1995). Thin-layer chromatographic procedures for lipid separation. *J. Chromatogr. B Biomed. Appl.* 671, 169–195.

Van Meer, G., Voelker, D. R., and Feigenson, G. W. (2008). Membrane lipids: where they are and how they behave. *Nat. Rev. Mol. Cell Biol.* 9, 112–124.

Vuckovic, D., Zhang, X., Cudjoe, E., and Pawliszyn, J. (2010). Solid-phase microextraction in bioanalysis: new devices and directions. *J. Chromatogr. A* 1217, 4041–4060.

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., and Bryant, S. H. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* 37, W623–W633.

Welti, R., Mui, E., Sparks, A., Wernimont, S., Isaac, G., Kirisits, M., Roth, M., Roberts, C. W., Botte, C., Marechal, E., and Mcleod, R. (2007a). Lipidomic analysis of *Toxoplasma gondii* reveals unusual polar lipids. *Biochemistry* 46, 13882–13890.

Welti, R., Shah, J., Li, W., Li, M., Chen, J., Burke, J. J., Fauconnier, M. L., Chapman, K., Chye, M. L., and Wang, X. (2007b). Plant lipidomics: discerning biological function by profiling plant complex lipids using mass spectrometry. *Front. Biosci.* 12, 2494–2506.

Welti, R., and Wang, X. (2004). Lipid species profiling: a high-throughput approach to identify lipid compositional changes and determine the function of genes involved in lipid metabolism and signaling. *Curr. Opin. Plant Biol.* 7, 337–344.

Wenk, M. R. (2005). The emerging field of lipidomics. *Nat. Rev. Drug Discov.* 4, 594–610.

Wenk, M. R. (2010). Lipidomics: new tools and applications. *Cell* 143, 888–895.

Williams, A. J. (2008). Public chemical compound databases. *Curr. Opin. Drug Discov. Devel.* 11, 393–404.

Wymann, M. P., and Schneiter, R. (2008). Lipid signalling in disease. *Nat. Rev. Mol. Cell Biol.* 9, 162–176.

Xu, Y., Heilier, J. F., Madalinski, G., Genin, E., Ezan, E., Tabet, J. C., and Junot, C. (2010). Evaluation of accurate mass and relative isotopic abundance measurements in the LTQ-orbitrap mass spectrometer for further metabolomics database building. *Anal. Chem.* 82, 5490–5501.

Yang, K., Cheng, H., Gross, R. W., and Han, X. (2009). Automated lipid identification and quantification by multidimensional mass spectrometry-based shotgun lipidomics. *Anal. Chem.* 81, 4356–4368.

Yang, K., Zhao, Z., Gross, R. W., and Han, X. (2007). Shotgun lipidomics identifies a paired rule for the presence of isomeric ether phospholipid molecular species. *PLoS ONE* 2, e1368. doi: 10.1371/journal.pone.0001368

Zhang, Y., Zhu, H., Zhang, Q., Li, M., Yan, M., Wang, R., Wang, L., Welti, R., Zhang, W., and Wang, X. (2009). Phospholipase dalpha1 and phosphatidic acid regulate NADPH oxidase activity and production of reactive oxygen species in ABA-mediated stomatal closure in *Arabidopsis*. *Plant Cell* 21, 2357–2377.

Conflict of Interest Statement: 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.

Received: 08 May 2011; accepted: 05 September 2011; published online: 12 October 2011.

Citation: Hummel J, Segu S, Li Y, Irgang S, Jueppner J and Giavalisco P (2011) Ultra performance liquid chromatography and high resolution mass spectrometry for the analysis of plant lipids. Front. Plant Sci. ***2****:54. doi: 10.3389/fpls.2011.00054*

This article was submitted to Frontiers in Plant Physiology, a specialty of Frontiers in Plant Science.

APPENDIX

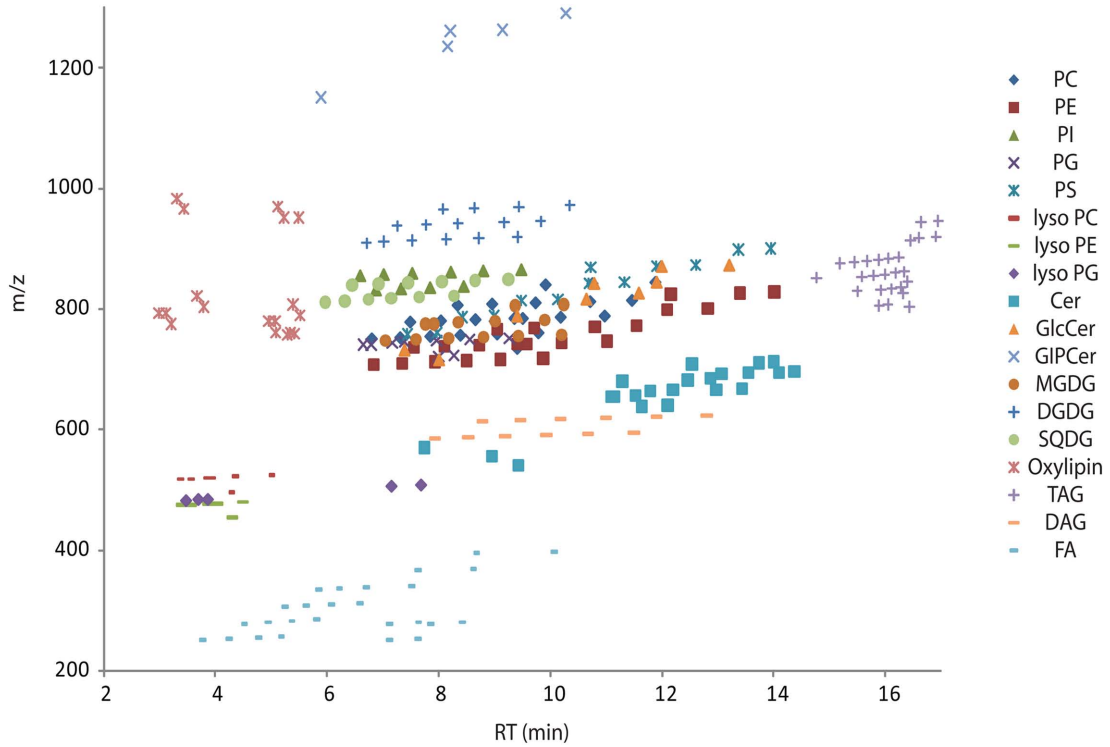


FIGURE A1 | Overview scatter plot of all lipids from Table S2 in Supplementary Material. The plot contains the measured retention time in minutes on the *x*-axes and the recorded mass of the $[M+H]^+$ adduct on the *y*-axes. Due to the modular building block structure of lipids within a homogenous class, systematic patterns of parallel lines should be observed. From bottom to top these lines should contain lipids with increasing fatty acid chain length, while the number of double bond should decrease from left to right. Abbreviations are as follows: PC, phosphatidylcholine; PE, phosphatidylethanolamine; PG, phosphatidylglycerol; PI, phosphatidylinositol; PS, phosphatidylserine; MGDG, monogalactosyldiacylglycerol; DGDG, digalactosyldiacylglycerol; SQDG, Sulfoquinovosyldiacylglycerol; Cer, ceramide; GlcCer, glucosylceramides; GIPC, glycosylinositolphosphoceramides; TAG, triacylglycerols; DAG, diacylglycerols; FA, fatty acids.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 12 July 2011
doi: 10.3389/fpls.2011.00029

Metabolite signature during short-day induced growth cessation in Populus

Miyako Kusano[1]*, Pär Jonsson[2], Atsushi Fukushima[1], Jonas Gullberg[3], Michael Sjöström[2], Johan Trygg[2] and Thomas Moritz[3]*

[1] *Metabolomics Research Division, RIKEN Plant Science Center, Yokohama, Japan*
[2] *Computational Life Science Cluster, Department of Chemistry, Umeå University, Umeå, Sweden*
[3] *Department of Forest Genetics and Plant Physiology, Umeå Plant Science Centre, Swedish University of Agricultural Sciences, Umeå, Sweden*

Edited by:
Alisdair Fernie, Max Planck Institute for Plant Physiology, Germany

Reviewed by:
Kris Morreel, University Ghent, Belgium
Joachim Selbig, University of Potsdam, Germany

****Correspondence:***
Miyako Kusano, RIKEN Plant Science Center, 1-7-22 Suehiro, Tsurumi, Yokohama 230-0045, Japan.
e-mail: mkusano005@psc.riken.jp;
Thomas Moritz, Department of Forest Genetics and Plant Physiology, Umeå Plant Science Centre, Swedish University of Agricultural Sciences, SE-901 87 Umeå, Sweden.
e-mail: thomas.moritz@genfys.slu.se

The photoperiod is an important environmental signal for plants, and influences a wide range of physiological processes. For woody species in northern latitudes, cessation of growth is induced by short photoperiods. In many plant species, short photoperiods stop elongational growth after a few weeks. It is known that plant daylength detection is mediated by *Phytochrome A* (*PHYA*) in the woody hybrid aspen species. However, the mechanism of dormancy involving primary metabolism remains unclear. We studied changes in metabolite profiles in hybrid aspen leaves (young, middle, and mature leaves) during short-day-induced growth cessation, using a combination of gas chromatography–time-of-flight mass spectrometry, and multivariate projection methods. Our results indicate that the metabolite profiles in mature source leaves rapidly change when the photoperiod changes. In contrast, the differences in young sink leaves grown under long and short-day conditions are less distinct. We found short daylength induced growth cessation in aspen was associated with rapid changes in the distribution and levels of diverse primary metabolites. In addition, we conducted metabolite profiling of leaves of *PHYA* overexpressor (PHYAOX) and those of the control to find the discriminative metabolites between PHYAOX and the control under the short-day conditions. The metabolite changes observed in PHYAOX leaves, together with those in the source leaves, identified possible candidates for the metabolite signature (e.g., 2-oxo-glutarate, spermidine, putrescine, 4-amino-butyrate, and tryptophan) during short-day-induced growth cessation in aspen leaves.

Keywords: aspen, *Phytochrome A*, metabolite profiling, GC–TOF-MS, dormancy, growth cessation

INTRODUCTION

The initiation of cold acclimation and dormancy for tree species in northern latitudes is synchronized with the end of the growth season and the onset of low temperatures in the autumn. Elongation growth stops in many woody species with indeterminate growth patterns after a few weeks under short photoperiods in controlled environments (Thomas and Vince-Prue, 1997). The photoperiodic timing of growth processes are dependent on photoreceptors that detect the day/night cycle, along with an endogenous circadian oscillator that perceives and resets the endogenous clock according to the environmental conditions (Eriksson and Millar, 2003; Schultz and Kay, 2003; Salome and Mcclung, 2005; Mcclung, 2008; Hoffman et al., 2010).

In *Populus*, the site of daylength detection and cessation of stem elongation is in the leaf-rib meristem area (Ruonala et al., 2008). One of the main players in the signaling pathways involved in short-day-induced growth cessation is the flowering locus T (FT) protein. However, flower induction generally occurs by transmission of FT and its ortholog from phloem to the shoot apex in *Arabidopsis* (Corbesier et al., 2007) and in rice (Tamaki et al., 2007). Studies of the *FT* in aspen species (*Populus* spp.), suggest that *FT* has dual roles in wood species (Bohlenius et al., 2006). Research indicates that FT is involved in the regulation of both flowering and short-day-induced growth cessation. Transgenic poplars overexpressing the *Populus FT*-gene under the cauliflower mosaic virus 35S-promotor do not form buds during short photoperiods (Bohlenius et al., 2006). This suggests that the FT protein is a key mobile regulator of daylength-controlled shoot elongation in *Populus*, similar to the FT protein role in flowering in *Arabidopsis* (Corbesier et al., 2007) and rice (Tamaki et al., 2007).

Previous research suggests that plant daylength detection is mediated by Phytochrome A (PHYA) in the woody hybrid aspen species, *Populus tremula* × *P. tremuloides* (Olsen et al., 1997; Kozarewa et al., 2010). Antisense PHYA hybrid aspen shows earlier bud-set in short winter photoperiod than in the corresponding wild-type (WT) plants (Kozarewa et al., 2010), and hybrid aspens expressing oat *PHYA* are severely dwarfed and insensitive to induction of dormancy by short days (Olsen et al., 1997). Furthermore, levels of the gibberellin group of plant hormones (GAs) are down-regulated in the *PHYA* overexpressing hybrid aspen (PHYAOX). Researchers have hypothesized that this is the reason for the loss of induced growth cessation under short photoperiods in hybrid aspen (Olsen et al., 1995; Eriksson and Moritz, 2002). Nevertheless, many other possible candidates acting as a transmittable metabolic signal that mediate photoperiod controlled elongation exist, including plant hormones (Baba et al., 2011) and primary and secondary metabolites (Ruttink et al., 2007).

In metabolomics, the goal is to identify and quantify every metabolite in a biological system (Fiehn, 2002; Fernie et al., 2004; Hall, 2006). Although this method is not technologically feasible (Saito and Matsuda, 2010), relevant metabolic profiles of different samples can be obtained and contrasted. Using a metabolomics approach, we investigated metabolite profiles in leaves at different developmental stages using hybrid aspen (*P. tremula* × *P. tremuloides*) during short-day-induced growth cessation. Our aim was to reveal at which developmental stage of the foliar metabolite responses to changes in photoperiod were most prominent. The metabolite profiles of those leaves were further explored in PHYAOX and the controls. The observed metabolic changes may provide candidates for the "metabolic signature" of short-day-induced growth cessation in hybrid aspen leaves. These candidates might belong to the PHYA-associated signaling pathways or primary metabolic responses. However, because PHYAOX are dwarfed, differential metabolites might be related with the developmental shift in these transgenic poplars rather than with their photoperiod insensitivity. Therefore, developmental shift- or age-dependent metabolites were annotated in a separate time-course study of wild-type leaves taking into account the photoperiod.

MATERIALS AND METHODS

PLANT GROWTH AND HARVESTING

Twenty-eight hybrid aspen trees were grown under long photoperiods (long-day conditions, LDs) of 18 h/6 h day/night cycle using photosynthetic active radiation (PAR) light for 12 h at 400 µmol $m^{-2}s^{-1}$, and extended for another 6 h at 30 µmol $m^{-2}s^{-1}$. After 3 months, 20 consecutive leaves (the length of the first leaf below the apex was approximately 1 cm) and the apex (defined as apical tissue from which all major leaf primordia had been removed) were sampled from seven plants (LD_0-samples). Two days later another seven plants were sampled (LD_2-samples), and the daylength was changed to short winter photoperiods (short-day conditions, SDs): 12/12 h day/night). After 2 and 6 days under short photoperiods (SD_2 and SD_6), seven plants were sampled. To examine the effect of 35S:: oat *PHYA*-overexpression, we sampled 10 leaves from 22-PHYAOX and 21-control plants. We used two different photoperiods [LD_0 and short photoperiod at day 7 (SD_7)] to obtain PHYOX and the control plants (wild-type, WT). After removal from the plant samples were dipped in liquid nitrogen and stored at –80°C until required.

METABOLITE PROFILING ANALYSIS

Leaf samples were crushed, extracted, and their metabolite profiles were analyzed according to (Gullberg et al., 2004). Stable isotope reference compounds (15 ng $µl^{-1}$ each of [$^{13}C_3$]-myristic acid, [$^{13}C_4$]-hexadecanoic acid, [$^{2}H_4$]-succinic acid, [$^{13}C_5$, ^{15}N]-glutamic acid, [$^{2}H_7$]-cholesterol, [$^{13}C_5$]-proline, [$^{13}C_4$]-disodium α-ketoglutarate, [$^{13}C_{12}$]-sucrose, [$^{2}H_4$]-putrescine, [$^{2}H_6$]-salicylic acid, and [$^{13}C_6$]-glucose) were added to an extraction mixture of chloroform:MeOH:H_2O (3:1:1). The samples (10 mg fresh weight each) were then extracted in 1 ml of the extraction mixture using a MM 301 Vibration Mill (Retsch GmbH & Co. KG, Haan, Germany) at a frequency of 30 Hz s^{-1} for 3 min using a 3-mm of tungsten carbide bead (Retsch GmbH & Co. KG, Haan, Germany) per tube to increase the extraction efficiency. After extraction samples were placed in an Eppendorf centrifuge (Model 5417C) for 10 min at 14,000 rpm. Following this, 200 µl of the supernatant was transferred to a GC-vial and evaporated to dryness. The samples were then derivatized by shaking them with 30 µl of methoxyamine hydrochloride (15 mg ml^{-1}) in pyridine for 10 min at 5°C. Samples were then incubated overnight at room temperature. The samples were then trimethylsilylated by adding 30 µl of MSTFA with 1% TMCS and incubating for 1 h at room temperature. After silylation, 30 µl of heptane was added.

The samples were analyzed according to Gullberg et al. (2004) using gas chromatography–time-of-flight mass spectrometry (GC–MS). We used blank control samples and a series of *n*-alkanes (C_{12}–C_{40}) to allow us to calculate retention indices (Schauer et al., 2005). One microliter of each derivatized sample was injected using a split/splitless injector in splitless mode of an Agilent 7683 autosampler (Agilent, Atlanta, GA, USA) into an Agilent 6890 gas chromatograph equipped with a 10-m × 0.18-mm i.d. fused silica capillary column with a chemically bonded 0.18 µm DB 5-MS stationary phase (J&W Scientific, Folsom, CA, USA). The injector temperature was 270°C, the septum purge flow rate was 20 ml min^{-1} and the purge was turned on after 60 s. The gas flow rate through the column was 1 ml min^{-1}, the column temperature was held at 70°C for 2 min, then increased by 40°C min^{-1} to 320°C, and held for 2 min. The column effluent was introduced into the ion source of a Pegasus III time-of-flight mass spectrometer, GC–MS (LECO Corp., St Joseph, MI, USA). The transfer line and the ion source temperatures were 250 and 200°C, respectively. Ions were generated by a 70-eV electron beam at an ionization current of 2.0 mA, and 30 spectra s^{-1} were recorded in the mass range 50–800 m/z. The acceleration voltage was turned on after a solvent delay.

All non-processed metabolite profile data were exported from the ChromaTOF software in NetCDF format to MATLAB™ software 7.0 (MathWorks, Natick, MA, USA), in which data pretreatment procedures such as base-line correction chromatogram alignment, data compression, and hierarchical multivariate curve resolution (H-MCR), were performed using custom scripts following Jonsson et al. (2005). All manual integrations were performed using ChromaTOF 2.00 software (LECO Corp., St Joseph, MI, USA) or custom scripts as described in Kusano et al. (2007).

STATISTICAL DATA ANALYSIS

Multivariate statistical investigations were performed using SIMCA-P + 12 software (Umetrics, Umeå, Sweden). All variables were log_{10}-transformed, centered, and scaled to unit variance for the analysis. To connect the information of two-block variables (*X* and *Y*) to each other, we used an orthogonal projection to latent structures (OPLS). OPLS is one of the supervised methods which is commonly applied in metabolomics. An OPLS regression model (Trygg and Wold, 2002) was calculated to investigate potential relationships between the metabolic compositions (*X*) of the aspen leaves and their positions (*Y*) on the stem. Peak areas under the resolved GC–MS peaks were used as descriptors (*X*) and the leaf positions as the response (*Y*) in the OPLS model. R^2X is the cumulative modeled variation in *X*, R^2Y is the cumulative modeled variation in *Y*, and Q^2Y is the cumulative predicted variation in *Y*, according to cross-validation. The range of these parameters is 0–1, where 1 indicates a perfect fit.

To determine metabolites which were affected only by time periods (i.e., day 0, day 2, day 4, and day 6), or only by daylength (i.e., LD or SD), which do not show any interaction, we used a two-way analysis of variance (ANOVA) as described in Pavlidis (2003). Here, we assume that the metabolite level was expressed as

$$X_{ijk} = \mu + T_i + L_j + (T \cdot L)_{ij} + \varepsilon_{ijk},$$
$$i = 1,\ldots,n,$$
$$j = 1,\ldots,m,$$
$$k = 1,\ldots,p.$$

This indicates a linear model of metabolite accumulation in replicate k of level i of factor T (time-period) and level j of factor L (daylength) with n and m levels, respectively; p represents replicates per group; μ is the mean metabolite level, and ε represents random error. The level of significance was set at $p < 0.05$ when corrected for the false-discovery rate (FDR) method (Benjamini and Hochberg, 1995). In PHYAOX samples, we calculated two-way ANOVA (factors: genotype × daylength). The ANOVA analyses were carried out using the R statistical environment (http://cran.r-project.org).

RESULTS

DESIGN AND EXPERIMENTAL SET-UP FOR SAMPLING

We investigated which positions on hybrid aspen trees were representative of young (sink), middle, and mature (source) leaves for this study, to obtain insights into the extent of differences of metabolite composition of leaf samples across different developmental stages (**Figure 1A**). To validate the sampling strategy, a preliminary study was conducted with an aspen plant grown in LDs (18 h). After 3 weeks, 20 consecutive leaves were sampled, from the first leaf below the apex approximately 1 cm long (which was numbered 1, **Figures 1A,B**). This sampling strategy provided a sequence of 20 leaves in different developmental stages, ranging from actively growing sink leaves to mature source leaves. Plant metabolites were extracted, derivatized, and analyzed from leaf samples using GC–MS (Gullberg et al., 2004; Jonsson et al., 2005). The OPLS model obtained showed a clear relationship between the leaf number and the corresponding metabolite profile (**Figure 1C**). Substantial differences between the leaves were anticipated from the differences in their developmental stages. Although similar age-related differences have previously been found in metabolite profiles of *Populus* leaves (Jeong et al., 2004), the validation result demonstrated that evaluation of the sampling strategy using OPLS is important when examining a large number of plants. Therefore, we chose leaf 2, leaf 10, and leaf 20 as representative of young sink, middle, and mature source leaves, respectively.

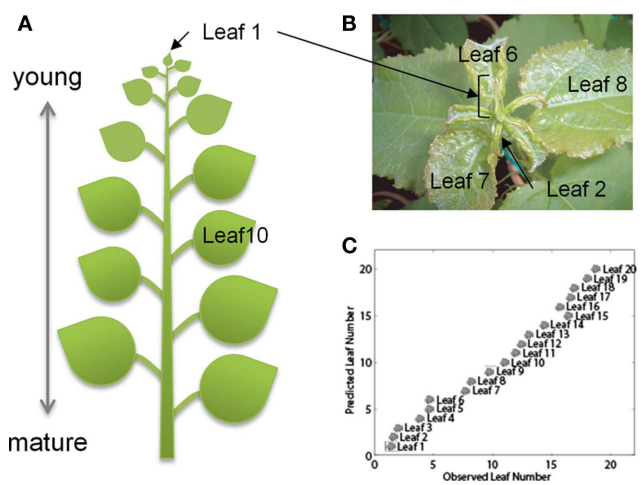


FIGURE 1 | The OPLS model of the metabolite profile of *Populus* leaves at different sampling positions. (A) The leaf numbers (leaf) refers to the sampling position, where leaf one is the first leaf longer than 1 cm. We sampled leaves from position 1 (leaf 1) to position 20 (leaf 20). **(B)** Expansion of upper side including apex, leaf 1 to leaf 8 of an aspen plant. **(C)** The number of components in the OPLS model was determined as two (one orthogonal and one predictive) according to seven-fold full cross-validation (Wold, 1978). The model explains 98.1% of the variation in Y ($R^2Y = 0.981$) and the estimated ability to predict 92.1% of the variation in Y ($Q^2Y = 0.921$) according to cross-validation. The model was able to model 47.8% of the variation in X, 22.9% of the variation is correlated to leaf position (Y) and 24.9% is uncorrelated.

MATURE SOURCE LEAVES SHOW RAPID RESPONSES TO CHANGES IN DAYLENGTH

The second step in the study was to investigate how the metabolite profiles of aspen leaves sampled from different developmental stages differ under long- and short-day conditions (18 and 12 h respectively). Metabolite profiling was conducted on leaf samples (see Materials and Methods above) from 28 aspen plants grown continuously in LD, or from plants grown first under LD and then SD (for 2 or 6 days). The data, including unknown and annotated peaks, were first evaluated by principal component analysis (PCA) in an unsupervised manner (**Figure A1** in Appendix). The PCA of leaf 20 samples shows a clear separation between the different photoperiods on the first component 1 (**Figure A1C** in Appendix). However, the PCA scores of leaf 2 and 10 samples revealed no clear photoperiod differences s (**Figures A1A,B** in Appendix). Therefore, any further analysis of metabolite data were carried out on samples from the leaf 20 position. The supervised method orthogonal projection to latent structures discriminant analysis (OPLS-DA) was used for metabolite profiling data of leaf 20 samples to maximize the information related to the differences in the four different photoperiods (**Figure 2**). For leaf 20, the LD_0-samples were predicted to be similar to LD_2, with the SD_2-samples to be intermediate between LD_2 and SD_6 (**Figure 2B**). This is consistent with the hypothesis that samples grown for two more days in LD (LD_2) should group with the LD_0-samples. Plants exposed to only two SD days should have intermediate profiles between those of LD_2 and SD_6. The loading plot suggests that the most of the detected peaks increased in their levels during short-day treatment (**Figure 2B**).

METABOLIC ALTERNATIONS BETWEEN LD AND SD CONDITIONS IN MATURE LEAVES

To identify the metabolites contributing the differences between the LD and SD samples in mature leaves, we conducted a two-way ANOVA (factors: time periods × daylength) to determine metabolites that showed significant changes between LD and SD. Among 454 peaks, 12 peaks showed significant changes in accordance with different photoperiods (LD and SD) after the FDR correction (**Figure A2** in Appendix; **Data Sheet 1** in Supplementary Material).

These changes were visualized in box plots (**Figure A2** in Appendix). Furthermore, 303 peaks, such as intermediates belonging to tricarboxylic acid (TCA) cycle, showed significant changes with different time periods (**Data Sheet 1** in Supplementary Material) and reflect rather developmental or age-associated changes in the leaf. Of the 12 metabolite peaks, serine, aspartate, pyroglutamate, glutamate, and three unknown peaks also showed significant alternations across different days of growth (**Data Sheet 1** in Supplementary Material).

METABOLITE PROFILING OF LEAVES OF PHYAOX GROWN UNDER LD AND SD

Metabolite profiling of WT leaves enabled us to find the metabolites that altered their levels between different photoperiods. To investigate which metabolites remain unchanged in the metabolite profiles of PHYAOX, but changed in those of WT between LD and SD, we performed metabolite profiling of leaves of PHYAOX and the control plants (WT). We conducted a PCA to investigate the distribution of PHYAOX and WT samples under LD and SD conditions (**Figure 3**). The PCA plot showed that metabolite profiles of PHYAOX and those of WT differ in the first component and difference between LD and SD appeared in the second component (**Figure 3**). The results of OPLS-DA suggested that genotype-dependent and photoperiod-dependent differences were likely to exist, though the latter differences were very small (**Figure A3** in Appendix). Since the PHYAOX plants do not induce growth cessation in SD (Olsen et al., 1997), we hypothesized that: candidate metabolite levels of which the levels can be regarded as "metabolite signature" for SD treatment should (1) differ in PHYAOX samples compared to WT, and (2) show no photoperiod-dependent changes in PHYAOX samples as PHYAOX is insensitive to short-day-induced dormancy. To identify such metabolites, an ANOVA (factors: daylength × genotype; **Data Sheet 2** in Supplementary Material) was conducted and the results presented in a Venn diagram. This revealed 97 metabolite peaks that are genotype-specific and, thus, candidates for this "metabolite signature" (**Figure A4** in Appendix). Of these, 39 peaks were annotated as known metabolites. In addition to the 97 genotype-specific peaks, the abundance of 20 peaks differed between PHYAOX and wild-type poplar, yet differences due to the photoperiod were also apparent; among these peaks were serine, aspartate, and glutamate. These metabolites showed an increase in their metabolite levels in mature WT leaves under SD (**Figure A2** in Appendix).

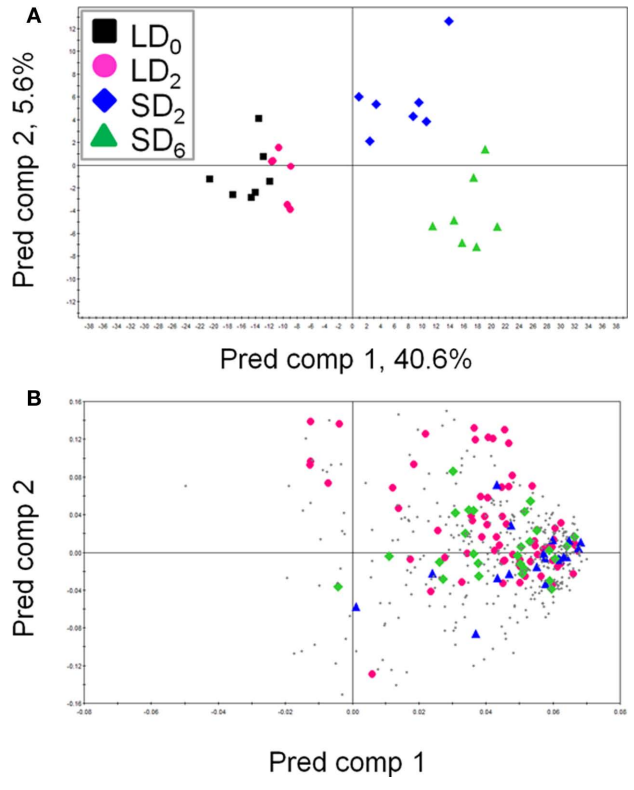


FIGURE 2 | The OPLS-DA score scatter plot (A) and loading plot (B) of leaf 20 samples under different photoperiods. Each point represents an independent plant in the score scatter plots and an individual peak in the loading plots. The OPLS-DA model for leaf 20 samples shows two significant components, with R^2X, R^2Y, and Q^2Y values of 0.58, 0.55, and 0.32, respectively. **(A)** Black square, LD_0; pink circle, LD_2; blue diamond, SD_2; green triangle, SD_6. **(B)** Pink diamond, identified metabolite peak; blue triangle, annotated metabolite peak; green square, MST; gray dot, unclassified peak. Abbreviations: Pred comp, predictive component; MST, mass spectral tag.

COMPARISON OF GENOTYPE-DEPENDENT METABOLITES WITH AGE-DEPENDENT METABOLITES

The result of the two-way ANOVA on the leaf 20 dataset demonstrated that WT leaves showed metabolite changes across the four different time periods (day 0, day 2, day 4, and day 6). These metabolites were thought to be involved in plant growth from day 0 to day 6 and are called age-dependent changes. Because PHYAOX have a dwarfed phenotype and is able to grow under short-day conditions, age-dependent changes might contribute to the observed genotype-specific differences. To reduce the number of the candidates for the "metabolite signature" of short-day-induced growth cessation, the age-dependent metabolites found in the leaf 20 dataset were filtered out from the genotype-specific

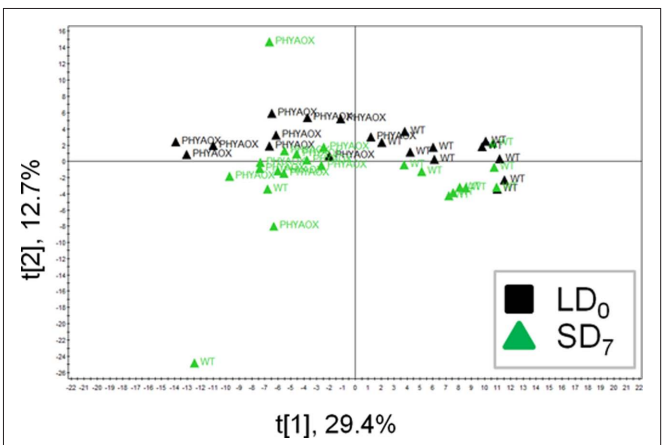


FIGURE 3 | The PCA score scatter plot of PHYAOX and WT samples under LD and SD conditions. Black square, LD_0; green triangle, SD_7. Abbreviations: PHYAOX, *PHYA* overexpressor; WT, wild-type.

metabolites in the PHYAOX dataset (**Figure A5** in Appendix). We focused on known metabolites to compare the two different datasets (**Data Sheet 1 and 2** in Supplementary Material). As visualized in the Venn diagram (**Figure A5** in Appendix), 14 of the 38 known genotype-specific metabolites were retained. Of these, the levels of 3-cyano-alanine, caffeate, 2-oxo-glutarate, spermidine, putrescine, and 4-amino-butyrate were increased in PHYAOX samples, whereas there was a significant decrease in the levels of threonate and tryptophan (**Table 1**; **Figure A5** in Appendix).

PATHWAY PROJECTION OF CHANGES OF THE CANDIDATE METABOLITES FOR METABOLITE SIGNATURE DURING GROWTH CESSATION

Among the candidate metabolites, several metabolites belong to the 4-aminobutyric acid (GABA) shunt and polyamine pathway. To visualize the changed peak levels on metabolic pathway, we projected the corresponding metabolites onto a metabolic map (**Figure 4**). In the GABA shunt pathway, the levels of 2-oxo-glutarate and 4-amino-butyrate in PHYAOX samples were higher than those in WT under LD and SD. However, there were no significant changes in the level of succinate (**Figure 4A**). Concerning the polyamine pathway, the levels of putrescine and spermidine showed a significant increase in PHYAOX, particularly under LD (**Figure 4A**; **Table 1**).

For another candidate metabolite, tryptophan, the level in PHYAOX was lower than that in WT under LD and SD (**Figure 4B**). Tryptophan is a known precursor of indole-3-acetic acid (IAA) which is a plant hormone in higher plants (Zhao, 2010).

Table 1 | The candidate metabolites for metabolite signature found in PHYAOX dataset.

Metabolite	Log_2 ratio (PHYAOX/WT in LD)	Log_2 ratio (PHYAOX/WT in SD)
Phosphoric acid, monomethyl ester*	–2.33	–0.44
Norvaline*	0.86	0.62
Alanine, 3-cyano-	1.13	0.84
Threonic acid	–1.03	–0.38
Caffeic acid, trans-	0.52	0.60
Tryptophan	–2.18	–1.99
Inositol-2-phosphate, myo-*	–0.54	–0.44
Quercetin*	–1.19	–0.87
beta-Alanine*	–0.69	–0.47
2-Oxo-glutaric acid	0.40	0.63
Spermidine	0.87	0.07
Putrescine	0.43	0.35
Dibutyl-sebacic acid*	0.56	0.64
4-Aminobutyric acid	0.44	0.50

These metabolites were determined by comparing ANOVA results of leaf 20 and PHYAOX datasets. We assayed the leaf 20 dataset using ANOVA to find metabolites that showed significant changes according to age-dependent differences (FDR < 0.05). We also assayed PHAOX dataset to find genotype-dependent metabolites (FDR < 0.05). We then compared known metabolite names in age-dependent and genotype-dependent metabolites.

**The metabolites were detected in PHYAOX and WT samples, but not in leaf 20 samples.*

DISCUSSION

METABOLITE PROFILING ANALYSIS CAN RECOGNIZE DIFFERENCE OF METABOLITE COMPOSITION OF ASPEN LEAVES IN THEIR POSITIONS, DIFFERENT TIME FRAMES AND PHOTOPERIODS

Our aim was to examine foliar metabolite alterations to changes in photoperiod at different developmental stages of the leaf. Metabolite composition of aspen leaves can be captured using GC–MS analysis with respect to the extent of leaf expansion (Jeong et al., 2004). In this study, we analyzed aspen leaf extracts from position 1 (leaf 1) to position 20 (leaf 20) using GC–TOF-MS by applying the definition of each leaf position as shown in **Figure 1A**. The OPLS analysis, which is one of the supervised methods, clearly showed that metabolite profiles of aspen leaves were well correlated with their leaf positions (**Figures 1A,C**).

A scatter plot of the score values provides an overview of the samples (observations) and their inter-relationships, e.g., groupings, trends, and deviating samples. The PCA score scatter plots of leaf 2 (young leaves), leaf 10 (middle), and leaf 20 (mature) showed that the profiles of mature leaves reflected better metabolite changes related to different time frames and change of photoperiods as compared to those of young and middle leaves, while young and middle leaves did not (**Figure A1** in Appendix). To interpret the patterns found in the score plots of the leaf 20 dataset, we examined the corresponding loading plots (**Figure 2**). This method revealed how each variable contributed to the separation among samples in the model plane, indicating the relative importance of each variable. Using multivariate projection methods, we validated our data. We used preliminary data sets in our models that enabled us to predict external sample data verifying the usefulness of our calculated models (**Figure 2**). This strategy is essential when multivariate projection methods are used to avoid problems associated with overfitting of the data (Eriksson et al., 2004). Our study demonstrated that after a few days in SD the mature source leaves form hybrid aspen trees showed a clear metabolic response.

DISSECTION OF THE CANDIDATE METABOLITES FOR METABOLITE SIGNATURE DURING GROWTH CESSATION FROM MULTIVARIATE DATASETS

Changes in the metabolome of plants grown under different photoperiods are complex phenomena. Various parts of the metabolome may be affected by other variables, depending on the experimental set-up, diurnal effects, differences in the time at which the lights turn on, differences in the PAR availability, or differences in the photoperiods *per se* (Thomas and Vince Prue, 1997). These aspects should be taken into account when interpreting our results. For example, large effects on the primary carbohydrate metabolism may indicate that some of the changes observed are from differences in photosynthesis, especially over a relatively short time-period. Our study used similar amounts of PAR in both LD and SD treatments, and extended the days in SD with low light conditions. Therefore, we reduced the possibility that the differences in metabolite changes are due to the amount of light that plants received. Furthermore, we emphasize that some of the alterations in metabolite profiles we observed after transferring hybrid aspen from LD to SDs are due to the aging of the plant, i.e., from first sampling time (LD_0) to the last sampling time (SD_6). Since PHYAOX plants grow under SD (Olsen et al., 1997; Ruonala et al., 2008), we need to consider

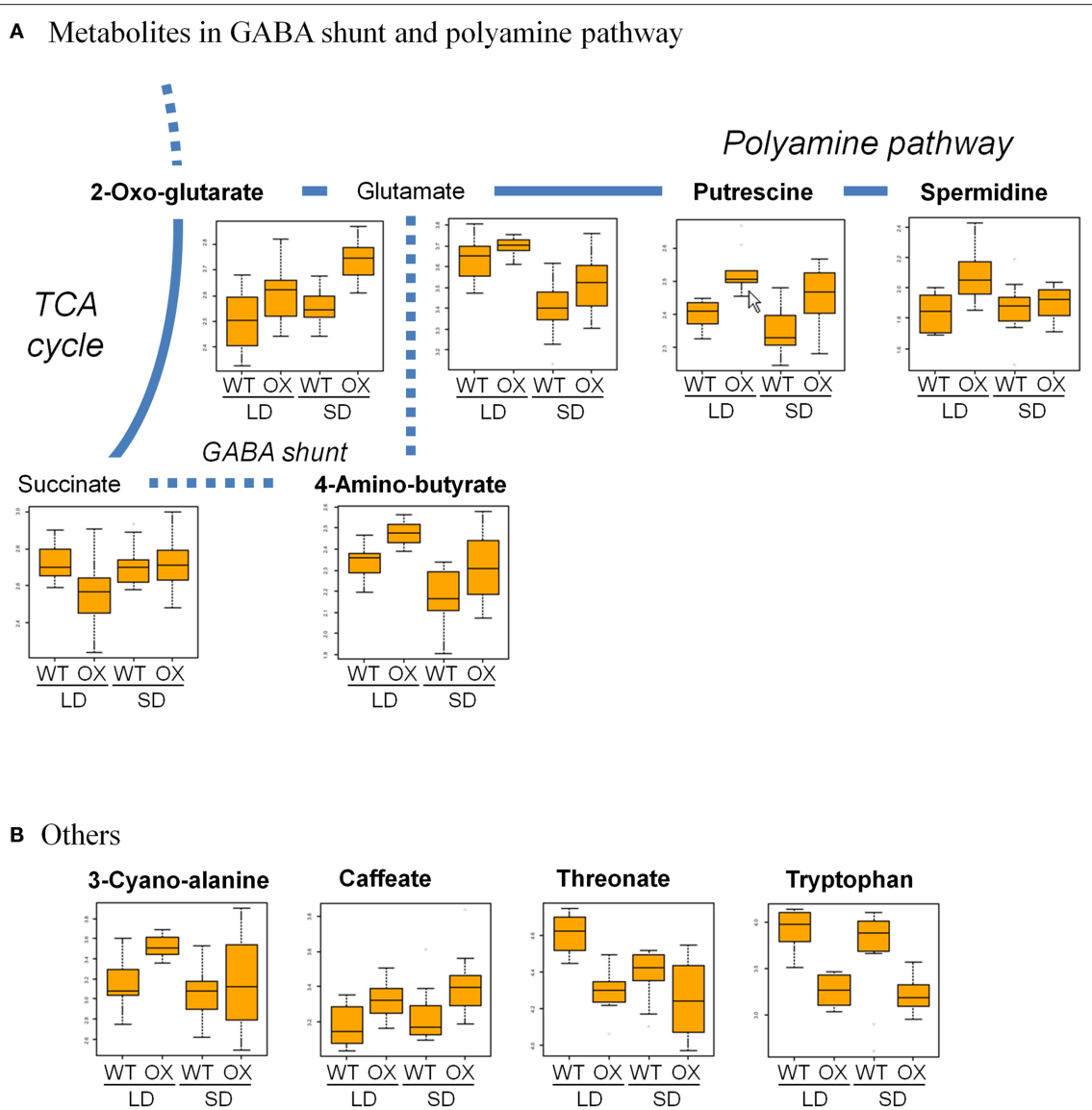


FIGURE 4 | Overlay of changes in the candidates for metabolite signature and related metabolites observed in PHYAOX and WT samples onto the metabolic map. (A) Changes in the levels of four candidate metabolites (in bold) and related metabolites in GABA shunt and polyamine pathway and **(B)** those of other cadidate metabolites for metabolite signature. In each box plot, *x*-axis represents metabolites detected in WT and PHYAOX (OX) in LD and SD, while *y*-axis shows the normalized response of metabolite levels after $\log_{10}$-transfromation.

the age-dependent differences. The results of the ANOVA of leaf 20 samples demonstrated that many metabolite peaks are likely to be involved in age-dependent differences even though the time frame is only 1 week (**Data Sheet 1** in Supplementary Material). By comparing the age-dependent metabolites found in the leaf 20 dataset with those in the PHYAOX dataset, we selected candidate metabolites of which the levels are a metabolite signature during growth cessation. Similar approaches could be applied to other datasets using this multivariate approach. For example, using a multivariate analyses may be useful when trying to detect changes that occur during growth cessation that involve transcript and metabolite levels in hybrid aspen (Bylesjo et al., 2007) and *Populus* (Ruttink et al., 2007) at a global scale.

POSSIBLE LINK BETWEEN GROWTH CESSATION AND NITROGEN METABOLISM IN HYBRID ASPEN

Levels of the candidate metabolites were altered in PHYAOX after 1 week with a short photoperiod. The expression level of *FT2* in WT source leaves was down-regulated after 1 week of SD (Ruonala et al., 2008). Constitutive expression of *Populus FT1* and oat *PHYA* in mature aspen leaves suppresses short-day-induced growth cessation. This is because the plants fail to down-regulation of *FT1* and *CONSTANS2* (*CO2*) within a week under SDs (Bohlenius et al., 2006). Furthermore, FT protein and the protein encoded by a rice ortholog of *FT* can mobilize to the apex via phloem as a long-distance signal for flowering in *Arabidopsis* (Corbesier et al., 2007) and rice (Tamaki et al., 2007). This suggests that a phloem-unloading

mechanism has to be well developed in leaves before any signal can be transmitted. Thus, our observations that sink leaves are not affected by daylength change in aspen may have a physiological explanation.

For the candidate metabolites belonging to the GABA shunt pathway, the level of 2-oxo-glutarate was increased in PHYAOX samples after transferring to the SD treatment while metabolites in WT remained unchanged (**Figure 4A**). In contrast, the GABA level decreased in the WT during short-day-induced growth cessation. We found no significant changes in PHYAOX samples (**Figure 4A**). 2-Oxo-glutarate is not only one of the intermediates in the TCA cycle but serves as a carbon assimilation precursor that is derived from nitrogen metabolism (Foyer et al., 2011; Millar et al., 2011). The significant increase in the level of 2-oxo-glutarate in PHYAOX samples may provide a source of carbon skeletons for macromolecules from source leaves that maintains their growth in short photoperiod (Ruonala et al., 2008).

GABA is an important component of signaling systems in both vertebrates and invertebrates, but its role in plants is largely unknown (Bouche and Fromm, 2004; Fait et al., 2008). GABA has been suggested to have various roles. For instance, in the regulation of nitrogen metabolism and transport, in oxidative stress, and in controlling pollen tube growth. Our results suggest that the level of GABA in WT decreased during short-day-induced growth cessation, implying that it may not only act as a protector from stress, but may also play a role during growth cessation. GABA can be detected in xylem sap and phloem exudates of walnut trees (Frak et al., 2002) and in a *Brassica* species (Beuve et al., 2004). GABA might be a signal molecule remobilized from source leaves to apex after a week of exposure to the short-day treatment. Like GABA, levels of its precursor glutamate were lower in WT under SD (**Figure 4A**), suggesting that short days induce changes in the flow through GABA shunt.

We also found the tryptophan level is lower in PHYAOX samples than that in WT under LD and SD (**Figure 4B**). Recently, Baba et al. (2011) described that IAA has a critical role for growth cessation in aspen (Baba et al., 2011). In the leaves of *Ricinus communis*, the level of tryptophan and IAA showed an inverse distribution in accordance with their position (Allen and Baker, 1980). In addition, the free tryptophan content was most abundant in mature leaves, while sink leaves showed less accumulation of tryptophan content in the *Ricinus* leaves. This suggests that mature leaves of PHAOX plants are likely to have a similar ability as sink leaves to produce IAA, although further experiments are required, e.g., measurement of IAA in sink and source leaves. Indeed, the level of IAA is more abundant in tobacco sink leaves than that in source leaves (Sitbon et al., 1990).

In summary, we could choose the candidate metabolites for metabolite signature in mature leaves during growth cessation. To investigate metabolic dynamics for these metabolites from mature leaves to apex, metabolite profiling of phloem and xylem sap and labeling experiments to trace the candidate metabolites should be conducted for future analysis.

ACKNOWLEDGMENTS

We thank K. Saito for fruitful discussions. This work was supported by grants from Swedish Research Council, FORMAS, SLU, the KEMPE foundation, and former Wallenberg Consortium North (Thomas Moritz).

SUPPLEMENTARY MATERIAL

The Data sheets 1 and 2 for this article can be found online at http://www.frontiersin.org/plant_physiology/10.3389/fpls.2011.00029/abstract

DATA SHEET 1 | ANOVA of leaf 20 samples.
DATA SHEET 2 | ANOVA of PHYAOX samples.

REFERENCES

Allen, J. R. F., and Baker, D. A. (1980). Free-tryptophan and indole-3-acetic-acid levels in the leaves and vascular pathways of *Ricinus-Communis* L. *Planta* 148, 69–74.

Baba, K., Karlberg, A., Schmidt, J., Schrader, J., Hvidsten, T. R., Bako, L., and Bhalerao, R. P. (2011). Activity-dormancy transition in the cambial meristem involves stage-specific modulation of auxin response in hybrid aspen. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3418–3423.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.* 57, 289–300.

Beuve, N., Rispail, N., Laine, P., Cliquet, J. B., Ourry, A., and Le Deunff, E. (2004). Putative role of gamma-aminobutyric acid (GABA) as a long distance signal in up-regulation of nitrate uptake in *Brassica napus* L. *Plant Cell Environ.* 27, 1035–1046.

Bohlenius, H., Huang, T., Charbonnel-Campaa, L., Brunner, A. M., Jansson, S., Strauss, S. H., and Nilsson, O. (2006). CO/FT regulatory module controls timing of flowering and seasonal growth cessation in trees. *Science* 312, 1040–1043.

Bouche, N., and Fromm, H. (2004). GABA in plants: just a metabolite? *Trends Plant Sci.* 9, 110–115.

Bylesjo, M., Eriksson, D., Kusano, M., Moritz, T., and Trygg, J. (2007). Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. *Plant J.* 52, 1181–1191.

Corbesier, L., Vincent, C., Jang, S., Fornara, F., Fan, Q., Searle, I., Giakountis, A., Farrona, S., Gissot, L., Turnbull, C., and Coupland, G. (2007). FT protein movement contributes to long-distance signaling in floral induction of *Arabidopsis*. *Science* 316, 1030–1033.

Eriksson, L., Antti, H., Gottfries, J., Holmes, E., Johansson, E., Lindgren, F., Long, I., Lundstedt, T., Trygg, J., and Wold, S. (2004). Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Anal. Bioanal. Chem.* 380, 419–429.

Eriksson, M. E., and Millar, A. J. (2003). The circadian clock. A plant's best friend in a spinning world. *Plant Physiol.* 132, 732–738.

Eriksson, M. E., and Moritz, T. (2002). Daylength and spatial expression of a gibberellin 20-oxidase isolated from hybrid aspen (*Populus tremula* L. x *P. tremuloides* Michx.). *Planta* 214, 920–930.

Fait, A., Fromm, H., Walter, D., Galili, G., and Fernie, A. R. (2008). Highway or byway: the metabolic role of the GABA shunt in plants. *Trends Plant Sci.* 13, 14–19.

Fernie, A. R., Trethewey, R. N., Krotzky, A. J., and Willmitzer, L. (2004). Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.* 5, 763–769.

Fiehn, O. (2002). Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* 48, 155–171.

Foyer, C. H., Noctor, G., and Hodges, M. (2011). Respiration and nitrogen assimilation: targeting mitochondria-associated metabolism as a means to enhance nitrogen use efficiency. *J. Exp. Bot.* 62, 1467–1482.

Frak, E., Millard, P., Le Roux, X., Guillaumie, S., and Wendler, R. (2002). Coupling sap flow velocity and amino acid concentrations as an alternative method to (15)N labeling for quantifying nitrogen remobilization by walnut trees. *Plant Physiol.* 130, 1043–1053.

Gullberg, J., Jonsson, P., Nordstrom, A., Sjostrom, M., and Moritz, T. (2004). Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of *Arabidopsis thaliana* samples in metabolomic studies with gas chromatography/mass spectrometry. *Anal. Biochem.* 331, 283–295.

Hall, R. D. (2006). Plant metabolomics: from holistic hope, to hype, to hot topic. *New Phytol.* 169, 453–468.

Hoffman, D. E., Jonsson, P., Bylesjo, M., Trygg, J., Antti, H., Eriksson, M. E., and Moritz, T. (2010). Changes in diurnal patterns within the *Populus* transcriptome and metabolome in response to photoperiod variation. *Plant Cell Environ.* 33, 1298–1313.

Jeong, M. L., Jiang, H., Chen, H. S., Tsai, C. J., and Harding, S. A. (2004). Metabolic profiling of the sink-to-source transition

in developing leaves of quaking aspen. *Plant Physiol.* 136, 3364–3375.

Jonsson, P., Johansson, A. I., Gullberg, J., Trygg, J., A, J., Grung, B., Marklund, S., Sjostrom, M., Antti, H., and Moritz, T. (2005). High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. *Anal. Chem.* 77, 5635–5642.

Kozarewa, I., Ibanez, C., Johansson, M., Ogren, E., Mozley, D., Nylander, E., Chono, M., Moritz, T., and Eriksson, M. E. (2010). Alteration of PHYA expression change circadian rhythms and timing of bud set in *Populus*. *Plant Mol. Biol.* 73, 143–156.

Kusano, M., Fukushima, A., Arita, M., Jonsson, P., Moritz, T., Kobayashi, M., Hayashi, N., Tohge, T., and Saito, K. (2007). Unbiased characterization of genotype-dependent metabolic regulations by metabolomic approach in *Arabidopsis thaliana*. *BMC Syst. Biol.* 1, 53. doi: 10.1186/1752-0509-1-53

Mcclung, C. R. (2008). Comes a time. *Curr. Opin. Plant Biol.* 11, 514–520.

Millar, A. H., Whelan, J., Soole, K. L., and Day, D. A. (2011). Organization and regulation of mitochondrial respiration in plants. *Annu. Rev. Plant Biol.* 62, 79–104.

Olsen, J. E., Junttila, O., and Moritz, T. (1995). A localised decrease of GA(1) in shoot tips of *Salix pentandra* seedlings precedes cessation of shoot elongation under short photoperiod. *Physiol. Plant* 95, 627–632.

Olsen, J. E., Junttila, O., Nilsen, J., Eriksson, M. E., Martinussen, I., Olsson, O., Sandberg, G., and Moritz, T. (1997). Ectopic expression of oat phytochrome A in hybrid aspen changes critical daylength for growth and prevents cold acclimatization. *Plant J.* 12, 1339–1350.

Pavlidis, P. (2003). Using ANOVA for gene selection from microarray studies of the nervous system. *Methods* 31, 282–289.

Ruonala, R., Rinne, P. L., Kangasjarvi, J., and Van Der Schoot, C. (2008). CENL1 expression in the rib meristem affects stem elongation and the transition to dormancy in *Populus*. *Plant Cell* 20, 59–74.

Ruttink, T., Arend, M., Morreel, K., Storme, V., Rombauts, S., Fromm, J., Bhalerao, R. P., Boerjan, W., and Rohde, A. (2007). A molecular timetable for apical bud formation and dormancy induction in poplar. *Plant Cell* 19, 2370–2390.

Saito, K., and Matsuda, F. (2010). Metabolomics for functional genomics, systems biology, and biotechnology. *Annu. Rev. Plant Biol.* 61, 463–489.

Salome, P. A., and Mcclung, C. R. (2005). What makes the *Arabidopsis* clock tick on time? A review on entrainment. *Plant Cell Environ.* 28, 21–38.

Schauer, N., Steinhauser, D., Strelkov, S., Schomburg, D., Allison, G., Moritz, T., Lundgren, K., Roessner-Tunali, U., Forbes, M. G., Willmitzer, L., Fernie, A. R., and Kopka, J. (2005). GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett.* 579, 1332–1337.

Schultz, T. F., and Kay, S. A. (2003). Circadian clocks in daily and seasonal control of development. *Science* 301, 326–328.

Sitbon, F., Sundberg, B., Olsson, O., and Sandberg, G. (1990). Free and conjugated indoleacetic acid (IAA) contents in transgenic tobacco plants expressing the iaam and iaah iaa biosynthesis genes from *Agrobacterium tumefaciens*. *Plant Physiol.* 95, 480–485.

Tamaki, S., Matsuo, S., Wong, H. L., Yokoi, S., and Shimamoto, K. (2007). Hd3a protein is a mobile flowering signal in rice. *Science* 316, 1033–1036.

Thomas, B., and Vince Prue, D. (1997). *Photoperiodism in Plants*, 2nd Edn. San Diego, CA: Academic Press, xv 428.

Thomas, B., and Vince-Prue, D. (1997). *Photoperiodism in Plants*. San Diego: Academic Press.

Trygg, J., and Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *J. Chemom.* 16, 119–128.

Wold, S. (1978). Cross-validatory estimation of number of components in factor and principal components models. *Technometrics* 20, 397–405.

Zhao, Y. (2010). Auxin biosynthesis and its role in plant development. *Annu. Rev. Plant Biol.* 61, 49–64.

Conflict of Interest Statement: 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.

Received: 01 May 2011; accepted: 29 June 2011; published online: 12 July 2011.

*Citation: Kusano M, Jonsson P, Fukushima A, Gullberg J, Sjöström M, Trygg J and Moritz T (2011) Metabolite signature during short-day induced growth cessation in Populus. Front. Plant Sci. **2**:29. doi: 10.3389/fpls.2011.00029*

This article was submitted to Frontiers in Plant Physiology, a specialty of Frontiers in Plant Science.

APPENDIX

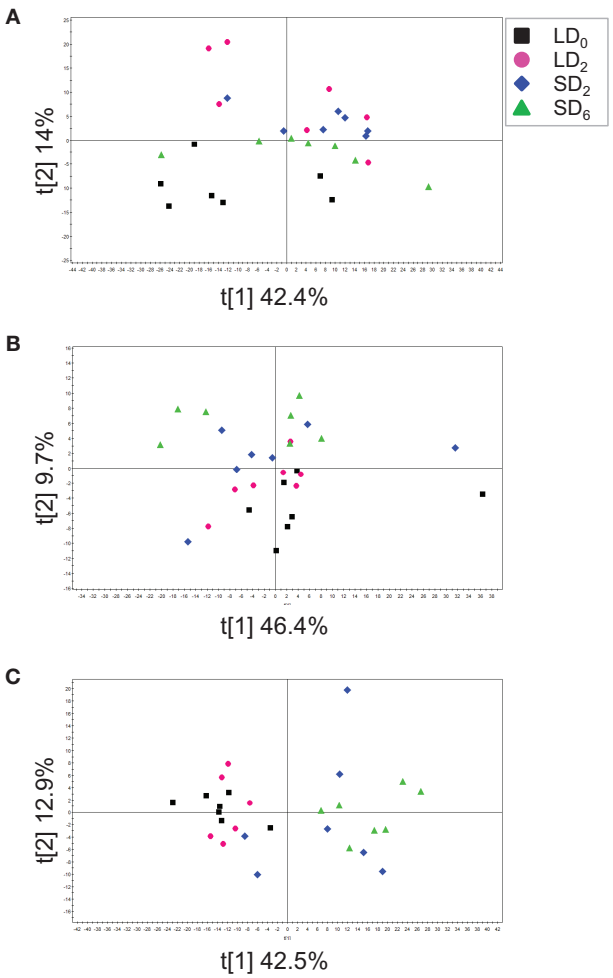


FIGURE A1 |The PCA score scatter plot of leaf 2 (A), leaf 10 (B), and leaf 20 (C) samples under different photoperiods. Black square, LD_0; pink circle, LD_2; blue diamond, SD_2; green triangle, SD_6. t[1] and t[2] present the first and second components and each percentage shows the fraction of the sum of squares for the selected component. We excluded a leaf 20 sample grown under the LD_2 condition because the sample was identified as an outlier in the score and Hotelling's T^2 plot.

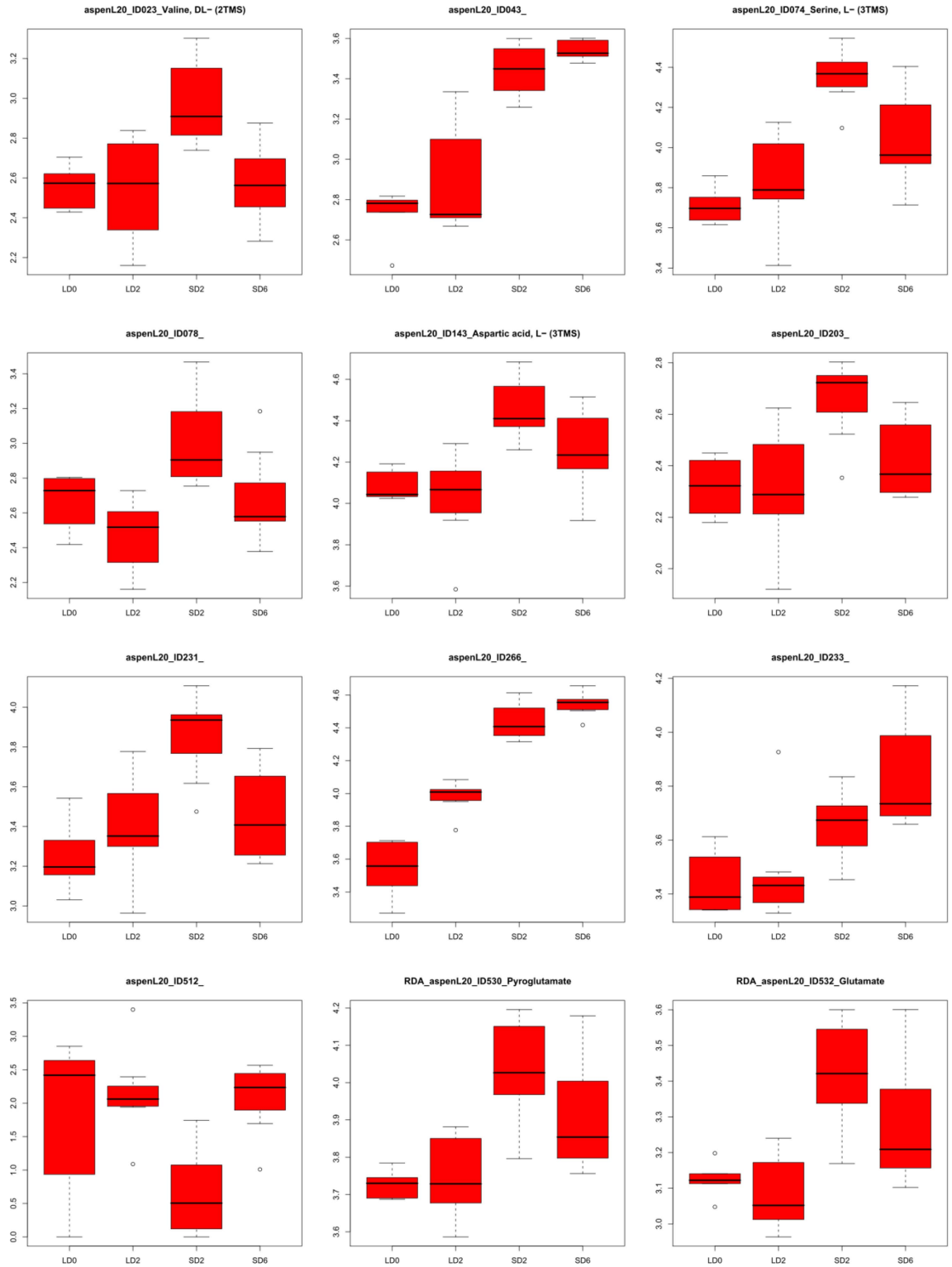


FIGURE A2 | Box plots of the 12 metabolites that showed significant changes from a two-way ANOVA. These metabolites were found to be significantly different in relation to our photoperiod treatments. The *x*-axis indicates sampling periods (LD_0, LD_2, SD_2, and SD_6).The *y*-axis shows an arbitrary unit of each metabolite level after $\log_{10}$-transformation. FDR < 0.05.

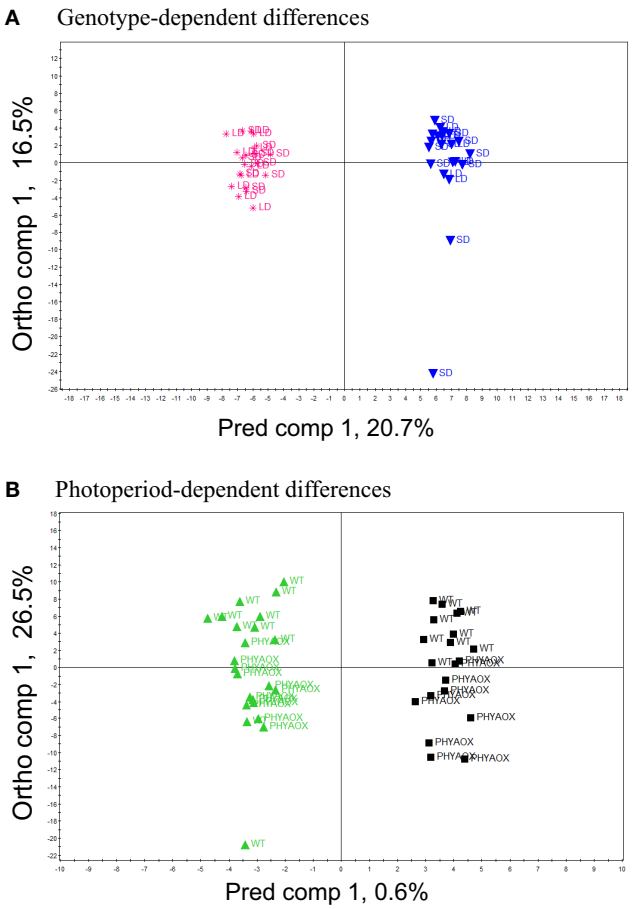


FIGURE A3 | The OPLS-DA score scatter plot of PHYAOX and WT samples. **(A)** Genotype-dependent differences. **(B)** Photoperiod-dependent differences. Pink star, PYHAOX; blue inverted triangle, SD_7; black square, LD_0; green triangle, SD_7.

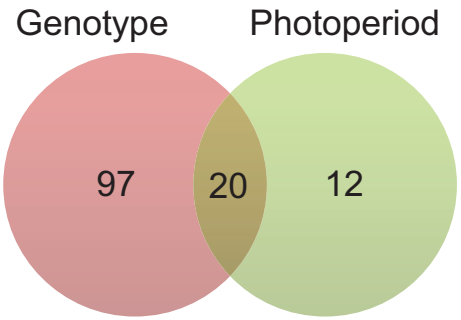


FIGURE A4 | The Venn diagram of the result of ANOVA using metabolite profile data of PHYAOX and WT samples. The number in the left-hand circle indicates how many metabolites showed significant changes according to genotype-dependent differences. The number in the right-hand circle represents the number of significant metabolites with respect to photoperiod-dependent differences. The number in the middle represents the common metabolites between PHYAOX and WT samples. FDR < 0.05.

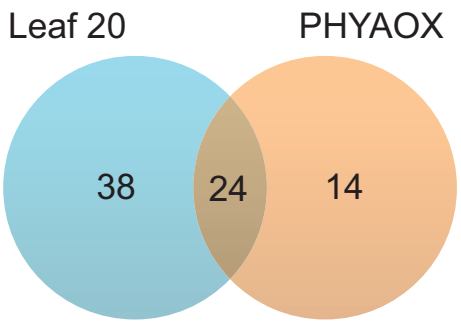


FIGURE A5 | The Venn diagram of known metabolites found as age-dependent (left) and genotype-dependent (right) metabolites in each dataset as determined by ANOVA. The number in the left-hand circle represents how many metabolites showed significant changes according to age-dependent differences. The number in the right-hand circle represents the number of significant metabolites with respect to genotype-dependent differences. The number in the middle represents the common metabolites between age-dependent and genotype-dependent metabolites. FDR < 0.05.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 12 September 2011
doi: 10.3389/fpls.2011.00043

SLocX: predicting subcellular localization of *Arabidopsis* proteins leveraging gene expression data

Malgorzata Ryngajllo[1], Liam Childs[1], Marc Lohse[1], Federico M. Giorgi[1], Anja Lude[1], Joachim Selbig[2] and Björn Usadel[1]*

[1] *Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany*
[2] *Department of Bioinformatics, Institute of Biochemistry and Biology, University of Potsdam, Potsdam, Germany*

Edited by:
Alisdair Fernie, Max Planck Institute for Plant Physiology, Germany

Reviewed by:
Joshua L. Heazlewood, Lawrence Berkeley National Laboratory, USA
Rainer Schwacke, University of Tromsø, Norway

****Correspondence:***
Björn Usadel, Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, Golm, 14476 Potsdam, Germany.
e-mail: usadel@mpmip-golm.mpg.de

Despite the growing volume of experimentally validated knowledge about the subcellular localization of plant proteins, a well performing *in silico* prediction tool is still a necessity. Existing tools, which employ information derived from protein sequence alone, offer limited accuracy and/or rely on full sequence availability. We explored whether gene expression profiling data can be harnessed to enhance prediction performance. To achieve this, we trained several support vector machines to predict the subcellular localization of *Arabidopsis thaliana* proteins using sequence derived information, expression behavior, or a combination of these data and compared their predictive performance through a cross-validation test. We show that gene expression carries information about the subcellular localization not available in sequence information, yielding dramatic benefits for plastid localization prediction, and some notable improvements for other compartments such as the mitochondrion, the Golgi, and the plasma membrane. Based on these results, we constructed a novel subcellular localization prediction engine, SLocX, combining gene expression profiling data with protein sequence-based information. We then validated the results of this engine using an independent test set of annotated proteins and a transient expression of GFP fusion proteins. Here, we present the prediction framework and a website of predicted localizations for *Arabidopsis*. The relatively good accuracy of our prediction engine, even in cases where only partial protein sequence is available (e.g., in sequences lacking the N-terminal region), offers a promising opportunity for similar application to non-sequenced or poorly annotated plant species. Although the prediction scope of our method is currently limited by the availability of expression information on the ATH1 array, we believe that the advances in measuring gene expression technology will make our method applicable for all *Arabidopsis* proteins.

Keywords: subcellular localization, support vector machine, prediction, gene expression

INTRODUCTION

In eukaryotic cells, the targeting of proteins to subcellular compartments is universally recognized to be important for proper protein function (Eisenhaber and Bork, 1998). In plants, several metabolic pathways either consist of enzymes residing in multiple compartments (e.g., the photorespiration pathway), or they occur in parallel in different compartments as is the case for the glycolysis. Therefore, detailed knowledge about protein localization is necessary to understand the plant metabolic network (Lunn, 2007). In addition, the presence of three compartments (nuclei, plastids, and mitochondria) harboring their own genetic information, makes a complex information flow necessary (for a recent overview see Pfannschmidt, 2010).

It is thus not surprising that many studies have focused on the experimental determination of protein subcellular localization in plants (Koroleva et al., 2005). Many of these have profited from the adoption of high-throughput proteomics (Schulze and Usadel, 2010; Wienkoop et al., 2010). These studies have revolutionized our understanding of the localization of proteins in organs (Baerenfaller et al., 2008) and individual subcellular compartments (van Wijk, 2004; Dunkley et al., 2006; Ito et al., 2010). In particular, the technique of organelle purification in combination with highly sensitive LC–MS/MS instruments has proven to be useful in providing a detailed experimental compendium of proteins localized in, e.g., the mitochondrion or the chloroplast (Heazlewood et al., 2004; Ferro et al., 2010). Several independent studies used relative protein concentration along density gradients (Dunkley et al., 2004, 2006) making use of statistical association methods similar to those for subcellular determination of metabolites (Gerhardt and Heldt, 1984; Krueger et al., 2011).

However, despite this avalanche of experimental data, experimentally determined subcellular information is only available for ca. 30% of all proteins for the well studied model organism *Arabidopsis* (SUBA database, Heazlewood et al., 2006; TAIR database, Rhee et al., 2003). Even in the case of the chloroplast, which is probably the most well studied organelle in terms of proteomics, only 30–60% of the estimated protein population has been found by proteomics methods (van Wijk and Baginsky, 2011). It has been suggested that this lack of information can be explained by

temporal, spatial, or experimental condition specificity of protein accumulation, or even by simple technical limitations (van Wijk and Baginsky, 2011). Furthermore, one must keep in mind that no fractionation is perfect and that some proteins might thus be wrongly tagged as belonging to a certain compartment. In part, this can be overcome by trusting high-throughput experimental evidence only if proteins have been associated with a particular compartment by multiple independent studies. Indeed, by combining different data sets an improved assignment can be reached (Trotter et al., 2010). Unfortunately, no matter how many studies are combined, it is still possible that certain wrong assignments can result from systematic problems in separation techniques. Furthermore, although some subcellular localization studies have been conducted for crop plants (Majeran et al., 2005; von Zychlinski et al., 2005; Huang et al., 2009), proteomics cannot yet keep up with the growth of genomic data for multiple plant species.

Therefore, it is still necessary to be able to accurately predict the subcellular localization of proteins. Traditionally, this was done by identifying protein sequence motifs such as signal peptides or targeting signals (see Emanuelsson et al., 2007 for an overview of these methods). Indeed, the widely used TAIR database relies on such predictions made by TargetP which only uses the N-terminal sequence information containing the signal peptide (von Heijne et al., 1989) to decide whether a protein is to be targeted to the chloroplast, the mitochondrion, the secretory pathway, or another location (Emanuelsson et al., 2000). Other widely applied prediction tools screening for N-terminal targeting signals are Predotar (Small et al., 2004) and iPSORT (Bannai et al., 2002). Since these tools have different strengths and weaknesses, a selection was combined in a meta-predictor using a naive Bayes approach (Schwacke et al., 2007). Although a wide variety of such N-terminal prediction systems has been developed throughout the years, some methods are limited in accuracy and/or in the breadth of coverage of subcellular compartments. More importantly, these methods fail to make a valid prediction when a protein is targeted to its final compartment through non-classical mechanisms of protein sorting (Herman and Schmidt, 2004; Nickel and Seedorf, 2008; Wienkoop et al., 2010) or contains a non-conventional targeting sequence (Brix et al., 1999; Diekert et al., 1999). Moreover, these predictors cannot operate in cases where only a partial protein sequence is known as might often be the case in projects relying on EST data to study a non-model plant organism.

To overcome the limitations of N-terminal-based predictions, tools employing a diverse range of other protein features have been developed. Due to the complexity of extracting protein localization, machine learning techniques such as neural networks, hidden Markov models or support vector machines (SVM) have been applied. As SVMs have yielded very good results, SVM based prediction tools based on diverse and robust protein features have gained in popularity (Hua and Sun, 2001; Gardy and Brinkman, 2006). Initially, the main features that were considered were simply derived from the amino acid composition of the whole protein (Nishikawa et al., 1983). Since then, many additional features have been employed to enhance the predictive power which has resulted in the development of systems which apply hybrid approaches using very diverse protein features in combination (Garg et al., 2005; Cui et al., 2011). Among the popular methods, some are homology-based (Kaundal et al., 2010), and others identify subcellular localization of proteins from phylogenetic profiles (Marcotte et al., 2000; Blum et al., 2009). Obviously though, the latter methods do not work on species-specific proteins.

Based on the expected avalanche of transcript data from next generation sequencing for non-model plants (Severin et al., 2010; Zhang et al., 2010), the need to develop robust methods for the prediction of protein subcellular localization is becoming more pressing. As a case study, we developed a novel tool to predict the subcellular localization of *Arabidopsis* proteins integrating protein amino acid composition with expression profiling data.

MATERIALS AND METHODS

GENERATION OF A WORKING AND AN INDEPENDENT TEST DATA SET

In order to construct a working data set, the GO Slim annotation was downloaded from the TAIR database[1] (ATH_GO_GOSLIM_02_01_11). Experimentally confirmed subcellular localizations were extracted by selecting only those records containing the IDA (i.e., "inferred from direct assay") evidence code. Afterward, all instances containing annotations for mitochondrion and plastid genome encoded proteins were removed from the data set. In cases where multiple splicing isoforms existed the "representative protein model" was downloaded from TAIR. In contrast to most previous approaches, proteins annotated to be localized in multiple localizations were retained. This yielded a total number of 6,188 unique protein identifiers having at least one experimentally confirmed subcellular localization. We further filtered this set based on available expression information yielding 5,429 unique proteins.

An independent test data set was created as follows: from all representative *Arabidopsis* proteins, those used to create the working data set were subtracted. Furthermore, all mitochondrion and plastid genome encoded proteins were removed giving a total number of 20,016 unique protein identifiers. From these, only proteins represented on the ATH1 chip where retained, yielding 13,104 proteins. For these proteins, the SUBA database was queried and 1,398 proteins with experimentally determined subcellular localization could be retrieved.

PREDICTIONS FROM STATE OF THE ART PREDICTORS

Sequences of 1,398 proteins from the independent test data set were downloaded from TAIR database (TAIR10_pep_20110103_representative_gene_model) and used to query: Predotar[2], MultiLoc2[3] [MultiLoc2-HighRes (Plant) method], and AtSubP[4] ("best hybrid" method). For the same proteins, predictions made by TargetP were downloaded from the TAIR database[5].

FEATURE SET GENERATION

For the proteins in the working and in the independent test data set, sequence data was downloaded from the TAIR database (TAIR10_pep_20110103_representative_gene_model). For

[1] http://www.arabidopsis.org/
[2] http://urgi.versailles.inra.fr/predotar/predotar.html
[3] http://abi.inf.uni-tuebingen.de/Services/MultiLoc2
[4] http://bioinfo3.noble.org/AtSubP
[5] http://www.arabidopsis.org/tools/bulk/protein/index.jsp

each protein the amino acid composition was calculated as the occurrence of each of the 20 amino acids in the sequence normalized to the protein length, as previously described in Garg et al. (2005). Additionally, for every protein in the working data set, its dipeptide and higher-order dipeptide composition was calculated (as in Garg et al., 2005). The dipeptide composition was calculated as the occurrence of two adjacent amino acids and pairs of amino acids separated by one, two, or three intervening residues normalized on the number of such dipeptides in the protein, yielding a total of 1,600 features.

The expression data set for *Arabidopsis* was the same as the one used in Giorgi et al. (2010). In brief, 3,707 *Arabidopsis thaliana* Affymetrix ATH1 (22,810 probe sets) microarray samples were obtained from the Gene Expression Omnibus database[6] (Edgar et al., 2002). The microarrays were normalized using the RMA (Robust Multi-Array Average) technique. The original data was further processed by removing *Arabidopsis* Gene Identifiers which matched more than one probeset or where one probeset matched multiple genes. Due to this reduction and absence of probesets for some genes on the ATH1 array, this data set provided expression information only for 5,429 and 1,398 experimentally annotated proteins in the working and in the independent test data set respectively. Subsequently, the whole microarray data matrix was linearly scaled between values of 0 and 1 (Eq. A1 in Appendix) as previously reported to be beneficial for SVM (Hsu et al., 2008).

The rice expression data set consisted of all non-redundant Affymetrix Rice Genome microarrays deposited in ArrayExpress (Parkinson et al., 2009) and GEO (Barrett et al., 2011). After quality filtration (as in Mutwil et al., 2011) and normalization using RMA, 487 arrays were retained.

FEATURE SELECTION AND PERFORMANCE MEASUREMENT

Features were selected in a stepwise manner using F-score and Spearman's correlation. The F-score (Eq. 1) is calculated as the ratio of the inter- and intra-group variation. Traits with a higher F-score have more separation between the positive and negative cases.

$$F(i) \equiv \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2} \quad (1)$$

where, $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$, $\bar{x}_i$, n_+, and n_- are the average of the positive data set, average of the negative data set, average of the whole data set, the total number of members of the positive dataset, and the total number of members of the negative data set for feature i, respectively. In each step, the feature with the next highest F-score was selected for addition to the set of selected features. The F-scores of the remaining features were then adjusted using the maximum Spearman's correlation coefficient of all features in the selected set (Eq. 2).

$$\text{adjFscr}_i = \text{Fscr}_i - \text{Fscr}_i \times \text{abs}(\max(\text{correlation}(\text{ftr}_i, \text{selected\_ftrs}))) \quad (2)$$

where adjFscr_i, Fscr_i, and selected_ftrs are for feature i (ftr_i): the adjusted F-score, the F-score and the features selected in previous steps, respectively.

To assess the performance of the prediction engine and to compare it with existing state of the art predictors, three common performance measures were applied: the Matthew's correlation coefficient, MCC (as in Matthews, 1975; Eq. 2 in Appendix), the sensitivity, SE (Eq. 3 in Appendix) and the precision (Eq. 4 in Appendix).

PREDICTION ENGINE CONSTRUCTION AND EVALUATION

The prediction engine constructed in this study is based on binary SVM classifiers. Each protein in the training data set of 5,429 proteins is characterized by a vector $\vec{x}_i$ ($i = 1,\ldots,5429$) that represents the chosen combination of features, along with the positive label "compartment" or the negative label "not compartment." The training of a classifier was conducted using a one-versus-rest (1-v-r SVM) strategy, where the nth SVM was trained with all the proteins in the nth class with a positive label and all other proteins with a negative label. The application of binary classifiers enabled training with proteins found in more than one compartment. The data was modeled by C-Support Vector Classification (as implemented in the libsvm library for python; Chang and Lin, 2011). The prediction engine construction and evaluation was performed on the entire working data set in two independent runs and using the same training procedure (**Figure 1**).

The training procedure first involved feature selection, when applicable, and then training of the classifiers on the given data based on the chosen features. The underlying training algorithm uses a cost parameter (C) that penalizes errors. The kernel used was the radial basis function (RBF), which requires a gamma parameter (γ) that determines the kernel bandwidth. To estimate the two parameters, we performed a grid search using fivefold cross-validation (CV) at each point in the grid to assess the performance of each parameter pair. The best performing parameter pair was then used to train an optimized classifier.

To assess the performance of the prediction engine, we used fivefold CV applying the training procedure described above to the training set of each fold and testing the resulting optimized classifier with the test set. The resulting performance measure distribution across five folds of CV is then used to estimate the performance of a prediction engine constructed using the applied training procedure (**Figure 1**). In both parameter estimation and performance evaluation, the proportion of positive and negative examples in the training and testing data sets was maintained.

TYPES OF PREDICTORS TESTED

In total, six types of predictors were built to compare different sets of features (**Table 1**). To investigate the predictive power of sequence and expression features separately, predictors based on either amino acid sequence or expression features were built. To test whether expression data provides additional information about subcellular localization that is not available in sequence data alone, further predictors using a combination of amino acid composition and expression features were built and the performance compared to the earlier predictors. The features were selected by

[6]www.ncbi.nlm.nih.gov/geo

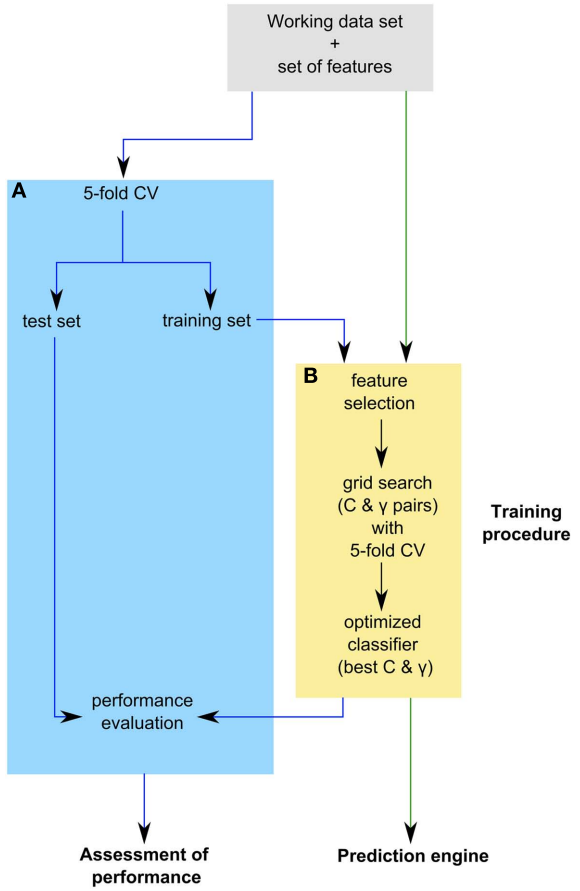


FIGURE 1 | Outline of prediction engine construction and evaluation. The prediction engine was constructed and its performance was evaluated in two independent runs on the entire working data set, as indicated by the green and blue arrows respectively. **(A)** Assessment of prediction engine performance through fivefold cross-validation (CV) loop. In each fold of CV an optimized classifier is built on the training set and its performance is assessed on the test set. The resulting performance measure distribution of the optimized classifiers was used to assess the performance of the prediction engine. **(B)** The training procedure used during prediction engine construction and evaluation. The training of a classifier involved feature selection and estimation of the best Cost (C) and gamma (γ) parameters using a grid search with fivefold CV loop. The best performing parameter pair was then used to train an optimized classifier.

using the above described method. The top 20 features were used as the stopping criterion to facilitate a fair comparison between predictors built on sequence, expression data, and mixed feature predictors. A further three types of predictors were built based on the top 1,000 expression features, the top 1,000 mixture of expression and amino acid composition features and the top 1,000 mixture of expression, amino acid composition and dipeptide features. Each predictor was tested using the above prediction engine evaluation procedure.

The final predictor, which was compared with the state of the art predictors, was built using top 1,000 features selected from a mixture of amino acid composition information and expression data. We found this number of features to be sufficient for our classifiers, as addition of a higher amount of features did not result in a noticeable improvement (data not shown).

Table 1 | Types of predictors tested and their underlying features.

Predictor	List of features
AA	Amino acid composition of 20 natural amino acids
T20 E	Top 20 expression features
T20 AA + E	Top 20 amino acid composition and expression features
T1000 E	Top 1000 expression features
T1000 AA + E	Top 1000 amino acid composition and expression features
T1000 AA + D + E	Top 1000 amino acid composition, dipeptide composition, and expression features

The top features were selected according to the rank given by adjusted F-score.

CATEGORY ENRICHMENT ANALYSIS

In order to search for enriched categories for the plastidial predictor, we tested for functional enrichment of the false negative and false positive set, using all "true" plastidial predictions and all proteins having an experimentally derived localization as backgrounds, respectively. The enrichment analysis was performed using the MapMan (Usadel et al., 2009) categories for TAIR9 and employing the online enrichment calculator based on Fisher's exact test (Usadel et al., 2006).

GENERATION OF CUSTOM VECTOR AND PROTEIN–GFP FUSION CONSTRUCTS

Two candidate genes, At1g16000.1 and At5g19540.1, whose subcellular localization was hitherto not experimentally determined (according to the SUBAII and TAIR database) were randomly selected. Our method predicted these to be localized in the mitochondrion and the plastid respectively. In order to validate our predictions, these two genes were cloned and the localization of their corresponding gene products investigated using protein–GFP fusions. Briefly, total RNA was isolated from entire *Arabidopsis* (Col-0) seedlings using the phenol–chloroform extraction method (as in Pant et al., 2009). Subsequently, the isolated RNA samples were digested with TURBO DNase (Ambion) and used as a template for reverse transcription using SuperScript® III Reverse Transcriptase Kit (Invitrogen) in the presence of the RNase inhibitor RNasion (Promega) as specified by the manufacturer. The coding sequence of the genes was amplified from this cDNA by PCR using Phusion DNA-Polymerase (Finnzymes). The primers used to obtain the final constructs are listed in **Table 2**. The pAM1 vector used for transient transformation was derived from pGreen0029 and pA7-GFP (Katrin Czempinski, Potsdam University, Germany) vectors. pGreen was digested at *Sma*I, *Ecl*136II, *Xho*I, *Sal*I, *Eco*RI, and *Hin*dIII restriction sites, to remove multiple cloning sites. The pA7-GFP vector was digested at *Eco*RI and *Hin*dIII restriction sites and this cassette, bearing GFP(S65T) under an enhanced version of CAMV35S promotor, was further cloned into the digested, as described above, pGreen0029 and relegated to give the pAM1 vector. Each candidate gene was inserted into pAM1 vector in two orientations, with respect to GFP sequence. By inserting the genes into pAM1 at either *Xba*I/*Bam*HI or Xho/*Nco*I restriction site,

Table 2 | Primers used for producing N-/C-terminal GFP fusion constructs together with their sequences.

Primer	Sequence
N-TERMINAL	
At1g16000N-fw	5′-ATCTAGAAATGGGAAATGAGACGAAGACCA-3′
At1g16000N-rev	5′-AGGATCCCTTGTTAGCTGATGAAGACGATGAG-3′
At5g19540N-fw	5′-AGCTAGCAATGGCGGTGAGCTCATTCGC-3′
At5g19540N-rev	5′-AGGATCCTACAATTTTTGTATTATCTATAAACT-3′
C-TERMINAL	
At1g16000C-fw	5′-ACTCGAGATGGGAAATGAGACGAAGACC-3′
At1g16000C-rev	5′-ATCCATGGCCTTGTTAGCTGATGAAGACGATGAG-3′
At5g19540C-fw	5′-ACTCGAGATGGCGGTGAGCTCATTCGC-3′
At5g19540C-rev	5′-ATCCATGGCTACAATTTTTGTATTATCTATAAACT-3′

N- and C-terminal GFP fusion constructs were obtained. The resulting inserts were sequenced to confirm correctness of the constructs.

TRANSIENT EXPRESSION IN TOBACCO

Five to 6-week-old tobacco protoplasts (cv. Petit havana) were generated and transformed via the polyethylene glycol-mediated (PEG) method adapted from Huang et al., 2002; Koop et al., 1996; Negrutiu et al., 1987. The transformed protoplasts were further incubated overnight in the dark. The protoplasts were transformed with the candidate gene–GFP constructs and control for the mitochondrion, pre101, and the plastid, TP101 (both controls, Renate Luhrs, personal communication) in parallel experiments. The protoplast cells, transformed with constructs and control for validation of At1g16000, were additionally stained with MitoTracker Orange (Invitrogen).

The transformed tobacco protoplasts were visualized 24 h after transformation using a confocal laser scanning microscope (TCS SP2/UV, Leica, Germany). The instrument was equipped with Argon and He/Ne lasers, and a 63× as well as a 20× planapo water objective. Two different filter settings were used: (i) for the GFP fluorescence excitation wave length: 488 nm, beam splitter: DD 488/568 (double dichroic, reflects at 488 and 568 nm), barrier filter: BP 530 (band pass, 515–545 nm); (ii) for the MitoTracker Orange, excitation wave length: 554 nm, beam splitter: DD 488/568, barrier filter: BP 590 (long pass > 590 nm). Autofluorescence of chlorophyll was detected at 580–600 nm. During image acquisition each line was scanned four times and averaged. Image analysis was performed with the Leica Confocal Software of TCS SP2 (version 2.61. build 1537).

RESULTS AND DISCUSSION

GENERATION OF A NOVEL SUBCELLULAR PREDICTION ENGINE

Many accurate subcellular localization predictors, including the one used by the TAIR database, rely on the targeting signal contained in the N-termini of proteins (Small et al., 2004; Emanuelsson et al., 2007). Therefore these predictors cannot estimate the correct subcellular localization if the N-terminus of proteins is absent. It had been shown, however, that the prediction of protein subcellular localization can be obtained by training a SVM employing the amino acid composition of a whole protein (Hua and Sun, 2001). Unfortunately, relying on amino acid composition alone has been shown to be insufficient for high accuracy predictions and consequently several predictors use additional information (Garg et al., 2005; Su et al., 2007; Blum et al., 2009; Kaundal et al., 2010).

We argue that in order to predict protein subcellular localization for plant species where no genome is available and thus full length transcript models are often lacking, one would need robust features that could be determined relatively quickly. It has previously been observed that *Arabidopsis* transcripts encoding for proteins localized in the plastid or in the mitochondrion are often highly correlated (Usadel et al., 2005, 2009; Cui et al., 2011) and that transcript accumulation in different experiments might therefore contain important information about protein localization.

To test whether expression data contained information about the subcellular localization, we extracted 3,707 slides from a compendium of *Arabidopsis* microarrays (Giorgi et al., 2010) and subjected them to principle component analysis (PCA). By using PCA we wanted to investigate whether a pattern in this expression data set exists, which would correlate with distribution of proteins in different subcellular localizations. The PCA revealed that over 80% of variance in the data could be explained by the first two principal components. Afterward, we projected the proteins in the coordinates of these two principle components and, to facilitate visual separation, we highlighted plastid proteins in green, leaving the proteins from the remaining compartments in black (**Figure 2**).

Within these projections most proteins lay on a somewhat diagonal line. However, it also became obvious that proteins separated off from this line by the second principal component tended to be enriched for plastid proteins (**Figure 2**). This observation indicated that expression data contains information that allows for

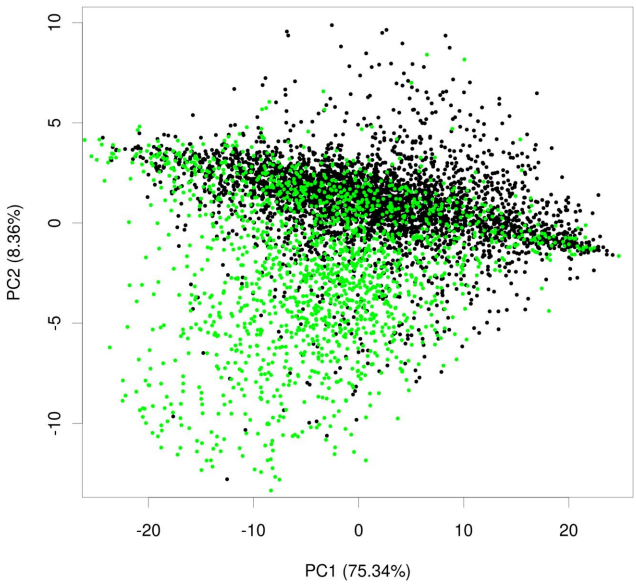


FIGURE 2 | Principle component analysis plot of plastid and non-plastid proteins. Exemplary principal component plot showing plastid proteins in green and proteins from other compartments in black.

a considerable degree of separation of plastid proteins from the background of proteins localized in the other compartments. We performed the same analysis for other compartments as well, but in no case did we see such a striking difference for the compartments for the first two principal components (**Figure A1** in Appendix). Furthermore, we wanted to check if this separation is conserved across species, and investigated if rice transcript data would also contain information that makes its plastid proteins distinguishable. To examine this, we performed PCA with the expression information from 487 experiments that used rice microarrays. Afterward, we projected the entire data in the coordinates of the first two principal components and highlighted the proteins, which were experimentally found in either the etioplast (von Zychlinski et al., 2005) or in the mitochondrion (Huang et al., 2009) in green and blue respectively, leaving the remaining proteins in black (**Figure A2** in Appendix). Here, we could also observe some degree of separation of plastid proteins (green) from other proteins (black). The separation from the rest of the proteins was much weaker for mitochondrial proteins (colored in blue), as in the case of *Arabidopsis*.

We therefore examined whether expression estimates could be combined with "traditional" data to predict the subcellular localization of plant proteins. To investigate this, we extracted only those proteins having an experimentally derived subcellular localization from the GO Slim annotation of the TAIR database. In total, this set comprised 6,188 proteins. After filtering for proteins, where we could find a unique probeset on the ATH1 chip, we were left with 5,429 proteins. These proteins were not evenly distributed between the different compartments. Here, as expected from the large organellar proteomics studies, a considerable portion was shown to be localized in the plastid or the mitochondrion (**Figure 3A**). Moreover, many proteins had been shown to be in the nucleus or the plasma membrane. Furthermore, for a significant proportion (24%) different experimentally determined localizations existed (**Table A1** in Appendix). Dual localization has probably been best studied for the plastid and the mitochondrion and Morgante et al. (2009) have already shown more than 50 *Arabidopsis* proteins to have these dual localization signals. This is reflected in the fact that most proteins from the plastid which have a second experimentally determined localization were also found in the mitochondrion (**Figure 3B**). However, for several other compartments such as the plasma membrane and the vacuole this was rather surprising and might indicate ambiguities in the data set or false positives in proteomic studies (**Figure 3B**; **Table A1** in Appendix).

Nevertheless we used the full experimentally determined protein set to train SVMs for the following compartments: the vacuole, the peroxisome, the cytosol, the ER, the plastid, the mitochondrion, the Golgi apparatus, the nucleus, the plasma membrane and the cell wall. It has to be noted that the latter is not representing any compartment but a training was attempted due to good experimentally derived evidence. In each case, we trained one SVM using only amino acid composition, one using the top 20 features selected from expression data, one incorporating the top 20 features chosen from the amino acid composition and expression behavior, one incorporating the top 1,000 expression features, one incorporating the top 1,000 features chosen from a mixture of amino acid composition and expression features and a final SVM, where the top 1,000 features were chosen from amino acid and dipeptide composition and transcript expression. The SVMs trained with the top 1,000 mixed features were used to gauge whether additional features beyond the amino acid composition could improve the SVM performance. On the other hand, the SVMs incorporating the top 20 mixed features were chosen to assess whether the inclusion of relatively few of these data sets would already increase prediction performance. Additionally, we

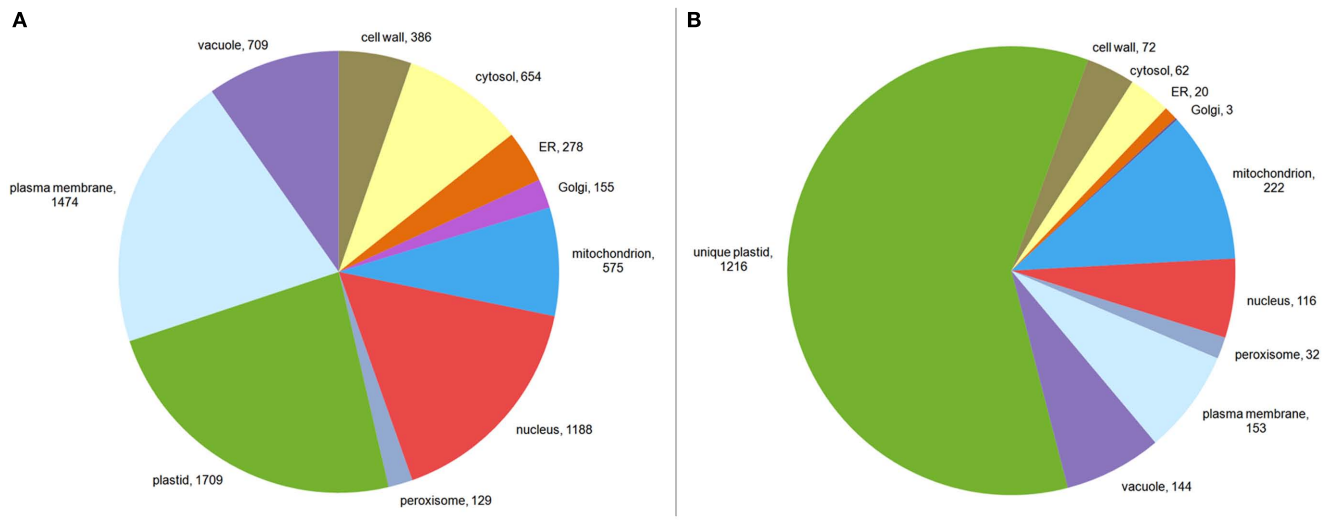


FIGURE 3 | Composition of the training set. The composition of the training set is shown in **(A)**. The size of each sector corresponds to all proteins in that compartment. Proteins localized to multiple compartments are counted toward each of their compartments. In **(B)** the data for plastid localized proteins from **Table A1** in Appendix is visualized, showing proteins having annotations for the plastid only, as well as for the plastid and other compartments.

wanted to test how informative the expression information on its own is and to investigate this we constructed SVMs based solely on expression features. The whole data set comprising 5,429 values by 5,327 features was then subjected to a model training procedure and subsequent evaluation using CV. We have performed feature selection by using an *F*-score based approach to identify features providing a high predictive power for the SVM (Chen and Lin, 2006). CV was used in two cases: once to estimate the parameters used to train the SVMs, and once to provide an unbiased assessment of prediction accuracy.

After evaluation of the prediction performance of the different SVMs, it became obvious that leveraging the expression of the underlying transcripts did not strongly improve the prediction, as judged by the MCC, for the cell wall, the cytosol, or the ER (**Figure 4**, upper panel and **Table 3**). In any case, for these compartments we only obtained a very low MCC (below 0.4) and therefore decided that these compartments could not be predicted solely based on these simple features. For the vacuole and the peroxisome we saw a slight increase of the MCC, but it stayed below a value of 0.4 (**Figure 4**, upper panel) and the predictive power was therefore also deemed to be not acceptable. In the case of the nucleus we did not observe any improvement in predictive power when incorporating expression data either (**Figure 4**, lower panel). However, here the addition of dipeptide composition elevated the MCC to nearly 0.5. Finally in the case of the plasma membrane, the Golgi and the mitochondrion we achieved an improvement of the predictive power by incorporating expression data, reaching MCC values slightly above 0.4 in every case (**Figure 4** lower panel). Strikingly, in accordance to the previous observations we saw a dramatic increase in MCC for the prediction of plastid proteins, where the MCC increased from below 0.4 to nearly 0.7 when about 1,000 array slides were incorporated (**Table 3**). Interestingly, when choosing as little as 20 features from the combined set of array slides and the amino acid composition the MCC rose to above 0.5 already, indicating that relatively few (targeted) expression arrays might be enough to significantly boost the predictive power for the plastid predictors.

Finally, when analyzing the performance of the SVMs based solely on the top 20 expression features we could notice that for the peroxisome, the cytosol, the ER, and the nucleus, the expression information alone is less informative than amino acid composition. However, with the same number of array slides, the predictor performance for the vacuole, the plastid, the Golgi apparatus, the plasma membrane, and the mitochondrion was already as good as, or sometimes even better than for those based on amino acid composition alone. In fact, what we have found characteristic for almost all compartments, except the nucleus and the cytosol, is that the 1,000 top expression features seemed to overlap with the

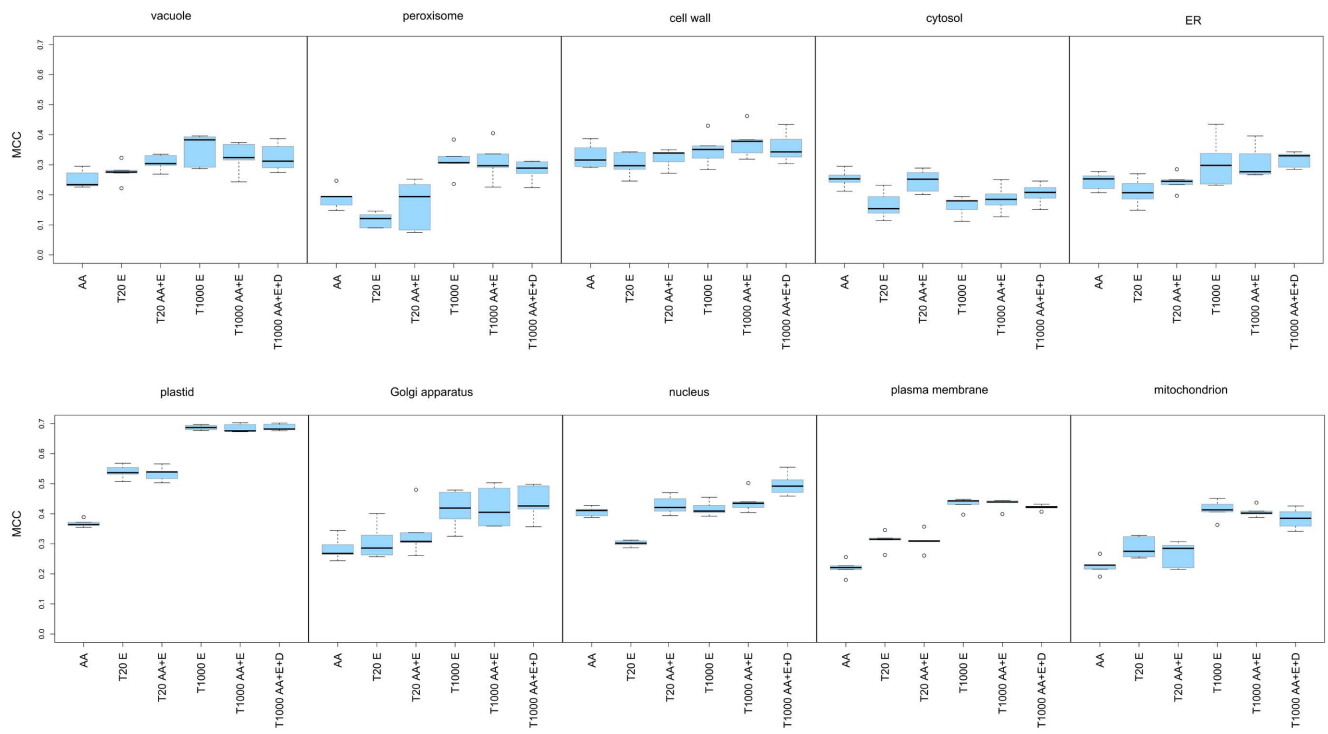


FIGURE 4 | Matthew's correlation coefficient plots presenting the performance of the predictors constructed for 10 subcellular compartments. The investigated compartments were: the vacuole, the peroxisome, the cell wall, the cytosol, the ER, the plastid, the Golgi apparatus, the nucleus, the plasma membrane, the mitochondrion. For each of the 10 compartments the prediction engines were built using: amino acid composition (AA), the top 20 expression features (T20 E), the top 20 mixed features selected from the amino acid composition and the expression data (T20 AA + E), the top 1,000 features selected from the expression features (T1000 E), the top 1,000 amino acid composition and expression features (T1000 AA + E) and the top 1,000 features selected from amino acid composition, dipeptide composition and expression data (T1000 AA + D + E). For each predictor the Matthews' correlation coefficients from the 5 cross-validation loops are visualized as a box plot.

Table 3 | Matthew's correlation coefficient values obtained using different features.

	AA	T20 E	T20 AA + E	T1000 E	T1000 AA + E	T1000 AA + E + D
Cell wall	0.33 ± 0.04	0.30 ± 0.04	0.32 ± 0.03	0.35 ± 0.05	0.38 ± 0.05	0.36 ± 0.05
Cytosol	0.25 ± 0.03	0.17 ± 0.05	0.25 ± 0.04	0.16 ± 0.03	0.19 ± 0.05	0.20 ± 0.04
ER	0.24 ± 0.03	0.21 ± 0.05	0.24 ± 0.03	0.31 ± 0.08	0.31 ± 0.06	0.32 ± 0.03
Golgi apparatus	0.28 ± 0.04	0.31 ± 0.06	0.34 ± 0.08	*0.42* ± 0.06	*0.42* ± 0.07	*0.44* ± 0.06
Mitochondrion	0.23 ± 0.03	0.29 ± 0.04	0.26 ± 0.04	*0.41* ± 0.03	*0.41* ± 0.02	0.38 ± 0.03
Nucleus	*0.41* ± 0.02	0.30 ± 0.01	*0.43* ± 0.03	*0.42 ± 0.02*	*0.44* ± 0.04	***0.50*** ± 0.04
Peroxisome	0.19 ± 0.04	0.06 ± 0.14	0.17 ± 0.08	0.31 ± 0.05	0.31 ± 0.07	0.28 ± 0.04
Plastid	0.37 ± 0.01	***0.54*** ± 0.02	***0.53*** ± 0.02	***0.69*** ± 0.01	***0.68*** ± 0.01	***0.69*** ± 0.01
Plasma membrane	0.22 ± 0.03	0.31 ± 0.03	0.31 ± 0.03	*0.43* ± 0.02	*0.43* ± 0.02	*0.42* ± 0.01
Vacuole	0.25 ± 0.03	0.28 ± 0.04	0.31 ± 0.03	0.35 ± 0.06	0.33 ± 0.05	0.32 ± 0.05

For each compartment the average MCC is given (± SD). The columns correspond to the amino acid composition as sole features (AA), the top 20 features chosen from the microarray slides (T20 E), the top 20 features chosen from the amino acid composition and the microarray slides (T20 AA + E), 1,000 top features chosen from the microarray slides (T1000 E), 1,000 top scoring features chosen from the amino acid composition and the microarray slides (T1000 AA + E) and finally the 1,000 top scoring features from the same set where dipeptide composition was added as an additional feature set (T1000 AA + E + D). Values above 0.4 are in italics and values above 0.5 in bold.

informative content of the protein sequence features (**Figure 4**), as the performance of predictors built on this data could not be further improved by incorporation of amino acid or dipeptide composition.

These results confirmed the initial findings from the PCA plots for the plastid. However, unlike in the PCA, we could show that expression profiling can provide useful information for half of the investigated compartments, albeit this improvement is not as dramatic as it is for the plastid. Furthermore, even the incorporation of relatively few expression sets increased the predictive power in the case of the plastid and for the plasma membrane (see **Figure 4** lower panel). This would suggest that, if one were to use expression information from crop or exotic plant species, a limited RNASeq profiling data set might be enough to provide an additional level of information for protein subcellular localization prediction, at the very least for plastid proteins.

IMPORTANCE OF INDIVIDUAL FEATURES FOR PLASTIDIAL PREDICTOR AS JUDGED BY AN ADJUSTED *F*-SCORE

We next set out to assess which data is most useful for the prediction of plastid proteins. We therefore investigated the ranking of the *F*-scores which were used for feature selection in the SVM training steps. As expected in the case of the plastid, microarray slides were residing at the top of the list (**Table A2** in Appendix). Interestingly, when assessing common themes amongst the microarrays providing most information about localization of plastid proteins, a set of microarrays studying a triose phosphate transporter mutant grown (Walters et al., 2004) under an 8-h light regime scored best. As even wild-type control arrays from this set were ranked amongst the most informative, it is likely that this might be rather due to the growth conditions and sampling time (2 h after light onset according to http://affymetrix.arabidopsis.info/narrays/experimentpage.pl?experimentid = 84) than the actual mutation, as many other top scoring arrays were from experiments investigating tissues grown under constant light (Schmid et al., 2005) or from the morning hours of carefully controlled diurnal cycles (Bläsing et al., 2005; Usadel et al., 2008). This might imply that one could tailor expression studies to be maximally beneficial for inferring protein subcellular localization, by choosing diurnal cycles or varying light intensities. This is not surprising, as many plastid proteins are obviously involved in light dependent processes and/or under the regulation of carbon status and react in response to either input. Consequently, when studying a carbon and light insensitive mutant, photosynthesis and plastid organization were the most significantly changed functional categories (Thum et al., 2008).

OVERREPRESENTED CATEGORIES

We next investigated whether we could detect any particular bias in the prediction accuracy for plastid localized genes. To investigate this, we used the proteins from our working data set and compared the set of false positives to all proteins contained in the working data set using the online MapMan enrichment tool (Usadel et al., 2006). In total, there were 23 false positive predictions, but we were not able to detect any meaningful enriched categories in this set (data not shown). Next we assessed the final false negative set which comprised 628 proteins for enriched categories by comparing it against the full set of 1,709 plastid proteins in the working data set. Interestingly, in this case we obtained many enriched categories pertaining to ribosomal proteins. However, it turned out that most of these were annotated as proteins constituting the eukaryotic ribosome. Furthermore, 10 proteins were classified as proteasome subunits. As in both cases plastid localization would be relatively unlikely, we concluded that these were either caused by experimental problems in high-throughput data sets or by a functional miss-annotation. We therefore revisited the underlying data by scrutinizing all 1,709 proteins from the plastid set manually without incorporating the novel predictions. We inferred subcellular localization based on experimental evidence and on textbook knowledge about processes and pathways. We further incorporated information about the occurrence of ribosomal subunits in cyanobacteria, algae, or bacteria derived from Interpro (Hunter et al., 2009) and by this checks we were indeed able to confirm the MapMan based annotations. We thus

concluded that 68 proteins were most likely not contained in plastids. After correcting our working data set based on these manual improvements, we did not seem to grossly improve SVM performance indicating that our training resulted in a relatively robust model despite the incorporation of false positives.

That said, the inclusion of at least ca. 5% false positive proteins in the plastid set shows that despite growing experimental evidence about the subcellular localization of proteins, these data have to be treated with caution. This is in agreement with the fact that organelle purification is not perfect (van Wijk and Baginsky, 2011). Furthermore, this observation is meaningful as it shows that – at least in the case of this novel plastid predictor – it is possible to find potential experimental errors by using *in silico* approaches. This further underlines the necessity for highly precise prediction tools even for well studies model organisms like *Arabidopsis*. It is likely that future studies will thus rely on intersected sets for training and testing and potentially weigh various experimental studies differently by assessing between-lab concordance.

COMPARISON OF PLASTIDIAL PREDICTORS PERFORMANCE USING AN INDEPENDENT TEST SET

We next compared the performance of our best performing classifier for the plastid with other state of the art predictors that could assess localization for this compartment. We chose TargetP, as this is being used by the TAIR database, Predotar, MultiLoc2, and AtSubP, as the latter represents another tool based on SVMs, which was specifically developed to annotate the *Arabidopsis* proteome and has been shown to have an excellent performance (Kaundal et al., 2010). Predictions made by Predotar and TargetP are based solely on the analysis of the N-terminal end of the protein sequence. Therefore these two predictors are tailored to predict mainly plastid or mitochondrial proteins. AtSubP and MultiLoc2 are another class of predictors which go beyond analysis of protein sequence and incorporate additional information. AtSubP leverages entire protein sequence composition and order, together with homology information using PSI–BLAST, to discriminate between proteins destined for seven plant compartments. MultiLoc2, apart from exhaustively analyzing protein sequence, incorporates additional protein information in the form of phylogenetic profiles and Gene Ontology terms to provide predictions for 10 plant subcellular compartments.

When comparing the performance of our predictor with that of other predictors according to the values from their internal performance validation tests, it became obvious that our MCC value estimated from CV was relatively low. However, this might be explained by the inclusion of many more proteins in our working data set or the inclusion of proteins which are hard to classify. We therefore composed an independent test data set, by querying the SUBAII subcellular localization database for proteins whose localization was experimentally confirmed. As the SUBAII database is curating protein subcellular localization independently from TAIR, we were thus able to obtain evidence for proteins not contained in our working data set. In total, we were able to retrieve experimentally derived subcellular localization annotations for 1,398 unique proteins for which expression information existed as well. Of these, 187 were from the plastid.

The compared predictors were queried with all proteins from the independent test data set and those predicted to be localized in the plastid were then selected for benchmarking. The Predotar predictions labeled as "possibly plastid" were not included. We next re-calculated the performance, for our SLocX predictor and the other four predictors, based on the independent test data set. As expected the performance dropped for all the predictors. Whilst it cannot be excluded that the independent test data set contains proteins which are harder to classify explaining the drop in MCC, the most likely explanation would be an overly optimistic estimation of MCC which might result from biases in CV (Jiang et al., 2008; Zervakis et al., 2009). However, we could show that on this independent test data set our plastidial predictor performed slightly better than Predotar and MultiLoc2. Generally, these three predictors performed better than the other two predictors by scoring MCC values of 0.48, 0.47, and 0.46 respectively (**Table 4**). Although Predotar and MultiLoc2 outcompeted SLocX in sensitivity, it still showed a higher precision. Even though TargetP made more true positive predictions than any of the three top predictors in **Table 4**, they were accompanied by almost the same number of false positive predictions and this was reflected in its very low precision (0.51). Interestingly, it can be noticed that the sensitivity of AtSubP, which is the highest of all classifiers, came at the cost of low precision as it made much more false positive predictions than true positive predictions. The low precision of AtSubP was also reflected in its MCC value of 0.32, which was the lowest among all the compared predictors. Additionally, we checked how the performance of Predotar would change after inclusion of its low confidence, "possible plastid," predictions. As expected, here we could observe a slight improvement

Table 4 | Benchmarking of predictions from SLocX, Predotar, MultiLoc2, TargetP, and AtSubP on the independent test set of 1,398 proteins.

Predictor	No. of predicted proteins	TP	FP	TN	FN	MCC	Precision	SE
SLocX	75	62	13	1198	125	**0.48**	0.83	0.33
Predotar	86	65	21	1190	122	**0.47**	0.76	0.35
MultiLoc2	90	66	24	1187	121	**0.46**	0.73	0.35
TargetP	144	74	70	1141	113	**0.38**	0.51	0.40
AtSubP	201	80	121	1090	107	**0.32**	0.40	0.43

According to SUBAII database, 187 proteins from the independent test data set were experimentally found in the plastid and 1211 in different compartments. The abbreviations mean: TP, true positive predictions; FP, false positive predictions; TN, true negative predictions; FN, false negative predictions; MCC, Matthew's correlation coefficient; SE, sensitivity. MCC values are given in bold.

in Predotar's sensitivity at the cost of lower precision (data not shown).

Given these differences, we investigated which proteins were correctly predicted by SLocX and the remaining classifiers and found these to have a relatively small overlap (data not shown). This might indicate that the protein sequence alone or enhanced with information derived from either homology, phylogenetic profiles, and GO annotations, does provide independent signals as compared to amino acid composition and expression data. Therefore, in the case of model species, where good gene models are known, it would thus likely make sense to combine such protein sequence-based prediction tools with the novel plastidial predictor.

EXPERIMENTAL VALIDATION OF CANDIDATE PROTEINS LOCALIZATION

As our results were comparing favorably to that of other prediction methods, we tried to validate two randomly chosen proteins by GFP fusions. The selected proteins were predicted by our method to be localized in the mitochondrion (At1g16000) and the plastid (At5g19540). We cloned the corresponding transcripts from seedling cDNA and transiently transformed tobacco leaf protoplasts. Each investigated protein was tagged with GFP either at its amino or carboxyl terminus. Tagging of the proteins in these two orientations was done to make sure that the observed localization was not due to the masking of a terminal signal peptide. We also queried publicly available prediction tools with the sequence of the investigated proteins.

The protein At1g16000 was predicted by our method to be located in the mitochondrion; however, both, Predotar and MitoProtII (Claros and Vincens, 1996) estimated the probability for an import of this protein into the mitochondrion at just 1 and 0.6% respectively. According to the specifications for interpretation of results of Predotar and MitoProtII, their predictions indicate that the protein is not localized in the mitochondrion. Furthermore, neither AtSubP nor TargetP were able to make any valid prediction for this protein whereas, MultiLoc2 predicted that this protein resides in the cytosol. The only prediction which overlapped with ours was the one made by Cui et al. (2011). After transforming the protoplasts with C-terminally tagged At1g16000 protein, we observed that the GFP signal overlaps with the cyan signal from MitoTracker (**Figures 5D–F**), which validates our prediction. This observation was additionally corroborated by the results obtained with the pre101(GFP) mitochondrial control (**Figures 5A-C**).

Interestingly enough, the cells expressing the N-terminally tagged version of this protein show a mitochondrial localization (**Figures 5G–I**). It came as a surprise to find both constructs in mitochondria, as it is known that proteins destined to this compartment usually contain an N-terminal mitochondrial transfer peptide (mTP) which should be blocked in case of the N-terminally tagged protein and therefore result in a different than mitochondrion localization. The reason for this behavior is unclear, but it might be explained by the presence of an alternative, not N-terminal, localization signal, which can reside inside of the protein sequence, as it was previously reported for a few mitochondrial proteins (Brix et al., 1999; Pfanner and Geissler, 2001). It could also be explained by the possibility that the available protein sequence is incomplete and its N-terminal part was wrongly assigned by gene prediction tools, thus making it impossible for the predictors based on N-terminal signal recognition to make a correct prediction. In order to exclude the possibility that At1g16000 is an incomplete gene model and to support the explanation that the observed localization was likely due to alternative localization signal, we filtered out the possible alternative starting sites and manually checked the 3,000-nucleotides upstream region of this gene. We found no putative N-terminal localization sequence (according to Predotar). The same result was achieved by checking for alternative starting codons in the first exon of this gene. The checked sequences are available in **Table A3** in Appendix. It appears that only the prediction methods which are not entirely based on protein sequence, but also on expression information, as ours and of Cui et al. (2011), can make a correct prediction in such cases.

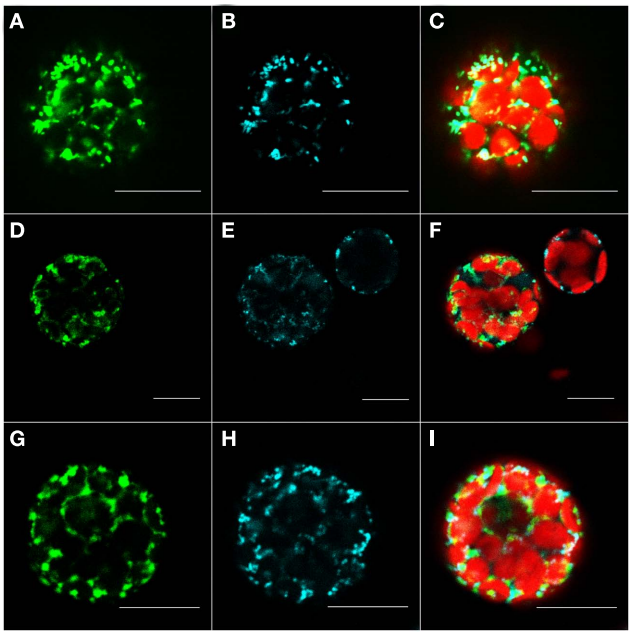


FIGURE 5 | Fluorescent microscopy analysis of tobacco protoplast cells transformed with At1g16000–GFP construct. Protoplast cells transformed with control for the mitochondrion – pre101(GFP) **(A–C)**, At1g16000 with C-terminally fused GFP **(D–F)**, and At1g16000 with N-terminally fused GFP **(G–I)**. Left panel – GFP (green fluorescence), middle panel – MitoTracker Orange (pseudo cyan fluorescence), right panel – channels overlay plus chlorophyll (red) autofluorescence. Bars in all pictures are 15 μm.

The second investigated protein, At5g19540, was predicted by our method to be localized in the plastid. In this case, Predotar, iPSORT and TargetP predicted that this protein contains a chloroplast transit peptide (cTP). Furthermore, the other prediction tools, such as MultiLoc2 and AtSubP also agreed with our verdict. The observed localization of C-terminally tagged At5g19540 protein indicated its localization to the plastid (**Figures 6C,D**). This observation was additionally validated by the results obtained with the TP101(GFP) plastidial control (**Figures 6A,B**).

As expected, the localization changed when the cells were transformed with an N-terminally GFP tagged protein. In this case our localization studies suggest a cytosolic location or a targeting to

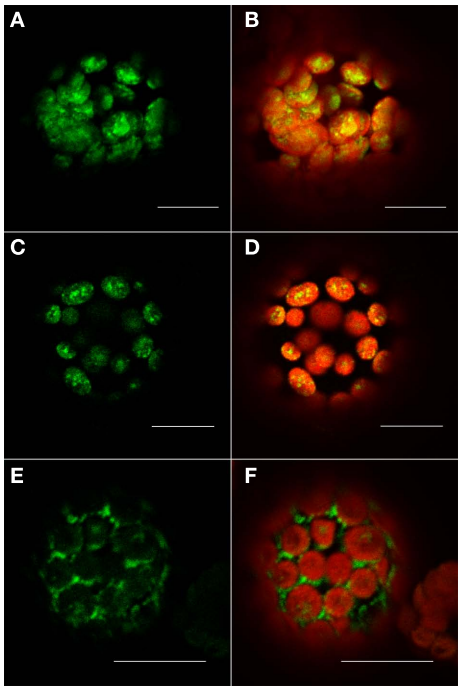


FIGURE 6 | Fluorescent microscopy analysis of tobacco protoplast cells transformed with At5g19540–GFP construct. Protoplast cells transformed with control for the plastid – TP101(GFP) **(A,B)**, At5g19540 with C-terminally fused GFP **(C,D)**, and At5g19540 with N-terminally fused GFP **(E,F)**. Left panel – GFP (green fluorescence), right panel – channels overlay plus chlorophyll (red) autofluorescence. Bars in all pictures are 15 μm.

the endoplasmic reticulum (**Figures 6E,F**). This two observations together demonstrate that this protein indeed contains a transit peptide at its N-terminus, as predicted by Predotar and iPSORT, which was masked in case of the N-terminally tagged version of this protein resulting in its possible mislocalization in cytosol/ER.

Taken together these experimental confirmations show that our novel predictor performs well on unknown proteins, and is indeed able to either correctly classify truncated mitochondrial proteins or to detect alternative localization signals for mitochondrial proteins.

LIMITATIONS OF THE METHOD AND FURTHER PERSPECTIVES

Given the performance of the SVM based predictor using simple amino acid and expression information it will be possible to combine these predictions with those stemming from N-terminal predictors for well studied model plants to (i) improve predictive power and in the case of conflicting predictions to (ii) potentially identify non-classically targeted proteins. Although, such leveraging of expression information for subcellular localization prediction appears promising, there are some limitations. Firstly, we could show that the compartments which can benefit from this information would be primarily the plastid and, to some extent, the mitochondrion and the plasma membrane. However, the main limitation is the need to have expression data for the protein to be studied. Therefore, our predictor requires that a protein's transcript must be represented on the ATH1 microarray. Generalizing this, repeating our methodology for other plant species would depend on the availability of data from experiments performed using microarrays designed for them. Moreover, it cannot be guaranteed that this would be as robust as for *Arabidopsis* and would depend of the quality of the microarrays, i.e., the number of transcripts that they measure. These limitations however, might no longer be a bottleneck of our methodology, since next generation sequencing can now provide expression measures for entire transcriptomes and this technique was already applied many times for *Arabidopsis* and other plant species (Jia et al., 2009; Eveland et al., 2010; Filichkin et al., 2010; Gilardoni et al., 2010; Zhang et al., 2010; Hsieh et al., 2011). As RNASeq projects can be used to infer (often incomplete) transcript and thus protein models at the same time, a prediction solely based on amino acid composition and expression information should be highly useful for these studies.

WEBSITE

In order to make the data available in a convenient form, we have set up a website of localizations predicted by SLocX. The website is available at the following URL: mapman.mpimp-golm.mpg.de/general/slocx/. Additional improvements will directly be incorporated into the database.

CONCLUSION

By leveraging gene expression information we could show that we can predict protein subcellular localization with a significantly higher accuracy than when using sequence data alone. Beyond simple CV and an independent test set, a subset of novel predictions was also shown to be correct using protein–GFP fusions.

ACKNOWLEDGMENTS

We want to thank Dr. Sandra Tanz for making data from the SUBA II database available. We are very grateful to Yvonne Weber for technical assistance with protoplast transformation. We further acknowledge Anthony Bolger for meticulous correction of the manuscript. Also we want to thank Eugenia Maximova for assistance with microscopy. Furthermore we want to acknowledge Diana Pese for general technical support. Last but not the least, we would like to acknowledge Paulina Troc for useful discussions.

REFERENCES

Baerenfaller, K., Grossmann, J., Grobei, M., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W., and Baginsky, S. (2008). Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics *Science* 320, 938–941.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S. (2002). Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* 18, 298–305.

Barrett, T., Troup, D., Wilhite, S., Ledoux, P., Evangelista, C., Kim, I., Tomashevsky, M., Marshall, K., Phillippy, K., Sherman, P., Muertter, R. N., Holko, M., Ayanbule, O., Yefanov, A., and Soboleva, A. (2011). NCBI GEO: archive for functional genomics data sets – 10 years on. *Nucleic Acids Res.* 39(Suppl. 1), D1005.

Bläsing, O. E., Gibon, Y., Günther, M., Höhne, M., Morcuende, R., Osuna, D., Thimm, O., Usadel, B., Scheible, W. R., and Stitt, M. (2005). Sugars and circadian regulation make major contributions to the

global regulation of diurnal gene expression in *Arabidopsis*. *Plant Cell* 17, 3257–3281.

Blum, T., Briesemeister, S., and Kohlbacher, O. (2009). MultiLoc 2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics* 10, 274. doi: 10.1186/1471-2105-10-274

Brix, J., Rüdiger, S., Bukau, B., Schneider-Mergener, J., and Pfanner, N. (1999). Distribution of binding sequences for the mitochondrial import receptors Tom20, Tom22, and Tom70 in a presequence-carrying preprotein and a non-cleavable preprotein. *J. Biol. Chem.* 274, 16522.

Chang, C., and Lin, C. (2011). *LIBSVM: a library for support vector machines*, ACM Transactions on Intelligent Systems and Technology (TIST), Taipei, 2, 27.

Chen, Y.-W., and Lin, C. J. (2006). "Combining SVMs with various feature selection strategies," in *Feature Extraction, Foundations and Applications,* eds I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh (Taipei: Springer), 315–324.

Claros, M. G., and Vincens, P. (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* 241, 779–786.

Cui, J., Liu, J., Li, Y., and Shi, T. (2011). Integrative identification of *Arabidopsis* mitochondrial proteome and its function exploitation through protein interaction network. *PLoS ONE* 6, e16022. doi: 10.1371/journal.pone.0016022

Diekert, K., Kispal, G., Guiard, B., and Lill, R. (1999). An internal targeting signal directing proteins into the mitochondrial intermembrane space. *Proc. Natl. Acad. Sci. U.S.A.* 196, 11752–11757.

Dunkley, T., Hester, S., Shadforth, I., Runions, J., Weimar, T., Hanton, S., Griffin, J., Bessant, C., Brandizzi, F., Hawes, C., Watson, R. B., Dupree, P., and Lilley, K. S. (2006). Mapping the *Arabidopsis* organelle proteome. *Proc. Natl. Acad. Sci. U.S.A.* 103, 6518–6523.

Dunkley, T. P., Watson, R., Griffin, J. L., Dupree, P., and Lilley, K. S. (2004). Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics* 3, 1128–1134.

Edgar, R., Domrachev, M., and Lash, A. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207.

Eisenhaber, F., and Bork, P. (1998). Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol.* 8, 169.

Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* 2, 953–971.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005–1016.

Eveland, A., Satoh-Nagasawa, N., Goldshmidt, A., Meyer, S., Beatty, M., Sakai, H., Ware, D., and Jackson, D. (2010). Digital gene expression signatures for maize development. *Plant Physiol.* 154, 1024–1039.

Ferro, M., Brugière, S., Salvi, D., Seigneurin-Berny, D., Court, M., Moyet, L., Ramus, C., Miras, S., Mellal, M., Le Gall, S., Kieffer-Jaquinod, S., Bruley, C., Garin, J., Joyard, J., Masselon, C., and Rolland, N. (2010). AT CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. *Mol. Cell. Proteomics* 9, 1063.

Filichkin, S., Priest, H., Givan, S., Shen, R., Bryant, D., Fox, S., Wong, W., and Mockler, T. (2010). Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.* 20, 45–58.

Gardy, J., and Brinkman, F. (2006). Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.* 4, 741–751.

Garg, A., Bhasin, M., and Raghava, G. (2005). Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J. Biol. Chem.* 280, 14427–14432.

Gerhardt, R., and Heldt, H. W. (1984). Measurement of subcellular metabolite levels in leaves by fractionation of freeze-stopped material in nonaqueous media. *Plant Physiol.* 75, 542–547.

Gilardoni, P., Schuck, S., Jüngling, R., Rotter, B., Baldwin, I., and Bonaventure, G. (2010). SuperSAGE analysis of the *Nicotiana attenuata* transcriptome after fatty acid-amino acid elicitation (FAC): identification of early mediators of insect responses. *BMC Plant Biol.* 10, 66. doi: 10.1186/1471-2229-10-66

Giorgi, F., Bolger, A., Lohse, M., and Usadel, B. (2010). Algorithm-driven artifacts in median polish summarization of microarray data. *BMC Bioinformatics* 11, 553. doi: 10.1186/1471-2105-11-553

Heazlewood, J., Tonti-Filippini, J., Gout, A., Day, D., Whelan, J., and Millar, A. (2004). Experimental analysis of the *Arabidopsis* mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. *Plant Cell* 16, 241–256.

Heazlewood, J., Verboom, R., Tonti-Filippini, J., Small, I., and Millar, A. (2006). SUBA: the Arabidopsis subcellular database. *Nucleic acids res.* 35(suppl 1), D213.

Herman, E., and Schmidt, M. (2004). Endoplasmic reticulum to vacuole trafficking of endoplasmic reticulum bodies provides an alternate pathway for protein transfer to the vacuole. *Plant Physiol.* 136, 3440–3446.

Hsieh, T., Shin, J., Uzawa, R., Silva, P., Cohen, S., Bauer, M., Hashimoto, M., Kirkbride, R., Harada, J., Zilberman, D., and Fischera, R. L. (2011). Regulation of imprinted gene expression in *Arabidopsis* endosperm. *Proc. Natl. Acad. Sci. U.S.A.* 108, 1755–1762.

Hsu, C., Chang, C., and Lin, C. (2008). *A practical guide to support vector classification.* National Taiwan University, Taipei 106, Taiwan.

Hua, S., and Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 17, 721–728.

Huang, F., Klaus, S., Herz, S., Zou, Z., Koop, H., and Golds, T. (2002). Efficient plastid transformation in tobacco using the aphA-6 gene and kanamycin selection. *Mol. Genet. Genomics* 268, 19–27.

Huang, S., Taylor, N. L., Narsai, R., Eubel, H., Whelan, J., and Millar, A. H. (2009). Experimental analysis of the rice mitochondrial proteome, its biogenesis, and heterogeneity. *Plant Physiol.* 149,719–734.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res.* 37, D211–D215.

Ito, J., Batth, T., Petzold, C., Redding-Johanson, A., Mukhopadhyay, A., Verboom, R., Meyer, E., Millar, A., and Heazlewood, J. (2010). Analysis of the *Arabidopsis* cytosolic proteome highlights subcellular partitioning of central plant metabolism. *J. Proteome Res.* 10, 1571–1582.

Jia, Y. ; Lisch, D., Ohtsu, K., Scanlon, M., Nettleton, D., and Schnable, P. (2009). Loss of RNA–dependent RNA polymerase 2 (RDR2) function causes widespread and unexpected changes in the expression of transposons, genes, and 24-nt small RNAs. *PLoS Genet.* 5, e1000737. doi: 10.1371/journal.pgen.1000737

Jiang, W., Varma, S., and Simon, R. (2008). Calculating confidence intervals for prediction error in microarray classification using resampling. *Stat. Appl. Genet. Mol. Biol.* 7, article 8.

Kaundal, R., Saini, R., and Zhao, P. (2010). Combining machine learning and homology-based approaches to accurately predict subcellular localization in *Arabidopsis*. *Plant Physiol.* 154, 36–54.

Koop, H., Steinmüller, K., Wagner, H., Rößler, C., Eibl, C., and Sacher, L. (1996). Integration of foreign sequences into the tobacco plastome via polyethylene glycol-mediated protoplast transformation. *Planta* 199, 193–201.

Koroleva, O. A., Tomlinson, M. L., Leader, D., Shaw, P., and Doonan, J. H. (2005). High-throughput protein localization in *Arabidopsis* using *Agrobacterium*-mediated transient expression of GFP-ORF fusions. *Plant J.* 41, 162–174.

Krueger, S., Giavalisco, P., Krall, L., Steinhauser, M. C., Bussis, D., Usadel, B., Flugge, U. I., Fernie, A. R., Willmitzer, L., and Steinhauser, D. (2011). A topological map of the compartmentalized *Arabidopsis thaliana* leaf metabolome. *PLoS ONE* 6, e17806. doi: 10.1371/journal.pone.0017806

Lunn, J. E. (2007). Compartmentation in plant metabolism. *J. Exp. Bot.* 58, 35–47.

Majeran, W., Cai, Y., Sun, Q., and van Wijk, K. J. (2005). Functional differentiation of bundle sheath and mesophyll maize chloroplasts determined by comparative proteomics. *Plant Cell* 17, 3111–3140.

Marcotte, E., Xenarios, I., Van Der Bliek, A., and Eisenberg, D. (2000). Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.* 97, 12115.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405, 442–451.

Morgante, C. V., Rodrigues, R. A., Marbach, P. A., Borgonovi, C. M., Moura, D. S., and Silva-Filho, M. C. (2009). Conservation of dual-targeted proteins in *Arabidopsis* and rice points to a similar pattern of gene-family evolution. *Mol. Genet. Genomics* 281, 525–538.

Mutwil, M., Klie, S., Tohge, T., Giorgi, F., Wilkins, O., Campbell, M., Fernie, A., Usadel, B., Nikoloski, Z., and Persson, S. (2011). PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* 23, 895.

Negrutiu, I., Shillito, R., Potrykus, I., Biasini, G., and Sala, F. (1987). Hybrid genes in the analysis of transformation conditions. *Plant Mol. Biol.* 8, 363–373.

Nickel, W., and Seedorf, M. (2008). Unconventional mechanisms of protein transport to the cell surface of eukaryotic cells. *Annu. Rev. Cell Dev. Biol.* 24, 287–308.

Nishikawa, K., Kubota, Y., and Ooi, T. (1983). Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. *J. Biochem.* 94, 981–995.

Pant, B., Musialak-Lange, M., Nuc, P., May, P., Buhtz, A., Kehr, J., Walther, D., and Scheible, W. (2009). Identification of nutrient-responsive *Arabidopsis* and rapeseed microRNAs by comprehensive real-time polymerase chain reaction profiling and small RNA sequencing. *Plant Physiol.* 150, 1541–1555.

Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I., Farne, A., Holloway, E., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rayner, T. F., Rezwan, F., Sharma, A., Williams, E., Bradley, X. Z., Adamusiak, T., Brandizi, M., Burdett, T., Coulson, R., Krestyaninova, M., Kurnosov, P., Maguire, E., Neogi, S. G., Rocca-Serra, P., Sansone, S. A., Sklyar, N., Zhao, M., Sarkans, U., and Brazma, A. (2009). ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.* 37(Suppl. 1), D868.

Pfanner, N., and Geissler, A. (2001). Versatility of the mitochondrial protein import machinery. *Nat. Rev. Mol. Cell Biol.* 2, 339–349.

Pfannschmidt, T. (2010). Plastidial retrograde signaling – a true "plastid factor" or just metabolite signatures? *Trends Plant Sci.* 15, 427–435.

Rhee, S., Beavis, W., Berardini, T., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J., and Zhang, P. (2003). The *Arabidopsis* information resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.* 31, 224–228.

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., and Lohmann, J. U. (2005). A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* 37, 501–506.

Schulze, W., and Usadel, B. (2010). Quantitation in mass-spectrometry-based proteomics. *Annu. Rev. Plant Biol.* 61, 491–516.

Schwacke, R., Fischer, K., Ketelsen, B., Krupinska, K., and Krause, K. (2007). Comparative survey of plastid and mitochondrial targeting properties of transcription factors in *Arabidopsis* and rice. *Mol. Genet. Genomics* 277, 631–646.

Severin, A. J., Woody, J. L., Bolon, Y. T., Joseph, B., Diers, B. W., Farmer, A. D., Muehlbauer, G. J., Nelson, R. T., Grant, D., Specht, J. E., Graham, M. A., Cannon, S. B., May, G. D., Vance, C. P., and Shoemaker, R. C. (2010). RNA-Seq atlas of glycine max: a guide to the soybean transcriptome. *BMC Plant Biol.* 10, 160. doi: 10.1186/1471-2229-10-160

Small, I., Peeters, N., Legeai, F., and Lurin, C. (2004). Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* 4, 1581–1590.

Su, E., Chiu, H., Lo, A., Hwang, J., Sung, T., and Hsu, W. (2007). Protein subcellular localization prediction based on compartment-specific features and structure conservation. *BMC Bioinformatics* 8, 330. doi: 10.1186/1471-2105-8-330

Thum, K. E., Shin, M. J., Gutiérrez, R. A., Mukherjee, I., Katari, M. S., Nero, D., Shasha, D., and Coruzzi, G. M. (2008). An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in *Arabidopsis*. *BMC Syst. Biol.* 2, 31. doi: 10.1186/1752-0509-2-31

Trotter, M. W., Sadowski, P. G., Dunkley, T. P., Groen, A. J., and Lilley, K. S. (2010). Improved sub-cellular resolution via simultaneous analysis of organelle proteomics data across varied experimental conditions. *Proteomics* 10, 4213–4219.

Usadel, B., Blasing, O. E., Gibon, Y., Retzlaff, K., Hohne, M., Gunther, M., and Stitt, M. (2008). Global transcript levels respond to small changes of the carbon status during a progressive exhaustion of carbohydrates in *Arabidopsis* rosettes. *Plant Physiol.* 146, 1834–1861.

Usadel, B., Nagel, A., Steinhauser, D., Gibon, Y., Blaesing, O. E., Redestig, H., Sreenivasulu, N., Krall, L., Hannah, M. A., Poree, F., Fernie, A. R., and Stitt, M. (2006). PageMan an interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. *BMC Bioinformatics* 18, 535. doi: 10.1186/1471-2105-7-535

Usadel, B., Nagel, A., Thimm, O., Redestig, H., Blaesing, O. E., Palacios-Rojas, N., Selbig, J., Hannemann, J., Piques, M. C., Steinhauser, D., Scheible, W. R., Gibon, Y., Morcuende, R., Weicht, D., Meyer, S., and Stitt, M. (2005). Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses. *Plant Physiol.* 138, 1195–1204.

Usadel, B., Poree, F., Nagel, A., Lohse, M., Czedik-Eysenberg, A., and Stitt, M. (2009). A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. *Plant Cell Environ.* 32, 1211–1229

van Wijk, K. J. (2004). Plastid proteomics. *Plant Physiol. Biochem.* 42, 963–977.

van Wijk, K. J., and Baginsky, S. (2011). Plastid proteomics in higher plants: current state and future goals. *Plant Physiol.* 155, 1578–1588.

von Heijne, G., Steppuhn, J., and Herrmann, R. G. (1989). Domain structure of mitochondrial and chloroplast targeting peptides. *Eur. J. Biochem.* 180, 535–545.

von Zychlinski, A., Kleffmann, T., Krishnamurthy, N., Sjölander, K., Baginsky, S., and Gruissem, W. (2005). Proteome analysis of the rice etioplast: metabolic and regulatory networks and novel protein functions. *Mol. Cell Proteomics* 4, 1072–1084.

Walters, R. G., Ibrahim, D. G., Horton, P., and Kruger, N. J. (2004). A mutant of *Arabidopsis* lacking the triose-phosphate/phosphate translocator reveals metabolic regulation of starch breakdown in the light. *Plant Physiol.* 135, 891–906.

Wienkoop, S., Baginsky, S., and Weckwerth, W. (2010). *Arabidopsis thaliana* as a model organism for plant proteome research. *J. Proteomics* 73, 2239–2248.

Zervakis, M., Blazadonakis, M. E., Tsiliki, G., Danilatou, V., Tsiknakis, M., and Kafetzopoulos, D. (2009). Outcome prediction based on microarray analysis: a critical perspective on methods. *BMC Bioinformatics* 10, 53. doi: 10.1186/1471-2105-10-53

Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., Zhuang, R., Lu, Z., He, Z., Fang, X., Chen, L., Tian, W., Tao, Y., Kristiansen, K., Zhang, X., Li, S., Yang, H., Wang, J., and Wang, J. (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res.* 20, 646–654.

Conflict of Interest Statement: 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.

Received: 18 May 2011; accepted: 12 August 2011; published online: 12 September 2011.

Citation: Ryngajllo M, Childs L, Lohse M, Giorgi FM, Lude A, Selbig J and Usadel B (2011) SLocX: predicting subcellular localization of Arabidopsis proteins leveraging gene expression data. Front. Plant Sci. 2:43. doi: 10.3389/fpls.2011.00043

This article was submitted to Frontiers in Plant Physiology, a specialty of Frontiers in Plant Science.

APPENDIX

The formula used to linearly scale the microarray data between values of 0 and 1. V, $V_{\min}$, and $V_{\max}$ are, respectively, the value to be scaled, the smallest, and the largest value in the expression data set.

$$\text{Scaled Value} = \frac{V - V_{\min}}{V_{\max} - V_{\min}} \tag{A1}$$

The formula used to calculate Matthews' correlation coefficient (MCC). Where, the true positive (TP) predictions is the total number of correctly predicted proteins which are localized in a particular compartment, the true negative (TN) predictions is the total number of proteins correctly predicted not to be localized in a particular compartment, the false positive (FP) predictions is the total number of proteins incorrectly predicted to be localized in a particular compartment, the false negative (FN) predictions is the total number of proteins incorrectly predicted not to be localized in a given compartment.

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}} \tag{A2}$$

The formula used to calculate sensitivity (SE).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{A3}$$

The formula used to calculate precision.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{A4}$$

Table A1 | Overlap between localizations for proteins representing 10 compartments.

	cw	Cytosol	ER	Golgi	Mitochondrion	Nucleus	Peroxisome	Plastid	pm	Vacuole
cw	**386**	35	15	5	40	47	5	72	96	84
Cytosol	35	**654**	11	6	23	343	6	62	133	34
ER	15	11	**278**	13	14	20	4	20	76	63
Golgi	5	6	13	**155**	1	3	0	3	24	18
Mitochondrion	4U	4	14	1	**575**	35	14	222	52	79
Nucleus	47	343	20	3	35	**1188**	10	116	130	64
Peroxisome	5	6	4	0	14	10	**129**	32	14	17
Plastid	72	62	20	3	222	116	32	**1709**	153	144
pm	96	133	76	24	52	130	14	153	**1474**	197
Vacuole	84	34	63	18	79	64	17	144	197	**709**

Proteins annotated to be localized to multiple compartments are shown. For each combination of compartments the total number of shared proteins is given. The numbers in the diagonal give the total number of proteins per compartment as a reference. Abbreviations: cw, cell wall; pm, plasma membrane.

Table A2 | Top Scoring Arrays for the plastid.

Array accession code	*F*-score	Title
GSM133833.CEL	0.620	Walters A-12-Kruger-MH3 REP3
GSM133831.CEL	0.613	Walters_A-10-Kruger-MH1_REP 1
GSM133826.CEL	0.610	Walters A-05-Kruger-WH2 REP2
GSM133828.CEL	0.588	Walters_A-07-Kruger-ML1_REP1
GSM133827.CEL	0.584	Walters A-06-Kruger-WH3 REP3
GSM133830.CEL	0.579	Walters_A-09-Kruger-ML3_REP3
GSM133832.CEL	0.570	Walters A-11-Kruger-MH2 REP2
GSM133825.CEL	0.567	Walters_A-04-Kruger-WH1_REP 1
GSM133823.CEL	0.565	Walters A-02-Kruger-WL2 REP2
GSM133824.CEL	0.558	Walters_A-03-Kruger-WL3_REP3
GSM318330.CEL	0.545	EL 14DAS 1
GSM183507.CEL	0.543	WT_for_ATR1/MYB51rep1
GSM131473.CEL	0.542	ATGE 7 C2
GSM133822.CEL	0.541	Walters_A-01-Kruger-WL 1 REP 1
GSM133829.CEL	0.538	Walters A-08-Kruger-ML 2 REP2
GSM131472.CEL	0.533	ATGE 7 B2
GSM131471.CEL	0.532	ATGE 7 A2
GSM131500.CEL	0.532	ATGE 5 C
GSM131499.CEL	0.530	ATGE 5 B
GSM45208.CEL	0.530	00304WT 1
GSM131501.CEL	0.529	ATGE 10 A
GSM131503.CEL	0.522	ATGE 10 C
GSM131502.CEL	0.527	ATGE 10 B
GSM131498.CEL	0.525	ATGE5A
GSM45278.CEL	0.522	00304AS12_2
AtGen_6-9512_Heatstress(3h) + 9hrecovery-Shoots-		
GSM131464.CEL	0.510	12.0h_Rep2
GSM318331.CEL	0.509	EL14DAS2
GSM183508.CEL	0.509	WT_for_ATRI/MYB5 I_rep2
AtGen 6-9511 Heatstress(3h) + 9hrecovery-Shoots-		
GSM131463.CEL	0.502	12.0h_Repl
GSM269488.CEL	0.501	mkk2, no-treatment, rep-A
gsm77059.CEL	0.500	04h Col-0 replicate B
GSM135552.CEL	0.499	syd-2_rep2
GSM135551.CEL	0.495	syd-2_repl
gsm77062.CEL	0.495	08h Col-0 replicate B
GSM265858.CEL	0.495	control shortB
GSM183516.CEL	0.494	MYB51_OE_repl
GSM268009.CEL	0.494	Col-0, Time 0, rep-B
GSM133084.CEL	0.492	JD AT + EO COL WT 24H UNINFECTED
GSM269490.CEL	0.491	mkk2, no-treatment, rep-C
GSM45209.CEL	0.491	00304WT_2
GSM133078.CEL	0.490	JD AT + EO COL WT 06H UNINFECTED
AtGen 6-9611 Heatstress(3h) + 21hrecovery-Shoots-		
GSM131467.CEL	0.489	24.0h_Repl
GSM265868.CEL	0.489	long 10B
GSM183512.CEL	0.486	MYB76_OE_rep2
AtGen_6-9612_Heatstress(3h) + 21 hrecovery-Shoots-		
GSM131468.CEL	0.485	24.0h_Rep2
GSM131252.CEL	0.483	AtGen_6-0512_Control-Shoots-12.0h_Rep2
GSM133079.CEL	0.481	JD AT + EO COL WT 12H INFECTED
GSM131260.CEL	0.481	AtGen_6-I 112_Cold(4°C)-Shoots-0.5h_Rep2
GSM131251.CEL	0.481	AtGen_6-051 I_Control-Shoots-12.0h_Repl

The adjusted F-score, the Arrays accession code as well as a title for the arrays series is given. Arrays from the same series are colored in the same color.

Table A3 | Sequences of upstream and downstream regions of At1g1600.

Seq id	Sequence
>5′3′ Frame 1-1	MKSPKLTCYKLQLFSFKSESLYFSQSLHCSCGRR
>5′3′ Frame 1-2	MAFDVSSEILR
>5′3′ Frame 1-3	MHPLF
>5′3′ Frame 1-4	MTMSCPRLT
>5′3′ Frame 1-5	MSCPRLT
>5′3′ Frame 1-6	MLLMIQCLKI
>5′3′ Frame 1-7	MIQCLKI
>5′3′ Frame 1-8	MCL
>5′3′ Frame 1-9	MSQNTN
>5′3′ Frame 1-10	MSFIDLKKTKKNIAIF
>5′3′ Frame 1-11	MYWDLYIILRNNHKLHAKINLTTSQQISII
>5′3′ Frame 1-12	MWESV
>5′3′ Frame 1-13	MIKENLGLEET
>5′3′ Frame 1-14	MRSVFTAYFDEARRVIIALFSSI
>5′3′ Frame 1-15	MGFKMLFNKKEILC
>5′3′ Frame 1-16	MLFNKKEILC
>5′3′ Frame 2-1	MKQEAQVLHC
>5′3′ Frame 2-2	MNIISLTGSPSRTM
>5′3′ Frame 2-3	MSLLRSLGKL
>5′3′ Frame 2-4	MFSSSTPLVSNHLY
>5′3′ Frame 2-5	MRSAKRRSPAIAIAMENKTSPGNVLVCSP
>5′3′ Frame 2-6	MENKTSPGNVLVCSP
>5′3′ Frame 2-7	MKTPKMSRVLCTSYRLNQ
>5′3′ Frame 2-8	MSRVLCTSYRLNQ
>5′3′ Frame 2-9	MMNKCLKTLIKKSHIYIETLTWLASIYQRR
>5′3′ Frame 2-10	MNKCLKTLIKKSHIYIETLTWLASIYQRR
>5′3′ Frame 2-11	MFNNAVFVGNTSDPLDP
>5′3′ Frame 2-12	MVLRVVVVTASFVSIPIQLLPELSTMGR
>5′3′ Frame 2-13	MGR
>5′3′ Frame 2-14	MTNNLFHTRSVLS
>5′3′ Frame 2-15	MQKLT
>5′3′ Frame 2-16	MTDE
>5′3′ Frame 2-17	MSESYHASTLICNKIWGLKCYSIKRKSYVDGP
>5′3′ Frame 3-1	MFVEPVDYEVS
>5′3′ Frame 3-2	MSEALHHKSLLLLTTLC
>5′3′ Frame 3-3	MYLSAALDLTCCS
>5′3′ Frame 3-4	MPEDIA
>5′3′ Frame 3-5	MYQNAPVICQNVFVKSESDQ
>5′3′ Frame 3-6	MLTE
>5′3′ Frame 3-7	MPFSLVTHPIL
>5′3′ Frame 3-8	MLSSFHLLGSLG
>5′3′ Frame 3-9	MK
>5′3′ Frame 3-10	MND
>5′3′ Frame 3-11	MKKLRVPT
>5′3′ Frame 3-12	MND
>5′3′ Frame 3-13	MSSEFTAYFLVKL
>5′3′ Frame 3-14	MLMGHNKAHLYMVLKPLMDKPC
>5′3′ Frame 3-15	MGHNKAHLYMVLKPLMDKPC
>5′3′ Frame 3-16	MVLKPLMDKPC
>5′3′ Frame 3-17	MDKPC
>At1g16000_down1	MAGGGGFRAKMEHYVYSGEKKHVLVGIGIVTIIFGVPWYLMTQG SKHQSHQDYMDKADKARKARLSSSSSANK
>At1g16000_down2	MEHYVYSGEKKHVLVGIGIVTIIFGVPWYLMTQGSKHQSHQDYM DKADKARKARLSSSSSANK

The sequences were searched for a N-terminal targeting signal for mitochondrion.

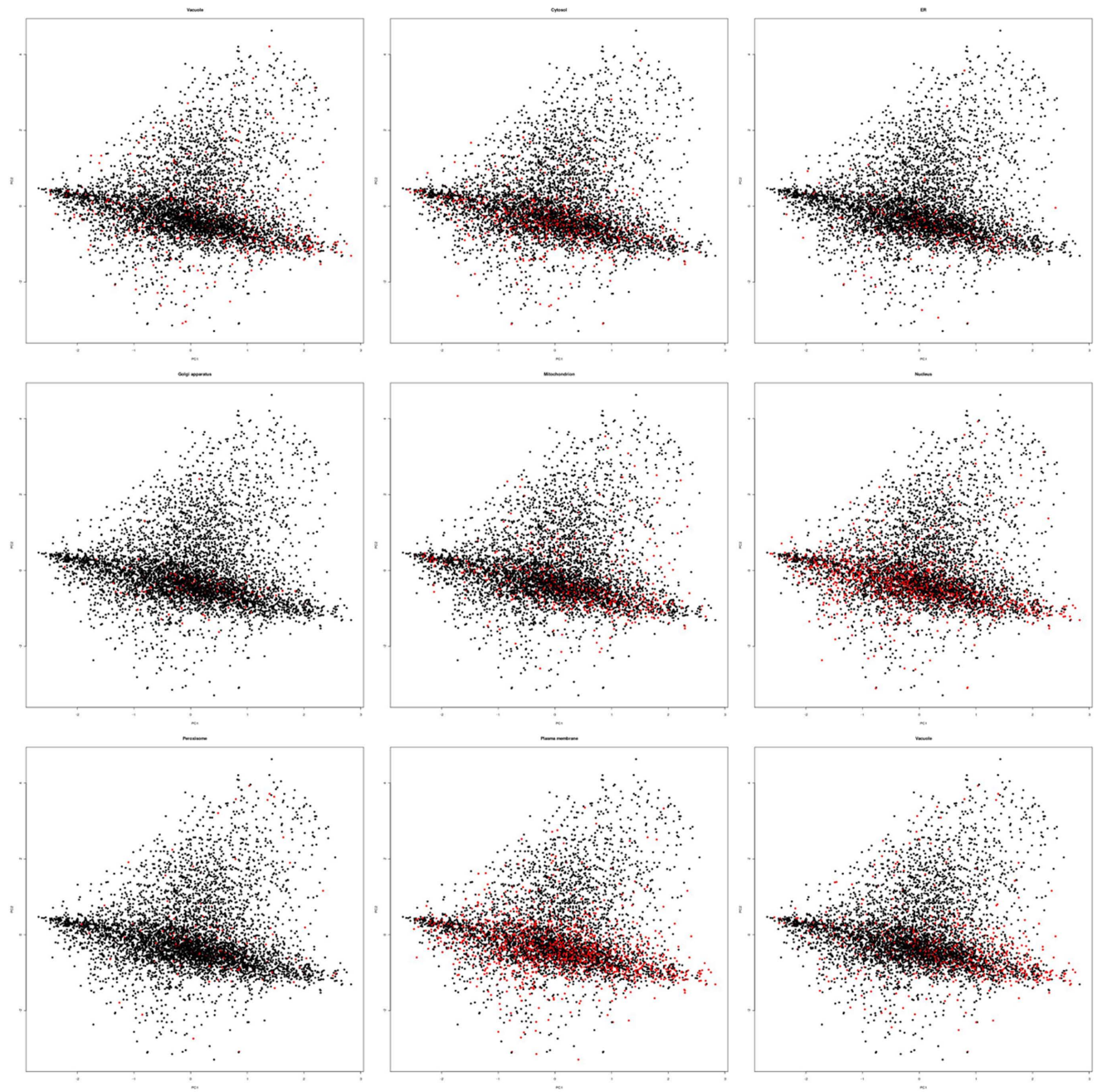


FIGURE A1 | Principle component analysis plots for all compartments. Exemplary principal component plots showing proteins belonging to a compartment in red and all other proteins in black. In each case 500 array slides were randomly sampled for the PCA plots and displayed exactly as in **Figure 1**.

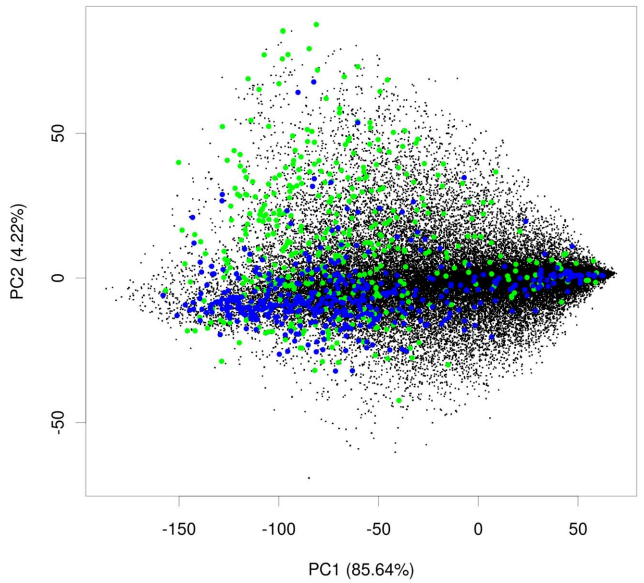


FIGURE A2 | PCA plot for rice. Principal component analysis plot showing plastid proteins in green, mitochondrial proteins in blue, and proteins from other compartments in black.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 22 September 2011
doi: 10.3389/fpls.2011.00055

Analysis of the compartmentalized metabolome – a validation of the non-aqueous fractionation technique

Sebastian Klie[1†], Stephan Krueger[2†], Leonard Krall[1], Patrick Giavalisco[1], Ulf-Ingo Flügge[2], Lothar Willmitzer[1] and Dirk Steinhauser[1*†]

[1] Department of Molecular Physiology, Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany
[2] Botanical Institute II, University of Cologne, Cologne, Germany

Edited by:
Alisdair Fernie, Max Planck Institute for Plant Physiology, Germany

Reviewed by:
Alisdair Fernie, Max Planck Institute for Plant Physiology, Germany
Lee Sweetlove, University of Oxford, UK

***Correspondence:**
Dirk Steinhauser, Department of Molecular Physiology, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany.
e-mail: steinhauser@mpimp-golm.mpg.de

[†] *Sebastian Klie, Stephan Krueger and Dirk Steinhauser have contributed equally to this work.*

With the development of high-throughput metabolic technologies, a plethora of primary and secondary compounds have been detected in the plant cell. However, there are still major gaps in our understanding of the plant metabolome. This is especially true with regards to the compartmental localization of these identified metabolites. Non-aqueous fractionation (NAF) is a powerful technique for the determination of subcellular metabolite distributions in eukaryotic cells, and it has become the method of choice to analyze the distribution of a large number of metabolites concurrently. However, the NAF technique produces a continuous gradient of metabolite distributions, not discrete assignments. Resolution of these distributions requires computational analyses based on marker molecules to resolve compartmental localizations. In this article we focus on expanding the computational analysis of data derived from NAF. Along with an experimental workflow, we describe the critical steps in NAF experiments and how computational approaches can aid in assessing the quality and robustness of the derived data. For this, we have developed and provide a new version (v1.2) of the *BestFit* command line tool for calculation and evaluation of subcellular metabolite distributions. Furthermore, using both simulated and experimental data we show the influence on estimated subcellular distributions by modulating important parameters, such as the number of fractions taken or which marker molecule is selected. Finally, we discuss caveats and benefits of NAF analysis in the context of the compartmentalized metabolome.

Keywords: subcellular metabolomics, analysis workflow, computational simulations, least squares algorithms, *BestFit* tool, visualization

INTRODUCTION

Although the main biochemical pathways in plants have been resolved by classical biochemical approaches in the last century (Fernie, 2007; Stitt et al., 2010a), many aspects of cellular metabolism and its regulatory functions are still not well understood, mostly due to technical limitations in gathering a more holistic insight into the cell's biochemistry. In recent years tremendous progress has been made in the establishment of high-throughput methods enabling the simultaneous analysis of a multitude of chemically diverse, small molecule metabolites from highly complex compound mixtures (Fiehn, 2001; Kopka et al., 2004; Brown et al., 2005; Pan and Raftery, 2007). Metabolomics, the comprehensive study of an organism's metabolite composition, has thus become an important tool in functional genomics and systems biology (Fernie et al., 2004; Saito and Matsuda, 2010). It has been widely used to study metabolic responses toward altered gene expression (Junker et al., 2006; Mugford et al., 2009; Albinsky et al., 2010), biotic and abiotic stresses (Kaplan et al., 2004; Bednarek et al., 2009), to characterize genetic and metabolic diversity (Schauer et al., 2006; Huege et al., 2011; Kusano et al., 2011), and has been combined with further Omic technologies in systems biology driven research (Kaplan et al., 2007; Hannah et al., 2010; Jozefczuk et al., 2010). While unexpected findings have yielded refined pathways as well as insights into their regulation and evolution (Zeeman et al., 2004; Eisenhut et al., 2008; Bednarek et al., 2009; Fettke et al., 2009), it has become evident that cellular metabolism needs to be considered as a highly integrative network bridging the genotype and ultimate phenotype or cellular responses (Meyer et al., 2007; Sweetlove et al., 2008; Sulpice et al., 2009; Stitt et al., 2010b). Even though the abovementioned studies provided major breakthroughs in the description of biological systems, we are still lacking information concerning the temporal and especially spatial regulation of the metabolome (Stitt and Fernie, 2003).

It is widely acknowledged that the compartmentalization of metabolism in eukaryotic cells represents a crucial factor for metabolic activity and functionality (Lunn, 2007). Consequently, the interrelation of metabolic networks within and between compartments needs to be deciphered. Whereas the subcellular localization of enzymes can be computationally predicted (Emanuelsson et al., 2000; Schwacke et al., 2003) or experimentally determined (Carter et al., 2004; Heazlewood et al., 2007; Taylor et al., 2011), the analysis of the subcellular localization of metabolites, the products and substrates of these enzymes, is more challenging due to redundant pathways, transport, and storage (Kruger and Von Schaewen, 2003; Büttner, 2007; Rébeillé et al., 2007; Krueger et al., 2010).

Further hurdles for reliable metabolite determinations in subcellular compartments are the fast turnover (Stitt et al., 1983; Stitt and Fernie, 2003) and the exceptionally rapid translocation of metabolites between compartments (Bowsher and Tobin, 2001; Martinoia et al., 2007; Weber and Fischer, 2007). Because of this, methods providing accurate information on the subcellular distributions of multiple metabolites are still limited.

Immunohistochemistry has been utilized to analyze the localization of non-protein molecules, such as cell wall polysaccharides and amino acids, permitting the analysis of metabolite compositions in compartments (e.g., Golgi, ER) which are normally not accessible by fractionation methods (Walker et al., 2001). However, dramatic losses of metabolites have been observed during tissue fixation which makes the interpretation of the results sometimes difficult (Peters and Ashley, 1967; Heinrich and Kuschki, 1978; Zechmann et al., 2011). Nuclear magnetic resonance (NMR) spectroscopy facilitates the determination of *in vivo* metabolite compositions in cells, tissues, and whole plants (Gout et al., 1993, 2000; Libourel et al., 2006). It requires distinct signals for the different compartments, which can be partly achieved by the pH dependency of the chemical shift of some molecules like inorganic phosphate, or organic and amino acids (Bligny and Douce, 2001). Compared to other spectroscopy/spectrometry methods, NMR is relatively insensitive and thus only feasible for metabolites which are highly abundant in the cell (Bligny and Douce, 2001). Genetically encoded molecular biosensors, proteins fused to two variants of the green fluorescent protein displaying conformational changes and fluorescence resonance energy transfer when a specific ligand binds, represent a promising molecular tool for temporal and spatial analyses of *in vivo* metabolite dynamics (Fehr et al., 2002; Lalonde et al., 2005; Chen et al., 2010). While this has been successfully applied for subcellular analysis of glucose and glutathione redox potentials (Deuschle et al., 2006; Gutscher et al., 2008), each targeted metabolite requires a unique sensor and therefore only a small number of metabolites might be simultaneously detectable in the same individual transgenic. Another widely used technique is protoplast fractionation. It is based on the fast purification of intact organelles through silicone oil or membrane filters followed by rapid quenching of metabolism (Wirtz et al., 1980; Lilley et al., 1982; Stitt et al., 1989). This method facilitates fractionation of plastids, mitochondria, and the cytosol from a single cell type, commonly mesophyll cells. However, digestion of the cell wall and the purification of protoplasts might substantially affect the metabolic state and therefore the obtained metabolite readout and the transferability of results.

Non-aqueous fractionation (NAF) is probably the most widely used technique to study metabolite compartmentalization, especially in plant science (Gerhardt and Heldt, 1984; Riens et al., 1991; Farre et al., 2001; Fettke et al., 2005; Krueger et al., 2009; Yamada et al., 2009). It separates fragments of subcellular compartments under non-aqueous conditions where biological activities, such as metabolite leakage, conversion, and translocation, are essentially completely arrested (Gerhardt and Heldt, 1984). Small subcellular particles, generated during lyophilization and ultrasonication of ground material, are separated by their composition-dependent density using equilibrium centrifugation in a gradient consisting of two differently dense, non-aqueous solvents (for details see Krueger et al., 2011). The abundance of metabolites and compartment-specific markers, which are also used as anchors to computationally estimate subcellular metabolite distributions, are analyzed throughout the collected gradient fractions. As non-aqueous fractionated material can be combined with a wide range of Omic technologies, it allows the determination of subcellular localizations for a large number of molecules including metabolites and lipids (Farre et al., 2001; Weise et al., 2004; Fettke et al., 2006; Krueger et al., 2011). In its routine application, the NAF technique allows for the separation of three distinct compartments – the cytosol, the plastids, and the vacuole (Riens et al., 1991; Farre et al., 2008; Krueger et al., 2009). However, it was recently shown that the resolution power of this technique has not yet been fully explored (Krueger et al., 2011).

As NAF results in continuous compartmental distributions due to variable and composition-dependent particle densities, computational methods need to be employed to analyze the obtained data. This and the interpretation of generated computational results reflect the main challenges for experimentalists. From the computational point of view as well, this type of data analysis is mostly underexplored.

Using both experimentally derived and simulated data, we investigated the effects of computationally modulating parameters important for the analysis of NAF gradients in order to address several technically and biologically relevant questions, such as: How many fractions are required to produce a good compartmental separation? Does the fraction number or the marker choice influence the estimated compartmental abundances? How good must the compartmental separation be in order to get reasonable estimates of compartmental abundances? How accurate must an estimate of compartmental abundances be in order to be considered valid? Taken together, the answers to these questions give a solid theoretical basis for the planning and execution of NAF experiments. Finally, we demonstrate and discuss alternative visualizations of NAF derived data in order to efficiently integrate additional knowledge to aid in the biological interpretation of the obtained results.

MATERIALS AND METHODS

The following sections introduce computational terminologies, used throughout the manuscript, denoted as *italicized* text along with their definition and/or abbreviation.

EXPERIMENTAL DATA

Experimental data and associated classifications used in this study were taken from Krueger et al. (2011). In brief, *Arabidopsis thaliana* leaf material was harvested 3 h after the onset of light and separated using an optimized NAF protocol (Krueger et al., 2009, 2011). A total of six fractions from three independent gradients were analyzed using mass-spectrometry (MS) – based metabolite profiling for primary and secondary metabolites as well as lipids in total comprising 3,921 mass spectrometric features. Three compartments, the plastids, the vacuole, and the cytosol were unambiguously delineated, each being represented by three compartment-specific markers. Although a clear trend was observed, the mitochondrial compartment was not considered to be unambiguously separated from the cytosol. However,

un-supervised clustering suggested the existence and contribution of yet unconsidered compartments (Krueger et al., 2011).

SIMULATED GRADIENTS

All simulation studies were performed using R 2.11.1 (R Development Core Team, 2010).

The distribution of a cellular constituent throughout a virtual gradient was simulated by 3,000 random deviates selected from a truncated normal distribution (Robert, 1995) in the interval between 0 and 30 units using the package "msm" (Jackson, 2010). Furthermore, 750 (25%) random deviates selected from a uniform distribution in the same interval were overlaid onto the random normal deviates to account for the fact that a cellular constituent is usually detectable throughout the entire gradient. The entire 3,750 random deviates were binned with a window width of 0.1 units, resulting in 300 bins. To place a simulated distribution at any bin position within the virtual gradient the mean parameter of the truncated normal distribution was changed from 0 to 30 in steps of 0.1 units. The SD, as the second parameter of the truncated normal distribution, was changed from 0.5 (as 0.0 is very unlikely to be achieved experimentally) to 30 in steps of 0.5 units to modify the degree of enrichment reflected by the amount of a cellular constituent observed at a certain gradient position. To simulate a non-enriched distribution, where the abundances throughout the gradient fractions are approximately equal, an SD of 35 and a mean of 15 (i.e., centrally positioned within the gradient) were used (*approximately uniform distribution*).

For the three-compartmental simulation model we assumed two compartments at the terminal positions at approximately 0 and 30 units of the gradient and a third, uniformly distributed compartment. For the four- and five-compartmental models, added compartments were positioned equidistant from each other, with exception of the uniformly distributed compartment, and from the terminal compartments (e.g., means of approximately 0, 10, 20, and 30 units in case of five compartments). In all simulations each compartment was represented by either 2, 3, or 5 compartment-specific marker distributions. To control the variation within compartments, the positions of markers reflecting the same compartment were varied (*marker spread, ms*) around the compartment center by shifting the means of their distribution by 0.1, 0.2 and further to 2.0 in steps of 0.2 units. Thus, for a compartment comprising a marker spread of 1.0, the gradient distance between the two most distant markers would be 2.0 units. For all simulations the aforementioned characteristic parameters (SD and marker spread) were changed identically for all compartmental distributions.

To simulate a systemic, technical, or experimental error on the abundances throughout the collected gradient fractions, we assume a uniform error model quantified as the normalized Manhattan distance (Eq. 2) between the initial (*error-free*) and modified (*error-containing*) distributions. The error was changed from 2 to 20% in steps of 2%.

DATA ANALYSIS

The abundances of cellular constituents throughout gradient fractions (*fraction abundances*) were expressed as percentages denoting the contribution of each fraction relative to the total amount (*scaled data*). Manhattan (Eq. 1) and Euclidean (Eq. 3) distances between the fraction abundances of cellular constituents were computed and normalized to fall within the range of 0–1 (*relative scale*; Eqs 2 and 4) and then multiplied by 100 to reflect percentages (*percentage scale*) (Krueger et al., 2011). A set of coordinates for each cellular constituent were derived by classical multidimensional scaling (*CMD*, Cox and Cox, 1994), such that the distances between the fraction abundances of those constituents are approximately equal to the normalized Euclidean distances. The within-compartment cohesion (*WCC*) was estimated as the average of all Manhattan distances between markers within the compartmental clusters. The between-compartment separation (*BCS*) was computed as the average of all Manhattan distances between markers of different compartmental clusters. Both parameters were computed using the package "fpc" (Hennig, 2010) on normalized Manhattan distances. *Silhouette information*, a combined measure of the WCC and BCS (Rousseeuw, 1987), was computed using the package "cluster" (Maechler et al., unpublished) and expressed as mean silhouette width for a clustering (*cluster solution*). Pearson's matrix correlation [also termed normalized gamma index (Halkidi et al., 2001) or non-parametric ANOVA using Mantel test (Sokal and Rohlf, 1995)] were computed between the initial distance matrix computed on fraction abundances and a binary (0, 1) matrix representing cluster assignments. Both cluster validity indices yield values in the interval of $[-1, 1]$, in which larger positive values reflect more favorable cluster solutions.

The percentage abundance of a cellular constituent in each of the resolved compartments (*compartmental abundances*), were computed on the basis of linear least squares methods with the *BestFit* (v1.1) command line tool using either the best fit (BFA, Riens et al., 1991) or non-negative least squares (NNLS, Lawson and Hanson, 1995) algorithms. The abundances of all markers delineating the same compartment were mean–averaged prior to computation (*compartmental center*). Due to run-time performance constraints, compartmental abundances on simulated data were estimated using NNLS while BFA was used for experimental data. The differences in compartmental abundances estimated using two different strategies, or on two different data sets for the same cellular constituent (*compartmental error*), were expressed as *maximum error* or *solution error*, i.e., only the maximum (identified on the absolute scale) or all observed differences among the considered compartments were taken into account. The 5th and 95th percentile of the observed differences are given, comprising the interval in which 90% of (non-extreme) differences lay.

The total percentage discrepancy (*TPD*, Krueger et al., 2011) was used as a quality measure for the estimated compartmental abundances derived from least squares solution (*LSS*). If not otherwise stated only LSS with a TPD $\leq 10\%$ were considered in comparisons to avoid bias in estimated parameters due to large discrepancies between two LSSs and to their respective fraction abundances. In some cases, thresholds to consider a LSS and thus the compartmental abundances as sufficiently explained were estimated as described in Krueger et al. (2011).

DATA VISUALIZATION

All figures were created with R 2.11.1 using the "graphics" (R Development Core Team, 2010) or "lattice" package (Sarkar, 2008).

Mean-difference (*MD*) plots were constructed to visualize the agreement of results derived from two different computational estimation strategies or on two different data sets. Generally, positive differences reflect larger estimates using strategy or data set A, while negative differences reflect larger values using strategy or data set B. The distribution of differences were visualized as box plots overlaid by violin plots to depict the data density. Level plots were generated to show the effects of two variables (x, y), represented as a two-dimensional grid, on a third variable (z) indicated by the coloring of every grid position. Contour lines were added to aid interpretation. As a convention throughout this manuscript for all constructed level plots, the values for the third variable z are smoothly colored using blue–yellow–red, where blue reflects a favorable measure while red reflects an unfavorable measure. Topological maps were created as scatter plots by depicting the first two principal coordinates (PCo's) derived from CMD analysis, which explain together about 98% of the total variance of the underlying distance matrices. Triangle (or ternary) plots were constructed to visualize the compartmental abundances for a three-compartmental estimation strategy using the "plotrix" package (Lemon, 2006).

EQUATIONS

Manhattan distance

$$d_M(x,y) = \sum_{i=1}^{n} |x_i - y_i| \tag{1}$$

Normalized Manhattan distance

$$d_m(x,y) = \frac{d_M(x,y)}{200} \tag{2}$$

Euclidean distance

$$d_E(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3}$$

Normalized Euclidean distance

$$d_e(x,y) = \frac{d_E(x,y)}{\sqrt{2 \cdot 100^2}} \tag{4}$$

RESULTS AND DISCUSSION

NON-AQUEOUS FRACTIONATION – DATA ANALYSIS WORKFLOW

The entire NAF procedure can be divided in experimental- and computational-driven analyses as illustrated in **Figure 1**. Although the main focus in this manuscript is targeted toward the computational analysis of NAF data, we include here, for completeness, a brief overview of the experimental analyses as well (for details see Krueger et al., 2011).

The experimental part encompasses the separation, discretization, and profiling of sample material (**Figure 1**). After sample processing, subcellular compartments are enriched at discrete positions within a continuous density gradient. The gradient can then be separated into a number of fractions of ideally equal volume for the subsequent determination of cellular constituents. The collected gradient fractions are analyzed with respect to both compartment-specific markers (to unambiguously designate a compartment) and cellular constituents (to estimate their compartmental abundances) either using targeted assays or high-throughput analytical technologies (Gerhardt and Heldt, 1984; Fettke et al., 2005; Benkeblia et al., 2007; Krueger et al., 2009, 2011).

The computational part comprises the validation, classification, visualization, and interpretation of the obtained data (**Figure 1**). As each step consists of various tasks achievable by using a multitude of computational approaches, we here only provide a short overview for comprehension.

Validation

While the fidelity of the obtained measurements first requires an evaluation, regardless the specific methodology employed, here validation focuses on the evaluation of the computed compartmental enrichments and separation. First, defined compartments must be delineated through the use of compartment-specific markers, which ideally represent compartments under investigation in an unambiguous manner. The enrichment of these markers is commonly depicted as bar plots (Riens et al., 1991; Winter et al., 1993; Farre et al., 2001) and, but less frequently, statistically supported by pairwise comparisons (e.g., Student's t-test) of the fraction abundances of markers (Krueger et al., 2009, 2011). However, while this shows the compartmental enrichment, it does not easily provide a parameter for the topological separation of all considered compartments. Normalized distances (Eqs 2 and 4) estimated on fraction abundances can be used to measure the separation between compartments designated using a single or multiple markers (Krueger et al., 2011). The subsequent use of clustering and associated cluster validation techniques (Halkidi et al., 2001), such as gap statistic (Tibshirani et al., 2001), or resampling approaches (Suzuki and Shimodaira, 2006), are powerful tools to statistically validate the cohesions within and separation between compartments, especially if multiples markers representing the same compartment are assayed (Krueger et al., 2011).

Classification

The main goal of this step is to estimate the compartmental abundance by computing the amount of cellular constituents in each of the previously defined compartments. While this can be achieved using simple linear regression for an individual compartment (Gerhardt and Heldt, 1984; Benkeblia et al., 2007), other linear least squares algorithms are more flexible for this purpose as estimates for all considered compartments can be computed simultaneously which also facilitates the assessment of the overall fit quality. These include the frequently used best fit algorithm (BFA, Riens et al., 1991), as well the non-negative least squares algorithm (NNLS, Lawson and Hanson, 1995). Both BFA and NNLS solve a system of linear equations defined by the marker-resolved compartments to determine the compartmental abundance by minimizing the discrepancy between the measured and fitted fraction abundances. Moreover, the iteratively (BFA), or by using the active-set method (NNLS), estimated compartmental abundances are constrained and thus restricted to yield always positive (and biological

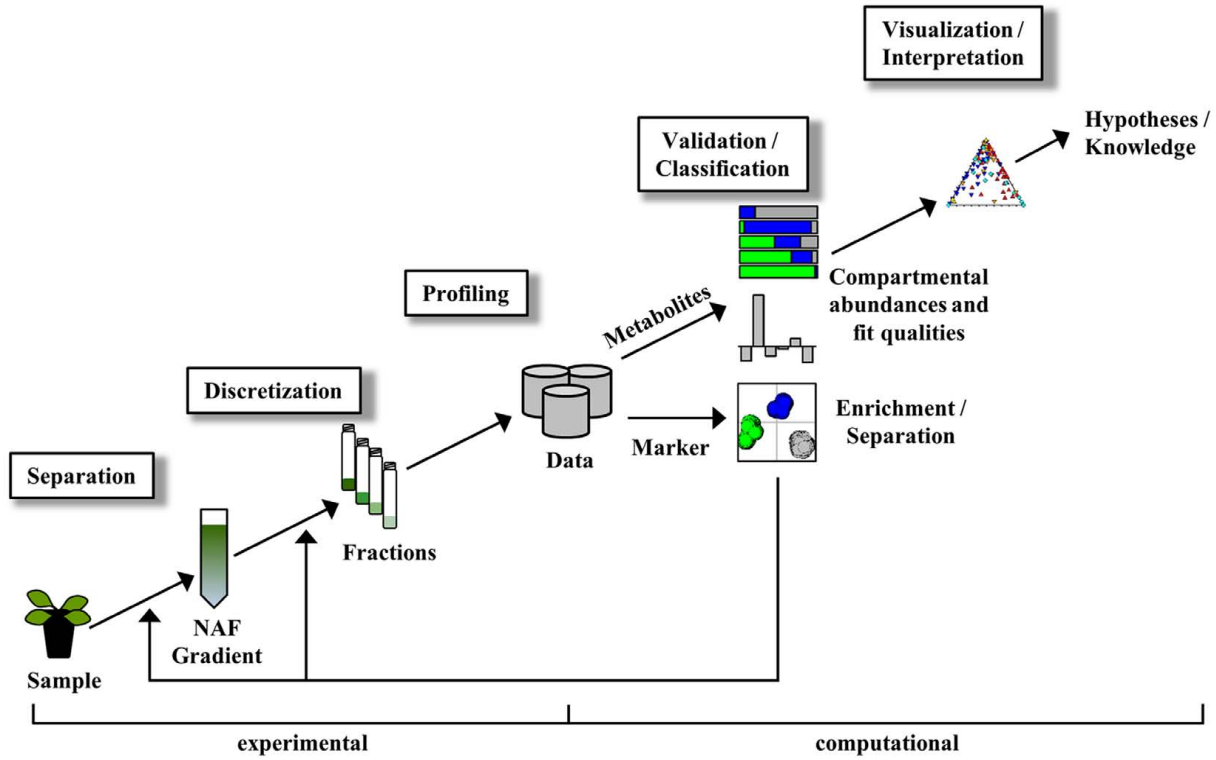


FIGURE 1 | Simplified scheme of non-aqueous fractionation procedure and associated data analysis steps. For details see manuscript text.

meaningful) estimates. However, the summed amount over all considered compartments does not need to equal 100% when using NNLS (Krueger et al., 2011). While estimates of compartmental abundances are computationally obtained, their qualities still need to be evaluated using the remaining associated discrepancy, commonly expressed as Euclidean distance or a derivate of it (Krueger et al., 2011).

Visualization and interpretation

A visual representation of the derived compartmentalization can be displayed through the combination of textual and graphical formats to aid in data interpretation. The choice of the visualization format depends on the questions addressed, the computational approach chosen, and the scientist's preferences. Typically the estimated compartmental distributions are provided in tabular format while the underlying fraction abundances are depicted as heat maps or are converted into cluster trees (Farre et al., 2001; Benkeblia et al., 2007; Krueger et al., 2009). A topological format may also be used to facilitate the inclusion of previous results or prior knowledge (Krueger et al., 2011). This can greatly aid in data interpretation, as the integration of enzyme localization or pathway properties can lead to new knowledge and facilitate hypotheses generation.

NON-AQUEOUS FRACTIONATION – A SIMULATION MODEL

Due to limited availability of experimental data, a simulation model was developed in order to study the influence of parameters on the compartmental separation and the estimation of compartmental abundances from NAF gradient data. To approximate the distribution of cellular constituents, such as compartment-specific markers or metabolites, we used random sampling from a truncated normal distribution, defined by the mean and SD within an interval ranging from 0 to 30 units. This choice was primarily motivated by four reasons: First, we used the abovementioned interval as a NAF gradient can comprise up to 30 mL (cf. Krueger et al., 2011) and therefore numerical parameters can be compared and interpreted in the context of existing experimental gradients. Secondly, by using the mean, simulated distributions can be easily placed at any position in the virtual gradient. Thirdly, by increasing the SD we can transform a cellular constituent from being highly enriched to being approximately equally distributed throughout the gradient. Finally, the effects produced through modifying the mean and SD are simple to understand as they are common parameters used to describe experimental results.

A model to illustrate the influence of distribution parameters on a virtual NAF gradient with four resolved compartments is shown as **Figure 2**, while a model for five compartments can be found as **Figure A1** in Appendix. In experimental data, the two terminal distributions would correspond to compartmental distributions observed for plastids and the vacuole in plant studies (Krueger et al., 2011), while the non-enriched compartment with approximately equal fraction abundances can be considered as "cytosol" (**Figure 2A**, **Figure A1A** in Appendix). Whereas the exact characteristics of the distributions are shown as line

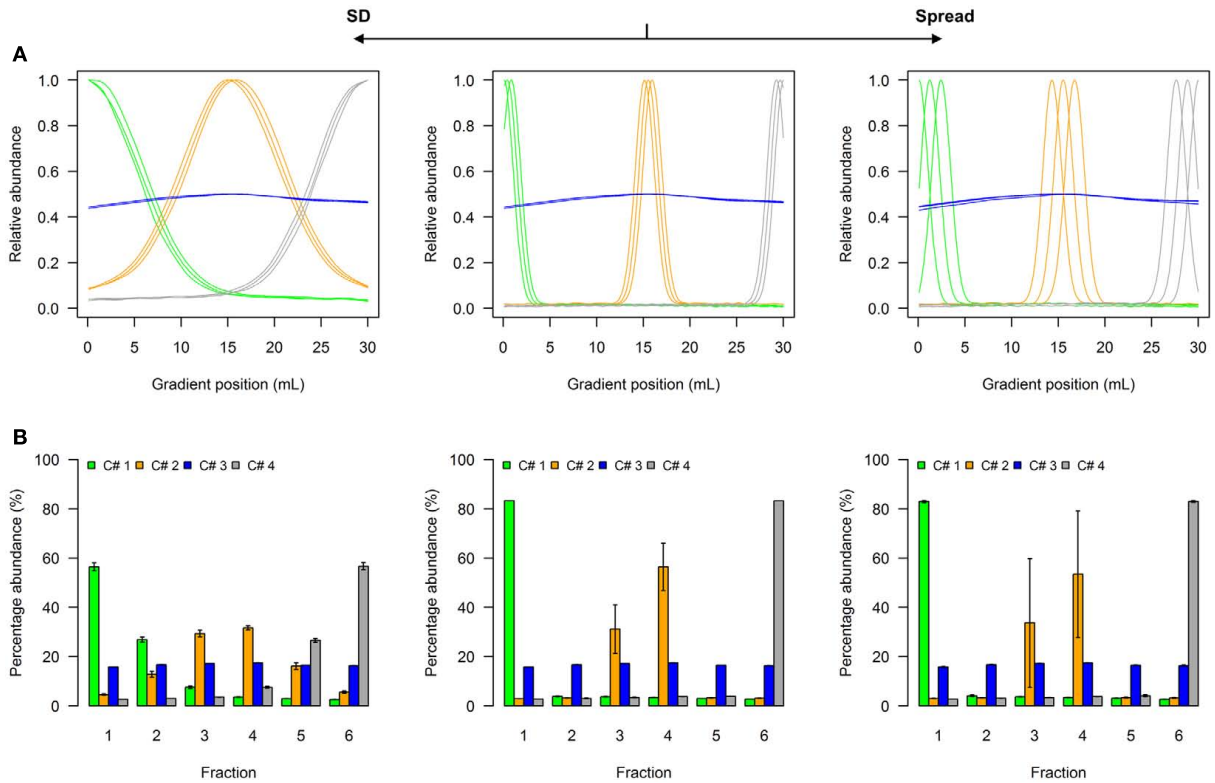


FIGURE 2 | Simplified scheme of the four-compartmental simulation model. (A) The continuous distributions are depicted by line plots for each of the 4 compartments represented by 3 individual marker distributions. To aid visualization the distributions are scaled to half-maximum (blue-colored compartment) or maximum (all other) observed values. **(B)** The bar plots show the mean–averaged fraction abundances including SD among compartment-specific markers for each compartment after the continuous distributions were discretized into 6 equally spaced fractions. The left–side graph illustrates the effect of increasing the SD (SD = 5, ms = 0.4), while the right-side graph shows the effect of increasing the marker spread ms (SD = 1, ms = 1.2) compared to a standard (middle graph with SD = 1, ms = 0.4).

plots (**Figure 2A**, **Figure A1A** in Appendix), experimentally only the discretized distributions can be assessed, i.e., the abundance of a cellular constituent throughout sampled gradient fractions (**Figure 2B**, **Figure A1B** in Appendix). When the SD is increased, the fraction abundances for the enriched compartments become closer to the non-enriched compartment (**Figure 2B**, **Figure A1B** in Appendix). Contrarily, increasing the marker spread by positioning markers representing the same compartment more distant from each other, the variation around the mean is increased (**Figure 2B**, **Figure A1B** in Appendix).

To show the behavior of the BCS and the WCC in dependence of changing SD and marker spreads, simulations were conducted individually for virtual NAF gradients containing either 3, 4, or 5 compartments each represented by 2, 3, or 5 markers, respectively. All combinations were tested using 2–30 fractions with 60 different SDs and 12 marker spread values each randomly repeated 99 times (see Materials and Methods). Both the BCS and the WCC were estimated on normalized Manhattan distances and visualized (**Figure 3**). **Figure 3A** clearly shows that with increasing SDs the BCS declines almost exponentially as the fraction abundances of compartmental markers become more similar to each other (**Figure 2B**, **Figure A1B** in Appendix), while the marker spread has minimal to no effect on the BCS. Contrarily, the WCC (**Figure 3B**) is influenced primarily by the marker spread, however, it is also affected by the compartmental separation modulated using the SD. Essentially the largest cohesion (i.e., the smallest distance) is observed at a high marker spread when the markers representing individual compartments are sharply focused at a specific gradient position (right bottom corner of **Figure 3B**).

HOW MANY FRACTIONS ARE NECESSARY TO PRODUCE A GOOD SEPARATION?

The answer is a trade-off between technical difficulties in generating highly reproducible gradients, the sampling of the same amount of liquid from gradients (fractionation), how much material is required for the down-stream analytical technologies to be used, and the analytical workload associated with taking more fractions. While this is truly a question of experimental constraints, from the computational point of view limits can be deduced.

For this we used the simulated data generated as described above. To quantitatively evaluate the clustering results we used the mean–average of the average Silhouette information and the Pearson's matrix correlation as the *NAF validity index*. **Figure 4** illustrates this index with respect to the BCS (modulated by the SD of the marker) and the WCC (modulated by the marker spread)

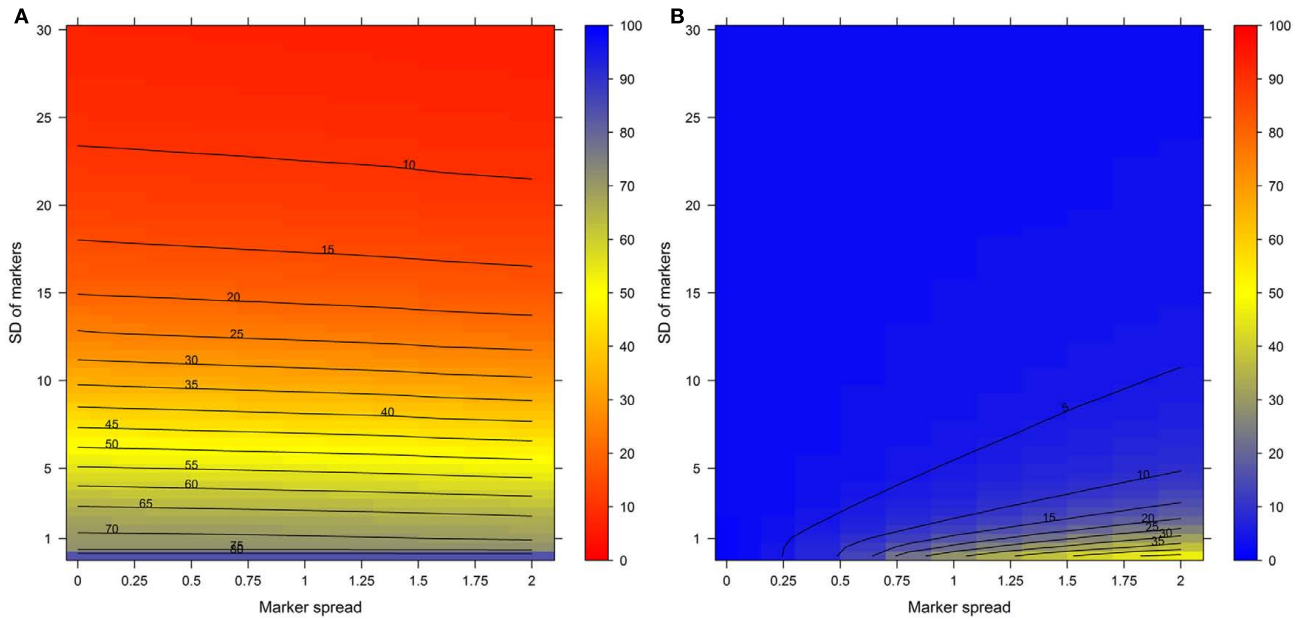


FIGURE 3 | (A) The between-compartment separation and **(B)** the within-compartment cohesion in dependence of the SD and marker spread. Both indices were estimated as mean–average of normalized, percentage-scaled Manhattan distances either **(A)** between or **(B)** within-compartmental clusters and are depicted as level plots color-coded according to the right-sided color bars. To aid interpretation, lines have been drawn at each 5% level. The closer the between-compartment separation is to 100% the better the separation while the closer the within-compartment cohesion is to 0 the tighter the compartmental clusters.

in dependence of the number of fractions collected. The figure was generated irrespective of the number of compartments (3–5) and markers (2, 3, or 5) per compartment considered by taking the mean value of the NAF validity index for different values of fractions and markers. For more detailed results see **Figures A2–A4** in Appendix depicting the individual results from simulations using 3, 4, or 5 compartments each represented by 3 independent compartmental markers.

While these figures first appear to be very complex, essentially the closer the values are to 1 (the bluer the color) the better the NAF validity index. Accordingly, we can conclude that the minimum number of fractions (n_F) necessary to be collected should equal the number of compartments (n_C) to be resolved, although similar cluster validities are also observed at $n_C - 1$ due to the inclusion of a non-enriched compartment (cf. **Figure 2**). The collection of a large number of small-volume fractions can result in the splitting of compartmental markers representing the same compartment, especially if markers are sharply focused (SD ≈ 1) and under increasing marker spreads. However, this can also happen when taking only a few fractions (see **Figure 4**, darker blue areas at SD ≈ 1 from left to right) but is more clearly evident when the number of fractions is increased. Finally, an increase in the number of fractions does not increase the cluster validity overall, i.e., there is no increase in the compartmental separation and thus the computationally estimated resolution of the gradient remains unchanged.

While this analysis only considers the compartmental separation, from the mathematical point of view there are constraints to the minimum number of fractions which should be taken. The least squares algorithms are usually applied to over-determined system of linear equations to yield an approximated solution, as otherwise an exact solution exists (Strang, 2009). In terms of the experimental system this means that there should be at least one more fraction used than the number of compartments which are to be determined, regardless if they are enriched or not. However, as it is desirable to apply additional constraints to the solution space (e.g., by restricting each variable to be positive and/or that all sum to 100%) even in cases where the number of fractions taken equals the number of compartments considered, a least squares approach, as opposed to an analytic solution, has to be employed as is the case for BFA or NNLS.

Therefore and in conclusion, the minimum number of fractions taken should ideally exceed the number of compartments. While the upper limit seems not to be definable, collecting a large number of fractions may led to compartment splitting which could be overcome by later combining and averaging the data from fractions known to be split. Nonetheless, it is rather important to determine the robustness of separation and compartmental boundaries in order to determine the experimentally optimal number of fractions and volumes to be collected. In some cases it may also be useful to take fractions of different volumes if the system under investigation is well defined or preliminary data with respect to compartmental boundaries is available.

DOES THE FRACTION NUMBER INFLUENCE THE COMPARTMENTAL ABUNDANCES?

To further address the influence of the number of fractions on the distribution of compartmental abundances we used the previously

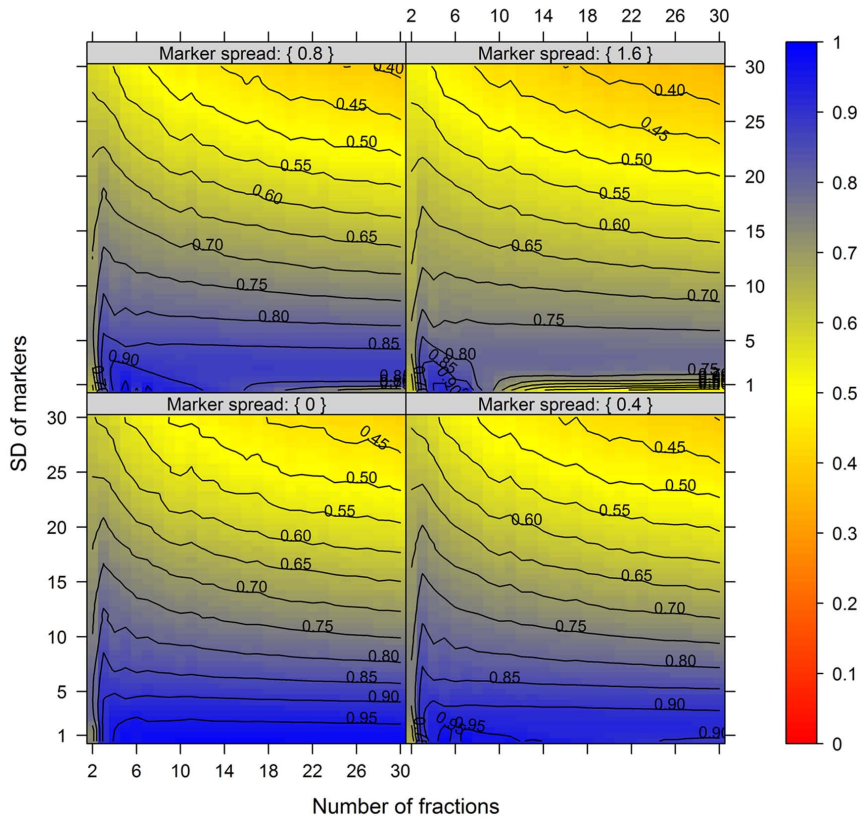


FIGURE 4 | The cluster validity in dependence of the number of collected fractions, the SD of markers, and the marker spread. While the SD modulates the between-compartmental separation, the marker spread modulates the within-compartmental cohesion (cf. **Figure 3**). The cluster validity index, estimated as mean–average of the Silhouette information and the Pearson's matrix correlation, is depicted independent of compartments (3–5) as well as markers (2, 3, or 5) per compartment considered. The closer the value is to 1 the better the observed cluster validity, color-coded as depicted in the right-side bar. To aid visualization negative cluster validity values were set to 0 and contour lines were drawn for each 0.05 unit. More detailed graphs can be found as **Figures A2–A4** in Appendix.

generated experimental data (Krueger et al., 2011). To modify the number of fractions we averaged either the terminal neighboring fractions $1+2$ (plastidic enriched) and $5+6$ (vacuolar enriched), resulting in 4 fractions (**Figure 5A**), or all neighboring fractions ($1+2$, $3+4$, and $5+6$) resulting in 3 fractions (**Figure A5A** in Appendix).

First, we evaluated the compartmental cluster solutions which revealed silhouette information of 0.63 ± 0.03, 0.68 ± 0.02, and 0.76 ± 0.01 and Pearson's matrix correlations of 0.74 ± 0.01, 0.75 ± 0.01, and 0.77 ± 0.02 for 6, 4, and 3 fractions, respectively (all as mean $\pm$ SD). Although an increase in both cluster validity indices were observed, indicating a better cohesion of and separation between compartments, the compartmental abundances of the individual markers were unchanged by reducing the fractions with, on average, 92.4 ± 8.9, 93.4 ± 7.5, and 93.3 ± 8.4% for 6, 4, and 3 fractions, respectively (all as mean $\pm$ SD) of the expected 100%. Comparing the differences of compartmental abundances for all metabolites based on 6 versus 4 or 3 fractions revealed rather narrow difference distributions (**Figure 5B**, **Figure A5B** in Appendix) independent of the observed mean values (**Figure 5C**, **Figure A5C** in Appendix) for both the maximum and the solution error. Using 4 fractions, 90% of all differences are within -8.7 to 10% and -6.7 to 7% for the maximum and solution error, respectively (**Figures 5B,C**). Similarly, using 3 fractions, 90% of all differences are in range from -8 to 13% and -8 to 9% for both the maximum and the solution error (**Figures A5B,C** in Appendix). While only minor effects were observed with respect to the compartmental abundances, reducing the fractions increased the number of sufficiently explained BFA estimates from 81.5% (TPD $\leq$ 10%), to 94.4% (TPD $\leq$ 9.4%) and 93.1% (TPD $\leq$ 7.4%) for 4 and 3 fractions, respectively. However, while this result is an enhancement, the reduced distribution space produced by shrinking the number of fractions effectively removes potential biologically meaningful intermediate distributions which are not delineated by the compartment-specific markers. Therefore, potential, yet-unresolved compartments can be overlooked. For instance, using 6 fractions, citrate synthase, a marker considered specific for the mitochondrial compartment (Stitt et al., 1982), has an insufficiently explained compartmental distribution (cf. Krueger et al., 2011). However, using 4 or 3 fractions, it is sufficiently explained with a 40% plastidic and 60% cytosolic distribution. Note that while the main isoform of citrate synthase is located in

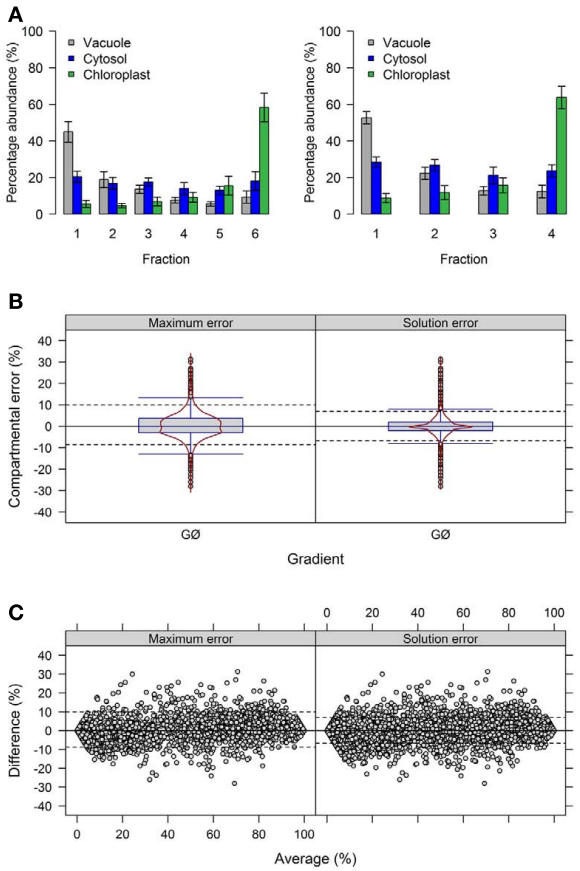


FIGURE 5 | Influence of the number of collected fraction on estimated compartmental abundances based on experimental data. (A) Bar plots depicting the mean–averages and SDs of the three compartment-specific markers for each of the three resolved compartments on the basis of 6 and 4 collected fractions. Neighboring terminal fractions, i.e., 1 + 2 and 5 + 6, were averaged to obtain results for 4 fractions. **(B)** Combined box- and violin plots (red lines) as well as **(C)** mean-difference plots showing the difference of the maximum and solution error based on compartmental abundances estimated using BFA for gradients comprising 6 and 4 fractions. For both, positive values indicate larger compartmental abundances for 6 fractions while negative values depict larger abundances for 4 fractions. The 5th and 95th percentiles are drawn as black dashed lines. The figures show the average of three independent gradients.

mitochondria, other isoforms are known to be present within the peroxisomes. These have been implicated in fatty acid respiration during seedling development and senescence (Pracharoenwattana et al., 2005; Kunz et al., 2009).

In conclusion, reducing the number of fractions only marginally influences the compartmental separation and the estimation of compartmental abundances. However, it does reduce the potential of detecting unknown, potentially fully resolved compartments. Therefore, increasing the number of fractions might enhance the detection of yet unassigned subcellular distribution and potential designation of a yet-unresolved or unconsidered compartment, especially if un-supervised "marker-free" approaches are employed (cf. Krueger et al., 2011).

DOES THE MARKER CHOICE INFLUENCE THE ESTIMATED COMPARTMENTAL ABUNDANCES?

Compartment-specific markers are central to NAF analyses as they anchor and establish the compartmental boundaries, and are ultimately used to estimate the compartmental abundances of the measured cellular constituents. Therefore the selection of certain cellular components as markers may influence the down-stream analysis and validity of the resulting data. Theoretically, the use of an unspecific marker (a marker that is shared between compartments or simply not localized to that compartment) would lead to erroneous conclusions. For example, although mannosidase enzyme activity has been widely used as a vacuolar marker in previously NAF studies (Gerhardt and Heldt, 1984; Riens et al., 1991; Winter et al., 1993; Farre et al., 2001; Benkeblia et al., 2007), recent experimental data from *Arabidopsis* showed activity in the Golgi (Strasser et al., 2006). This result questions the validity of using this marker specifically in *Arabidopsis*.

The use of multiple markers designating the same compartment can balance for a potential non-specificity of markers or their measurement errors. To test the importance of this factor we used the experimental NAF data where each of the resolved compartments (cytosol, plastid, and vacuole) is represented by three compartment-specific markers. First, we estimated how strong the compartmental abundance would be influenced by using jackknife approaches where we deleted a single marker, or used a marker twice. Although there are influences with respect to the cluster validity and BCS, for both the majority of observed values lay in a range of ±4% (**Figures A6A,B** in Appendix). Furthermore, all combinations of removing a marker or considering a marker twice do not largely influence the estimated compartmental abundances, compared to using all markers (**Figures A6C,D** in Appendix). Essentially, using one marker twice (**Figure A6D** in Appendix) led to very similar compartmental abundance, as 90% of all estimates were in range of −2.7 to 3% for the maximum error and −2.3 to 2.3% for the solution error. A similar, but slightly larger bias was observed with the omission of a marker with −5 to 5.7% for the maximum error and −4 to 4.3% for the solution error. Interestingly, the deletion or addition of a cytosolic marker influenced the estimates more strongly as compared to vacuolar or plastid-specific markers (**Figures A6C,D** in Appendix). This is in agreement with previous analyses, as it has been noted that the cytosolic compartment will cluster separately into three sub-clusters (Krueger et al., 2011). To further illustrate the effects of marker combinations and thus marker choice on compartmental abundances, we computed the compartmental abundances for each non-redundant marker combination. In contrast to the jackknife approaches, the distributions of differences are more heterogeneous. Here, 90% of the values over all marker combinations are in a range of −16 to 21.7% for the maximum error and −13.7% to 15.3% for the solution error (**Figure 6**). This illustrates that there is a clear influence on compartmental abundance estimation depending on the marker selected to represent a compartment. Therefore, by using multiple markers which ideally comprise the entire compartmental distribution space, the bias toward individual combinations can be reduced either by averaging prior to the estimation of compartmental abundances, or by averaging the

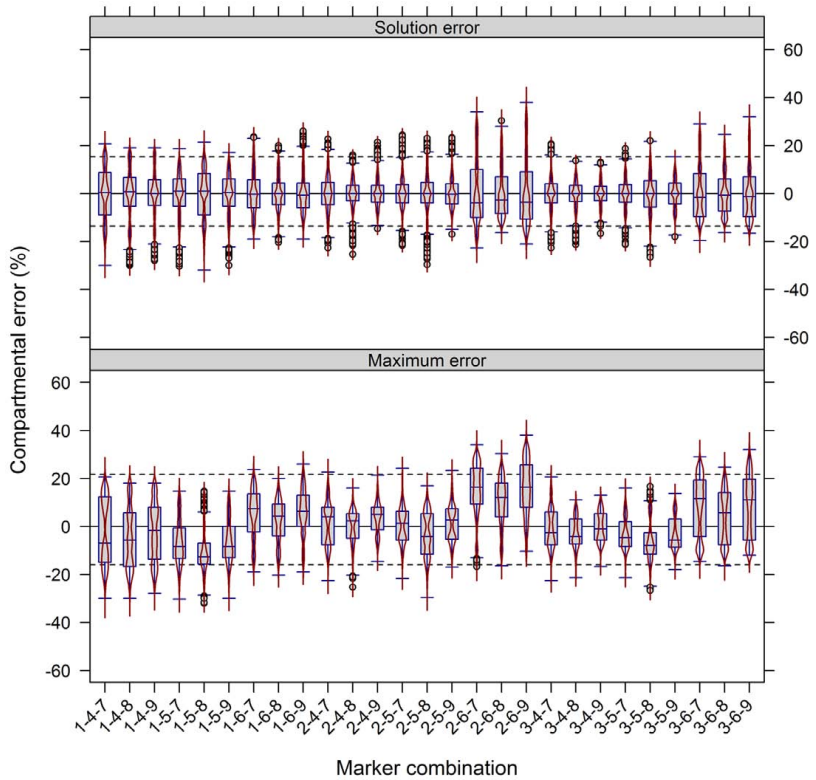


FIGURE 6 | Influence of compartment-specific marker combinations on estimated compartmental abundances based on experimental data. Combined box- and violin plots (red lines) showing the difference of the maximum and solution error based on compartmental abundances estimated using BFA as mean–average difference over the three independent gradients. While positive values indicate larger compartmental abundances when all nine markers were used, negative values depict larger abundances for a certain 3-marker combination. The 5th and 95th percentiles among all combinations are drawn as black dashed lines. Marker 1–3, 4–6, and 7–9 represent the plastids, the cytosol, and the vacuole, respectively. Plastids: 1 – glyceraldehyde-3-phosphate dehydrogenase (GAPDH), 2 – starch, 3 – Digalactosyldiacylglycerols (DGDGs); Cytosol: 4 – Uridine diphosphate – glucose-pyrophosphorylase (UGPase), 5 – glyceroceramids (GlcCer), 6 – triacylglycerides (TAGs); Vacuole: 7 – nitrate, 8 – glucosinolates, and 9 – flavonoids.

compartmental abundances estimated using all non-redundant combinations (cf. Krueger et al., 2011).

HOW GOOD MUST THE SEPARATION BE TO GET REASONABLE ESTIMATES OF COMPARTMENTAL ABUNDANCES?

In order to show the influence of the BCS and the metabolite error on compartmental abundance estimation, simulations were conducted individually for virtual NAF gradients containing either 3, 4, or 5 compartments each represented by two markers with 6, 9, or 12 considered fractions, respectively, and for all 60 SDs. The BCSs were estimated with normalized Manhattan distances (percentage scale) and binned in steps of 5% from 5 to 75. Differences in compartmental abundances were compared for each of the 300 gradient positions separately between the error-free and the error-containing distribution if the TPD did not exceed 15%. For all models and combinations of metabolite error and BCS bins, 200–1800 random estimates of compartmental abundances were considered.

Figure 7 illustrates the compartmental error with respect to the BCS and the cellular constituent error irrespective of the number of compartments (3–5) and fractions (6, 9, or 12) considered by taking the mean of the cellular constituent error for different values of the number of compartments and fractions. For more detailed results see **Figures A7** and **A8** in Appendix depicting the individual results of the compartmental errors for the compartments and fractions considered. As expected, the error in compartmental abundance estimation shows a clear dependence on both variables – the BCS and cellular constituent error (diagonal contours). Generally, the error of compartmental abundances increases rapidly with small increases of the cellular constituent error and small decreases of the BCS (**Figures 7A,B**). When considering the mean (**Figure 7A**) of all errors of compartmental abundances obtained for each binned value of BCS and cellular constituent error, both variables seem to contribute almost equally. However when considering the magnitude in the error of compartmental abundance, here derived by taking the 99th percentile (**Figure 7B**) of all computed values, a stronger influence of the cellular constituent error can be seen, illustrated as the much larger area of high errors of compartmental abundances, indicated by the dark red shading. This shows that the risk of obtaining high errors of compartmental abundances is more likely when the individual error of a cellular constituent is high, rather than with a low BCS. For experimentally

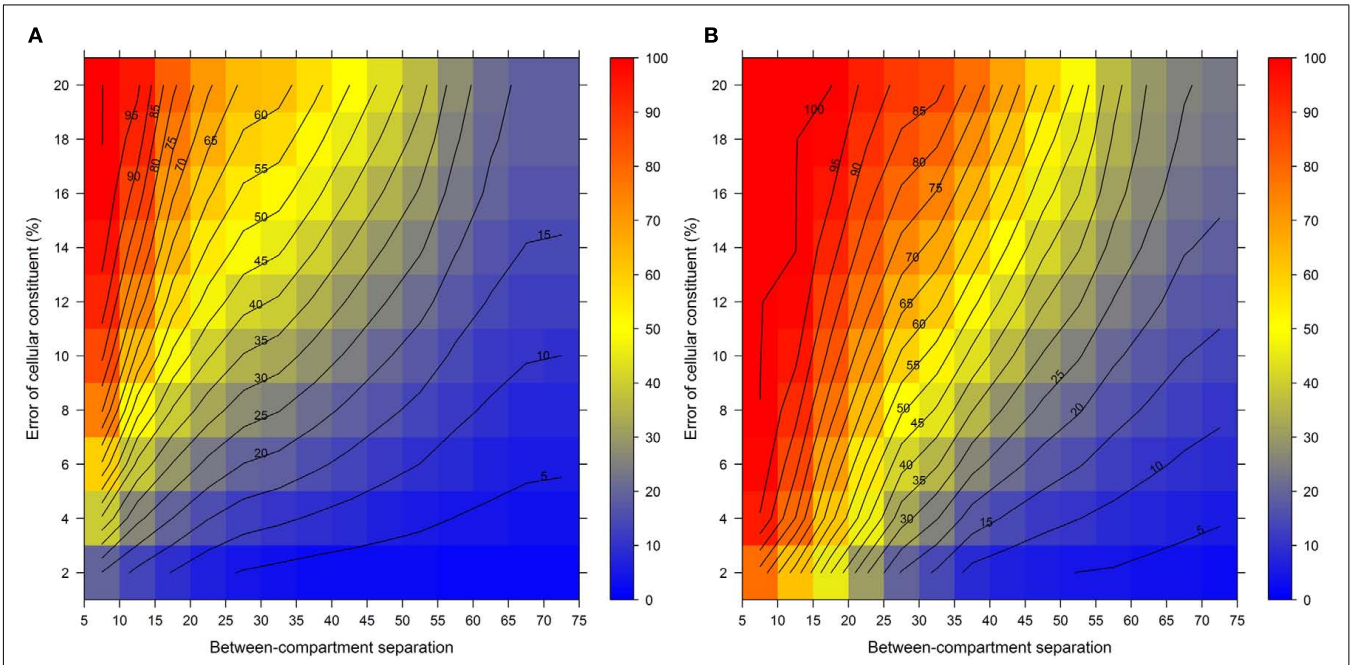


FIGURE 7 | Calculation error of compartmental abundances in dependence of the between-compartment separation and the metabolite error based on simulated data. The estimated error, the discrepancies to the estimates of the error-free distributions, are depicted as **(A)** mean–average and **(B)** 99th percentile for each combination of between-compartment separation and metabolite error independent of the number of compartments (3–5) and fractions (6, 9, and 12) considered. The closer the value is to 0% the smaller the maximum error of the compartmental abundances, color-coded as depicted in the right-side bar. Contour lines have been drawn for each 5% error. More detailed graphs can be found as **Figures A7** and **A8** in Appendix.

obtained data, this would mean that the individual measurement accuracy of fraction abundances and the associated measurement error influence the compartmental abundance estimation stronger than the overall compartmental separation.

To validate this, we performed a similar analysis on the experimental data (**Figure 8**). Similarly, the error in compartmental abundances increased almost exponentially with the increasing metabolite error, irrespective of either the maximum or the solution error. At the 10% metabolite error level the majority of differences in estimated compartmental abundances revealed a ∼10% compartmental error. When the metabolite error was increased further, the compartmental error rose rapidly to the point where the majority of absolute values reside in the range of up to 30%.

In conclusion, both simulated and experimental data revealed a larger effect of the measurement error of a cellular constituent on estimated compartmental abundances compared to the overall separation of compartments. To balance for these effects, the number of technical (ideally biological) replicates could be increased to obtain more robust compartmental estimates, thus increasing the confidence of the estimated subcellular metabolite distributions.

HOW ACCURATE MUST A LEAST SQUARES SOLUTION BE TO BE CONSIDERED STATISTICALLY VALID?

The quality or the "goodness of fit" of any statistical model describes how well a set of observed data points fit to the estimated values the model returns. In our case this corresponds to how well the measured fraction abundances match the fitted ones determined by BFA or NNLS. Classically, the fit quality is quantified as the distance between the measured and estimated model data. Here, the residual sum of squares (RSS) or the Euclidean distance (the square-rooted RSS) is used where a small, closer to zero value indicates a "good fit." However, since both measures are unscaled (unadjusted), it impedes interpretation of the fit quality. While TPD employs a normalized Manhattan distance to map distances on a percentage scale (Krueger et al., 2011), here we additionally suggest the use of a normalization of the Euclidean distance (Eq. 4), which is very closely related to the distance parameter a least squares approach tries to minimize.

Conceptually, one wants to use the discrepancy to decide whether a cellular constituent can be partitioned with confidence into the delineated subcellular compartments. The reason for investigating the fit quality by any of the mentioned measures is because a cellular constituent can display, compared to the fraction abundances of the marker, intermediate fraction abundances or a unique pattern that does not coincide with any considered marker/compartmental pattern. Both BFA and NNLS try to derive a predictive model assigning this cellular constituent into one of the defined subcellular compartments by using the observed marker distributions. Both algorithms will therefore result in a relatively high discrepancy and thus an incorrect classification of the cellular constituent with respect to its abundances in the considered compartments. However, in order to detect such cases,

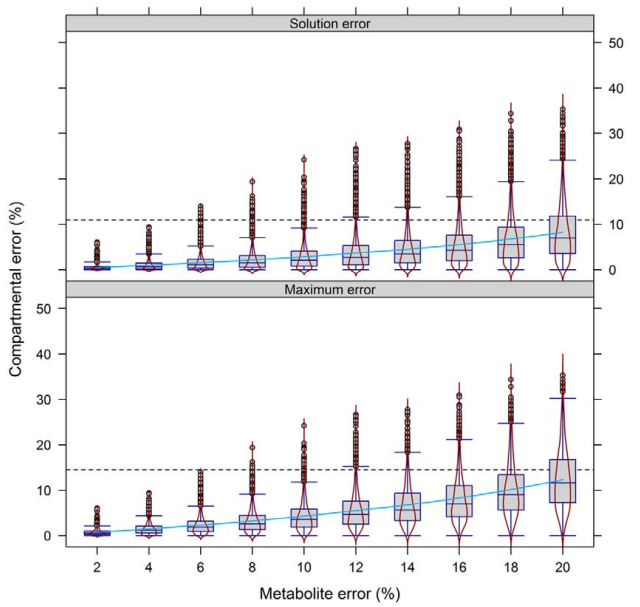


FIGURE 8 | Calculation error of compartmental abundances in dependence of the metabolite error based on experimental data. Compartmental abundances are estimated using BFA and visualized as absolute mean–average difference over the three independent gradients. The 95th percentile among all metabolite error levels is drawn as black dashed lines. The light-blue solid line shows a fitted smooth curve to illustrate the increase of the estimated compartmental error with increasing metabolite error.

one needs to qualify or statistically quantify such a relatively high discrepancy.

While this could be achieved by arbitrarily defining a threshold to consider discrepancies as being acceptable, the threshold could also be adaptively inferred by employing a topological measure. For this purpose we use the degree of cohesion, defined by the spatial distance of the markers to their respective compartmental centers, either between independent gradients (Krueger et al., 2011) or within a gradient (applicable when multiple markers are assayed). We assume that the topological space a cellular constituent occupies cannot be larger than the one observed for the resolved compartments. Therefore, if the observed discrepancy for a cellular constituent exceeds the topological space of the compartments, the marker does not encompass its compartmentalization either due to potential transport processes or due to an unconsidered compartment. By assuming normality of the cohesion, one can express the divergence of an observed fit-discrepancy to this topological measure by standard (z-) scores, indicating how many SD a particular fit-discrepancy differs from the mean of the distances as defined by the compartmental cohesion. While negative values reflect discrepancies below the mean compartmental cohesion, positive values show discrepancies above the mean. Also, by employing a standard normal distribution function (e.g., pnorm in R), one can devise a right-tailed test to obtain p-values to further assess the statistical significance of the measured distributions.

BESTFIT – FURTHER DEVELOPMENTS AND CURRENT IMPLEMENTATION

As previously mentioned (Krueger et al., 2011), for fast computation on large data sets both least squares algorithms were implemented (BFA) or compiled (NNLS; Fortran 77 routine from R's "nnls" package; Katharine and Van Stokkum, 2010) into the *BestFit* C-language command line tool (v1.1; Steinhauser et al., unpublished). In order to further enhance the calculation and evaluation of subcellular metabolite distributions from NAF data we have restructured and added further statistical analyses routines. In the current version (v1.2), *BestFit* supports the automatic calculation of compartmental cluster statistics, such as BCS, WCC, silhouette information, z-score estimation, and Pearson's matrix correlation, based on both normalized Euclidean as well as Manhattan distances. Using the $-A$ option (if multiple markers designating the same compartment are assayed) the user can control if the compartmental center or all marker combinations should be used to compute compartmental abundances for cellular constituents. Per default ($-M$ option), markers are included in this analysis, i.e., treated as cellular constituents. To evaluate the fit quality the user can specify the cutoff ($-T$ option; default to "max") adaptively estimated either using the distance to the compartmental center (default) or the WCC using the $-W$ option.

We observed when using NNLS that the sum of compartmental abundances for a solution equals the sum of fitted fraction abundances, even though the compartmental abundances do not need to sum to 100%. Interestingly, both the sums of fitted fractions and the compartmental abundances are perfectly correlated with a coefficient of determination of $R^2 = 1$ (data not shown). Rescaling of the NNLS fitted fraction abundances followed by re-calculation using NNLS bound the sum of estimated compartmental abundances to 100% (termed *NNLSs*). To compare the difference in compartmental abundances estimated using BFA and NNLS or NNLSs we computed the LSS using a three- or four-compartmental model (the mitochondrial compartment was for this purpose considered as being unambiguously resolved) on the experimental data (**Figure 9** and **Figure A9** in Appendix). When 3 compartments are considered, 90% of all differences (TPD $\leq$ 10%) are within -1 to 0.9% and -0.8 to 0.8% for NNLSs while reveling for NNLS a larger spread, namely from -4.7 to 2.6% and -1.4 to 2%, for the maximum and solution error, respectively, compared to BFA (**Figure 9**). Similarly, using 4 compartments, 90% of all differences fall in the range from -0.8 to 1.6% and -0.7 to 0.7% for both the maximum and the solution error (**Figure A9** in Appendix). As BFA uses a 1% interval to iteratively compute compartmental abundances, a large fraction of the observed differences fall within this range of $\pm$ 1% or can be the result of error propagation. Compared to BFA, which is limited to 5 compartments due to run-time constrains, NNLS is applicable to more than 5 compartments and is guaranteed to find the optimal solution that satisfies the conditions (non-negative solution). Also, using NNLSs the compartmental estimates can be scaled. This can be advantageous for some visualization formats (see below). Although there might be more sophisticated algorithms for constrained-based LSS to obtain non-negative values that sum up to 100%, we find it useful to implement NNLSs, an iterative NNLS, where the user can

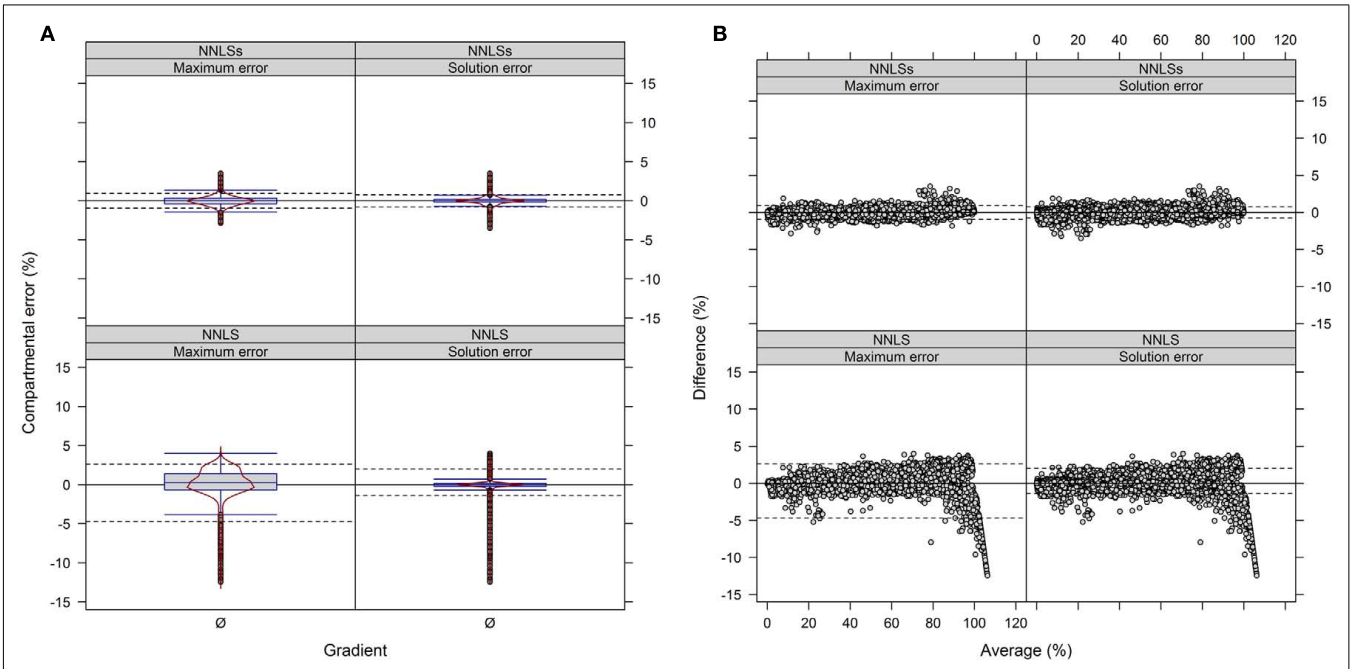


FIGURE 9 | Diagnostic plots showing the differences in estimated compartmental abundances using (A,B) a three-compartmental estimation strategy based on experimental data. (A) The distributions of observed differences in compartmental abundances are shown as box- and violin plots (red lines). **(B)** The mean-difference plots depict the differences in dependence of the averages of compartmental abundances. For all graphs the difference in compartmental abundances between BFA and NNLS or iterative (seed-based) NNLS (NNLSs) are visualized based on the mean–average of compartmental abundances for the three independent gradients. The 5th and 95th percentiles are drawn as black dashed lines. Only values are shown if they are considered as sufficiently explained, i.e., TPD $\leq$ 10% based on the BFA solutions.

decide to choose this using $-I$ option (default is set to 1 iteration, i.e., NNLS).

BestFit (v1.2) is available from CSB.DB website (Steinhauser et al., 2004) at http://csbdb.mpimp-golm.mpg.de/bestfit.html.

VISUALIZATION-AIDED INTERPRETATION OF DATA

While metabolic data has traditionally been visualized as cluster trees and their associated heat maps, we have attempted to focus on alternative types of visualization which can be easily overlaid or integrated with additional knowledge in order to achieve a more holistic overview of the data produced from NAF.

The use of PCo space, based on normalized Euclidean distances, is an excellent method to show the localization of the markers and associated metabolite classes or markers and compartment assignments through a visually appealing and easily interpretable figure (**Figure 10**). Here, the spatial spread of the markers clearly illustrates the heterogeneity of the considered classes across the entire space for metabolites from primary metabolism (**Figure 10A**), or the enrichment of the metabolite classes from secondary metabolism associated with specific compartments, such as the galactolipids in the chloroplast, the flavonoids or glucosinolates in the vacuole, or the triacylglycerides in the cytosol (**Figure 10B**). Using PCo space, the specificity for the metabolites assigned into a certain compartment or even between the compartments (for details on assignments see Krueger et al., 2011) can also be easily visualized (**Figures 10C,D**). Theoretically there is no limit to the number of compartments which may thus be shown. As this method greatly reduces the complexity of the data, the aid in biological interpretations is greatly increased.

However, the data simplification for visualization using PCo space does not show the absolute compartmental enrichment or to what extent metabolites or classes of metabolites are shared between the analyzed compartments. Therefore the use of triangle plots is another useful way to present NAF derived data (**Figure 11**). They present the percent distribution of a certain metabolite, or group of metabolites shared between the different compartments, in an easily interpretable figure. In essence, this is the graphical equivalent of a tabulation of the data, as the estimated fraction amounts can be directly determined from the figure.

For example, it is more obvious in a triangle plot that amines are closer associated to the chloroplast and cytosol than to the vacuole, that carbohydrates are more closely associated with the cytosol and vacuole, and that organic acids are more associated to the cytosol (**Figure 11A**). Just by eye, additional important information can be extracted. For example, only very few metabolites are located in the clp|cyt|vac space within the triangle plot, indicating that only few metabolites are equally shared between all three compartments (**Figures 11A–D**) and only a minor amount of metabolites show a distinct enrichment within the vacuole and the chloroplast, without being present in the cytosol. This is even more pronounced when depicted for all metabolites (**Figure 11D**). This is most likely due to the cytosol being the transit route between the other two compartments. One caveat for use of the triangle

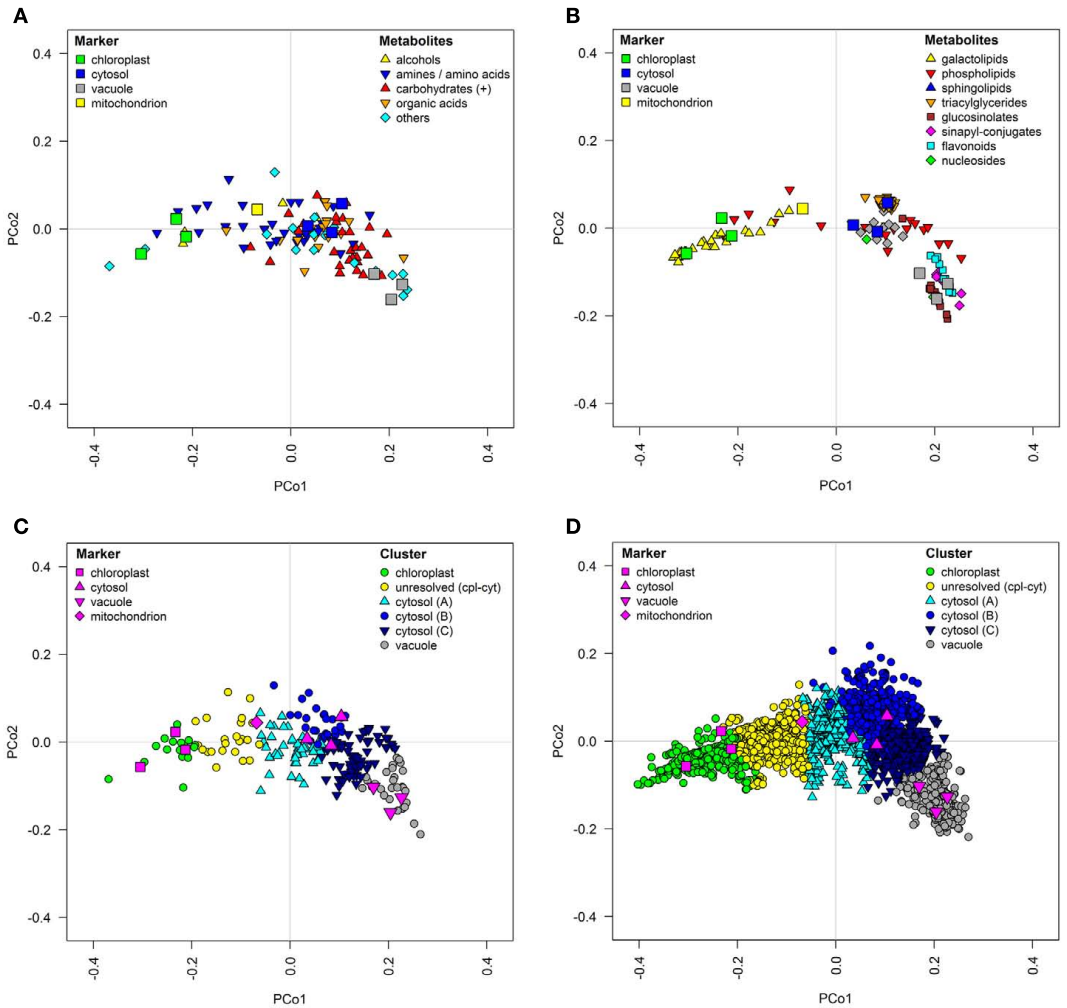


FIGURE 10 | Topological maps of (A,C) primary metabolites as well as (B,D) secondary metabolites and lipids based on experimental data in proximity to resolved compartments. All graphs depict classification results visualized in principal coordinates (PCo) space on the basis of averaged normalized Euclidean distances among metabolites for the three independent gradients. **(A,B)** Chemical superclass assignments were overlaid on structurally identified metabolites and **(C,D)** k-medoids cluster assignments (cf. Krueger et al., 2011) were overlaid on all metabolites.

plot as a visualization tool is that the sum of the compartmental abundances must total 100%. Furthermore, it is only feasible for a three-compartmental separation.

CAVEATS AND BENEFITS OF NON-AQUEOUS FRACTIONATION

Although several different approaches exist to study metabolite composition on the subcellular level, none can be referred to as "the method of choice" as every method has specific advantages and disadvantages. The best method to use depends on the experimental question.

First, the caveats: NAF is a generally labor intensive process and requires technical precision to produce consistent gradients. Secondly, analysis of the data from NAF gradients is critically dependent on the use of compartment-specific markers. The more markers used to define the compartmental space, and the more specific the markers are for a compartment, regardless of their biochemical nature, the better the resulting designation of the compartments. Until suitable markers are determined for the mitochondria or other unconsidered compartments, such as the peroxisome, or organelle sub-compartments, these structures must be considered unresolved. Finally, while the absolute purity of the isolated compartments in NAF gradients is not as tight as seen with protoplast fractionation or the perfusion technique, using statistical tests (as we have shown in this manuscript), high confidence data can clearly be produced from NAF gradients.

As for the benefits, the main one is that metabolism is effectively stopped immediately after harvesting. This prevents metabolite conversion or translocation, unlike protoplast fractionation (Robinson and Walker, 1979; Wirtz et al., 1980; Lilley et al., 1982; Stitt et al., 1989) and intracellular perfusion techniques (Takeshige and Tazawa, 1989).

Non-aqueous fractionation also produces an enrichment of the compartmental constituents, allowing for the potential

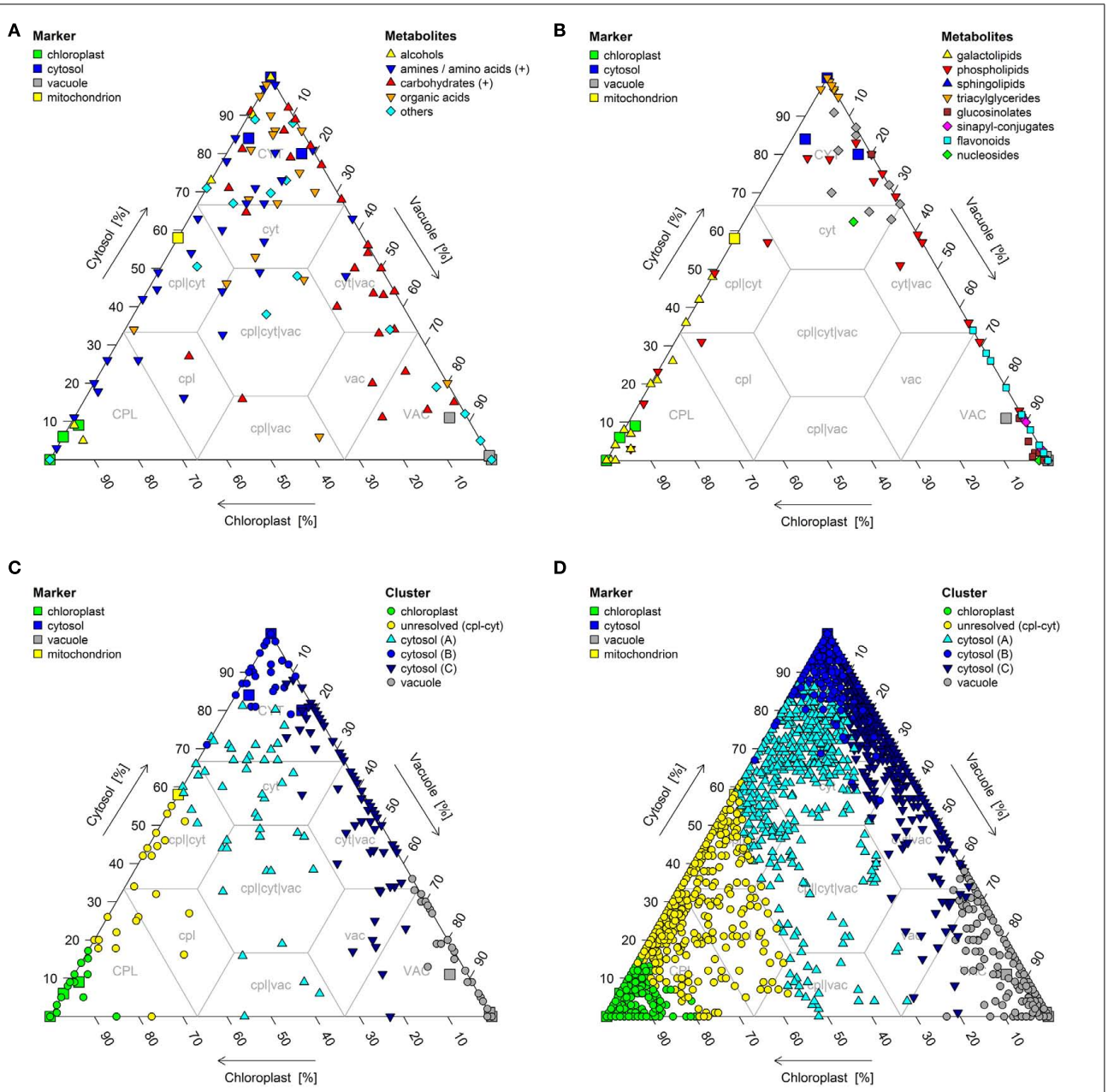


FIGURE 11 | Triangle plots of (A,C) primary metabolites as well as (B,D) secondary metabolites and lipids based on estimated compartmental abundances for the three resolved compartments – plastids, the cytosol, and the vacuole. (A,B) Chemical superclass assignments were overlaid on structurally identified metabolites and **(C,D)** k-medoids cluster assignments (cf. Krueger et al., 2011) were overlaid on all metabolites.

detection of low-abundant compounds, such as hormones. As well, the relatively large amount of material used permits multiple down-stream analysis techniques to be applied, based on the number and volume of fractions taken. While we have routinely used GC/and LC/MS – based metabolomic approaches, this can easily be expanded toward the measurements of enzyme activities (Gibon et al., 2004; Steinhauser et al., 2010), proteomic-based technologies (Giavalisco et al., 2006), or to NMR (Weise et al., 2004). In the current age of systems biology the combination of comprehensive Omics technologies with classical NAF and modern computational biology approaches can dramatically increases the knowledge about the spatial and also temporal changes of metabolism on the subcellular level.

As NAF has been generally applied to whole organs or tissues, there has also been a concern of the contribution of the different cell types to the detected metabolite pools. As the *Arabidopsis* leaf is composed mainly of mesophyll cells, it can be assumed that these cells are the major contributor to the observed metabolite pool sizes. However, as shown previously, with a comprehensive enough or even specific analysis metabolites known to be

spatially separated in different cell types can be localized to their experimentally proven compartments (Krueger et al., 2011).

Interestingly, because NAF separates not only intact organelles but also fragments of organelles, it might be also possible that identification of sub-organelle compartments may be achievable, such as the thylakoids from the stroma in chloroplasts, or dissecting the sub-compartments present in the plant vacuole (Paris et al., 1996), however this would require specific markers to delineate these compartments. For example, using NAF Riewe et al. (2008) could demonstrate that the apoplast of potato tuber is similarly, but not identically distributed as the vacuole in potato tubers. For *Arabidopsis* leaves, unassigned subcellular compartments could thus far only be identified by metabolite distributions that could not be explained by the three compartment-specific markers, strongly indicating the presence of additional subcellular compartments (Krueger et al., 2011).

ACKNOWLEDGMENTS

This work was supported by the Max Planck Society and the University of Cologne. We thank the R Foundation for Statistical Computing and the R community for the continued development of the free software environment for statistical computing and graphics R. We would like to thank Dr. Björn Usadel for critical reading of this manuscript.

REFERENCES

Albinsky, D., Kusano, M., Higuchi, M., Hayashi, N., Kobayashi, M., Fukushima, A., Mori, M., Ichikawa, T., Matsui, K., Kuroda, H., Horii, Y., Tsumoto, Y., Sakakibara, H., Hirochika, H., Matsui, M., and Saito, K. (2010). Metabolomic screening applied to rice FOX *Arabidopsis* lines leads to the identification of a gene-changing nitrogen metabolism. *Mol. Plant* 3, 125–142.

Bednarek, P., Pislewska-Bednarek, M., Svatos, A., Schneider, B., Doubsky, J., Mansurova, M., Humphry, M., Consonni, C., Panstruga, R., Sanchez-Vallet, A., Molina, A., and Schulze-Lefert, P. (2009). A glucosinolate metabolism pathway in living plant cells mediates broad-spectrum antifungal defense. *Science* 323, 101–106.

Benkeblia, N., Shinano, T., and Osaki, M. (2007). Metabolite profiling and assessment of metabolome compartmentation of soybean leaves using non-aqueous fractionation and GC-MS analysis. *Metabolomics* 3, 297–305.

Bligny, R., and Douce, R. (2001). NMR and plant metabolism. *Curr. Opin. Plant Biol.* 4, 191–196.

Bowsher, C. G., and Tobin, A. K. (2001). Compartmentation of metabolism within mitochondria and plastids. *J. Exp. Bot.* 52, 513–527.

Brown, S. C., Kruppa, G., and Dasseux, J. L. (2005). Metabolomics applications of FT-ICR mass spectrometry. *Mass Spectrom. Rev.* 24, 223–231.

Büttner, M. (2007). The monosaccharide transporter(-like) gene family in *Arabidopsis*. *FEBS Lett.* 581, 2318–2324.

Carter, C., Pan, S., Zouhar, J., Avila, E. L., Girke, T., and Raikhel, N. V. (2004). The vegetative vacuole proteome of *Arabidopsis thaliana* reveals predicted and unexpected proteins. *Plant Cell* 16, 3285–3303.

Chen, L. Q., Hou, B. H., Lalonde, S., Takanaga, H., Hartung, M. L., Qu, X. Q., Guo, W. J., Kim, J. G., Underwood, W., Chaudhuri, B., Chermak, D., Antony, G., White, F. F., Somerville, S. C., Mudgett, M. B., and Frommer, W. B. (2010). Sugar transporters for intercellular exchange and nutrition of pathogens. *Nature* 468, 527–532.

Cox, T. F., and Cox, M. A. A. (1994). *Multidimensional Scaling. Monographs on Statistics and Applied Probability*. Boca Raton: Chapman and Hall/CRC.

Deuschle, K., Chaudhuri, B., Okumoto, S., Lager, I., Lalonde, S., and Frommer, W. B. (2006). Rapid metabolism of glucose detected with FRET glucose nanosensors in epidermal cells and intact roots of *Arabidopsis* RNA-silencing mutants. *Plant Cell* 18, 2314–2325.

Eisenhut, M., Ruth, W., Haimovich, M., Bauwe, H., Kaplan, A., and Hagemann, M. (2008). The photorespiratory glycolate metabolism is essential for cyanobacteria and might have been conveyed endosymbiontically to plants. *Proc. Natl. Acad. Sci. U.S.A.* 105, 17199–17204.

Emanuelsson, O., Nielsen, H., Brunak, S., and Von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005–1016.

Farre, E. M., Fernie, A. R., and Willmitzer, L. (2008). Analysis of subcellular metabolite levels of potato tubers (*Solanum tuberosum*) displaying alterations in cellular or extracellular sucrose metabolism. *Metabolomics* 4, 161–170.

Farre, E. M., Tiessen, A., Roessner, U., Geigenberger, P., Trethewey, R. N., and Willmitzer, L. (2001). Analysis of the compartmentation of glycolytic intermediates, nucleotides, sugars, organic acids, amino acids, and sugar alcohols in potato tubers using a nonaqueous fractionation method. *Plant Physiol.* 127, 685–700.

Fehr, M., Frommer, W. B., and Lalonde, S. (2002). Visualization of maltose uptake in living yeast cells by fluorescent nanosensors. *Proc. Natl. Acad. Sci. U.S.A.* 99, 9846–9851.

Fernie, A. R. (2007). The future of metabolic phytochemistry: larger numbers of metabolites, higher resolution, greater understanding. *Phytochemistry* 68, 2861–2880.

Fernie, A. R., Trethewey, R. N., Krotzky, A. J., and Willmitzer, L. (2004). Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.* 5, 763–769.

Fettke, J., Chia, T., Eckermann, N., Smith, A., and Steup, M. (2006). A transglucosidase necessary for starch degradation and maltose metabolism in leaves at night acts on cytosolic heteroglycans (SHG). *Plant J.* 46, 668–684.

Fettke, J., Eckermann, N., Tiessen, A., Geigenberger, P., and Steup, M. (2005). Identification, subcellular localization and biochemical characterization of water-soluble heteroglycans (SHG) in leaves of *Arabidopsis thaliana* L.: distinct SHG reside in the cytosol and in the apoplast. *Plant J.* 43, 568–585.

Fettke, J., Hejazi, M., Smirnova, J., Hochel, E., Stage, M., and Steup, M. (2009). Eukaryotic starch degradation: integration of plastidial and cytosolic pathways. *J. Exp. Bot.* 60, 2907–2922.

Fiehn, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp. Funct. Genomics* 2, 155–168.

Gerhardt, R., and Heldt, H. W. (1984). Measurement of subcellular metabolite levels in leaves by fractionation of freeze-stopped material in nonaqueous media. *Plant Physiol.* 75, 542–547.

Giavalisco, P., Kapitza, K., Kolasa, A., Buhtz, A., and Kehr, J. (2006). Towards the proteome of *Brassica napus* phloem sap. *Proteomics* 6, 896–909.

Gibon, Y., Blaesing, O. E., Hannemann, J., Carillo, P., Hohne, M., Hendriks, J. H., Palacios, N., Cross, J., Selbig, J., and Stitt, M. (2004). A Robot-based platform to measure multiple enzyme activities in *Arabidopsis* using a set of cycling assays: comparison of changes of enzyme activities and transcript levels during diurnal cycles and in prolonged darkness. *Plant Cell* 16, 3304–3325.

Gout, E., Aubert, S., Bligny, R., Rebeille, F., Nonomura, A. R., Benson, A. A., and Douce, R. (2000). Metabolism of methanol in plant cells. Carbon-13 nuclear magnetic resonance studies. *Plant Physiol.* 123, 287–296.

Gout, E., Bligny, R., Pascal, N., and Douce, R. (1993). 13C nuclear magnetic resonance studies of malate and citrate synthesis and compartmentation in higher plant cells. *J. Biol. Chem.* 268, 3986–3992.

Gutscher, M., Pauleau, A. L., Marty, L., Brach, T., Wabnitz, G. H., Samstag, Y., Meyer, A. J., and Dick, T. P. (2008). Real-time imaging of the intracellular glutathione redox potential. *Nat. Methods* 5, 553–559.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *J. Intell. Inf. Syst.* 17, 107–145.

Hannah, M. A., Caldana, C., Steinhauser, D., Balbo, I., Fernie, A. R., and Willmitzer, L. (2010). Combined transcript and metabolite profiling of *Arabidopsis* grown under widely variant growth conditions facilitates the identification of novel metabolite-mediated regulation of gene expression. *Plant Physiol.* 152, 2120–2129.

Heazlewood, J. L., Verboom, R. E., Tonti-Filippini, J., Small, I., and Millar, A. H. (2007). SUBA: the *Arabidopsis* subcellular database. *Nucleic Acids Res.* 35, D213-218.

Heinrich, G., and Kuschki, B. (1978). Verluste radioaktiv markierter Substanzen aus Pisum-Wurzeln nach Verfütterung von D-Glucose-14C im Verlauf unterschiedlicher Präparationsmethoden für die Elektronenmikroskopie. *Histochem. Cell Biol.* 319–328.

Hennig, C. (2010). *fpc: Flexible Procedures for Clustering*. R package version 2.0–2. Available at: http://CRAN.R-project.org/package=fpc

Huege, J., Krall, L., Steinhauser, M. C., Giavalisco, P., Rippka, R., Tandeau De Marsac, N., and Steinhauser, D. (2011). Sample amount alternatives for data adjustment in comparative cyanobacterial metabolomics. *Anal. Bioanal. Chem.* 399, 3503–3517.

Jackson, C. (2010). *msm: Multi-State Markov and Hidden Markov Models in Continuous Time*. R package version 0.9.7. Available at: http://CRAN.R-project.org/package=msm

Jozefczuk, S., Klie, S., Catchpole, G., Szymanski, J., Cuadros-Inostroza, A., Steinhauser, D., Selbig, J., and Willmitzer, L. (2010). Metabolomic and transcriptomic stress response of *Escherichia coli*. *Mol. Syst. Biol.* 6, 364.

Junker, B. H., Wuttke, R., Nunes-Nesi, A., Steinhauser, D., Schauer, N., Bussis, D., Willmitzer, L., and Fernie, A. R. (2006). Enhancing vacuolar sucrose cleavage within the developing potato tuber has only minor effects on metabolism. *Plant Cell Physiol.* 47, 277–289.

Kaplan, F., Kopka, J., Haskell, D. W., Zhao, W., Schiller, K. C., Gatzke, N., Sung, D. Y., and Guy, C. L. (2004). Exploring the temperature-stress metabolome of *Arabidopsis*. *Plant Physiol.* 136, 4159–4168.

Kaplan, F., Kopka, J., Sung, D. Y., Zhao, W., Popp, M., Porat, R., and Guy, C. L. (2007). Transcript and metabolite profiling during cold acclimation of *Arabidopsis* reveals an intricate relationship of cold-regulated gene expression with modifications in metabolite content. *Plant J.* 50, 967–981.

Katharine, M. M., and Van Stokkum, I. H. M. (2010). *nnls: The Lawson-Hanson Algorithm for Non-Negative Least Squares (NNLS)*. R package version 1.3. Available at: http://CRAN.R-project.org/package=nnls

Kopka, J., Fernie, A., Weckwerth, W., Gibon, Y., and Stitt, M. (2004). Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.* 5, 109.

Krueger, S., Donath, A., Lopez-Martin, M. C., Hoefgen, R., Gotor, C., and Hesse, H. (2010). Impact of sulfur starvation on cysteine biosynthesis in T-DNA mutants deficient for compartment-specific serine-acetyltransferase. *Amino Acids* 39, 1029–1042.

Krueger, S., Giavalisco, P., Krall, L., Steinhauser, M. C., Bussis, D., Usadel, B., Flugge, U. I., Fernie, A. R., Willmitzer, L., and Steinhauser, D. (2011). A topological map of the compartmentalized *Arabidopsis thaliana* leaf metabolome. *PLoS ONE* 6, e17806. doi: 10.1371/journal.pone.0017806

Krueger, S., Niehl, A., Lopez Martin, M. C., Steinhauser, D., Donath, A., Hildebrandt, T., Romero, L. C., Hoefgen, R., Gotor, C., and Hesse, H. (2009). Analysis of cytosolic and plastidic serine acetyltransferase mutants and subcellular metabolite distributions suggests interplay of the cellular compartments for cysteine biosynthesis in *Arabidopsis*. *Plant Cell Environ.* 32, 349–367.

Kruger, N. J., and Von Schaewen, A. (2003). The oxidative pentose phosphate pathway: structure and organisation. *Curr. Opin. Plant Biol.* 6, 236–246.

Kunz, H. H., Scharnewski, M., Feussner, K., Feussner, I., Flugge, U. I., Fulda, M., and Gierth, M. (2009). The ABC transporter PXA1 and peroxisomal beta-oxidation are vital for metabolism in mature leaves of *Arabidopsis* during extended darkness. *Plant Cell* 21, 2733–2749.

Kusano, M., Tohge, T., Fukushima, A., Kobayashi, M., Hayashi, N., Otsuki, H., Kondou, Y., Goto, H., Kawashima, M., Matsuda, F., Niida, R., Matsui, M., Saito, K., and Fernie, A. R. (2011). Metabolomics reveals comprehensive reprogramming involving two independent metabolic responses of *Arabidopsis* to ultraviolet-B light. *Plant J.* 67, 354–369.

Lalonde, S., Ehrhardt, D. W., and Frommer, W. B. (2005). Shining light on signaling and metabolic networks by genetically encoded biosensors. *Curr. Opin. Plant Biol.* 8, 574–581.

Lawson, C. L., and Hanson, R. J. (1995). *Solving Least Squares Problems. Classics in Applied Mathematics*. Philadelphia: SIAM.

Lemon, J. (2006). Plotrix: a package in the red light district of R. *R. News* 6, 8–12.

Libourel, I. G., Van Bodegom, P. M., Fricker, M. D., and Ratcliffe, R. G. (2006). Nitrite reduces cytoplasmic acidosis under anoxia. *Plant Physiol.* 142, 1710–1717.

Lilley, R. McC., Stitt, M., Mader, G., and Heldt, H. W. (1982). Rapid fractionation of wheat leaf protoplasts using membrane filtration: the determination of metabolite levels in the chloroplasts, cytosol, and mitochondria. *Plant Physiol.* 70, 965–970.

Lunn, J. E. (2007). Compartmentation in plant metabolism. *J. Exp. Bot.* 58, 35–47.

Martinoia, E., Maeshima, M., and Neuhaus, H. E. (2007). Vacuolar transporters and their essential role in plant metabolism. *J. Exp. Bot.* 58, 83–102.

Meyer, R. C., Steinfath, M., Lisec, J., Becher, M., Witucka-Wall, H., Torjek, O., Fiehn, O., Eckardt, A., Willmitzer, L., Selbig, J., and Altmann, T. (2007). The metabolic signature related to high plant growth rate in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 104, 4759–4764.

Mugford, S. G., Yoshimoto, N., Reichelt, M., Wirtz, M., Hill, L., Mugford, S. T., Nakazato, Y., Noji, M., Takahashi, H., Kramell, R., Gigolashvili, T., Flugge, U. I., Wasternack, C., Gershenzon, J., Hell, R., Saito, K., and Kopriva, S. (2009). Disruption of adenosine-5'-phosphosulfate kinase in *Arabidopsis* reduces levels of sulfated secondary metabolites. *Plant Cell* 21, 910–927.

Pan, Z., and Raftery, D. (2007). Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics. *Anal. Bioanal. Chem.* 387, 525–527.

Paris, N., Stanley, C. M., Jones, R. L., and Rogers, J. C. (1996). Plant cells contain two functionally distinct vacuolar compartments. *Cell* 85, 563–572.

Peters, T. Jr., and Ashley, C. A. (1967). An artefact in radioautography due to binding of free amino acids to tissues by fixatives. *J. Cell Biol.* 33, 53–60.

Pracharoenwattana, I., Cornah, J. E., and Smith, S. M. (2005). *Arabidopsis* peroxisomal citrate synthase is required for fatty acid respiration and seed germination. *Plant Cell* 17, 2037–2048.

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0. Available at: http://www.R-project.org

Rébeillé, F., Alban, C., Bourguignon, J., Ravanel, S., and Douce, R. (2007). The role of plant mitochondria in the biosynthesis of coenzymes. *Photosyn. Res.* 92, 149–162.

Riens, B., Lohaus, G., Heineke, D., and Heldt, H. W. (1991). Amino acid and sucrose content determined in the cytosolic, chloroplastic, and vacuolar compartments and in the phloem sap of spinach leaves. *Plant Physiol.* 97, 227–233.

Riewe, D., Grosman, L., Fernie, A. R., Wucke, C., and Geigenberger, P. (2008). The potato-specific apyrase is apoplastically localized and has influence on gene expression, growth, and development. *Plant Physiol.* 147, 1092–1109.

Robert, C. P. (1995). Simulation of truncated normal variables. *Stat. Comput.* 5, 121–125.

Robinson, S. P., and Walker, D. A. (1979). Rapid separation of the chloroplast and cytoplasmic fractions from intact leaf protoplasts. *Arch. Biochem. Biophys.* 196, 319–323.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65.

Saito, K., and Matsuda, F. (2010). Metabolomics for functional genomics, systems biology, and biotechnology. *Annu. Rev. Plant Biol.* 61, 463–489.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. New York: Springer.

Schauer, N., Semel, Y., Roessner, U., Gur, A., Balbo, I., Carrari, F., Pleban, T., Perez-Melis, A., Bruedigam, C., Kopka, J., Willmitzer, L., Zamir, D., and Fernie, A. R. (2006). Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nat. Biotechnol.* 24, 447–454.

Schwacke, R., Schneider, A., Van Der Graaff, E., Fischer, K., Catoni, E., Desimone, M., Frommer, W. B., Flugge, U. I., and Kunze, R. (2003). ARAMEMNON, a novel database for *Arabidopsis* integral membrane proteins. *Plant Physiol.* 131, 16–26.

Sokal, R. R., and Rohlf, F. J. (1995). *Biometry: The Principles and Practice of Statistics in Biological Research*. New York: W. H. Freeman and Company.

Steinhauser, D., Usadel, B., Luedemann, A., Thimm, O., and Kopka, J. (2004). CSB.DB: a comprehensive systems-biology database. *Bioinformatics* 20, 3647–3651.

Steinhauser, M. C., Steinhauser, D., Koehl, K., Carrari, F., Gibon, Y., Fernie, A. R., and Stitt, M. (2010). Enzyme activity profiles during fruit development in tomato cultivars and *Solanum pennellii*. *Plant Physiol.* 153, 80–98.

Stitt, M., and Fernie, A. R. (2003). From measurements of metabolites to metabolomics: an "on the fly" perspective illustrated by recent studies of carbon-nitrogen interactions. *Curr. Opin. Biotechnol.* 14, 136–144.

Stitt, M., Lilley, R. M., Gerhardt, R., and Heldt, H. W. (1989). Determination of metabolite levels in specific cells and subcellular compartments of plant leaves. *Meth. Enzymol.* 174, 518–552.

Stitt, M., Lilley, R. M., and Heldt, H. W. (1982). Adenine nucleotide levels in the cytosol, chloroplasts, and mitochondria of wheat leaf protoplasts. *Plant Physiol.* 70, 971–977.

Stitt, M., Lunn, J., and Usadel, B. (2010a). *Arabidopsis* and primary photosynthetic metabolism – more than the icing on the cake. *Plant J.* 61, 1067–1091.

Stitt, M., Sulpice, R., and Keurentjes, J. (2010b). Metabolic networks: how to identify key components in the regulation of metabolism and growth. *Plant Physiol.* 152, 428–444.

Stitt, M., Wirtz, W., and Heldt, H. W. (1983). Regulation of sucrose synthesis by cytoplasmic fructosebisphosphatase and sucrose phosphate synthase during photosynthesis in varying light and carbon dioxide. *Plant Physiol.* 72, 767–774.

Strang, G. (2009). *Introduction to Linear Algebra.* Wellesley: Wellesley-Cambridge Press.

Strasser, R., Schoberer, J., Jin, C., Glossl, J., Mach, L., and Steinkellner, H. (2006). Molecular cloning and characterization of *Arabidopsis thaliana* Golgi alpha-mannosidase II, a key enzyme in the formation of complex N-glycans in plants. *Plant J.* 45, 789–803.

Sulpice, R., Pyl, E. T., Ishihara, H., Trenkamp, S., Steinfath, M., Witucka-Wall, H., Gibon, Y., Usadel, B., Poree, F., Piques, M. C., Von Korff, M., Steinhauser, M. C., Keurentjes, J. J., Guenther, M., Hoehne, M., Selbig, J., Fernie, A. R., Altmann, T., and Stitt, M. (2009). Starch as a major integrator in the regulation of plant growth. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10348–10353.

Suzuki, R., and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22, 1540–1542.

Sweetlove, L. J., Fell, D., and Fernie, A. R. (2008). Getting to grips with the plant metabolic network. *Biochem. J.* 409, 27–41.

Takeshige, K., and Tazawa, M. (1989). Determination of the inorganic pyrophosphate level and its subcellular localization in *Chara corallina*. *J. Biol. Chem.* 264, 3262–3266.

Taylor, N. L., Heazlewood, J. L., and Millar, A. H. (2011). The *Arabidopsis thaliana* 2-D gel mitochondrial proteome: refining the value of reference maps for assessing protein abundance, contaminants and post-translational modifications. *Proteomics* 11, 1720–1733.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a dataset via the gap statistic. *J. R. Stat. Soc. Ser. B* 63, 411–423.

Walker, R. P., Chen, Z. H., Johnson, K. E., Famiani, F., Tecsi, L., and Leegood, R. C. (2001). Using immunohistochemistry to study plant metabolism: the examples of its use in the localization of amino acids in plant tissues, and of phosphoenolpyruvate carboxykinase and its possible role in pH regulation. *J. Exp. Bot.* 52, 565–576.

Weber, A. P., and Fischer, K. (2007). Making the connections–the crucial role of metabolite transporters at the interface between chloroplast and cytosol. *FEBS Lett.* 581, 2215–2222.

Weise, S. E., Weber, A. P., and Sharkey, T. D. (2004). Maltose is the major form of carbon exported from the chloroplast at night. *Planta* 218, 474–482.

Winter, H., Robinson, D. G., and Heldt, H. W. (1993). Subcellular volumes and metabolite concentrations in barley leaves. *Planta* 191, 180–190.

Wirtz, W., Stitt, M., and Heldt, H. W. (1980). Enzymic determination of metabolites in the subcellular compartments of spinach protoplasts. *Plant Physiol.* 66, 187–193.

Yamada, K., Norikoshi, R., Suzuki, K., Imanishi, H., and Ichimura, K. (2009). Determination of subcellular concentrations of soluble carbohydrates in rose petals during opening by nonaqueous fractionation method combined with infiltration-centrifugation method. *Planta* 230, 1115–1127.

Zechmann, B., Stumpe, M., and Mauch, F. (2011). Immunocytochemical determination of the subcellular distribution of ascorbate in plants. *Planta* 233, 1–12.

Zeeman, S. C., Thorneycroft, D., Schupp, N., Chapple, A., Weck, M., Dunstan, H., Haldimann, P., Bechtold, N., Smith, A. M., and Smith, S. M. (2004). Plastidial alpha-glucan phosphorylase is not required for starch degradation in *Arabidopsis* leaves but has a role in the tolerance of abiotic stress. *Plant Physiol.* 135, 849–858.

Conflict of Interest Statement: 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.

Received: 01 May 2011; accepted: 05 September 2011; published online: 22 September 2011.

*Citation: Klie S, Krueger S, Krall L, Giavalisco P, Flügge U-I, Willmitzer L and Steinhauser D (2011) Analysis of the compartmentalized metabolome – a validation of the non-aqueous fractionation technique. Front. Plant Sci. **2**:55. doi: 10.3389/fpls.2011.00055*

This article was submitted to Frontiers in Plant Physiology, a specialty of Frontiers in Plant Science.

APPENDIX

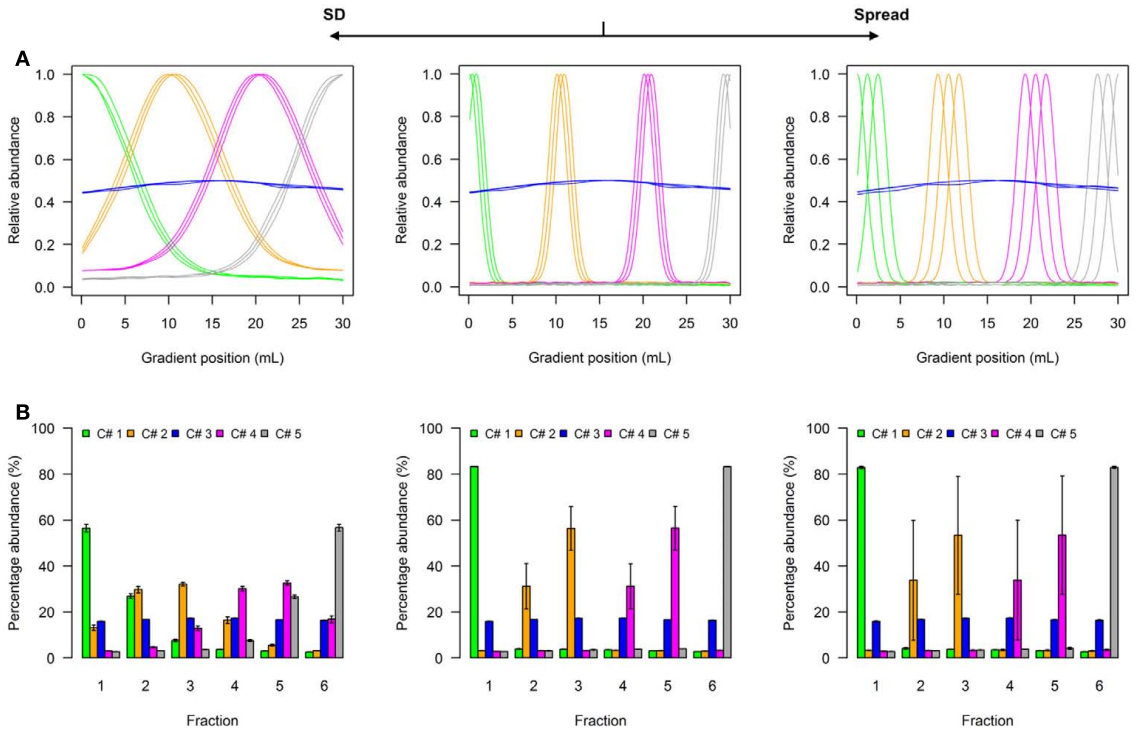


FIGURE A1 | Simplified scheme of the five-compartmental simulation model. (A) The continuous distributions are depicted by line plots for each of the 5 compartments represented by 3 individual marker distributions. To aid visualization the distributions are scaled to half-maximum (blue-colored compartment) or maximum (all other) observed values. **(B)** The bar plots show the mean–averaged fraction abundances including SD among compartment-specific markers for each compartment after the continuous distributions were discretized into 6 equally spaced fractions. The left–side graph illustrates the effect of increasing the SD (SD = 5, ms = 0.4), while the right-side graph shows the effect of increasing the marker spread ms (SD = 1, ms = 1.2) compared to a standard (middle graph with SD = 1, ms = 0.4).

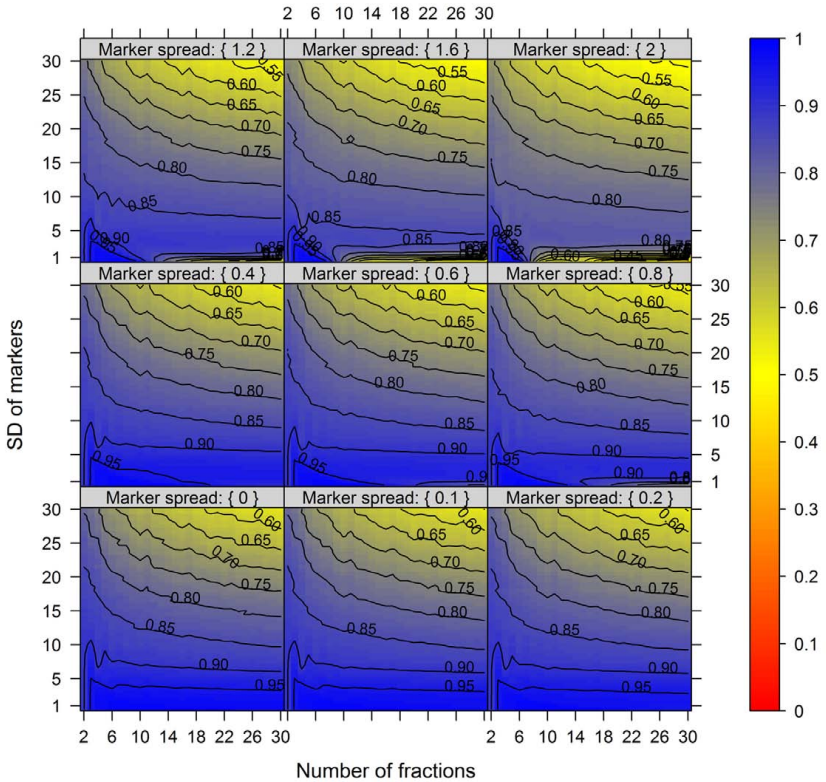


FIGURE A2 | The cluster validity in dependence of the number of collected fractions, the SD of markers, and the marker spread for the three-compartmental model. While the SD modulates the between-compartmental separation, the marker spread modulates the within-compartmental cohesion (cf. Figure 3). The cluster validity index estimated as mean–average of the Silhouette information and the Pearson's matrix correlation is depicted for 3 compartments each represented by 3 markers. The closer the value is to 1 the better the observed cluster validity, color-coded as depicted in the right-side bar. To aid visualization negative cluster validity values were set to 0 and contour lines were drawn for each 0.05 unit.

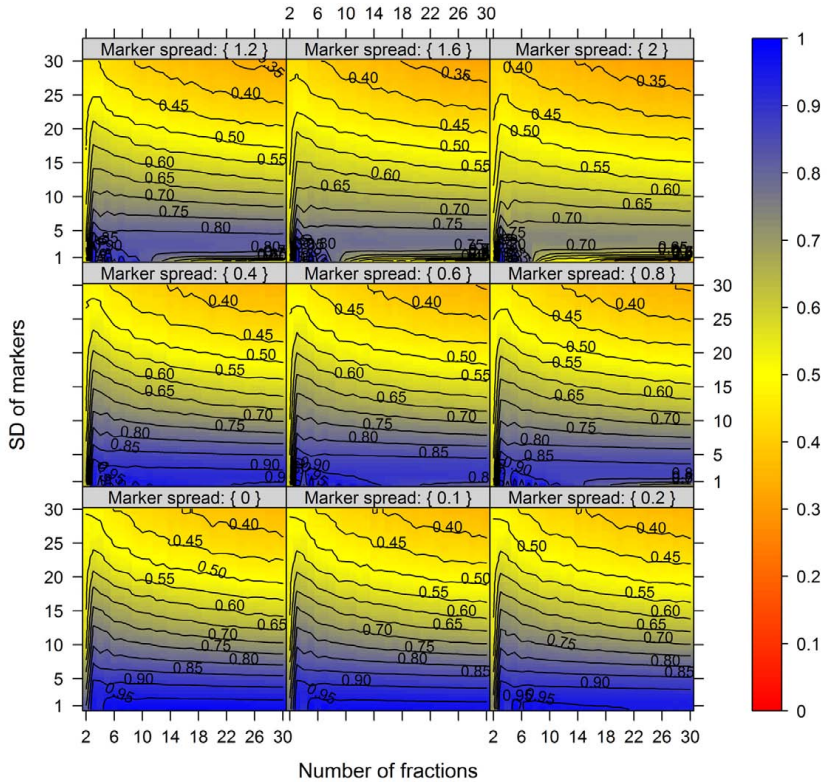


FIGURE A3 | The cluster validity in dependence of the number of collected fractions, the SD of markers, and the marker spread for the four-compartmental model. The cluster validity index is depicted for 4 compartments each represented by 3 markers. For further details see **Figure A2**.

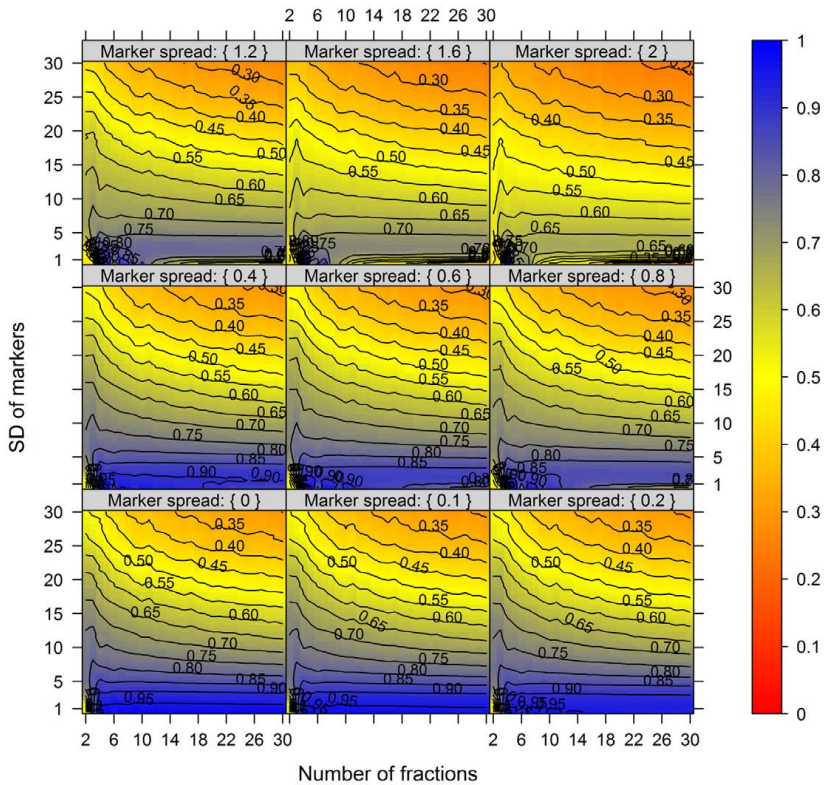


FIGURE A4 | The cluster validity in dependence of the number of collected fractions, the SD of markers, and the marker spread for the five-compartmental model. The cluster validity index is depicted for 5 compartments each represented by 3 markers. For further details see **Figure A2**.

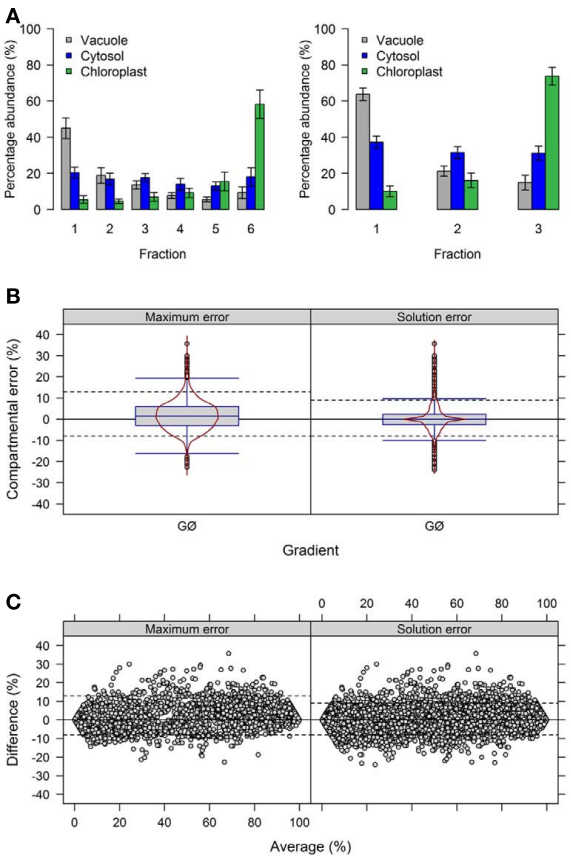


FIGURE A5 | Influence of the number of collected fraction on estimated compartmental abundances based on experimental data. (A) Bar plots depicting the mean–averages and SDs of the three compartment-specific markers for each of the three resolved compartments on the basis of 6 and 3 collected fractions. Neighboring fractions, i.e., $1+2$, $3+4$, and $5+6$, were averaged to obtain results for 3 fractions. **(B)** Combined box- and violin plots (red lines) as well as **(C)** mean-difference plots showing the difference of the maximum and solution error based on compartmental abundances estimated using BFA for gradients comprising 6 and 3 fractions. For both, positive values indicate larger compartmental abundances for 6 fractions while negative values depict larger abundances for 3 fractions. The 5th and 95th percentiles are drawn as black dashed lines. The figures show the average of three independent gradients.

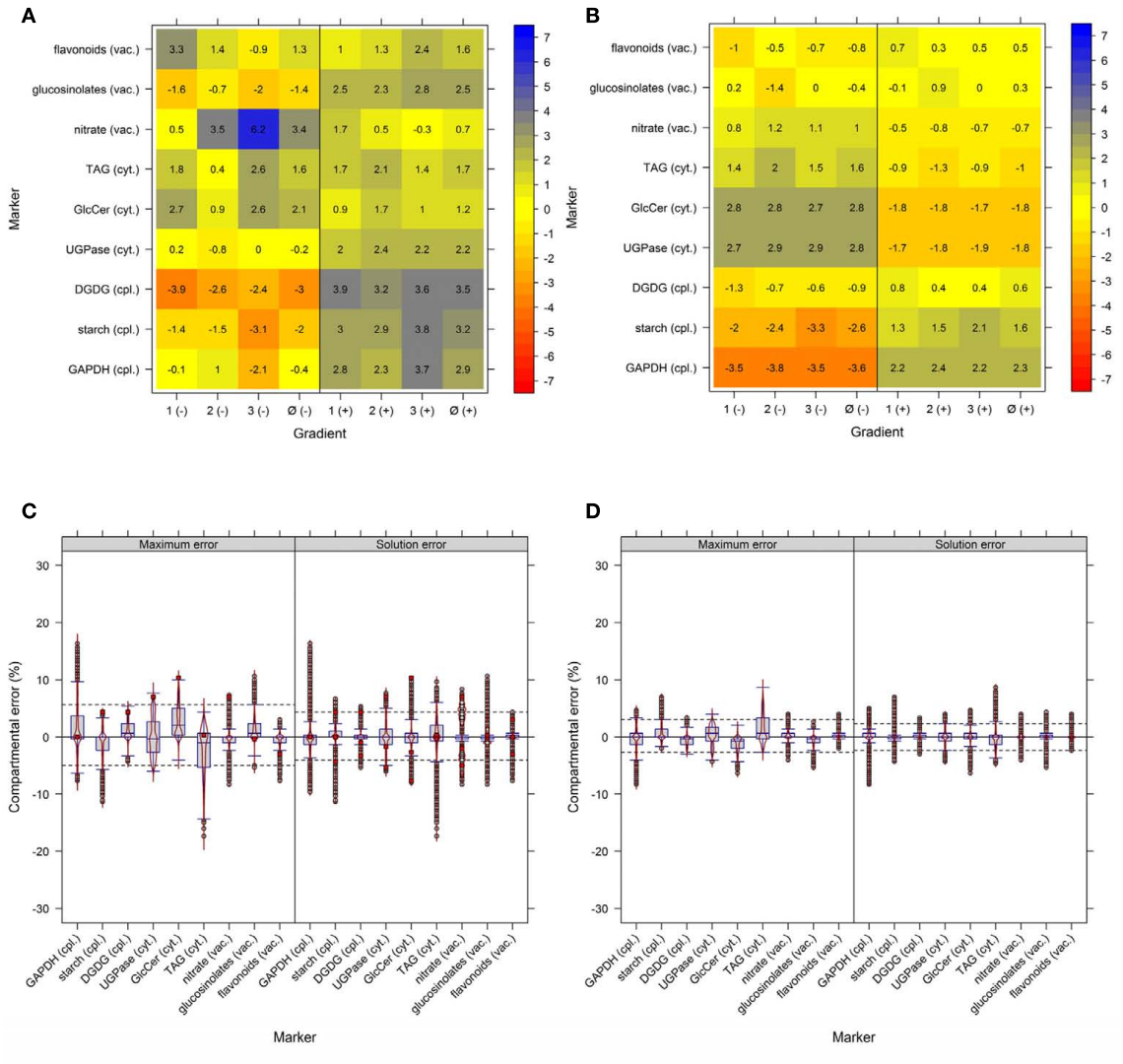


FIGURE A6 | Influence of compartment-specific marker combinations on (A) cluster validity, (B) between-compartment separation, and (C,D) estimated compartmental abundances based on experimental data. (A) The cluster validity index, estimated as mean–average of the Silhouette information and the Pearson's matrix correlation, and **(B)** the between-compartment separation are depicted as percentage difference from the cluster solution using all nine compartmental markers. The values are provided for all three independent gradients (1–3) and as mean–average (∅) by deleting one marker (−; jackknife−) or taking one marker twice (+; jackknife+). **(C,D)** Combined box- and violin plots (red lines) showing the difference of the maximum and solution error based on compartmental abundances estimated using BFA when **(C)** deleting one marker or **(D)** considering one marker twice. All estimates are based on the difference observed after mean–average of the three independent gradients. Red squares in **(C)** depict the difference in the marker that was deleted. The 5th and 95th percentiles are drawn as black dashed lines. For both, positive values indicate larger compartmental abundances when all nine markers were used while negative values depict larger abundances when a marker was deleted or considered twice.

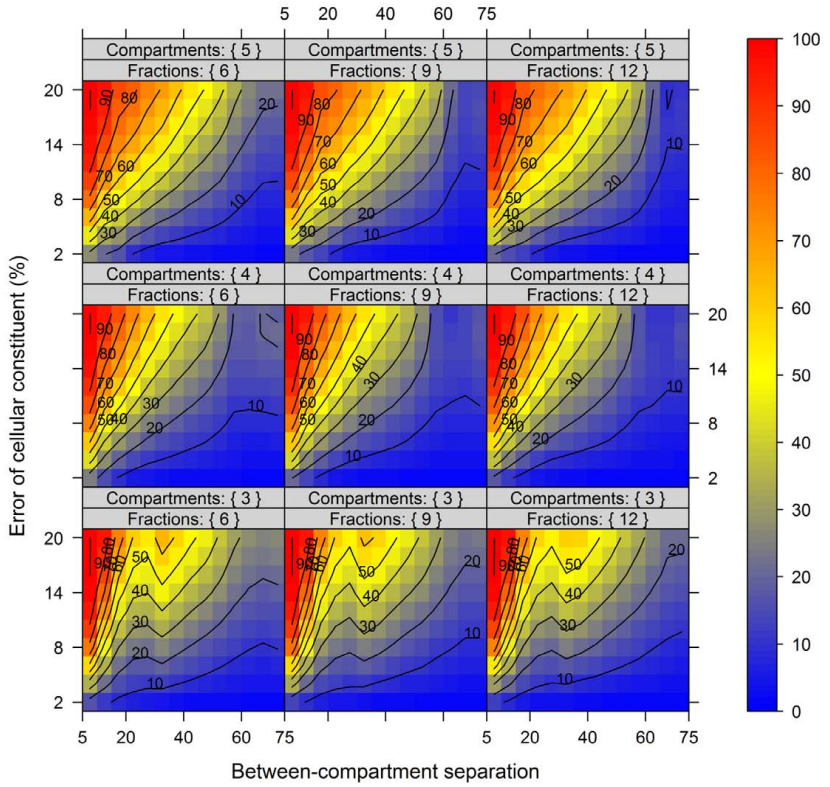


FIGURE A7 | Mean–averaged calculation error of compartmental abundances in dependence of the between-compartment separation and the metabolite error based on simulated data. The estimated maximum error is depicted as mean–average for each combination of between-compartment separation and metabolite error in dependence of the number of compartments (3–5) and number of fraction (6, 9, and 12) considered. The closer the value is to 0% the smaller the error of the compartmental abundances (estimated using NNLS), color-coded as depicted in the right-side bar. Contour lines are drawn for each 10% error.

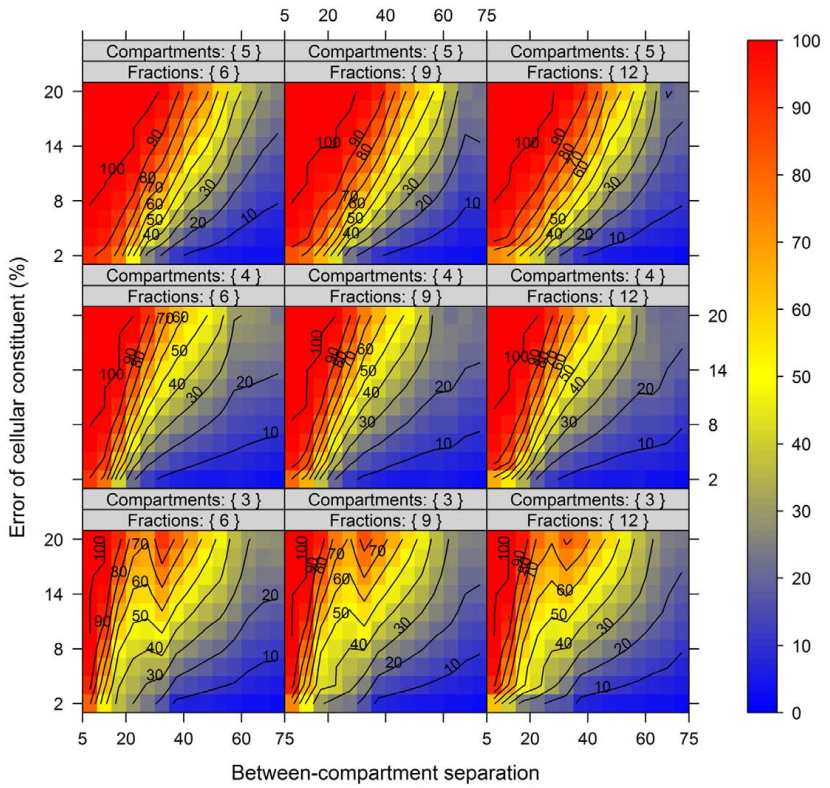


FIGURE A8 | Percentile (99%) calculation error of compartmental abundances in dependence of the between-compartment separation and the metabolite error based on simulated data. The error is depicted as 99th percentile for each combination of between-compartment separation and metabolite error in dependence of the number of compartments (3–5) and number of fraction (6, 9, and 12) considered. For further details see **Figure A7**.

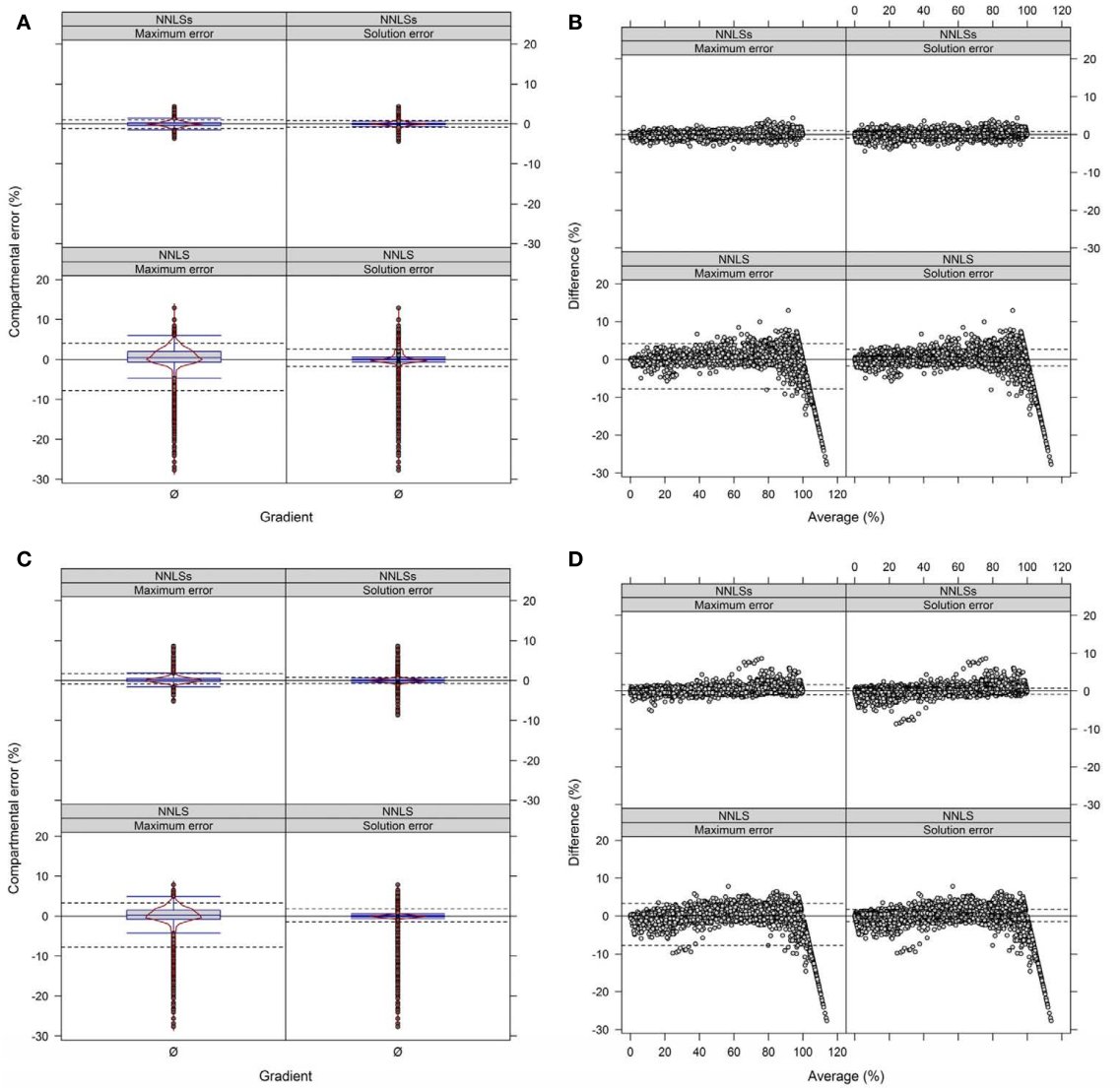


FIGURE A9 | Diagnostic plots showing the differences in estimated compartmental abundances using (A,B) a three- and (C,D) a four-compartmental estimation strategy based on experimental data. **(A,C)** The distributions of observed differences in compartmental abundances are shown as box- and violin plots (red lines). **(B,D)** The mean-difference plots depict the differences in dependence of the averages of compartmental abundances. For all graphs the difference in compartmental abundances between BFA and NNLS or iterative (seed-based) (NNLSs) are visualized based on the mean–average of compartmental abundances for the three independent gradients. The 5th and 95th percentiles are drawn as black dashed lines. All values are shown, regardless if they are considered as sufficiently explained or not.

frontiers in
PLANT SCIENCE

PERSPECTIVE ARTICLE
published: 10 October 2011
doi: 10.3389/fpls.2011.00063

Experimental flux measurements on a network scale

Jörg Schwender*

Department of Biology, Brookhaven National Laboratory, Upton, NY, USA

Edited by:
Alisdair Fernie, Max Planck Institute for Plant Physiology, Germany

Reviewed by:
Alisdair Fernie, Max Planck Institute for Plant Physiology, Germany
Lee Sweetlove, University of Oxford, UK
Ganesh Sriram, University of Maryland, USA

****Correspondence:***
Jörg Schwender, Department of Biology, Brookhaven National Laboratory, Building 463, Upton, NY 11973, USA.
e-mail: schwend@bnl.gov

Metabolic flux is a fundamental property of living organisms. In recent years, methods for measuring metabolic flux in plants on a network scale have evolved further. One major challenge in studying flux in plants is the complexity of the plant's metabolism. In particular, in the presence of parallel pathways in multiple cellular compartments, the core of plant central metabolism constitutes a complex network. Hence, a common problem with the reliability of the contemporary results of ^{13}C-Metabolic Flux Analysis in plants is the substantial reduction in complexity that must be included in the simulated networks; this omission partly is due to limitations in computational simulations. Here, I discuss recent emerging strategies that will better address these shortcomings.

Keywords: ^{13}C-metabolic flux analysis, primary metabolism, flux balance analysis, carbon partitioning, constraint-based model

INTRODUCTION

Isotopic tracers have different but important uses in metabolic research. Among the various approaches to stoichiometrical modeling of cell metabolism (Llaneras and Pico, 2008), ^{13}C-Metabolic Flux Analysis (^{13}C-MFA) is a method that combines a knowledge of cell metabolism with ^{13}C-tracer experiments to analyze the *in vivo* flux distribution in the network of central cellular primary metabolism. It affords us a quantitative integrated view of core metabolism (Koschutzki et al., 2010) that unravels the *in vivo* function of biochemical pathways under different physiological conditions, or reveals the effect of manipulation by transgenic approaches. In plants, ^{13}C-MFA mostly is applied to cultures of cells or tissue, growing heterotrophically or photoheterotrophically on ^{13}C-labeled substrates. The increasing number of studies on different species over the last 10- to 15-years documents the development of researches with ^{13}C-MFA in plants. Maize root tips, detached from germinating seeds, were first used as a model to study energy metabolism in non-photosynthetic tissues (Dieuaide-Noubhani et al., 1995; Alonso et al., 2005, 2007b,c). Other studies used the hairy root cultures of *Catharanthus roseus*, the Madagascar periwinkle (Sriram et al., 2007), and cell-suspension cultures of tomato or *Arabidopsis thaliana* (Rontein et al., 2002; Williams et al., 2008; Masakapalli et al., 2010). Various studies focused on the distribution of flux in central metabolism in the developing seeds and embryos of rapeseed and *Arabidopsis* (Schwender and Ohlrogge, 2002; Schwender et al., 2003, 2004a, 2006; Junker et al., 2007; Lonien and Schwender, 2009), soybean (Sriram et al., 2004; Iyer et al., 2008; Allen et al., 2009b), sunflower (Alonso et al., 2007a), and in developing maize endosperm or embryos (Ettenhuber et al., 2005; Spielbauer et al., 2006; Alonso et al., 2010, 2011). Several studies also began to assess the effect of physiological- or genotypical-perturbations of central metabolism (Alonso et al., 2007b; Junker et al., 2007; Iyer et al., 2008; Williams et al., 2008; Lonien and Schwender, 2009). Recent studies began to explore the synergy between plant ^{13}C-MFA and the more predictive modeling approach of flux balance analysis (FBA; Williams et al., 2010; Hay and Schwender, 2011a,b).

A great deal of biological knowledge about an organism is needed to construct a model of its biochemical network. Even in the post-genomic age, the definition of metabolic networks is not straightforward (Sweetlove et al., 2008). Yet the results of the analytic process critically depend upon having a realistic network (van Winden et al., 2001a; Schwender et al., 2004b; Masakapalli et al., 2010). Due to the practices of computational analysis, the typical scale of a ^{13}C-MFA network (**Table 1**) results from tailoring to a smaller size the detailed topology inferred from literature, e.g., by lumping the metabolite pools present in different subcellular compartments.

This paper offers some insights into the experimental- and computational-modeling practices of ^{13}C-MFA to highlight the typical assumptions built into such models, and to discuss how their constructions and their general reliability can be improved. I discuss modeling related to applying 13CFLUX (Wiechert et al., 2001; www.13cflux.net), a software used by many groups in the field. The paper is not intended to be a comprehensive review of all recent work, but rather, to give my personal perspective based on the practice of experimental- and computational- modeling of plant central metabolism.

PRINCIPLE OF STEADY-STATE ^{13}C-MFA

Several recent detailed reviews summarize experimental procedures, the modeling process, as well as discuss important biological insights that have resulted from plant ^{13}C-MFA studies, e.g., Libourel and Shachar-Hill (2008), Schwender (2008), Kruger and Ratcliffe (2009), Allen et al. (2009a), and Schwender (2009). **Figure 1** illustrates a general experimental workflow. Zamboni et al. (2009) gave a very detailed and useful description of ^{13}C-MFA, including a tutorial for 13CFLUX. In short, an organism is grown on a minimal culture medium with well-defined composition of organic- and inorganic-substrates. While

Table 1 | Characteristics of example networks used in ^{13}C-MFA and FBA of higher plants.

	E. coli[1]	*B. napus*[2]	*A. thaliana*[3]	*B. napus*[4]	*A. thaliana*[5]	*A. thaliana*[6]
Modeling approach	^{13}C-MFA	^{13}C-MFA	^{13}C-MFA	FBA	FBA	FBA
Reconstruction	Bibliomic, lumped	Bibliomic, lumped	Bibliomic, lumped	Bibliomic, large-scale	Genome-scale	Genome-scale
Network type	Carbon label network	Carbon label network	Carbon label network	Stoichiometric network	Stoichiometric network	Stoichiometric network
Intracellular compartments	1	3	3	9		4
Metabolic pools	37	86[10]	82	376	1253	1748
Reactions	68	146[10]	125	572	1406	1567
Uptake/exchange reactions [7]	2	4	4	14	6	18
Biomass drain fluxes[8]	10	15	19	41	36	47
Total carbon positions in network	186	354[10]	387	–	–	–
Full network simulation (cumomers)[9]	3183	1974[10]	4045	–	–	–
Reduced network simulation (EMU's)[9]	438	672[10]	514	–	–	–
MS measurement groups/number of total signals	35/193	37/165	29/160			

Data were obtained from different ^{13}C-MFA models available in executable 13CFLUX model format, with consideration of outputs of the function "Benchmark" in 13CFLUX2 (re-implementation of 13CFLUX, www.13cflux.net). For FBA models, data were obtained from respective publications. [1]Zamboni et al. (2009). [2]Schwender et al. (2006). [3]Lonien and Schwender (2009). [4]Hay and Schwender (2011a,b). [5]Poolman et al. (2009). [6]de Oliveira Dal'Molin et al. (2010a). [7]Includes inorganic uptakes, CO_2 and O_2 exchanges, or light flux. [8]Number of metabolites that are accumulated in biomass. [9]The number of labeling state variables (cumomers or EMU's) largely determines computational speed. [10]The network size actually reflects the presence of three metabolic networks simulated simultaneously to evaluate data from three experiments with different ^{13}C-tracers.

a ^{13}C-labeled carbon source (e.g., [1-^{13}C]glucose) is being metabolized, ^{12}C- and ^{13}C-atoms are distributed throughout the organisms' metabolic network. The fate of a ^{13}C-labeled carbon position of the carbon source, or of pairs of adjacent ^{13}C-atoms (^{13}C–^{13}C bond label) is traced through the network by detecting the labeling signatures of the intermediary metabolites by the techniques of mass spectrometry (MS; Dauner and Sauer, 2000; Schwender and Ohlrogge, 2002) or nuclear magnetic resonance (NMR; Dieuaide-Noubhani et al., 1995; Szyperski, 1995). For the widely used approach of metabolic- and isotopic-stationary ^{13}C-MFA, the essential prerequisite is that the labeling state of each metabolite attains a steady-state before the cells are harvested, the metabolites extracted, and the labeling signatures analyzed (Wiechert, 2001). Thus, information is gained about intracellular fluxes for alternative pathways converging to the same metabolite (Szyperski, 1995; van Winden et al., 2001b), meaning that different labeling signatures are generated and mixed in one metabolite at the convergent node. For example, oxaloacetate (OxA) may be labeled differentially depending on whether it is formed by the carboxylation of phosphoenol pyruvate, or from α-ketoglutarate via the reactions of the tricarboxylic acid (TCA) cycle (**Figure 2B**). Whether we can evaluate the flux ratio at the OxA node rests upon the particular ^{13}C-substrate label used in the culture. Other nodes of this kind are pyruvate, α-ketoglutarate, and 3-phosphoglyceric acid. Often the labeling pattern in these intermediates is not measured directly, but accessed indirectly through their anabolic products (Szyperski, 1995, 1998). Asp, for example, accumulates in protein and represents the labeling signature in OxA (**Figure 2A**). For studying plant flux, the analyses of protein-bound amino acids by NMR, or by gas chromatography/MS (GC/MS) methods, have emerged as standard practices (Allen and Ratcliffe, 2009).

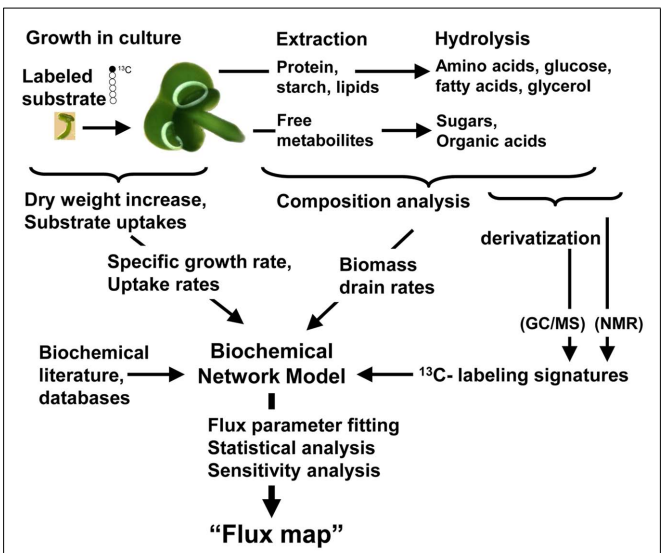


FIGURE 1 | Basic workflow in ^{13}C-MFA. For details see text.

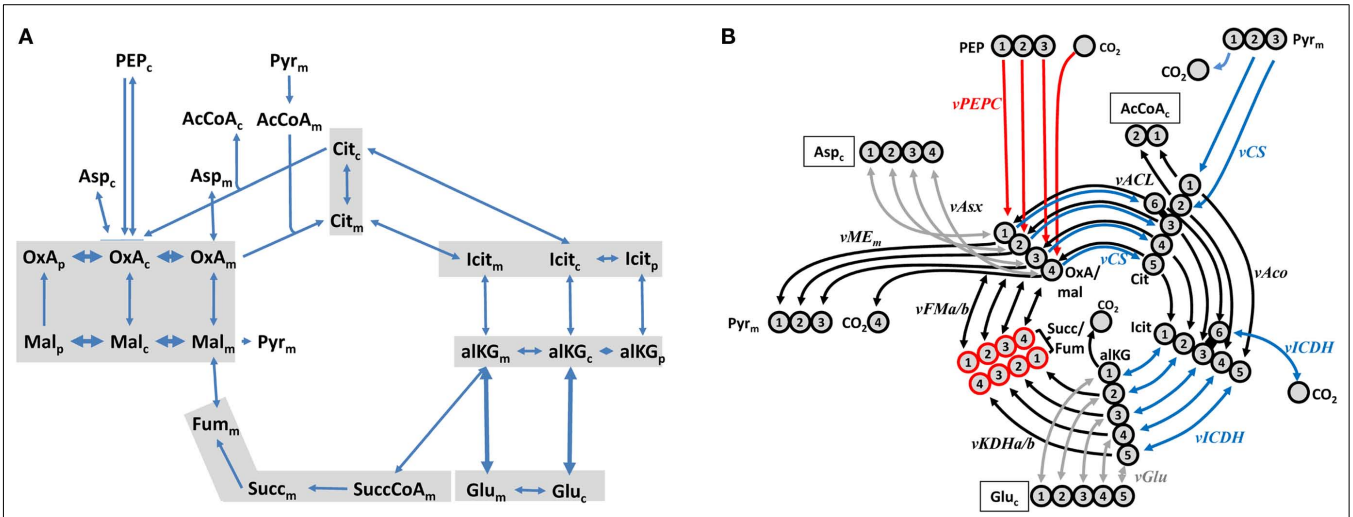


FIGURE 2 | Difference in network topology between large-scale stoichiometric models and the models used in ^{13}C-MFA. Two representations are shown of a sub-network representing the TCA cycle and some associated reactions in *B. napus* developing embryos. Indices *c, p,* and *m* indicate whether the localization of a metabolite is specified to be cytosol, plastid, or mitochondrium, respectively. Metabolites for which ^{13}C-labeling signatures are measured are boxed. **(A)** Representation of the network in *bna572* (Hay and Schwender, 2011a). Arrows depict individual reactions that are formulated in *bna572* by a complete reaction stoichiometry. To make the topology understandable, in most cases only one substrate to product transition is shown for each reaction. Thick arrows indicate that two pools are inter-converted by multiple reactions. Sets of metabolite pools that are lumped in the ^{13}C-MFA model are highlighted in gray. **(B)** Representation of the network for a related ^{13}C-MFA model (Schwender et al., 2006) showing the carbon backbones of the metabolites. Biochemical reactions are carbon transitions as connecting arrows (double-headed arrows for reversible reactions). For succinate (Succ), the symmetric randomization of carbon atoms is indicated that is accomplished in the model by two reactions (e.g., KDHa: aIKG$_{(\#ABCDE)}$ > Succ$_{(\#ABCD)}$ + CO2$_{(\#E)}$; KDHb: aIKG$_{(\#ABCDE)}$ > Succ$_{(\#DCBA)}$ + CO2$_{(\#E)}$) occurring at the same rate. Metabolites with measured ^{13}C-label signatures are boxed. Metabolite abbreviations: AcCoA, acetyl coenzyme A; aIKG, α-ketoglutarate; Cit, citrate; Fum, fumarate; Icl, isocitrate; Mal, malate; OxA, oxaloacetate; Pyr, pyruvate; Succ, succinate; SuccCoA, succinyl coenzymeA.

DEFINING THE MODEL BOUNDARIES

A ^{13}C-MFA experiment allows to explore the distribution of *in vivo* flux under a particular physiological condition. For all organic substrates present in the medium, such as sucrose or glutamine, their uptake reactions must be defined. Furthermore, by quantitatively breaking down the cell components, we can identify the most abundant compounds to result from biosynthetic fluxes (**Figure 1**). This approach defines several biomass drain fluxes that are responsible for cell growth. Typically neglected are the growth demands for synthesizing a multitude of low-abundance free intermediary metabolites, as well as enzyme cofactors, pigments, and phytohormones. The inclusion of such minor compounds into the metabolic network would not significantly affect the flux distribution in central metabolism. Finally, measurements of growth kinetics can serve to scale the model fluxes relative to a specific growth rate.

ENCODING BIOCHEMICAL REACTIONS

In ^{13}C-MFA all reaction stoichiometries must be augmented by carbon transitions. Any particular biochemical reaction may be formulated as a set of carbon-atom transitions, defining how each one moves between the main substrates and products. For example, a textual notation following the style of Wiechert and de Graaf (1996) for the carboxylation of phosphoenol pyruvate (PEP) is

$$\mathrm{PEP}_{(\#\,\mathrm{ABC})} + \mathrm{CO2}_{(\#a)} > \mathrm{oxaloacetate}_{(\#\,\mathrm{ABCa})}$$

with A, B, and C respectively denoting the carbons one, two, and three of PEP being converted into carbons one, two, and three of OxA. The carbon chain #ABC joins with #a (CO_2), becoming carbon four of OxA (**Figure 2B**). Co-substrates, such as ATP, phosphate, or H_2O are not considered; hence, both PEP carboxylase (EC 4.1.1.31) and PEP carboxykinase (EC 4.1.1.32) would be encoded by the above equation. In addition to carbon transitions, we must decide if the reaction is a unidirectional- or bidirectional-one. Several plant models define the above reaction as unidirectional, assuming it to be PEP carboxylase, which reportedly is unidirectional. Such reactions with very large standard enthalpy can safely be assumed to be unidirectional in any organism under any condition. Yet, for reactions with smaller standard enthalpy a highly reliable definition of a reaction's directionality would require knowledge of organism- or tissue-specific *in vivo* concentrations of all enzyme substrates (Heinrich and Schuster, 1996).

ESSENTIAL COMPUTATIONAL ASPECTS OF ^{13}C-MFA

Based on the definition of all reactions in the network outlined above, the modeling framework 13CFLUX (Wiechert et al., 2001) automatically generates the necessary equation systems to simulate the distribution of the ^{13}C-label in the network. Labeling state variables can be encoded as the relative abundance of isotope isomers (isotopomers; Schmidt et al., 1997). Accordingly, in the above example, PEP would be represented by the fractional abundances of the eight isotopomer species #000, #100, #010, #001, #110, #011, #101, and #111 (with "1" denoting ^{13}C, and "0" denoting ^{12}C). Recently, the efficiency of computations increased, based on derived concepts like *cum*ulated isotop*omers* (cumomers; Wiechert, 2001), bond isomers (bondomers; van Winden et al.,

2002; Sriram et al., 2007), and elementary metabolic units (EMUs; Antoniewicz et al., 2007). The network examples in **Table 1** range between about 2000- and 4000-simulated cumomer species, while the EMU approach reduces by several-fold the number of labeling state variables (**Table 1**); in future, this should support the simulation of much larger networks.

The goal of the computational analysis is to determine the unknown fluxes that best explain the experimental data. While various studies can differ considerably in their details, a general outline is given here. To determine the fluxes in the system, we use a search algorithm (iterative least-squares fitting procedure). Starting from an initial guess, a set of flux values is sought wherein the fluxes and labeling signatures predicted by the model are the closest to the experimentally determined flux and labeling measurements. Various studies repeated this optimization process between about 200- and 1000-times (Allen et al., 2009b; Lonien and Schwender, 2009; Masakapalli et al., 2010) to assure an adequate exploration of the possible existence of alternative solutions. In addition, the search algorithm might converge repeatedly to flux values that represent only a local optimum. The more often the search is done, the more confident can the modeler be that the global optimum solution is uncovered. For the network definition process discussed below, it is important to note that the required computation time per optimization run increases with the network's size and complexity, as does the computational effort necessary to analyze a solution space of increasing complexity (alternative optima). For modeling four genotypes (model variants) of a larger network (Lonien and Schwender, 2009), access to cluster computing has proven indispensable.

Finally, an important part of computational analysis is to determine the statistical confidence in the obtained flux values (statistical analysis). Flux values that are not obtained consistently by multiple optimizations and have large statistical uncertainty are called "not resolved." Additionally, to assess how strong is the foundation of the flux results in the experimental data, we can determine in a sensitivity analysis how the values for optimal flux depend on the label measurements and other model parameters. By further considering the choice of substrate labeling, experiments can be optimized to resolve particular fluxes of interest (experimental design; Libourel et al., 2007).

NETWORK DEFINITION AND VALIDATION IN ^{13}C-MFA

The core of a formulated network in ^{13}C-MFA typically consists of the reactions in glycolysis, the pentose-phosphate pathway and Calvin cycle, the TCA cycle, and in anaplerosis. This formulation could be based on a consensus from the biochemical literature on the plant's central metabolic pathways. For example, while the presence of mitochondrial- and plastidic-isoforms of pyruvate dehydrogenase in higher plants is well established (Tovar-Méndez et al., 2003), including a cytosolic isoform in a model would be unrealistic unless there was good evidence from the particular plant being studied. Beyond a consensus, experimental evidence in the literature or databases about a particular plant species and cell type typically also is considered.

A key to resolving fluxes of pathways organized in parallel in different compartments is to obtain compartment-specific labeling information. For example, Val, Leu, and Ile are formed from pyruvate in the plastid, i.e., their carbon chains store label information of plastidic pyruvate (Singh and Shaner, 1995). Furthermore, protein-bound Asp, Ala, and Glu, respectively, are assumed to represent cytosolic oxaloacetate, pyruvate, and α-ketoglutarate, provided that the following two assumptions are valid: (1) Asp, Ala, and Glu in the cytosol are isotopically equilibrated with their respective corresponding α-ketocarboxylic acids due to the high activity of the reversible aminotransferases, and, (2) Most of the analyzed protein is synthesized from cytosolic amino acids, i.e., in the analyzed biomass only very small fractions of proteins originated from plastidic or mitochondrial protein synthesis.

In ^{13}C-MFA models, the intrinsic complexity of the metabolic network often is reduced extensively by lumping metabolic pools (van Winden et al., 2001b), as demonstrated for the highly connected sub-network of the TCA cycle in *B. napus* (**Figure 2**). Pools of OxA and malate (Mal), localized in the cytosol and mitochondria (**Figure 2A**), are lumped into one pool (Oxa/Mal, **Figure 2B**). This combination was justified mainly by observations made in labeling signatures in Asp, derived from storage protein (Schwender et al., 2006). Symmetries in the labeling pattern suggested that OxA, after its derivation from the carboxylation of PEP in the cytosol, undergoes a randomization, attributed to the symmetry in succinate (Succ) and fumarate (Fum; **Figure 2B**). Therefore, the equilibration of the carbon-labeling signatures of the C-4 carboxylic acids OxA, Mal, Succ, and Fum supposedly reflects the large fluxes that inter-convert those pools within cytosol and mitochondria, and across the mitochondrial membrane (**Figure 2A**). Therefore, for the ^{13}C-MFA model, the complexity of the C-4 carboxylic acids inter-conversions was reduced by defining two lumped pools, i.e., OxA/Mal and Succ/Fum, and condensing the various reversible inter-conversions (**Figure 2A**) into one reversible reaction (vFM, **Figure 2B**). The consequence of this network reduction is that the net and exchange flux of vFM can be determined with good precision, although the parallel reactions in the cytosol and mitochondria cannot be resolved.

Typically, the modeling process also considers whether adding or removing particular reactions in an existing model might generate a model that better fits the labeling measurements (Schwender et al., 2006; Williams et al., 2008; Lonien and Schwender, 2009; Masakapalli et al., 2010). For example, the isocitrate dehydrogenase reaction is often considered unidirectional from citrate to α-ketoglutarate. Yet, in *Brassica napus* (rapeseed) and soy embryos, the labeling pattern in citrate is explained only if the model also allows for conversion of α-ketoglutarate back to citrate (Schwender et al., 2006; Allen et al., 2009b). This finding showed that, in contrast to the common assumption in the literature, the isocitrate dehydrogenase reaction (**Figure 2B**) must be reversible *in vivo*. Other observations on labeling signatures in *B. napus* justified the assumption that the conversions of PEP to OxA, or PEP to Pyr are *in vivo* irreversible (Schwender et al., 2006).

In conclusion, the topology of published ^{13}C-MFA networks often reflects several assumptions and circumstantial experimental evidence used to justify using lumped networks. Often the underlying unreduced (large-scale) network, and the reduction process are not documented fully and transparently. Lumped metabolic models might depend in part on intuition, and only somewhat result from a transparent process to reduce network complexity.

Yet, in ^{13}C-MFA, the resulting values for flux and their interpretation critically depend upon the network's topology (van Winden et al., 2001a). In addition, once flux results are obtained, projecting the lumped metabolic models on to large-scale models involves substantial ambiguity. This means that mapping fluxes to pathways from pathway databases is problematic.

Generally, more organized, transparent, and reproducible workflows might improve model reconstruction; this is a major topic in other fields of biological computational research (Dalman et al., 2010; Goecks et al., 2010; Mesirov, 2010). With this in mind, we can employ some recently published genome-scale plant metabolic models used for FBA (**Table 1**; Poolman et al., 2009; de Oliveira Dal'Molin et al., 2010a,b; Williams et al., 2010; Saha et al., 2011) as a reference for a more unbiased and more clearly defined network reconstruction in ^{13}C-MFA. Yet, although the genome-scale networks claim to be unbiased representations of the whole genome (Covert et al., 2001), they suffer from incompleteness and from the limited accuracy of gene annotation; certainly, for eukaryotes (plants) they reflect the very limited availability and reliability of predictions of the subcellular localization of the gene products (Poolman et al., 2006; Sweetlove et al., 2008). A particular problem arising in deriving compartmentalized networks is that many of the intracellular transporters functionally required remain unidentified and uncharacterized. Also, there is the ambiguous affinity of many of the known transporters to different substrates of similar structure (Linka and Weber, 2010). Furthermore, if a whole plant-genome is the template for network reconstruction, the result must be a generalized network rather than a network specific for a certain cell type. In addition, despite the recent comprehensive atom mapping of an *E. coli* genome-scale model (Ravikirthi et al., 2011), the carbon transitions in such large plant networks cannot yet be straightforwardly derived from databases.

Consequently, deriving reliable networks from plant-genome databases should require an enormous amount of manual curation. Alternatively, more useful may be the well- documented "bottom-up" reconstructions of large-scale plant models based on published biochemical- and tissue-specific-evidence (**Table 1**; Grafahrend-Belau et al., 2009; Hay and Schwender, 2011a,b). These models might be developed into large-scale ^{13}C-MFA models. While current ^{13}C-MFA models encompass between ∼50 and 100 reactions (**Table 1**), Suthers et al. (2007, 2010) modeled a large-scale *E. coli* network with 238 reactions. Recent advances in the mathematical formulation of isotope models, like the simulation of EMU support the representation of such networks with substantially less computation time than presently required (Antoniewicz et al., 2007). If large-scale plant ^{13}C-MFA models are to be simulated, certain aspects must be dealt with as detailed for the *E. coli* large-scale ^{13}C-MFA model (Suthers et al., 2007). No single optimal flux solution is obtained, and a complex analysis of the solution space is necessary, implying that, for many fluxes, a range of optimum values will be obtained rather than a discrete one. This problem can be attributed largely to parallel pathways that produce redundant labeling patterns and cannot be resolved. Some redundant solutions involving parallel pathways can contain substrate cycles that expectedly are of little biological relevance; thus Suthers et al. (2007) suggested a multi-step reduction in network size. They verified that each time metabolic pools are merged or a parallel pathway is removed, the model fit is not worsened, i.e., simplifying the model does not introduce bias. This kind of approach could replace the more intuitive "classical" model definition of lumped ^{13}C-MFA networks.

A further improvement of the definition of large-scale metabolic networks could lie in using quantitative analysis of the transcriptome by deep-sequencing technologies (RNA-seq; Wang et al., 2009). This technology requires having a genome sequence but should assure a more precise definition about which gene products are present in a particular cell type under specific conditions. The definition of central core metabolism would be improved, in particular since the subcellular localization of core metabolism enzymes can differ between cell types or species. For example, phosphoglyceromutase is only present in plastids of certain cell types (Stitt and ap Rees, 1979). The subcellular localization of ADP-Glucose Pyrophosphorylase differs between gramineous and non-gramineous species (Beckles et al., 2001).

CONCLUSION

In plant-specific ^{13}C-MFA studies published to date lumped network topologies are required. These networks represent a substantial simplification relative to the real complexity inherent to plant central metabolism. Often the validity of network simplifications has to be justified by vague assumptions or circumstantial experimental evidence. Constructing large-scale metabolic models can provide fully detailed networks, useful as a clearly defined reference point for deriving lumped ^{13}C models. Moreover, without lumping, ^{13}C-MFA with plant models of about 500 reactions in size should become computationally feasible, as indicated by recent microbial studies using large-scale ^{13}C-MFA (Suthers et al., 2007, 2010).

The large-scale reference models also offer the potential to develop approaches that combine FBA with ^{13}C-MFA (Blank et al., 2005). Some explorations of the synergies between the two approaches were reported (Williams et al., 2010; Chen et al., 2011; Hay and Schwender, 2011b). With FBA, different physiological conditions can be simulated *in silico* to analyze situations in which steady-state ^{13}C-tracer experiments are impossible.

Another important goal in plant ^{13}C-MFA is to improve the precision of the flux estimates. This can be achieved by simulation of different experiments with differently ^{13}C-labeled tracers in one flux model (Schwender et al., 2006; Alonso et al., 2007b, 2011; Junker et al., 2007; Masakapalli et al., 2010).

Furthermore, analysis of how the distribution of cellular flux changes in response to targeted perturbations can help to unravel the kinetic- and regulatory-controls in metabolism (Lonien and Schwender, 2009). Such approaches should be particularly promising if for experimental systems that have been well established for ^{13}C-MFA, metabolomic, transcriptomic, and proteomic data are recorded in parallel.

ACKNOWLEDGMENTS

Current funding from the U.S. Department of Energy (Division of Chemical Sciences, Geosciences, and Biosciences, Office of Basic Energy Sciences, Field Work Proposal BO-133) as well as by Bayer Bioscience is much appreciated. I like to thank Avril Woodhead (Brookhaven National Laboratory) for English language edits.

REFERENCES

Allen, D. K., Libourel, I. G. L., and Shachar-Hill, Y. (2009a). Metabolic flux analysis in plants: coping with complexity. *Plant Cell Environ.* 32, 1241–1257.

Allen, D. K., Ohlrogge, J. B., and Shachar-Hill, Y. (2009b). The role of light in soybean seed filling metabolism. *Plant J.* 58, 220–234.

Allen, D. K., and Ratcliffe, R. G. (2009). "Quantification of isotope label," in *Plant Metabolic Networks,* ed. J. Schwender (New York: Springer), 105–149.

Alonso, A. P., Dale, V. L., and Shachar-Hill, Y. (2010). Understanding fatty acid synthesis in developing maize embryos using metabolic flux analysis. *Metab. Eng.* 12, 488–497.

Alonso, A. P., Goffman, F. D., Ohlrogge, J. B., and Shachar-Hill, Y. (2007a). Carbon conversion efficiency and central metabolic fluxes in developing sunflower (*Helianthus annuus* L.) embryos. *Plant J.* 52, 296–308.

Alonso, A. P., Raymond, P., Hernould, M., Rondeau-Mouro, C., De Graaf, A., Chourey, P., Lahaye, M., Shachar-Hill, Y., Rolin, D., and Dieuaide-Noubhani, M. (2007b). A metabolic flux analysis to study the role of sucrose synthase in the regulation of the carbon partitioning in central metabolism in maize root tips. *Metab. Eng.* 9, 419–432.

Alonso, A. P., Raymond, P., Rolin, D., and Dieuaide-Noubhani, M. (2007c). Substrate cycles in the central metabolism of maize root tips under hypoxia. *Phytochemistry* 68, 2222–2231.

Alonso, A. P., Val, D. L., and Shachar-Hill, Y. (2011). Central metabolic fluxes in the endosperm of developing maize seeds and their implications for metabolic engineering. *Metab. Eng.* 13, 96–107.

Alonso, A. P., Vigeolas, H., Raymond, P., Rolin, D., and Dieuaide-Noubhani, M. (2005). A new substrate cycle in plants. Evidence for a high glucose-phosphate-to-glucose turnover from in vivo steady-state and pulse-labeling experiments with [13C]glucose and [14C]glucose. *Plant Physiol.* 138, 2220–2232.

Antoniewicz, M. R., Kelleher, J. K., and Stephanopoulos, G. (2007). Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metab. Eng.* 9, 68–86.

Beckles, D. M., Smith, A. M., and ap Rees, T. (2001). A cytosolic ADP-glucose pyrophosphorylase is a feature of graminaceous endosperms, but not of other starch-storing organs. *Plant Physiol.* 125, 818–827.

Blank, L. M., Kuepfer, L., and Sauer, U. (2005). Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biol.* 6, R49.

Chen, X., Alonso, A. P., Allen, D. K., Reed, J. L., and Shachar-Hill, Y. (2011). Synergy between 13C-metabolic flux analysis and flux balance analysis for understanding metabolic adaptation to anaerobiosis in *E. coli*. *Metab. Eng.* 13, 38–48.

Covert, M. W., Schilling, C. H., Famili, I., Edwards, J. S., Goryanin Ii Selkov, E., and Palsson, B. O. (2001). Metabolic modeling of microbial strains in silico. *Trends Biochem. Sci.* 26, 179–186.

Dalman, T., Droste, P., Weitzel, M., Wiechert, W., and Nöh, K. (2010). "Workflows for metabolic flux analysis: data integration and human interaction," in *Proceedings of the 4th International Conference on Leveraging Applications of Formal Methods, Verification, and Validation – Volume Part I* (Heraklion: Springer-Verlag).

Dauner, M., and Sauer, U. (2000). GC-MS analysis of amino acids rapidly provides rich information for isotopomer balancing. *Biotechnol. Prog.* 16, 642–649.

de Oliveira Dal'Molin, C. G., Quek, L. E., Palfreyman, R. W., Brumbley, S. M., and Nielsen, L. K. (2010a). AraGEM, a genome-scale reconstruction of the primary metabolic network in *Arabidopsis*. *Plant Physiol.* 152, 579–589.

de Oliveira Dal'Molin, C. G., Quek, L. E., Palfreyman, R. W., Brumbley, S. M., and Nielsen, L. K. (2010b). C4GEM, a genome-scale metabolic model to study C4 plant metabolism. *Plant Physiol.* 154, 1871–1885.

Dieuaide-Noubhani, M., Raffard, G., Canioni, P., Pradet, A., and Raymond, P. (1995). Quantification of compartmented metabolic fluxes in maize root tips using isotope distribution from 13C- or 14C-labeled glucose. *J. Biol. Chem.* 270, 13147–13159.

Ettenhuber, C., Spielbauer, G., Margl, L., Hannah, L. C., Gierl, A., Bacher, A., Genschel, U., and Eisenreich, W. (2005). Changes in flux pattern of the central carbohydrate metabolism during kernel development in maize. *Phytochemistry* 66, 2632–2642.

Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11, R86.

Grafahrend-Belau, E., Schreiber, F., Koschützki, D., and Junker, B. H. (2009). Flux balance analysis of barley seeds: a computational approach to study systemic properties of central metabolism. *Plant Physiol.* 149, 585–598.

Hay, J. O., and Schwender, J. (2011a). Metabolic network reconstruction and flux variability analysis of storage synthesis in developing oilseed rape (*Brassica napus* L.) embryos. *Plant J.* 67, 526–541.

Hay, J. O., and Schwender, J. (2011b). Computational analysis of storage synthesis in developing *Brassica napus L.* (oilseed rape) embryos: flux variability analysis in relation to ^{13}C metabolic flux analysis. *Plant J.* 67, 513–525.

Heinrich, R., and Schuster, S. (1996). *The Regulation of Cellular Systems.* New York: Chapman and Hall.

Iyer, V. V., Sriram, G., Fulton, D. B., Zhou, R., Westgate, M. E., and Shanks, J. V. (2008). Metabolic flux maps comparing the effect of temperature on protein and oil biosynthesis in developing soybean cotyledons. *Plant Cell Environ.* 31, 506–517.

Junker, B. H., Lonien, J., Heady, L. E., Rogers, A., and Schwender, J. (2007). Parallel determination of enzyme activities and in vivo fluxes in *Brassica napus* embryos grown on organic or inorganic nitrogen source. *Phytochemistry* 68, 2232–2242.

Koschutzki, D., Junker, B. H., Schwender, J., and Schreiber, F. (2010). Structural analysis of metabolic networks based on flux centrality. *J. Theor. Biol.* 265, 261–269.

Kruger, N. J., and Ratcliffe, R. G. (2009). Insights into plant metabolic networks from steady-state metabolic flux analysis. *Biochimie* 91, 697–702.

Libourel, I. G., Gehan, J. P., and Shachar-Hill, Y. (2007). Design of substrate label for steady state flux measurements in plant systems using the metabolic network of *Brassica napus* embryos. *Phytochemistry* 68, 2211–2221.

Libourel, I. G. L., and Shachar-Hill, Y. (2008). Metabolic flux analysis in plants: from intelligent design to rational engineering. *Annu. Rev. Plant Biol.* 59, 625–650.

Linka, N., and Weber, A. P. (2010). Intracellular metabolite transporters in plants. *Mol. Plant* 3, 21–53.

Llaneras, F., and Pico, J. (2008). Stoichiometric modelling of cell metabolism. *J. Biosci. Bioeng.* 105, 1–11.

Lonien, J., and Schwender, J. (2009). Analysis of metabolic flux phenotypes for two *Arabidopsis* mutants with severe impairment in seed storage lipid synthesis. *Plant Physiol.* 151, 1617–1634.

Masakapalli, S. K., Le Lay, P., Huddleston, J. E., Pollock, N. L., Kruger, N., and Ratcliffe, R. G. (2010). Subcellular flux analysis of central metabolism in a heterotrophic *Arabidopsis* cell suspension using steady-state stable isotope labeling. *Plant Physiol.* 152, 602–619.

Mesirov, J. P. (2010). Computer science. Accessible reproducible research. *Science* 327, 415–416.

Poolman, M. G., Bonde, B. K., Gevorgyan, A., Patel, H. H., and Fell, D. A. (2006). Challenges to be faced in the reconstruction of metabolic networks from public databases. *Syst. Biol. (Stevenage)* 153, 379–384.

Poolman, M. G., Miguet, L., Sweetlove, L. J., and Fell, D. A. (2009). A genome-scale metabolic model of *Arabidopsis* and some of its properties. *Plant Physiol.* 151, 1570–1581.

Ravikirthi, P., Suthers, P. F., and Maranas, C. D. (2011). Construction of an *E. Coli* genome-scale atom mapping model for MFA calculations. *Biotechnol. Bioeng.* 108, 1372–1382.

Rontein, D., Dieuaide-Noubhani, M., Dufourc, E. J., Raymond, P., and Rolin, D. (2002). The metabolic architecture of plant cells. Stability of central metabolism and flexibility of anabolic pathways during the growth cycle of tomato cells. *J. Biol. Chem.* 277, 43948–43960.

Saha, R., Suthers, P. F., and Maranas, C. D. (2011). *Zea mays* iRS1563: a comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLoS ONE* 6, e21784. doi:10.1371/journal.pone.0021784

Schmidt, K., Carlsen, M., Nielsen, J., and Villadsen, J. (1997). Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices. *Biotechnol. Bioeng.* 55, 831–840.

Schwender, J. (2008). Metabolic flux analysis as a tool in metabolic engineering of plants. *Curr. Opin. Biotechnol.* 19, 131–137.

Schwender, J. (2009). "Isotopic steady-state flux analysis," in *Plant Metabolic Networks,* ed. J. Schwender (New York: Springer), 245–284.

Schwender, J., Goffman, F., Ohlrogge, J. B., and Shachar-Hill, Y. (2004a). Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. *Nature* 432, 779–782.

Schwender, J., Ohlrogge, J., and Shachar-Hill, Y. (2004b). Understanding flux in plant metabolic networks. *Curr. Opin. Plant Biol.* 7, 309–317.

Schwender, J., and Ohlrogge, J. B. (2002). Probing in vivo metabolism by stable isotope labeling of storage lipids and proteins in developing *Brassica napus* embryos. *Plant Physiol.* 130, 347–361.

Schwender, J., Ohlrogge, J. B., and Shachar-Hill, Y. (2003). A flux model of glycolysis and the oxidative pentose phosphate pathway in developing *Brassica napus* embryos. *J. Biol. Chem.* 278, 29442–29453.

Schwender, J., Shachar-Hill, Y., and Ohlrogge, J. B. (2006). Mitochondrial metabolism in developing embryos of *Brassica napus*. *J. Biol. Chem.* 281, 34040–34047.

Singh, B. K., and Shaner, D. L. (1995). Biosynthesis of branched chain amino acids: from test tube to field. *Plant Cell* 7, 935–944.

Spielbauer, G., Margl, L., Hannah, L. C., Romisch, W., Ettenhuber, C., Bacher, A., Gierl, A., Eisenreich, W., and Genschel, U. (2006). Robustness of central carbohydrate metabolism in developing maize kernels. *Phytochemistry* 67, 1460–1475.

Sriram, G., Fulton, D. B., Iyer, V. V., Peterson, J. M., Zhou, R., Westgate, M. E., Spalding, M. H., and Shanks, J. V. (2004). Quantification of compartmented metabolic fluxes in developing soybean embryos by employing biosynthetically directed fractional 13C labeling, two-dimensional [13C, 1H] nuclear magnetic resonance, and comprehensive isotopomer balancing. *Plant Physiol.* 136, 3043–3057.

Sriram, G., Fulton, D. B., and Shanks, J. V. (2007). Flux quantification in central carbon metabolism of *Catharanthus roseus* hairy roots by 13C labeling and comprehensive bondomer balancing. *Phytochemistry* 68, 2243–2257.

Stitt, M., and ap Rees, T. (1979). Capacities of pea chloroplasts to catalyze the oxidative pentose phosphate pathway and glycolysis. *Phytochemistry* 18, 1905–1911.

Suthers, P. F., Burgard, A. P., Dasika, M. S., Nowroozi, F., Van Dien, S., Keasling, J. D., and Maranas, C. D. (2007). Metabolic flux elucidation for large-scale models using 13C labeled isotopes. *Metab. Eng.* 9, 387–405.

Suthers, P. F., Chang, Y. J., and Maranas, C. D. (2010). Improved computational performance of MFA using elementary metabolite units and flux coupling. *Metab. Eng.* 12, 123–128.

Sweetlove, L. J., Fell, D., and Fernie, A. R. (2008). Getting to grips with the plant metabolic network. *Biochem. J.* 409, 27–41.

Szyperski, T. (1995). Biosynthetically directed fractional 13C-labeling of proteinogenic amino acids. An efficient analytical tool to investigate intermediary metabolism. *Eur. J. Biochem.* 232, 433–448.

Szyperski, T. (1998). 13C-NMR, MS and metabolic flux balancing in biotechnology research. *Q. Rev. Biophys.* 31, 41–106.

Tovar-Méndez, A., Miernyk, J. A., and Randall, D. D. (2003). Regulation of pyruvate dehydrogenase complex activity in plant cells. *Eur. J. Biochem.* 270, 1043–1049.

van Winden, W., Verheijen, P., and Heijnen, S. (2001a). Possible pitfalls of flux calculations based on 13C-labeling. *Metab. Eng.* 3, 151–162.

van Winden, W. A., Heijnen, J. J., Verheijen, P. J., and Grievink, J. (2001b). A priori analysis of metabolic flux identifiability from (13)C-labeling data. *Biotechnol. Bioeng.* 74, 505–516.

van Winden, W. A., Heijnen, J. J., and Verheijen, P. J. (2002). Cumulative bondomers: a new concept in flux analysis from 2D [13C,1H] COSY NMR data. *Biotechnol. Bioeng.* 80, 731–745.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63.

Wiechert, W. (2001). 13C metabolic flux analysis. *Metab. Eng.* 3, 195–206.

Wiechert, W., and de Graaf, A. A. (1996). In vivo stationary flux analysis by 13C labeling experiments. *Adv. Biochem. Eng. Biotechnol.* 54, 109–154.

Wiechert, W., Möllney, M., Petersen, S., and De Graaf, A. A. (2001). A universal framework for 13C metabolic flux analysis. *Metab. Eng.* 3, 265–283.

Williams, T. C., Miguet, L., Masakapalli, S. K., Kruger, N. J., Sweetlove, L. J., and Ratcliffe, R. G. (2008). Metabolic network fluxes in heterotrophic *Arabidopsis* cells: stability of the flux distribution under different oxygenation conditions. *Plant Physiol.* 148, 704–718.

Williams, T. C., Poolman, M. G., Howden, A. J., Schwarzlander, M., Fell, D. A., Ratcliffe, R. G., and Sweetlove, L. J. (2010). A genome-scale metabolic model accurately predicts fluxes in central carbon metabolism under stress conditions. *Plant Physiol.* 154, 311–323.

Zamboni, N., Fendt, S. M., Ruhl, M., and Sauer, U. (2009). (13)C-based metabolic flux analysis. *Nat. Protoc.* 4, 878–892.

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

Received: 16 May 2011; accepted: 14 September 2011; published online: 10 October 2011.

Citation: Schwender J (2011) Experimental flux measurements on a network scale. Front. Plant Sci. ***2****:63. doi: 10.3389/fpls.2011.00063*

This article was submitted to Frontiers in Plant Physiology, a specialty of Frontiers in Plant Science.

frontiers in
PLANT SCIENCE

REVIEW ARTICLE
published: 11 August 2011
doi: 10.3389/fpls.2011.00038

Flux-balance modeling of plant metabolism

Lee J. Sweetlove* and R. George Ratcliffe

Department of Plant Sciences, University of Oxford, Oxford, UK

Edited by:
Alisdair Fernie, Max Planck Institute for Plant Physiology, Germany

Reviewed by:
Jörg Schwender, Brookhaven National Laboratory, USA
John Morgan, Purdue University, USA

****Correspondence:***
Lee J. Sweetlove, Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK.
e-mail: lee.sweetlove@plants.ox.ac.uk

Flux-balance modeling of plant metabolic networks provides an important complement to ^{13}C-based metabolic flux analysis. Flux-balance modeling is a constraints-based approach in which steady-state fluxes in a metabolic network are predicted by using optimization algorithms within an experimentally bounded solution space. In the last 2 years several flux-balance models of plant metabolism have been published including genome-scale models of *Arabidopsis* metabolism. In this review we consider what has been learnt from these models. In addition, we consider the limitations of flux-balance modeling and identify the main challenges to generating improved and more detailed models of plant metabolism at tissue- and cell-specific scales. Finally we discuss the types of question that flux-balance modeling is well suited to address and its potential role in metabolic engineering and crop improvement.

Keywords: flux-balance modeling, metabolism, flux

INTRODUCTION

Metabolism is a prerequisite for life. Hundreds of chemical reactions, mostly catalyzed by enzymes, define a metabolic network that supports all biological activity. In particular, the coupling of energy-releasing processes to energy-consuming anabolic reactions drives the biosynthesis of the polymers and metabolites that constitute the fabric of the cell. The rates of all the enzyme-catalyzed reactions, including the associated relocation of ions and metabolites across membranes, are tightly controlled through the regulation of enzyme activity, allowing metabolic outputs to be adjusted according to varying environmental conditions and growth patterns.

Plant metabolic networks are arguably the most complex of any organism, both because of the tremendous variation in their metabolic output and because of the range of environmental conditions that they encounter. Nevertheless, because the growth and survival of plants is intimately connected to metabolism (Smith and Stitt, 2007; Stitt et al., 2010) there is a need to understand and predict metabolic behavior. In particular, there is a need to connect genotype to specific metabolic outputs so that plant breeders and metabolic engineers can generate new varieties of crops with increased yield or altered chemical composition (Fernie and Schauer, 2009).

Although there have been some notable successes in engineering plant metabolism (Butelli et al., 2008; Naqvi et al., 2009), these are mainly related to the production of secondary metabolites. In contrast, there are few examples where the synthesis of the main biomass polymers has been manipulated in a predictive manner. At the heart of this contrasting ability to engineer the metabolic network is the difference in connectivity of primary and secondary metabolism. Many secondary metabolites are synthesized by reactions that occur at the periphery of the metabolic network with relatively few interconnections to other parts of the network. As a result, there are fewer regulatory constraints on the flux through these essentially linear pathways and control may be disproportionately resident in a single enzyme, providing a single target for genetic manipulation (Fraser et al., 2002; Enfissi et al., 2005). Where control is more evenly distributed, it is often possible to identify transcription factors that co-ordinately affect the expression of genes encoding enzymes in a pathway (Schwinn et al., 2006; Memelink and Gantet, 2007). In contrast, the main growth components of the cell are synthesized through a highly connected set of reactions that is commonly referred to as "central metabolism." Not only is control of flux generally shared amongst many, if not all, enzymes of central metabolism (Raines, 2003; Geigenberger et al., 2004; Araujo et al., 2011), but because of the degree of connectivity, a perturbation of one part of the network has consequences for other parts of the network.

So in central metabolism, the functioning of individual enzymes and even pathways is dependent on the operational state of the whole metabolic network (Kruger and Ratcliffe, 2008; Sweetlove et al., 2008). As a result, efforts to understand the regulation of the central metabolic network in the last few years have focused on measuring the metabolic phenotype – flux – at a network level. Much of this work has centered on the development of approaches to determine network fluxes based on the steady-state redistribution of isotopically labeled carbon (Libourel and Shachar-Hill, 2008; Allen et al., 2009a; Kruger and Ratcliffe, 2009). This approach, known as steady-state metabolic flux analysis (MFA), has matured into a powerful technique whereby it is possible to reliably quantify both net and exchange fluxes of tens of reactions across the central metabolic network. MFA has been applied to several different plant species and tissue types and has yielded some significant insights into the behavior and organization of plant metabolism. For example, MFA was used to establish that Rubisco can function without the Calvin cycle to recycle carbon lost as CO_2 during lipid synthesis in green oilseeds in the light, substantially increasing the carbon conversion efficiency (Schwender et al., 2004; Allen et al., 2009b). MFA has also revealed a variety of different flux modes in the TCA cycle (Sweetlove et al., 2010) as well as demonstrating the inherent stability of central metabolism to environmental perturbation (Iyer et al., 2008; Williams et al., 2008), and the complex, non-intuitive relationship between fluxes, metabolite levels, and enzyme activities (Junker et al., 2007; Kruger and Ratcliffe, 2009).

However, despite the undoubted power of steady-state MFA for defining the metabolic phenotype, it does have some limitations. In particular, the requirement to supply a labeled organic carbon substrate to isotopic steady-state limits the approach to heterotrophic or mixotrophic tissues in culture. The method is also entirely dependent on the correctness of the user-defined metabolic network because it is often possible to fit the labeling data to more than one network structure with little statistical power to discriminate between the alternatives (Masakapalli et al., 2010). MFA is, moreover, a relatively low throughput technique and this limits its use as a comparative tool since comparison of multiple samples (e.g., different genotypes) requires substantial effort (Lonien and Schwender, 2009).

These limitations have driven the search for alternative, complementary approaches to characterize and explore the plant metabolic network. Following the lead of the microbial field (Borodina and Nielsen, 2005), flux-balance modeling has emerged as an alternative to MFA. Like MFA, flux-balance analysis (FBA) is a constraints-based modeling approach in which steady-state fluxes in a metabolic network are predicted by applying mass-balance constraints to a model of the network based on the matrix of reaction stoichiometries. Typically, simple and easy to measure mass-balance information, such as growth rate, biomass composition, and substrate-consumption rate, is used to place boundaries on the flux solution space (Reed and Palsson, 2003). However, in contrast to MFA, isotopic labeling information is not used. As a result, the network fluxes are underdetermined and a range of feasible flux solutions are obtained that satisfy the constraints. Within this range, flux solutions that are optimal with respect to a specific objective function (such as maximizing growth rate or minimizing substrate consumption) can be identified with optimization algorithms such as linear programming (Edwards and Palsson, 2000).

Several flux-balance models of different plant species have been published in the last 2 years. These include models for *Arabidopsis* (Poolman et al., 2009; de Oliveira Dal'Molin et al., 2010a; Radrich et al., 2010), barley seeds (Grafahrend-Belau et al., 2009), *Brassica napus* seeds (Hay and Schwender, 2011b; Pilalis et al., 2011), maize (Saha et al., 2011), *Chlamydomonas* (Boyle and Morgan, 2009; Cogne et al., 2011), and photoautotrophic bacteria (Knoop et al., 2010; Montagud et al., 2010). The aim of this article is to review what has been learnt from these models, to discuss the advantages and limitations of flux-balance modeling and to look to the future. What insights into plant metabolic networks can we expect to obtain from flux-balance modeling and what are the main challenges for the biologically informative application of flux-balance modeling?

GENOME-SCALE METABOLIC MODELING

One of the main advantages of flux-balance modeling is that it is relatively easy to scale up to cover very large networks. Indeed, metabolic models can be constructed at a genome-scale, using all the reactions catalyzed by the enzymes encoded in an annotated genome. However this remains a non-trivial task: *Arabidopsis* and maize are the only higher plants with genome-scale metabolic models (Poolman et al., 2009; de Oliveira Dal'Molin et al., 2010a; Radrich et al., 2010; Saha et al., 2011) – all the other plant models have been constructed using metabolic databases, biochemical textbooks, and the primary literature, and are essentially confined to the well known pathways of central metabolism. Several problems arise in the construction of metabolic models from genome-annotation databases, including network gaps caused by incomplete or imprecise genome annotation, mass-balance errors caused by reaction stoichiometry errors in the annotation database, and the presence of excess, non-functional reactions. However, working practices and computational approaches are emerging to help deal with such issues (Fell et al., 2010; Henry et al., 2010; Soh and Hatzimanikatis, 2010).

An additional challenge is that genome-annotation databases contain no information about reaction directionality. In smaller models of primary metabolism, it is possible to manually constrain reactions to a defined direction based on standard Gibbs free energy changes (and sometimes the *in planta* concentration of the reaction substrates and products). However, in genome-scale models, reaction directionality is often left unconstrained, with the result that flux solutions may contain thermodynamically infeasible reactions. A comprehensive standard Gibbs free energy of formation database is urgently required for metabolites to allow thermodynamic constraints to be included in genome-scale FBA. However, because experimentally measured free energies are not available for many reactions, theoretical approaches for estimating standard free energies such as the group contribution method (Jankowski et al., 2008) will need to be implemented.

Given the challenges inherent in constructing and analyzing such large models (the current *Arabidopsis* genome-scale models contain around 1500 reactions), it is relevant to ask whether this effort is worthwhile. Indeed, only 232 of the available 1406 reactions in the *Arabidopsis* genome-scale model constructed by Poolman et al. (2009) are required to synthesize the main biomass components and account for maintenance costs of heterotrophic *Arabidopsis*. The model may be genome-scale, but the flux solution is of similar size and considers similar reactions to FBA models of primary metabolism. It is also worth pointing out that most flux-balance models to date consider a similar span of the metabolic network to previous plant MFA models, although due to reaction lumping and network simplification the actual number of reactions in the MFA models is generally considerably lower.

A genome-scale metabolic network is, of course, not a biological reality. It is unlikely that every enzyme is expressed in a single cell type and under a single condition. Much of secondary metabolism, for example, is induced upon abiotic or biotic stress. Nevertheless, a genome-scale metabolic network has significant value as a foundation for investigating condition-specific scenarios. Thus, cell type-specific sub-models can be constructed based on transcriptomic or proteomic datasets (Lewis et al., 2010) although this has not yet been done to any significant degree for plant metabolism. Similarly, with the inclusion of appropriate constraints, it should be possible to model the consequences of the synthesis of a range of secondary metabolites. For example, in a recent genome-scale flux-balance model of maize, lignin metabolism was explicitly included as part of the biomass function (Saha et al., 2011).

THE ISSUE OF MULTIPLE FLUX SOLUTIONS

Although boundaries are imposed on the flux solution space, it will often still be possible to accommodate multiple solutions that satisfy the chosen objective function. Thus, when linear programming

provides a flux solution that minimizes or maximizes a particular objective function, it is not necessarily a unique solution (Lee et al., 2000; Mahadevan and Schilling, 2003). The ability to uniquely define fluxes is dependent both on the structure of the metabolic network and the objective function being considered. For example, the objective function "minimization of total intracellular fluxes" will select metabolic routes that contain the fewest steps since this will result in a lower sum of fluxes. However, the metabolic network may contain equivalent alternative routes for the production of a given metabolite. For example in **Figure 1A**, both routes lead to the conversion of input substrate to output metabolite via an equal number of steps, meaning that there is no basis by which to select one over the other when minimization of total intracellular flux is used as an objective function. Other commonly used objective functions such as maximization of biomass per unit substrate (and the equivalent minimization of substrate consumed per unit biomass produced) which optimize the molar yield of the system would also fail to discriminate between the two routes if they are stoichiometrically equivalent with respect to carbon.

In contrast, if the two routes contain a different number of steps (**Figure 1B**) then the route with the fewest steps will be utilized under the minimization of flux objective function. Other differences between parallel pathways relate to energy production (**Figure 1C**). If the objective function is to maximize ATP yield then the objective function would select the ATP-producing pathway in **Figure 1C**. Another source of alternative solutions can be the presence of substrate cycles (**Figure 1D**). The minimization of flux objective function will eliminate such cycles, but other objective functions, such as maximization of biomass production, will not because substrate cycles do not influence the net flux from input to output. Accordingly the flux through the cycle is not defined by the objective function and it can hold any value. Subcellular compartmentation, especially the presence of equivalent pathways in different compartments (**Figure 1E**), can also lead to alternative flux solutions.

This issue of alternative optima can be dealt with in two ways: either, additional optimization criteria can be applied such that a unique flux solution is reached; or, flux variability can be viewed as a potentially informative aspect of network behavior that can be explicitly quantified and explored. A model of barley seed metabolism took the former approach (Grafahrend-Belau et al., 2009). First a conventional linear optimization was applied (with the objective function to maximize growth) and then a non-linear quadratic optimization was applied using the objective value (growth rate) of the first optimization as an additional constraint to the second optimization (with the objective function to minimize the overall sum of fluxes). This two-step procedure provided a unique solution because of the nature of the quadratic optimization. Similarly, a two-step linear optimization procedure can be used in which a "maximization" objective function (e.g., maximization of biomass) is followed by a "minimization" objective function (e.g., minimization of photon use). This approach led to a unique flux solution in an flux-balance model of photoautotrophic metabolism in *Synechocystis* (Shastri and Morgan, 2005, 2007).

In contrast, a recent FBA analysis of oilseed rape seed metabolism (Hay and Schwender, 2011a,b) made a virtue of flux variability. An explicit analysis of the extent of variability was performed using a linear programming routine based on a secondary minimization and maximization of the flux through each reaction (Mahadevan and Schilling, 2003). In a 572-reaction network of primary metabolism solved by minimization of substrate consumption, it was found that 75 reactions, mainly in the central core of the network, were variable. Flux variability was classified according to the direction and magnitude of the flux change, and it was found that the variability type of 57 reactions altered when different external substrates

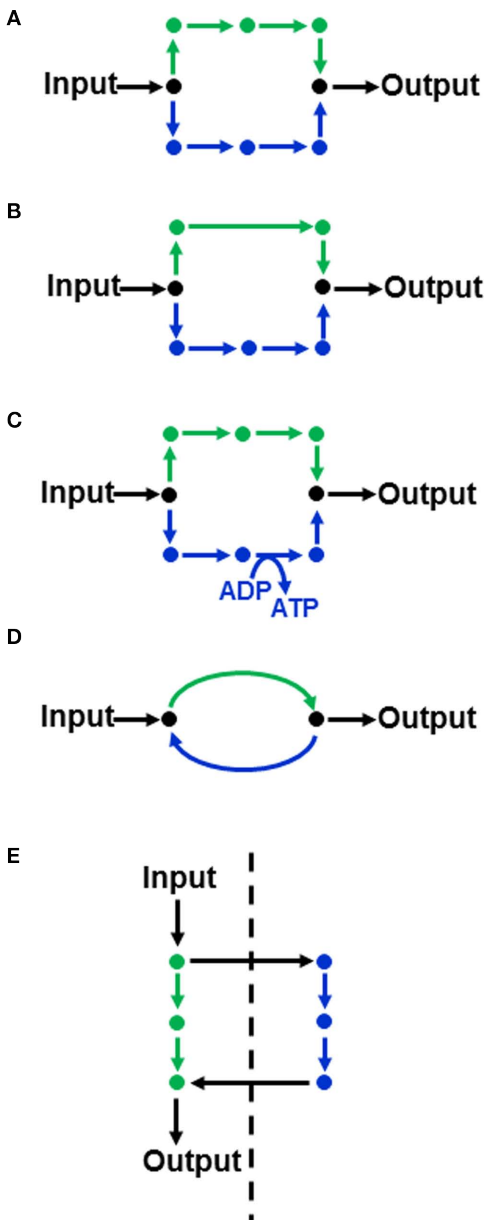


FIGURE 1 | Features of a metabolic network than can lead to flux variability in FBA. (A) Two equivalent routes (shown in green and blue) for converting an input substrate into an output metabolite. **(B,C)** illustrate non-equivalent routes that may be discriminated in FBA, depending on the objective function. **(D)** A substrate cycle. **(E)** Equivalent routes in different subcellular compartments (the dashed line indicating a membrane separating two subcellular compartments).

were used in the model. Flux variability is essentially a modeling issue that arises because the available constraints do not produce a unique solution with the chosen objective function. Nevertheless, changes in variability type can supplement the information that can be deduced from the changes that occur in the fluxes that are uniquely defined. It was also found that 51 reactions varied with infinite bounds and these were largely due to metabolite cycles in which there was no net consumption of carbon or energy. The flux through these reactions can hold any value if the constraints applied only relate to carbon or energy use. Most of the variable fluxes are substitutable, meaning that a solution to the flux optimization problem can be found using alternative reactions. This is a clear demonstration of metabolic redundancy. This work illustrates the utility of flux variability analysis in providing an additional layer of information about the behavior of the network and the nature of the flux solution, and this is particularly valuable when dealing with large networks in which it is not possible to manually inspect the entire flux solution. A similar analysis of flux variability was used to ascertain flux differences between bundle sheath and mesophyll cells in a model of C4 photosynthesis, although in this case, only four reactions were not uniquely defined by the optimality criterion and imposed constraints (de Oliveira Dal'Molin et al., 2010b). To date, only these two studies and a model of *Synechocystis* (Knoop et al., 2010) have explicitly analyzed flux variability in plant flux-balance models, but one would expect it to be a standard component of FBA in future work.

VALIDATION OF FLUX-BALANCE MODELS

When a flux-balance solution is generated it is important to establish how closely it reflects the actual behavior of the metabolic network. One way of doing this is to look for the operation of metabolic pathways that are known to be physiologically important. Thus, when photosynthetic metabolism was modeled in an *Arabidopsis* genome-scale model (de Oliveira Dal'Molin et al., 2010a), the classical photorespiratory cycle was observed to support a flux when a 3:1 ratio of the carboxylation:oxygenation reaction of Rubisco was imposed and the photon-use efficiency was optimized. Moreover, the model predicted that 30–50% of the carbon fixed by photosynthesis would be lost through the photorespiratory cycle, a range consistent with experimental measurements.

However, while the recapitulation of known metabolic behavior is reassuring, FBA is unlikely to give a completely faithful representation of the actual flux distribution, and consideration of areas in which the flux-balance model diverges from known metabolic behavior is potentially more informative. For example, the oxidative reactions of the oxidative pentose phosphate pathway typically carry no flux in flux-balance solutions of heterotrophic plant metabolism (Williams et al., 2010; Hay and Schwender, 2011a). This is in contrast to the known importance of the oxidative pentose phosphate pathway in heterotrophic tissues (Averill et al., 1998) and to the fact that the oxidative reactions carry considerable flux in MFA-based flux maps (Kruger and von Schaewen, 2003; Schwender et al., 2003; Sriram et al., 2004; Alonso et al., 2007, 2010; Masakapalli et al., 2010). The discrepancy arises because in the flux-balance models the provision of NADPH, a likely role for the oxidative pentose phosphate pathway, can be met by plastidial NADP-dependent dehydrogenases, such as glyceraldehyde 3-phosphate dehydrogenase, malate dehydrogenase, and malic enzyme. Production of NADPH by these routes is marginally more efficient in terms of either carbon use or overall flux, and so the models predict that the oxidative branch of the pentose phosphate pathway is not required. Two points are worth making here. First, neither of the flux-balance models include thermodynamic constraints beyond specification of reaction directions, and there may be thermodynamic limitations to the establishment of a high NADPH:NADP ratio using these dehydrogenases. Secondly, the NADPH requirement specified in both models is only that required for synthesis of biomass components. There are several other known NADPH requirements in the cell including antioxidant enzyme activity (Apel and Hirt, 2004) and membrane NADPH oxidase activity (Torres, 2010). Thus, the actual NADPH demand in the cell is certainly higher than that specified in the model and would require NADPH to be produced in different subcellular compartments.

Perhaps the most rigorous way to validate an FBA flux solution is to compare it to fluxes estimated independently by ^{13}C-MFA, and this has been done for both *B. napus* and *Arabidopsis* flux-balance models (Williams et al., 2010; Hay and Schwender, 2011b). Two points need to be borne in mind in making such comparisons. First, the metabolic networks that are used during ^{13}C-MFA are projections of the real network that provide an explanation of the redistribution of ^{13}C that occurs during the steady-state labeling experiment (Roscher et al., 2000). Steps between branch-points are represented as single steps, since the net flux carried by each intermediate step must be the same, and the model is usually constructed to eliminate indeterminable fluxes, i.e., fluxes that cannot be defined from the labeling data and which would therefore show infinite flux variability if included in the model. In contrast, a flux is assigned to every step in the complete network in FBA, even though many of them are the same, and indeterminable fluxes are not grouped to eliminate flux variability. Thus, many reactions in the FBA solution do not have a direct counterpart in the MFA model. It is also simpler to restrict the comparison to reactions that do not show flux variability in the FBA solutions, although statistical comparisons are possible that include a weighting factor to account for flux variability (Schuetz et al., 2007). Secondly, both FBA and MFA constrain the fluxes that lead directly to synthesis of biomass in the same manner, so these reactions will necessarily hold the same values in the FBA and MFA solutions. When these factors are taken into account, only a relatively small number of fluxes can be directly compared between the two approaches (24 reactions in the *Arabidopsis* genome-scale model and 19 reactions in the *B. napus* seed primary metabolism model). Nevertheless, for these few reactions (which are mainly from the core backbone of the network) a reasonable correlation with the predicted fluxes and those estimated by ^{13}C-MFA was found, suggesting that FBA is able to predict realistic values for plant metabolic network fluxes. Moreover, FBA was able to successfully predict changes in flux under two environmental stress conditions in *Arabidopsis* (Williams et al., 2010).

SUBCELLULAR COMPARTMENTATION IN FLUX-BALANCE MODELS

Most of the published flux-balance models of plant metabolism make some attempt to take subcellular compartmentation into account, which is clearly desirable if the model is to reflect

biological reality (Lunn, 2007). However, inclusion of subcellular compartmentation is problematic and at present, there is insufficient information to assess whether inclusion of compartmentation improves the models. One of the biggest problems is how to place reactions in the correct compartment. While the compartmentation of the core pathways of central metabolism is well established, this only accounts for a small percentage of a genome-scale network. Similarly, while databases such as SUBA (Heazlewood et al., 2007) are excellent inventories of subcellular compartmentation supported by experimental evidence mainly drawn from organellar proteome studies, they only represent a relatively small proportion of the metabolic network. Ideally subcellular location of reactions would be assigned automatically in a genome-scale model, perhaps on the basis of predicted protein sequences, but current algorithms are too unreliable (Heazlewood et al., 2004) and there is currently no alternative to manual curation.

Thus, the assignment of subcellular compartmentation is usually done manually. As a result, particularly in genome-scale models, the extent of compartmentation is patchy and may contain errors. For example, in the AraGEM genome-scale model of *Arabidopsis* metabolism (de Oliveira Dal'Molin et al., 2010a), the vast majority of reactions are assigned to the cytosol (1265 reactions in the cytosol, with 60, 159, and 98 reactions assigned to mitochondria, plastid, and peroxisome, respectively). This is almost certainly not a true reflection of the situation in the cell, and indeed many reactions assigned to cytosol in the model are known to occur in other compartments. For example, most reactions of amino acid biosynthesis and secondary metabolism were assigned a cytosolic location even though it is well established that both occur extensively in the plastid. By way of comparison, only 20% of the reactions in a model of *Chlamydomonas* primary metabolism were cytosolic and nearly half were plastidic (Boyle and Morgan, 2009). This was despite the use of the cytosol as a "default" location where the subcellular localization of an enzyme was unclear.

Another issue with introducing compartmentation into metabolic models is the lack of information about metabolite transport. This means that intracellular transporters are often added to metabolic models based on their necessity to allow the synthesis of biomass within the compartmented model (de Oliveira Dal'Molin et al., 2010a). It is also generally the case that no attempt is made to account explicitly for the energetic cost of transport, so by default this is included in the energy cost attributed to cell maintenance.

The functional significance of subcellular compartmentation is not necessarily obvious, and steady-state MFA has drawn attention to the importance of transmembrane metabolite exchange rates in determining the extent to which the intermediates in physically compartmented pathways are able to function indistinguishably from an uncompartmented pathway (Schwender et al., 2003; Ratcliffe and Shachar-Hill, 2006; Masakapalli et al., 2010). Thus in principle it would be useful if the problems identified above could be resolved to allow FBA to explore the functionality of subcellular compartmentation. However, it might be well to consider whether flux-balance models are sufficiently constrained to define compartmented fluxes. The addition of compartmentation, and especially the addition of parallel pathways in more than one compartment, effectively increases the solution space and it seems likely that increased compartmentation in a model will simply increase the number of alternate solutions to the optimization problem. There have been too few attempts to include the necessary level of subcellular compartmentation in plant models to reach a conclusion on this point, and the extent to which flux-balance models can usefully analyze highly compartmented networks requires further investigation.

FLUX-BALANCE MODELING OF SPECIFIC CELL TYPES OR TISSUES

Almost all of the flux-balance models of plant metabolism, compartmented or not, consider only a single cell type, either by modeling single-celled organisms such as *Chlamydomonas* (Boyle and Morgan, 2009; Cogne et al., 2011) or *Synechocystis* (Knoop et al., 2010; Montagud et al., 2010), modeling cell suspension cultures (Poolman et al., 2009; Williams et al., 2010) or simply ignoring the presence of multiple cell types in models of specific organs or tissues (Grafahrend-Belau et al., 2009; de Oliveira Dal'Molin et al., 2010a; Hay and Schwender, 2011a,b; Pilalis et al., 2011; Saha et al., 2011). To an extent this is justifiable because the measured biomass composition used to constrain these models are whole tissue biomass compositions. However, it follows that the resulting metabolic network flux solution represents an average of the different cell types in that tissue or organ. Given that different cell types have very different metabolic capacities (Brady et al., 2007; Lee et al., 2011), it is likely that there will be major differences in both the structure of their metabolic networks and in the fluxes through them. Ultimately, if a flux model is going to be useful in explaining the metabolic phenotype in detail, it will be necessary to provide information about flux at a cell-type level (Sweetlove et al., 2010).

The challenge goes beyond simply constructing specific models for specific cell types, but also in joining up these models to form multi-layered representations of complete tissues. This has been achieved in a sophisticated model of metabolic interactions between different cell types in the human brain (Lewis et al., 2010). This study used transcriptomic and proteomic data to define cell type specific metabolic networks for three different neuronal cell types, astrocytes, and blood/endothelium. Subcellular compartmentation was introduced into each cell-type model and transporters were included to allow transport of specific metabolites between the cell types. The model was able to generate possible explanations for the differential effects of Alzheimer's disease on different cell types and regions of the brain.

This flux-balance model of brain metabolism is state-of-the-art and in principle there is no reason why such detailed large-scale models should not be constructed for plant metabolism. To date, only one study has attempted to account for the interaction of more than one cell type. Based on their previous genome-scale model of *Arabidopsis* metabolism, de Oliveira Dal'Molin et al. (2010b) constructed a flux-balance model describing the interaction between bundle sheath and mesophyll cells in C4 photosynthesis. The metabolic network was restricted in each cell type to reflect the known distribution of carbon fixation in C4 photosynthesis, with primary fixation of carbon through PEP carboxylase in mesophyll cells, transport of aspartate or malate to the bundle sheath cells, and subsequent decarboxylation by NADP-malic enzyme, NAD-malic enzyme or phospho*enol*pyruvate carboxykinase. Flux solutions were generated using optimization of photon use as an objective

function. While this objective function could not reproduce every aspect of C4 metabolism, for example the preferential accumulation of starch in bundle sheath cells, the model could be used to examine the energetic implications of the three C4 sub-types. For example, the ATP/NADPH ratio required in NAD-malic enzyme species is higher in mesophyll cells than in bundle sheath cells, but the opposite is true for NADP-malic enzyme species. The flux distribution in the models of the different sub-types confirmed the hypothesis that the additional ATP demand in the different cell types is met by cyclic photophosphorylation.

Another interesting observation from this study was that the relative fluxes between bundle sheath and mesophyll cells correlated well with the relative abundance of the enzymes estimated from proteomic studies (de Oliveira Dal'Molin et al., 2010b). It is clear that enzyme abundance does not relate directly to flux, partly because of post-translational regulation of enzyme activity and partly because of the impact of the thermodynamic poise of a reaction on the ability of an enzyme to support a net flux. However, what this correlation shows is that changes in flux are reflected in proteome-wide adjustments in enzyme amount. The implication is that relative enzyme abundance might be a useful proxy for the change in flux between two conditions or cell types. That said, there are many reasons why such a correlation might break down. For example, it has been shown in yeast that while some V_{max} values correlate positively with flux changes, others show an inverse correlation and some show no correlation at all (Rossell et al., 2006). And during stress, many enzymes are inhibited by oxidative damage (Taylor et al., 2004; Lehmann et al., 2009), but this is not necessarily reflected at the protein level. An alternative way of exploiting the correlation between changes in flux and enzyme amount would be to use the enzyme abundance data as a constraint when predicting changes in flux, although the effort required to establish proteomic measurements of sufficient enzymes to cover a significant proportion of the metabolic network would be substantial.

ACCOUNTING FOR CELL MAINTENANCE ENERGY COSTS

In modeling the central metabolic network, the published flux-balance models have exclusively considered the conversion of carbon and nitrogen inputs into biomass. Biosynthesis of the precursors that constitute the main biomass polymers (cell wall, protein, lipid, starch) requires both ATP and NAD(P)H and thus, the energy costs of biomass synthesis are explicitly taken into account. However, there are several other cellular drains on ATP and NAD(P)H apart from the synthesis of biomass. These other energy costs are often bracketed together under the term "maintenance," implying that these are growth-independent costs that are required just to keep the cell ticking over. This distinction is not strictly accurate because the maintenance costs in metabolic models often include some growth-associated costs.

Other energy costs are associated with the need to replace polymers as they turn over, with the costs of maintaining plasma membrane and tonoplast electrochemical potential gradients through ATP- and PP_i-dependent proton pumps, and with the consumption of reductant during antioxidant metabolism (Amthor, 2000). The usual approach to dealing with these maintenance costs in flux-balance modeling is to include a fixed value for maintenance costs based on experimental measures (e.g., Grafahrend-Belau et al., 2009). However, there is a wide variation in reported values for maintenance respiration for plant tissues, and maintenance costs are likely to increase during environmental stress, which may be one explanation for the observed reduction in carbon conversion efficiency under stress (Williams et al., 2010). The importance of an accurate measure of maintenance costs is revealed by the observation that the maximum yield of ATP generated by the metabolic network of heterotrophic *Arabidopsis* cells is over seven times that required for the synthesis of the main biomass components (Masakapalli et al., 2010), the implication being that maintenance costs account for the majority of ATP consumed by the cell.

Poolman et al. (2009) used an alternative approach for the estimation of maintenance costs. Initially, fluxes in the network were estimated, with the synthesis of biomass components as constraints and minimization of total flux as an objective function, without taking maintenance into account. Subsequently, a generic ATPase reaction was added to the model to represent the maintenance ATP requirement. The flux value of this ATPase was iteratively varied and the linear programming optimization repeated. As the ATPase was increased, glucose consumption, glycolysis, and oxidative phosphorylation increased to meet the increased ATP demand. This allowed the maintenance ATP cost to be set to the ATPase reaction flux that led to a glucose consumption rate equal to the value measured experimentally in the cell suspension cultures (Williams et al., 2010). Effectively, the maintenance cost was estimated from the carbon balance of the system by assuming that consumed carbon that was not accounted for by biomass synthesis must have been for maintenance. A similar approach was used in a recent FBA model of oilseed rape, with the slight modification that carbon conversion efficiency was used as the parameter to set the value for the generic ATPase flux (Hay and Schwender, 2011b). The use of a generic ATPase flux in this way provides a convenient method for accounting for ATP costs that are additional to those required for biosynthesis of biomass components. However, it should be pointed out that the accuracy of the predicted maintenance ATP cost will be dependent on how close the flux-balance solutions are to the actual metabolic flux state.

EXPLORING METABOLIC EFFICIENCY WITH FLUX-BALANCE MODELS

Flux-balance modeling is well equipped for the analysis of metabolic efficiency because FBA is based on the discovery of flux solutions that are optimal with respect to a specific objective function. Several of the published FBA studies of plant metabolism explore issues that relate to metabolic efficiency. For example photorespiration was non-zero in a model of photoautotrophic metabolism in *Synechocystis* optimized for maximal biomass production (Knoop et al., 2010). This is surprising because Rubisco oxygenase was not forced to operate and one would expect that reactions leading to loss of carbon as CO_2 would have zero flux when maximization of biomass production is the objective function This is because when the carbon input rate is fixed, maximization of biomass equates to a maximization of carbon conversion efficiency. The fact that photorespiration carried flux under these circumstances means that the requirement for production of intermediates by this route outweighed the loss of carbon. Part of the explanation appears to be a lack of alternative routes to serine in the *Synechocystis* model.

However, it is noteworthy that a flux-balance model of heterotrophic *Arabidopsis* metabolism also contains a non-zero flux for the Rubisco oxygenase reaction and subsequent reactions of photorespiration as far as glycine. (Poolman et al., 2009). In this model, this was the main route for synthesis of glycine. Transcript and proteomic data both suggest the presence of photorespiratory enzymes in non-photosynthetic tissues in *Arabidopsis* (Zimmermann et al., 2004; Baerenfaller et al., 2008). The precise role of the photorespiratory reactions in non-photosynthetic tissues, and their requirement for optimal growth of photosynthetic tissues, requires further investigation, highlighting the power of FBA in the identification of non-intuitive flux behavior in metabolic networks.

Flux-balance modeling can also be used to explore the efficiency of different modes of carbon assimilation within realistic growth constraints. Because the carbon conversion efficiency of photosynthesis is directly related to crop yield, there is a great deal of interest in the possibility of alternative photo-assimilatory pathways that might operate at higher efficiency (Bar-Even et al., 2010). The efficiency of six carbon assimilation pathways (Calvin–Benson–Bassham cycle, reductive TCA cycle, 3-hydroxypropionate/malyl-CoA cycle, reductive acetyl-CoA pathway, 3-hydroxypropionate/4-hydroxybutyrate cycle, and the dicarboxylate/4-hydroxybutyrate cycle) was compared by establishing flux-balance solutions for six different bacteria (Boyle and Morgan, 2011). Based on comparisons of either photon requirement or the energy demand for conversion of photoassimilate into biomass, it was found that photoautotrophic pathways are more efficient than chemoautotrophic carbon assimilation pathways (unless there is a free source of hydrogen) and that the reductive TCA cycle is the most efficient way of generating biomass from solar energy. However, the reductive TCA cycle is only marginally more efficient than the Calvin–Benson–Bassham cycle (25.3 and 24.9% efficiency, respectively, where efficiency is calculated as the heat of combustion of biomass divided by the total amount of energy used to create biomass).

The calculation of theoretical optimal yields of metabolic networks is relatively straightforward from flux-balance models, but more biologically informed assessments of metabolic efficiencies can be made by comparison of computed optimal flux distributions against those that actually occur. Two studies have found that flux-balance models can replicate experimentally determined flux distributions in heterotrophic *Arabidopsis* cells (Williams et al., 2010) and *B. napus* seeds (Hay and Schwender, 2011b). In both of these studies objective functions were used that equate to carbon conversion efficiency: minimization of total intracellular flux or minimization of substrate consumption, per unit biomass produced. The fact that these objective functions produce flux solutions that match the measured *in vivo* flux distributions suggests that the metabolic network in these tissues is functioning close to optimal carbon conversion efficiency. A similar conclusion can be reached from a flux-balance model of barley seed in which maximization of growth rate for a fixed substrate-consumption rate was able to predict the growth rate of barley seeds (Grafahrend-Belau et al., 2009). Maximization of growth (biomass) for a fixed amount of substrate is effectively a maximization of molar yield (Schuster et al., 2008). In other words, this objective function is closely related to objective functions that minimize substrate consumption or overall intracellular flux for a fixed biomass output.

The conclusion that plant metabolic networks may already be operating close to maximal carbon conversion efficiency is important, because improvement of carbon conversion efficiency is often cited as a key breeding target for improved crop yield (Hauben et al., 2009; Parry et al., 2010). However, the conclusion, as it stands, requires substantial qualification. The main issue is that the carbon conversion efficiency in the *Arabidopsis* and *B. napus* models is forced to match the measured value and this to a large extent dictates the good match between the modeled and measured fluxes in the central metabolic network. Moreover, only a small fraction of the flux distribution can be legitimately validated for the reasons discussed earlier, and systematic assessments of different objective functions (Schuetz et al., 2007) have not yet been reported for plant models. It is entirely possible that the differences in the flux solutions obtained with different objective functions may fall within the statistical error of the flux determinations, and thus provide no real discriminatory power to investigate metabolic network efficiency. Nevertheless the conclusion is in line with previous estimates of the theoretical efficiencies of plant energy metabolism based on less sophisticated pathway analysis (Penning de Vries, 1974; Penning de Vries et al., 1974).

FUTURE CHALLENGES FOR FLUX-BALANCE MODELING OF PLANT METABOLISM

The flux-balance models of plant metabolism that have been published in the last 2 years have been steadily gaining in sophistication. The inclusion of subcellular compartmentation, the analysis of multiple cell types, and the analysis of flux variability are significant developments that increase the utility and predictive capacity of the models. Technical challenges remain, for example in the analysis of subcellular compartmentation, but it is clear that FBA is a useful addition to the toolbox for analyzing plant metabolic networks. Moreover the ease of implementation in comparison to stable isotope-based MFA, and the availability of metabolic compendia (Zhang et al., 2010) based on genomic information, suggest that flux-balance models will continue to be developed for a variety of plant species and tissue types. Given the growing popularity of the approach, and the potential for genome-scale models to be used in tandem with computational analysis of genomes (Bekaert et al., 2011) it is pertinent to try and identify the areas in which FBA could be most usefully deployed as the technique develops.

Ultimately, the goal toward which metabolic modeling must advance is a reconstruction of metabolism at the whole-plant level. While in principle FBA is well suited to dealing with interacting cell types, considering whole plants raises the problem of temporal differences in metabolism during the development of the tissues (Walton and de Jong, 1990). A particular issue is that the pattern of growth of most plant tissues is not uniform with time: cells are initiated by division at the meristem and subsequently grow by expansion. This represents two very different modes of growth that will not be fully captured by constraints derived from biomass composition of mature cells. This is because such constraints assume that each component of biomass accumulates in a linear fashion and in the same proportion over the history of the cell. This is unlikely to be true since the nature of biomass accumulation during cell expansion is different to that during division (Thornley and Johnson, 1990). Moreover mature organs can make a significant

contribution to whole-plant metabolism while not growing at all. Thus, the growth-based objective functions that currently tend to dominate flux-balance modeling would not be appropriate. On the other hand, differentiation and secondary growth can occur after organs have reached maturity, especially the synthesis of lignins and hemicelluloses prior to senescence (Amthor, 2000); and at the whole-plant scale, remobilization of resources during senescence can be quantitatively important for sink metabolism (Taylor et al., 2010). There is also the issue of environment to consider: acclimation to changing conditions is often associated with altered growth and composition (Armstrong et al., 2006) and leaves, in particular, are especially sensitive to diurnal fluctuations in light intensity. While none of these problems are insurmountable, it is obvious that careful consideration of objective functions and constraints that can usefully bound the flux space will be required. Additional experimental measures of growth and composition on a finer-grained scale will surely be necessary (Pramanik and Keasling, 1997). Computational approaches, such as dynamic FBA (Mahadevan et al., 2002), will also have to be developed to allow steady-state models to be concatenated along a developmental time axis.

Given the considerable effort that will be required to develop such sophisticated metabolic models, it is worth establishing at the outset the type of biological insights that might be expected from the approach. As we have already discussed, flux-balance modeling can highlight non-intuitive metabolic routes that may be worthy of subsequent experimental investigation. Flux-balance modeling can also be a useful means of predicting changes in flux under different conditions and with different nutrient sources, providing a deeper understanding of the metabolic demands of a varying environment. And because of the optimization algorithms at the heart of FBA, it is inherently good at examining the theoretical yield limits and energetic efficiencies of the metabolic networks under consideration.

Perhaps the ultimate driving force of most metabolic modeling is the lure of metabolic engineering; the aim being to identify the most appropriate target enzymes for genetic manipulation with respect to a desired metabolic output. However, in contrast to enzyme-kinetic modeling which provides a quantitative measure of the extent to which each enzyme controls pathway flux (Heinzle et al., 2007; Schallau and Junker, 2010), flux-balance models contain no intrinsic information about enzyme regulation or the control of flux. Nevertheless, flux-balance modeling can be used to identify optimal flux distributions that maximize the synthesis of a desired metabolic end-product. Commonly this takes the form of a systematic analysis of reaction deletions (either singly or combinations of multiple reaction deletions) while setting maximization of product yield as the objective function (Burgard et al., 2003; Alper et al., 2005). Unfortunately the solution usually involves a significant reduction in growth, reflecting the diversion of carbon into the desired product, and this is not particularly useful from a biotechnological viewpoint. Therefore, more recent attempts have incorporated combined objective functions of maximizing the production of the target compound while still maximizing biomass production (Montagud et al., 2010). This allows flux distributions to be identified that lead to over-production of the target compound without a drastic reduction in growth rate. Identifying the ideal flux distribution is only the first step and engineering that flux state into organisms, even bacteria, is a formidable challenge. But it can be done. In an impressive demonstration of metabolic engineering prowess, information from flux analysis was used to guide a total of 12 genetic interventions in *Corynebacterium glutamicum* (Becker et al., 2011) to generate a lysine-overproducing strain of similar performance to those produced industrially.

SUMMARY

In summary, flux-balance modeling of plant metabolic networks provides an important complement to ^{13}C-based MFA and an alternative to smaller scale mechanistic models based on enzyme kinetics. While flux-balance modeling has its limitations, stemming from the underdetermined nature of the problem and the lack of enzyme regulatory information or kinetic responses, it is capable of generating novel insights into metabolic behavior, capacities, and efficiency and it can be used as a framework for metabolic engineering. Future efforts toward multi-cell, multi-tissue, and ultimately whole-plant models will form an important component of computational models of plant growth and development (Christophe et al., 2008) and are likely to play a major role in efforts to improve crop yield and quality.

REFERENCES

Allen, D. K., Libourel, I. G. L., and Shachar-Hill, Y. (2009a). Metabolic flux analysis in plants: coping with complexity. *Plant Cell Environ.* 32, 1241–1257.

Allen, D. K., Ohlrogge, J. B., and Shachar-Hill, Y. (2009b). The role of light in soybean seed filling metabolism. *Plant J.* 58, 220–234.

Alonso, A. P., Dale, V. L., and Shachar-Hill, Y. (2010). Understanding fatty acid synthesis in developing maize embryos using metabolic flux analysis. *Metab. Eng.* 12, 488–497.

Alonso, A. P., Goffman, F. D., Ohlrogge, J. B., and Shachar-Hill, Y. (2007). Carbon conversion efficiency and central metabolic fluxes in developing sunflower (*Helianthus annuus* L.) embryos. *Plant J.* 52, 296–308.

Alper, H., Jin, Y. S., Moxley, J. F., and Stephanopoulos, G. (2005). Identifying gene targets for the metabolic engineering of lycopene biosynthesis in *Escherichia coli*. *Metab. Eng.* 7, 155–164.

Amthor, J. S. (2000). The McCree-de Wit-Penning de Vries-Thornley respiration paradigms: 30 years later. *Ann. Bot.* 86, 1–20.

Apel, K., and Hirt, H. (2004). Reactive oxygen species: metabolism, oxidative stress, and signal transduction. *Annu. Rev. Plant Biol.* 55, 373–399.

Araujo, W. L., Nunes-Nesi, A., Nikoloski, Z., Sweetlove, L. J., and Fernie, A. R. (2011). Metabolic control and regulation of the tricarboxylic acid cycle in photosynthetic and heterotrophic plant tissues. *Plant Cell Environ.* doi: 10.1111/j.1365-3040.2011.02332.x. [Epub ahead of print].

Armstrong, A. F., Logan, D. C., Tobin, A. K., O'Toole, P., and Atkin, O. K. (2006). Heterogeneity of plant mitochondrial responses underpinning respiratory acclimation to the cold in *Arabidopsis thaliana* leaves. *Plant Cell Environ.* 29, 940–949.

Averill, R. H., Bailey-Serres, J., and Kruger, N. J. (1998). Co-operation between cytosolic and plastidic oxidative pentose phosphate pathways revealed by 6-phosphogluconate dehydrogenase-deficient genotypes of maize. *Plant J.* 14, 449–457.

Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W., and Baginsky, S. (2008). Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* 320, 938–941.

Bar-Even, A., Noor, E., Lewis, N. E., and Milo, R. (2010). Design and analysis of synthetic carbon fixation pathways. *Proc. Natl. Acad. Sci. U.S.A.* 107, 8889–8894.

Becker, J., Zelder, O., Hafner, S., Schroder, H., and Wittmann, C. (2011). From zero to hero-design-based systems metabolic engineering of *Corynebacterium glutamicum* for L-lysine production. *Metab. Eng.* 13, 159–168.

Bekaert, M., Edger, P. P., Pires, J. C., and Conant, G. C. (2011). Two-phase resolution of polyploidy in the *Arabidopsis* metabolic network gives rise to relative

and absolute dosage constraints. *Plant Cell* 23, 1719–1728.

Borodina, I., and Nielsen, J. (2005). From genomes to *in silico* cells via metabolic networks. *Curr. Opin. Biotechnol.* 16, 350–355.

Boyle, N. R., and Morgan, J. A. (2009). Flux balance analysis of primary metabolism in *Chlamydomonas reinhardtii*. *BMC Syst. Biol.* 3, 4. doi: 10.1186/1752-0509-3-4

Boyle, N. R., and Morgan, J. A. (2011). Computation of metabolic fluxes and efficiencies for biological carbon dioxide fixation. *Metab. Eng.* 13, 150–158.

Brady, S. M., Orlando, D. A., Lee, J. Y., Wang, J. Y., Koch, J., Dinneny, J. R., Mace, D., Ohler, U., and Benfey, P. N. (2007). A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science* 318, 801–806.

Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003). OptKnock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* 84, 647–657.

Butelli, E., Titta, L., Giorgio, M., Mock, H. P., Matros, A., Peterek, S., Schijlen, E. G. W. M., Hall, R. D., Bovy, A. G., Luo, J., and Martin, C. (2008). Enrichment of tomato fruit with health-promoting anthocyanins by expression of select transcription factors. *Nat. Biotechnol.* 26, 1301–1308.

Christophe, A., Letort, V., Hummel, I., Cournede, P. H., de Reffye, P., and Lecoeur, J. (2008). A model-based analysis of the dynamics of carbon balance at the whole-plant level in *Arabidopsis thaliana*. *Funct. Plant Biol.* 35, 1147–1162.

Cogne, G., Rugen, M., Bockmayr, A., Titica, M., Dussap, C. G., Cornet, J. F., and Legrand, J. (2011). A model-based method for investigating bioenergetic processes in autotrophically growing eukaryotic microalgae: application to the green algae *Chlamydomonas reinhardtii*. *Biotechnol. Prog.* 27, 631–640.

de Oliveira Dal'Molin, C. G., Quek, L. E., Palfreyman, R. W., Brumbley, S. M., and Nielsen, L. K. (2010a). Aragem, a genome-scale reconstruction of the primary metabolic network in *Arabidopsis*. *Plant Physiol.* 152, 579–589.

de Oliveira Dal'Molin, C. G., Quek, L. E., Palfreyman, R. W., Brumbley, S. M., and Nielsen, L. K. (2010b). C4gem, a genome-scale metabolic model to study C_4 plant metabolism. *Plant Physiol.* 154, 1871–1885.

Edwards, J. S., and Palsson, B. O. (2000). The *Escherichia coli* mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. U.S.A.* 97, 5528–5533.

Enfissi, E. M., Fraser, P. D., Lois, L. M., Boronat, A., Schuch, W., and Bramley, P. M. (2005). Metabolic engineering of the mevalonate and non-mevalonate isopentenyl diphosphate-forming pathways for the production of health-promoting isoprenoids in tomato. *Plant Biotechnol. J.* 3, 17–27.

Fell, D. A., Poolman, M. G., and Gevorgyan, A. (2010). Building and analysing genome-scale metabolic models. *Biochem. Soc. Trans.* 38, 1197–1201.

Fernie, A. R., and Schauer, N. (2009). Metabolomics-assisted breeding: a viable option for crop improvement? *Trends Genet.* 25, 39–48.

Fraser, P. D., Romer, S., Shipton, C. A., Mills, P. B., Kiano, J. W., Misawa, N., Drake, R. G., Schuch, W., and Bramley, P. M. (2002). Evaluation of transgenic tomato plants expressing an additional phytoene synthase in a fruit-specific manner. *Proc. Natl. Acad. Sci. U.S.A.* 99, 1092–1097.

Geigenberger, P., Stitt, M., and Fernie, A. R. (2004). Metabolic control analysis and regulation of the conversion of sucrose to starch in growing potato tubers. *Plant Cell Environ.* 27, 655–673.

Grafahrend-Belau, E., Schreiber, F., Koschutzki, D., and Junker, B. H. (2009). Flux balance analysis of barley seeds: a computational approach to study systemic properties of central metabolism. *Plant Physiol.* 149, 585–598.

Hauben, M., Haesendonckx, B., Standaert, E., Van Der Kelen, K., Azmi, A., Akpo, H., Van Breusegem, F., Guisez, Y., Bots, M., Lambert, B., Laga, B., and De Block, M. (2009). Energy use efficiency is characterized by an epigenetic component that can be directed through artificial selection to increase yield. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20109–20114.

Hay, J., and Schwender, J. (2011a). Computational analysis of storage synthesis in developing *Brassica napus* L. (oilseed rape) embryos: flux variability analysis in relation to ^{13}C-metabolic flux analysis. *Plant J.* 67, 513–525.

Hay, J., and Schwender, J. (2011b). Metabolic network reconstruction and flux variability analysis of storage synthesis in developing oilseed rape (*Brassica napus* L.) embryos. *Plant J.* 67, 526–541.

Heazlewood, J. L., Tonti-Filippini, J. S., Gout, A. M., Day, D. A., Whelan, J., and Millar, A. H. (2004). Experimental analysis of the Arabidopsis mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. *Plant Cell* 16, 241–256.

Heazlewood, J. L., Verboom, R. E., Tonti-Filippini, J., Small, I., and Millar, A. H. (2007). SUBA: the Arabidopsis subcellular database. *Nucleic Acids Res.* 35, D213–D218.

Heinzle, E., Matsuda, F., Miyagawa, H., Wakasa, K., and Nishioka, T. (2007). Estimation of metabolic fluxes, expression levels and metabolite dynamics of a secondary metabolic pathway in potato using label pulse-feeding experiments combined with kinetic network modelling and simulation. *Plant J.* 50, 176–187.

Henry, C. S., Dejongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* 28, 977–982.

Iyer, V. V., Sriram, G., Fulton, D. B., Zhou, R., Westgate, M. E., and Shanks, J. V. (2008). Metabolic flux maps comparing the effect of temperature on protein and oil biosynthesis in developing soybean cotyledons. *Plant Cell Environ.* 31, 506–517.

Jankowski, M. D., Henry, C. S., Broadbelt, L. J., and Hatzimanikatis, V. (2008). Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.* 95, 1487–1499.

Junker, B. H., Lonien, J., Heady, L. E., Rogers, A., and Schwender, J. (2007). Parallel determination of enzyme activities and in vivo fluxes in *Brassica napus* embryos grown on organic or inorganic nitrogen source. *Phytochemistry* 68, 2232–2242.

Knoop, H., Zilliges, Y., Lockau, W., and Steuer, R. (2010). The metabolic network of *Synechocystis* sp. PCC 6803: systemic properties of autotrophic growth. *Plant Physiol.* 154, 410–422.

Kruger, N. J., and Ratcliffe, R. G. (2008). Metabolic organization in plants: a challenge for the metabolic engineer. *Adv. Plant Biochem. Mol. Biol.* 1, 1–27.

Kruger, N. J., and Ratcliffe, R. G. (2009). Insights into plant metabolic networks from steady-state metabolic flux analysis. *Biochimie* 91, 697–702.

Kruger, N. J., and von Schaewen, A. (2003). The oxidative pentose phosphate pathway: structure and organisation. *Curr. Opin. Plant Biol.* 6, 236–246.

Lee, C. P., Eubel, H., O'Toole, N., and Millar, A. H. (2011). Combining proteomics of root and shoot mitochondria and transcript analysis to define constitutive and variable components in plant mitochondria. *Phytochemistry* 72 1092–1108.

Lee, S., Chan, P., Domach, M. M., and Grossman, I. E. (2000). Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Comput. Chem. Eng.* 24, 711–716.

Lehmann, M., Schwarzlander, M., Obata, T., Sirikantaramas, S., Burow, M., Olsen, C. E., Tohge, T., Fricker, M. D., Moller, B. L., Fernie, A. R., Sweetlove, L. J., and Laxa, M. (2009). The metabolic response of arabidopsis roots to oxidative stress is distinct from that of heterotrophic cells in culture and highlights a complex relationship between the levels of transcripts, metabolites, and flux. *Mol. Plant* 2, 390–406.

Lewis, N. E., Schramm, G., Bordbar, A., Schellenberger, J., Andersen, M. P., Cheng, J. K., Patel, N., Yee, A., Lewis, R. A., Eils, R., König, R., and Palsson, B. Ø. (2010). Large-scale *in silico* modeling of metabolic interactions between cell types in the human brain. *Nat. Biotechnol.* 28, 1279–1285.

Libourel, I. G. L., and Shachar-Hill, Y. (2008). Metabolic flux analysis in plants: from intelligent design to rational engineering. *Annu. Rev. Plant Biol.* 59, 625–650.

Lonien, J., and Schwender, J. (2009). Analysis of metabolic flux phenotypes for two *Arabidopsis* mutants with severe impairment in seed storage lipid synthesis. *Plant Physiol.* 151, 1617–1634.

Lunn, J. E. (2007). Compartmentation in plant metabolism. *J. Exp. Bot.* 58, 35–47.

Mahadevan, R., Edwards, J. S., and Doyle, F. J. III (2002). Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophys. J.* 83, 1331–1340.

Mahadevan, R., and Schilling, C. H. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* 5, 264–276.

Masakapalli, S. K., Le Lay, P., Huddleston, J. E., Pollock, N. L., Kruger, N. J., and Ratcliffe, R. G. (2010). Subcellular flux analysis of central metabolism in a heterotrophic *Arabidopsis* cell suspension using steady-state stable isotope labeling. *Plant Physiol.* 152, 602–619.

Memelink, J., and Gantet, P. (2007). Transcription factors involved in terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. *Phytochem. Rev.* 6, 353–362.

Montagud, A., Navarro, E., Fernandez de Cordoba, P., Urchueguia, J. F., and Patil, K. R. (2010). Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. *BMC Syst. Biol.* 4, 156. doi: 10.1186/1752-0509-4-156

Naqvi, S., Zhu, C., Farre, G., Ramessar, K., Bassie, L., Breitenbach, J., Perez Conesa, D., Ros, G., Sandmann, G., Capell, T., and Christou, P. (2009). Transgenic multivitamin corn through biofortification of endosperm with three vitamins representing three

distinct metabolic pathways. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7762–7767.
Parry, M. A., Reynolds, M., Salvucci, M. E., Raines, C., Andralojc, P. J., Zhu, X. G., Price, G. D., Condon, A. G., and Furbank, R. T. (2010). Raising yield potential of wheat. II. Increasing photosynthetic capacity and efficiency. *J. Exp. Bot.* 62, 453–467.
Penning de Vries, F. W. T. (1974). Substrate utilization and respiration in relation to growth and maintenance in higher plants. *Neth. J. Agric. Sci.* 22, 40–44.
Penning de Vries, F. W. T., Brunsting, A. H. M., and Van Laar, H. H. (1974). Products, requirements and efficiency of biosynthesis – quantitative approach. *J. Theor. Biol.* 45, 339–377.
Pilalis, E., Chatziioannou, A., Thomasset, B., and Kolisis, F. (2011). An in silico compartmentalized metabolic model of *Brassica napus* enables the systemic study of regulatory aspects of plant central metabolism. *Biotechnol. Bioeng.* 108, 1673–1682.
Poolman, M. G., Miguet, L., Sweetlove, L. J., and Fell, D. A. (2009). A genome-scale metabolic model of *Arabidopsis* and some of its properties. *Plant Physiol.* 151, 1570–1581.
Pramanik, J., and Keasling, J. D. (1997). Stoichiometric model of *Escherichia coli* metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnol. Bioeng.* 56, 398–421.
Radrich, K., Tsuruoka, Y., Dobson, P., Gevorgyan, A., Swainston, N., Baart, G., and Schwartz, J. M. (2010). Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Syst. Biol.* 4, 114. doi: 10.1186/1752-0509-4-114
Raines, C. A. (2003). The Calvin cycle revisited. *Photosynth. Res.* 75, 1–10.
Ratcliffe, R. G., and Shachar-Hill, Y. (2006). Measuring multiple fluxes through plant metabolic networks. *Plant J.* 45, 490–511.
Reed, J. L., and Palsson, B. O. (2003). Thirteen years of building constraint-based in silico models of *Escherichia coli*. *J. Bacteriol.* 185, 2692–2699.
Roscher, A., Kruger, N. J., and Ratcliffe, R. G. (2000). Strategies for metabolic flux analysis in plants using isotope labelling. *J. Biotechnol.* 77, 81–102.
Rossell, S., van der Weijden, C. C., Lindenbergh, A., van Tuijl, A., Francke, C., Bakker, B. M., and Westerhoff, H. V. (2006). Unraveling the complexity of flux regulation: a new method demonstrated for nutrient starvation in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.* 103, 2166–2171.
Saha, R., Suthers, P., and Maranas, C. (2011). *Zea mays* irs1563: a comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLoS ONE* 6, e21784. doi: 10.1371/journal.pone.0021784
Schallau, K., and Junker, B. H. (2010). Simulating plant metabolic pathways with enzyme-kinetic models. *Plant Physiol.* 152, 1763–1771.
Schuetz, R., Kuepfer, L., and Sauer, U. (2007). Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol.* 3, 119.
Schuster, S., Pfeiffer, T., and Fell, D. A. (2008). Is maximization of molar yield in metabolic networks favoured by evolution? *J. Theor. Biol.* 252, 497–504.
Schwender, J., Goffman, F., Ohlrogge, J. B., and Shachar-Hill, Y. (2004). Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. *Nature* 432, 779–782.
Schwender, J., Ohlrogge, J. B., and Shachar-Hill, Y. (2003). A flux model of glycolysis and the oxidative pentosephosphate pathway in developing *Brassica napus* embryos. *J. Biol. Chem.* 278, 29442–29453.
Schwinn, K., Venail, J., Shang, Y., Mackay, S., Alm, V., Butelli, E., Oyama, R., Bailey, P., Davies, K., and Martin, C. (2006). A small family of myb-regulatory genes controls floral pigmentation intensity and patterning in the genus antirrhinum. *Plant Cell* 18, 831–851.
Shastri, A. A., and Morgan, J. A. (2005). Flux balance analysis of photoautotrophic metabolism. *Biotechnol. Prog.* 21, 1617–1626.
Shastri, A. A., and Morgan, J. A. (2007). A transient isotopic labeling methodology for ^{13}C metabolic flux analysis of photoautotrophic microorganisms. *Phytochemistry* 68, 2302–2312.
Smith, A. M., and Stitt, M. (2007). Coordination of carbon supply and plant growth. *Plant Cell Environ.* 30, 1126–1149.
Soh, K. C., and Hatzimanikatis, V. (2010). DREAMS of metabolism. *Trends Biotechnol.* 28, 501–508.
Sriram, G., Fulton, D. B., Iyer, V. V., Peterson, J. M., Zhou, R., Westgate, M. E., Spalding, M. H., and Shanks, J. V. (2004). Quantification of compartmented metabolic fluxes in developing soybean embryos by employing biosynthetically directed fractional ^{13}C labeling, two-dimensional [^{13}C, ^{1}H] nuclear magnetic resonance, and comprehensive isotopomer balancing. *Plant Physiol.* 136, 3043–3057.
Stitt, M., Sulpice, R., and Keurentjes, J. (2010). Metabolic networks: how to identify key components in the regulation of metabolism and growth. *Plant Physiol.* 152, 428–444.
Sweetlove, L. J., Beard, K. F., Nunes-Nesi, A., Fernie, A. R., and Ratcliffe, R. G. (2010). Not just a circle: flux modes in the plant TCA cycle. *Trends Plant Sci.* 15, 462–470.
Sweetlove, L. J., Fell, D., and Fernie, A. R. (2008). Getting to grips with the plant metabolic network. *Biochem. J.* 409, 27–41.
Taylor, L., Nunes-Nesi, A., Parsley, K., Leiss, A., Leach, G., Coates, S., Wingler, A., Fernie, A. R., and Hibberd, J. M. (2010). Cytosolic pyruvate, orthophosphate dikinase functions in nitrogen remobilization during leaf senescence and limits individual seed growth and nitrogen content. *Plant J.* 62, 641–652.
Taylor, N. L., Day, D. A., and Millar, A. H. (2004). Targets of stress-induced oxidative damage in plant mitochondria and their impact on cell carbon/nitrogen metabolism. *J. Exp. Bot.* 55, 1–10.
Thornley, J. H. M., and Johnson, I. R. (1990). *Plant and Crop Modeling* Oxford: Oxford University Press.
Torres, M. A. (2010). ROS in biotic interactions. *Physiol. Plant.* 138, 414–429.
Walton, E. F., and de Jong, T. M. (1990). Estimating the bioenergetic cost of a developing kiwifruit berry and its growth and maintenance respiration components. *Ann. Bot.* 66, 417–424.
Williams, T. C. R., Miguet, L., Masakapalli, S. K., Kruger, N. J., Sweetlove, L. J., and Ratcliffe, R. G. (2008). Metabolic network fluxes in heterotrophic *Arabidopsis* cells: stability of the flux distribution under different oxygenation conditions. *Plant Physiol.* 148, 704–718.
Williams, T. C. R., Poolman, M. G., Howden, A. J., Schwarzlander, M., Fell, D. A., Ratcliffe, R. G., and Sweetlove, L. J. (2010). A genome-scale metabolic model accurately predicts fluxes in central carbon metabolism under stress conditions. *Plant Physiol.* 154, 311–323.
Zhang, P., Dreher, K., Karthikeyan, A., Chi, A., Pujar, A., Caspi, R., Karp, P., Kirkup, V., Latendresse, M., Lee, C., Mueller, L. A., Muller, R., and Rhee, S. Y. (2010). Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiol.* 153, 1479–1491.
Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., and Gruissem, W. (2004). Genevestigator. *Arabidopsis* microarray database and analysis toolbox. *Plant Physiol.* 136, 2621–2632.

Conflict of Interest Statement: 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.

Received: 16 May 2011; accepted: 28 July 2011; published online: 11 August 2011.
Citation: Sweetlove LJ and Ratcliffe RG (2011) Flux-balance modeling of plant metabolism. Front. Plant Sci. ***2****:38. doi: 10.3389/fpls.2011.00038*
This article was submitted to Frontiers in Plant Physiology, a specialty of Frontiers in Plant Science.