Metabolite-based genome-wide association studies in plants

Jie Luo

The plant metabolome is the readout of plant physiological status and is regarded as the bridge between the genome and the phenotype of plants. Unraveling the natural variation and the underlying genetic basis of plant metabolism has received increasing interest from plant biologists. Enabled by the recent advances in high-throughput profiling and genotyping technologies, metabolite-based genome-wide association study (mGWAS) has emerged as a powerful alternative forward genetics strategy to dissect the genetic and biochemical bases of metabolism in model and crop plants. In this review, recent progress and applications of mGWAS in understanding the genetic control of plant metabolism and in interactive functional genomics and metabolomics are presented. Further directions and perspectives of mGWAS in plants are also discussed.

Addresses
National Key Laboratory of Crop Genetic Improvement and National Center of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan 430070, China.

Corresponding author: Luo, Jie (jie.luo@mail.hzau.edu.cn)

Introduction
Metabolites are intermediate or end products that are regarded as the readouts of developmental or physiological status of an organism. The total number of plant metabolites is estimated to exceed 200 000 [1,2], reflecting the extraordinary diversity of functions that natural products may serve. Plant metabolites play essential roles in growth, cellular replenishment and whole-plant resource allocation as well as in their interactions with the environment. In addition, they provide indispensable resources for human nutrition, energy and medicine [3]. Plant metabolites can be conceptually divided into two categories: primary and secondary. Primary metabolites are thought to be crucial for proper growth and development while secondary metabolites are regarded as more connected to stress responses to the constantly changing environment [4].

The extreme diversity of metabolites has made plants the ideal models for dissecting the biosyntheses of metabolites and the regulation of metabolic pathways. In recent years, investigations of the natural variation of plant metabolism and its underlying genetic basis have been increasingly conducted using both reverse and forward genetics approaches [5–7]. Benefiting from the completion of Arabidopsis genome and the huge amount of transcriptome data, tremendous progress has been achieved by reverse genetics approaches in elucidating biosynthesis and regulation of plant metabolites, especially secondary metabolites in this species [8]. The advances of recent profiling technologies have enabled quantitative analysis of the variation of metabolites between species and within natural accessions of a single species [8–10]. Forward genetic studies integrating metabolic profiling and quantitative genetics have begun to reveal the genetic regulation of the metabolome in model and crop species. So far, most metabolic studies of this kind have been carried out based on linkage mapping of populations such as F₂ or recombinant inbred line sets derived from crosses between two or more parental accessions, using low-density markers, such as restriction fragment length polymorphism and simple sequence repeat markers [11–13]. Although the introduction of high-density maps generated by next-generation sequencing could substantially increase the mapping resolution of metabolite-based quantitative loci (mQTLs), and a large number of mQTLs have been disclosed by this approach [14*–16], it is clearly not scalable to explore the tremendous variation of abundant diverse germplasm [17]. Moreover, owing to the relatively low resolution, resulting from both the limited recombination with the linkage population and the low-density markers, and the need to set up individual populations, the identification of genes or nucleotides underlying quantitative trait loci (QTL) is still laborious and time-consuming [18,19].

With the advent of high-throughput genotyping technologies such as microarray and next-generation sequencing [20–22], association mapping has been adopted in plant genetic research and has also been used for a few metabolic traits. More recent study combining genome-wide association with metabolomics has demonstrated that it is a powerful forward genetics strategy to dissect the genetic and biochemical bases of plant metabolism [23**,24*].
In this review, I summarize recent progress in metabolite-based genome-wide association study (mGWAS) in plants and the application of mGWAS in understanding the genetic basis of plant metabolism and in functional genomics and metabolomics. Further directions and perspectives of mGWAS in plants are also presented.

**mGWAS in plants**

Genome-wide association studies (GWASs) in plants have identified many loci for complex traits and the number of associations is increasing rapidly as more such trials are conducted in different species for various traits under multiple stresses [25–27]. However, the effect sizes of genetic associations with complex traits are often small [17,28,29] and information on the underlying biological processes is in most cases lacking [30]. Therefore, increasing interest has focused on studying associations with intermediate traits that might be related to biochemical and physiological status of the plants [23**,24**]. A similar situation exists for mGWAS in humans [31–33]; however, mGWAS in plants gets extra benefit from both huge diversity [34] and high inheritability of the majority of the metabolites [23**,24**].

mGWAS in plants was initially applied in the model species A. thaliana, and then successfully performed and extended in a number of other species, especially important crops (Table 1).

A number of mGWASs were based on a targeted profiling strategy in which a small number of metabolites were detected. Chan et al. [35] conducted an mGWAS study using 96 Arabidopsis accessions, 43 glucosinolate (GSL) phenotypes and ~230,000 single nucleotide polymorphisms (SNPs). They identified the two major polymorphic loci controlling GSL variation in natural populations within large blocks of positive associations encompassing dozens of genes. Also in Arabidopsis, an mGWAS carried out on nine branched-chain amino acid (BCAA) traits using 170 344 SNPs among 313 accessions from a diverse panel demonstrated that BRANCHED-CHAIN AMINO ACID TRANSFERASE 2 (BCAT2) contributes to natural variation in BCAA levels, glutamate recycling and free amino acid homeostasis in seeds in an allelic-dependent manner [36]. Examination of maize oil biosynthesis in a GWAS using one million SNPs characterized in 368 maize inbred lines identified 74 loci significantly associated with kernel oil concentration and fatty acid composition (P < 1.8 × 10⁻⁸) [37**]. More recently, an mGWAS on the natural variation of tocochromanolcs with an association panel of 181 maize inbred lines genotyped for 591 822 markers provided insight into the association between ZmVTE4 haplotypes and α-tocopherol content and revealed a novel association between ZmVTE1 and tocotrienol composition [38].

Targeted metabolic profiling is limited to analyzing only a subset of preselected compounds. Recent advances in non-targeted metabolomics based on either nuclear magnetic resonance or mass spectrometry (MS) analysis allow broad-range metabolic profiles to be obtained in high-throughput. Riedelsheimer et al. [39**] assayed 118 biochemical compounds in the leaves of young plants, as well as the agronomic traits of mature plants in field trials for a set of 289 diverse maize inbred lines. Genome-wide association mapping identified strong associations with SNPs for 26 distinct metabolites, explaining up to 32.0% of the observed genetic variance [39**]. Matsuda et al. carried out mGWAS using non-targeted profiling of 175 Japanese rice accessions, leading to identification of 323 associations among 143 SNPs and 89 metabolites [40].

In addition, widely-targeted metabolomics technology based on LC–MS has been recently developed for quantifying metabolites in a high-throughput manner [41,42],

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Summary of mGWAS in plants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>Tissue</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>Leaf</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>Leaf</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>Seedling</td>
</tr>
<tr>
<td>Maize</td>
<td>Grain</td>
</tr>
<tr>
<td>Maize</td>
<td>Kernel</td>
</tr>
<tr>
<td>Rice</td>
<td>Leaf</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>Leaf</td>
</tr>
<tr>
<td>Maize</td>
<td>Leaf</td>
</tr>
<tr>
<td>Maize</td>
<td>Grain</td>
</tr>
<tr>
<td>Maize</td>
<td>Leaf</td>
</tr>
<tr>
<td>Tomato</td>
<td>Fruit</td>
</tr>
<tr>
<td>Rice</td>
<td>Leaf</td>
</tr>
<tr>
<td>Maize</td>
<td>Kernel</td>
</tr>
<tr>
<td>Rice</td>
<td>Leaf</td>
</tr>
</tbody>
</table>
combining the advantages of precise measurement conferred by the targeted profiling and the wide metabolite coverage by the non-targeted analysis [41]. We published what was probably the most comprehensive evaluation of genetic variation in plant metabolism so far [23**]. We profiled a matrix of 840 metabolite features by LC–MS-based widely-targeted metabolomics approach in a worldwide collection of 524 diverse accessions of rice (Oryza sativa). Subsequent mGWAS identified hundreds of significant loci for various metabolites of physiological and nutritional importance, both of large effects and at high resolution [23**]. We also applied the same metabolic profiling strategy in mGWAS using a panel of 368 diverse maize inbred lines and identified 1459 significant locus–trait associations across three environments [24*].

mGWAS for understanding the genetic basis of the plant metabolome

The importance of plant metabolites to both plants and humans makes it a long-standing goal for plant biologists to unravel the genetic basis of plant metabolism [50]. So far, most genetic studies on plant metabolism have been based on structured mapping populations derived mainly from two parent genomes [11,12,13,14*]. mGWASs that allow a wide sampling of the genotypes present within a species have the ability to identify common variants and also have the potential to identify a greater proportion of the variable loci controlling metabolic traits (Figure 1).

mGWASs in a number of species have shown that plant metabolism as a whole is moderately inheritable and shows polygenic inheritance. For example, among the 840 metabolite features detected in rice leaf, 58.7% displayed broad-sense heritability greater than 0.5. Over 70% of the metabolic features had at least one significant association, with an average of 4.9 associations per metabolite feature [23**]. Different to complex traits such as flowering time and grain yield that are controlled by numerous loci of small effects [28,51], levels of metabolites, especially secondary metabolites, are in general controlled by a small number of loci with large effects [23**]. Similar results were also reported for mGWASs in human blood and urine [33,52,53], demonstrating the ubiquity of major mQTLs in nature. However, natural variation in primary metabolites tends to be controlled by loci of smaller effects [16,46,54**].

Genome-wide analysis of significant loci identified a significant deviation from random distribution across the chromosomes for genotype–metabolite associations, revealing the ‘hotspots’ of major genes controlling the levels of large sets of metabolites within the genome [23**,46]. Further study showed that some of the hotspots are within the regions of the Arabidopsis genome previously identified as subject to recent strong positive selection (selective sweeps) and regions showing trans-linkage to these putative sweeps, suggesting that selective forces have impacted genome-wide control of Arabidopsis metabolism [46].

The effects of interaction between genotypes and environment or development on the accumulation of secondary metabolites have been well documented within structured mapping populations. mGWAS on the naturally occurring variation of GSL accumulation in Arabidopsis showed a significant bias toward identifying different causal genes for the GSL phenotypes in the two different tissues and under different developmental stages [35**], which suggests that natural variation of GSLs is genetically controlled in a spatiotemporal manner. Interestingly, distinct genetic control of metabolism was also observed at subspecies level [23**]. mGWAS hotspots were located on different chromosomes in the indica and japonica subspecies of rice. Notably, the genetic architecture was quite different between most of the subspecies differentiation metabolites such as some C-glycosyl flavonoids and phenolamides [23**].

mGWAS for functional genomics and metabolomics

Functional genomics in plants aims to identify the functions of all genes in plant genomes. In addition to the application in elucidating the genetic architecture of plant metabolism, mGWAS can also apply in functional genomics (Figure 1).

The corroboration of candidate associations with biological and functional arguments in mGWAS makes it possible to generate testable hypotheses that can be verified in follow-up experiments [23**,24*,33]. Candidate genes can be mined by looking for a protein or protein cluster that is biochemically related to the associated metabolic trait encoded at the associated loci, by performing cluster analysis of candidate genes relative to homologous genes with known function; and by cross-referencing with results from linkage mapping. Benefiting from the rich knowledge regarding many plant metabolic pathways, mGWAS has allowed the identification of 36 candidate genes underlying metabolites that are of physiological and nutritional importance. Follow-up experiments using multiple genetic and molecular approaches successfully characterized or annotated five genes encoding a methyltransferase, a glucosyltransferase and three putative acyltransferases [23**]. Detailed examination of spatiotemporal accumulation and natural variation of phenolamides in rice followed by subsequent mGWAS also identified two spermidine hydroxycinnamoyl transferases that might be underlying the natural variation of levels of spermidine conjugates in rice [45*]. Understanding natural variation at metabolic level by mGWAS also facilitates the reconstruction of biosynthetic pathways [14*,23**,43], which in turn will benefit synthetic
A schematic view of mGWAS and its applications in plants. The diverse association panel was genotyped with next-generation sequencing (NGS), RNA sequencing (RNA seq) or SNP chip. Candidate genes can be identified through gene annotation, expression and co-expression pattern, prior knowledge of metabolic pathways and common variation detection.
biology and metabolic engineering of desirable compounds in plants.

Integrating GWAS results with additional forms of genome-scale data, such as transcript profiling or proteomics datasets allows rapid identification of novel genes and potential networks of plant metabolism. Combining the mGWAS results with the gene expression data generated from RNA sequencing on the association panel, i.e. expression QTL (eQTL), we were able to obtain novel biochemical insights of maize kernels and also identified metabolite features associated with kernel weight that could be used as biomarkers for genetic improvement of maize [24•]. To facilitate the identification of candidate genes with natural variation not occurring at transcription level, another network approach was generated in Arabidopsis. In this approach, mGWAS candidates were refined with the gene–co-expression data derived from transcript accumulation within a single Arabidopsis accession (Col-0) across a wide range of developmental and environmental states [35•].

So far, the chemical identity of the majority of metabolites quantified by metabolomics techniques remains unknown, which hinders their further usability either as biomarkers of metabolic processes or targets for metabolic engineering. By linking the metabolites of unknown identity with functionally identified genes, mGWASs in both humans and plants have facilitated the large-scale identification or annotation of these unidentified metabolites (Figure 1). Integrating the data from genetic associations, metabolic networks and knowledge-based pathway information allowed deduction of the biochemical identities of 106 unknown metabolites in humans [55]. We recently applied a similar approach to experimentally identify or annotate over 160 metabolites covering a wide range of biochemical entries on the basis of their associations with the corresponding characterized genes, together with the information obtained from the fragmentation pattern of the metabolites [23••].

Concluding remarks and future perspectives

mGWAS in plants extends our knowledge on the genetic contribution to plant metabolism by depicting which DNA variants have significant impact on metabolic changes. mGWAS greatly facilitates large-scale interactive gene–metabolite annotation and identification in plants and provides insights into the genetic and biochemical basis of the plant metabolome [23••]. mGWAS also extends our knowledge on so far undiscovered biochemical pathways or pathway interactions, and provides a powerful tool for knowledge-based genetic improvement of desirable traits in plants. The past decade has seen the completion of genome sequencing of an increasing number of plants [56–62] and the constant decreasing cost of genotyping by next-generation sequencing. With the extension of metabolomics [63] it will not be long before we see mGWAS efficiently applied to a wide range of other plant species.

Although remarkable progress in mGWAS has been made in model and crop plants, to better understand the vast biochemical diversity of plant metabolomes and their interactions with other factors, advances in both metabolomics techniques and plant genetics are still needed.

In terms of metabolomics techniques, further improvements are needed to cover the wide array of metabolites with increased spatial and temporal resolution [64,65]. Another upcoming challenge is to develop databases [66–69] and bioinformatics tools for accurate annotation of metabolites. Despite the gains in understanding of genetics underlying complex traits, bridging this apparent genotype–phenotype gap remains a big challenge. A promising approach is to investigate the genetic basis of intermediate phenotypes such as metabolites and link these results back with the complex trait of interest [50]. Regarded as a readout of plant physiological status, a metabolic trait may either be a functional intermediate or a correlated biomarker. Although associations between metabolic and phenotypic traits have been reported, direct linkage between them is still lacking. Systematically overlaying mGWAS data with associations from GWAS or QTL mapping for phenotypic traits may allow the identification of new candidate SNPs and provide new genetic and biochemical interpretation of mechanisms underlying the phenotypic traits or identify potential biomarkers of the phenotypic traits.

Most of the mGWASs in humans were conducted using sample aliquots collected and stored from previous epidemiological studies [70]. Considering the huge diversity of plant metabolomes and that no single technique allows the measurement of all metabolites in one go, the strategy of collecting and stored samples for further analysis with a combination of different profiling platforms could also be applied to mGWAS in plants. Using ratios between metabolite concentrations as traits in mGWAS metabolite pairs that have genetic underpinning could substantially increase the strength of association, up to several orders of magnitude stronger than that of levels of the two individual metabolites [52]. Therefore, although computationally expensive, it is recommended that all ratios between metabolite pairs be tested for association. Till date, mGWASs have mostly focused on common variants within a population. Although it is suitable for screening a large number of accessions for genetic variation, joint mapping with association panels and multiple bi-parental crosses [14•,23••,24•] is likely to be more powerful in identifying alleles with low frequency or small effects in the population.
Acknowledgements
I apologize to those authors whose work could not be discussed due to the space limitations. I would like to thank Wei Chen for helping to prepare the manuscript and my colleagues for contributing to the crop mGWAS projects. Research in my laboratory was supported by the Ministry of Science and Technology of China (2013CB127001, 2012AA10A304 and 2012AA10A303) and by the National Natural Science Foundation of China (31070267 and 30970258).

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:

● of special interest
●● of outstanding interest

This report demonstrated that the incorporation of a high-density map in mQTL analyses with a widely-targeted profiling strategy resulted in substantially increased mapping resolution and allowed for genetic and biochemical inferences in agronomic model systems.

Using rice as a model, the authors presented the most comprehensive mGWAS in plants so far. The authors revealed hundreds of common variants with large effects for metabolites of physiological and nutritional importance. A number of candidate genes were identified combining mGWAS with reverse genetics and biochemical approaches. This strategy provides novel genetic and biochemical insights into rice metabolism and can also be used for metabolomics-assisted breeding.

Combining mGWAS with mQTL and eQTL, this paper revealed novel biochemical insights of metabolism in maize kernels and inferred the usage of metabolite features as biomarkers for important agronomical traits.
46. A case study of the distribution and variation of phenolamides in rice by GWAS with targeted profiling that revealed a number of significant loci, followed by the verification of two candidate genes underlying hydroxycinnamoyl spermidine by a transgenic approach.


