Chapter 3

Guidelines for Sample Normalization to Minimize Batch Variation for Large-Scale Metabolic Profiling of Plant Natural Genetic Variance

Saleh Alseekh, Si Wu, Yariv Brotman, and Alisdair R. Fernie

Abstract

Recent methodological advances in both liquid chromatography–mass spectrometry (LC-MS) and gas chromatography–mass spectrometry (GC-MS) have facilitated the profiling highly complex mixtures of primary and secondary metabolites in order to investigate a diverse range of biological questions. These techniques usually face a large number of potential sources of technical and biological variation. In this chapter we describe guidelines and normalization procedures to reduce the analytical variation, which are essential for the high-throughput evaluation of metabolic variance used in broad genetic populations which commonly entail the evaluation of hundreds or thousands of samples. This chapter specifically deals with handling of large-scale plant samples for metabolomics analysis of quantitative trait loci (mQTL) in order to reduce analytical error as well as batch-to-batch variation.

Key words Large-scale metabolomics, Batch normalization, Variation, Natural genetic variation, QTL mapping, LC-MS, GC-MS

1 Introduction

The metabolites of the plant kingdom are extremely diverse; a commonly quoted estimate is that plants produce somewhere in the order of 200,000 unique chemical structures [1]. Recently, there has been an increasing use the analytical technologies such as metabolomics for comprehensive profiling of metabolites in biological samples and its subsequent application in several related research areas such as human nutrition, drug discovery and plant breeding [2, 3]. Given the diversity of structural classes of metabolites, ranging from primary metabolites such as carbohydrates, amino acids, and organic acids to very complex secondary metabolites such as phenolics, alkaloids, and terpenoids, there is no single methodology that can measure the complete metabolome in one step. It is, therefore, often necessary to combine different techniques to detect (even a significant proportion of) all metabolites

Carla António (ed.), *Plant Metabolomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1778,
https://doi.org/10.1007/978-1-4939-7819-9_3,

within a complex mixture [4]. Both liquid chromatography–mass spectrometry (LC-MS) and gas chromatography–mass spectrometry (GC-MS) have been intensively used to profile a broad natural variance in the form of recombinant inbred lines (RILs), introgression lines (ILs) and, more recently, genome-wide association mapping panels in order to boost our understanding of the regulation of plant primary and secondary metabolite levels [5–7].

In all metabolomics applications, it is important to understand and control factors that contribute to sources of variation within the datasets. The variability between samples can arise from multiple sources including natural biological variation itself and that which occurs on sample collection and storage [8, 9]. In addition, analytical variation caused by suboptimal performance of the chosen apparatus, and instrument drift over time, are two major issues in large-scale metabolomics studies [10].

Batch-to-batch variation is a technical source of variation arising from the sum of both manual and robotic samples handling [11]. The presence of batch-to-batch variation makes it difficult to integrate data from independent batches of samples. This issue is particularly problematic when dealing with large number of samples such as is the case when analyzing structured plant populations.

To counter this, several normalization methods have been developed and suggested to overcome these issues and to minimize nonbiological variation [11–13]. For example normalizations by a single or multiple internal or external standard compounds based on empirical rules, such as specific regions of retention time have been used [2]. Similarly, isotope-labeled internal standard approaches were developed to monitor analytical error [14]. While there is no single best way to conduct metabolomics studies, there are a number of pitfalls and known problems that need to be carefully avoided. Detailed guidelines and normalization protocols have been previously published for this purpose [15–17].

In this chapter, we describe a workflow to minimize analytical errors and provide guidelines for handling large sample numbers for the specific purpose of metabolic quantitative trait loci (mQTL) approaches which utilize sources of broad natural genetic variance. We solely concentrate on aspects pertinent to the large-scale analysis of genetic populations and normalization aspects that need to be adopted to ensure proper cross-sample comparability as well as the downstream analysis of the data within the framework of quantitative loci and association mapping analyses.

2 Experimental Design for Large-Scale mQTL Approaches

In order to correctly evaluate such large sample sets it is important to manipulate variables under strictly controlled conditions while taking precise measurements. Therefore, the precision of an

experiment critically depends on the size of the experiment and the homogeneity of experimental materials. In large genomics experiments, the next step after choosing the population is to determine the number of lines and associated biological replicates. This is then followed by choosing the statistical approach to link genotype with phenotype. Here we neither focusing on the choices of the number of lines nor the population structure needed to obtain a complete genotype-to-phenotype matrix to identify all possible QTL [18] but rather on how many biological replicates are required per line for acceptable statistical analysis and data normalization.

The key to minimize technical sources of variation involves designing an experiment whereby several samples are taken per plant with multiple independent plants per parental genotype per replicate. Multiple independent replicates are conducted and all samples independently analyzed via metabolomics. Analysis of variance for this experiment will allow one to estimate the variation due to spatial differences within a plant, from differences between plants, from differences between replicate experiments and from differences between genotypes as well as any interactions between these different features. The optimum result is that most of the variance is due to genetic factors with the rest of the error being split between within replicates from the same plant or replicates between plants of the same genotype. If this is the case, it is best to take one measurement per plant with each line being represented by two or more plants per replicate.

In QTL analysis the number of replicates profiled will have a major influence on the reliability and reproducibility of the data and consequently on the QTL mapping results. Therefore, the ability to make broad conclusions or identify causal genes using quantitative studies of metabolic variation is greatly influenced by the fact that metabolic abundances measured in these studies are highly dependent on the environmental, developmental, and genetic variations present within the experiment as well as the experimental error. For these reasons and based on our own experience, it is recommended to use at least six independent biological replicates for each line (genotype) and many more control plants in a completely randomized design to overcome unavoidable effects associated with variation in microenvironmental factors such as light intensities, temperature and air humidity. This should be planed carefully in advance and he population size and time needed for collecting the samples should also take into account in order to ensure that harvesting is carried out in as homogeneous a manner as possible.

2.1 Plant Material and Sampling

Plant sampling (harvesting) is a crucial step in sample preparation for metabolomics, and much care needs to be afforded to it (*see* Chapter 1). The total variation in the dataset is a function of different sources of variation including variation introduced by differences in sample collection. Large scale experiments with vast

sample size and genotypes (e.g., ILs, RILs, or GWAS) which might slightly different in their developmental age adding yet another source of variation. However, the experimental design is key to any metabolomics experiment and having a large number of biological replicates is an essential means to minimize metabolite variation during sample preparation. In the case of introgression lines (ILs) a reasonable number of biological replicates is six independent plants the best strategy being to collect several different plant organs [3–5] per biological replicate pool them and treat them as a single sample. In the case of other population such as RILs, BILs and GWAs less replication is needed than in the ILs since in these populations genetic variance is represented in multiple lines, as opposed to a single line, within the population.

Most metabolomics studies are carried out in the laboratory under highly controlled conditions. However, most mQTL studies have been carried out for crop species such as maize, tomato, and rice have been conducted in the field. For this reason and in order to minimize the variation there are several crucial points to take into consideration during harvest.

Given that the levels of metabolites vary through the day, and that some experiments are too large to allow harvest in a single day it is essential to harvest control samples for each temporally separate harvest. Also as mentioned above plant metabolomics experiments are generally performed at the organ level (developing fruit, whole leaf, root, etc.), and it is recommended to have pooled samples per replicate to reduce the level of within genotype variation. These issues are especially important when the harvest sessions of a given experiment are numerous or when each session requests several people harvesting to limit its duration. The age, or preferably the developmental stage, of the plants or their organs needs to be defined relative to standardized growth conditions and/or phenology descriptors, by using dedicated ontology's (Plant Ontology at http://www.plantontology.org/ for phenology) or reference articles for Arabidopsis [19] or tomato [20] when available.

2.2 Sample Processing and Extraction

After harvesting, plant organs (e.g., leaves, flowers, or fruits) or dissected tissues, plant should be immediately frozen in liquid nitrogen and stored at −80 °C, or immediately ground to a powder and extracted. Sample grinding is usually required to optimize solvent extraction and additionally aids in the homogenization of the sample material [21]. It is recommended that all samples for a given experiment follow exactly the same procedure before, during, and after grinding. For further reading on extraction protocols available for plant metabolomics we suggest the work of Shimizu et al. (*see* Chapter 12) for LC-MS and the comprehensive work of Osorio and colleagues for GC-MS [22].

However, there are some important points at which these protocols should be adapted when handling the large number of plant

samples required for QTL analysis. First, quality control (QC) is necessary throughout the entire sample preparation process, from the field to the sample storage location and through distribution to chemical analysts for data normalization to reduce the analytical errors. The quality control (QC) samples should qualitatively and quantitatively representative the entire collection of samples included in the study, providing an average of all of the metabolomes analyzed in the study. Sample prepared by pooling aliquots of individual study samples, either all or a subset representative for the study. The QC sample has (should have) an identical or a very similar (bio) chemical diversity as the study samples. The QC samples are evenly distributed over all the batches and are extracted, derivatized, and analyzed at the same time as the individual study samples as part of the total sequence order. The data from the QC samples is used to monitor drift, separate high- and low-quality data, equilibrate the analytical platform, correct for drift in the signal and allow the integration of multiple analytical experiments. The data analysis technique such as principal component analysis can be used to quickly assess the reproducibility of the QC samples in an analytical run. The QC samples are used to determine the variance of a metabolite feature.

Before extraction QC samples should prepared by pooling aliquots of individual study samples, the QC samples should then be distributed across all machine-batches and aliquots thereof should be extracted, derivatized, and analyzed at the same time as the individual study samples (Fig. 1).

2.3 MS-Based Metabolomics Analysis (LC-MS, GC-MS)

Once the extraction has been made, extracts must be subsequently prepared for MS-based analysis. In the case of LC-MS, once the samples are extracted aliquots of the extract can be directly introduced into the LC-MS apparatus (*see* Shimizu et al., Chapter 12). In GC-MS-based metabolomics, however, additional preparation steps are necessary to confer volatility to the metabolites via silylation and to simplify chromatography of sugars via methoxyamination [16, 17, 22]. We recommend dividing the samples in batches so that each batch contains 50–80 samples with ample QC samples distributed across the sequence run (Fig. 1). Metabolite profiling via GC-MS involves several general steps [23]. After derivatization, automated sample injection robotics and separated in GC in highly standardized conditions of gas flow, temperature programming, and standardized capillary column material. Electron impact (EI) is the most widely used ionization technique applied in GC-MS. Mass separation and detection is achieved preferably by TOF detectors that can be tuned to fast scanning rates, and finally acquisition and evaluation of GC–MS data files.

In the case of LC-MS-based metabolomics approaches, the most frequently used protocols use C18-based reversed phase columns coupled to soft ionization techniques, such as electrospray

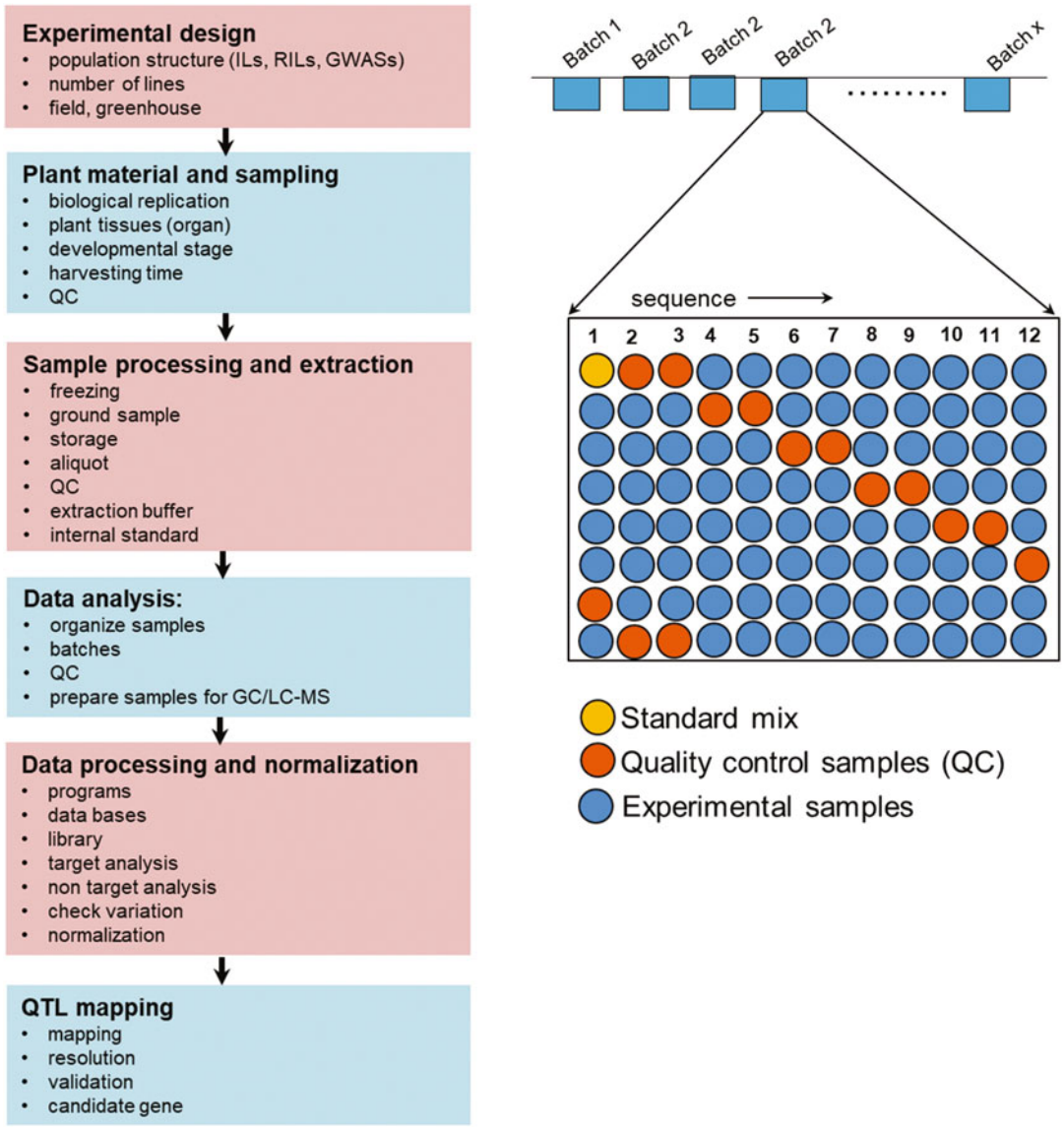

Fig. 1 Flowchart of the metabolomics study in plants. Left panel represent the different steps for experimental design, sample preparation and process for QTL experimental study. The left panel shows sample organization and suggested sequence running in GC-MS or LC-MS

ionization (ESI) or atmospheric pressure chemical ionization (APCI), resulting in protonated (in positive mode) or deprotonated (in negative mode) molecular ions. Modern high resolution instruments with exact mass detection, such as TOF-MS, ion cyclotron FT-MS, or Orbitrap FT-MS, nowadays enable the profiling of hundreds to thousands of compounds in plant extracts, combined with elemental formulae calculations of the detected masses [24, 25].

3 Data Processing

Once samples are analyzed, automatic data processing tools are required for peak picking and mass peak alignment. In GC-MS several tools, software and databases have been established and used for this purpose [23, 26, 27]. For further details on data processing from LC-MS metabolomics data see Shimizu et al. (Chapter 12). Chromatograms obtained from e.g., UPLC-FT-MS runs can be analyzed and processed with REFINER MS® 10.0 (GeneData, http://www.genedata.com), where molecular masses, retention time (RT), and associated peak intensities for each sample are extracted from the .raw files. The chemical noise is subtracted automatically. The chromatogram alignments are performed using a pairwise alignment-based tree using *m/z* windows of five points and RT windows of five scans within a sliding frame of 200 scans. Further processing of the MS data includes isotope clustering, adduct detection, and library searches. Resulting data matrices with peak ID, RT, and peak intensities in each sample are generated. However, for both LC- and GC-MS methods, manual checking of peaks is strongly recommended.

4 Data Normalization

The goal of metabolomics as a phenotyping platform depends on its ability to detect biologically related metabolite changes in complex biological samples. As with any high-throughput technology, systematic biases are often observed in LC-MS and GC-MS metabolomics data [26, 27]. As the number of samples in the dataset increases there is a corresponding time-dependent variation in the metabolite data. The variability in samples can arise from multiple sources including physiological differences and variability from the analytical method itself. Removing platform-specific sources of variability such as systematic errors is one of the top priorities in metabolomics data preprocessing. However, metabolite diversity leads to different responses to variations at given experimental conditions, making normalization a very demanding task [27]. For the effective elimination of different sources of analytical variation, preprocessing steps should follow a specific sequence.

The first step in data normalization is using an internal standard (IS); a compound added to the sample before a critical step in the analysis. An IS is not necessarily an isotope-labeled version of an analyte. However, it can be structurally related to one or more analytes, but not naturally occurring in the samples of interest. This normalization step reduces the differences in sample extraction (which can be caused by slight differences in the composition of the samples and also differences in the volumes injected). The

second step is the removal of between-batch and within-batch variations and machine drifts. The final steps consist of the combination of data from replicate sample analysis and removal of noise and biomass correction. The biomass correction neutralizes differences in response due to sample weight or volume.

Here the QC samples are of key importance, and these are best prepared by pooling equal volumes of material from all of the biological samples to be analyzed. Alternatively, a chemically defined mixture of authenticated reference compounds [28] that mimics the metabolic composition of the investigated biological material can be employed. Both the synthetic mixtures and biological QC samples are then subjected to the same sample extraction, instrumental analyses (ideally distributed across the analytical run), and data processing, thus providing quality checks for technical and analytical error, and quantitative calibration to eliminate batch effects for the final processed data. This normalization is a crucial step for minimizing the batch-to-batch data variability across extended periods. As such this is a crucial requirement for large-scale phenotyping and facilitates interbatch data integration.

5 QTL Mapping

The principle of quantitative trait locus (QTL) mapping is based of detecting association of molecular genetic markers with the phenotype of interest in the resultant offspring [29]. Markers are used to partition the mapping population into different genotypic groups based on the presence or absence of a particular marker locus and to determine whether significant differences exist between groups with respect to the trait being measured [27]. If a QTL is linked to a marker locus, then the individuals with different marker locus genotypes will have different mean values of the quantitative trait. In plants, the use of such mapping populations is highly useful since the use of stable populations permits the growth of clonal replicates and, additionally, multiple analyses of genetically identical individuals across multiple harvests. There are several structural populations and methods have been used to detect the QTL and mapping. Therefore, choosing the proper population for such experiments is a key determinant in the success of any given project. There are several factors influencing the detection of QTL detections that should be considered in advance of planning such experiments.

Factors influencing QTL mapping: the genetic properties of QTL controlling traits include the magnitude of the effect of individual QTL. Only QTL with sufficiently large phenotypic effects will be detected; and the QTL with small effects may fall below the significance threshold of detection. Another genetic property is the distance between linked QTL; QTL that are closely linked will usually be detected as a single QTL in typical population

sizes (<500) [30–32]. The environmental effects may have a large influence on the expression of quantitative traits. The size of the population used in the mapping study is also highly important; the larger the population, the more accurate the mapping study and the more likely it is to allow detection of QTL with smaller effects.

Owing to the factors and variations described above, QTL mapping studies should be independently confirmed or validated. Such confirmation studies (referred to as validation or replication studies) can be achieved by repeating the experiment and the QTL mapping at different sites, seasons, or years. The conserved detected QTL throughout several repeated experiment most likely the QTL that have strong genetic effect (high heritability) and that can be chosen as a region to focus on in further analysis. A second type of validation may involve independent populations constructed from the same parental genotypes or closely related genotypes used in the primary QTL mapping study. Once an association between a particular SNP and variation in a trait of interest has been established, a crucial but yet too often overlooked step is to replicate the association in an independent mapping population. As the number of studies documenting significant associations between SNPs and variation in quantitative traits of interest accumulates, increasing emphasis should be placed on replicating studies to validate effects of significant associations. In the following sections we briefly define some of the commonly used structural populations for the QTL mapping.

5.1 RIL Mapping

In plant species, the uses of immortal mapping populations consisting of homozygous individual have been used to map loci for complex traits. Recombinant inbred lines (RILs) (Fig. 2) can be obtained relatively easily and are produced by successively selfing the progeny of individual F2 plants (single seed descent method), from which the F8 generation and onward are practically homozygous lines that will produce further progeny that is essentially identical to the previous generation. Such a population can also be produced by induced chromosomal doubling of haploids, such as for doubled haploids (DHs) [33–35]. RILs are likely advantageous over DHs since they are characterized by a higher frequency of recombination within the population, resulting from multiple meiotic events occurred during repeated selfing [36]. Candidate mutations (such as a SNP, illustrated by the red dots) are then identified.

5.2 IL Mapping

Another type of immortal population consists of introgression lines (IL) (Fig. 2) which are obtained through repeated backcrossing and extensive genotyping. These can also be referred to as near isogenic lines (NILs) [37] or backcross inbred lines (BILs) [38, 39]; although the latter are slightly different in nature. These lines

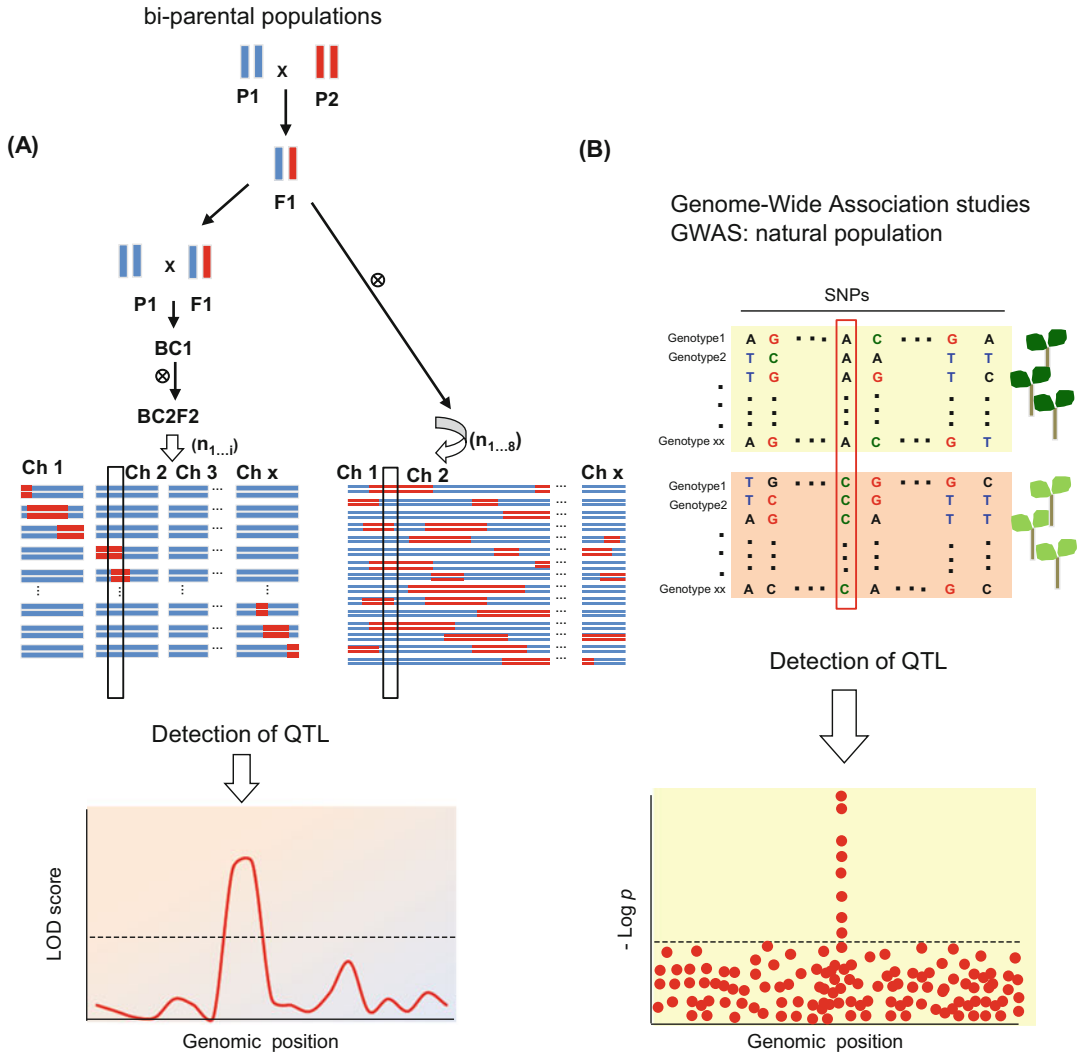

Fig. 2 Quantitative trait locus mapping. (**a**) Comparison of introgression and recombinant inbred lines; which created by backcrossing an F1 of a cross between two parental lines to a recurrent parent for several times. Recombinant inbred lines are generated by selfing an F1 for at least eight generations when full homozygosity is reached. (**b**) Genome-Wide Association Study (GWAS) for Identification of QTL using natural population; GWAS makes use of natural genotypic variation and enables the analysis of associations between hundreds of thousands of single nucleotide polymorphisms (SNPs) and specific traits

contain a single or a small number of genomic introgression fragments from a donor parent into an otherwise homogeneous genetic background.

5.3 GWAs Mapping

The IL and RILs have historically been the most common types of experimental populations used for the analysis of quantitative traits and powerful method to identify regions of the genome that cosegregate with a given trait. However, they suffer from some limitations; only allelic diversity that segregates between the parents of

the particular F2 cross or within the RIL population can be assayed [27], and second, the amount of recombination that occurs during the creation of the RIL population places a limit on the mapping resolution [40]. The basic principle of genome-wide association studies (GWAS) (Fig. 2), which was initially developed for use in medical genetics, is that the incidence of nucleotide polymorphisms is associated with the presence of variance is overcome the limitations of using the IL and RILs. This approach has several major advantages over conventional QTL mapping. First, a much larger and more representative gene pool can be surveyed. Second, it bypasses the expense and time of mapping studies and enables the mapping of many traits in one set of genotypes. Third, a much finer mapping resolution can be achieved, resulting in small confidence intervals of the detected loci compared to classical mapping, where the identified loci need to be fine-mapped. Finally, it has the potential not only to identify and map QTLs but also to identify the causal polymorphism within a gene that is responsible for the difference in two alternative phenotypes [27]. A major issue with association studies is false positives, and the main sources of such false positives are linkage between causal and noncausal sites [41, 42].

QTL analysis is predicated by looking for associations between the quantitative trait and the marker alleles segregating in the population. Classical mapping approaches use segregating populations, such as recombinant inbred lines (RIL) and introgression lines (IL), to shed light on the genetic contributions to diverse phenotypes. Both are obtained from a cross (F1) of two parental accessions (P1 and P2), through repeated selfing (RIL) or backcrossing of the initial hybrid with one of its parents (IL) followed by a selfing until a homozygous state is reached. Alternate alleles at loci in homozygous lines with distinct genetic basis in such structured populations potentially influence the trait of interest and allow to map the genomic loci responsible for the observed intraspecific or interspecific variation (Fig. 2a).

GWAS makes use of natural genotypic variation and enables the analysis of associations between hundreds of thousands of single nucleotide polymorphisms (SNPs) and specific traits. For Identification of QTL; DNA obtained from hundreds of natural genetic accessions and tested for genetic variations, like SNPs. If certain SNPs are found significantly more frequently in group of genotypes (accessions) with a certain phenotype (trait) than in the general population, the mutations are said to be "associated" with the trait. The GWAS analysis, represented in a Manhattan plot with significance ($-\log_{10}$ (P value) on the y-axis, and genomic position shown as chromosomes in the x-axis, is done to look for genetic variants that are associated with certain trait in a group of genotypes (leaf color for example) but not found among the other group. Significant variants (positive z-scores) are represented by the red dots. The

dot that rises above background variation and is significantly associated with the phenotype is represented in the Manhattan plot at chromosomal position (Fig. 2b).

6 Conclusions

Both GC-MS and LC-MS are widely used analytical tools for profiling highly complex mixtures of primary and secondary metabolites. High-throughput use of these techniques is faced with a large number of potential sources of nonbiological variation that can compromise the interpretation of the results. However, by following several recommendations prior to and during the conductance of large-scale genomics and QTL mapping experiments such problems can be circumvented in a relatively facile manner.

Acknowledgments

This wok was in part supported by the PlantaSYST project by the European Union's Horizon 2020 Research and Innovation Programme (SGA-CSA Number 664621 and Number 739582 under FPA Number 664620).

References

1. Dixon RA, Strack D (2003) Phytochemistry meets genome analysis, and beyond. Phytochemistry 62:815–816
2. Bijlsma S, Bobeldijk I, Verheij ER et al (2006) Large-scale human metabolomics studies: a strategy for data (pre-)processing and validation. Anal Chem 78:567–574
3. Schauer N, Semel Y, Roessner U et al (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nat Biotechnol 24:447–454
4. Fiehn O (2002) Metabolomics - the link between genotypes and phenotypes. Plant Mol Biol 48:155–171
5. Fu J, Keurentjes JJ, Bouwmeester H et al (2009) System-wide molecular evidence for phenotypic buffering in Arabidopsis. Nat Genetics 41:166–167
6. Rowe HC, Hansen BG, Halkier BA et al (2008) Biochemical networks and epistasis shape the Arabidopsis thaliana metabolome. Plant Cell 20:1199–1216
7. Wentzell AM, Rowe HC, Hansen BG et al (2007) Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. PLoS Genet 3:1687–1701
8. Biais B, Bernillon S, Deborde C et al (2012) Precautions for harvest, sampling, storage, and transport of crop plant metabolomics samples. In: Hardy N, Hall R (eds) Plant Metabolomics, Methods in Molecular Biology (Methods and Protocols), vol 860. Humana Press, New York, pp 51–63
9. Gibon Y, Rolin D (2012) Aspects of experimental design for plant metabolomics experiments and guidelines for growth of plant material. Methods Mol Biol 860:13–30
10. Sysi-Aho M, Katajamaa M, Yetukuri L, Oresic M (2007) Normalization method for metabolomics data using optimal selection of multiple internal standards. BMC Bioinformatics 15:8–93
11. van der Kloet FM, Bobeldijk I, Verheij ER, Jellema RH (2009) Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping. J Proteome Res 8:5132–5141
12. van der Greef J, Martin S, Juhasz P et al (2007) The art and practice of systems biology in

medicine: Mapping patterns of relationships. J Proteome Res 6:1540–1559

13. Dunn WB, Broadhurst D, Brown M et al (2008) Metabolic profiling of serum using ultra performance liquid chromatography and the LTQ-orbitrap mass spectrometry system. J Chromatogr B Anal Technol Biomed Life Sci 871:288–298
14. Chen MJ, Rao RP, Zhang Y et al (2014) A modified data normalization method for GC-MS-based metabolomics to minimize batch variation. Spring 3:439
15. Fiehn O, Kopka J, Dörmann P et al (2001) Metabolite profiling for plant functional genomics. Nat Biotechnol 18:1157–1161
16. Lai Z, Fiehn O (2016) Mass spectral fragmentation of trimethylsilylated small molecules. Mass Spectrom Rev 9999:1–13
17. Lisec J, Schauer N, Kopka J et al (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. Nat Protocols 1:387–396
18. Joseph B, Corwin JA, Kliebenstein DJ (2015) Genetic variation in the nuclear and organellar genomes modulates stochastic variation in the metabolome, growth, and defense. PLoS Genet 11:e1004779
19. Boyes DC, Zayed AM, Ascenzi R et al (2001) Growth stage-based phenotypic analysis of arabidopsis: a model for high throughput functional genomics in plants. Plant Cell 13:1499–1510
20. Brukhin V, Hernould M, Gonzalez N et al (2003) Flower development schedule in tomato Lycopersicon esculentum cv. sweet cherry. Sex Plant Reprod 15:311–320
21. Markert B (1995) Sample preparation (cleaning, drying, homogenization) for trace element analysis in plant matrices. Science Total Environ 176:45–61
22. Osorio S, Do PT, Fernie AR (2012) Profiling primary metabolites of tomato fruit with gas chromatography-mass spectrometry. In: Hardy N, Hall R (eds) Plant Metabolomics, Methods in Molecular Biology (Methods and Protocols), vol 860. Humana Press, New York, pp 101–109
23. Kopka J, Fernie A, Weckwerth W et al (2004) Metabolite profiling in plant biology: platforms and destinations. Genome Biol 5:109
24. Allwood JW, De Vos RC, Moing A et al (2011) Plant metabolomics and its potential for systems biology research background concepts, technology, and methodology. In: Jameson D, Verma M, Westerhoff HV (eds) Methods in Enzymology, vol 500. Academic Press, Amsterdam, pp 299–33623
25. Allwood JW, Clarke A, Goodacre R, Mur LA (2010) Dual metabolomics: a novel approach to understanding plant-pathogen interactions. Phytochemistry 71:590–597
26. Karpievitch YV, Nikolic SB, Wilson R et al (2014) Metabolomics data normalization with EigenMS. PLoS One 9:e116221
27. Sehgal D, Singh R, Rajpal VR (2016) Quantitative trait loci mapping in plants: concepts and approaches. In: Rajpal V, Rao S, Raina S (eds) Molecular breeding for sustainable crop improvement. sustainable development and biodiversity, vol 11. Springer, Cham
28. Strehmel N, Hummel J, Erban A et al (2008) Retention index thresholds for compound matching in GC-MS metabolite profiling. J Chromatogr B AnalTechnol Biomed Life Sci 871:182–190
29. Broman KW (2001) Review of statistical methods for QTL mapping in experimental crosses. Lab Anim 30:44–52
30. Tanksley SD (1993) Mapping polygenes. Ann Rev Genetics 27:205–233
31. Collard BCY, Pang ECK, Taylor PWJ (2003) Selection of wild Cicer accessions for the generation of mapping populations segregating for resistance to ascochyta blight. Euphytica 130:1–9
32. Soltis NE, Kliebenstein DJ (2015) Natural variation of plant metabolism: genetic mechanisms, interpretive caveats, and evolutionary and mechanistic insights. Plant Physiol 169:1456–1468
33. Han F, Ullrich SE, Kleinhofs A et al (1997) Fine structure mapping of the barley chromosome-1 centromere region containing malting-quality QTLs. Theoretical Applied Genetics 95:903–910
34. Rae AM, Howell EC, Kearsey MJ (1999) More QTL for flowering time revealed by substitution lines in Brassica oleracea. Heredity 83:586–596
35. von Korff M, WJ LK, Pillen K (2004) Development of candidate introgression lines using an exotic barley accession (Hordeum vulgare ssp spontaneum) as donor. Theoretical Applied Genetics 109:1736–1745
36. Balding DJ, Bishop M, Cannings C, Jansen RC (2004) Quantitative Trait Loci in Inbred Lines. In: Balding DJ, Bishop M, Cannings C (eds) Handbook of Statistical Genetics, 3rd edn. John Wiley & Sons Ltd, Chichester, UK
37. Monforte AJ, Tanksley SD (2000) Development of a set of near isogenic and backcross recombinant inbred lines containing most of the Lycopersicon hirsutum genome in a L-esculentum genetic background: A tool for

gene mapping and gene discovery. Genome 43:803–813

38. Jeuken MJW, Lindhout P (2004) The development of lettuce backcross inbred lines (BILs) for exploitation of the Lactuca saligna (wild lettuce) germplasm. Theoretical Applied Genetics 109:394–401
39. Blanco A, Simeone R, Gadaleta A (2006) Detection of QTLs for grain protein content in durum wheat. Theoretical Applied Genetics 113:563–565
40. Jamann TM, Balint-Kurti PJ, Holland JB (2015) QTL mapping using high-throughput sequencing. In: Alonso J, Stepanova A (eds) Plant functional genomics, Methods in molecular biology, vol 1284. Humana Press, New York
41. Platt A, Vilhjalmsson BJ, Nordborg M (2010) Conditions under which genome-wide association studies will be positively misleading. Genetics 186:1045–1052
42. Larsson SJ, Lipka AE, Buckler ES (2013) Lessons from Dwarf8 on the strengths and weaknesses of structured association mapping. PLoS Genet 9:e1003246