Max-Planck-Institut für Molekulare Pflanzenphysiologie
Arbeitsgruppe von Thomas Altmann

Identification and characterization of metabolic Quantitative Trait Loci (QTL) in *Arabidopsis thaliana*

Dissertation
zur Erlangung des akademischen Grades
"doctor rerum naturalium" (Dr. rer. nat.)

eingereicht im
Institut für Biochemie und Biologie an der
Mathematisch-Naturwissenschaftlichen Fakultät der
Universität Potsdam

von Jan Lisec
Potsdam, den 09.05.2008

Online published at the
Institutional Repository of the Potsdam University:
http://opus.kobv.de/ubp/volltexte/2008/2590/
urn:nbn:de:kobv:517-opus-25903
[http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-25903]

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und unter Verwendung keiner anderen als den von mir angegebenen Quellen und Hilfsmitteln verfasst habe.
Ferner erkläre ich, dass ich bisher weder an der Universität Potsdam noch anderweitig versucht habe, eine Dissertation einzureichen oder mich einer Doktorprüfung zu unterziehen.

Jan Lisec
Potsdam, den 09.05.2008

Table of Contents

Summary

Plants are the primary producers of biomass and thereby the basis of all life. Many varieties are cultivated, mainly to produce food, but to an increasing amount as a source of renewable energy. Because of the limited acreage available, further improvements of cultivated species both with respect to yield and composition are inevitable. One approach to further progress in developing improved plant cultivars is a systems biology oriented approach.

This work aimed to investigate the primary metabolism of the model plant *A.thaliana* and its relation to plant growth using quantitative genetics methods. A special focus was set on the characterization of heterosis, the deviation of hybrids from their parental means for certain traits, on a metabolic level. More than 2000 samples of recombinant inbred lines (RILs) and introgression lines (ILs) developed from the two accessions Col-0 and C24 were analyzed for 181 metabolic traces using gas-chromatography/ mass-spectrometry (GC-MS). The observed variance allowed the detection of 157 metabolic quantitative trait loci (mQTL), genetic regions carrying genes, which are relevant for metabolite abundance. By analyzing several hundred test crosses of RILs and ILs it was further possible to identify 385 heterotic metabolic QTL (hmQTL).

Within the scope of this work a robust method for large scale GC-MS analyses was developed. A highly significant canonical correlation between biomass and metabolic profiles ($r = 0.73$) was found. A comparable analysis of the results of the two independent experiments using RILs and ILs showed a large agreement. The confirmation rate for RIL QTL in ILs was 56 % and 23 % for mQTL and hmQTL respectively. Candidate genes from available databases could be identified for 67 % of the mQTL. To validate some of these candidates, eight genes were re-sequenced and in total 23 polymorphisms could be found. In the hybrids, heterosis is small for most metabolites (< 20%). Heterotic QTL gave rise to less candidate genes and a lower overlap between both populations than was determined for mQTL. This hints that regulatory loci and epistatic effects contribute to metabolite heterosis.

The data described in this thesis present a rich source for further investigation and annotation of relevant genes and may pave the way towards a better understanding of plant biology on a system level.

1 Introduction

1.1 Metabolomics

The metabolome is the entirety of small molecules present in an organism and can be regarded as the ultimate expression of its genotype in response to environmental changes. Since the term was coined in 1998 (Oliver *et al.*, 1998), the technology to qualify and quantify an ever increasing part of the metabolome was developed rapidly along with the complementary 'omics' approaches measuring transcript (transcriptomics) and protein (proteomics) abundances. Compared to the latter, information obtained in metabolomics analyses is regarded to closer mirror the biological endpoint (Lindon *et al.*, 2004) and, as such, is perhaps more relevant to our understanding of how a plant exists, functions and responds within its own environment (Hall, 2006).

While single metabolites have been measured by spectrophotometric assays or simple chromatographic separation for a long time (Fernie *et al.*, 2004), the analysis of several hundreds to thousands of compounds only started to become feasible with the hyphenation of separation methods to various detection systems. The separation methods which are commonly applied include gas chromatography (GC), liquid chromatography (LC) and capillary electrophoresis (CE). Different types of mass spectrometry (MS), nuclear magnetic resonance (NMR) and ultraviolet light spectroscopy (UV/VIS) devices are utilized for detection. Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) as a special case is often used in direct infusion (DI) mode, as its high mass accuracy allows a separation solely based on this parameter. All available methods exhibit various advantages and disadvantages and the method of choice highly depends on the biological question to be answered.

NMR surpasses other methods if the determination of metabolite structures for highly abundant compounds in a non-destructive way is required. However, compared to mass spectrometry, NMR systems have a low sensitivity and are cost intensive. Mass spectrometry based methods are very suitable for plant metabolomics due to their robustness and high sensitivity, which is necessary given the complexity of samples and the dynamic range of metabolite concentrations.

Liquid chromatography if compared to gas chromatography allows the analysis of a broader range of metabolites from plant extracts including secondary metabolites. Still, problems with ion suppression effects for the often utilized electron spray ioniza-

tion (ESI), a low chromatographic reproducibility and the lack of reference databases have to be considered.

Gas chromatography, by principle, is limited to the fraction of small molecules which are volatile or can be rendered volatile through chemical derivatization. This includes most of the metabolites involved in primary metabolism, such as - but not limited to - hydroxy- and amino acids, sugars, sugar alcohols, organic monophosphates, (poly)amines, sterols and fatty acids (Fiehn *et al.*, 2000a).

The biochemical diversity in the plant kingdom is estimated to well exceed 100,000 distinct compounds (Weckwerth, 2003) and 4,000 to 20,000 metabolites per species seem likely (Fernie *et al.*, 2004). None of the above mentioned methods is capable of measuring all the metabolites that can be expected to be present in a plant sample. The chemical complexity and range of concentrations render a comprehensive analysis impossible. Published studies reported on identifying from 76 up to more than 2,000 metabolites or metabolic mass traces with several hundreds of samples per experiment (Keurentjes *et al.*, 2006; Meyer *et al.*, 2007a; Schauer *et al.*, 2006), demonstrating recent technological advances. A variety of software tools aiding in the unbiased or reference based evaluation of metabolomics experiments are developed to keep up with that progress (Fiehn *et al.*, 2005; Luedemann *et al.*, 2008; Smith *et al.*, 2006; Styczynski *et al.*, 2007; Vos *et al.*, 2007).

Principal strategies in metabolomics include (i) fingerprinting approaches, where a high number of metabolic traces of known and unknown structure are evaluated and used for sample classification; (ii) metabolic profiling, which aims to annotate and possibly quantify as many metabolites as possible and (iii) targeted analysis, which include a rather limited number of metabolites chosen on prior knowledge (Fiehn, 2002; Hall, 2006).

1.2 The use of immortal populations and natural variation in plant quantitative genetics

The observed phenotype of any organism can be assumed to be the product of its genetic disposition and the influence of its environment. Likewise, the observed variation (V_P) of any phenotypic trait within a population of organisms can be partitioned into the variance components attributable to genotype (V_G), environment (V_E) and their interaction ($V_G \times V_E$). Quantitative genetics provides a mathematical foun-

dation which allows to elucidate the mechanisms underlying V_G by extending the principles of Mendelian genetics to quantitative traits and populations.

Quantitative traits are determined on a genetic level by two to potentially hundreds of loci. Single contributions are therefore small and can not be identified by their segregation in individual lines but rather due to gene frequencies in populations. Recombinant inbred lines (RILs) and introgression lines (ILs) also termed near isogenic lines (NILs) are two population types which are frequently used to investigate these traits and the underlying genes. (Salvi and Tuberosa, 2005) Both are developed from a cross (F1) of two parental accessions (P1, P2) which are preferably homozygous and genetically distinct to ensure alternate alleles at loci potentially influencing the trait of interest.

Measuring genetic diversity can be achieved by classical methods like pedigree analysis and morphological, physiological or cytological markers (Melchinger, 1999). With the advent of modern screening technologies restriction fragment length polymorphisms (RFLPs) and simple sequence repeats (SSRs or microsatellites) allowed to increase the number of markers accessible and ensured an evenly distribution over the genome. Today, single nucleotide polymorphisms (SNPs) became the method of choice due to the ease of genotyping (Wang *et al.*, 1998) and their low mutation rates (Kruglyak, 2008). The HapMap project, which aims to identify all polymorphisms present in humans, was for instance able to identify more than 3 million SNPs (International HapMap Consortium, 2007). In Arabidopsis a resequencing effort led to the prediction of over 1 million SNPs (Clark *et al.*, 2007). The progress in SNP detection technologies facilitated the generation and genotyping of large RIL and IL populations.

RILs are generated from a cross (F1) through repeated selfing. The number of selfing steps determines the degree of inbreeding and is usually continued until a homozygous state is reached. Since each line is the progeny of an individual initial hybrid, different sections of the parental genomes are fixed during the inbreeding process, so that all lines, on average, bear half of the genetic information from each parent but are otherwise genetically diverse. Consequently, each allele for genes where both parents differed (and each marker) will be represented in the resulting RIL population with a frequency of 0.5. If such a population is cultivated under equal conditions and analyzed for a certain trait, a significant difference may be found between two

subpopulations that can be formed based on any of these markers. It can then be concluded that this marker is linked to one or several genes causing this difference.

Introgression lines (ILs) are generated by repeated backcrossing of the initial hybrid with one of its parents (e.g. P1). The number of backcrossing steps determines the amount of the genetic information from the other parent (P2) still present in the introgression line. A final selfing step produces a homozygous line, carrying one or several short fragments of a donor genotype (here P2) in the background of a recipient (here P1). A useful IL population for genome wide analysis would comprise two reciprocal populations of distinct lines which inherit introgression that together cover the full genome. If any of these ILs are grown together with the corresponding parent, thus minimizing environmental influences, observed significant differences can be contributed to the introgressed fragment alone.

Because RILs and ILs are 'immortal' and seeds can be produced easily by selfing they are well suited for replicated experiments under different environmental conditions or analyzing different traits of interest. Backcrosses (BC) or testcrosses (TC) between any RIL_i or IL_i with both of its parents and the parent hybrid can be generated. These experimental designs are known as North Carolina Design III (Comstock and Robinson, 1952) when only backcrosses with both parents are produced, and triple testcross design (TTC) (Kearsey and Jinks, 1968), when an additional backcross to the parent hybrid is generated. They allow to estimate additive, dominant and, in case of TTC, epistatic variation.

RIL and IL populations are available for several plant species like tomato, rice, maize and Arabidopsis (Alonso-Blanco *et al.*, 1998c; Burr *et al.*, 1988; Eshed and Zamir, 1995; Li *et al.*, 1995) and have been widely examined to investigate quantitative traits (cf. 1.3).

An alternative with respect to genome wide analyses is association mapping, which is based on linkage disequilibrium (LD) in wild strains. In Arabidopsis LD decays rapidly (<10kb) in a sample of accessions selected for maximum genetic diversity possibly allowing near-gene-level resolution in mapping approaches (Kim *et al.*, 2007). However, spurious associations which can arise due to an underlying population structure have to be taken into account.

1.3 Quantitative trait analyses

A particular case which renders the aforementioned populations of recombinant inbred lines and introgression lines useful is the investigation of quantitative traits which are genetically determined by multiple loci (quantitative trait loci, QTL). As was pointed out in the previous paragraph it is a straight forward approach to identify a chromosomal fragment influencing a trait of interest in ILs. The quantitative nucleotide (QTN) which is causing the observed difference can be confirmed by complementation tests or positional cloning once the size of the fragment has been narrowed down to a few kb using subILs generated from the initial candidate line.

In RIL populations QTL mapping was classically carried out as single marker analysis, testing the association between each marker and a trait, assuming linkage of this marker to a gene influencing this trait if association was confirmed in a test statistic. Interval mapping (IM) based on maximum likelihood (Lander and Botstein, 1989) or multiple regression (Haley and Knott, 1992) allowed estimating QTL positions and effects in continuous intervals throughout the whole genome, taking the distance and expected recombination frequencies between markers into account. Composite interval mapping (CIM) (Zeng, 1994) further increased precision of estimates by incorporating determined QTL as cofactors into the calculations. Multiple interval mapping (MIM) (Kao *et al.*, 1999) was introduced to analyze all potential QTL in a full model hereby allowing to include epistatic effects between loci.

Permutation tests as suggested by Churchill and Doerge (1994) are widely used to compute significance thresholds and a variety of computational tools to carry out the actual calculations is publicly available (MAPMAKER/QTL (Lincoln *et al.*, 1992), QTL Cartographer (Basten *et al.*, 1994), PLABQTL (Utz and Melchinger, 1996), R/QTL (Broman *et al.*, 2003)).

Some inherent limitations of QTL analyses using RILs exist. Confidence intervals reported in literature rarely drop below ~10 cM (equivalent to ~300 kbp in Arabidopsis), that is they contain hundreds of genes. Effects are likely to be overestimated, since only significant effects are retained in the model (Kearsey and Farquhar, 1998). Gene clustering of several loci influencing a trait, a situation which is known for genes controlling e.g. floral traits (Bernacchi and Tanksley, 1997), hampers the mapping process in two ways: If genes act synergistically several small effects may appear as a single, strong QTL but if they bear opposing effects, no QTL at all may be detectable. Another concern is that detected QTL could be caused by

environmental interactions and thus impede their confirmation at different conditions. It is still a debate to which extent QTL detected in one population can be confirmed in a second or, even more severe, in a different species, a feature that would be necessary to allow general conclusions.

The most important factors determining the success in QTL analyses are the heritability of the trait, the number of genes involved, their individual contribution and the size of the mapping population.

The number of genes identified based on QTL studies is steadily increasing (Korstanje and Paigen, 2002), with focus on single trait analyses like disease predisposition, plant yield and yield related traits. With the advent of multi parallel techniques, gene expression, protein and metabolite abundances started to become accessible for QTL analyses.

1.4 Heterosis

The injurious effects of self fertilization are well known (Darwin, 1876). The reverse phenomenon of an increased fitness of a hybrid cross compared to its homozygous parents was described by Shull in 1908 and later termed heterosis. The observed fitness often refers to increased biomass, size, yield, fertility, speed of development and stress resistance. It is described for animals and throughout the plant kingdom with maize being the most prominent example ever since East and Hayes (1912) suggested to use hybrid vigor in crop breeding. Some maize hybrids exhibit a more than 100% increase in grain yield (Becker, 1993) over the better parent (better parent heterosis, BPH), highlighting the agronomical importance of the biological phenomenon. More formally, every deviation of a hybrid trait value from the average of its parents for this trait can be regarded as heterosis (mid parent heterosis, MPH).

Two classic quantitative genetic explanations have been derived (Crow, 1948).

(i) The dominance theory (Bruce, 1910; Davenport, 1908), builds on the concept that deleterious alleles which are present in both parents are expressed in the hybrid to a lesser extend due to complementation. If true, this should in principle allow to combine the proposed superior alleles in inbred lines. Therefore, a part of the heterotic effect would be fixed and the absolute amount of heterosis should decline. However, this is not observed. While inbred lines have been improved steadily the amount of heterosis has slightly increased (Duvick, 1999). Further, for each individual gene the dominance theory can explain only hybrid

values to increase up to the better parent level. Consequently, cumulative action of many genes is thought to result in best parent heterosis.

(ii) The second classic explanation postulates the heterozygous state to be beneficial per se. Overdominance (Crow, 1948; Hull, 1945) (and dominance) with respect to the quantitative genetic meaning refer to a non-additive expression in the hybrid, where overdominance leads to values lying outside the parental range. This is thought to be caused by allelic interactions at one locus, still, it can not be distinguished from the situation if two contributing loci are linked in repulsion (e.g. as dominant and recessive alleles on opposite homologues) which was termed pseudo-overdominance.

Finally, the epistasis theory (Powers, 1944; Williams, 1959) attributes heterosis to the interactions between non-allelic genes at two or more loci in hybrids.

The methods and tools described in chapters 1.1 to 1.3 and further developments permitted to investigate heterosis on a molecular level. On the way to identify the causal genes Stuber *et al* (1992) analyzed QTL for grain yield and 5 other traits using 264 F3 lines developed from a cross between the two maize elite lines B73 and Mo17 which were backcrossed to both parents and had been genotyped with a high number of markers (76). Based on the finding that phenotypic traits in hybrids are higher if compared to the better parent for most QTL the authors suggested overdominance (or pseudo-overdominance) as the mode of action. A re-evaluation of the same experimental data using novel methods Cockerham and Zheng (Cockerham and Zeng, 1996) provided evidence for epistatic effects between linked QTL.

Similar studies investigating backcrosses of RILs or ILs to elucidate QTL, mostly for growth related traits were conducted for rice (Li *et al.*, 2001; Luo *et al.*, 2001; Mei *et al.*, 2005; Xiao *et al.*, 1995; Yu *et al.*, 1997), tomato (Monforte and Tanksley, 2000; Semel *et al.*, 2006), soybean (Lark *et al.*, 1995), maize (Frascaroli *et al.*, 2007; Lu *et al.*, 2003; Yan *et al.*, 2006) and Arabidopsis (Kusterer *et al.*, 2007; Melchinger *et al.*, 2007a; Syed and Chen, 2005). All studies differ with respect to the organism, developmental stage, evaluated traits, experimental design and the conclusions drawn regarding the predominant mechanism underlying heterosis.

Several groups investigated heterotic effects in physiological processes like water use efficiency in Ipomopsis (Campbell *et al.*, 2005), leaf morphology and CO_2 exchange rate in maize (Ahmadzadeh *et al.*, 2004; Tollenaar *et al.*, 2004) or the

contents of adenine nucleotides and nicotinamide coenzymes in fiber flax (Titok *et al.*, 2005).

Since the analyses of gene expression for single genes and whole genomes became feasible, techniques like RT-PCR and microarrays have been widely applied to study heterosis on different levels. Some studies on genome organization strengthened the view that complementation may contribute to the particularly high levels of heterosis in maize (Brunner *et al.*, 2005; Fu and Dooner, 2002; Song and Messing, 2003). Evaluation of gene expression for selected genes (Brunner *et al.*, 2005; Meyer *et al.*, 2007b) and on the full genome (Guo *et al.*, 2003; Stupar and Springer, 2006; Swanson-Wagner *et al.*, 2006) identified numerous heterotic loci in maize but neither a consensus gene set nor a consensus trend with respect to the amount of additive and non-additive behavior was revealed. Vuylsteke *et al* (2005) found approximately 9% of all analyzed genes to be expressed in a heterotic manner in a particular Arabidopsis hybrid. Huang *et al* (2006) could identify only 2.4% of 5771 expressed sequence tags in rice to show heterosis. However, there are arguments raised by Bancroft in a recently filed patent (Bancroft *et al.*, 2007) that transcript abundance changes in hybrids, two-fold or greater, may not be related to heterosis but to hybrid formation itself.

On the way to characterize the action at single loci Wittkopp *et al.* (2004) started to investigate allele specific gene expression. This approach was in the following applied to maize hybrids, where allele specific expression could be shown to exist in response to abiotic stresses and being predominantly *cis*-regulated (Guo *et al.*, 2004; Stupar and Springer, 2006).

The amount of heterosis can be predicted to a certain extent for crosses between related lines based on the genetic distance of both parents. However, this does not hold true for inter-group crosses between parents from genetically diverse heterotic groups (Melchinger, 1999) which are of highest interest to breeders due to the high level of heterosis often found.

1.5 Aim of the Thesis

To broaden our knowledge of plant primary metabolism and its relation to biomass using quantitative genetics a robust method for metabolic profiling of large sample populations was established. RIL and IL populations developed based on the cross of the two *Arabidopsis thaliana* accessions C24 and Col-0 and test crosses thereof

were analyzed for the abundance of 181 metabolic traces. This data allowed to explore the primary metabolism of Arabidopsis by means of a comprehensive quantitative trait locus analysis, mapping metabolic and heterotic metabolic loci. These loci were further characterized according to their distribution and co-location with each other, with biomass QTL and with a set of possible candidate genes retrieved from the AraCyc database (http://www.arabidopsis.org).

The parallel examination of the corresponding biomass for all samples was related to the metabolic profiles by multivariate statistics. Methods seeking to predict biomass or biomass heterosis based on metabolite measurements were applied.

2 Gas chromatography mass spectrometry-based metabolite profiling in plants

Jan Lisec[1,*,§], Nicolas Schauer[1,*], Joachim Kopka[1], Lothar Willmitzer[1] and Alisdair R Fernie[1]

[1] Max-Planck-Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476 Golm, Germany

[*] These authors contributed equally to this work.

[§] JL contributed to method development and preparation of the manuscript

2.1 Abstract

The concept of metabolite profiling has been around for decades, but technical innovations are now enabling it to be carried out on a large scale with respect to the number of both metabolites measured and experiments carried out. Here we provide a detailed protocol for gas chromatography mass spectrometry (GC-MS)-based metabolite profiling that offers a good balance of sensitivity and reliability, being considerably more sensitive than NMR and more robust than liquid chromatography–linked mass spectrometry. We summarize all steps from collecting plant material and sample handling to derivatization procedures, instrumentation settings and evaluating the resultant chromatograms. We also define the contribution of GC-MS–based metabolite profiling to the fields of diagnostics, gene annotation and systems biology. Using the protocol described here facilitates routine determination of the relative levels of 300–500 analytes of polar and nonpolar extracts in ~400 experimental samples per week per machine.

2.2 Introduction

Although metabolite measurements have been carried out for decades owing to the fundamental regulatory importance of these molecules as components of metabolic pathways, the importance of some metabolites in the human diet and their use as diagnostic markers for a wide range of biological conditions, including disease and response to chemical treatment, is only now being recognized (Fernie *et al.*, 2004). Historically, the measurement of metabolites was achieved either by spectrophotometric assays capable of detecting single metabolites or by simple chromatographic separation of mixtures of low complexity. Over the past decade, however, several methods offering both high accuracy and sensitivity for the analysis of highly complex mixtures of compounds have been established (Fiehn *et al.*, 2000a; Harrigan and Goodacre, 2003; Hirai and Saito, 2004; Kopka *et al.*, 2004; Roessner *et al.*, 2001a; Soga *et al.*, 2003; Sumner *et al.*, 2003). These methods

include GC-MS, liquid chromatography mass spectrometry (LC-MS), capillary electrophoresis mass spectrometry (CE-MS) and Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS). In addition, chromatographically coupled NMR technologies have found great utility in addressing specific issues, particularly in the medical field (Lindon, 2003; Wasim *et al.*, 2003) and perhaps most importantly with respect to the unequivocal determination of metabolite structures (Meiler and Will, 2002). Nevertheless, NMR shows relatively low sensitivity and thus can be used for highly abundant metabolites when profiling complex mixtures.

GC-MS facilitates the identification and robust quantification of a few hundred metabolites in a single plant extract (Fiehn *et al.*, 2000a; Halket and Zaikin, 2003; Roessner *et al.*, 2001a), resulting in fairly comprehensive coverage of the central pathways of primary metabolism. The main advantages of this technology are that it has long been used for metabolite profiling and thus there are therefore stable protocols for machine setup and maintenance, and chromatogram evaluation and interpretation. Although no single analytical system can cover the whole metabolome, GC-MS has a relatively broad coverage of compound classes (Sumner *et al.*, 2003), including organic and amino acids, sugars, sugar alcohols, phosphorylated intermediates and lipophilic compounds. Recovery experiments of all measurable classes of compounds have been done during method validation. For unknown compounds, recovery rates can be determined by recombination experiments in which extracts of two plant species are evaluated both independently and after mixing (Roessner-Tunali *et al.*, 2003a; Roessner *et al.*, 2000).

Although liquid chromatography–based methods offer distinct advantages, such as the broader range of metabolites detectable (Aharoni *et al.*, 2002; Kopka *et al.*, 2004; Plumb *et al.*, 2003; Swart *et al.*, 1993), they suffer from the lower reproducibility of retention times in liquid chromatography; in addition, owing to the predominant use of electron spray ionization, they are more susceptible to ion suppression effects, which render accurate quantification more difficult. Two alternative mass spectrometry technologies, FT-ICR-MS and CE-MS, are worth mentioning. FT-ICR-MS has unrivalled mass accuracy, thereby enabling the researcher to obtain directly a good idea about the chemical composition of the respective compound; however, a robust documentation of the validity of this technology, specifically with respect to quantification, is lacking for broad metabolite profiling. More data are available for CE-MS, a technology that detects low-abundance metabolites and affords good

chromatographic separation. Despite robust validation of this procedure (6), however, only a few reports document its use (Sato *et al.*, 2004; Unger *et al.*, 2004).

Gas chromatography time-of-flight mass spectrometry (GC-TOF-MS) technology has been developed (Saito *et al.*, 2006; Taylor *et al.*, 2002; Wagner *et al.*, 2003) and offers several advantages over the previously used quadrupole technology (GC-quad-MS)—notably fast scan times, which give rise to either improved deconvolution or reduced run times for complex mixtures and higher mass accuracy. For these reasons, the protocol described here is based on GC-TOF-MS technology; however, GC-quad-MS could be alternatively used in combination with published mass spectral alignment tools such as XCMS (Smith *et al.*, 2006), MSFACTs (Duran *et al.*, 2003), MetAlign (http://www.metalign.nl), AnalyzerPro(http://www.spectralworks.com) and BinBase (http://fiehnlab.ucdavis.edu). A detailed protocol for GC-quad-MS can be obtained by contacting the authors.

This protocol uses a MDN-35 or equivalent column with fatty acid methylesters (FAMEs) as retention time standards and was chosen because of the relative ease of application and fast chromatographic times (Kaplan *et al.*, 2004; Weckwerth *et al.*, 2004). *N*-Methyl-*N*-(trimethylsilyl)trifluoroacetamide (MSTFA) and methoxyamine hydrochloride are used as derivatization reagents because initial studies indicated that these compounds were the most appropriate for profiling of plant metabolites (Fiehn *et al.*, 2000b; Roessner *et al.*, 2000). Despite this, it has been shown that derivatization time and temperature influence the outcome of the results (Gullberg *et al.*, 2004). Derivatization of compounds often results in more than one peak for a metabolite of interest, owing to either partial silylation or isomerization in the case of methoxyaminated compounds such as sugars. In this protocol we identify all peaks of one compound, calculate their response independently, and pick the more reliable one using the statistical methods described here, but other methods such as summation of the peaks of one compound could be used as an alternative strategy. Further developments are ongoing and deal with this issue, in addition to degradation and partial silylation effects (J.K. and N.S., unpublished data).

2.2.1 Application of metabolite profiling

Improvements in metabolite profiling have rendered it an important tool for addressing biological problems. Previously, its main applications have been in the areas of diagnostics and descriptive analysis of metabolic response to various

experimental perturbations, but increasingly examples of its use in gene function annotation and systems biology are being reported. Metabolite profiles have been widely used in conjunction with statistical tools for diagnosis: they have been used to infer the mode of action of various herbicides on barley seedlings (Sauter *et al.*, 1991), and to discriminate *Arabidopsis*, potato and tomato genotypes (Fiehn *et al.*, 2000a; Roessner *et al.*, 2001a), various tissues of *Lotus japonicus* (Desbrosses *et al.*, 2005) and different stages of tomato fruit ripening (Roessner-Tunali *et al.*, 2003a).

In combination with a second round of experimental perturbation, diagnostic tools have been used to identify the principle metabolic change leading to metabolic shifts apparent after genetic perturbation (Junker *et al.*, 2004; Roessner *et al.*, 2001b). Metabolite profiling has also been used in the process of testing whether genetically modified plants are substantially equivalent to conventional crops (Catchpole *et al.*, 2005; Defernez *et al.*, 2004) and in understanding the complex shifts in metabolism that occur under nutrient limitation (Hirai *et al.*, 2004a; Nikiforova *et al.*, 2005; Urbanczyk-Wochniak and Fernie, 2005) and biotic stress (Broeckling *et al.*, 2005; Schnee *et al.*, 2006). Taken together, these examples show that metabolite profiling has important applications in the diagnostic characterization of different genetic and environmental conditions and can also aid in understanding the complex changes apparent under such circumstances.

In addition to its above-mentioned utility in diagnostics, metabolic profiling provides direct functional information on metabolic phenotypes and indirect information on a range of phenotypes that are determined by small molecules, such as stress tolerance or disease manifestations (Fernie *et al.*, 2004). Given this, there is great potential for metabolite profiling as a tool for functional genomics. Indeed, gain-of-function analysis by the transgenomic expression of every gene of the *Escherichia coli* and yeast genomes independently in *Arabidopsis thaliana* both confirmed expected functions and facilitated the assignment of gene function to unannotated open reading frames (Fernie *et al.*, 2004). This experiment was reliant on the fact that metabolite profiling can be used in a high-throughput format. Indeed, of all of the genomics technologies, it offers one of the best combinations of practical performance and cost per sample.

Metabolite profiling has also been used to demonstrate gene function by comparison of profiles derived form knockout mutants of Arabidopsis to their respective wild-type

ecotypes, facilitating the annotation of genes associated with isoflavonoid, triterpenoid, pyridine alkaloid glucosinolate, flavonoid and sterol metabolism (Goossens *et al.*, 2003; Hirai *et al.*, 2005; Morikawa *et al.*, 2006; Suzuki *et al.*, 2005; Tohge *et al.*, 2005). The approach of focusing on individual genes can be extended to exploring the phenotypic relevance of genomic regions (Schauer *et al.*, 2006; Tagashira *et al.*, 2005). A GC-MS profiling study of breeding populations of tomato, wherein genomic sequences from the wild tomato species *Solanum pennellii* were introgressed into the elite cultivated species *Solanum lycopersicum*, identified nearly 900 quantitative trait loci for fruit metabolite accumulation and ultimately, through the study of progressively smaller recombinant introgressions, should facilitate the identification of genes that regulate metabolite content in a species of nutritional significance (Schauer *et al.*, 2006). Similarly, the integration of metabolite and transcript profiling data has proved effective for identifying candidate genes for biotechnology (Askenazi *et al.*, 2003; Urbanczyk-Wochniak *et al.*, 2003).

In all technologies for metabolite profiling, the main limitation is the number of metabolites that can be detected and quantified. As ~200,000 metabolites are estimated to exist in the plant kingdom, it is clear that we are a long way from detecting the complement of plant small molecules. The availability of a full complement of isotopically labeled standards could greatly aid metabolite quantification, and further progress is undoubtedly required in determining the chemical identity of the peaks that can be resolved by current metabolite profiling methods. The use of metabolic profiling as a diagnostic tool is largely independent of the abovementioned limitations, but its application to gene function analysis and systems biology depends largely on technological improvements. The fact that the phenotype of any biological system is largely dependent on its metabolite composition (Fernie *et al.*, 2004), however, gives ample reason to invest resources in attaining this goal. Although the protocol described here was developed for the analysis of *Arabidopsis* leaf samples, its use has been validated for plant heterotrophic tissues, highlighting its broad utility.

2.2.2 Considerations for the procedure

Metabolite profiling, like any technique concerned with measuring metabolites, requires the immediate inactivation of metabolism because the turnover of metabolites, as compared with proteins and DNA or RNA, is extremely rapid. Quenching of metabolism is generally achieved by rapidly freezing samples (at a

constant temperature of –60 °C or less). In addition, the whole procedure critically requires materials of the highest purity to prevent contamination, which can easily influence the outcome of the experiment. Given this hazard, it is necessary to run quality control samples frequently alongside each experiment. Basic requirements for an experiment should be considered a priori for a generalized standard design (Jenkins *et al.*, 2004). The following details are critically important. Given that experiments generally comprise sets of samples of interest and their respective controls, it is a prerequisite that these samples are comparable to one another. For large sample sets it is imperative that there are a sufficient number of control samples, particularly because it is sometimes not possible to measure all samples in a single GC-MS run. Alongside each experiment, blank samples should be run for to identify contaminants. Blank samples should be derivatized alongside the other samples—the only difference is that this sample vial contains no metabolite extract. Another important detail is the reproduction of biological data—a minimum number of six biological replicates is sufficient (Roessner *et al.*, 2001a), but where possible the number of replicates should be even higher, especially if in-depth statistical analysis of the data sets emerging from the analysis is intended (Scholz *et al.*, 2004; Scholz *et al.*, 2005; Steuer *et al.*, 2003). Indeed, consideration of elementary statistic suggests that the sample number requirements can be determined by power analysis determined from the degree of variance within populations.

2.2.3 Sampling and extraction

Before beginning the sampling process, the time point for sampling must be carefully considered. As a general rule, we harvest photosynthetic leaf tissue in the middle of the light period because experiments in our own laboratory have indicated that almost all metabolites that we can detect and quantify are subject to strong diurnal rhythms (Urbanczyk-Wochniak *et al.*, 2005). We also tend to take samples from plants before emergence of the first inflorescence, always harvesting from the same internode and using fully developed, nonsenescent leaves. Experience dictates that these factors are crucial (Desbrosses *et al.*, 2005; Ishizaki *et al.*, 2005; Ishizaki *et al.*, 2006), as is the rapidity of the process because many metabolites show turnover times of a fraction of a second (Stitt and Fernie, 2003). Although those metabolites that can be readily detected by GCMS methods generally turnover less quickly, rapid quenching is still critical (Kopka *et al.*, 2004).

Cut samples should be rapidly weighed and then immediately frozen to quench metabolism. Before homogenization of the sample, all laboratory material to be used should be cooled down to prevent thawing of the biological sample. The amount of tissue (100 mg) used in this protocol differs from those used by other groups, but the solvent-to-tissue ratio is conserved (Gullberg *et al.*, 2004). The main reason for taking samples of relatively high mass is that lesser amounts are more difficult to handle and small errors in weighing can propagate to produce large changes in the final evaluation. If the user chooses to use less tissue, however, the extraction volume can be readily adapted. The first three steps of the extraction procedure (Fig. 1) are particularly crucial in terms of avoiding thawing and its associated problems.

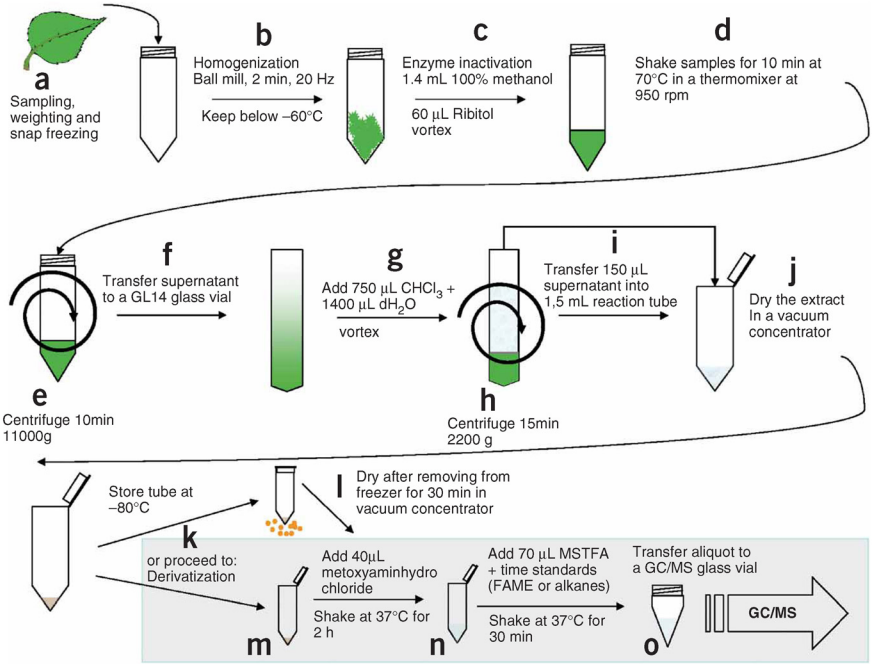


Figure 1 (a) Sampling, weighting and snap-freezing. (b) Homogenization in a ball mill for 2 min at 20 Hz. Keep the temperature below -60 °C. (c) Enzyme inactivation. Add 1.4 ml of 100 % methanol, vortex sample, add 60 µl of Ribitol, and vortex sample. (d) Shake samples for 10 min at 70 °C in a thermomixer at 950 r.p.m. (e) Centrifuge sample for 10 min at 11,000 *g*. (f) Transfer supernatant to a glass vial. (g) Add 750 µl of chloroform and 1,400 µl of dH_2O to the sample, and vortex. (h) Centrifuge sample for 15 min at 2,200 *g*. (i) Transfer 150 µl of supernatant into a 1.5-ml reaction tube. (j) Dry the extract in a vacuum container. (k) Store the tube at -80 °C or proceed to derivatization. (l) If storing the tube, dry the sample in a vacuum concentrator for 30 min after removing from the freezer. (m) Add 40 µl of methoxyamination reagent (see REAGENT SETUP) to sample and shake at 37 °C for 2 h. (n) Add 70 µl MSTFA reagent and time standards (such as FAMEs or alkanes) to sample, and shake at 37 °C for 30 min. (o) Transfer aliquot to a GC-MS glass vial and analyze by GC-MS.

This protocol is suitable for both polar and apolar extraction of metabolites. Although there is considerable experience in the analysis of polar metabolites, far less is known about apolar compounds, owing, at least in part, to carryover and contamination effects, which require more sophisticated knowledge, equipment and

methodology. For this reason, here we concentrate on only the polar phase. Although we supply precise information pertaining to the instrumentation used, it should be noted that this protocol is broadly applicable to all machines of this type.

2.3 Materials

2.3.1 Reagents

- Argon 5.0 (Messer-Griesheim)
- Chloroform for liquid chromatography (Merck, cat. no. 67-66-3) ***Caution*** **Chloroform is toxic and should be handled under a fume hood**
- Helium 5.0 carrier gas (Air Liquide)
- Methanol, gradient grade for liquid chromatography (Merck, cat. no. 67-56-1)
- *N*-Methyl-*N*-(trimethylsilyl)trifluoroacetamide (MSTFA reagent; Macherey-Nagel, cat. no. 24589-78-4); prepared in 1-ml vials and stored at 4 °C. ***Caution*** **Reagent is extremely toxic and should be handled under the fume hood**
- Orange silica gel, no. 77.1 (Carl Roth)
- Ribitol (Sigma, cat. no. 488-81-3)

2.3.2 Equipment

- Autosampler and software (CTC Combi PAL and PAL cycle composer software version 1.5.0; CTC Analytics); the configuration comprises an agitator-incubator oven, a 98-sample tray for 2.0-ml vials, a 32-sample tray for 10–20-ml vials, three 100-ml solvent reservoirs (i.e. a syringe wash station and a liquid version 25-µl syringe kit mounted on the robotic autosampler arm)
- Conical single taper split/splitless liner (Agilent)
- Gas chromatograph, 6890N, split/splitless injector with electronic pressure control up to 150 psi (Agilent)
- GL14 glass vials (Schott)
- MDN-35 capillary column, 30 m length, 0.32 mm inner diameter, 0.25 µm film thickness (e.g. Macherey-Nagel or equivalent (Schad *et al.*, 2005))
- Micro-vials: 1.5-ml, safe-lock, tapered bottom, and 2.0-ml, screw-cap, round bottom (Eppendorf)
- Oscillating ball mill, MM200 (Retsch)
- Pegasus III time-of-flight mass spectrometer (Leco Instruments)
- Steel balls, VA5mm (Th. Geyer Berlin)

- Screw caps for GL14 glass vials (Schott, cat. no. 29 990 12 04)
- Teflon adaptor for 1.5–2.0-ml micro-vials (Retsch)
- Automated mass spectral deconvolution and identification system (AMDIS; National Institute of Standards and Technology)
- ChromaTOF chromatography processing and mass spectral deconvolution software, version 1.00 or higher, driver 1.61 or higher (LECO Instrumente), running on a state-of-the-art computer with a minimum of 512-MB RAM and an 1.0G-Hz Pentium IV processor or equivalent
- R: a Language and Environment for Statistical Computing (R Foundation for Statistical Computing)

2.3.3 Reagent Setup

Methoxyamination reagent Dissolve methoxyamine hydrochloride (Sigma, cat. no. 593-56-6) at 20 mg ml^{-1} in pure pyridine (Merck, cat. no. 110-86-1) at 20–25 °C. This reagent needs to be prepared freshly before the experiment. ***Caution*** **Reagents are extremely toxic and should be handled under the fume hood.**

Retention time index standard mixture Dissolve FAMEs in chloroform at a final concentration of 0.4 ml ml^{-1} or 0.8 mg ml^{-1} for liquid or solid standards. Reagent can be stored at -4 °C. Esters included are methylcaprylate (Sigma, cat. no. 111-11-5), methyl pelargonate (Sigma, cat. no. 1731-84-6), methylcaprate (Sigma, cat. no. 110-42-9), methyllaurate (Sigma, cat. no. 111-82-0), methylmyristate (Sigma, cat. no. 124-10-7), methylpalmitate (Sigma, cat. no. 112-39-0), methylstearate (Sigma, cat. no. 112-61-8), methyleicosanoate (Sigma, cat. no. 1120-28-1), methyldocosanoate (Sigma, cat. no. 929-77-1), lignoceric acid methylester (Sigma, cat. no. 2442-49-1), methylhexacosanoate (Sigma, cat. no. 5802-82-4), methyloctacosanoate (Sigma, cat. no. 55682-92-3), and triacontanoic acid methylester (Weckwerth *et al.*, 2004) (Sigma, cat. no. 629-83-4). Alternatively, alkanes (Roessner-Tunali *et al.*, 2003a) or fatty acids (Roessner *et al.*, 2001a) have been, and can be, used.

2.3.4 Equipment Setup

Standardization As a rule, metabolite profiling studies compare two or more states of a given biological system; thus, absolute quantification is not necessary and relative quantifications of the level of metabolites of interest per tissue mass (i.e. per gram of fresh weight) is sufficient. In such instances, the challenge of quantification is

reduced to comparison between one or many samples, which essentially transforms the problem of quantification into a problem of standardization. Any standardization has to correct for the following.

(i) Experimental errors during sample preparation (determination of the sample amount and subsequent liquid handling): this is corrected for by the critical inclusion of a compound not present in biological samples (i.e. Ribitol or ^{13}C Sorbitol), directly after homogenization of the sample. For normalization with Ribitol, the unique mass *m*/*z* 319 is used, whereas for ^{13}C Sorbitol *m*/*z* 323 is used.

(ii) Overall machine sensitivity.

(iii) Changes in sensitivity towards specific compounds owing to differences in the matrix. Although this aspect in the form of ion suppression is a major problem in any mass spectrometry technique relying on electrospray ionization, as a rule it is a lesser problem in GC-coupled mass spectrometers.

Machine sensitivity Given the variability in the overall machine, sensitivity is most probably the crucial factor to correct for during quantification procedures. In principle, the following standardization protocols can be applied to correct for machine sensitivity.

(i) The expression of every identifier mass trace used for quantification as a proportion of the total ion intensity of all identified compounds of that sample. In our experience this is only a very gross correction for machine performance and not overly reliable.

(ii) The evaluation of a large number of controls (~20 % of the total samples) in a random order between experimental samples is highly preferable. This control should be as similar as possible with respect to chemical complexity to the experimental samples. In practice, this means that one need only to prepare a large batch of pooled extracts from, for example, leaves of a given species and aliquot these for subsequent use as machine sensitivity controls. This quantification is the most reliable; however, it requires a large number of control samples to be run and thus leads to increased costs and reduced throughput. An alternative that is used in experiments where absolute quantification is a prerequisite, such as analyses of metabolic fluxes (Roessner-Tunali *et al.*, 2004; Tieman *et al.*, 2006), is the evaluation of calibration curves of authentic standards. Such standards are analyzed in replicate over a dilution series after

derivatization and handling in the exact same way as described above. This approach is also cost- and labor-intensive but offers the advantage of facilitating comparison with published metabolite data obtained with other analytic techniques (Roessner-Tunali *et al.*, 2003b).

(iii) Given that machine sensitivity is generally sufficiently stable over a day, the median of distribution of each metabolite across the samples measured in a day can be calculated and the content of each metabolite can be subsequently expressed in comparison to its daily median. This is a very cost-effective approach because it does not require as many controls to be run as in the previous approach. In addition, it allows a metabolite-by-metabolite correction for machine sensitivity. An example of this approach is given in Figure 2: measurement of glycine over five independent days shows some variance; when samples are related to the daily median, however, the variance between biological replicates is comparable between measurement days. It is important to note that this approach is valid only when the chemical composition is very similar and the distributions of concentration in the different samples are similar in the samples measured on different days. If these prerequisites are fulfilled, this approach is both a robust and reliable one.

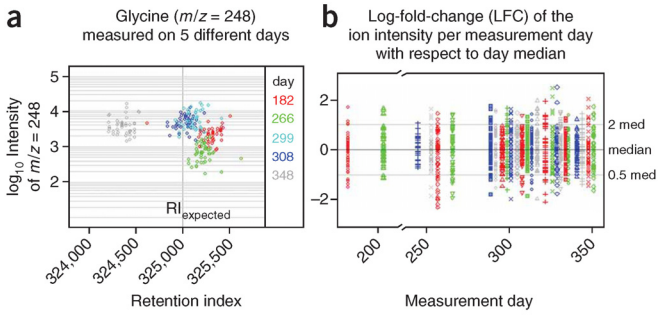


Figure 2 Although we convert the data from each sample individually from retention time to RI, there is still daily variation over long measurement periods. Detector sensitivity is another factor that we have to take into account to enable us to perform large-scale experiments. (a) Glycine (m/z = 248) measured on five different days: intensity is clearly dependent on the day of measurement (samples came from the same experiment and were completely randomized). (b) Log fold-change of the ion intensity per measurement day with respect to day median: if we normalize our results on the median value for a specific metabolite per measurement day, the distribution of the samples is comparable.

2.4 Procedure

2.4.1 Sampling and Extraction

1. Sample leaf material in 2-ml, screw cap, round bottom tubes. Define the exact mass of plant sample (~100 mg of fresh weight) and rapidly freeze the

sample-containing vial using liquid nitrogen or an equivalent low-temperature liquid.

2. To homogenize the tissue, place steels balls into the sample tubes and insert samples into precooled Teflon adaptors. Homogenize in ball mill for 2 min at 20 Hz.

Pause Point **Frozen homogenate can be stored at -80 °C for up to 3 months.**

3. Add 1,400 µl of 100 % methanol (pre-cooled at -20 °C) and vortex for 10 s.
4. Add 60 µl of Ribitol (0.2 mg ml^{-1} stock in dH_2O) as an internal quantitative standard and vortex for 10 s.
5. Shake for 10 min at 70 °C in a thermomixer at 950 r.p.m.
6. Centrifuge for 10 min at 11,000 *g*.
7. Transfer supernatant to a Schott GL14 glass vial.
8. Add 750 µl of chloroform (-20 °C).
9. Add 1,500 µl dH_2O (4 °C) and vortex for 10 s.
10. Centrifuge 15 min at 2,200 *g*.
11. Transfer 150 µl from the upper phase (polar phase) into a fresh 1.5-ml tube.
12. As a backup (in case you lose a sample), take a second aliquot into a new 1.5-ml tube.
13. Dry in a vacuum concentrator without heating.
14. Before freezing the aliquots at -80 °C, fill the tubes with argon gas and place them inside a plastic bag containing silica bead desiccant. Argon-filled sample vials prevent the extract from oxidization and degradation by reactions through components of atmospheric air. **CAUTION** Halogenic reagents and solutions should be disposed with halogenic waste.

Pause Point **Samples can be stored at -80 °C for up to 3 months.**

2.4.2 Derivatisation

Critical step **Steps 15–22 have been shown to be very critical. In this protocol we use derivatization reagent in supersaturated concentrations to ensure completion of derivatization.**

15. Place samples stored at -80 °C in a vacuum concentrator for 30 min before derivatization.

16. Add 40 µl of methoxyamination reagent (see REAGENT SETUP) to the aliquots. **CAUTION** Derivatization reagents are extremely toxic. Handle with absolute care. Work with gloves and under the fume hood.

Critical step **In the process of derivatization, condensation of reagents appears on the wall and lid of the reaction tubes; therefore, centrifugation of the reaction mixture is essential after every incubation step.**

17. Also prepare one derivatization reaction using an empty reaction tube as a control.
18. Shake for 2 h at 37 °C.
19. Prepare MSTFA reagent with 20 µl ml^{-1} of retention time index standard mixture (see REAGENT SETUP).
20. Add 70 µl of the solution prepared in Step 19 to the sample aliquots.
21. Shake for 30 min at 37 °C.
22. Transfer into glass vials suitable for GC-MS analysis.

2.4.3 GC-TOF/MS metabolite profiling

23. *Injection parameters* Inject 1 µl of sample at 230 °C in splitless mode with helium carrier gas flow set to 2 ml min^{-1} by using the autosampler setup (see EQUIPMENT) The flow rate is kept constant with electronic pressure control enabled. Optionally, but especially recommended in cases of high metabolite concentrations, injection can be done in split mode with the split ratio adjusted to 1:25.
24. *Chromatography parameters* Perform chromatography with a 30 m MDN-35 capillary column. The temperature program should be isothermal for 2 min at 80 °C, followed by a 15 °C per min ramp to 330 °C, and holding at this temperature for 6 min. Cooling should be as rapid as instrument specifications allow. Set the transfer line temperature to 250 °C and match ion source conditions (Schad *et al.*, 2005).
25. *Mass Spectrometer parameters* Set the ion source to maximum instrument specifications, 250 °C. The recorded mass range should be *m*/*z* 70 to *m*/*z* 600 at 20 scans per s. Proceed the remaining monitored chromatography time with a 170 s solvent delay with filaments turned off. Manual mass defect should be set to 0, filament bias current should be -70 V, and detector voltage should be

~1700–1850 V. Automatically tune the instrument according to the manufacturer's instructions.

2.5 Anticipated Results

Figure 3 visualizes the estimated amount of time from collecting 50 biological samples to final results. Sampling, extraction and derivatization takes less than 50% of the time, but more effort and time, approximately between 2 and 5 days, is needed for the comprehensive data analysis.

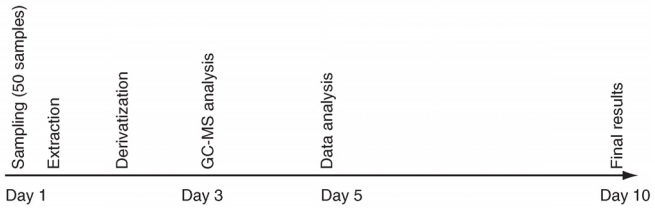


Figure 3 Timeline of standard operating procedure. Note that five days are assigned to final results despite the use of an instantaneous algorithm, because manual inspection of chromatograms is a highly advisable quality control.

2.5.1 Deconvolution

Following the general philosophy of metabolic profiling, the extraction procedure used should introduce as little bias as possible with respect to the complexity of the compounds extracted from the biological sample. Thus, metabolic profiling leads to complex chromatograms characterized by coeluting compounds and vast differences in the relative abundance of the different compounds. Although problematic, these issues can be partially resolved by deconvolution of the chromatograms. The machine supplier’s software, e.g. ChromaTOF, offers a build-in deconvolution algorithm. Deconvoluted spectra can be exported as plain text files for further processing. We suggest the following parameters for the deconvolution process (with acceptable range in parentheses). In all instances we used the machine manufacturer’s recommended approaches, which we have found to be highly appropriate.

- Baseline offset = 1 (0.5–1)
- Smoothing = 5 data points (3–7)
- Peak width = 3 s (3–4 s)
- S/N (signal-to-noise ratio) = 10 (2–15)

2.5.2 Retention time index

Retention time index (RI) is probably the most important parameter for peak assignment. In our experience it is absolutely crucial that each chromatogram is

corrected for retention times separately, as even within a day absolute retention times show variance that, combined with the fact that the complex mixtures apparent in plant extracts result in highly complex chromatograms, can lead to false peak annotations (see Figure 4 for an example of the variation of retention times of the FAME retention time standards). To minimize this problem, we apply an algorithm (R-Script 1; available from J.L.) comprising the following steps:

(i) Identification of the retention time for each of the internal markers (see 'REAGENT SETUP for internal retention time standards) and assign a 'fixed RI' to the respective peaks.

(ii) Calculation of the RI for all compounds eluting between two standards using a linear interpolation.

(iii) Extension of the linear correction for all compounds eluting prior to the first or after the last standard.

To be able to narrow considerably the time window to search for a certain compound, each file needs to be corrected independently. Although this can be theoretically achieved by the original software, it is highly impractical on a high throughput scale.

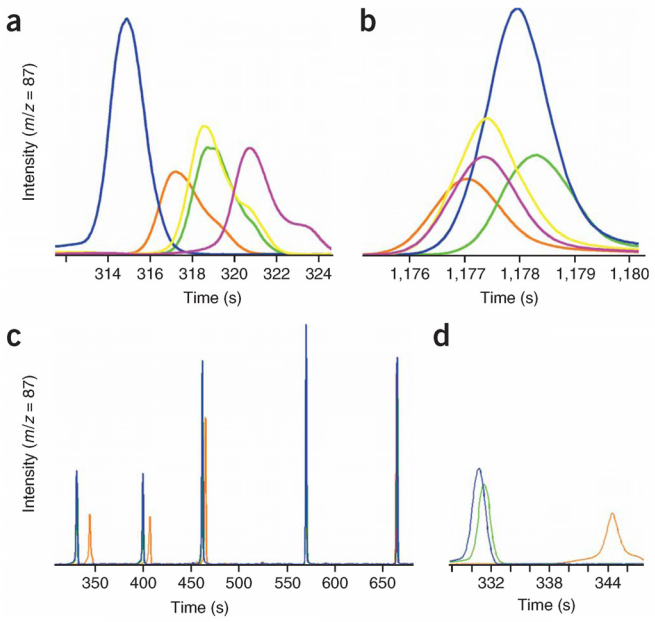


Figure 4 (a,b) General variation of retention time of RI standards within a day. Generally we observe stronger variation (up to 6 s) for early eluting standards (a), whereas later standards are robust (b). (c,d) Outlier behavior at early elution times. Outliers may occur in the early elution phase, showing differences of up to 12 s if compared with other files from the same day, but lining up with those during the chromatographic run. Intensities of specific FAMEs of the retention index standards within a batch of chromatograms are shown. Selected files of an authentic data set are used for illustrative purposes. Different colors represent different exemplary data sets.

2.5.3 Peak annotation

The R-script 1 algorithm developed in our laboratories facilitates annotation of a given peak to a compound (with known or unknown chemical structure), a process that is reliant on two known factors, namely the RI and the mass spectrum.

2.5.4 Unique masses as identifiers

In principle, it should be possible to annotate each compound based on its unique mass spectrum and RI (Kopka *et al.*, 2005; Schauer *et al.*, 2005; Wagner *et al.*, 2003).

In metabolite profiling, however, the presence of coeluting compounds present in high dynamic range (Sumner *et al.*, 2003) can mean that reliance on these parameters proves to be difficult. This is even more pronounced when the coeluting compounds have one or more masses in common. Many commercial and publicly available mass spectral evaluation tools exist. These tools are largely similar in function, if not execution, and offer distinct advantages and disadvantages. Because a broad comparison of these various algorithms is currently lacking, we do not discuss them in detail here, but rather concentrate on an algorithm developed specifically for the metabolite profiling method that we describe. For this purpose, we decided to use a combination of a very precise relative retention time as described above and one or more mass traces unique within this retention time window for the assignment of a given peak to a compound (R-Script 1). The RI-corrected spectra are processed by a second bespoke R-Script (R-Script 2, available from J.L. according to a prepared reference list using an algorithm that achieves the following:

(i) All peaks within the specified time window are evaluated.

(ii) The peak showing the maximum intensity for the predefined unique ion is chosen.

The reference list—containing name, expected RI, allowed RI variation and unique mass for a number of metabolites—can be initially prepared by evaluating the GC-MS spectra resulting from the evaluation of a mixed sample pooled from aliquots of the whole measurement set in conjunction with available and constantly expanding GC-MS library sets (Kopka *et al.*, 2005; Schauer *et al.*, 2005; Wagner *et al.*, 2003), when necessary following the troubleshooting procedure to ensure authenticity of peak identification (see TROUBLESHOOTING).

2.5.5 Identification of novel compounds and contaminants

Compound identification is essentially performed by running authentic standards and determining RI and specific masses. In many cases, however, specifically in the case of unknown metabolites for which some mass spectral properties are clear, it is highly desirable to obtain a mass spectrum as an aid for their further identification.

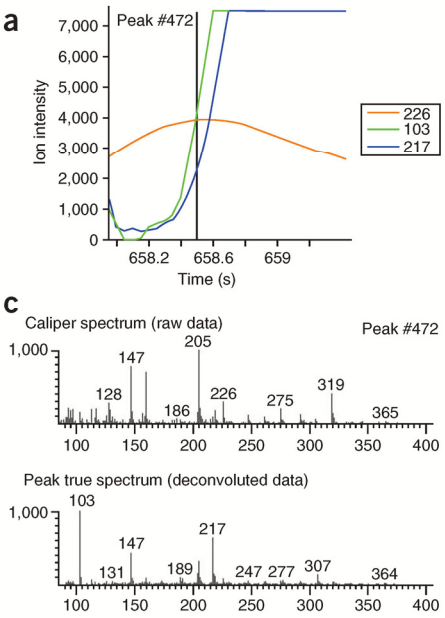
Median spectra (as used for error correction) may be computed from a number of samples and can be exported in NIST format for comparison to external databases. Given that many such databases report data for nonderivatized metabolites, the derivatization and subsequent analysis of authentic standards represents one way to identify unknown peaks via GC-MS. Other ways to tackle this difficult problem are largely reliant on analytical techniques such as LC-MS (Tolstikov and Fiehn, 2002) and NMR; the exceedingly high mass accuracy of FT-ICR-MS make it this technology seem likely that this technology, when coupled to chromatography, will have a considerable role in the future development of plant metabolomics.

The direct experimental output from this protocol is a list of metabolite contents of the experimental conditions in comparison to the control. The number of compounds detected in polar leaf extracts depends on the RI and mass spectral information annotated by the experimenter. In general, taking mass spectral tag information from publicly available libraries (Kopka *et al.*, 2005; Schauer *et al.*, 2005; Wagner *et al.*, 2003) into account, this should lead to ~150–500 compounds of known and unknown origin; however this number varies on the species and tissue type. The direct output described above can be subsequently evaluated with respect to the biological question of the research either in a metabolite-by-metabolite manner or by using one of the many available statistical packages for multivariate analysis. The former analysis would be suggested in studies of metabolic regulation, for example coordinated metabolic responses to nutrient deprivation and for gene annotation, whereas the later retains great utility in diagnostics-based approaches. As a general rule, only 40 % of the compounds are annotated to a specific metabolite, so if a particularly interesting trend in an unannotated metabolite is found in an experiment, it is recommended that a second analytical technique is used to determine the chemical structure of this unknown metabolite. A note of caution is necessary here, however, because this is generally a far from trivial task.

2.5.6 TROUBLESHOOTING

Deconvolution. Using the manufacturer's deconvolution software we have, in a few instances, encountered errors that can be defined either as errors of multiple deconvolution of a single peak, or as errors in which deconvoluted spectra contain the wrong ion intensities (Fig. 5). Although these errors occur infrequently, their frequency is high enough to preclude over-reliance on the machine manufacturer's own software. For this reason, we recommend that correction of these errors be

carried out by collection all data files of a sample set (generally 40–50 samples) and processing these files together in the framework of the open source software package R using the designed scripts (available from J.L.).



b

Spectra table
Peak #472

Mass	Caliper	Peak true
103	2028	49843
217	1229	31597
147	10906	21370
205	14034	15716
204	2079	7222
218	60	6905
307	0	6567
219	2359	5378
189	659	4671
206	2825	4578
104	410	4446
148	2214	4356
160	9740	4351
226	4040	3889
191	182	3808
84	147	3107
113	2637	2972
277	459	2633
275	2761	2394
207	1268	2304
93	2870	2083
173	38	2063
149	666	2030
247	1566	2001
308	0	1896
131	1207	1870

Figure 5 (a) Unique mass and two high-abundance ion intensities from a coeluting peak. (b) Raw data (Caliper, the triangular visible at the base of a, indicates a single data point on the timescale, in this instance, identical to the time point of the deconvoluted peak) and deconvoluted data (Peak true) at peak position; ions from a are boxed. (c) Spectrum representation of raw and deconvoluted data at peak position. Note that deconvoluted data obviously do not always represent actual values. Spectral comparison would fail to identify this peak because it will be highly variable owing to deconvolution errors. Nevertheless, intensity values for the unique mass are sensible.

Annotation. Within a given matrix (e.g. *Arabidopsis* leaf or root, potato tuber, tomato pericarp tissue, *Lotus japonicus* nodule), the above procedure for compound annotation is highly reliable. When a new matrix is analyzed or when a given matrix is analyzed where the organism was exposed to widely different environmental condition or a genetic variant shows a marked visual phenotype, however, we strongly suggest a manual inspection of chromatograms to avoid the erroneous use of a unique mass for compound identification owing to the appearance of a novel compound with the same mass trace in the same retention time window. Manual inspection is a highly time-consuming and laborious process. To speed up the process of error identification, however, a graphical overview of all analyzed spectra for a specified metabolite per data set can be used, as exemplified in Figure 6. Such a display can indicate dubiously annotated spectra. The box-plot spectra (Fig. 6b), represents the statistical comparison of the samples of a given set of chromatograms to the median standard spectra (Fig. 6a).

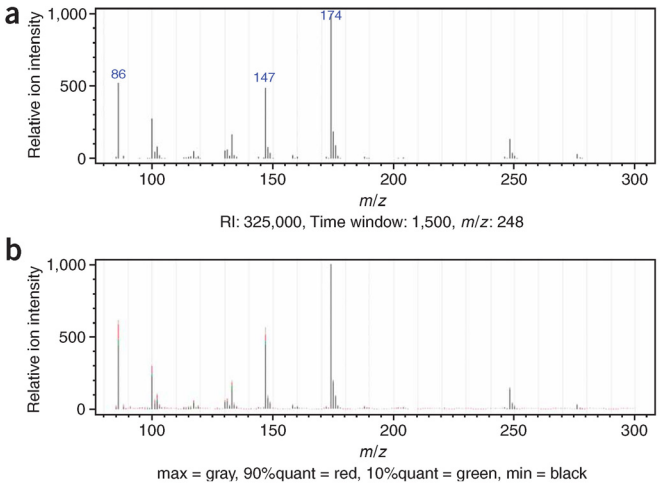


Figure 6 These plots are regularly used to check the quality of search results quickly. Here glycine was searched for with the following reference conditions and data set: RI, 325,000; time window, ± 1,500 RI units (~1.5 s); unique mass, 248; data set, measurement day 182, 43 samples, no missing values. (a) Median value for each ion calculated from all 43 relative spectra chosen. It should be close to the recorded library spectrum for this metabolite. (b) Box-plot of the candidate spectra, indicating by color minimum, 10% quantile, median, 90% quantile and maximum of all extracted values. In this example, one could observe the biggest (but negligible) variation for the ion of mass 86. Only if this box-plot spectrum reveals strong variation between the chosen spectra do we calculate the mean squared difference for all single spectra from the median spectrum, thus quickly identifying the outliers, plotting them as an overview plot (see Fig. 5) and then re-evaluating the chromatogram where necessary.

Where variance is great across the chromatograms, additional plotting possibilities enable the user to trace back to the offending samples or peaks, thus facilitating a rapid verification or falsification of peak annotation and an acceleration of the process of manual correction. Only when the data are validated either algorithmically or manually are they deemed acceptable for publication and/or storage in publicly accessible databases. An overview of the complete process is outlined in Figure 7.

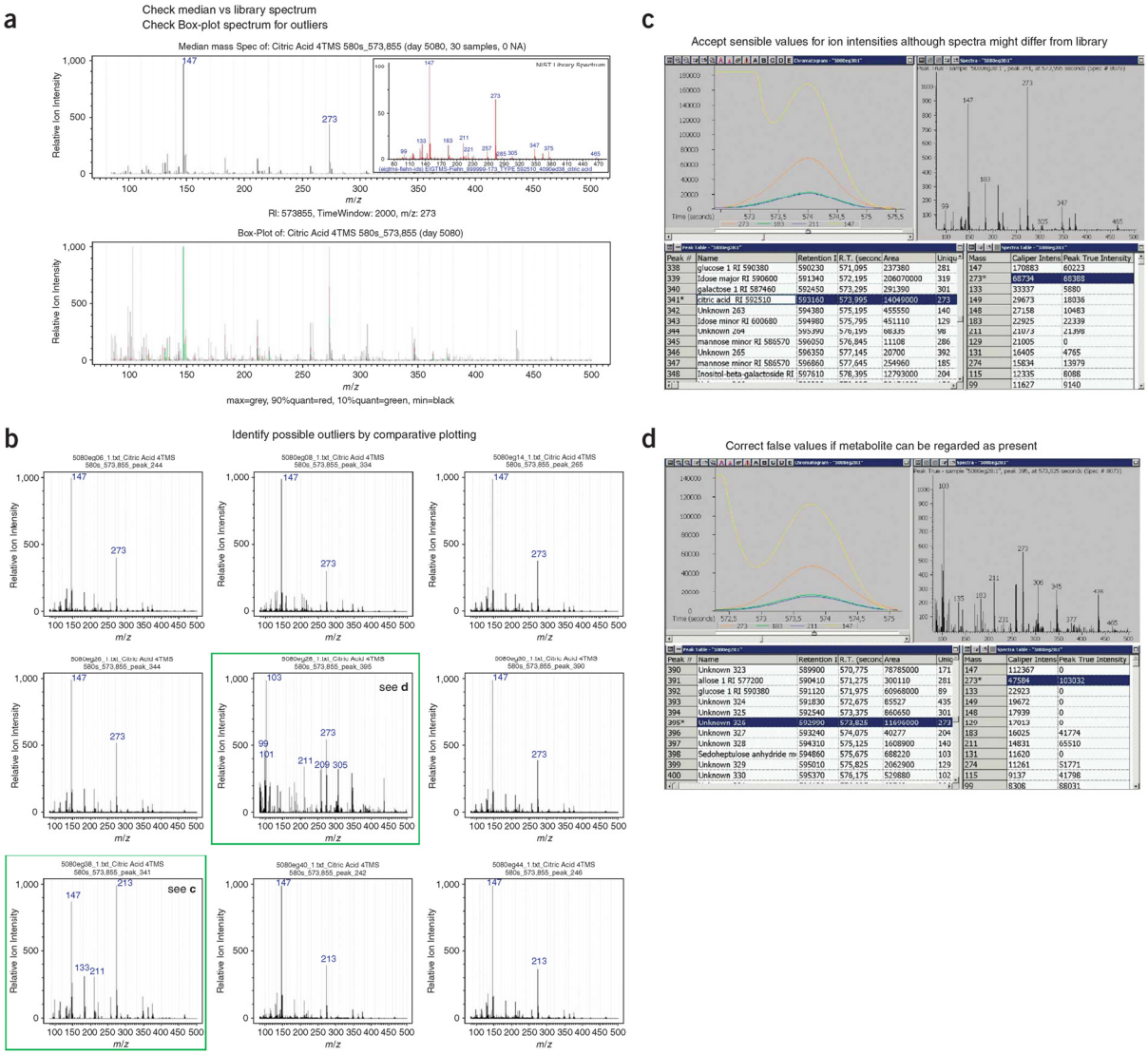


Figure 7 After comparing the median spectrum with a library, outliers (indicated by the box-plot) are visually checked by comparative plotting (a,b). Checking back in the original software (c,d) shows that although both spectra look dubious they represent the correct metabolite. Nevertheless, only in the first case the calculated value may be used (intensity of m/z_{273} = 68,388), whereas the second one seems to be too high and would need to be annotated manually.

3 The metabolic signature related to high plant growth rate in *Arabidopsis thaliana*

Rhonda C. Meyer[1,*], Matthias Steinfath[2,*], Jan Lisec[3,*,§], Martina Becher[1], Hanna Witucka-Wall[1], Ottó Törjék[1], Oliver Fiehn, Änne Eckardt[3], Lothar Willmitzer[3], Joachim Selbig[2,3] and Thomas Altmann[1,3]

[1,2] Departments of [1]Genetics and [2]Bioinformatics, Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Strasse 24–25, 14476 Potsdam, Germany

[3] Max-Planck-Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476 Golm, Germany

[*] These authors contributed equally to this work.

[§] JL contributed to method development, data analysis and preparation of the manuscript

3.1 Abstract

The decline of available fossil fuel reserves has triggered worldwide efforts to develop alternative energy sources based on plant biomass. Detailed knowledge of the relations of metabolism and biomass accumulation can be expected to yield powerful novel tools to accelerate and enhance energy plant breeding programs. We used metabolic profiling in the model *Arabidopsis* to study the relation between biomass and metabolic composition using a recombinant inbred line (RIL) population. A highly significant canonical correlation (0.73) was observed, revealing a close link between biomass and a specific combination of metabolites. Dividing the entire data set into training and test sets resulted in a median correlation between predicted and true biomass of 0.58. The demonstrated high predictive power of metabolic composition for biomass features this composite measure as an excellent biomarker and opens new opportunities to enhance plant breeding specifically in the context of renewable resources.

3.2 Introduction

Multicellular organisms have to optimize the use of available resources to fit their needs in terms of energy, biosynthetic building blocks, and reserves. Green plants unlike animals produce their own organic compounds. Their ability to grow thus solely depends on their own photosynthetic and metabolic capacity. Biomass accumulation in the vegetative growth phase of a plant can therefore be regarded as the ultimate expression of its metabolic performance. Plants function as integrated systems, in which metabolic and developmental pathways draw on common resource pools and respond to changes in environmental energy and resource supplies (Tonsor *et al.*, 2005). The distribution of metabolites between growth, production of defense compounds and storage compounds therefore has to be very tightly

regulated. Growth and the concomitant drain of metabolites into cellular components has to be adjusted to the metabolic capacity of the system, i.e., the ability to supply sufficient amounts of organic compounds. This regulation is demonstrated by several observations of growth depression upon reduction of primary metabolism such as sucrose synthesis (Chen *et al.*, 2005; Fernie *et al.*, 2002). Growth ceases upon severe starvation caused by an extended dark period and is reinitiated only after a lag period of several hours after relief from the starvation by reillumination (Gibon *et al.*, 2004). Recent observations of the roles of the DELLA proteins in plants indicate that plant growth is limited to a submaximum level to enable plants to cope with unfavorable conditions (Achard *et al.*, 2006). Thus, growth rate has to be adjusted to the metabolic status of a plant that needs to be translated into an appropriate response. This interaction between metabolism and the growth regulatory mechanisms may operate in two ways: either a high supply of metabolites triggers growth, or growth drains metabolites to a minimum tolerable level upon which growth is restricted. Metabolites may exert control on growth either by acting as substrates for the synthesis of cellular components, that become limiting under conditions of maximum tolerable growth, or by acting as signals that are sensed leading to subsequent changes in growth. Sugars such as glucose and sucrose have been shown to act as metabolic signals and to be involved in the control of plant growth and development (Gibson *et al.*, 2004). Trehalose-6-phosphate has recently been shown to be involved in signaling of the plant sugar status and in control of growth and development (Kolbe *et al.*, 2005; Schluepmann *et al.*, 2003).

Metabolic profiling is a mass-spectrometry (MS)- or NMR-based technology for an unbiased analysis of the metabolome of a given biological system with a high diagnostic power (Fernie *et al.*, 2004). Thus, in case of e.g. yeast or plants, metabolic analysis allows to distinguish between different genotypes, developmental status or environmental conditions (Allen *et al.*, 2003; Keurentjes *et al.*, 2006; Tarpley *et al.*, 2005). In the case of humans, metabolomic approaches allow us to predict the response of individuals to drugs opening aspects of personalized drug treatments (Clayton *et al.*, 2006). In addition, single or a small number of metabolites can be extracted from metabolic profiling studies that have the potential to be developed into rapidly accessible biomarkers (Lindon *et al.*, 2004).

Based on the above considerations between the metabolic status of a plant system and growth and the proven high diagnostic power of metabolic profiling approaches,

we decided to test whether biomass of a plant is correlated with and can thus be predicted by its metabolic composition. To this end we took advantage of a recombinant inbred line (RIL) population of *Arabidopsis thaliana* derived from a cross between the *Arabidopsis* lines Col-0 and C24 (Törjék *et al.*, 2006), which in previous studies showed strong transgressive segregation for biomass (Meyer *et al.*, 2004). RILs represent permanent segregating populations of homozygous lines, which allow to reduce the environmental variance in replicated experiments (Alonso-Blanco *et al.*, 1998b). The extensive biochemical variation in *Arabidopsis* is largely under genetic control (Keurentjes *et al.*, 2006). Therefore, the use of such a population for an exploratory analysis of relations between growth and metabolite levels is particularly advantageous (over e.g., using environmental perturbations to modulate growth and metabolism) as it offers the opportunity to identify the genetic determinants of all studied traits in addition to the determination of correlations.

As shown below, when applying multivariate analysis to the combined data sets of biomasses and metabolic profiles, a statistically highly significant correlation between metabolic composition and biomass was obtained. We believe this result to be of high relevance for our basic understanding of plant growth and metabolism and to have obvious implications for breeding of high plant biomass producers, an aspect which in recent years has become of increasing importance regarding renewable resources as energy supply (Schubert, 2006; Somerville, 2006). It furthermore provides precedence for the utility of molecular profiling data to extract biomarkers with high predictive power for a complex trait.

3.3 Results

3.3.1 Biomass and Metabolite Profile Determination of Col-0/C24 RILs

The combined analysis of biomass and metabolic profile was performed on a total of 1,144 genotypes. Of these lines, 429 genotypes were derived from a RIL population from the reciprocal crosses Col-0×C24 (228 lines) and C24×Col-0 (201 lines) and 715 lines were derived from crosses of the RILs to parents Col-0 and C24. All plants were grown under controlled conditions in six replicated experiments. Plants were harvested 15 days after sowing and used for shoot biomass determination or were pooled and frozen for metabolite profiling by gas-chromatography/ mass-spectrometry (GC-MS). The distribution of mean biomass within the population clearly shows transgressive segregation (Fig. 8). We detected no significant

differences in biomass (*t* test, $P = 0.238$) between the two subpopulations, and therefore treated the RILs as one population in subsequent analyses.

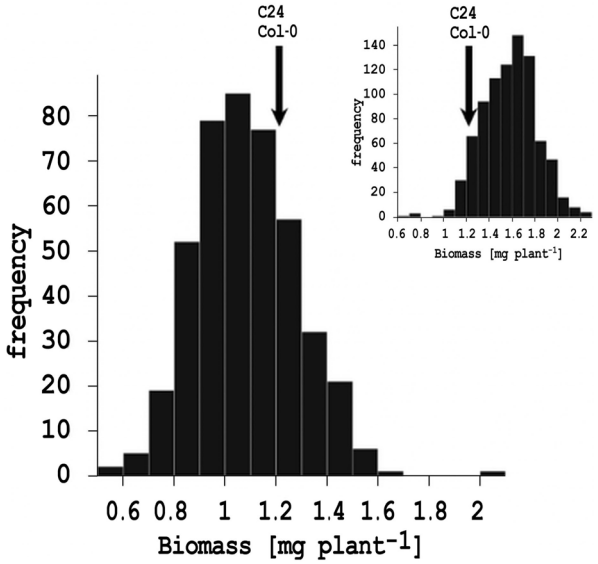


Figure 8 Distribution of shoot biomass in the recombinant inbred line (RIL) population. Shown is the mean biomass (milligrams per plant) estimated by REML. The arrow indicates the biomass determined for the parental lines C24 (1.265 mg per plant) and Col-0 (1.254 mg per plant). The histogram of the shoot biomass of the RIL crosses to the parents is shown in *Inset*.

From the metabolic profiling data we took only those metabolites into account, which were detected reproducibly in at least 85 % of the samples analyzed. Major groups among these metabolites are organic acids, sugars, sugar phosphates, polyols, amines and amino acids. Concentrations could be determined for a set of 181 compounds, 84 of which were assigned a chemical structure by comparison with a library (Kopka *et al.*, 2005; Schauer *et al.*, 2005). The remaining compounds were classified into chemical groups by using representative masses.

3.3.2 Canonical Correlation Reveals a Close Link Between Biomass and a Specific Combination of Metabolites.

In a first approach distributions of single metabolites were queried for their predictive power with respect to the biomass distribution by calculating pairwise correlations between all 181 measured metabolite levels and biomass (Supplemental Table 1). Because a normal distribution cannot be assumed for all variables rank correlation was used as a robust estimation of the correlation coefficient. The highest absolute correlation found was for a carbohydrate, which yielded a value of 0.266. Although the correlation is statistically highly significant (P value of $5.17 \cdot 10^{-20}$), it can only explain 7.07 % of the variance. Other significantly correlated compounds are ethanolamine (0.238; $P = 3.87 \cdot 10^{-16}$), fructose-6-phosphate (-0.177; $P = 1.65 \cdot 10^{-9}$, glutamine (-0.177, $P = 1.81 \cdot 10^{-9}$), glucose-6-phosphate (-0.175; $P = 2.44 \cdot 10^{-9}$ and citric acid (-0.175; $P = 2.80 \cdot 10^{-9}$). Their individual contribution to the explained variance is smaller than 5.64 %.

In the second approach we applied multivariate tools to analyze the relationships between the two large groups of metabolite and biomass variables. Canonical correlation analysis (CCA) is a multivariate technique often used in psychological, climate and ecological studies to quantify the associations between two separate data sets measured on the same experimental units (Gittins, 1985; Hotelling and Gittins, 1935; Laudadio *et al.*, 2005). In contrast to the aforementioned pairwise correlation analysis, CCA yielded a much stronger correlation of 0.73. This value corresponds to 53.29 % of variance explained by the linear combination of metabolites, almost 10 times more than explained by any individual metabolite. To test the significance of this result, the biomass vector was permutated 50,000 times. At this point the maximum correlation did not increase significantly with additional permutations. This maximum value is 0.46. The distance between the median of the random correlations and the estimated value amounts to 17 standard deviations (Fig. 9*A*), which for normal distributions corresponds to a *P* value of $4.1 \cdot 10^{-65}$ strongly suggesting that the model is statistically highly significant.

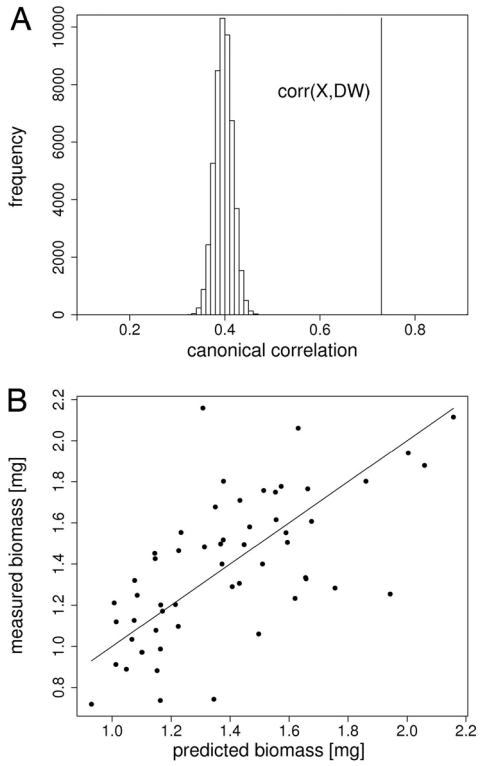


Figure 9 Significance (*A*) and predictive power (*B*) of the multiplicative model. (*A*) Histogram of canonical correlations between the metabolite matrix and random permutations of the biomass vector. The vertical line on the right corresponds to the canonical correlation between the actual biomass vector (DW) and the metabolite matrix (X). The distance to the median of the random correlations amounts to 17 standard deviations. (*B*) Prediction of the biomass by the metabolite matrix. Shown is one representative example of 20 repeats in the cross-validation. Size of the training set was 1,086, the 58 data points of the test set are displayed. The straight line represents the exact prediction.

3.3.3 Predictive Power of Metabolic Composition for Biomass

In a final step we wanted to test the predictive power of metabolite composition for biomass. To this end, we decided to apply the partial least square (PLS) approach, because CCA yields the maximum correlation and thus an upper limit for the true

correlation, but is notoriously inferior to other methods, especially PLS, for cross-validation (Frank and Friedman, 1993). (compare also *Supplemental Text*). Thus, the metabolite matrix and biomass vector were divided into training and test sets. The PLS coefficients estimated in the training set explaining 90 % of the variance of the training data were used to predict the biomass in the test set. This procedure was repeated 20 times. For a size of the training set of 1086 genotypes we obtained a median correlation between predicted and true biomass of 0.58 in the remaining 58 genotypes (representing the test set) confirming a strong predictive power of metabolic composition for biomass (Fig. 9*B*). To evaluate the significance the same permutation as for CCA was applied. For each of the 500 permutations a cross-validation was performed. The median of the corresponding correlations was -0.001 ± 0.052, thus, using the same assumption as above, we estimate a *P* value of $3.4 \cdot 10^{-29}$.

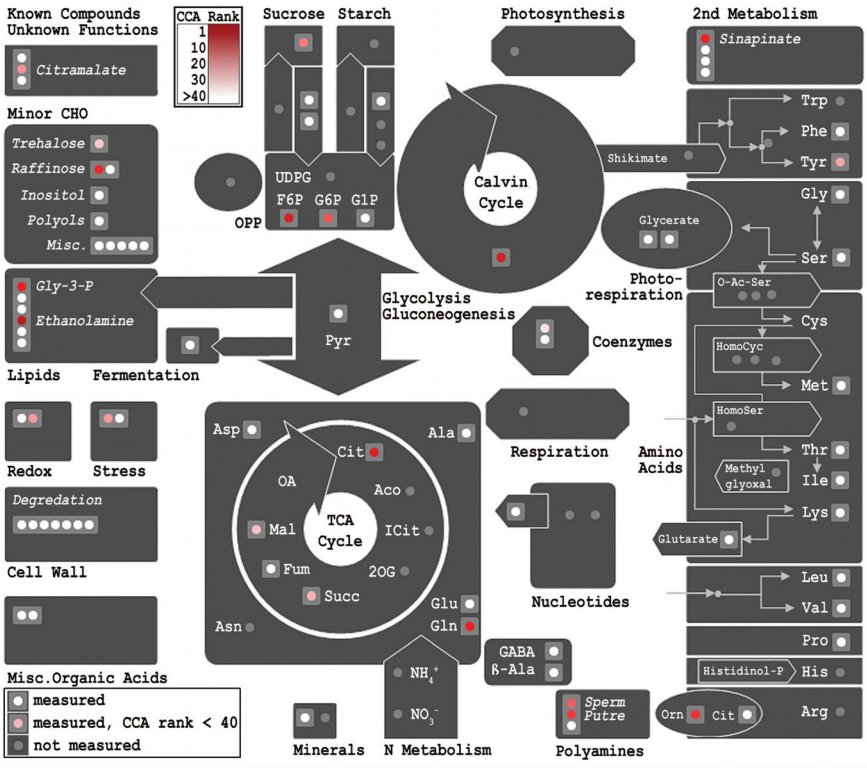


Figure 10 Representation of the most important metabolites known by structure according to CCA on biochemical pathways. This representation of metabolism indicates all known metabolites we analyzed by using GC/MS that could be annotated in MapMan (28). Red color visualizes metabolites which are high ranked in CCA (positions 1–44), with ranking according to the color-coded scale bar.

3.3.4 Metabolites Most Relevant for Biomass Accumulation

As a next step in our analysis we extracted the metabolites most relevant for biomass accumulation by their correlation to the canonical variate (Razavi *et al.*, 2005). The first 44 metabolites with significant correlations are listed in Table 1 and displayed on biochemical pathways (Thimm *et al.*, 2004) in Fig. 10. Strongly represented are

central metabolism derived compounds such as glucose-6-phosphate and fructose-6-phosphate, members of the tricarboxylic acid (TCA) cycle such as succinate, citrate and malate, members of the membrane/ phospholipid biosynthesis such as glycerol-3-phosphate, ethanolamine and sinapine, or sucrose. A list of all relevant metabolites is given in Supplemental Table 2.

METABOLITE	COR	PV
Unknown_038*	0.3688	0.00E+00
Unknown_035*	0.311	0.00E+00
Ethanolamine	0.296	0.00E+00
Unknown_086*	-0.2738	1.51E-24
Fructose 6-phosphate	-0.2449	3.66E-16
Citric acid	-0.2373	6.12E-18
Unknown_078*	0.237	1.01E-12
Unknown_061*	0.2241	4.52E-13
Glutamine	-0.2227	1.68E-12
Glycerol-3-phosphate	-0.2222	1.82E-13
Sinapic acid (cis)	-0.205	3.29E-10
Raffinose	-0.1964	7.04E-08
Glucose 6-phosphate	-0.192	4.92E-14
Putrescine	0.1918	1.49E-12
Ornithine	0.1905	2.85E-13
Unknown_074	0.1875	9.58E-08
Sucrose	-0.1857	7.01E-10
Unknown_051	0.1851	4.81E-08
Unknown_048	-0.1835	8.13E-09
Spermidine (major)	0.175	7.15E-10
Sinapic acid (trans)	-0.1737	7.09E-07
Citramalic acid	-0.1699	1.06E-10
Ascorbic acid	-0.1656	4.50E-07
Tyrosine	-0.1585	1.29E-03
Unknown_062	-0.1544	1.67E-08
Unknown_071*	-0.1497	1.25E-06
Succinic acid	-0.1472	5.00E-05
Trehalose	0.1418	3.44E-05
Malic acid	-0.1402	2.08E-10
Unknown_091	0.1377	2.20E-06
Unknown_060	0.1335	3.06E-07
Unknown_063	0.1301	6.77E-06
Unknown_033	0.1284	5.46E-06
Unknown_054	-0.1264	2.25E-06
Nicotinic acid	0.1235	2.58E-06
Unknown_043	0.1188	3.51E-05
Propanoic acid	-0.1159	5.97E-04
Maleic acid	-0.1154	1.71E-06
Unknown_079	0.1154	4.21E-05
Unknown_011	0.114	4.34E-03
Unknown_021	-0.1132	7.28E-06
Phenylalanine	-0.1105	2.16E-04
Unknown_084	0.1103	1.35E-04
Unknown_056	0.1099	9.45E-05

Table 1 List of top 44 signature metabolites ranked according to the strength of the canonical correlation. Given are correlation (COR) and corresponding *P* value (PV). * MassSpectrum indicates following chemical classes for these unknown compounds: 038, sugar; 035, glucopyranoside; 086, lactobionic acid; 078, pyranoside; 061, polyol; 071, sugar phosphate.

3.4 Discussion

We took advantage of an *Arabidopsis thaliana* RIL population for a parallel and integrative analysis of vegetative biomass accumulation and metabolic composition to answer the question whether or not biomass can be described as a function of metabolic composition.

As outlined in the Results section, pairwise correlation analysis of biomass and single metabolites could explain a maximum of 7 % of the total variance observed in biomass. These data strongly suggest that there is no single ‘‘magic’’ compound detectable, which could explain the biomass variance in a satisfying way. In contrast, canonical correlation analysis yielded a highly significant (the estimated *P* value based on permutations is lower than 10^{-64}) canonical correlation of 0.73 (compare Fig. 9*A*). Furthermore, in cross-validations a median correlation of 0.58 between the predicted and the observed biomass was observed (compare Fig. 9*B*).

This result demonstrates that a combination of the levels of a large number of metabolites rather than few individual metabolites show a close correlation with growth. It indicates that variation in growth coincides with characteristic combinatorial changes of metabolite levels, whereas individual metabolites may fluctuate largely independently of alterations in growth. To exclude the possibility that the strong correlation between biomass and metabolic composition is simply due to coincidental overlap of quantitative trait loci (QTL) for biomass and metabolites, we performed a QTL analysis on the RIL data set (429 lines) and detected a total of 157 QTL for 84 metabolites and six QTL for biomass (data not shown). Of the latter only two co-locate with significantly more metabolite QTL than expected by random, thus making this explanation highly unlikely.

Inspection of the metabolites highly ranked in CCA and thus representing the main drivers of the correlation shows that central metabolism derived metabolites are strongly represented. Of high relevance are the three metabolic intermediates of the hexose phosphate pool, fructose-6-phosphate, glucose-6-phosphate, and glucose-1-phosphate, which link carbon flow from photosynthesis and starch and sucrose metabolism with cell wall formation, the oxidative pentose phosphate pathway (it provides substrates for nucleic acid synthesis and for lignin, polyphenol and amino acid synthesis) and glycolysis. Members of the TCA cycle such as succinate, citrate, and malate are highly ranked. This finding underpins the central importance of this pathway which together with reactions of the glycolysis pathway and the oxidative

phosphorylation constitutes a key process delivering carbon skeletons, reduction equivalents, and energy for the vast majority of biochemical pathways. Also highly ranked is sucrose, the major transport form of carbon from source to sink tissue and which is central to the export from the sources and the import to the sinks. It thus represents the interface between carbohydrate production and utilization at the whole plant level Other metabolites such as glycerol-3-phosphate or ethanolamine play a major role in membrane/phospholipid biosynthesis. The anti-oxidant ascorbic acid (vitamin C) has been implicated in cell division (Liso *et al.*, 1984) and plant growth regulation by means of its role as enzyme cofactor (Smirnoff, 2000). Glutamine as a central metabolite in nitrogen assimilation and the major primary donor of reduced nitrogen is also found amongst the most important metabolites. This observation is contrasted by the fact that nearly all other amino acids analyzed are of rather low contribution based on the CCA. Further highly ranking metabolites can be assigned to general stress metabolites such as sinapine as the major phenylpropanoid in *Brassicaceae*, ornithine, the polyamines putrescine and spermidine, and trehalose. Thus, a link between the metabolites ranked high in the CCA and biomass accumulation is plausible because central metabolism and stress response are of utmost importance to plant growth, and thus biomass.

Another noteworthy observation is that the canonical variate determined by means of a multiplicative model resulted in closer correlations between the predicted and the observed biomass values than by means of an additive model (data not shown). It indicates that the involved metabolites act synergistically rather than additively which is very plausible as the aforementioned closely interlinked pathways of carbon metabolism are required for different cellular components that all are crucial for growth/biomass formation. The strong reciprocal interrelation between nitrogen and carbon assimilation would also strongly argue for synergistic and not additive effects between key metabolites representing these classes of biochemical compounds as observed in our case. Similar arguments can be made for e.g., ethanolamine synthesized via serine as a major constituent of membranes or sinapic acid as the major phenylpropanoid component in *Arabidopsis*.

A surprising observation from our data are the occurrence of both positive and negative correlations between metabolites and biomass. The large majority of known metabolites displaying a negative correlation to the biomass vector are the aforementioned intermediates of central metabolic pathways including sucrose,

glucose- and fructose-6-phosphate, the TCA cycle members citric acid, succinate or malic acid, as well as the amino acids glutamine and phenylalanine. On the other hand, amongst the positively correlated metabolites are a large fraction of unknown chemical structure as well as some metabolites discussed in stress response such as nicotinic acid (Hageman and Stierum, 2001) or putrescine (Tkachenko *et al.*, 2001), or the stress metabolite trehalose discussed in connection with drought resistance (Garg *et al.*, 2002). A negative correlation suggests that pool sizes of these metabolites are reduced to low levels when strong growth occurs. It is conceivable that this process involves mostly metabolites providing the major building blocks for growth such as the central metabolites mentioned. In conclusion, this observation would suggest that growth drives metabolism and not vice versa. This finding would indicate that high growth rates cause a depletion of central metabolite pools rather than growth being enhanced through increased supply of substrates for the synthesis of cellular components. A similar conclusion of metabolism driven by growth has been derived from a study of the relationship between tomato fruit size and metabolites (Schauer *et al.*, 2006). In this scenario, the positively correlated metabolites could play a role in plant defense against abiotic and biotic stress and it is comprehensible that higher concentrations of these metabolites would coincide with better armed plants. For both groups of substances, however, the relation with growth may be nonlinear. On the one hand, the reduction of central metabolite levels below a certain minimum necessary to sustain high flux rates may result in growth limitation and thus a breakdown of a linear negative relationship. Similarly, a positive effect on growth because of elevated stress tolerance may be achieved in a certain range of stress metabolite levels above which no further beneficial or even detrimental effects may occur. As the procedures applied here determine linear correlations, it is not unexpected that no tighter relationships (stronger correlations) were detected. A complementary hypothesis regards metabolites not primarily as chemicals for growth and defense but rather as signals. Under this assumption positively correlated metabolites are positive signals regulating plant growth and the contrary would be true for negatively correlated metabolites. In the context of signal molecules the large number of positively correlated compounds of as yet unknown structure is worth noting and stresses the need for identification of their chemical nature. They might constitute unusual products of metabolic side reactions that are derived from primary metabolites generated for signaling purposes and which can

move to sites of perception without further conversion along the major metabolic reactions or transport pathways. Further studies querying some testable predictions from such models (e.g., the presence of receptors/ sensors or the elicitation of specific responses in case of signaling metabolites) are needed to validate these models.

3.5 Conclusion

Using an *Arabidopsis thaliana* RIL population and conducting a combined analysis of biomass and metabolite profiles allowed the prediction of biomass as a function of metabolic composition providing a direct proof for the hypothesis that metabolic composition is related to biomass and thus growth. The observations made here further extend this hypothesis toward the notion that major global changes in metabolism are the result of variation in growth rather than vice versa. In addition to fostering our basic understanding, these data are of immediate potential for a number of applied purposes. The possibility to predict biomass on the basis of the metabolic signature of a plant presents a first precedence for the use of metabolite profiles as biomarkers with high predictive power and could potentially revolutionize the selection and thus breeding process for biomass producers such as trees that are cultivated for decades before harvest. Identification of highly productive genotypes already at an early growth stage would result in enormous time and cost-savings. In the light of reduced availability of fossil fuels and increasing reliance on bio-derived energy, the importance of such an opportunity can hardly be overestimated.

3.6 Material and Methods

3.6.1 Creation of Recombinant Inbred Line (RIL) Population

Two reciprocal sets of RILs were developed from a cross between the two *Arabidopsis thaliana* accessions C24 and Col-0 as described elsewhere (Törjék *et al.*, 2006). The population consisted of 228 Col-0×C24 F_8 and 201 C24×Col-0 F_8 individual lines.

3.6.2 Plant Cultivation

The RILs were planted in a split plot design with 54 incomplete blocks and four replicates, repeated six times. Plants were grown in 1:1 mixture of GS 90 soil and vermiculite in 96-well-trays. Six plants of the same line were grown per well. Seeds were germinated in a growth chamber at 6 °C for 2 days before transfer to a long-day

regime (16 h fluorescent light [120 µmol·m^{-2}·s^{-1}] at 20 °C and 60 % relative humidity/ 8 h dark at 18 °C and 75 % relative humidity). To avoid position effects, trays were rotated around the growth chamber every 2 days.

3.6.3 Shoot Dry Biomass

Shoot dry biomass was determined 15 days after sowing. Plants from the same well were harvested together and placed in a vacuum oven at 80 °C for 48 h. Dry biomass was recorded by using an analysis balance. Mean shoot dry biomass in mg per plant was estimated by using the linear mixed model (Piepho *et al.*, 2003) G + E:E·G + E·GC + E·GC·T where E is experiment, G is genotype, GC is growth chamber and T is tray (REML procedure in Genstat). Biomass in the two subpopulations was compared with a two-sided *t* test. We detected no significant differences in biomass (P = 0.238) between the two subpopulations, and treated the RILs as one population in subsequent analyses.

3.6.4 Metabolite Data. Sample Preparation, Measurement, and Data Processing

Samples for the analysis of metabolic composition were collected together with the material for dry biomass analysis at 15 days after sowing. Harvested material (shoot and leaf) was cooled below -80 °C immediately and kept at this temperature until further processing. Derivatization, GC-MS analysis, and data processing were done as described elsewhere (Lisec *et al.*, 2006). All 181 metabolic signatures that have been evaluated within this experiment are listed in Supplemental Table 3. The GC-MS spectra of evaluated metabolites that are unknown with respect to their actual chemical formula but can be repeatedly found in *Arabidopsis* are available in Supplemental Table 3.

The extracted metabolite data consist of unique mass intensity values for each referenced compound and measurement respectively. These raw data were normalized and otherwise directly used for analysis. This method allows between sample comparisons but no quantitative statements about single metabolites.

Normalization. Metabolite data were normalized by dividing each raw value by the median of all measurements of a day for one metabolite.

Missing Value Estimation. For the canonical correlation analysis (CCA) missing value replacement is necessary. The 6 % missing values in the metabolite matrix were imputed with a selforganizing map (SOM) algorithm (Kuss and Graepel, 2003). The mean square error was estimated by the comparison of known values with those

calculated from the SOM algorithm. The coefficient of variation (root mean square error divided by the mean) was 0.3.

3.6.5 Integrated Analysis of Phenotypic and Metabolite Data

Linear models for the relation between metabolite profile and biomass. The relation between biomass and metabolite profile was measured by simple Spearman correlation between the dry biomass and relative abundances of all metabolites, and by a more complex multiplicative model. The first corresponds to the following model, referred to as model 1:

$$B = c_i x_i$$

The second model can be described by:

$$B = \prod_i x_i^{c_i}$$

B denotes the biomass, *x* the relative metabolite abundance and *c* the corresponding constants for all *i* metabolites.

Multivariate linear analysis. Canonical correlation analysis (CCA) calculates the highest possible correlation between linear combinations of the columns from two matrices with the same number of rows. If the second matrix has only one column, this procedure corresponds to a ordinary least square (OLS) regression. The correlation thus found is called canonical correlation, the corresponding linear combination canonical variate. The mathematical foundation is described in the literature (Hotelling and Gittins, 1935; Kuss and Graepel, 2003).

The *R* function *cancor* was used to calculate the canonical correlation between metabolites and biomass. For crossvalidation a partial least square (PLS) regression was performed. This method (Wold, 1975) seeks to maximize the covariance instead of the correlation between the matrices. To carry out the procedure the *R* function *plsr* was used. These functions are publicly available (www.r-project.org). All procedures were applied after missing value estimation followed by normalization of the metabolic matrix. To test the robustness of the selection of the signature metabolites, we applied the following procedure: with 90 % of the 1,144 genotypes chosen at random the canonical variate was calculated and the important metabolites selected as described above. Selected metabolites, which were not in the original list of 44 metabolites, were regarded as false. This procedure was repeated 100 times. We obtained a median ‘‘false positive rate’’ of 0.048 (± 0.034).

4 Identification of metabolic and biomass QTL in *Arabidopsis thaliana* in a parallel analysis of RIL and IL populations

Jan Lisec[1,*,§], Rhonda C. Meyer[2,*], Matthias Steinfath[3,*], Henning Redestig[1], Martina Becher[2], Hanna Witucka-Wall[2], Oliver Fiehn[1,4], Ottó Törjék[2], Joachim Selbig[1,3], Thomas Altmann[1,2] and Lothar Willmitzer[1]

[1] Max-Planck-Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476 Golm, Germany

[2,3] Departments of [2]Genetics and [3]Bioinformatics, Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Strasse 24–25, 14476 Potsdam, Germany

[4] Present address: University of California at Davis Genome Center, GBSF Building Room 1315, 451 East Health Sciences Drive, Davis, CA 95 616-8816, USA

[*] These authors contributed equally to this work.

[§] JL contributed to method development, data analysis and preparation of the manuscript

4.1 Abstract

Plant growth and development are tightly linked to primary metabolism and are subject to natural variation. In order to obtain an insight into the genetic factors controlling biomass and primary metabolism and to determine their relationships, two *Arabidopsis thaliana* populations [429 recombinant inbred lines (RIL) and 97 introgression lines (IL), derived from accessions Col-0 and C24] were analyzed with respect to biomass and metabolic composition using a mass spectrometry-based metabolic profiling approach. Six and 157 quantitative trait loci (QTL) were identified for biomass and metabolic content, respectively. Two biomass QTL coincide with significantly more metabolic QTL (mQTL) than statistically expected, supporting the notion that the metabolic profile and biomass accumulation of a plant are linked. On the same basis, three out the six biomass QTL can be simulated purely on the basis of metabolic composition. QTL based on analysis of the introgression lines were in substantial agreement with the RIL-based results: five of six biomass QTL and 55% of the mQTL found in the RIL population were also found in the IL population at a significance level of $P \leq 0.05$, with >80% agreement on the allele effects. Some of the differences could be attributed to epistatic interactions. Depending on the search conditions, metabolic pathway-derived candidate genes were found for 24–67% of all tested mQTL in the database AraCyc 3.5. This dataset thus provides a comprehensive basis for the detection of functionally relevant variation in known genes with metabolic function and for identification of genes with hitherto unknown roles in the control of metabolism.

4.2 Introduction

The phenotype displayed by an organism is the result of interaction between its genotype and the environment. Natural genetic variation is usually due to effects of multiple genes detectable as quantitative trait loci (QTL), and the expression of complex traits is the result of the contribution and interaction of numerous genes. One particular example of this is the growth of multicellular organisms, which has been shown to be governed by many genes that each contribute a small portion to the overall phenotype, for example in mouse (Rocha *et al.*, 2004), chicken (Jacobsson *et al.*, 2005), Arabidopsis (El-Lithy *et al.*, 2004) or rice (Li *et al.*, 2006).

In plants, numerous transgenic single-gene-driven attempts have been described with the goal of modifying growth and/or biomass. Many of these have targeted the production and/or distribution of primary metabolites within various parts of the plant such as the source and sink organs, i.e. the growing areas and storage organs, respectively (Sonnewald *et al.*, 1994). However, it is fair to say that the success rate has been rather limited. On the other hand, numerous transgenic approaches have been utilized in an attempt to improve the metabolic composition of plants to meet requirements with respect to human food and animal feed. In such cases, the success rate has varied with the pathway targeted. Transgenic approaches have shown an impressively high success rate when applied to secondary metabolites such as carotenoids or flavonoids, or when applied to polymer quality (Lorberth *et al.*, 1998; Mann *et al.*, 2000; Muir *et al.*, 2001). As a rule, the intended biochemical changes were achieved and were not accompanied by any major pleiotropic effects concerning growth and development. In contrast, when attempting to modify primary metabolism, such as sucrose biosynthesis or the tricarboxylic acid (TCA) cycle, major and mostly negative effects at the whole-plant level, specifically impaired growth and development, were observed in many cases (Trethewey *et al.*, 1998).

Variation of growth and metabolic traits has been detected for a series of natural accessions and recombinant inbred lines (Cross *et al.*, 2006; Meyer *et al.*, 2007a). Correlation analyses showed weak relationships between growth and the levels of individual metabolites, but a close and highly significant link between biomass and a specific combination of metabolites has been shown (Meyer *et al.*, 2007a). The observation of positive correlations of rosette weight with several enzyme activities indicated the importance of the catalytic activity of enzymes in central carbon and nitrogen metabolism and their effects on metabolic fluxes (Cross *et al.*, 2006).

Taken together, these data indicate that primary metabolism, in contrast to secondary metabolism, is a network that is closely linked to plant growth and development, and that major perturbation of this network has strong detrimental effects on plant performance.

In order to obtain further insight into the genetic factors that control growth and metabolic traits and to elucidate their relationships, we performed a parallel QTL analysis for biomass and metabolic composition. To this end, metabolic profiling using GC-TOF mass spectrometry was applied to recombinant inbred line (RIL) and introgression line (IL) populations of *Arabidopsis thaliana* according to the concept of genetical genomics (Jansen and Nap, 2001). All plants were derived from a cross between the *Arabidopsis thaliana* accessions Col-0 and C24 (Törjék *et al.*, 2006). Using the data obtained with regard to growth and metabolite composition, we addressed the following questions:

(i) Is the heritable variation in these populations and its genetic basis sufficient to allow identification of biomass QTL and metabolic QTL?

(ii) Are metabolic QTL randomly distributed over the genome?

(iii) Can links between metabolism and growth be established on the basis of a statistically significant co-localization of shoot biomass and metabolic QTL?

(iv) What fraction of metabolic QTL regions contain candidate genes with related known or proposed metabolic function, and how many of these candidate genes show sequence variation leading to changes in the encoded protein?

As the Arabidopsis genome is fully sequenced (Arabidopsis Genome Initiative, 2000), well annotated (Haas *et al.*, 2005) and was recently very thoroughly analyzed for its genetic diversity across 20 accessions (Clark *et al.*, 2007), we were able to investigate 39 of the 85 metabolites of known chemical nature. An analysis was also performed to answer the question of whether analyses of RILs and ILs lead to similar or different results, and thus to what extent the two approaches can be considered complementary or redundant.

The results we present here demonstrate that, at least for a subset of the biomass QTL, there is substantial and significant overlap with metabolic QTL, suggesting a strong link between biomass and primary metabolism. In addition, QTL have been identified for multiple metabolites, and a candidate gene was identified for up to 67 % of them. Five biomass QTL were identified in both the RIL and IL populations, and

55 % of the mQTL identified in the RIL population were confirmed in the IL population.

4.3 Results

4.3.1 Analysis of the RIL population for biomass and metabolic QTL

Description of the RIL population and QTL mapping. The analyzed RIL population (Törjék *et al.*, 2006) consisted of 429 lines from the reciprocal crosses Col-0 × C24 (n = 228) and C24 × Col-0 (n = 201) grown under controlled conditions in six consecutive experiments, in which each line was replicated at least three times. Plants harvested 15 days after sowing were used for shoot biomass determination, or were pooled and frozen, and subsequently subjected to metabolite profiling by GC–MS. We did not find significant differences in marker distribution between the two sub-populations (association between marker matrices estimated by Mantel test, $P < 0.001$). As we could not detect a significant difference in biomass between the two sub-populations either (Kolmogorov–Smirnov test, $P = 0.180$), we treated the RILs as one population in subsequent analyses.

The shoot biomass and metabolite data were used to map QTL based on a linkage map of 105 markers established for the Col-0/C24 RIL population (Törjék *et al.*, 2006) by application of the software packages PLABQTL (Utz and Melchinger, 1996) and QTL Cartographer (Basten *et al.*, 1994).

To identify the fraction of variation that is genetically determined and can potentially be mapped into mQTL, we estimated broad- and narrow-sense heritability for all metabolic traits as described in Experimental procedures. Broad-sense heritability was determined as $H^2 = 0.40$, on average. Narrow-sense heritability was $h^2 = 0.08$, with $h^2 = 0.16$ for metabolites showing at least one QTL and $h^2 = 0.02$ for the remaining metabolites. For biomass, h^2 was determined to be 0.70.

Six biomass QTL explain 18 % of the phenotypic variation. A complete list and description of QTL detected for shoot biomass is given in Supplemental Table S4. The explained phenotypic (denoted by R^2) and genotypic variation were determined from the final simultaneous fit of all putative QTL using PLABQTL. For biomass, six QTL explain 18.5 ± 3.4 % of the phenotypic and 26.8 ± 4.9 % of the genotypic variation. Individual QTL contributions range from 1.5 to 6.0 % of the total variance. The mean R^2 after cross-validation was 16.01 % in the calibration and 8.92 % in the validation, for a mean of six QTL.

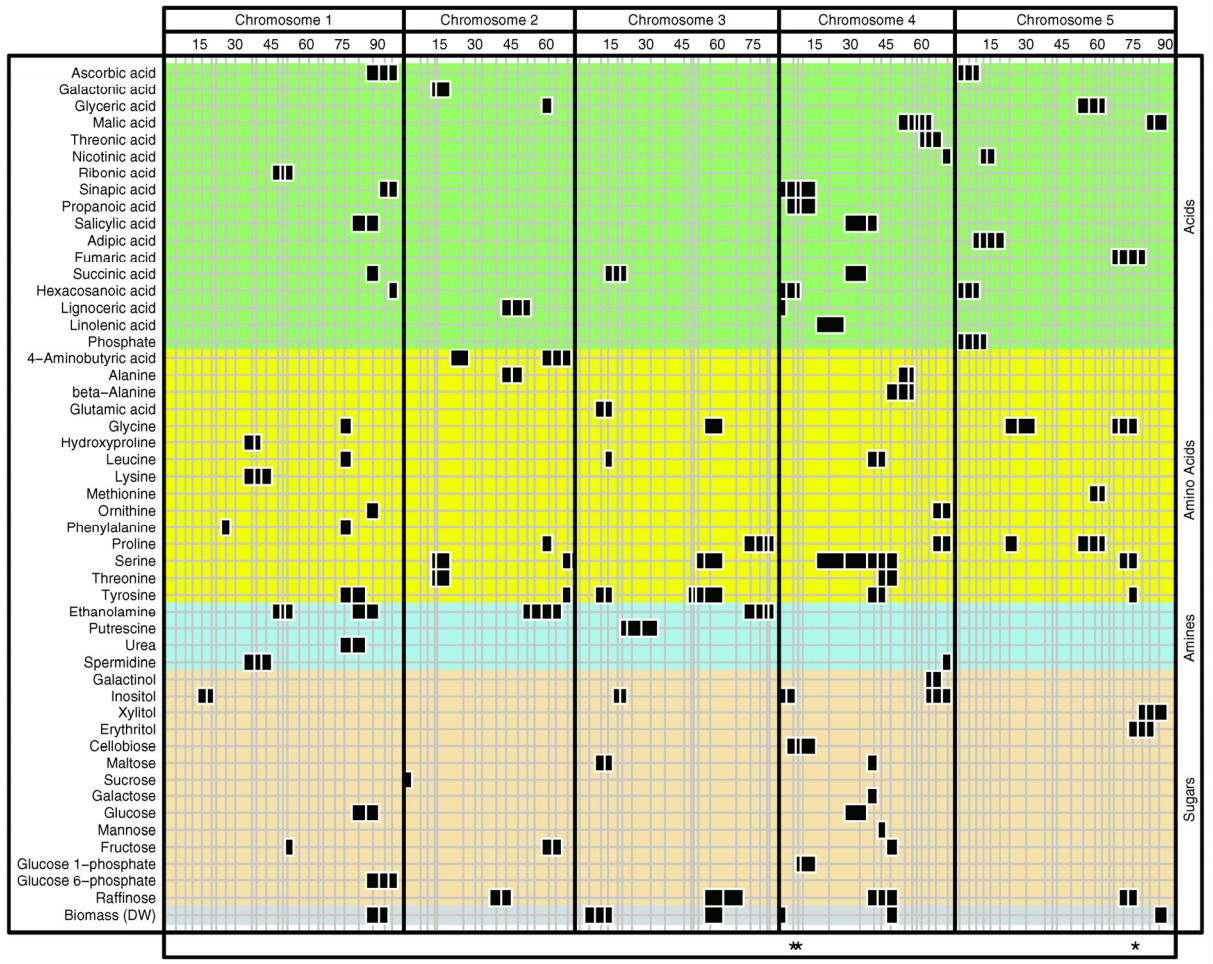


Figure 11 Distribution of metabolic and biomass QTL. Significant metabolic QTL of metabolites known by structure are shown as black boxes at marker positions if covered by the support interval. For simplicity, the QTL of metabolites of unknown structure are omitted here. Information on all detected QTL is given in Supplemental Table S4. Metabolites are color-coded according to their chemical group as shown on the right. Vertical lines indicate marker positions, several of which are labeled with approximate distance in cM (top). Asterisks indicate QTL 'hot spots' (as determined using 1,000 permutations at a 0.05 level).

Identification of mQTL for 84 metabolites. Samples taken from 369 of the 429 RILs were analyzed for their metabolic composition. A total of 181 compounds could be detected in more than 85 % of all samples, and only those metabolites were taken into further consideration. The chemical nature is known for 85 of these compounds.

In total, we found 157 metabolic QTL for 84 metabolites, 50 of which are of known chemical structure. For 42 metabolites, only one QTL was identified, but a maximum of six QTL was found for tyrosine. The QTL are distributed unequally over marker positions, indicating 'hot spots' and empty regions (no metabolic QTL at 10 marker positions). The contribution of individual QTL to the phenotypic variation varied between 1.7 (unknown_092) and 52.1 % (cellobiose).

A comparative overview of QTL for known metabolites and biomass is presented in Figure 11.

Shared mQTL enriched for metabolites showing strong pairwise correlation. As a large fraction of the observed variation is due to genetic effects, concentrations of metabolites with shared QTL are expected to correlate, and with increasing numbers of co-located QTL the correlation may increase. Metabolite correlations may be caused by common genetic factors, e.g. regulatory or pathway genes. On the other hand, even if co-located QTL exist, the corresponding metabolites may be weakly correlated if their QTL show strong interaction with other loci that are different for the two metabolites in question. Alternatively, the metabolites may be subject to differential metabolic control, or may be differently affected by environmental influence. To test this, we plotted the number of QTL shared between two metabolites against the value for the Pearson correlation determined between the concentrations of the two metabolites measured in all RILs (Figure 12). The chance of sharing at least one QTL increases with stronger correlations, and overall the correlation increases with the number of shared QTL. However, examples of two deviant scenarios were found: (i) metabolite concentrations are highly correlated but are controlled by QTL at different positions [e.g. glucose and fructose ($r = 0.849$) show all together five individual QTL but none are shared], and (ii) metabolite concentrations are weakly correlated but they share common QTL (e.g. ethanolamine and fructose ($r = 0.058$) show three and four individual QTL, respectively, with two QTL in common).

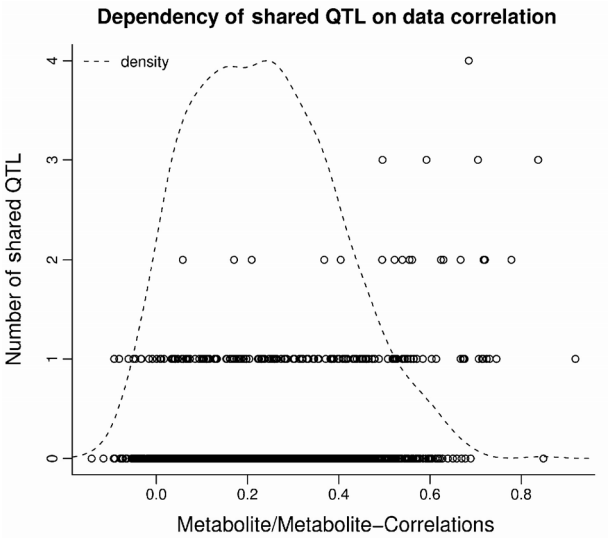


Figure 12 Dependency of shared QTL on data correlation. The number of overlapping QTL between two metabolites is plotted against the Pearson correlation value for the data vectors used for QTL calculation. Higher numbers of shared QTL are predominantly found for more strongly correlated traits. No normalization was applied with respect to the total number of determined QTL per trait.

Candidate genes involved in the biochemical pathways of the respective metabolite are identified for 24–67 % of the metabolic QTL. Initial analyses of detected metabolic QTL with respect to underlying biochemical pathways show that it is possible to identify candidate genes even at the rather low mapping resolution that can be achieved using an RIL population. For example, inspection of the available information on pathways involving *myo*-inositol suggested candidate genes for three of four identified QTL (Supplemental Table S4 and Figure 13).

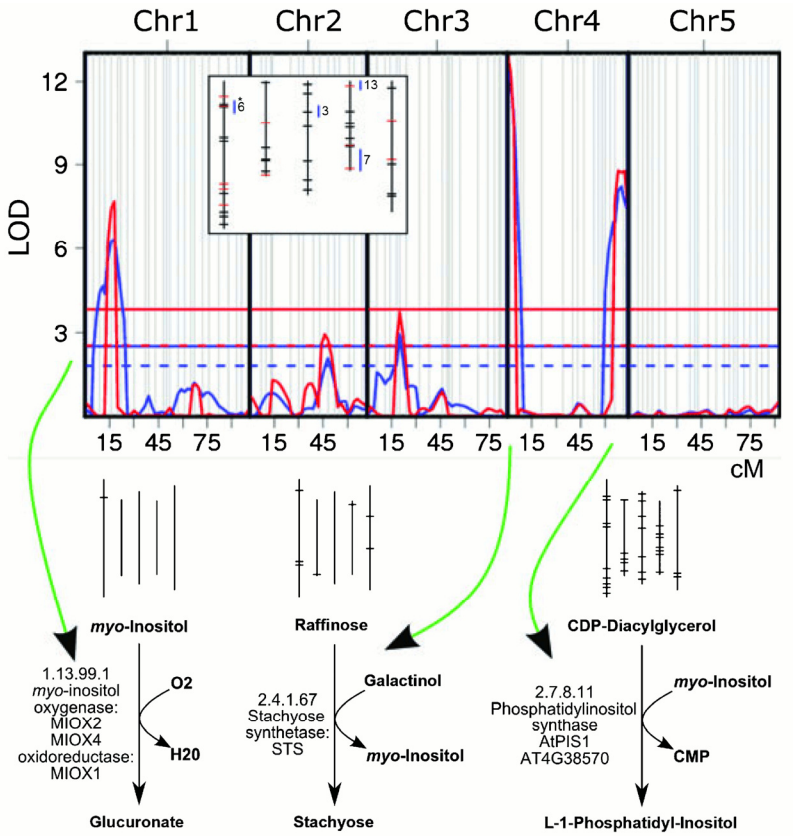


Figure 13 *myo*-inositol QTL analysis reveals direct candidate genes for three of four determined QTL (1/18, 4/0 and 4/65). A LOD curve calculated using two independent programs (PLABQTL, red lines; QTL Cartographer, blue lines) is shown at the top. Horizontal lines indicate 0.05 (solid) and 0.25 (dotted) significance thresholds calculated based on 5000 permutations. Vertical lines indicate marker positions. At the bottom, the three relevant reaction steps according to the mQTL as connected by arrows are presented (pathways from left to right are inositol oxidation, stachyose biosynthesis and phospholipids biosynthesis). The pictograms in the center indicate the total number and location of genes known per pathway. Twelve genes (from six pathways) for enzymes catalyzing reactions in which *myo*-inositol is involved directly are known. The insert shows a comprehensive view of all AGI codes associated with *myo*-inositol (red, direct; black, pathway), indicating mQTL support intervals (blue), approximate LOD (number) and IL confirmation threshold reached (asterisk). A similar plot for all known metabolites is shown in Supplemental Figure S1.

The AraCyc section of the TAIR database lists only 12 loci representing enzymes that catalyze reactions on *myo*-inositol. Three of these loci co-locate with determined mQTL: a *myo*-inositol oxygenase (AT1G14520, inositol oxidation pathway), a phosphatidyltransferase (AT4G38570, phospholipid biosynthesis pathway) and a

stachyose synthase (AT4G01970, stachyose biosynthesis pathway). If all genes from pathways involving *myo*-inositol are considered, it is possible to find candidates for the remaining mQTL of this metabolite.

We extended this analysis to all metabolites of known chemical structure, considering either (i) only the genes encoding enzymes that participate in a direct reaction with the respective metabolite, or (ii) all proteins assigned to pathways involving the metabolite. We were able to identify at least one candidate gene for 24 % (direct reaction) and 67 % (pathway enzymes) of all tested mQTL (Supplemental Table S4).

In order to assess how much this coincidence of mQTL and enzyme genes relating to the respective metabolites deviates from the situation expected based on a random distribution of mQTL, we performed a permutation test in which we distributed for each of the 38 metabolites with assigned AGI codes all determined mQTL randomly over all chromosomes. We analyzed the overlap with potential candidate genes from AraCyc as described above, and compared the outcome with the results based on the measurement data. For 13 metabolites, we found more candidate genes than found on average in permutations, while three showed fewer genes. In most of the remaining cases, no candidate gene was assigned either in our experiment or in permutations. The number of identified candidate genes for *myo*-inositol, maltose and ethanolamine exceed the 95th perentile of the respective permutation results. However, although this analysis suggests that experimentally determined mQTL are enriched for corresponding pathway genes, the test statistic is not significant if multiple testing is considered.

For 10 of 20 mQTL, at least one of the direct candidate genes contains a polymorphism in the protein coding region leading to an amino acid exchange, according to recently published data on single nucleotide polymorphisms (SNPs) between C24 and Col-0 (Supplemental Table S5; Clark *et al.* [2007]).

Co-localization of biomass QTL with mQTL. One of the aims of this project was to determine the relationship between biomass QTL and mQTL, i.e. to what extent they co-localize. Inspection of the overlap between mQTL and biomass QTL in the RIL population shows that each biomass QTL coincides with several mQTL (with the number of mQTL per biomass QTL ranging from 5 to 12). However, due to the limited resolution of the QTL mapping, a considerable number of overlaps are expected to occur by chance. We therefore used a permutation test to identify statistically significant overlaps. This analysis showed that two of the six biomass QTL (1/88 and

4/0) co-locate with significantly more mQTL than expected by chance. Some metabolites (raffinose, tyrosine, serine, succinic acid) display up to two QTL co-localized with any of the biomass QTL. However, no enrichment for single compound classes or certain biochemical pathways was found amongst these metabolites.

Epistasis. Due to the nature of metabolic networks consisting of multiple interconnected metabolic pathways, prevalent epistatic interactions are expected to occur among mQTL, influencing various steps within single or among multiple pathways. Therefore, we conducted a full scan for all possible digenic maker interactions using the PLABBIC version of PLABQTL, but no significant digenic marker epistasis for the biomass and metabolite traits was detected. To reduce the multiple-testing problem inherent in this approach, we adjusted the procedure to test only previously determined mQTL of known metabolites against markers located elsewhere in the genome. From the resulting likelihood profile, we kept the maxima and evaluated *per se* mQTL and epistatic effect maxima in a final model. In this last step, non-significant effects were dropped according to a Bayesian information criterion. Following this procedure, one significant epistatic effect was detected for biomass (1/88 × 3/82, R^2 = 2.17%) and a further 38 such interactions were identified for 27 of the 50 known metabolites taken into consideration. However, these effects are rather small if compared to *per se* mQTL, explaining only 2.72 % of the phenotypic variation on average. The strongest interaction (R^2 = 4.92%) was determined between two tyrosine QTL (4/42 and 5/74). Other substantial effects were identified for glycerate (2/61, R^2 = 4.43 %) and maltose (4/38, R^2 = 3.69 %), which exhibited additive interactions with genomic positions to which no mQTL had been previously assigned (Supplemental Table S6).

4.3.2 Analysis of the IL populations and comparison with the RIL-based data

Five common biomass QTL detected in IL and RIL populations. A QTL analysis was also carried out using 97 lines of two corresponding reciprocal IL populations of the crosses Col-0×C24 and C24×Col-0 (Törjék *et al.*, unpublished data) in order to verify the QTL detected in the RIL population. Biomass data were analyzed using the appropriate contrasts in anova. Twenty-six IL were significantly different ($P < 0.05$) from the recurrent parental line and thus identified biomass QTL. The six biomass QTL regions previously identified by the RIL analysis were covered by another set of 26 ILs. The intersection of the two sets consists of 13 ILs. To compare this result to a random intersection, we calculated the probability of identifying 13 or more of 26

fixed ILs when 26 are drawn randomly from 97. This probability is given by the hypergeometric distribution. Its value of < 0.003 demonstrates the significance of the finding. By means of the IL analysis, five (of six) biomass QTL detected by the RIL analysis were verified at positions 1/88, 3/13, 3/59, 4/47 and 5/86 (Supplemental Table S4). In addition, the IL analysis revealed a further four regions with an effect on biomass at positions 1/10, 2/72, 3/46 and 5/62-67. Detailed analyses of the individual ILs indicated complex situations, e.g. on chromosome 1 with potentially two QTL of opposing effects located very closely to each other.

RIL-based metabolic QTL are also detected in the introgression lines. Detection of changes in metabolite concentrations due to introgression of a donor genotype in the background was used to confirm mQTL determined in the RIL population. For 94 of the ILs, six replicate GC–MS measurements were carried out, and compared against the metabolite values determined in up to 30 measurements for the respective parental lines. Due to the different population size, *P*-values were estimated as described in Experimental procedures.

At a level of $P \leq 0.05$ (not multiple-testing-corrected), 55 % of the RIL mQTL are confirmed within the IL population. In 82 % of the cases, the positive-effect allele was also confirmed, i.e. if the C24 genotype in the RIL population showed an increased metabolite level compared to Col-0, the same was true for the respective IL. This high level of allele confirmation is independent of the *P*-value applied (see Table 2).

Significance level	Number of significant changes	FDR (%)	Number of confirmed RIL QTL	Confirmed RIL QTL (%)	Average R^2 of confirmed RIL QTL (%)	Average R^2 of non confirmed RIL QTL (%)	Confirmed allelic effect	Confirmed allelic effect (%)
0.001	177	9.61	17	11.33	11.62	6.67	16	94
0.01	773	22	41	27.33	10.17	6.12	38	93
0.05	2511	33.9	83	55.33	7.79	6.54	68	82
0.1	3941	43.2	99	66	7.45	6.79	80	81

Table 2 Estimated *P*-values for IL–parent comparisons. Significant results and RIL QTL confirmation at various threshold levels. The false dicovery rate (FDR) is defined as the expectation of the ratio of false positives to the sum of false and true positives. We estimated the FDR by (significance level × number of observations)/(number of significant changes). R^2, phenotypic explained variance.

RIL-based mQTL explaining a large part of the phenotypic variance were confirmed preferentially in the IL population. This becomes even more evident when the significance threshold is lowered. At *P*-values ≤ 0.001, differences between ILs and parents were detected for 177 metabolite/ IL combinations (equivalent to 177 mQTL).

Of these, 17 had been observed previously in the RILs (11 % confirmation). The mean contribution to phenotypic variance of confirmed QTL is 11.6 %, compared with 6.4 % for non-confirmed QTL.

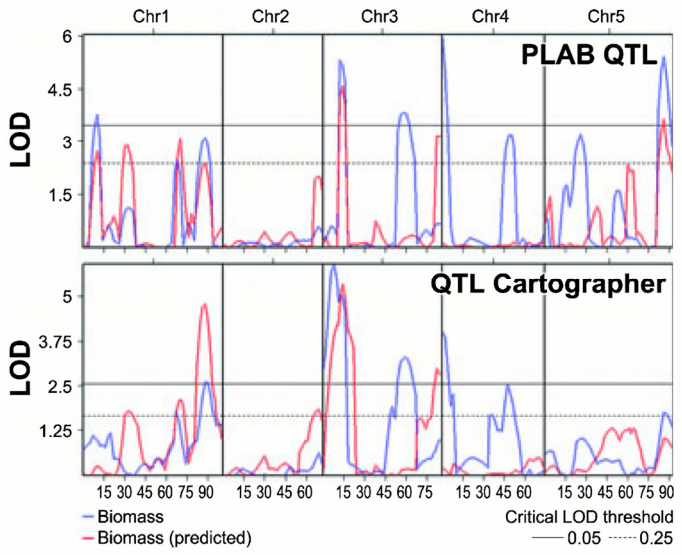


Figure 14 Meta-QTL analysis. Meta-QTL analysis using the measured biomass (blue line) and the canonical variate (predicted biomass, red line) calculated from the metabolic profiles as described by Meyer *et al.* (2007). Horizontal lines indicate 0.05 (solid) and 0.25 (dotted) significance thresholds calculated based on 5,000 permutations. Chromosomal length is given in cM.

Prediction of biomass QTL via a combination of mQTL. We are interested in determining the link between biomass QTL and mQTL/ metabolic composition. We have previously shown that, for this RIL population, a canonical combination of metabolites can be used to predict the biomass (Meyer *et al.*, 2007a). We therefore performed a meta-QTL search using the predicted dry biomass vector (canonical variate) as a new trait to determine whether or not any of the biomass QTL can be predicted based on the metabolic composition (Figure 14). Of the six biomass QTL, three (1/88, 3/13 and 5/86) could be predicted by the metabolic composition, three were not predicted, and two new QTL appeared in the predicted pattern (1/10 and 3/82). One of these new peaks (1/10) corresponds to a QTL that was also identified in ILs.

4.4 Discussion

Several successful studies have been conducted to date to identify novel genes based on QTL analysis (Kliebenstein *et al.*, 2001; Kroymann *et al.*, 2003; Werner *et al.*, 2005; Zhang *et al.*, 2006). However, with a few exceptions (Keurentjes *et al.*, 2006; Schauer *et al.*, 2006; West *et al.*, 2007), only a limited number of traits (usually less than 20) have been assessed.

The parallel analysis of IL and RIL populations of *Arabidopsis thaliana* for biomass and 181 individual metabolites (for which the chemical structure is known for 85)

described here is a unique data set. Together with the available information from the fully sequenced and thoroughly annotated Arabidopsis genome, it provides a direct method for detection of functionally relevant variation in known genes with metabolic function and the identification of genes hitherto not assigned to metabolic functions, and emphasizes the link between metabolism and growth/ biomass accumulation.

4.4.1 Comparison of ILs versus RILs

Introgression lines are widely used to test for changes in various traits when compared to a parental line. Lines of an introgression population have a common genetic background and various short donor segments from another line, thus allowing focus on a specific region of the genome (Eshed and Zamir, 1995). Several studies have demonstrated the possibility of fine-mapping single RIL QTL in Arabidopsis using ILs (Alonso-Blanco *et al.*, 1998a; Bentsink *et al.*, 2003; Swarup *et al.*, 1999). Recently Keurentjes *et al.* (2007a) described an exhaustive analysis on the overlap between QTL based on RIL and IL populations derived from the *Arabidopsis thaliana* accessions Cvi and L*er*. Comparing QTL detected for six developmental traits analyzed in a RIL population of 167 lines and an IL population of 92 lines, 58 % of 33 RIL QTL were confirmed. Although these authors tested up to 116 replicas per IL in a BIN approach, allowing a stricter significance level, this figure is largely in agreement with our findings. Applying a *P*-value threshold of 0.05, 83 metabolic QTL detected in the RIL population (equivalent to 55 %) were confirmed in the IL population. In 82 % of the cases, the direction of the effect was the same in RILs and ILs, independent of the applied threshold.

However, not all QTL identified in the RIL population were confirmed using the ILs, and additional QTL were detected using the latter. The differences between RIL- and IL-based QTL (QTL_{RIL} and QTL_{IL}, respectively) can be explained to some extent by epistatic effects. Although no significant digenic marker epistasis was found when all possible marker interactions were considered, probably due to the high number of hypotheses to test, we did identify 38 epistatic interactions using the more targeted approach of testing only the previously detected QTL against the genetic background. With respect to epistasis and effect confirmation, four possible scenarios can be distinguished.

(i) An $mQTL_{RIL}$ is not confirmed by an $mQTL_{IL}$ and shows significant epistatic interactions. This is the case for many of the stronger epistatic effects (e.g.

tyrosine 4/42 × 5/74, R^2 = 4.92 %; glycerate 2/61 × 3/67, R^2 = 4.43 %), and is consistent with the assumption that loss of the epistatic interaction prevented its identification in ILs.

(ii) An $mQTL_{RIL}$ is confirmed by an $mQTL_{IL}$ and lacks epistasis (as observed for threonic acid 4/63 and urea 1/76). This suggests that such mQTL act as single-effect loci independently of other genetic factors. The existence of examples for these two cases supports the hypothesis that epistasis is a major cause of differences between sets of QTL_{RIL} and QTL_{IL}.

(iii) $mQTL_{RIL}$ that are confirmed by $mQTL_{IL}$ but show significant epistasis. This was true for three of four raffinose QTL. However, here the variance explained by the epistatic effects is low (R^2 = 1.5–2.6 %) especially with respect to the variance explained by the QTL *per se* (R^2 = 4.7–6.4 %). Furthermore, epistatic interactions between a C24 and Col-0 allele will be retained in ILs and hence contribute to effects detectable in both analyses.

(iv) An $mQTL_{RIL}$ is not confirmed by an $mQTL_{IL}$ but no epistasis can be detected, as exemplified by the five serine mQTL. Here, a more complex situation such as multi-way interactions between several loci, which escape detection in the epistasis analysis, can be assumed.

Very complex epistatic interactions may strongly interfere with QTL detection in RIL populations. Depending on the specific allele combination of the interacting loci necessary to elicit a strong effect, the uniformity of the genetic background in ILs may be an advantage, allowing identification of QTL that are not detectable in a RIL population. In addition, opposing-effect QTL present in close vicinity to each other may also interfere with QTL detection in RILs. Such arrangements have been shown to exist by Kroymann and Mitchell-Olds (2005), and may by chance be broken up upon creation of a particular IL through recombination between the QTL linked in repulsion. Events such as this will result in QTL detection in ILs but not in RILs, with the IL approach being favored by the fact that more replication can be afforded in IL analyses (due to a more limited number of individual lines), with a concomitant increase in the precision of the trait expression measurement. Thus, while QTL detected in both analyses may be preferred for follow-up studies, QTL detected in only one of the two populations should not be generally dismissed, and the two approaches may be considered complementary. Taking into account the amount of work that we invested in the generation and evaluation of both populations, they

yielded a comparable level of information. However, if it is not the genetic architecture of a trait that is of major interest but its modification in a more applied approach, ILs would be favored for use because advantageous genome segments may be identified in a genetic background close to an elite variety, und thus may be integrated into a breeding program more quickly.

4.4.2 Number and contribution of biomass and mQTL compared with other studies

As described in the Results section, the variance of both the RIL and IL populations allows identification of QTL for shoot biomass. Six biomass QTL explaining 18 % of the total variation, with individual contributions varying from 1.5 to 6 %, were identified in the RIL population, and nine biomass QTL were observed in the IL population. Five of the biomass QTL were detected in both populations. These results for biomass are similar to those for other Arabidopsis RIL populations used to detect QTL for aerial/shoot mass, with up to eight QTL detected (El-Lithy *et al.*, 2004; Loudet *et al.*, 2003; Rauh *et al.*, 2002; Ungerer and Rieseberg, 2003). Both RIL QTL on chromosome 3 and two effects in ILs on chromosomes 2 and 4 overlap with results obtained by Loudet *et al.(2003)* (3/49, 2/72 and 5/62) and El-Lithy *et al.* (2004) (3/13).

In a comparable study of biomass at an early developmental stage in Aegilops tauschii (ter Steege *et al.*, 2005), only two putative QTL were detected. In seedling-stage maize, three QTL for shoot dry weight, each explaining 11–15 % of phenotypic variance, were detected in a F2:F3 population of 226 families (Jompuk *et al.*, 2005). Further biomass QTL analyses, e.g. of poplar (Wullschleger *et al.*, 2005), rice (Hittalmani *et al.*, 2002; Li *et al.*, 2006) and Miscanthus sinensis (Atienza *et al.*, 2003), each revealed a limited number of QTL, usually with a restricted fraction of the phenotypic variance explained. Even in a very large QTL mapping experiment in maize (Schön *et al.*, 2004), in which more than 30 growth-related QTL were identified, only about 50 % of the genetic variance was explained. The effects of individual QTL on the phenotypic variance were generally small. Thus, the individual contribution of shoot biomass QTL in the Arabidopsis RIL/ IL populations analyzed here is very similar to the situation described for other species, including crops such as maize. The observed heritability (0.71) of the biomass trait in the analyzed RIL population and the rather limited fraction of the genetic variation explained jointly or individually by the detected QTL indicate that biomass accumulation is probably

affected by a very large number of small-effect QTL. This is consistent with the conclusions of Kroymann and Mitchell-Olds (2005).

In the RIL population, a total of 157 metabolic QTL were identified, with at least one metabolic QTL for approximately half of the compounds analyzed (181 compounds were analyzed, and at least one QTL was identified for 84 compounds). The contribution of individual QTL to the total phenotypic variance ranged from 1.7 % to more than 50 %. Analysis of the IL population resulted in numerous QTL, the specific numbers ranging from 177 for $P \leq 0.001$ to 2511 for $P \leq 0.05$. These numbers are in the same range as described in two previous reports on identification of metabolic QTL using RIL and IL populations. Schauer *et al.* (2006) identified 889 mQTL in a tomato population of 76 ILs, monitoring 74 metabolites at a significance level of 0.05. Using 2129 mass signals (with an unknown number of underlying chemical compounds), Keurentjes *et al.* (2006) identified 4213 metabolic QTL at a P-value threshold of 0.0001 in an Arabidopsis Cvi×Ler RIL population. In a recent QTL study using a Bay-0×Sha Arabidopsis RIL population, Calenge *et al.* (2006) detected a total of 39 QTL for starch, glucose, fructose and sucrose contents at 14 distinct loci, which co-localize with QTL for other physiological traits.

The findings that two biomass QTL co-locate with significantly more mQTL than expected from a random distribution, and, furthermore, that some of the biomass QTL can be simulated by QTL mapping of a certain linear combination of metabolite levels, fit into the emerging picture that metabolic composition is related to growth/biomass accumulation, as also shown previously (Meyer *et al.*, 2007a). While some metabolites such as ethanolamine, raffinose and tyrosine, which contributed strongly to the metabolic signature identified in the previous work, also show co-localized QTL with biomass, others do not. However, this finding is not unexpected considering the small amount of variation that is explained by any individual metabolite or metabolic QTL.

4.4.3 Derivation of metabolites sharing mQTL from either the same or widely divergent pathways

A number of mQTL are shared between metabolites. Two principal classes can be distinguished:

(i) Metabolites sharing a QTL are derived from the same biochemical pathway or from related pathways, as observed for ornithine/ proline (position 4/66;

common pathway: proline biosynthesis). This identifies the shared QTL as a candidate for a pathway QTL, which could be either a gene controlling the formation of a rate-limiting precursor or a higher-hierarchy controller of the entire pathway such as a transcription factor.

(ii) In other cases (e.g. position 3/14), metabolites with common QTL are derived from widely divergent pathways, which could be due to a major controller of several pathways or a small molecule produced in one pathway and controlling the other pathway. However, it should be kept in mind that, at present, the limited genetic resolution does not allow exclusion of the much more trivial possibility of the shared genomic regions actually being composed of several linked genes with enzymatic functions in different pathways.

4.4.4 mQTL cover both biosynthetic and regulatory genes

A comprehensive overview of all mQTL observed in the RIL population for known metabolites, including their effect, confirmation in the IL population and the chromosomal localization of all associated genes is shown in Figure S1.

The observation that a pathway-associated gene could be localized in the mQTL region for 24–67 % of all metabolic QTL can be exploited in a number of ways. One exciting possibility is the use of this dataset as a source for identifying novel functionally relevant polymorphisms in the genome by comparative sequencing of both alleles of candidate genes. In agreement with this, comparison with recently published data on SNPs between C24 and Col-0 (Clark *et al.*, 2007) showed that, for 10 out of 20 mQTL, at least one of the direct candidate genes contains a polymorphism in the protein coding region leading to an amino acid exchange. Obviously this is only a first indication, and does not prove that this amino acid exchange is responsible for the mQTL. Furthermore, it should be kept in mind that we did not observe significant enrichment of pathway genes within experimentally determined mQTL (see Results). This is mainly due to the fact that, for 18 of 39 metabolites, 15 or more (up to 130) direct candidate genes are known, which appear to be uniformly distributed by visual inspection (Supplemental Figure S1). Due to the relatively large confidence intervals, random distribution of mQTL over chromosomes will, in such cases, always lead to successful candidate gene identification hampering a permutation test. If we exclude these 18 metabolites, a *P*-value of 0.08 is obtained in the permutation test, indicating that mQTL are possibly enriched for

pathway-related genes, and that comparative analyses of the alleles would be worthwhile.

A level of up to 67 % coverage of mQTL by biosynthetic candidate genes also implies that at least 33 % of the mQTL probably harbor genes of hitherto unknown metabolic functions (e.g. as regulators), a rather large and at first unexpectedly high fraction. Although the presence of biosynthetic genes in C24 only (and thus not in AraCyc) could explain some of these mQTL, the above conclusion is supported by a seemingly unrelated observation, i.e. the unequal distribution of mQTL over the genome. The chromosomal distribution of the total 157 mQTL differs in a statistically significant manner from a random distribution, with two significant hot spots (up to 16 QTL at the top of chromosome 4, and 12 QTL at 5/75) and other areas lacking mQTL (no mQTL detected for 38 marker positions). There are two possible explanations for this uneven distribution: either it is a reflection of the uneven distribution of biosynthetic genes in the Arabidopsis genome, or a larger proportion of mQTL detected do not correspond to genes with known metabolic function (mostly enzyme genes) but represent regulatory genes of a higher hierarchical order that thus control more than one metabolite. To distinguish these two possibilities, we compared the distribution of metabolic genes in the genome with the mQTL distribution. The results of this analysis showed that the clustering of mQTL does not correlate significantly with the distribution of metabolic genes over the Arabidopsis genome, irrespective whether all metabolic genes or only metabolic genes from biosynthetic pathways covered in our analysis are taken into account (data not shown). This suggests that the uneven distribution of mQTL is due to the second explanation, i.e. a large proportion of mQTL detected identify hitherto unknown metabolic functions, most likely regulatory genes controlling primary metabolism and thus probably having a strong influence on biomass formation. The available ILs that confirmed such mQTL enable positional cloning of the corresponding novel metabolic function genes.

4.5 Material and Methods

4.5.1 Creation of recombinant inbred (RIL) and introgression line (IL) populations

Two reciprocal sets of RILs were developed from a cross between the two *A. thaliana* accessions C24 and Col-0. F_2 plants were propagated by controlled self-pollination using the single-seed descent method to the F_8 generation, at which stage

genotyping and bulk amplification was performed. The mapping population consisted of 228 Col-0×C24 F_8 and 201 C24×Col-0 F_8 individual lines. The RIL population was genotyped using a set of 110 framework SNP markers (Törjék *et al.*, 2003) as described previously (Törjék *et al.*, 2006). Marker distributions per chromosome in the two sub-populations were compared using Mantel tests (1000 permutations) of the corresponding similarity matrices obtained by simple matching, using the statistical software package Genstat for Windows version 6.1 (Payne *et al.*, 2002).

As a base population for IL development, two sets of reciprocal BC_3 F_1 lines were created from the F_2 of a reciprocal cross between the two *A. thaliana* accessions C24 and Col-0, through three cycles of backcrossing followed by one cycle of selfing using the single-seed descent method (Törjék *et al.*, 2008). The BC_3 F_1 lines were genotyped using the same set of 110 framework markers (Törjék *et al.*, 2003). Lines with positive-effect segments were subjected to further cycles of backcrossing and selfing to produce substitutions in both the Col-0 and C24 genomic backgrounds using marker-assisted selection. The average introgression lengths are 17.3 and 19.3 cM in ILs with Col-0 and C24 backgrounds, respectively.

4.5.2 Plant cultivation

The RIL population was cultivated in at least three experiments using a split-plot design. The growth room was declared as the whole plot with two factors (chamber 1 and chamber 2). Each sub-plot contained the entire RIL population and the controls (C24, Col-0, C24×Col F_1, Col×C24 F_1). Plants were grown in a 1:1 mixture of GS 90 soil (Gebrüder Patzer, Sinntal-Jossa, Germany) and vermiculite (Deutsche Vermiculite Dämmstoff-GmbH; http://www.vermiculite.de) in 96-well trays. Six plants of the same line were grown per well. Seeds were germinated in a growth chamber at 6°C for 2 days before transfer to a long-day regime [16 h fluorescent light (120 µmol m^{-2} sec^{-1}) at 20 °C and 60 % relative humidity/ 8 h dark at 18 °C and 75 % relative humidity]. To avoid position effects, trays were rotated around the growth chamber every 2 days.

ILs were selected to cover the QTL regions determined in the RIL experiment (26 ILs for the six biomass QTL, 16 for other traits), and plants were grown in two blocks with 12 sub-plots each. Each subplot contained 42 ILs, 42 test crosses (IL TCs) to the recurrent parent, and the controls twice (C24, Col-0, C24×Col, Col×C24). The position within the sub-plot was random. In addition, ‘unselected’ ILs without IL TCs

were grown in the same experiment. In this case, each sub-plot consisted of 56 ILs and 36 controls. Growing conditions were identical to those used for the RILs.

4.5.3 Shoot dry biomass

Shoot dry biomass was determined 15 days after sowing. Plants from the same well were harvested together and placed in a vacuum oven at 80 °C for 48 h. Dry biomass was measured using an analysis balance. Mean shoot dry biomass (mg per plant) was estimated using a linear mixed model (Piepho *et al.*, 2003) as described by Meyer *et al.* (2007a). Biomass in the two RIL sub-populations was compared by a Kolmogorov–Smirnov test using Genstat for Windows version 6.1 (Payne *et al.*, 2002).

4.5.4 Metabolite data

Sample preparation, measurement and data processing. Samples for the analysis of metabolic composition were collected together with the material for dry biomass analysis at 15 days after sowing. Harvested material (shoot and leaf) was frozen at -80 °C immediately, and kept at this temperature until further processing. Between two and six plants were pooled per sample. One replicate for each RIL and six replicates for each IL were measured. Extraction, derivatization, GC–MS analysis and data processing were performed as described previously (Lisec *et al.*, 2006). A targeted metabolomics approach was used, based on a reference library containing 181 compounds.

The resulting data consist of intensity values for each referenced compound and measurement, respectively. These raw data were normalized (see below) before QTL analysis.

Normalization. All samples were measured in groups of 30–50, equivalent to one measurement day. The huge number of samples led to measurement periods of several weeks per experiment. It is therefore necessary to correct for variation in detector sensitivity over this time, which otherwise causes artificial differences in absolute intensity depending on the measurement day. The samples for the two experiments (RILs and ILs) were measured using different set-ups (see below), and were therefore normalized using different strategies.

For the RIL experiment, all samples were measured over a measurement period of 26 days. Samples from different genetic backgrounds were distributed in equal proportions per day and otherwise completely randomized. Hence we assumed that

the genetic and phenotypic variance covered by all samples of a set (approximately 40) is comparable between days. Therefore, metabolite data were normalized by dividing the intensity of the metabolite *i* by the median of all measurements of *i* per measurement day.

IL samples were measured in groups consisting of genotypes related to either of the parents [C24 and M lines (C24 with Col-0 introgressions) or Col-0 and N lines (Col-0 with C24 introgressions)] in an attempt to reduce the technical error for our comparisons of interest (parent versus corresponding ILs). Six replicates per genotype were measured in total on 6 days, always together with the same set of genotypes including four to six replicates of the respective parent. Samples were randomized within days.

Two normalization steps were applied to IL samples. To account for intensity differences, we normalized each metabolite profile (sample) by its mean trimmed 20 % (the vector sum between the 10th and 90th percentiles, $k = 0.1\,n$ [n = metabolite number]).

$$x_i' = \frac{1}{n-2k}\frac{x_i}{\sum_{i=k+1}^{n-k} x_{(i)}}$$

This was determined to be more robust than using an internal standard (data not shown). However, it assumes that the same amount of material is applied to the column for each sample, and that the variation in total peak area of the analyzed metabolic subset is low. As only a low correlation ($r < 0.05$) between trimmed mean and biomass was observed for the IL data, this seems to be a fair assumption. In a second step, we improved normality by dividing each metabolite intensity value by the median of all values for this metabolite *i* from the same measurement set *j* and applying the logarithm:

$$x_{ij}'' = \log_2\left(\frac{x_{ij}'}{median(x_j)}\right)$$

Candidate gene identification. To identify possible candidate genes for mQTL, the AraCyc 3.5 database was downloaded from TAIR (Arabidopsis Information Resource, http://www.arabidopsis.org). For each mQTL, a search window was determined according to the presence of markers within its 1-LOD support interval (Supplemental Table S4). The resulting AGI codes were tested for either direct

association with the metabolite or association with one of the pathways in which the metabolite is involved.

To compare the distribution of metabolic genes over the Arabidopsis genome against the mQTL distribution, we counted all genes around each marker, i.e. the interval

$$\left\{ x \mid M_k - \frac{M_k - M_{k-1}}{2} \le x \le M_k + \frac{M_{k+1} - M_k}{2} \right\}$$

where M_k is the position of marker k (in bp). This approach was followed for the complete AraCyc data set and for a selection containing only information on pathways in which metabolites measured in this study are present. A Pearson correlation value was calculated between the mQTL distribution and both gene distributions separately.

For permutation tests in the candidate gene approach, all QTL of a single metabolite were combined and randomly distributed over the five chromosomes 10,000 times. The total number of overlapping candidate genes was recorded in each permutation, and the final distribution of these values was compared against the outcome for the actual data.

Estimation of heritability. Broad-sense heritability (H^2) is defined as the part of phenotypic variation that is explained by the genotype. We used a similar approach to that of Keurentjes *et al.* (2007a), and estimated within-line variance (V_P) based on replicate measurements of both parents and the reciprocal F_1 hybrids (10 replicates each). To account for various measurement levels, we normalized the calculated variance (s^2) of each genotype (G_k) using the squared mean before averaging:

$$V_P = \frac{1}{n} \sum_{k=1}^{n} \frac{s^2(G_k)}{\overline{G_k}^2} \qquad G = \{C24,\ Col-0,\ C24\ x\ Col-0,\ Col-0\ x\ C24\}$$

After an equivalent transformation of the RIL values, we calculated broad-sense heritability as:

$$H^2 = \frac{V_{RIL} - V_P}{V_{RIL}}$$

To prevent over-estimation, we removed outliers more than three standard deviations away from the mean. In the case of negative values, we assumed the heritability to be zero.

Narrow-sense heritability (h^2) was estimated by parent–offspring regression according the method described by Falconer and Mackay (1996). Here we made use

of available RIL parent test crosses, which were measured together with the RIL samples.

4.5.5 QTL analyses

Recombinant inbred lines. For QTL analyses, a map containing 105 markers was used, on which only one representative (with fewest missing values) of very tightly linked markers was integrated. Two software packages implementing different detection algorithms [PLABQTL, multiple regression (Utz and Melchinger, 1996); QTL Cartographer, maximum-likelihood methods (Basten *et al.*, 1994)] were combined to obtain robust QTL estimates. Composite interval mapping (CIM) was performed on an RIL population of 429 lines (dry biomass) or 369 lines (metabolites) with 1 cM increments. Co-factors were automatically selected by forward stepwise regression. Significant LOD thresholds were determined using 5,000 permutations. QTL were regarded as significant if they were detected using $LOD_{0.05}$ in one package, and reached at least $LOD_{0.25}$ in the other. QTL location and partial R^2 were further validated using 1000 runs of the fivefold cross-validation procedure implemented in PLABQTL. Given a population size of 429 and a significance level of 0.05, it can be shown (Hackett, 2002) that 99 % of all QTL that contribute more than 5 % to the total variance and more than 50 % of those that contribute more than 1 % will be detected. Most of the undetected QTL will be below the 1 % line. To identify 50 % of QTL that have a contribution of 0.5 %, we would have to double our population size.

Co-localizations of QTL from different traits are expected given the high number of traits and the limited number of markers. The deviation from the random number of co-localizations was calculated as follows. The QTL of each metabolite were randomly distributed over the 105 marker positions. We then counted the number of co-localizations with each of the dry biomass QTL or with other metabolite QTL. This procedure was repeated 1,000 times, yielding a distribution of the maximum numbers of co-localizations. The 95 % quantile of the distribution for metabolite–biomass QTL co-localization was eight, hence eight or more QTL at one genome position are regarded as significantly co-localized. The corresponding 95 % quantile for the metabolite–metabolite QTL co-localization was ten.

Introgression lines. To identify metabolites with a significantly altered intensity in a certain genotype, we compared metabolite values of ILs (six replicates, *IL*) against all

parental line samples measured within the same set (approximately 30 replicates, *P*). To estimate a *P*-value empirically, we compared the true mean difference $x = \left|\overline{x_{IL}} - \overline{x_P}\right|$ in *k* permutations (*k* = 10,000) with the calculated difference $y_k = \left|\overline{x_{IL,k}} - \overline{x_{P,k}}\right|$, where $\overline{x_{IL,k}}$ is the mean of a sample (of size six) drawn from the set union of *IL* and *P* and $\overline{x_{P,k}}$ is the mean of the remaining values of this set union:

$$p_{est} = \frac{1}{n}\sum_{k=1}^{n} F(x, y_k) \qquad F(x,y) = \begin{cases} 0 \; for \; x \le y_k \\ 1 \; else \end{cases}$$

Hence, if the measured mean difference was higher than the mean differences calculated for 9,500 of the 10,000 permutations, we obtain a *P*-value estimate of 0.05.

4.5.6 Epistasis

The software package PLABQTL (version 1.2BIC was used to estimate epistatic interactions. In an initial screening for digenic epistatic effects by two-way anova between all pairs of marker loci, no significant effects were determined using the integrated scanning function.

In the following analysis, every mQTL of a known metabolite was tested for additive × additive effects against the genetic background at intervals of 2 cM. A range of 10 cM around the actual QTL position was blocked during this analysis. The resulting likelihood profiles for all mQTL of a metabolite were overlaid and inspected visually for maximum LOD estimates. A full model for each metabolite containing all *per se* QTL and their putative epistatic interactions was set up (one epistatic interaction per mQTL was usually included, none if the maximum effect coincided with another *per se* QTL, and two if equivalent interactions were present). From this full model, non-significant effects were omitted in a backward elimination step using a Bayesian information criterion (Kusterer *et al.*, 2007) before re-estimating all remaining parameters simultaneously.

5 Heterotic metabolic QTL analyses of *Arabidopsis thaliana* RIL and IL populations

Jan Lisec[1,§], Rhonda C. Meyer[2], Matthias Steinfath[3], Joachim Selbig[1,3], Thomas Altmann[1,2] and Lothar Willmitzer[1]

[1] Max-Planck-Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476 Golm, Germany

[2,3] Departments of [2]Genetics and [3]Bioinformatics, Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Strasse 24–25, 14476 Potsdam, Germany

[§] JL contributed to method development, data analysis and preparation of the manuscript

(The manuscript is currently in preparation for submission to Genetics.)

5.1 Abstract

Two mapping populations of a cross between the *Arabidopsis thaliana* accessions Col-0 and C24 were cultivated and analyzed for the level of 181 metabolites to elucidate the biological phenomenon of heterosis on a metabolic level. Initially the metabolite profiles of 369 Recombinant Inbred Lines (RILs) and their test cross progeny with both parents allowed us to determine the position and effect of 147 Quantitative Trait Loci (QTL) for metabolite Absolute Mid Parent Heterosis (AMPH). Furthermore we found 153 and 83 QTL for augmented additive and dominance effects respectively.

In a second experiment we investigated the potential overlap of these QTL with significant effects determined for 41 Introgression Lines (ILs) and their test crosses with the respective parent. A confirmation rate of 23.3 % was reached.

These findings as well as a candidate gene search, mode of inheritance analyses for IL-QTL, average degree of dominance estimations for RIL-QTL, the comparison with results for corresponding biomass data recorded and an attempt to predict biomass heterosis of a hybrid based on the metabolic profile of its parents are discussed within this manuscript.

5.2 Introduction

Despite being known, researched and applied for a hundred years, the often observed advantage of heterozygous offspring over its homozygous parents (heterosis) is still not well understood with respect to the underlying molecular mechanisms. As its advantageous effects are lost during the inverse process of inbreeding, it is well accepted that hybrid vigor is based, to some extent, on the combined action of heterozygous alleles. Three main theories which seek to explain the phenomenon on the genetic level have been developed in the past and

experimental evidence for all of them was published in literature (see Lippman and Zamir (2007) for a review).

The Dominance hypothesis coined by Davenport as early as 1908 assumes a complementation mechanism where for each gene the more favorable allele inherited from either of the homozygous parents will be dominant within the hybrid, thus masking deleterious effect alleles. However, an exclusively dominant mechanism should in principal allow the complete fixation of beneficial effects from both parents into a homozygous offspring. As this is partly contradictory with observed results (Birchler *et al.*, 2003; Duvick, 1999) overdominance was suggested as a mechanism to explain heterosis (Hull, 1945). Overdominance postulates the superiority of the heterozygous state due to allelic interactions which can not occur in either homozygous state. Thus, a number of loci inherited as heterotic Mendelian factors are thought to cause the observed phenotypes.

The Epistasis theory (Powers, 1944) suggests that the highly increased number of possible epistatic interactions within a hybrid contributes mainly to hybrid vigor.

Maize, tomato, rice and other crop plants are intensively studied for heterotic effects in plant science due to their agronomic importance and the sometimes exceptionally high levels of heterosis reached for yield related traits. On the other hand, Arabidopsis is well suited as a model organism for quantitative genetics and development because of a fully annotated genome (Arabidopsis Genome Initiative, 2000), a high natural diversity (Alonso-Blanco and Koornneef, 2000) and short generation times allowing to perform large scale experiments under controlled environmental conditions. Mid parent heterosis levels of up to 161 % for biomass under high light conditions have been reported (Meyer *et al.*, 2004) and recent improvements of molecular marker technologies facilitated the generation of several RIL and IL populations (Alonso-Blanco *et al.*, 1998c; Keurentjes *et al.*, 2007a; Törjék *et al.*, 2008; Törjék *et al.*, 2006) covering the whole genome with a high number of lines.

With the advent of the metabolic profiling technology it became feasible to investigate the metabolome of such populations by applying a targeted analysis of several hundred metabolites covering predominantly the primary (GC-MS) or secondary (LC-MS) metabolism. Several studies have followed this approach focusing on natural variation (Keurentjes *et al.*, 2006), the connection between metabolism and yield

associated traits or biomass (Meyer *et al.*, 2007b; Schauer *et al.*, 2006) and the identification of metabolic quantitative trait loci (mQTL) (Lisec *et al.*, 2008).

The ability to map QTL contributing to heterosis has been shown for individual phenotypic traits in maize, rice and Arabidopsis (Frascaroli *et al.*, 2007; Kusterer *et al.*, 2007; Li *et al.*, 2001; Luo *et al.*, 2001; Melchinger *et al.*, 2007a; Semel *et al.*, 2006; Stuber *et al.*, 1992; Xiao *et al.*, 1995). However, to unravel the genetic basis of heterosis the more detailed investigation of its components contributing to the observed phenotypes will become necessary. A step in this direction could be multi-parallel analyses.

A few studies embarked on tackling this problem on a genome wide scale by analyzing expression levels but did not yield consistent results so far (Auger *et al.*, 2005; Song and Messing, 2003; Swanson-Wagner *et al.*, 2006). Only very recently Schauer *et al.* (2008) published the first QTL analysis on a metabolic level for a tomato IL population examining the mode of inheritance for all effects.

While such multi-parallel studies provide excellent insight into general trends and are a rich source for further research, the desire to functionally annotate the causal genes requires additional fine-mapping approaches. A few QTL were successfully examined in this sense to date (Fridman *et al.*, 2004; Konishi *et al.*, 2006; Steinmetz *et al.*, 2002), however, no individual gene involved in heterosis has hitherto been identified and characterized at the molecular level in plants (Hochholdinger and Hoecker, 2007).

In this study we aimed to (i) identify and characterize heterotic metabolic QTL (hmQTL) in two mapping populations of 369 RILs and 41 ILs and (ii) searched for candidate genes being present within the QTL support intervals. We wanted to (iii) compare results between both populations and with our previous assessment of mQTL which are unrelated to heterosis. (iv) We tried to gain insight into the mode of gene action with respect to heterosis on the level of the primary metabolism and (v) searched for a link between the metabolic profiles of two homozygous plants and the amount of heterosis fixed in their hybrid expressed by the integrative trait biomass.

5.3 Results

5.3.1 Heterotic metabolic effects between the two parental genotypes

To test to which extent biomass heterosis is reflected on a metabolic level we took advantage of the large number of replicates of the two parental accessions (C24,

Col-0) and their offspring (C24×Col-0, Col-0×C24) measured during the IL experiment. The medians of about 50 replicates per genotype were used to calculate the amount of mid-parent heterosis (MPH) for each of the 181 analyzed metabolites. As shown in Figure 15a, most metabolites are expressed in the offspring on levels close to the expected parental mean. Two thirds show less than 10 % heterosis and only a few (24 out of 181) show heterosis values above 20 %. Maximum values (> 50 %) were determined in only three cases, all of which concern low abundant metabolites which are more likely to give rise to outliers despite the high number of replicates and were therefore discarded from the plot.

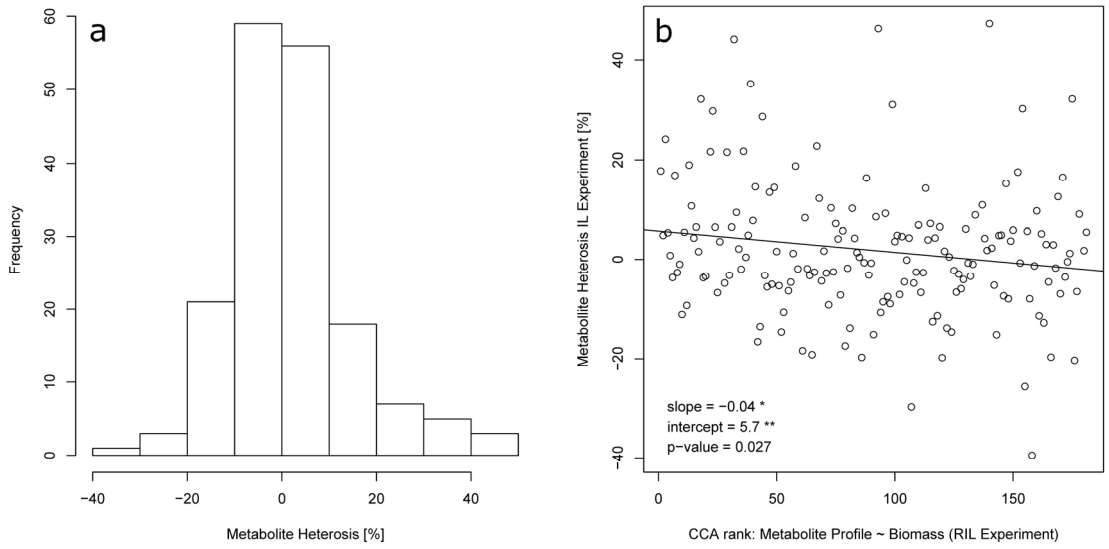


Figure 15 Histogram of metabolite heterosis values between accessions C24 and Col calculated based on parental and hybrid samples from the IL experiment (a). Regression of metabolite heterosis values on the rank of each metabolite in a canonical correlation analysis linking metabolite profiles to biomass determined in the RIL experiment (b).

Positive and negative effects are balanced; no dependencies on chemical classes – being predominantly positive or negative – could be verified. However, there is a significant correlation between the MPH value of a metabolite and its rank in a canonical correlation between metabolite profiles and biomass as published in a previous manuscript (Figure 15b).

5.3.2 Heterotic metabolic QTL (hmQTL)

Recombinant Inbred Lines: Using the procedure described in the Methods section of this manuscript we could determine 385 metabolic QTL (mQTL) for 136 out of 181 metabolites at a 5 % threshold level (Supplemental Table 7). On average, these QTL explain 5 % of the phenotypic variation and their support intervals (1-LOD) are

approximately 10 cM wide. Using the methods suggested by Melchinger *et al.* (2007a) we partitioned the observed variation into average mid-parent heterotic effects for Col-0 ($AMPH_{Col}$, 63 QTL in total) and C24 ($AMPH_{C24}$, 86), augmented additive effects (ADD, 153) and augmented dominance effects (DOM, 83). A QTL overview for metabolites of known chemical structure is given in Figure 16. Maximum values of eleven and eight heterotic effects for one unknown metabolite (#074) and cellobiose, respectively, have to be treated with caution as these metabolites whilst showing very high heterosis values per se (173 % and 214 %) are generally low abundant.

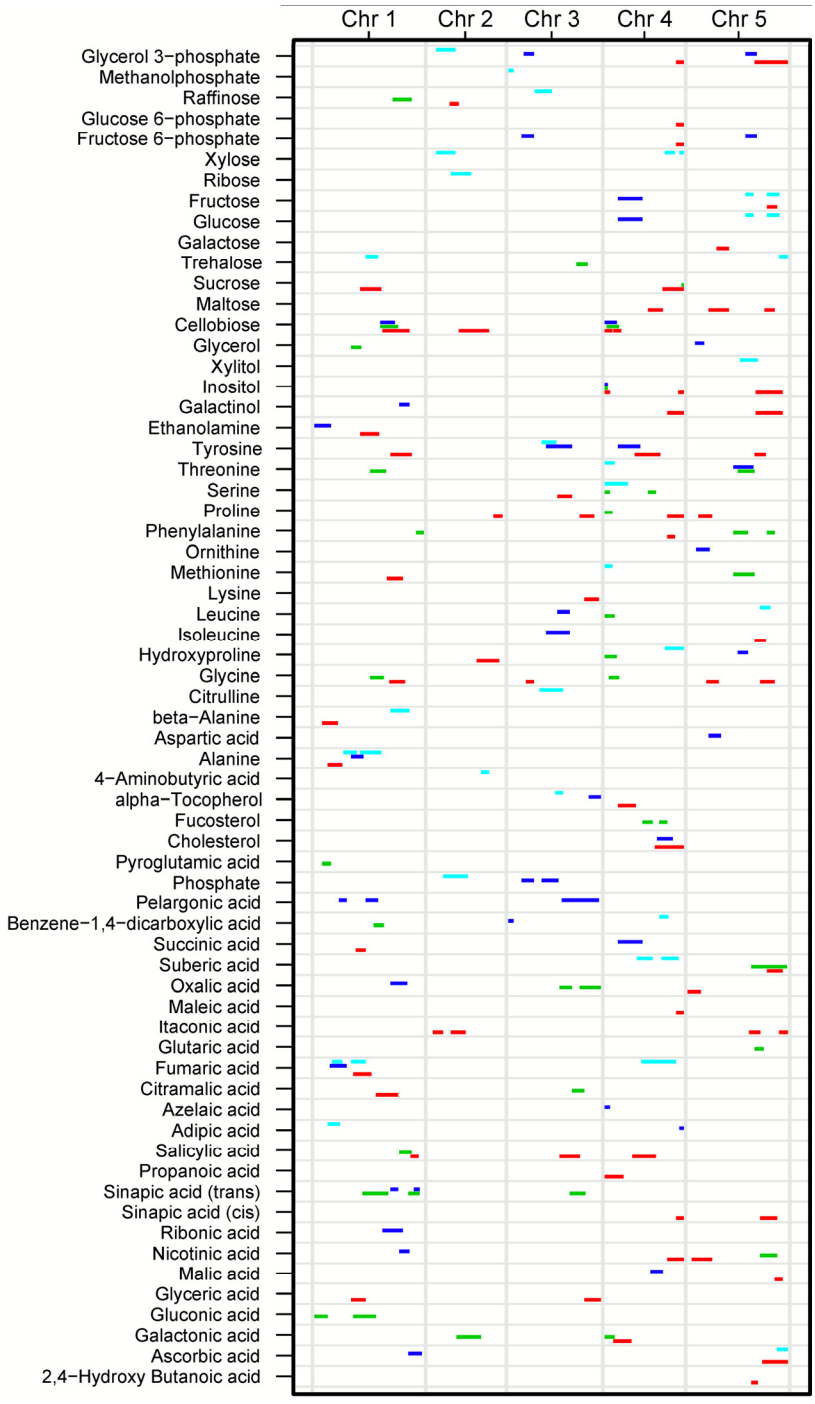


Figure 16 QTL for known metabolites. Different colors indicate QTL for absolute MPH (dark and light blue), augmented additive (red) or dominance (green) effects. Metabolites are ordered approximately to chemical function.

The distribution of these segments bearing heterotic effects follows largely the distribution of metabolic QTL (mQTL_RIL) as shown in Figure 17. A permutation test reveals three significant ($P < 0.05$) 'hot spots' at 1/75 (Chromosome 1, Position 75 cM), 4/3 and 5/74, two of which are co-located with the hot spots previously determined. This is reflected as well by the fact that 36 % of the previously identified mQTL do overlap with at least one (and up to four) heterotic QTL for the same metabolite.

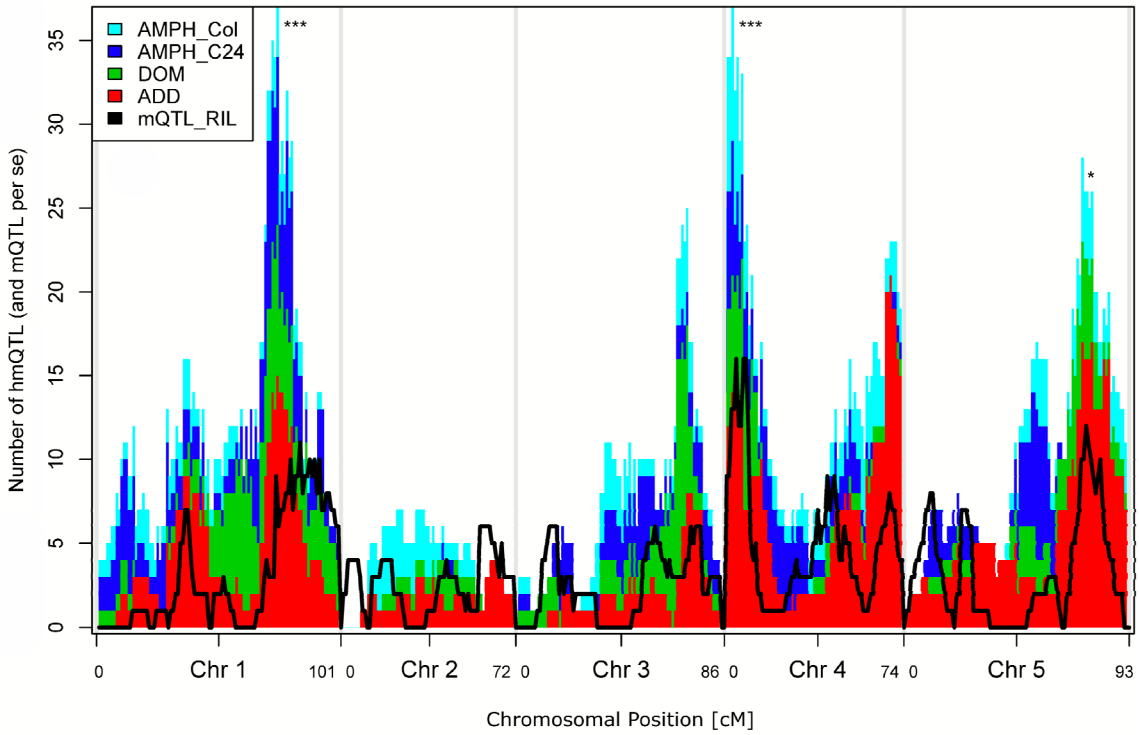


Figure 17 Distribution of heterotic metabolic QTL (hmQTL) for absolute mid parent heterosis (light and dark blue), augmented additive (ADD) and augmented dominance (DOM) effects. Metabolic QTL determined based on RILs (black line) are shown for comparison. Asterisks indicate QTL hotspots which were determined to be significant (* $P < 0.001$; * $P < 0.05$) using permutations.**

Introgression Lines: To confirm and potentially fine map heterotic metabolic effects determined within a RIL population, we conducted an independent experiment using an IL population which covered ~73 % of the parental genome. Depending on the significance threshold used up to 23.3 % of the RIL effects could be confirmed (Table 3). This confirmation rate is approximately half of the confirmation which was achieved for mQTL on a comparable significance level.

Mode of Inheritance: The mode of inheritance for effects in ILs was estimated according to Semel *et al.* (2006). To be able to detect general trends we applied the

decision tree to classify the QTL on all effects which were determined using a less stringent significance threshold ($P \leq 0.05$).

P	Number of significant Changes	FDR [%]	Number of confirmed RIL QTL	confirmed RIL QTL [%]	average R2 of confirmed RIL QTL [%]	confirmed allelic effect	confirmed allelic effect [%]
0.001	357	4.16	14	4.59	5.72	10	71
0.01	1071	13.86	49	16.07	5.34	32	65
0.05	2155	34.44	71	23.28	5.07	41	58
0.1	2966	50.04	82	26.89	4.91	50	61

Table 3 Number of IL effects and onfirmation rate of RIL QTL at different significance levels.

The majority of QTL fall into the dominant (67 %) and additive (19 %) bin with only a minor proportion of QTL exhibiting over-dominant (9 %) or recessive (5 %) gene action (Figure 18). Interestingly the results are biased in two ways:

(i) If a larger number of ILs is detected to be significantly different to the respective parent, these effects in most of the cases tend to change the metabolite level in the same direction, either increasing or decreasing it. This is surprising as theoretically we would expect a distribution of IL metabolite values around its parent mean and hence more balanced effects. We hypothesize that metabolites which show very different levels for both parents initially are prone to exhibit in many introgressions significant changes in metabolite level towards the level of the parental donor. That means, the stronger two parental genotypes differ for the level of any metabolite *i*, the more likely it is to observe a shift in the distribution of the levels of *i* in ILs derived from the respective parent. To test this we calculated the IL effect bias (expressed as the absolute difference of the sum of all positive and the sum of all negative mode of inheritance annotations) and the parental difference (expressed as the absolute value of the log2-ratio of the parental mean values) for each metabolite. A regression of the parent difference on the IL effect bias supports our hypothesis. (Supplemental Figure S2).

(ii) If analyzed separately, ILs carrying a C24 introgression in the Col-0 background predominantly reveal metabolite increasing effects (Supplemental Figure S3). This is independent of the mode of inheritance. This finding is caused by the fact that the majority of the 181 metabolites analyzed in this study have higher values in C24 than in Col-0. Consequently, IL values for N-Lines (Col-0 background) show predominantly increased metabolite values if compared to Col-0.

With respect to the different chemical classes it is noteworthy that most of the amino acids are relatively stable with only a few ILs changing their level significantly. In contrast, many organic acids and sugars are changed in a high number of ILs. Organic acids are found to be predominantly increased if compared to the respective parent level while most sugars are decreased.

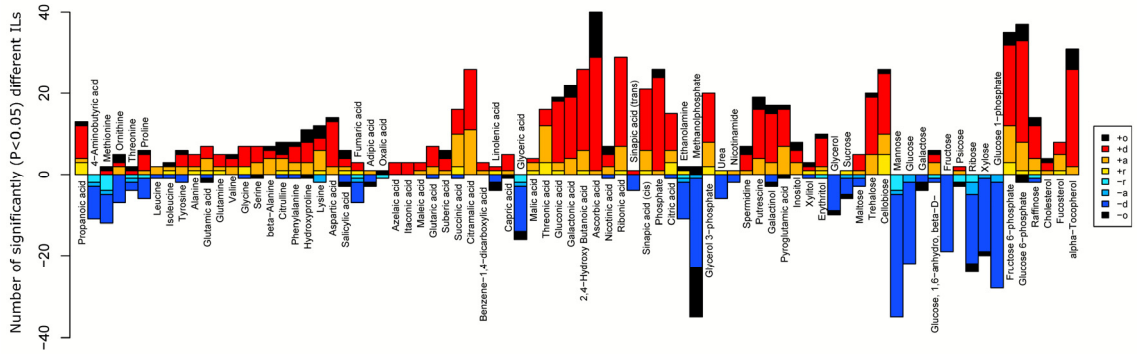


Figure 18 Mode of inheritance distribution for known metabolites. Colors encode for overdominant (black), dominant (red), additive (orange) and recessive (yellow) gene action. Positive effects represent an increased metabolite level in the IL or IL-hybrid compared to the respective parent. Negative effects represent an equivalent decrease. The size of the bars indicate the number of ILs being significantly different.

Per se heterosis in relation to population effects: The amount of individual metabolite heterosis between P_1 and P_2 (expressed by the levels in $F_{1\text{-}a}$ and $F_{1\text{-}b}$) should be reflected in the number and significance of individual effects calculated using RILs or ILs. To test this we grouped metabolites into bins according to the absolute value of their heterosis. We then plotted for each of the resulting groups the median number of significant effects (RILs and ILs), the minimal *P*-value determined (for ILs) and the maximum explained phenotypic variation (for RILs) (Figure 19). As expected, the number of determined significant effects is increased for metabolites showing higher per se heterosis. They tend to explain more of the phenotypic variation of the trait and give rise to more significant *P*-values. Interestingly this holds true for both experiments (RILs and ILs) although heterosis values were estimated based solely on the IL experiment data.

Candidate Genes: For all RIL QTL we tried to identify candidate genes catalyzing known chemical reactions of this metabolite or being annotated within pathways where the metabolite is involved. For 18 % and 55 % of all QTL which could be tested (that is, where at least some information was available within the public

database AraCyc4.0) we found direct and pathway candidate genes, respectively (Supplemental Figure S4). Contrary to our previous experiment (Lisec *et al.*, 2008) we could not confirm this to be a significant increase over a similar search using randomly positioned QTL (being equal in number and size to the observed data).

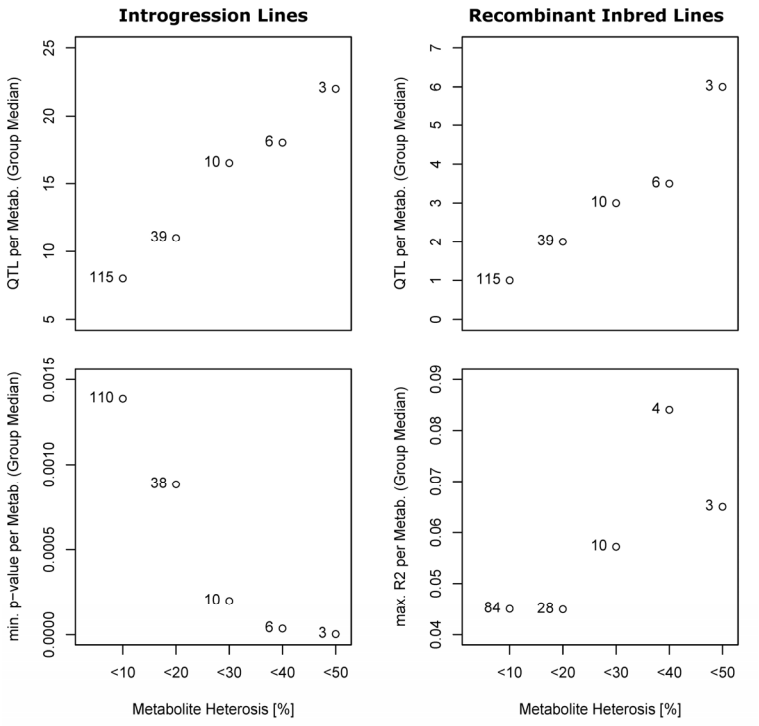


Figure 19 Parameter dependency of QTL analyses from metabolite heterosis. Metabolites were grouped according to per se heterosis calculated based on the measurements of C24, Col-0 and their hybrids (see Figure 15). For each bin the median value with respect to number of determined QTL and minimal *P* value (ILs) or number of determined QTL and highest explained phenotypic variance (RILs) is plotted. Numbers indicated the number of metabolites within each bin.

Comparison with heterotic biomass QTL: In our companion study focusing on biomass as the trait of interest we found four QTL for biomass heterosis and further seven QTL for the linear transformations Z1 and Z2, the augmented additive and dominance effects (Meyer, unpublished data).

The strongest heterotic effect was mapped on position 4/4 explaining 15.75 % of the phenotypic variation. It co-locates not only with two further effects of Z1 and Z2 for biomass but also with a cluster of 37 mQTL. All other biomass effects which explain individually less than 8 % of phenotypic variation do not co-locate with further mQTL hot spots except for one augmented dominance effect on chromosome five (5/74).

5.3.3 Average degree of dominance:

The average degree of dominance $\overline{D}$ was calculated according to Kearsey and Pooni (1996). It is a weighted mean of the level of dominance over all segregating loci and can be obtained as the ratio of the estimated variance components V_D and V_A (see Material and Methods). Surprisingly, all of our metabolites lie within a range of 1 to 1.5 thus classified exclusively as dominant (if lower than 1.2) and

overdominant (if higher than 1.2) according to Stuber and Wendel (1987). This is at least partly in line with the mode of inheritance estimations for IL effects being to a large extend classified as dominant as well.

5.3.4 Heterosis Prediction

Similar to the previously identified relation of a metabolic profile and its corresponding biomass (Meyer *et al.*, 2007a) we aimed to connect the metabolic profiles of two parental lines (C24 or Col-0 and RIL_i) and the biomass heterosis value of their progeny (TC-$C24_i$ or TC-Col_i). Hence, we formed a metabolite profile matrix dividing each RIL_i profile by either an averaged C24 or Col-0 profile and associated this matrix with the vector of biomass heterosis values of the corresponding TC_is. The correlation between the canonical variate and the observed data was 0.64. However, as a permutation test (n = 10,000) yielded on average correlation values of 0.52 we conclude that, although the relation between parental metabolic profiles and biomass heterosis of their progeny is clearly non-random, we can not meaningfully predict heterosis using this approach (Supplemental Figure S5).

5.4 Discussion

In our previous work we were able to show that the variation in the metabolite levels of both our genetic populations, RILs and ILs, is sufficient to map metabolic QTL with broad sense heritability being on average 0.40 (Lisec *et al.*, 2008). Furthermore we found heterosis for biomass (Meyer *et al.*, 2004) and a metabolic signature related to high plant growth (Meyer *et al.*, 2007a). In this study we successfully analyzed RIL and IL populations of a cross between the two *Arabidopsis thaliana* accessions Col-0 and C24 for heterotic QTL on a metabolic level. The generally modest metabolite heterosis (median value of 6.54%) was sufficient to map a high number of significant effects within both approaches.

Regarding the fact that we partitioned the observed metabolite variation into four linear transformations per metabolite, we observed a comparable number of metabolic QTL in both populations.

The moderate overlap (36%) of previously detected effects and particularly the co-localization of hot spots from both experiments hint that loci which contribute to differences between the two parental genotypes are also involved in heterosis for the respective traits. This is in agreement with the observation that mQTL which overlap with heterotic effects do not differ from non-overlapping mQTL with respect to their

independent confirmation via ILs but differ significantly ($P < 0.05$) in their phenotypic explained variance which is higher for the group of overlapping mQTL.

Candidate Gene Approach: An interesting finding is that the search for candidate genes from metabolic pathways which are annotated within the QTL support intervals did not yield a statistical significant increase of positive hits for the observed QTL data compared with a set of randomly distributed QTL of similar size. This is an important hint towards the nature of genes involved in heterosis. As pointed out by Milborrow (1998): a large increase in plant size (or biomass) in F1 hybrids will derive from much smaller percentage increases in relative growth rate or duration of growth, which can be further partitioned in even smaller differences between the components of growth. Consequently, for our analysis of metabolism it would be a surprise to map predominantly enzyme encoding genes contributing to heterotic effects. It seems more likely that the observed deviation from the parental mean is caused by differences in regulatory elements.

Along this line goes the enrichment of pathway related genes which we detected in our previous study for non-heterotic mQTL. While in some cases differences between two parental genotypes fixed in a RIL population can be caused by enzyme encoding genes this can not be confirmed for such differences representing heterotic behavior. The idea of picturing heterosis as a lessening of the tight regulation of growth (Milborrow, 1998) is also supported by the mainly negative correlations of metabolite levels and biomass for those metabolites which were highly ranked in a canonical correlation between metabolic profiles and biomass (Meyer *et al.*, 2007a).

Mode of inheritance: On the level of gene action our results for the IL population show, that only a minor fraction of the determined effects (9 %) were categorized as over-dominant. This ratio is similar to the results of Schauer *et al.* (2008) for a tomato IL population where only 5 % of all QTL were over-dominant and further in agreement with the moderate levels of metabolite heterosis determined in the F1 hybrids. A likely explanation could be that metabolites are not as closely related to reproduction as the traits usually under study in heterosis research. In that sense, Semel *et al*. (2006) found an enrichment for overdominant mode of inheritance only for traits related to reproductive fitness. Interestingly, Schauer *et al.* (2008) observed a similar bias towards increasing metabolite content, as was determined in our study for Col-0 related ILs. This bias is caused by lower levels in Col-0 compared with C24 for the majority of the metabolites. As a similar amount of sample was used for

extraction and derivatization a possible explanation could be that C24 has higher levels of primary metabolites – which we predominantly detect with our method – while the Col-0 samples contain a higher fraction of secondary metabolites. While this idea has to be proven in a further study it has no impact on the differences between parental lines and ILs observed in this work. The bias indicates the polygenic nature of metabolic traits because for strong parental differences many introgressed segments bear alleles which lead to significantly different metabolic levels in ILs and IL-hybrids.

Contradictory to this IL-QTL classification, the average degree of dominance which was estimated based on the RIL population reveals that overdominance is the major contributor of the explained variation. However, this estimation using the variance components based on the test crosses has to be treated with some caution (Melchinger *et al.*, 2007b).

Epistasis: Using comparable significance thresholds the confirmation rate of heterotic RIL-QTL in the IL population is approximately 50 % of what we found for non-heterotic effects. This could indicate that epistatic effects play an important role in heterosis.

Epistatic effects can be expected to be broken up to a large extent in ILs. While mQTL, which were found using the RIL population, seemed to be robust enough to allow their confirmation in ILs heterotic mQTL appear to be much more dependent on epistatic interactions. A possible explanation would be the prevalence of trans-regulated genes among heterotic loci. However, for biomass all heterotic QTL could be confirmed in ILs.

Biomass: Taken the results of our companion study on biomass into account we clearly determined the top of chromosome four to be of highest interest for heterosis. For biomass three out of eleven QTL are mapped to this location including the QTL bearing the strongest effect. For metabolites, not only a high number (52 mQTL from 4/0–4/15 with 37 mQTL at 4/4) of significant effects was determined, but the explained variances of these QTL are significantly higher than for the remaining effects if compared in a t-test ($P < 0.001$). We therefore currently undergo the process of sub-IL generation and hope to investigate this striking result in more detail in the near future.

Heterosis prediction: Regarding the promising results from our previous work (Meyer *et al.*, 2007a), where we could identify a metabolic signature to be strongly linked to

the integrative trait biomass we thought it an exciting possibility to predict biomass heterosis based on the metabolic profiles of two homozygous parents. This would potentially allow the use of cost and time efficient GC-MS technique to screen a wide range of homozygous elite lines and facilitate the selection of potential crosses to be validated for heterotic effects in field trials. However, our current approach is limited to a single RIL population and only a rather low improvement of our observed data if compared to permuted data sets was found (Supplemental Figure S5). The approach may be improved using a higher number of lines and different mathematical strategies. It seems clear to us that GC-MS along with the other -omics technologies due to their broad coverage of parameters promises successful application in predicting heterosis and thus breeding.

5.5 Material and Methods

5.5.1 Plant cultivation and metabolite analysis

Plant materials: All plant materials analyzed in this study were established as progeny of the two *Arabidopsis thaliana* accessions C24 and Col-0 (P_1 and P_2). The generation of the two homozygous mapping populations of (i) 422 Recombinant Inbred Lines (RILs) and (ii) 97 Introgression Lines (ILs) is described in more detail elsewhere (Törjék *et al.*, 2008; Törjék *et al.*, 2006). Both mapping populations were genotyped with a set of 110 framework SNP markers (Törjék *et al.*, 2003) as described elsewhere (Törjék *et al.*, 2006).

To allow analyses for heterotic effects backcrosses with testers P_1 and P_2 were produced for 41 ILs (20 IL-TC_{P1} and 21 IL-TC_{P2}) and 369 RILs (368 RIL-TC_{P1} and 363 RIL-TC_{P2}). The average introgression length is 19.3 and 17.3 cM in ILs with C24 and Col-0 backgrounds, respectively. ILs were chosen based on previously determined biomass QTL and cover ~73 % of the parental genome.

Experimental design: We used the North Carolina Design III as proposed by Comstock and Robinson (1952).

Plant cultivation: All plants were grown in 1:1 mixture of GS 90 soil and vermiculite in 96-well-trays under a long-day regime (16 hours fluorescent light [120 µmol m-2 s-1] at 20°C and 60% relative humidity / 8 hours dark at 18°C and 75% relative humidity). Six plants of the same line were grown per well. To avoid position effects, trays were rotated around the growth chamber every two days.

Within the first experiment all RILs and RIL-TCs were cultivated together with both parents (P_1 and P_2) and their reciprocal hybrids C24×Col-0 and Col-0×C24 ($F_{1\text{-}a}$ and $F_{1\text{-}b}$) in a split plot design. Although at least three replicates per line were grown, these replicates were pooled into one sample due to a limited number of measurements feasible. Controls (P_1, P_2, $F_{1\text{-}a}$ and $F_{1\text{-}b}$) were measured in 10-12 replicates originating from different pooled plant samples grown together with the RILs.

The second experiment contained the 41 ILs and their IL-TCs in two blocks and six subplots per block. The position within the subplot was random. For metabolomics analyses we pooled six times two subplots (three samples per block). All controls (P_1, P_2, $F_{1\text{-}a}$ and $F_{1\text{-}b}$) were grown together with the ILs but in higher replicate numbers.

Metabolite analysis: Harvested plant material was processed as described elsewhere (Lisec *et al.*, 2006) and analyzed using Gas-Chromatography Time of Flight Mass-Spectrometry (GC-ToF-MS). Within the first experiment, metabolite profiles were recorded for P_1, P_2, $F_{1\text{-}a}$, $F_{1\text{-}b}$ (10 to 12 replicates each), 369 RILs and 731 distinct RIL-TCs with P_1 or P_2 (single measurements).

Within the second experiment, metabolite profiles were recorded for P_1, P_2, $F_{1\text{-}a}$, $F_{1\text{-}b}$ (47 to 57 replicates each), 41 ILs and 41 IL-TCs with either P1 or P2 (six replicates each).

Each metabolite profile consists of 181 intensity values which represent the levels of 82 known (with respect to comparison to a reference) and further 98 unknown chemical compounds.

The data was normalized as described previously (Lisec *et al.*, 2008).

5.5.2 QTL analysis and statistical methods

Statistical analyses: Mid-parent heterosis (MPH) for metabolites was calculated based on the median values of all P_i and F_i measurements (approx. 50 samples per genotype) conducted during the IL experiment using the formula: $MPH = 100 \cdot \left(\overline{F_1} - \overline{P}\right) / \overline{P}$, where $\overline{F_1} = \left(F_{1-a} + F_{1-b}\right)/2$ and $\overline{P} = \left(P_1 + P_2\right)/2$

QTL analysis: QTL analyses were performed using the software tool QTL Cartographer (Basten et al. 1994). From the normalized metabolite data we calculated for each metabolite absolute mid parent heterosis of the RIL-TCs with C24 as $AMPH_{P_1} = TC_{P_1,i} - 0.5\left(RIL_i + \overline{P_1}\right)$ and Col-0 as $AMPH_{P_2} = TC_{P_2,i} - 0.5\left(RIL_i + \overline{P_2}\right)$.

Furthermore we used the transformations $ADD = TC_{P_1,i} + TC_{P_2,i}$ and $DOM = TC_{P_1,i} - TC_{P_2,i}$ to characterize additive and dominant effects (Frascaroli *et al.*, 2007). Composite Interval Mapping (CIM) was conducted for each of the four above mentioned variables using automatic co-factor selection by forward stepwise regression. Significant LOD thresholds (for $P < 0.05$) were determined by 1,000 permutations for each trait individually. Support intervals were calculated using the 1-LOD method.

Hotspots in the QTL distribution were computed by permutations as described in Lisec *et al.* (2008).

Candidate Gene Search: To search for candidate genes which are annotated in AraCyc 4.0 and co-locate with determined QTL, we followed exactly the procedure described in Lisec *et al.* (Lisec *et al.*, 2008) with the exception that we used the updated version of the database.

Mode of inheritance: The mode of inheritance for significant effects in ILs was determined with a decision tree as described in Semel *et al.* (2006).

Average degree of dominance: The average degree of dominance $\overline{D}$ was calculated as $\overline{D} = \sqrt{2V_D / V_A}$ according to Kearsey and Pooni (1996). The variance components V_A and V_D were estimated as the variance of ADD and DOM (see above).

Canonical correlation analysis: Canonical correlation analysis was performed using the function *cancor* built in the statistical software package R (http://www.R-project.org). By permuting the MPH vector we computed the distribution of canonical correlations for random datasets.

6 Discussion and Outlook

As each chapter has a discussion focusing on the achievement and the impact for the research contained within it, this final chapter will primarily extend the aforementioned points regarding thesis integrity and detail its findings in a broader context. In addition it will give an outlook of ongoing and future research planned using the work described herein as its foundation.

6.1 Discussion

6.1.1 Metabolomics on a large scale

Despite the fact that the first approaches in metabolomics were conducted less than ten years ago, GC-MS can be considered as a mature technique. However, when this work began, no metabolite profiling approach of comparable scale was published. Thus, we had to answer the questions: Is it feasible to perform metabolomics on a large scale and which results can be expected?

We conducted GC-MS based metabolomics by analyzing the levels of 181 metabolites in more than 2000 Arabidopsis samples. A higher number of compounds can be expected if a broader natural diversity is under research or if stress conditions are applied. For such diverse sample sets a fraction of detectable compounds are likely to occur only in a few samples requiring non-targeted analytical tools to be applied, in contrast to the reference based methods used throughout this work. Such tools are currently under development and will prove useful in the future (Fiehn *et al.*, 2005; Luedemann *et al.*, 2008; Smith *et al.*, 2006; Styczynski *et al.*, 2007; Vos *et al.*, 2007). However, a number of issues should be considered.

(i) *Quantification.* Although it is in principle possible to conduct an absolute quantification, metabolite analysis on this scale rather aims to compare compound levels in different samples. Problems which render absolute quantification difficult are the occurrence of multiple peaks per metabolite as derivatization artifacts, the fragmentation of the metabolite molecule into a mass spectrum and the huge amount of necessary standard curves which are impossible to establish for the fraction of unknown metabolites which nonetheless can be reliably measured. Those issues may be solved if a full complement of isotopically labeled metabolites of known concentrations is added to each sample (Fernie *et al.*, 2004).

(ii) *Variation.* Only changes exceeding the measurement variation can be observed. While this statement is trivial and generally true for every method it should not be overlooked in metabolomics. The technical variation is found to be ~10 to 15 % for most of the metabolites (Fiehn *et al.*, 2000a; Strelkov, 2004). However, a thousand samples in GC-MS usually require several thousand plants to be sawn, cultivated, harvested and analyzed. Hence, some environmental variation can be expected even if protocols are followed most carefully. Statistical methods and an appropriate experimental design will help to separate this additional variation from the changes which correspond to the effect under study but it limits the expectations put on metabolomics if pathways and metabolic networks are investigated.

(iii) *Comprehensiveness.* Based on the estimated number of metabolites in biological samples metabolomics is still far from accessing the full metabolome and the annotation of detected peaks heavily depends on available reference databases. Improvements in available tools and the application of promising techniques like two dimensional gas chromatography (GC×GC-MS) and FT-ICR-MS (Aharoni *et al.*, 2002; Blumberg *et al.*, 2008; Hirai *et al.*, 2004b) will enlarge the fraction of small molecules which are possible to investigate simultaneously in the future.

Apart from the issues mentioned above several robust metabolomics techniques exist today which are well capable of processing large data sets from time course experiments or segregating populations at moderate costs and with high throughput.

How can the results obtained in metabolomics experiments be used? As was pointed out by Hall (2006), metabolomics is a particularly suitable initial approach as a hypothesis generator to use to provide early leads for future research. This reflects exactly what was achieved in this study. The path to verify some of these leads, such as QTL, using forward genetics approaches is quite clear and has been successfully applied in a number of cases (Fridman *et al.*, 2004; Konishi *et al.*, 2006; Steinmetz *et al.*, 2002). Metabolomics is already used in diagnostics and gene-function analysis (Fernie *et al.*, 2004), can provide a more global picture of the molecular organization of multicellular organisms and help to investigate a part of the still unexploited biodiversity (Hall, 2006).

Large scale metabolite profiling most certainly will complement proteomics and genomics approaches. Especially the determination of gene expression has been

used to answer similar questions, as this work, with respect to the identification of QTL (Keurentjes *et al.*, 2007b; Kliebenstein *et al.*, 2006; West *et al.*, 2007) and the elucidation of the basis of heterosis (Swanson-Wagner *et al.*, 2006). However, as genes and gene expressions are the cause of changes which propagate via proteins to metabolites and ultimately growth, a future target will be to apply the 'omics' approaches in an integrative way (Hirai *et al.*, 2004b; Tohge *et al.*, 2005).

6.1.2 Comparing results of RIL and IL populations

Mapping approaches based on segregating populations of immortalized homozygous genotypes have been used for many years and numerous RIL and IL populations are available for divergent species such as tomato, rice, maize, and Arabidopsis (Alonso-Blanco *et al.*, 1998c; Burr *et al.*, 1988; Eshed and Zamir, 1995; Li *et al.*, 1995). While ILs have been used to fine map QTL detected in RIL populations in some cases, this study is amongst the first to explore two of such populations which cover the full Arabidopsis genome in parallel. Therefore we could ask the question whether or not RILs and ILs fulfill the expectations with respect to observed differences and overlaps. Although being conceptually quite similar, a few properties which distinguish RILs and ILs influence the anticipated results.

(i) QTL detected in composite interval mapping with RILs can be assigned precise estimates for the genetic position and effect. For ILs, the QTL position is defined by the boundaries of the introgression alone. Only in case that some overlap exists between one or several ILs the interval bearing the QTL may be narrowed down applying a binning approach (Schauer *et al.*, 2006). Further, each substitute chromosome segment may harbor more than one QTL hampering the estimation of the QTL effect. If two or more QTL are in coupling phase the chance of detection and the effects significance will be increased. If neighboring QTL within an introgression are in repulsion phase no significant difference to the recurrent parent may be detectable anymore.

(ii) Epistatic interactions between any two genes in a parental genome will be masked in an IL progeny if one of these genes is substituted with a donor segment of another genotype. New interactions may occur. While all epistatic interactions can be theoretically mapped in a RIL population (given a sufficient number of individuals) estimates for these effects will be different in an IL analysis.

(iii) RIL populations have a lower power than IL populations to detect small effect QTL due to segregation of multiple QTL in the genetic background (Keurentjes *et al.*, 2007a).

Further differences are related to multiple testing corrections – which depend on the number of markers for RILs and the number of lines for ILs – and the number of measurements necessary to detect a QTL.

The only two other studies which aimed to compare RILs and ILs in Arabidopsis investigated developmental traits and determined an overlap of approximately 50 % (Keurentjes *et al.*, 2007a; Kusterer *et al.*, 2007; Melchinger *et al.*, 2007a). We found a comparable confirmation rate (55 %) for RIL-QTL in ILs with respect to metabolic QTL per se. However, the confirmation rate was much lower (23 %) in the analysis of heterotic effects, indicating the contribution of epistasis to heterosis.

Is any of the populations preferred for metabolite profiling? Both population types are accessible for metabolite profiling but an appropriate experimental design has to be carefully considered. The amount of samples to be processed requires long measurement periods for the complete dataset. Thus, a well established and robust protocol is essential, as is a sufficient amount of controls (~10 %) to be measured along with the samples. Metabolite profiling of ILs has the advantage that parental samples can efficiently serve as a control for machine performance in parallel, thus, reducing the total number of samples to process. Furthermore, the statistical power in RIL analyses is increased with the number of lines investigated implying that the focus is on analyzing more lines rather than several replicates of each genotype. In contrast, the statistical power in IL analyses increases with the number of replicates measured per line. The results obtained by Keurentjes *et al.* (2007a) indicated that the strongest gain in power is reached until up to 6 replicates are processed while additional replicates allowed only moderate improvements in QTL detection. However, in GC-MS analyses these replicates can serve to investigate machine variance and reveal possible flaws. We conclude that while it is possible to detect numerous difference in both types of populations on the metabolomics level, ILs are preferable because they allow for additional control and provide excellent material to narrow down observed QTL through the generation of subILs. With respect to QTL mapping in general both population types can be regarded as complementary.

Association mapping will be a future alternative to RILs and ILs in QTL mapping approaches. Recently, the decrease in costs for haplotyping based on SNPs allowed

the prediction of more than 1 million non-redundant SNPs in 20 Arabidopsis accessions (Clark *et al.*, 2007). This paves the way to use the natural genetic variation in whole genome mapping approaches. Clear advantages are the easy access to large sample populations not limited to a particular cross and potentially high power and resolution in QTL detection (Buckler and Thornsberry, 2002). Association mapping has been successfully used in human genetics (Eerdewegh *et al.*, 2002; Ozaki *et al.*, 2002) and is currently adopted to plant genetics (Aranzana *et al.*, 2005; Wilson *et al.*, 2004).

6.1.3 Heterosis for metabolic traits

One goal of this study was to shed some light on the molecular processes underlying heterosis. To this end, hundreds of QTL for most of the evaluated metabolites have been determined and further characterized. This is an interesting finding per se as heterosis is most often associated with increased fitness related to biomass, stress resistance, fertility and likewise traits. To confirm the presence of heterosis down to the metabolic level may encourage conducting pathway and systems oriented approaches in the future. The present data allow drawing some conclusions already. Heterotic effects in primary metabolism are unequally distributed over the Arabidopsis genome and a hotspot on top of chromosome 4 is co-located with heterotic QTL mapped for biomass. Further, there is no evidence that pathway related genes are a major contributor to hmQTL. These findings indicate that a substantial amount of hybrid vigor detectable on the metabolic level is attributed by a small number of loci which are likely to be involved in regulatory processes. We have to be cautious, however, to extend this reasoning to other plants or phenotypic traits. Hochholdinger and Hoecker (2007) showed in their review of heterosis related gene expression studies conducted in maize, rice and Arabidopsis, that detected global trends – favoring dominance, overdominance or epistasis – are controversy and may depend on developmental stage, genetic background, analyzed tissue or the technical method applied.

Given the beauty and lucidity of the Mendelian laws it is tempting to hope that heterosis might be explainable by an equally simple mechanism. Current results, however, hint that the advantage of a heterozygous state operates in many ways. With respect to the primary metabolism in Arabidopsis cultivated under controlled conditions we detected predominantly dominant effects.

That said, one exciting question remains to be addressed: Will it be possible to predict heterosis based on any measure obtained from the two homozygous parents? Current approaches mainly focus on the genetic distance between the parents which is accessed based on genomic markers (Cho *et al.*, 2004; Schrag *et al.*, 2007; Yu *et al.*, 2005) but are still rather of academic interest and not yet applied by breeders (JC Reif, personal communication). As became apparent from our analysis of biomass and metabolic profiles, a multivariate approach could be more appropriate to predict an integrative trait. Promising results of transcript abundance correlating with heterosis in Arabidopsis encouraged Bancroft and colleges to file a patent (WO/2007/113532). In their approach a linear regression of the number of genes which were remodeled within a hybrid at a 1.5-fold change level on the magnitude of heterosis observed revealed a positive correlation of r = 0.738.

Clearly, gene expression values, protein content and metabolite levels are the likely candidates to be used as a multivariate measure. The metabolic state of a plant as the ultimate expression of its genotype and interaction with the environment could be closest related to heterosis. We already found a strong relation between metabolism and biomass and a less exposed connection to biomass heterosis. We hope to extend this approach in the future.

No individual gene involved in heterosis has hitherto been identified and characterized at the molecular level in plants (Hochholdinger and Hoecker, 2007). Can we expect to find any Mendelian locus exhibiting an effect strong enough to be detected? If at all, this should be the case for less integrative traits than are usually under study. Bancroft's results point in the direction that for highly integrative traits like biomass a rather large amount of small effects could be responsible on the level of gene expression. Therefore, it will be difficult to map any single locus effect and metabolomics might aid in elucidating the first heterotic gene.

6.2 Outlook

6.2.1 Resequencing of eight mQTL candidate genes

One of the objectives of this study was the characterization of metabolic QTL. Ultimately this means to disclose the underlying genes. As our results indicated that a large part of these genes may encode for enzymes from pathways where the metabolite is involved we decided to investigate possible polymorphisms between both parents. Here, we sequenced the coding region of accession C24 for eight of

our candidate genes. The candidate genes were chosen according to the following criteria:

- LOD value in RIL analysis (preferably high)
- Confirmation of the effect in ILs (preferably low *P*-value)
- Total number of genes associated with this metabolite in AraCyc (preferably low)
- Catalysed reaction (preferably directly acting on the metabolite)

Primers were designed according to the complementary Col-0 sequence within 200 bp up- and downstream of the promoter and the stop codon respectively.

According to the Perlegen data published by Clark *et al.* (2007) for three of these candidate genes substitutions were predicted. In total the authors annotated three substitutions and further six synonymous polymorphisms. However, as these annotations were based on hybridization efficiency applying a moderate false discovery rate (FDR) we expected to find additional changes. Altogether, five out of nine predicted polymorphisms were confirmed by resequencing of the respective genes. Additionally we found 23 polymorphisms to a large extent in intron regions (14) but also four of them causing substitutions which were not predicted beforehand (Table 4).

AGI	Substitution	Synonymus	Intron	bp	Metabolite	Gene function	LOD_{RIL}	P_{IL}
AT1G14520	3	6	7	2169	Inositol	MIOX1 (myo-inositol oxygenase)	6.5	0.047
AT5G53970	1	1	0	2240	Tyrosine	tyrosine aminotransferase	9.6	0.054
AT1G43710	0	1	1	2064	Ethanolamine	glutamate decarboxylase	8.7	0.000
AT3G44740	1	0	0	1446	Glycine	glycyl-tRNA synthetase	8.0	0.015
AT4G15210	0	0	4	3132	Maltose	beta-amylase activity	9.9	-
AT2G38400	0	0	2	2660	4-Aminobutyric acid	glyoxylate aminotransferase	3.6	0.004
AT4G05632	0	0	4	747	Glucose 1-phosphate	unknown (G3P DH)	10.7	0.013
AT5G15600	1	0	0	891	Nicotinic acid	unknown (Nitrilase)	13.2	0.013

Table 4 Polymorphisms between Col-0 and C24 identified in candidate genes of metabolic QTL by resequencing of C24. Length of the gene sequence (bp) and annotated function is given. Additionally the QTL LOD (RIL experiment) and the *P* value of the best confirming IL is shown.

Thus, for four of eight candidate genes determined by a comparison of the positions of metabolic QTL and present knowledge about metabolic pathways we found in total six polymorphisms changing the amino acid chain of the encoded protein. One of the remaining genes (AT4G15210, maltose) showed a significantly different expression profile between C24 and Col-0 at four days after sowing in a different experiment,

suggesting that a polymorphism might be located in the regulatory region. For candidate AT2G38400 (4-aminobutyric acid) a rather low LOD compared to the other testers was found in the RIL experiment. The genes AT1G43710 and AT4G05632 did exhibit one polymorphism per gene both located within intron regions and were otherwise not differentially expressed.

We are currently investigating possible changes in the tertiary structure of the enzymes caused by the determined substitutions. Furthermore, forward genetics approaches will be conducted. As we determined polymorphisms in all of the eight candidate genes under consideration, it seems promising to measure allele specific expression following the approach described by Wittkop *et al.* (2004).

Taking the rapid development in sequencing technology during recent years into account, the methods applied in this study will clearly not serve as a tool to reveal new polymorphisms, but rather aid in the functional characterization of present natural variation.

6.2.2 Metabolite flux analysis as a complement to investigate heterosis

A full understanding of the molecular mechanisms underlying the biological phenomenon heterosis has not been obtained so far. As this may be attributable to its complex nature a successful experimental strategy needs to approach the problem from different sides.

A possible criticism of the strategy followed throughout this work could be that plant material was sampled only at a single time point, thus, neglecting developmental changes as well as differences in metabolic fluxes.

The first point could be problematic in case that heterosis effects cause differences between homozygous and heterozygous plants before (or after) the sampling point. Hence, metabolite profiling might analyze only a subset of the responsible effects characteristic for a single developmental stage. It could even solely detect the differences in growth caused by heterosis without monitoring the initial processes leading to these differences.

The second point would hamper detection of heterotic effects in case that these effects purely influence metabolic fluxes. Increased fluxes may lead to increased biomass without changing the metabolite levels detectable by GC-MS per se.

With respect to the first point, it is clear that no time series data can be generated for large RIL or IL populations due to the necessary experimental effort. We chose

15 days after sowing (DAS) as a harvesting point to allow a substantial fixation of heterotic effects in biomass while differences in relative growth rate (biomass heterosis) still are detectable. In previous experiments it was shown that heterosis can be found as early as 10 DAS and until 15 DAS for low to intermediate light intensities as used in our study (Meyer *et al.*, 2004). Differences in the contribution to heterosis due to developmental changes are investigated in parallel in our group.

To account for the importance of metabolic fluxes, preliminary experiments using both parental accessions and their reciprocal hybrids were performed. We employed inverse $^{13}CO_2$-isotope dilution experiments (Huege *et al.*, 2007) harvesting at four time points within intervals of two hours during the light period 15 DAS. Prior to this, we investigated the number of carbon atoms (and hence the expected mass shifts) of all unique masses from our reference. We monitored the enrichment for 67 compounds and found significant differences between the genotypes for 14 metabolites (data not shown). We intend to use these data to narrow down the number of mQTL to be included in future investigations. Furthermore, we believe that such a 'fluxomics' approach will be a necessary complement in the exploration of heterosis.

6.2.3 Analysis of metabolite heterosis in *Zea mais*

It was pointed out earlier (Chapter 5) that Arabidopsis is an excellent model organism to investigate the molecular mechanisms underlying heterosis. It is clear however, that more applied approaches would focus on crop plants due to their agronomic importance. To account for that we currently investigate a maize root dataset using the methods developed within this work. Significant heterosis in primary root length was detected as early as three days after germination (DAG) in six out of twelve hybrids developed from maize inbred lines UH002, UH005, UH250 and UH301 (Hoecker *et al.*, 2006). We analyzed metabolite profiles of these well characterized crosses and could confirm MPH of up to 250% for e.g. leucine in samples harvested at 3.5 DAG (data not shown). Therefore, root metabolism may be used as an early indicator for heterosis and could be linked in a similar fashion as described in Chapter 5.

6.2.4 An extended metabolite GC-MS library based on KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a well known and widely used online database resource for biological systems (Kanehisa and Goto, 2000). If

compared to the Arabidopsis specific AraCyc database which was used to identify candidate genes for metabolic QTL (Chapter 4 and 5) it covers a broader range of species and consequently pathways and metabolites.

As became apparent during the study, many interesting effects – mQTL and explained variation of biomass – are contributed by metabolites of unknown chemical structure. Although it is in principle possible to elucidate the structure of some of these compounds by MS-MS experiments or NMR, a continuous extension of the available mass spectral libraries used in metabolomics experiments seems favorable.

In a collaborative approach in the Max-Planck-Institute of Molecular Plant Physiology more than 1,000 compounds were purchased and evaluated in single measurements and as mixtures (n=20) using the established GC-MS technique. The decision which compounds to include was based mainly on the annotation in KEGG. Hereby, a biological importance was ensured.

The R-Scripts developed for mass evaluation of metabolomics data (Chapter 2) were – with some slight adjustments – of great help in evaluating this dataset in an automatic fashion. A spectral library with automatically suggested reference spectra could be provided in a short time (two weeks). However, a manual inspection of a data subset revealed that the error rate of the automatic approach may be as high as 5 to 10 %. Thus a manual validation was initiated based on the present data to ensure a high quality final result.

It is intended to include this resource into the Golm Metabolome Database (Kopka *et al.*, 2005) to provide an open access for the scientific community in the future. A publication is currently in preparation.

6.3 Conclusion

The work presented in this thesis is the first large scale metabolite profiling analysis to integrate RIL and IL populations of *Arabidopsis thaliana*. The development of a robust method (Chapter 2) was a prerequisite for a stable annotation of 181 metabolite levels in more than 2000 measured samples. Within these data a metabolic signature related to plant biomass (Chapter 3) was found.

A large number of metabolite QTL (in total 157; Chapter 4) and heterotic metabolite QTL (in total 385; Chapter 5) were identified using quantitative genetics methods. These QTL were characterized with respect to their distribution, effect size, position and the co-localization with possible candidate genes based on present knowledge.

Primary efforts have been made to detect the underlying polymorphisms by comparative sequencing.

Moderate heterosis was found on a metabolic level and attributed mainly to dominance effects. The results further indicate that pathway related genes do not play a major role in hybrid vigor.

All together, the obtained results will serve as a rich data source for the identification of novel functional polymorphisms.

References

Achard, P., Cheng, H., Grauwe, L.D., Decat, J., Schoutteten, H., Moritz, T., Straeten, D.V.D., Peng, J. and Harberd, N.P. (2006) Integration of plant responses to environmentally activated phytohormonal signals. *Science*, **311**, 91-94.

Aharoni, A., de Vos, C.H.R., Verhoeven, H.A., Maliepaard, C.A., Kruppa, G., Bino, R. and Goodenowe, D.B. (2002) Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry. *OMICS*, **6**, 217--234.

Ahmadzadeh, A., Lee, E.A. and Tollenaar, M. (2004) Heterosis for Leaf CO2 Exchange Rate during the Grain-Filling Period in Maize. *Crop Sci.*, **44**, 2095--2100.

Allen, J., Davey, H.M., Broadhurst, D., Heald, J.K., Rowland, J.J., Oliver, S.G. and Kell, D.B. (2003) High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol.*, **21**, 692--696.

Alonso-Blanco, C., El-Assal, S.E., Coupland, G. and Koornneef, M. (1998a) Analysis of natural allelic variation at flowering time loci in the Landsberg erecta and Cape Verde Islands ecotypes of Arabidopsis thaliana. *Genetics*, **149**, 749-764.

Alonso-Blanco, C. and Koornneef, M. (2000) Naturally occurring variation in Arabidopsis: an underexploited resource for plant genetics. *Trends Plant Sci.*, **5**, 22--29.

Alonso-Blanco, C., Koornneef, M. and Stam, P. (1998b) The use of recombinant inbred lines (RILs) for genetic mapping. *Methods Mol. Biol.*, **82**, 137-146.

Alonso-Blanco, C., Peeters, A.J., Koornneef, M., Lister, C., Dean, C., van den Bosch, N., Pot, J. and Kuiper, M.T. (1998c) Development of an AFLP based linkage map of Ler, Col and Cvi Arabidopsis thaliana ecotypes and construction of a Ler/Cvi recombinant inbred line population. *Plant J.*, **14**, 259-271.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796--815.

Aranzana, M.J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., Toomajian, C., Traw, B., Zheng, H., Bergelson, J., Dean, C., Marjoram, P. and Nordborg, M. (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genet.*, **1**, e60.

Askenazi, M., Driggers, E.M., Holtzman, D.A., Norman, T.C., Iverson, S., Zimmer, D.P., Boers, M.-E., Blomquist, P.R., Martinez, E.J., Monreal, A.W., Feibelman, T.P., Mayorga, M.E., Maxon, M.E., Sykes, K., Tobin, J.V., Cordero, E., Salama, S.R., Trueheart, J., Royer, J.C. and Madden, K.T. (2003) Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nat. Biotechnol.*, **21**, 150--156.

Atienza, S.G., Satovic, Z., Petersen, K.K., Dolstra, O. and Martín, A. (2003) Identification of QTLs influencing agronomic traits in Miscanthus sinensis Anderss. I. Total height, flag-leaf height and stem diameter. *Theor. Appl. Genet.*, **107**, 123-129.

Auger, D.L., Gray, A.D., Ream, T.S., Kato, A., Coe, E.H. and Birchler, J.A. (2005) Nonadditive gene expression in diploid and triploid hybrids of maize. *Genetics*, **169**, 389--397.

Bancroft, I., Stokes, R.D., Morgan, L.C., Fraser, F. and O'Neill, M.C. (2007) Prediction Of Heterosis And Other Traits By Transcriptome Analysis. WO/2007/113532.

Basten, C.J., Weir, B.S. and Zeng, Z.B. (1994) *Zmap - a QTL cartographer*: Organizing Committee, 5th World Congress on Genetics Applied to Livestock Production, Guelph, Ontario, Canada.

Becker, H.C. (1993) *Pflanzenzüchtung*. Stuttgart, Germany: Eugen Ulmer Verlag.

Bentsink, L., Yuan, K., Koornneef, M. and Vreugdenhil, D. (2003) The genetics of phytate and phosphate accumulation in seeds and leaves of Arabidopsis thaliana, using natural variation. *Theor. Appl. Genet.*, **106**, 1234-1243.

Bernacchi, D. and Tanksley, S.D. (1997) An interspecific backcross of Lycopersicon esculentum x L. hirsutum: linkage analysis and a QTL study of sexual compatibility factors and floral traits. *Genetics*, **147**, 861--877.

Birchler, J.A., Auger, D.L. and Riddle, N.C. (2003) In Search of the Molecular Basis of Heterosis. *Plant Cell*, **15**, 2236-2239.

Blumberg, L.M., David, F., Klee, M.S. and Sandra, P. (2008) Comparison of one-dimensional and comprehensive two-dimensional separations by gas chromatography. *Journal of chromatography*, **1188**, 2-16.

Broeckling, C.D., Huhman, D.V., Farag, M.A., Smith, J.T., May, G.D., Mendes, P., Dixon, R.A. and Sumner, L.W. (2005) Metabolic profiling of Medicago truncatula cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. *J. Exp. Bot.*, **56**, 323--336.

Broman, K.W., Wu, H., Sen, S. and Churchill, G.A. (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, **19**, 889-890.

Bruce, A.B. (1910) The mendelian theory of heredity and the augmentation of vigor. *Science*, **32**, 627--628.

Brunner, S., Fengler, K., Morgante, M., Tingey, S. and Rafalski, A. (2005) Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell*, **17**, 343--360.

Buckler, E.S. and Thornsberry, J.M. (2002) Plant molecular diversity and applications to genomics. *Curr. Opin. Plant Biol.*, **5**, 107--111.

Burr, B., Burr, F.A., Thompson, K.H., Albertson, M.C. and Stuber, C.W. (1988) Gene mapping with recombinant inbreds in maize. *Genetics*, **118**, 519--526.

Calenge, F., Saliba-Colombani, V., Mahieu, S., Loudet, O., Daniel-Vedele, F. and Krapp, A. (2006) Natural variation for carbohydrate content in Arabidopsis. Interaction with complex traits dissected by quantitative genetics. *Plant Physiol.*, **141**, 1630-1643.

Campbell, D.R., Galen, C. and Wu, C.A. (2005) Ecophysiology of first and second generation hybrids in a natural plant hybrid zone. *Oecologia*, **144**, 214--225.

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.

Chen, S., Hajirezaei, M., Peisker, M., Tschiersch, H., Sonnewald, U. and Börnke, F. (2005) Decreased sucrose-6-phosphate phosphatase level in transgenic tobacco inhibits photosynthesis, alters carbohydrate partitioning, and reduces growth. *Planta*, **221**, 479--492.

Cho, Y.-I., Park, C.-W., Kwon, S.-W., Chin, J.-H., Ji, H.-S., Park, K.-J., McCouch, S. and Koh, H.-J. (2004) Key DNA Markers for Predicting Heterosis in F1 Hybrids of japonica Rice. *Breeding Sci.*, **54**, 389--397.

Churchill, G.A. and Doerge, R.W. (1994) Empirical Threshold Values for Quantitative Trait Mapping. *Genetics*, **138**, 963-971.

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.

Clayton, T.A., Lindon, J.C., Cloarec, O., Antti, H., Charuel, C., Hanton, G., Provost, J.-P., Net, J.-L.L., Baker, D., Walley, R.J., Everett, J.R. and Nicholson, J.K. (2006) Pharmaco-metabonomic phenotyping and personalized drug treatment. *Nature*, **440**, 1073--1077.

Cockerham, C.C. and Zeng, Z.B. (1996) Design III with marker loci. *Genetics*, **143**, 1437-1456.

Comstock, R.E. and Robinson, H.F. (1952) *Estimation of average dominance of genes*: Iowa State College Press, Ames, IA.

Cross, J.M., von Korff, M., Altmann, T., Bartzetko, L., Sulpice, R., Gibon, Y., Palacios, N. and Stitt, M. (2006) Variation of enzyme activities and metabolite levels in 24 Arabidopsis accessions growing in carbon-limited conditions. *Plant Physiol.*, **142**, 1574--1588.

Crow, J.F. (1948) Alternative Hypotheses of Hybrid Vigor. *Genetics*, **33**, 477--487.

Darwin, C. (1876) *The effects of cross- and self-fertilization in vegetable kingdom*: Appleton, New York.

Davenport, C.B. (1908) Degeneration, albinism and inbreeding. *Science*, **28**, 454-455.

Defernez, M., Gunning, Y.M., Parr, A.J., Shepherd, L.V.T., Davies, H.V. and Colquhoun, I.J. (2004) NMR and HPLC-UV profiling of potatoes with genetic modifications to metabolic pathways. *J. Agric. Food Chem.*, **52**, 6075--6085.

Desbrosses, G.G., Kopka, J. and Udvardi, M.K. (2005) Lotus japonicus metabolic profiling. Development of gas chromatography-mass spectrometry resources for the study of plant-microbe interactions. *Plant Physiol.*, **137**, 1302--1318.

Duran, A.L., Yang, J., Wang, L. and Sumner, L.W. (2003) Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics*, **19**, 2283-2293.

Duvick, D.N. (1999) *Heterosis: Feeding People and Protecting Natural Resources*. Madison, WI: American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc.

East, E.M. and Hayes, H.K. (1912) Heterozygosis in evolution and plant breeding. *U.S. Dept. Agric. Bur. Plant Industr. Bull.*, **243**, 1-58.

Eerdewegh, P.V., Little, R.D., Dupuis, J., Mastro, R.G.D., Falls, K., Simon, J., Torrey, D., Pandit, S., McKenny, J., Braunschweiger, K., Walsh, A., Liu, Z., Hayward, B., Folz, C., Manning, S.P., Bawa, A., Saracino, L., Thackston, M., Benchekroun, Y., Capparell, N., Wang, M., Adair, R., Feng, Y., Dubois, J., FitzGerald, M.G., Huang, H., Gibson, R., Allen, K.M., Pedan, A., Danzig, M.R., Umland, S.P., Egan, R.W., Cuss, F.M., Rorke, S., Clough, J.B., Holloway, J.W., Holgate, S.T. and Keith, T.P. (2002) Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature*, **418**, 426--430.

El-Lithy, M.E., Clerkx, E.J.M., Ruys, G.J., Koornneef, M. and Vreugdenhil, D. (2004) Quantitative trait locus analysis of growth-related traits in a new Arabidopsis recombinant inbred population. *Plant Physiol.*, **135**, 444-458.

Eshed, Y. and Zamir, D. (1995) An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. *Genetics*, **141**, 1147--1162.

Falconer, D.S. and Mackay, T.F.C. (1996) *Quantitative Genetics* 4th edn: Pearson Education Limited.

Fernie, A.R., Tauberger, E., Lytovchenko, A., Roessner, U., Willmitzer, L. and Trethewey, R.N. (2002) Antisense repression of cytosolic phosphoglucomutase in potato (Solanum tuberosum) results in severe growth retardation, reduction in tuber number and altered carbon metabolism. *Planta*, **214**, 510-520.

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.

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R.N. and Willmitzer, L. (2000a) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, **18**, 1157-1161.

Fiehn, O., Kopka, J., Trethewey, R.N. and Willmitzer, L. (2000b) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.*, **72**, 3573--3580.

Fiehn, O., Wohlgemuth, G. and Scholz, M. (2005) Setup and Annotation of Metabolomic Experiments by Integrating Biological and Mass Spectrometric Metadata. *LNCS*, **3615**, 224-239.

Frank, I.E. and Friedman, J.H. (1993) A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109--135.

Frascaroli, E., Canè, M.A., Landi, P., Pea, G., Gianfranceschi, L., Villa, M., Morgante, M. and Pè, M.E. (2007) Classical genetic and QTL analyses of heterosis in a maize hybrid between two elite inbred lines. *Genetics*, **176**, 625–644.

Fridman, E., Carrari, F., Liu, Y.-S., Fernie, A.R. and Zamir, D. (2004) Zooming In on a Quantitative Trait for Tomato Yield Using Interspecific Introgressions. *Science*, **305**, 1786 - 1789.

Fu, H. and Dooner, H.K. (2002) Intraspecific violation of genetic colinearity and its implications in maize. *Proc. Natl. Acad. Sci. U S A*, **99**, 9573--9578.

Garg, A.K., Kim, J.-K., Owens, T.G., Ranwala, A.P., Choi, Y.D., Kochian, L.V. and Wu, R.J. (2002) Trehalose accumulation in rice plants confers high tolerance levels to different abiotic stresses. *Proc. Natl. Acad. Sci. U S A*, **99**, 15898--15903.

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.

Gibson, R.W., Aritua, V., Byamukama, E., Mpembe, I. and Kayongo, J. (2004) Control strategies for sweet potato virus disease in Africa. *Virus Res.*, **100**, 115--122.

Gittins, R. (1985) *Canonical Analysis - A review with applications in ecology*. Berlin: Springer.

Goossens, A., Häkkinen, S.T., Laakso, I., Seppänen-Laakso, T., Biondi, S., Sutter, V.D., Lammertyn, F., Nuutila, A.M., Söderlund, H., Zabeau, M., Inzé, 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.

Gullberg, J., Jonsson, P., Nordström, A., Sjöström, 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.

Guo, M., Rupe, M.A., Danilevskaya, O.N., Yang, X. and Hu, Z. (2003) Genome-wide mRNA profiling reveals heterochronic allelic variation and a new imprinted gene in hybrid maize endosperm. *Plant J.*, **36**, 30--44.

Guo, M., Rupe, M.A., Zinselmeier, C., Habben, J., Bowen, B.A. and Smith, O.S. (2004) Allelic variation of gene expression in maize hybrids. *Plant Cell*, **16**, 1707--1716.

Haas, B.J., Wortman, J.R., Ronning, C.M., Hannick, L.I., Smith, R.K., Maiti, R., Chan, A.P., Yu, C., Farzad, M., Wu, D., White, O. and Town, C.D. (2005) Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. *BMC Biol.*, **3**, 7.

Hackett, C.A. (2002) Statistical methods for QTL mapping in cereals. *Plant. Mol. Biol.*, **48**, 585-599.

Hageman, G.J. and Stierum, R.H. (2001) Niacin, poly(ADP-ribose) polymerase-1 and genomic stability. *Mutat. Res.*, **475**, 45--56.

Haley, C.S. and Knott, S.A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, **69**, 315--324.

Halket, J.M. and Zaikin, V.G. (2003) Derivatization in mass spectrometry--1. Silylation. *Eur. J. Mass Spectrom.*, **9**, 1--21.

Hall, R.D. (2006) Plant metabolomics: from holistic hope, to hype, to hot topic. *New Phytol.*, **169**, 453-468.

Harrigan, G.G. and Goodacre, R. (2003) *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*: Springer.

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. and Saito, K. (2004) Post-genomics approaches for the elucidation of plant adaptive mechanisms to sulphur deficiency. *J. Exp. Bot.*, **55**, 1871--1879.

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. (2004a) Transcriptome and metabolome analyses reveal a whole adaptive process of plant to sulfur deficiency. *Plant Cell Physiol.*, **45**, S122-S122.

Hirai, M.Y., Yano, M., Goodenowe, D.B., Kanaya, S., Kimura, T., Awazuhara, M., Arita, M., Fujiwara, T. and Saito, K. (2004b) 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.

Hittalmani, S., Shashidhar, H.E., Bagali, P.G., Huang, N., Sidhu, J.S., Singh, V.P. and Khush, G.S. (2002) Molecular mapping of quantitative trait loci for plant growth,

yield and yield related traits across three diverse locations in a doubled haploid rice population. *Euphytica*, **125**, 207-214.

Hochholdinger, F. and Hoecker, N. (2007) Towards the molecular basis of heterosis. *Trends Plant Sci.*, **12**, 427--432.

Hoecker, N., Keller, B., Piepho, H.P. and Hochholdinger, F. (2006) Manifestation of heterosis during early maize (Zea mays L.) root development. *Theor. Appl. Genet.*, **112**, 421-429.

Hotelling, H. and Gittins, R. (1935) The most predictable criterion. *J. Educational Psychol.*, **26**, 139--143.

Huang, Y., Zhang, L., Zhang, J., Yuan, D., Xu, C., Li, X., Zhou, D., Wang, S. and Zhang, Q. (2006) Heterosis and polymorphisms of gene expression in an elite rice hybrid as revealed by a microarray analysis of 9198 unique ESTs. *Plant. Mol. Biol.*, **62**, 579--591.

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.

Hull, F.H. (1945) Recurrent selection for specific combining ability in corn. *J. Am. Soc. Agron.*, **37**, 134-135.

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851--861.

Ishizaki, K., Larson, T.R., Schauer, N., Fernie, A.R., Graham, I.A. and Leaver, C.J. (2005) The critical role of Arabidopsis electron-transfer flavoprotein:ubiquinone oxidoreductase during dark-induced starvation. *Plant Cell*, **17**, 2587--2600.

Ishizaki, K., Schauer, N., Larson, T.R., Graham, I.A., Fernie, A.R. and Leaver, C.J. (2006) The mitochondrial electron transfer flavoprotein complex is essential for survival of Arabidopsis in extended darkness. *Plant J.*, **47**, 751--760.

Jacobsson, L., Park, H.-B., Wahlberg, P., Fredriksson, R., Perez-Enciso, M., Siegel, P.B. and Andersson, L. (2005) Many QTLs with minor additive effects are associated with a large difference in growth between two selection lines in chickens. *Genet. Res.*, **86**, 115-125.

Jansen, R.C. and Nap, J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet.*, **17**, 388--391.

Jenkins, H., Hardy, N., Beckmann, M., Draper, J., Smith, A.R., Taylor, J., Fiehn, O., Goodacre, R., Bino, R.J., Hall, R., Kopka, J., Lane, G.A., Lange, B.M., Liu, J.R., Mendes, P., Nikolau, B.J., Oliver, S.G., Paton, N.W., Rhee, S., Roessner-Tunali, U., Saito, K., Smedsgaard, J., Sumner, L.W., Wang, T., Walsh, S., Wurtele, E.S. and Kell, D.B. (2004) A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotechnol.*, **22**, 1601-1606.

Jompuk, C., Fracheboud, Y., Stamp, P. and Leipner, J. (2005) Mapping of quantitative trait loci associated with chilling tolerance in maize (Zea mays L.) seedlings grown under field conditions. *J. Exp. Bot.*, **56**, 1153-1163.

Junker, B.H., Wuttke, R., Tiessen, A., Geigenberger, P., Sonnewald, U., Willmitzer, L. and Fernie, A.R. (2004) Temporally regulated expression of a yeast invertase in potato tubers allows dissection of the complex metabolic phenotype obtained following its constitutive expression. *Plant. Mol. Biol.*, **56**, 91--110.

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nuc. Ac. Res.*, **28**, 27--30.

Kao, C.H., Zeng, Z.B. and Teasdale, R.D. (1999) Multiple interval mapping for quantitative trait loci. *Genetics*, **152**, 1203-1216.

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.

Kearsey, M.J. and Farquhar, A.G. (1998) QTL analysis in plants; where are we now? *Heredity*, **80 (Pt 2)**, 137-142.

Kearsey, M.J. and Jinks, J.L. (1968) A general method of detecting additive, dominance and epistatic variation for metrical traits. *Heredity*, **23**, 403--409.

Kearsey, M.J. and Pooni, H.S. (1996) *The Genetical Analysis of Quantitative Traits*: Chapman & Hall, London.

Keurentjes, J.J.B., Bentsink, L., Alonso-Blanco, C., Hanhart, C.J., Vries, H.B.-D., Effgen, S., Vreugdenhil, D. and Koornneef, M. (2007a) Development of a near-isogenic line population of Arabidopsis thaliana and comparison of mapping power with a recombinant inbred line population. *Genetics*, **175**, 891-905.

Keurentjes, J.J.B., Fu, J., de Vos, C.H.R., Lommen, A., Hall, R.D., Bino, R.J., van der Plas, L.H.W., Jansen, R.C., Vreugdenhil, D. and Koornneef, M. (2006) The genetics of plant metabolism. *Nat. Genet.*, **38**, 842-849.

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. (2007b) Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proc. Natl. Acad. Sci. U S A*, **104**, 1708-1713.

Kim, S., Plagnol, V., Hu, T.T., Toomajian, C., Clark, R.M., Ossowski, S., Ecker, J.R., Weigel, D. and Nordborg, M. (2007) Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nat. Genet.*, **39**, 1151--1155.

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.

Kliebenstein, D.J., West, M.A.L., van Leeuwen, H., Loudet, O., Doerge, R.W. and Clair, D.A.S. (2006) Identification of QTLs controlling gene expression networks defined a priori. *BMC Bioinformatics*, **7**, 308.

Kolbe, A., Tiessen, A., Schluepmann, H., Paul, M., Ulrich, S. and Geigenberger, P. (2005) Trehalose 6-phosphate regulates starch synthesis via posttranslational redox activation of ADP-glucose pyrophosphorylase. *Proc. Natl. Acad. Sci. U S A*, **102**, 11118--11123.

Konishi, S., Izawa, T., Lin, S.Y., Ebana, K., Fukuta, Y., Sasaki, T. and Yano, M. (2006) An SNP caused loss of seed shattering during rice domestication. *Science*, **312**, 1392--1396.

Kopka, J., Fernie, A., Weckwerth, W., Gibon, Y. and Stitt, M. (2004) Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.*, **5**, 109.

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. and Steinhauser, D. (2005) GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics*, **21**, 1635-1638.

Korstanje, R. and Paigen, B. (2002) From QTL to gene: the harvest begins. *Nat. Genet.*, **31**, 235-236.

Kroymann, J., Donnerhacke, S., Schnabelrauch, D. and Mitchell-Olds, T. (2003) Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. *Proc. Natl. Acad. Sci. U S A*, **100 Suppl 2**, 14587--14592.

Kroymann, J. and Mitchell-Olds, T. (2005) Epistasis and balanced polymorphism influencing complex trait variation. *Nature*, **435**, 95-98.

Kruglyak, L. (2008) The road to genome-wide association studies. *Nat. Rev. Genet.*, **9**, 314--318.

Kuss, M. and Graepel, T. (2003) The geometry of kernel canonical correlation analysis. Max Planck Institute for Biological Cybernetics.

Kusterer, B., Piepho, H.-P., Utz, H.F., Schön, C.C., Muminovic, J., Meyer, R.C., Altmann, T. and Melchinger, A.E. (2007) Heterosis for biomass-related traits in Arabidopsis investigated by quantitative trait Loci analysis of the triple testcross design with recombinant inbred lines. *Genetics*, **177**, 1839--1850.

Lander, E.S. and Botstein, D. (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185-199.

Lark, K.G., Chase, K., Adler, F., Mansur, L.M. and Orf, J.H. (1995) Interactions between quantitative trait loci in soybean in which trait variation at one locus is conditional upon a specific allele at another. *Proc. Natl. Acad. Sci. U S A*, **92**, 4656--4660.

Laudadio, T., Pels, P., Lathauwer, L.D., Hecke, P.V. and Huffel, S.V. (2005) Tissue segmentation and classification of MRSI data using canonical correlation analysis. *Magn. Reson. Med.*, **54**, 1519--1529.

Li, S.-B., Zhang, Z.-H., Ying, H., Li, C.-Y., Xuan, J., Ting, M., Li, Y.-S. and Zhu, Y.-G. (2006) Genetic dissection of developmental behavior of crop growth rate and its relationships with yield and yield related traits in rice. *Plant Sci.*, **170**, 911-917.

Li, Z., Pinson, S.R.M., Stansel, J.W. and Park, W.D. (1995) Identification of quantitative trait loci (QTLs) for heading date and plant height in cultivated rice (Oryza sativa L.). *Theor. Appl. Genet.*, **91**, 374--381.

Li, Z.K., Luo, L.J., Mei, H.W., Wang, D.L., Shu, Q.Y., Tabien, R., Zhong, D.B., Ying, C.S., Stansel, J.W., Khush, G.S. and Paterson, A.H. (2001) Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. I. Biomass and grain yield. *Genetics*, **158**, 1737--1753.

Lincoln, S., Daly, M. and Lander, E. (1992) Constructing Genetics Maps with Mapmaker/Exp. 3.0. Whitehead Inst. Tech. Rep. (Whitehead Inst., Cambridge, MA).

Lindon, J.C. (2003) HPLC-NMR-MS: past, present and future. *Drug Discov. Today*, **8**, 1021--1022.

Lindon, J.C., Holmes, E., Bollard, M.E., Stanley, E.G. and Nicholson, J.K. (2004) Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. *Biomarkers*, **9**, 1--31.

Lippman, Z.B. and Zamir, D. (2007) Heterosis: revisiting the magic. *Trends Genet.*, **23**, 60-66.

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.

Lisec, J., Schauer, N., Kopka, J., Willmitzer, L. and Fernie, A.R. (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat. Protoc.*, **1**, 387--396.

Liso, R., Calabrese, G., Bitonti, M.B. and Arrigoni, O. (1984) Relationship between ascorbic acid and cell division. *Exp. Cell Res.*, **150**, 314--320.

Lorberth, R., Ritte, G., Willmitzer, L. and Kossmann, J. (1998) Inhibition of a starch-granule-bound protein leads to modified starch and repression of cold sweetening. *Nat. Biotechnol.*, **16**, 473-477.

Loudet, O., Chaillou, S., Merigout, P., Talbotec, J.I. and Daniel-Vedele, F.o. (2003) Quantitative Trait Loci Analysis of Nitrogen Use Efficiency in Arabidopsis. *Plant Physiol.*, **131**, 345-358.

Lu, H., Romero-Severson, J. and Bernardo, R. (2003) Genetic basis of heterosis explored by simple sequence repeat markers in a random-mated maize population. *Theor. Appl. Genet.*, **107**, 494--502.

Luedemann, A., Strassburg, K., Erban, A. and Kopka, J. (2008) TagFinder for the quantitative analysis of gas chromatography--mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics*, **24**, 732--737.

Luo, L.J., Li, Z.K., Mei, H.W., Shu, Q.Y., Tabien, R., Zhong, D.B., Ying, C.S., Stansel, J.W., Khush, G.S. and Paterson, A.H. (2001) Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. II. Grain yield components. *Genetics*, **158**, 1755--1771.

Mann, V., Harker, M., Pecker, I. and Hirschberg, J. (2000) Metabolic engineering of astaxanthin production in tobacco flowers. *Nat. Biotechnol.*, **18**, 888-892.

Mei, H.W., Li, Z.K., Shu, Q.Y., Guo, L.B., Wang, Y.P., Yu, X.Q., Ying, C.S. and Luo, L.J. (2005) Gene actions of QTLs affecting several agronomic traits resolved in a recombinant inbred rice population and two backcross populations. *Theor. Appl. Genet.*, **110**, 649--659.

Meiler, J. and Will, M. (2002) Genius: a genetic algorithm for automated structure elucidation from 13C NMR spectra. *J. Am. Chem. Soc.*, **124**, 1868--1870.

Melchinger, A.E. (1999) *Genetic diversity and heterosis*: American Society of Agronomy: Crop Science Society of America: Soil Science Society of America, Madison, WI.

Melchinger, A.E., Piepho, H.-P., Utz, H.F., Muminovic, J., Wegenast, T., Törjék, O., Altmann, T. and Kusterer, B. (2007a) Genetic basis of heterosis for growth-related traits in Arabidopsis investigated by testcross progenies of near-isogenic lines reveals a significant role of epistasis. *Genetics*, **177**, 1827--1837.

Melchinger, A.E., Utz, H.F., Piepho, H.P., Zeng, Z.B. and Schön, C.C. (2007b) The role of epistasis in the manifestation of heterosis: a systems-oriented approach. *Genetics*, **177**, 1815--1825.

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. (2007a) The metabolic signature related to high plant growth rate in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U S A*, **104**, 4759-4764.

Meyer, R.C., Törjek, O., Becher, M. and Altmann, T. (2004) Heterosis of Biomass Production in Arabidopsis. Establishment during Early Development. *Plant Physiol.*, **134**, 1813-1823.

Meyer, S., Pospisil, H. and Scholten, S. (2007b) Heterosis associated gene expression in maize embryos 6 days after fertilization exhibits additive, dominant and overdominant pattern. *Plant. Mol. Biol.*, **63**, 381-391.

Milborrow, B. (1998) A biochemical mechanism for hybrid vigour. *J. Exp. Bot.*, **49**, 1063-1071.

Monforte, A.J. and Tanksley, S.D. (2000) Fine mapping of a quantitative trait locus (QTL) from Lycopersicon hirsutum chromosome 1 affecting fruit characteristics and agronomic traits: breaking linkage among QTLs affecting different traits and dissection of heterosis for yield. *Theor. Appl. Genet.*, **100**, 471--479.

Morikawa, T., Mizutani, M., Aoki, N., Watanabe, B., Saga, H., Saito, S., Oikawa, A., Suzuki, H., Sakurai, N., Shibata, D., Wadano, A., Sakata, K. and Ohta, D. (2006) Cytochrome P450 CYP710A encodes the sterol C-22 desaturase in Arabidopsis and tomato. *Plant Cell*, **18**, 1008--1022.

Muir, S.R., Collins, G.J., Robinson, S., Hughes, S., Bovy, A., Vos, C.H.R.D., van Tunen, A.J. and Verhoeyen, M.E. (2001) Overexpression of petunia chalcone isomerase in tomato results in fruit containing increased levels of flavonols. *Nat. Biotechnol.*, **19**, 470-474.

Nikiforova, V.J., Kopka, J., Tolstikov, V., Fiehn, O., Hopkins, L., Hawkesford, M.J., Hesse, H. and Hoefgen, R. (2005) Systems rebalancing of metabolism in response to sulfur deprivation, as revealed by metabolome analysis of Arabidopsis plants. *Plant Physiol.*, **138**, 304--318.

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.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y. and Tanaka, T. (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.*, **32**, 650--654.

Payne, R.W., Baird, D.B., Cherry, M., Gilmour, A.R., Harding, S.A., Kane, A.F., Lane, P.W., Murray, D.A., Soutar, D.M., Thompson, R., Todd, A.D., Tunnicliffe, W.G., Webster, R. and Welham, S.J. (2002) *Genstat Release 6.1 Reference Manual*. Oxford, UK: VSN International.

Piepho, H.P., Büchse, A. and Emrich, K. (2003) A Hitchhiker's Guide to Mixed Models for Randomized Experiments. *J. Agron. Crop Sci.*, **189**, 310-322.

Plumb, R.S., Stumpf, C.L., Granger, J.H., Castro-Perez, J., Haselden, J.N. and Dear, G.J. (2003) Use of liquid chromatography/time-of-flight mass spectrometry and multivariate statistical analysis shows promise for the detection of drug metabolites in biological fluids. *Rapid Commun. Mass Spectrom.*, **17**, 2632--2638.

Powers, L. (1944) An expansion of Jones' theory for the explanation of heterosis. *Am. Nat.*, **78**, 275-280.

Rauh, L., Basten, C. and Buckler, S. (2002) Quantitative trait loci analysis of growth response to varying nitrogen sources in Arabidopsis thaliana. *Theor. Appl. Genet.*, **104**, 743-750.

Razavi, A.R., Gill, H., Ståal, O., Sundquist, M., Thorstenson, S., Ahlfeldt, H., Shahsavar, N. and Group, S.-E.S.B.C.S. (2005) Exploring cancer register data to find risk factors for recurrence of breast cancer--application of Canonical Correlation Analysis. *BMC Med. Inform. Decis. Mak.*, **5**, 29.

Rocha, J.L., Eisen, E.J., Vleck, L.D.V. and Pomp, D. (2004) A large-sample QTL study in mice: I. Growth. *Mamm. Genome*, **15**, 83-99.

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

Roessner-Tunali, U., Liu, J., Leisse, A., Balbo, I., Perez-Melis, A., Willmitzer, L. and Fernie, A.R. (2004) Kinetics of labelling of organic and amino acids in potato tubers by gas chromatography-mass spectrometry following incubation in (13)C labelled isotopes. *Plant J.*, **39**, 668-679.

Roessner-Tunali, U., Urbanczyk-Wochniak, E., Czechowski, T., Kolbe, A., Willmitzer, L. and Fernie, A.R. (2003b) De novo amino acid biosynthesis in potato tubers is regulated by sucrose levels. *Plant Physiol.*, **133**, 683--692.

Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L. and Fernie, A. (2001a) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*, **13**, 11-29.

Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N. and Willmitzer, L. (2000) Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.*, **23**, 131--142.

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

Saito, K., Dixon, R.A. and Willmitzer, L. (2006) *Plant metabolomics*: Springer Verlag, Berlin Heidelberg.

Salvi, S. and Tuberosa, R. (2005) To clone or not to clone plant QTLs: present and future challenges. *Trends Plant Sci*, **10**, 297--304.

Sato, S., Soga, T., Nishioka, T. and Tomita, M. (2004) Simultaneous determination of the main metabolites in rice leaves using capillary electrophoresis mass spectrometry and capillary electrophoresis diode array detection. *Plant J.*, **40**, 151--163.

Sauter, H., Lauer, M. and Fritsch, H. (1991) Metabolic profiling of plants. A new diagnostic technique. *ACS Symp. Ser.*, **443**, 288-299.

Schad, M., Mungur, R., Fiehn, O. and Kehr, J. (2005) Metabolic profiling of laser microdissected vascular bundles of Arabidopsis thaliana. *Plant Methods*, **1**, 2.

Schauer, N., Semel, Y., Balbo, I., Steinfath, M., Repsilber, D., Selbig, J., Pleban, T., Zamir, D. and Fernie, A.R. (2008) Mode of Inheritance of Primary Metabolic Traits in Tomato. *Plant Cell*, (online).

Schauer, N., Semel, Y., Roessner, U., Gur, A., Balbo, I., Carrari, F., Pleban, T., PerezMelis, 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., 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.

Schluepmann, H., Pellny, T., van Dijken, A., Smeekens, S. and Paul, M. (2003) Trehalose 6-phosphate is indispensable for carbohydrate utilization and growth in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U S A*, **100**, 6849--6854.

Schnee, C., Köllner, T.G., Held, M., Turlings, T.C.J., Gershenzon, J. and Degenhardt, J. (2006) The products of a single maize sesquiterpene synthase form a volatile defense signal that attracts natural enemies of maize herbivores. *Proc. Natl. Acad. Sci. U S A*, **103**, 1129--1134.

Scholz, M., Gatzek, S., Sterling, A., Fiehn, O. and Selbig, J. (2004) Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics*, **20**, 2447-2454.

Scholz, M., Kaplan, F., Guy, C.L., Kopka, J. and Selbig, J. (2005) Non-linear PCA: a missing data approach. *Bioinformatics*, **21**, 3887--3895.

Schön, C.C., Utz, H.F., Groh, S., Truberg, B., Openshaw, S. and Melchinger, A.E. (2004) Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. *Genetics*, **167**, 485-498.

Schrag, T.A., Maurer, H.P., Melchinger, A.E., Piepho, H.-P., Peleman, J. and Frisch, M. (2007) Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield. *Theor. Appl. Genet.*, **114**, 1345--1355.

Schubert, C. (2006) Can biofuels finally take center stage? *Nat. Biotechnol.*, **24**, 777--784.

Semel, Y., Nissenbaum, J., Menda, N., Zinder, M., Krieger, U., Issman, N., Pleban, T., Lippman, Z., Gur, A. and Zamir, D. (2006) Overdominant quantitative trait loci for yield and fitness in tomato. *Proc. Natl. Acad. Sci. U S A*, **103**, 12981-12986.

Shull, G.H. (1908) *The composition of a field of maize*: Rept. Amer. Breeders Assoc.

Smirnoff, N. (2000) Ascorbic acid: metabolism and functions of a multi-facetted molecule. *Curr. Opin. Plant Biol.*, **3**, 229--235.

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.

Soga, T., Ohashi, Y., Ueno, Y., Naraoka, H., Tomita, M. and Nishioka, T. (2003) Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J. Proteome Res.*, **2**, 488--494.

Somerville, C. (2006) The billion-ton biofuels vision. *Science*, **312**, 1277.

Song, R. and Messing, J. (2003) Gene expression of a gene family in maize based on noncollinear haplotypes. *Proc. Natl. Acad. Sci. U S A*, **100**, 9055–9060.

Sonnewald, U., Lerchl, J., Zrenner, R. and Frommer, W. (1994) Manipulation Of Sink-Source Relations In Transgenic Plants. *Plant Cell Environ.*, **17**, 649-658.

Steinmetz, L.M., Sinha, H., Richards, D.R., Spiegelman, J.I., Oefner, P.J., McCusker, J.H. and Davis, R.W. (2002) Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, **416**, 326-330.

Steuer, R., Kurths, J., Fiehn, O. and Weckwerth, W. (2003) Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, **19**, 1019-1026.

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.

Strelkov, S. (2004) Entwicklung und Anwendung einer Methode zur Metabolomanalyse von Corynebacterium glutamicum. Universität Köln.

Stuber, C.W., Lincoln, S.E., Wolff, D.W., Helentjaris, T. and Lander, E.S. (1992) Identification of Genetic Factors Contributing to Heterosis in a Hybrid From Two Elite Maize Inbred Lines Using Molecular Markers. *Genetics*, **132**, 823-839.

Stuber, C.W.M.D.E. and Wendel, J.F. (1987) Molecular marker-facilitated investigation of quantitative trait loci in maize. II. Factors influencing yields and its component traits. *Crop Sci.*, **27**, 639-648.

Stupar, R.M. and Springer, N.M. (2006) Cis-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. *Genetics*, **173**, 2199--2210.

Styczynski, M.P., Moxley, J.F., Tong, L.V., Walther, J.L., Jensen, K.L. and Stephanopoulos, G.N. (2007) Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery. *Anal. Chem.*, **79**, 966-973.

Sumner, L.W., Mendes, P. and Dixon, R.A. (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry*, **62**, 817--836.

Suzuki, H., Reddy, M.S.S., Naoumkina, M., Aziz, N., May, G.D., Huhman, D.V., Sumner, L.W., Blount, J.W., Mendes, P. and Dixon, R.A. (2005) Methyl jasmonate and yeast elicitor induce differential transcriptional and metabolic re-programming in cell suspension cultures of the model legume Medicago truncatula. *Planta*, **220**, 696--707.

Swanson-Wagner, R.A., Jia, Y., DeCook, R., Borsuk, L.A., Nettleton, D. and Schnable, P.S. (2006) All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. *Proc. Natl. Acad. Sci. U S A*, **103**, 6805-6810.

Swart, P.J., Bronner, G.M., Bruins, A.P., Ensing, K., Tepper, P.G. and De Zeeuw, R.A. (1993) HPLC-UV atmospheric pressure ionization mass spectrometric determination of the dopamine D2 agonist N-0923 and its major metabolites after oxidative metabolism by rat-, monkey-, and human liver microsomes. *Toxicology Methods*, **3**, 279-290.

Swarup, K., Alonso-Blanco, C., Lynn, J.R., Michaels, S.D., Amasino, R.M., Koornneef, M. and Millar, A.J. (1999) Natural allelic variation identifies new genes in the Arabidopsis circadian system. *Plant J.*, **20**, 67-77.

Syed, N.H. and Chen, Z.J. (2005) Molecular marker genotypes, heterozygosity and genetic interactions explain heterosis in Arabidopsis thaliana. *Heredity*, **94**, 295--304.

Tagashira, N., Plader, W., Filipecki, M., Yin, Z., Wiśniewska, A., Gaj, P., Szwacka, M., Fiehn, O., Hoshi, Y., Kondo, K. and Malepszy, S. (2005) The metabolic profiles of transgenic cucumber lines vary with different chromosomal locations of the transgene. *Cell. Mol. Biol. Lett.*, **10**, 697--710.

Tarpley, L., Duran, A.L., Kebrom, T.H. and Sumner, L.W. (2005) Biomarker metabolites capturing the metabolite variance present in a rice plant developmental period. *BMC Plant Biol.*, **5**, 8.

Taylor, J., King, R.D., Altmann, T. and Fiehn, O. (2002) Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics*, **18 Suppl 2**, S241--S248.

ter Steege, M.W., den Ouden, F.M., Lambers, H., Stam, P. and Peeters, A.J.M. (2005) Genetic and physiological architecture of early vigor in Aegilops tauschii, the D-genome donor of hexaploid wheat. A quantitative trait loci analysis. *Plant Physiol.*, **139**, 1078-1094.

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.

Tieman, D., Taylor, M., Schauer, N., Fernie, A.R., Hanson, A.D. and Klee, H.J. (2006) Tomato aromatic amino acid decarboxylases participate in synthesis of the flavor volatiles 2-phenylethanol and 2-phenylacetaldehyde. *Proc. Natl. Acad. Sci. U S A*, **103**, 8287--8292.

Titok, V.V., Iurenkova, S.I., Titok, M.V. and Khotyleva, L.V. (2005) Characterization of developmental changes in energy metabolism of fiber flax in heterosis. *Genetika*, **41**, 668--675.

Tkachenko, A.G., Pshenichnov, M.R. and Nesterova, L.I. (2001) [Putrescine as a oxidative stress protecting factor in Escherichia coli]. *Mikrobiologiia*, **70**, 487--494.

Tohge, T., Nishiyama, Y., Hirai, M.Y., Yano, M., Nakajima, J.-i., 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.

Tollenaar, M., Ahmadzadeh, A. and Lee, E.A. (2004) Physiological Basis of Heterosis for Grain Yield in Maize. *Crop Sci.*, **44**, 2086--2094.

Tolstikov, V.V. and Fiehn, O. (2002) Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal. Biochem.*, **301**, 298--307.

Tonsor, S.J., Alonso-Blanco, C. and Koornneef, M. (2005) Gene function beyond the single trait: natural variation, gene effects, and evolutionary ecology in Arabidopsis thaliana. *Plant Cell Environ.*, **28**, 2-20.

Törjék, O., Berger, D., Meyer, R.C., Müssig, C., Schmid, K.J., Sörensen, T.R., Weisshaar, B., Mitchell-Olds, T. and Altmann, T. (2003) Establishment of a high-efficiency SNP-based framework marker set for Arabidopsis. *Plant J.*, **36**, 122-140.

Törjék, O., Meyer, R.C., Zehnsdorf, M., Teltow, M., Strompen, G., Witucka-Wall, H., Blacha, A. and Altmann, T. (2008) Construction and Analysis of 2 Reciprocal Arabidopsis Introgression Line Populations. *J. Hered.*, (online).

Törjék, O., Witucka-Wall, H., Meyer, R.C., von Korff, M., Kusterer, B., Rautengarten, C. and Altmann, T. (2006) Segregation distortion in Arabidopsis C24/Col-0 and Col-0/C24 recombinant inbred line populations is due to reduced fertility caused by epistatic interaction of two loci. *Theor. Appl. Genet.*, **113**, 1551-1561.

Trethewey, R.N., Geigenberger, P., Riedel, K., Hajirezaei, M.-R., Sonnewald, U., Stitt, M., Riesmeier, J.W. and Willmitzer, L. (1998) Combined expression of glucokinase and invertase in potato tubers leads to a dramatic reduction in starch accumulation and a stimulation of glycolysis. *Plant J.*, **15**, 109-118.

Unger, M., Dreyer, M., Specker, S., Laug, S., Pelzing, M., Neusüss, C., Holzgrabe, U. and Bringmann, G. (2004) Analytical characterisation of crude extracts from an African Ancistrocladus species using high-performance liquid chromatography and capillary electrophoresis coupled to ion trap mass spectrometry. *Phytochem. Anal.*, **15**, 21--26.

Ungerer, M.C. and Rieseberg, L.H. (2003) Genetic architecture of a selection response in Arabidopsis thaliana. *Evolution Int. J. Org. Evolution*, **57**, 2531-2539.

Urbanczyk-Wochniak, E., Baxter, C., Kolbe, A., Kopka, J., Sweetlove, L.J. and Fernie, A.R. (2005) Profiling of diurnal patterns of metabolite and transcript abundance in potato (Solanum tuberosum) leaves. *Planta*, **221**, 891--903.

Urbanczyk-Wochniak, E. and Fernie, A.R. (2005) Metabolic profiling reveals altered nitrogen nutrient regimes have diverse effects on the metabolism of hydroponically-grown tomato (Solanum lycopersicum) plants. *J. Exp. Bot.*, **56**, 309--321.

Urbanczyk-Wochniak, E., Luedemann, A., Kopka, J., Selbig, J., Roessner-Tunali, U., Willmitzer, L. and Fernie, A.R. (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.*, **4**, 989--993.

Utz, H.F. and Melchinger, A.E. (1996) PLABQTL: A program for composite interval mapping of QTL. *J. Quant. Trait Loci*, **2**, (online).

Vos, R.C.H.D., Moco, S., Lommen, A., Keurentjes, J.J.B., 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.

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

Wagner, C., Sefkow, M. and Kopka, J. (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry*, **62**, 887--900.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lipshutz, R., Chee, M. and Lander, E.S. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077--1082.

Wasim, M., Hassan, M.S. and Brereton, R.G. (2003) Evaluation of chemometric methods for determining the number and position of components in high-performance liquid chromatography detected by diode array detector and on-flow 1H nuclear magnetic resonance spectroscopy. *Analyst*, **128**, 1082-1090.

Weckwerth, W. (2003) Metabolomics in systems biology. *Annu. Rev. Plant Biol.*, **54**, 669-689.

Weckwerth, W., Loureiro, M.E., Wenzel, K. and Fiehn, O. (2004) Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. U S A*, **101**, 7809--7814.

Werner, J.D., Borevitz, J.O., Warthmann, N., Trainer, G.T., Ecker, J.R., Chory, J. and Weigel, D. (2005) Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. *Proc. Natl. Acad. Sci. U S A*, **102**, 2460--2465.

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

Williams, W. (1959) Heterosis and the genetics of complex characters. *Nature*, **184**, 527--530.

Wilson, L.M., Whitt, S.R., Ibáñez, A.M., Rocheford, T.R., Goodman, M.M. and Buckler, E.S. (2004) Dissection of maize kernel composition and starch production by candidate gene association. *Plant Cell*, **16**, 2719--2733.

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

Wold, H. (1975) *Soft modelling by latent variables*. London: Academic Press.

Wullschleger, S.D., Yin, T.M., Difazio, S.P., Tschaplinski, T.J., Gunter, L.E. and Davis, M.F. (2005) Phenotypic variation in growth and biomass distribution for two advanced-generation pedigrees of hybrid poplar. *Can. J. For. Res.*, **35**, 1779-1789.

Xiao, J., Li, J., Yuan, L. and Tanksley, S.D. (1995) Dominance Is the Major Genetic Basis of Heterosis in Rice as Revealed by QTL Analysis Using Molecular Markers. *Genetics*, **140**, 745-754.

Yan, J.-b., Tang, H., Huang, Y.-q., Zheng, Y.-l. and Li, J.-s. (2006) Quantitative trait loci mapping and epistatic analysis for grain yield and yield components using molecular markers with an elite maize hybrid. *Euphytica*, **149**, 121--131.

Yu, C.Y., Hu, S.W., Zhao, H.X., Guo, A.G. and Sun, G.L. (2005) Genetic distances revealed by morphological characters, isozymes, proteins and RAPD markers and their relationships with hybrid performance in oilseed rape (Brassica napus L.). *Theor. Appl. Genet.*, **110**, 511--518.

Yu, S.B., Li, J.X., Xu, C.G., Tan, Y.F., Gao, Y.J., Li, X.H., Zhang, Q. and Maroof, M.A.S. (1997) Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid. *Proc. Natl. Acad. Sci. U S A*, **94**, 9226--9231.

Zeng, Z.B. (1994) Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457-1468.

Zhang, Z., Ober, J.A. and Kliebenstein, D.J. (2006) The gene controlling the quantitative trait locus EPITHIOSPECIFIER MODIFIER1 alters glucosinolate hydrolysis and insect resistance in Arabidopsis. *Plant Cell*, **18**, 1524-1536.

Supplemental Information

The supplemental information can be accessed online using the following links for Chapter 3 (http://www.pnas.org/cgi/content/full/0609709104/DC1) and Chapter 4 (http://www3.interscience.wiley.com/journal/119410765/suppinfo).

Sup.Tab.S1 List of significantly correlated metabolites resulting from pairwise correlations (ordered by correlation)

Sup.Tab.S2 List of all relevant metabolites determined by the correlation between them and the canonical variate (ordered by absolute correlation)

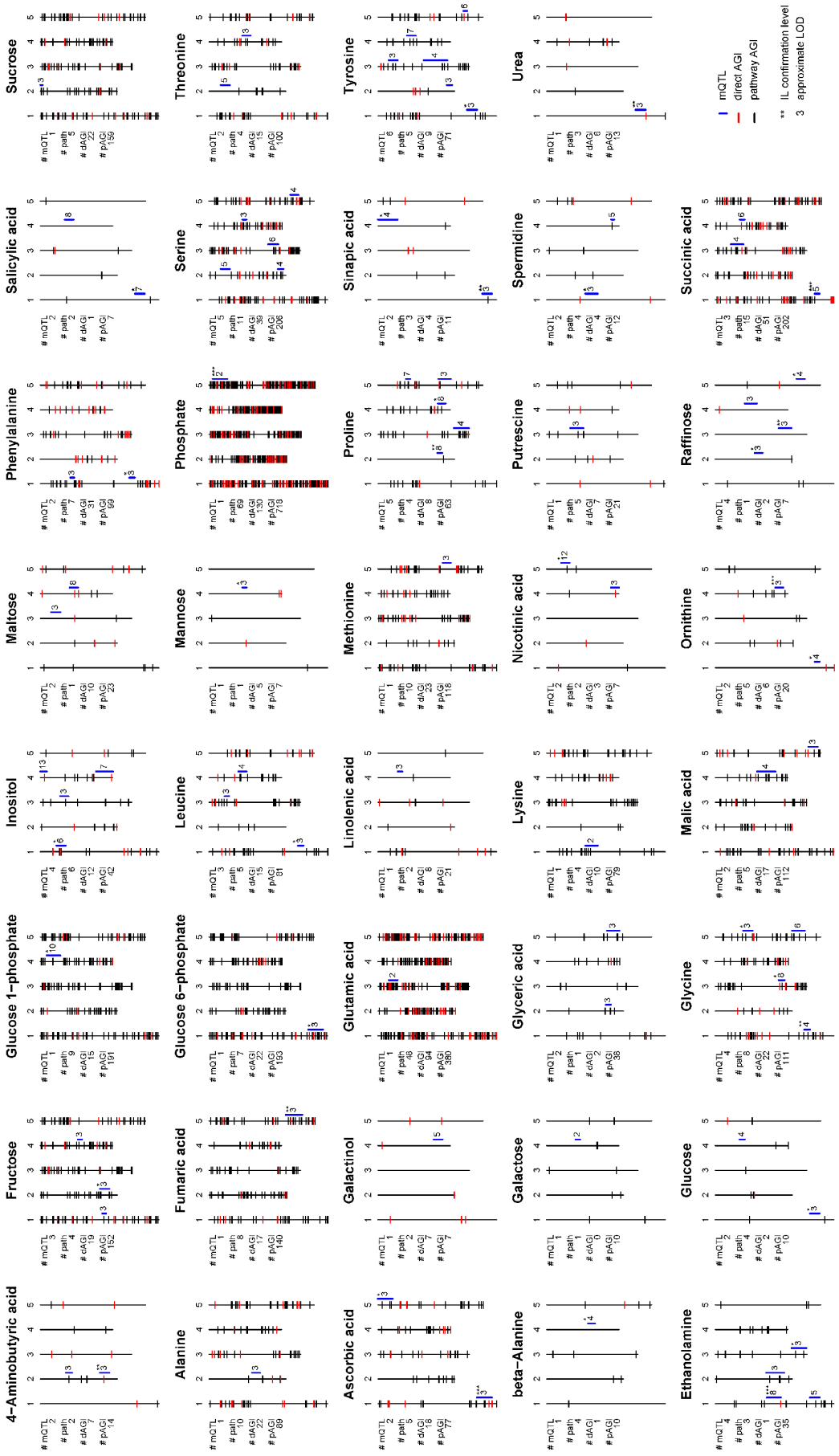
Sup.Tab.S3 List of all 181 metabolic signatures that have been evaluated within this experiment

Sup.Tab.S4 QTL analysis results for biomass and metabolic traits

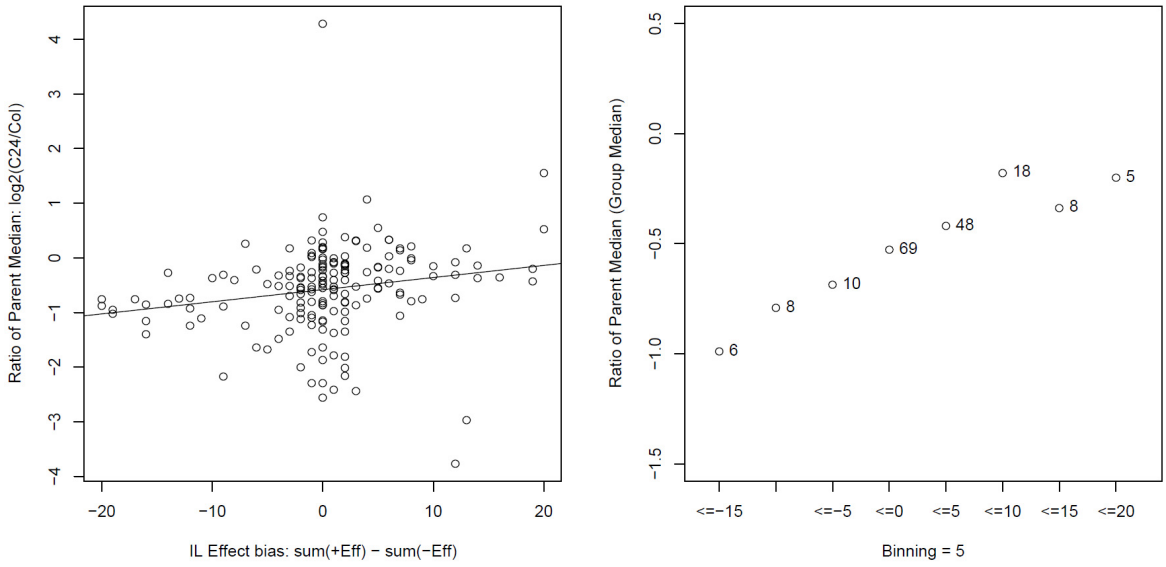
Sup.Tab.S5 SNP positions and resulting amino acid changes for mQTL candidate genes according to data published by Clark *et al.* (2007)

Sup.Tab.S6 Epistatic interactions (additive×additive) for mQTL of 50 known metabolites

As Chapter 5 is not yet published the intended supplemental information will be included here instead of an online source.



Supplemental Figure S 1 Comprehensive mQTL overview for known metabolites



Supplemental Figure S 2 Regression of the parental difference on the effect bias for mode of inheritance classifications (a). Simplification of this plot using a binning approach (b). For clearity only M-Lines are included in the plot to prevent any counfounding with the second bias only present in N-Lines (Supplemental Figure S3)

A	M	N
+r	8	8
−r	0	0

B	M	N
+a	79	291
−a	20	12

C	M	N
+d	63	209
−d	56	28

D	M	N
+d	27	41
−d	82	29

E	M	N
+o	21	30
−o	1	0

F	M	N
+d	216	421
−d	200	62

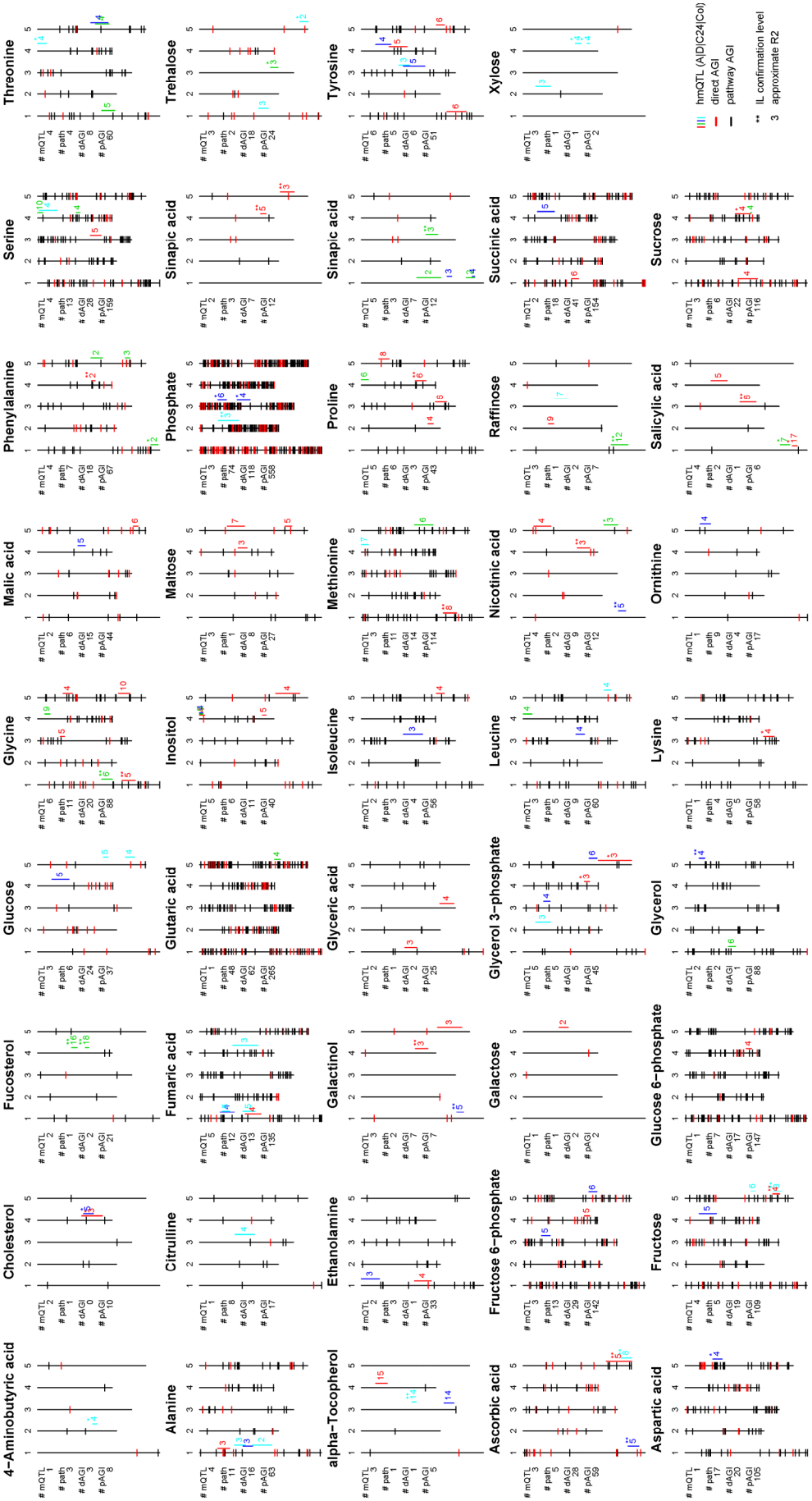
G	M	N
+r	28	31
−r	25	17

H	M	N
+o	9	32
−o	15	4

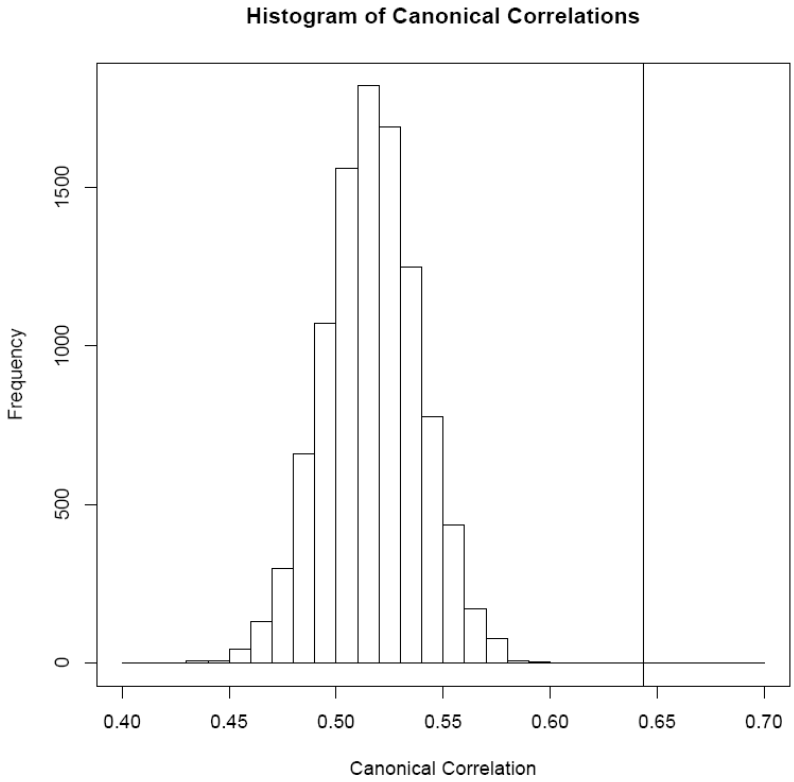
I	M	N
+o	23	27
−o	25	5

J	M	N
+o	7	0
−o	3	0

Supplemental Figure S 3 Mode of inheritance. Depicted are the different branches of the decision tree according to Semel *et al.* (2006). Small letters indicate the classification of the QTL to be recessive (r), additive (a), dominant (d) or over-dominant (o) with ILs being significantly lower (-) or higher (+) than the respective parent. Results are separated by the subpopulationns of M-Lines (C24 with Col-0 introgression) and N-Lines (Col-0 with C24 introgression). A clear bias towards positive effects for N-Lines is present.



Supplemental Figure S 4 Candidate gene search for heterotic metabolic QTL.



Supplemental Figure S 5 Canonical Correlation for 10,000 permuted data sets (Histogram) and Observed Data (Single Line).

Deutsche Zusammenfassung

Pflanzen sind die Primärproduzenten von Biomasse und damit Grundlage allen Lebens. Sie werden nicht nur zur Gewinnung von Nahrungsmitteln, sondern zunehmend auch als Quelle erneuerbarer Energien kultiviert. Aufgrund der Begrenztheit der weltweit zu Verfügung stehenden Anbaufläche ist eine zielgerichtete Selektion und Verbesserung der verwendeten Sorten unabdingbar. Um solch eine kontinuierliche Verbesserung zu gewährleisten, ist ein grundlegendes Verständnis des biologischen Systems Pflanze nötig.

Diese Arbeit hatte zum Ziel, den Primärmetabolismus der Modellpflanze *A. thaliana* mit Methoden der quantitativen Genetik zu untersuchen und in Beziehung zu Wachstum und Biomasse zu stellen. Insbesondere sollte Heterosis, die Abweichung von Hybriden in ihren Merkmalen vom Mittelwert der Eltern, auf Stoffwechselebene charakterisiert werden. Mit Hilfe der Gas Chromatographie/ Massen Spektrometrie (GC-MS) wurden über 2000 Proben von rekombinanten Inzucht Linien (RIL) und Introgressions Linien (IL) der Akzessionen Col-0 und C24 bezüglich des Vorkommens von 181 Metaboliten untersucht. Die beobachtete Varianz erlaubte die Bestimmung von 157 metabolischen QTL (mQTL), genetischen Regionen, die für die Metabolitkonzentrationen relevante Gene enthalten. Durch die Untersuchung von Testkreuzungen der RILs und ILs konnten weiterhin 385 heterotische metabolische QTL (hmQTL) identifiziert werden.

Im Rahmen dieser Arbeit wurde eine robuste Methode zur Auswertung von GC-MS Analysen entwickelt. Es wurde eine hoch signifikante kanonische Korrelation (r=0.73) zwischen Biomasse und Metabolitprofilen gefunden. Die unterschiedlichen Ansätze zur QTL Analyse, RILs und ILs, wurden verglichen. Dabei konnte gezeigt werden, daß die Methoden komplementär sind, da mit RILs gefundene mQTL zu 56% und hmQTL zu 23% in ILs bestätigt wurden. Durch den Vergleich mit Datenbanken wurden für 67% der mQTL Kandidatengene identifiziert. Um diese zu überprüfen wurden acht dieser Gene resequenziert und insgesamt 23 Polymorphismen darin bestimmt. Die Heterosis in den Hybriden ist für die meisten Metabolite gering (<20%). Für hmQTL konnten weniger Kandidatengene als für mQTL bestimmt werden und sie zeigten eine geringere Übereinstimmung in den beiden Populationen. Dies deutet darauf hin, daß regulatorische Loci und epistatische Effekte einen wichtigen Beitrag zur Heterosis besteuern.

Die gewonnenen Daten stellen eine reiche Quelle für die weitergehende Untersuchung und Annotation relevanter Gene dar und ebnen den Weg für ein besseres Verständnis des Systems Pflanze.