Genetical Genomics in Arabidopsis:

from natural variation to regulatory networks

Joost Keurentjes

Promotoren:

Prof. dr. ir. M. Koornneef
Persoonlijk hoogleraar bij het Laboratorium voor Erfelijkheidsleer, Wageningen Universiteit

Prof. dr. L.H.W. van der Plas
Hoogleraar in de Plantenfysiologie, Wageningen Universiteit

Copromotor:

Dr. D. Vreugdenhil
Universitair hoofddocent bij het Laboratorium voor Plantenfysiologie, Wageningen Universiteit

Promotiecommissie:

Prof. dr. W.J. Stiekema, Wageningen Universiteit
Prof. dr. R.G.F. Visser, Wageningen Universiteit
Prof. dr. J.C.M. Smeekens, Universiteit Utrecht
Prof. dr. M. Stitt, Max Planck Instituut voor Moleculaire Plantenfysiologie, Golm, Duitsland

Dit onderzoek is uitgevoerd binnen de onderzoekschool voor Experimentele Plantwetenschappen.

Joost Keurentjes

Genetical Genomics in Arabidopsis: from natural variation to regulatory networks

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit,
Prof. dr. M.J. Kropff,
in het openbaar te verdedigen
op vrijdag 7 september 2007
des namiddags te 13:30 uur in de Aula

Joost J.B. Keurentjes

Genetical Genomics in Arabidopsis: from natural variation to regulatory networks (2007)

PhD thesis, Wageningen University, Wageningen, The Netherlands.
With references – with summaries in English and Dutch.

ISBN 978-90-8504-704-9

Contents

Chapter 1

General introduction

Natural variation and quantitative traits

For most organisms variation between individuals can be observed in nature. Plants are no exception to this and naturally occurring variation can be observed between and within species. Although part of the within-species variation observed in nature can be attributed to environmental influences, genetic variation can be observed when plants of different origins are grown together in the same environment (Nordborg *et al.*, 2005). The contribution of genetic factors to the totally observed variation between different genotypes is often expressed as the heritability of a trait.

Natural variation exhibited by genotypically different accessions can be classified as qualitative or quantitative. Qualitative traits are characterized by distinct phenotypic classes, *e.g.* presence or absence of a property, often a result from genetic differences at single genes. Such traits can relatively easily be dissected genetically due to their clear segregation pattern in the progeny of crosses. Quantitative traits on the other hand, often display a more continuous variation in phenotypes due to a multiplicity of genes involved and a relatively large effect of environmental factors on the expression of the trait. Because different genes can contribute positively or negatively to a quantifiable trait, recombination of genes results in a large number of phenotypic classes which can not unambiguously be associated with genotypic classes (Kearsey *et al.*, 2003; Weigel and Nordborg, 2005; Holland, 2007). The complexity of quantitative traits is further enhanced by the presence of epistatic interactions and interactions between genes and the environment (Carlborg and Haley, 2004; Kroymann and Mitchell-Olds, 2005).

Although much more difficult to dissect, quantitative variation is found for many agronomical important traits like biomass formation, plant height, flowering time, reproductive yield and seed dormancy (Koornneef *et al.*, 2004; Ross-Ibarra, 2005; Ashikari and Matsuoka, 2006; Semel *et al.*, 2006; Zhao *et al.*, 2006). Furthermore, quantitative natural variation controls adaptive strategies to cope with biotic and abiotic influences and its understanding can provide insight in ecological mechanisms and the evolutionary history of plants (Tonsor *et al.*, 2005; Mitchell-Olds and Schmitt, 2006).

Arabidopsis thaliana as a model plant

The study of quantitative traits is often contrasted with the analysis of qualitative traits, which are mostly represented by single gene mutants or single gene natural variants. For the study of such single genes *Arabidopsis thaliana* has proven to be a very efficient model plant because of a number of biological properties that make genetic analyses very efficient (Somerville and Koornneef, 2002). Although it is self-fertilizing it can easily be out-crossed and it combines short generation times with high reproductive yield. Moreover, it contains a fully sequenced small genome (120 Mbp) made up of only five chromosomes and approximately 30,000 genes (The Arabidopsis Genome Initiative, 2000). The accumulation of knowledge, biological resources and available molecular tools adds up to the attractiveness of Arabidopsis as a model system (Alonso and Ecker, 2006).

These advantages also make Arabidopsis very suitable for the genetic analysis of natural variation. The plant shows a broad global distribution throughout the northern hemisphere at different continents, including America, Africa, Europe and Asia (Schmid *et al.*, 2006). Moreover it is found at different latitudes and altitudes ranging from Scandinavian sea level to high up in the Asian Himalayas. At many locations, accessions or ecotypes, have been collected displaying a broad spectrum of natural variation for numerous traits (Alonso-Blanco and Koornneef, 2000; Koornneef *et al.*, 2004). Many of those accessions are deposited to stock centers making them publicly available for genetic analyses.

Genetic analysis of quantitative traits

Despite the complexity in genetic regulation of quantitative traits much progress has been made over the past decades in dissecting these traits by the use of molecular markers. The increasing ease by which molecular markers can be generated (Borevitz and Chory, 2004) in combination with the application of sophisticated mapping methods (Jansen, 1993) has led to a strong interest in the use of natural variation for studying quantitative traits (Slate, 2005). Mutant screens, often directed to a specific trait, and the subsequent mapping and cloning of the affected gene, have been a very effective strategy to analyze the function of genes in Arabidopsis (Meinke *et al.*, 2003). However, specific advantages are associated with the study of multiple natural perturbations in the same mapping population. This allows for the analysis of an almost infinite number of traits (Doerge, 2002). For this type of study so-called immortal mapping populations, consisting in most cases of homozygous genotypes that can be tested in replicates and in different experiments, have proven very useful.

Although various types of such mapping populations have been developed for a variety of species (Eshed and Zamir, 1995; Rae *et al.*, 1999; Yoon *et* *al.*, 2006), the relative ease of generating recombinant inbred lines (RILs) has led to their favorable use for quantitative trait locus (QTL) analysis in Arabidopsis and many other plants (Jansen, 2003b). RILs are produced by crossing two distinct genotypes and using single seed descent propagation of the inbred lines obtained by selfing a random set of F_2 individuals. While the accuracy of QTL mapping depends on statistical factors such as the size of the mapping population, it has been shown to be quite accurate in many cases (Price, 2006). However, there is often still a need for confirmation and further fine mapping (Paran and Zamir, 2003; Weigel and Nordborg, 2005). For these aspects, which are the basis of the cloning of genes underlying QTLs, near isogenic lines (NILs) are often used to isolate a QTL. A set of NILs consists of lines with identical genetic background but differing in genotype at the position of a limited number of loci. NILs are generally constructed by introgressing a donor accession into the genetic background of another accession by crossing and repeated back-crossing with the recurrent accession. NILs allow studying the effect of Mendelized QTLs and can refine the position of a QTL by varying position and size of introgressions.

Despite the fact that RIL populations have been developed for an increasing number of different genotypes the development of NILs has lagged behind. Upon the detection of a QTL, NILs are often not available for the confirmation and finemapping of those QTLs. Valuable time is often lost in developing NILs before the necessary follow-up experiments can be continued. The Landsberg *erecta* (L*er*) x Cape Verde Islands (Cvi) RIL population (Alonso-Blanco *et al.*, 1998b) is one of the most frequently used populations in quantitative genetics and several NILs have been developed at distinct loci for these genotypes (Alonso-Blanco *et al.*, 1998a; Swarup *et al.*, 1999; Alonso-Blanco *et al.*, 2003; Bentsink *et al.*, 2003; Edwards *et al.*, 2005; Juenger *et al.*, 2005; Teng *et al.*, 2005). However, most of these NILs were developed after the detection of QTLs in the RIL population and these studies could have benefited much from the direct availability of NILs. To increase the efficiency from the mapping of quantitative traits to the actual cloning of the causal genes it would therefore be advantageous to have a NIL at every possible genomic location at one's disposal. Moreover, collections of NILs with genome-wide coverage can serve as mapping populations, which differ in effectiveness from RILs, mainly because the complexity of epistasis is strongly reduced (Eshed and Zamir, 1995).

Genetical genomics: variation in genome sequence and expression

In Arabidopsis as well as in other species, genome-wide analyses of genomic polymorphisms in a large collection of accessions have revealed extensive sequence variation (Borevitz *et al.*, 2003; Han and Xue, 2003; Schmid *et al.*, 2003;

Nordborg *et al.*, 2005; Vigouroux *et al.*, 2005). Polymorphisms, when converted to molecular markers, are indispensable for (fine) mapping of quantitative traits in experimental populations. When surveyed in natural populations at high density, polymorphisms even enable high resolution mapping through linkage disequilibrium (Remington *et al.*, 2001; Nordborg *et al.*, 2002; Aranzana *et al.*, 2005; Kim *et al.*, 2006). The best marker, however, is the polymorphism causal for the observed variation. By definition natural genetic variation is a result of genomic differences and therefore the extent of variation in quantitative traits is largely dependent on the level of DNA sequence variation. Although many of the polymorphisms will be neutral, it leaves little doubt that the study of quantitative traits can benefit enormously from genomic analyses (Borevitz and Nordborg, 2003; Maloof, 2003; Gilad and Borevitz, 2006). Non-synonymous polymorphisms in coding sequences of genes might alter protein function or stability, introducing phenotypic variation. Polymorphisms in regulatory sequences on the other hand might result in differences in transcriptional efficiency of genes. It is conceivable that expression differences, or variation in mRNA stability caused by coding sequence polymorphisms, contribute heavily to natural variation in Arabidopsis (Chen *et al.*, 2005). Given the extensive variation in phenotype and genomic sequence within Arabidopsis, it is therefore not surprising that for many genes expression differences can be observed between accessions (Vuylsteke *et al.*, 2005; Kliebenstein *et al.*, 2006a; West *et al.*, 2006).

The genetic regulation of natural variation in gene expression is presumably not different from any other 'classical' quantitative trait. Therefore, gene expression can be treated like any other quantitative trait, on which all statistical tools of quantitative genetics can be applied. However, the effect of this variation may be reflected at the phenotypic level, thereby explaining the genetic component of natural phenotypic variation. This combination of linkage analysis (genetics) and expression profiling (genomics) was coined 'genetical genomics' (Jansen and Nap, 2001) and experiments were first reported in yeast (Brem *et al.*, 2002), soon followed by data of higher eukaryotes (Schadt *et al.*, 2003). Because of the available high quality mapping populations and the commercially available genome-wide microarrays, Arabidopsis is ideally suited for these kinds of analyses. However, upon publication of the first genetical genomics studies no genome-wide data for Arabidopsis were available yet and only recently a number of studies in various RIL populations have indicated extensive genetic regulation of gene expression (DeCook *et al.*, 2006; Vuylsteke *et al.*, 2006; Keurentjes *et al.*, 2007; West *et al.*, 2007).

Genome-wide expression analysis of fully sequenced genomes, like Arabidopsis, offers the unique possibility to compare genomic positions of genes with the map position(s) of their detected expression QTL(s) (eQTL). Such comparative analyses reveal either local or distant regulation of gene expression. Local regulatory variation is observed when genes and their respective eQTLs co-locate and distant regulatory variation is observed when genes and their respective eQTLs are positionally separated on the genome (Rockman and Kruglyak, 2006). Local regulatory variation is often a result of polymorphisms within the gene for which the eQTL was observed. When such polymorphisms reside in *cis*-acting regulatory elements this might affect transcriptional activity. Regulation in *cis* could also act post-transcriptionally by altering mRNA stability when polymorphisms reside in coding sequences of the gene. However, polymorphisms within the gene itself might also act in *trans* by altering auto-regulation and feedback loops. Furthermore, occasionally local regulatory variation might act in *trans* due to polymorphisms in a tightly linked gene that regulates the gene for which the eQTL was detected. To determine whether local regulatory variation acts in *cis* or *trans* further experimentation, like allele specific expression analysis, is necessary (Ronald *et al.*, 2005; Zhang *et al.*, 2007). Distant regulatory variation most likely acts in *trans* when polymorphisms in another gene (*e.g.* a transcription factor) affect transcription of the gene for which the distant eQTL was detected. Nonetheless, other mechanisms of distant regulation, both in *cis* and *trans*, are imaginable (Rockman and Kruglyak, 2006).

Genetic regulation of plant metabolic content

The impact of gene expression variation on quantitative traits is now widely acknowledged and the use of high throughput genomic analyses has become an important tool in genetic analyses of natural variation (Gibson and Weir, 2005). Transcription however, is only a first link in the chain from genotype to phenotype and successive entities like proteins and metabolites (quality and quantity) are expected as causal sources for natural phenotypic variation but have been largely under-exploited. Yet, high-throughput technologies, *i.e.* proteomics and metabolomics, have shown that much variation is observed upon physiological perturbation and between genetic variants (Fiehn *et al.*, 2000; Chevalier *et al.*, 2004). Moreover, small-scale targeted analyses and subsequent QTL analysis revealed strong genetic regulation in a number of studies (Kliebenstein *et al.*, 2001; Consoli *et al.*, 2002).

Analogous to genetical genomics, the combination of high-throughput proteomics and metabolomics and multifactorial genetic analyses would therefore allow studying the functional consequences of natural genetic variation at a much larger scale (Jansen, 2003a). However, full-scale analyses for proteins and metabolites, equivalent to genome-wide expression analysis, are not available yet.

This is mainly because proteins and metabolites are much more diverse in their properties than nucleic acids, making it difficult to extract and analyze all different classes using a single protocol. Even based on a fully sequenced genome one cannot predict all protein variants and metabolites that a plant can contain. Moreover, the dynamic range of protein and metabolite abundance is far greater than for nucleic acids and no amplification techniques are available for these entities, making sample volume and detection range (sensitivity *vs.* saturation) critical limitations. Nevertheless, several complementing high-throughput technologies have been developed covering together a large part of the proteome (Peck, 2005) and metabolome (Ward *et al.*, 2003; Lisec *et al.*, 2006; De Vos *et al.*, 2007).

The progress made in proteomics and metabolomics now also enables the large-scale genetic analysis of these entities, which has only recently be demonstrated for primary metabolites (Schauer *et al.*, 2006). However, variation in secondary metabolism is probably more extensive and determines much of the phenotypic variation that can be observed. Plants are especially rich in the number of secondary metabolites, possibly as a consequence of their sessile nature. Since plants are unable to move away from biotic and abiotic threats they have adapted to cope with many environmental influences. In Arabidopsis alone already hundreds of secondary metabolites representing numerous chemical classes have been discovered (D'Auria and Gershenzon, 2005). Given the wide global distribution range of Arabidopsis and the diverge range of sites plants have been collected, it is conceivable that metabolites play an important role in local adaptation strategies. It is therefore likely that the high level of natural variation in Arabidopsis is also reflected in metabolite composition and content (Fiehn, 2002).

A large drawback of metabolomic analyses is the lack of compound identification. Unlike microarrays, where each signal can be reduced to a specific gene, most large-scale metabolomic techniques are untargeted. The output of a metabolic sample analysis typically consists of a complex chromatogram of many, often anonymous peaks, where compounds can be represented by multiple peaks depending on adduct formation, fragmentation and isotopes. For genetic analysis it is essential that chromatograms of different genotypes are qualitatively comparable. This alignment problem can be solved by adding reference compounds, standardization and proper alignment software (Lisec *et al.*, 2006; De Vos *et al.*, 2007). Although each peak represents a specific chemical compound, the order, retention time and intensity of peaks can differ substantially depending on analytical differences and sample properties. Such inconsistencies in data output make it difficult to compare analyses performed in different labs or experiments. Although some efforts have been made in constructing identification libraries

(Schauer *et al.*, 2005; Moco *et al.*, 2006; Ward *et al.*, 2007), such libraries do not cover entirely the still expanding number of detected compounds. Moreover, the different methodologies applied in various labs make it difficult to implement such libraries. The scientific community would therefore benefit much from a commonly adopted standard for metabolomic analyses (Jenkins *et al.*, 2004).

Regulatory network construction

To functionally link the large data sets obtained in 'omic' experiments as an order of events that ultimately result in a specific phenotype, network construction provides a useful tool. Biological networks describe relationships between individual components of a biological process (Barabasi and Oltvai, 2004). Such components can either be genes, proteins, metabolites or a combination thereof. Depending on the data source, networks can be constructed in various ways but all of them serve to elucidate the, often complex, regulation of biological processes.

A special type of networks does not rely on experimental data but rather predicts *in silico* connections based on genome-wide sequence information. Most notably are genome-scale metabolic connectivity networks, where metabolites are connected when the genome contains a gene encoding an enzyme able to catalyze the conversion of one of the metabolites into the other (Jeong *et al.*, 2000). However, genetic networks have also been predicted *in silico* by analyzing regulatory elements of genes for binding sites of known transcription factors (Palaniswamy *et al.*, 2006). Although powerful in hypothesis formation such studies require empirical data for confirmation of predicted pathways and interactions. Therefore, many approaches for network construction are based on experimental data, which also allows the identification of relationships unable to be predicted from genomic information only. Protein-protein interactions for instance, are difficult to deduce from sequence information but require immuno precipitation or two-hybrid screens. Similar analyses, like chromatine immuno precipitation (ChIP-chip), can also be used to identify and confirm transcriptional regulation of target genes by transcription factors or other known regulators (Lee *et al.*, 2002). In yeast, much progress in regulatory network construction has been made by expression and metabolic profiling of deletion strains (Forster *et al.*, 2002; Hu *et al.*, 2007) and genetic interaction analyses of double mutants (synthetic lethals) (Tong *et al.*, 2004). However, for most higher eukaryotes such genome-wide analyses are not realistic because of the much higher gene number, the presumably more complex genetic architecture, and aspects of sub-cellular and tissue specific compartmentation. Many attempts in regulatory network construction therefore rely on more indirect approaches of establishing associations between network components.

A straightforward approach is correlation analysis over a large set of data compiled from numerous perturbation experiments (de la Fuente *et al.*, 2004). Exemplary are the widely applied gene co-expression analyses, where correlation in gene expression patterns is surveyed under a large number of diverse conditions (Stuart *et al.*, 2003; Gachon *et al.*, 2005). The rationale for this kind of analysis is that genes participating in the same biological process are often co-regulated and hence exhibit similar expression patterns. Following the same line of reasoning, metabolic correlation networks have been constructed (Steuer *et al.*, 2003). However, correlation does not necessarily imply functional relatedness nor does it address causality issues. The reliability of, and information contained in constructed networks would therefore gain much strength from integrated analyses of interdisciplinary approaches (Fiehn *et al.*, 2001; Winnacker, 2003). Such integrated studies can either combine experimental data with *in silico* analyses (Segal *et al.*, 2003) or benefit from multi-parallel analyses of diverse biological samples (Urbanczyk-Wochniak *et al.*, 2003; Hirai *et al.*, 2005; Joosen *et al.*, 2007).

Although demonstrably effective, correlation analyses depend on large compendia of publicly available data or suffer from the limited number of physiological conditions that can be analyzed in dedicated experiments. However, sometimes co-regulation is displayed only in particular conditions (Gachon *et al.*, 2005) which may even remain undiscovered in large data sets due to dilution effects. The largest drawback of correlation analyses, however, is that no information can be retrieved about the nature of the underlying genetic regulation. Correlation may be a result from co-regulation by a common regulator or due to independent pathways that occur in parallel, possibly due to developmental or spatial control. A highly correlated cluster of biological elements, such as genes, proteins and metabolites, can also result from downstream effects of the regulation of a single member but no information about cause and consequence can be extracted from genetic correlations.

Mapping populations combine a high number of genetic perturbations by which numerous quantitative traits can segregate in a single experiment. Moreover, genetic analysis offers the unique possibility of identifying genomic loci causal for observed variation in, and possible correlation between traits. When applied to genome-wide expression analysis or other large-scale 'omic' analyses this therefore allows the identification of true gene-to-gene or gene-to-function regulation. Unfortunately, mapping resolution is often not high enough to identify directly causal genes underlying detected QTLs and will require further analysis such as fine mapping, the study of overexpressors and mutants of candidate genes, etc. However, *cis*-regulated genes are obvious candidates and co-regulated traits can effectively be identified through co-location of detected QTLs. Still, not all coinciding QTLs necessarily represent the same causal gene because effects of closely linked genes are difficult to distinguish from true pleiotropic effects of a single gene. Without further experimentation genetic interactions can be predicted computationally by comparing QTL profiles and correlation analyses (Zhu *et al.*, 2004; Bing and Hoeschele, 2005; Li *et al.*, 2005; Lan *et al.*, 2006; Fu *et al.*, 2007). However, the accuracy of constructed networks can benefit tremendously from the integration of additional information like gene ontology (Kliebenstein *et al.*, 2006b; Keurentjes *et al.*, 2007), sequence data (Hitzemann *et al.*, 2003) and related quantitative trait data (Consoli *et al.*, 2002; Hubner *et al.*, 2005).

Although much progress has been made in the construction of regulatory networks, any information inferred from such networks should be interpreted with caution. Where many studies have shown the identification of correct interactions, most approaches can not exclude the assignment of false positives. Predicted interactions and regulatory steps should therefore be considered as hypothesis formation only and confirmation of such relationships should come from additional experimentation.

Scope of the thesis

In Arabidopsis natural variation exists for many quantitative traits. The genetic regulation of quantitative traits can effectively be analyzed in mapping populations by way of quantitative trait locus (QTL) analyses. This thesis describes the large-scale genetic analysis of 'omics' data and their use in dissecting the genetic regulation of quantitative traits.

Chapter two describes the development of a near isogenic line (NIL) population and its use in mapping and fine-mapping of QTLs. NILs are widely used in the confirmation of QTLs, detected in recombinant inbred line (RIL) populations. However, when a population of NILs with genome-wide coverage is available, such a population can also be used for mapping purposes. A genome-wide NIL population was generated by introgressing genomic regions of an accession from the Cape Verde Islands (Cvi) into the genetic background of the commonly used laboratory accession Landsberg *erecta* (L*er*). Mapping power and resolution of this population was compared with the previously developed L*er* x Cvi RIL population.

Chapter three describes the genome-wide expression analysis of the L*er* x Cvi RIL population. Similar to 'classical' quantitative traits, natural variation also exists for expression levels of many genes. QTL mapping of expression variation therefore reveals genomic loci controlling the expression of genes. This information can then be used to construct genetic regulatory networks and help elucidating the genetic control of many physiological traits.

Chapter four describes the large-scale untargeted metabolomic analyses in the L*er* x Cvi RIL population. Subsequent mapping revealed substantial genetic control for metabolite composition and content. Identification of anonymous mass peaks enabled the reconstruction of metabolic pathways and revealed novel biosynthetic steps.

Chapter five describes the integrated analysis of gene expression, enzyme activities and metabolite content in primary carbohydrate metabolism. QTL and correlation analyses identified different modes of control of primary carbohydrate metabolism, including regulation of structural gene expression and metabolic control.

Finally, in chapter six, the work described in this thesis is summarized and discussed.

REFERENCES

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

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., Bentsink, L., Hanhart, C.J., Blankestijn-de Vries, H. and Koornneef, M. (2003). Analysis of natural allelic variation at seed dormancy loci of *Arabidopsis thaliana*. *Genetics* **164,** 711-729.

Alonso, J.M. and Ecker, J.R. (2006). Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in Arabidopsis. *Nat Rev Genet* **7,** 524-536.

Aranzana, M.J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C. *et al.* (2005). Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genet* **1,** e60.

Ashikari, M. and Matsuoka, M. (2006). Identification, isolation and pyramiding of quantitative trait loci for rice breeding. *Trends Plant Sci* **11,** 344-350.

Barabasi, A.L. and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5,** 101-113.

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.

Bing, N. and Hoeschele, I. (2005). Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* **170,** 533-542.

Borevitz, J.O., Liang, D., Plouffe, D., Chang, H.S., Zhu, T., Weigel, D., Berry, C.C., Winzeler, E. and Chory, J. (2003). Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res* **13,** 513-523.

Borevitz, J.O. and Nordborg, M. (2003). The impact of genomics on the study of natural variation in Arabidopsis. *Plant Physiol* **132,** 718-725.

Borevitz, J.O. and Chory, J. (2004). Genomics tools for QTL analysis and gene discovery. *Curr Opin Plant Biol* **7,** 132-136.

Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* **296,** 752-755.

Carlborg, O. and Haley, C.S. (2004). Epistasis: too often neglected in complex trait studies? *Nat Rev Genet* **5,** 618-625.

Chen, W.J., Chang, S.H., Hudson, M.E., Kwan, W.K., Li, J., Estes, B., Knoll, D., Shi, L. and Zhu, T. (2005). Contribution of transcriptional regulation to natural variations in Arabidopsis. *Genome Biol* **6,** R32.

Chevalier, F., Martin, O., Rofidal, V., Devauchelle, A.D., Barteau, S., Sommerer, N. and Rossignol, M. (2004). Proteomic investigation of natural variation between Arabidopsis ecotypes. *Proteomics* **4,** 1372-1381.

Consoli, L., Lefevre, A., Zivy, M., de Vienne, D. and Damerval, C. (2002). QTL analysis of proteome and transcriptome variations for dissecting the genetic architecture of complex traits in maize. *Plant Mol Biol* **48,** 575-581.

D'Auria, J.C. and Gershenzon, J. (2005). The secondary metabolism of *Arabidopsis thaliana*: growing like a weed. *Curr Opin Plant Biol* **8,** 308-316.

de la Fuente, A., Bing, N., Hoeschele, I. and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20,** 3565-3574.

De Vos, R.C., 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.

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

Doerge, R.W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet* **3,** 43-52.

Edwards, K.D., Lynn, J.R., Gyula, P., Nagy, F. and Millar, A.J. (2005). Natural allelic variation in the temperature-compensation mechanisms of the *Arabidopsis thaliana* circadian clock. *Genetics* **170,** 387-400.

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.

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

Fiehn, O., Kloska, S. and Altmann, T. (2001). Integrated studies on plant biology using multiparallel techniques. *Curr Opin Biotechnol* **12,** 82-86.

Fiehn, O. (2002). Metabolomics--the link between genotypes and phenotypes. *Plant Mol Biol* **48,** 155-171.

Forster, J., Gombert, A.K. and Nielsen, J. (2002). A functional genomics approach using metabolomics and in silico pathway analysis. *Biotechnol Bioeng* **79,** 703-712.

Fu, J., Swertz, M.A., Keurentjes, J.J.B. and Jansen, R.C. (2007). MetaNetwork: a computational protocol for the genetic study of metabolic networks. *Nat Protoc* **2,** 685-694.

Gachon, C.M., Langlois-Meurinne, M., Henry, Y. and Saindrenan, P. (2005). Transcriptional co-regulation of secondary metabolism enzymes in Arabidopsis: functional and evolutionary implications. *Plant Mol Biol* **58,** 229-245.

Gibson, G. and Weir, B. (2005). The quantitative genetics of transcription. *Trends Genet* **21,** 616-623.

Gilad, Y. and Borevitz, J. (2006). Using DNA microarrays to study natural variation. *Curr Opin Genet Dev* **16,** 553-558.

Han, B. and Xue, Y. (2003). Genome-wide intraspecific DNA-sequence variations in rice. *Curr Opin Plant Biol* **6,** 134-138.

Hirai, M.Y., Klein, M., Fujikawa, Y., Yano, M., Goodenowe, D.B., Yamazaki, Y., Kanaya, S., Nakamura, Y., Kitayama, M., Suzuki, H. *et al.* (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.

Hitzemann, R., Malmanger, B., Reed, C., Lawler, M., Hitzemann, B., Coulombe, S., Buck, K., Rademacher, B., Walter, N., Polyakov, Y. *et al.* (2003). A strategy for the integration of QTL, gene expression, and sequence analyses. *Mamm Genome* **14,** 733-747.

Holland, J.B. (2007). Genetic architecture of complex traits in plants. *Curr Opin Plant Biol* **10,** 156-161.

Hu, Z., Killion, P.J. and Iyer, V.R. (2007). Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet* **39,** 683-687.

Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V. *et al.* (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet* **37,** 243-253.

Jansen, R.C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135,** 205-211.

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

Jansen, R.C. (2003a). Studying complex biological systems using multifactorial perturbation. *Nat Rev Genet* **4,** 145-151.

Jansen, R.C. (2003b). Quantitative trait loci in inbred lines. In Handbook of Statistical Genetics, D.J. Balding, M. Bishop and C. Cannings, eds (Chichester, UK: John Wiley & Sons), pp. 445-476.

Jenkins, H., Hardy, N., Beckmann, M., Draper, J., Smith, A.R., Taylor, J., Fiehn, O., Goodacre, R., Bino, R.J., Hall, R. *et al.* (2004). A proposed framework for the description of plant metabolomics experiments and their results. *Nat Biotechnol* **22,** 1601-1606.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000). The large-scale organization of metabolic networks. *Nature* **407,** 651-654.

Joosen, R., Cordewener, J., Supena, E.D., Vorst, O., Lammers, M., Maliepaard, C., Zeilmaker, T., Miki, B., America, T., Custers, J. *et al.* (2007). Combined transcriptome and proteome analysis identifies pathways and markers associated with the establishment of *Brassica napus* microspore-derived embryo development. *Plant Physiol* **144,** 155-172.

Juenger, T.E., McKay, J.K., Hausmann, N., Keurentjes, J.J.B., Sen, S., Stowe, K.A., Dawson, T.E., Simms, E.L. and Richards, J.H. (2005). Identification and characterization of QTL underlying whole-plant physiology in *Arabidopsis thaliana*: delta13C, stomatal conductance and transpiration efficiency. *Plant Cell Environ* **28,** 697-708.

Kearsey, M.J., Pooni, H.S. and Syed, N.H. (2003). Genetics of quantitative traits in *Arabidopsis thaliana*. *Heredity* **91,** 456-464.

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

Kim, S., Zhao, K., Jiang, R., Molitor, J., Borevitz, J.O., Nordborg, M. and Marjoram, P. (2006). Association mapping with single-feature polymorphisms. *Genetics* **173,** 1125-1133.

Kliebenstein, D.J., Kroymann, J., Brown, P., Figuth, A., Pedersen, D., Gershenzon, J. and Mitchell-Olds, T. (2001). Genetic control of natural variation in Arabidopsis glucosinolate accumulation. *Plant Physiol* **126,** 811-825.

Kliebenstein, D.J., West, M.A., van Leeuwen, H., Kim, K., Doerge, R.W., Michelmore, R.W. and St Clair, D.A. (2006a). Genomic survey of gene expression diversity in *Arabidopsis thaliana*. *Genetics* **172,** 1179-1189.

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

Koornneef, M., Alonso-Blanco, C. and Vreugdenhil, D. (2004). Naturally occurring genetic variation in *Arabidopsis Thaliana*. *Annu Rev Plant Physiol Plant Mol Biol* **55,** 141-172.

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

Lan, H., Chen, M., Flowers, J.B., Yandell, B.S., Stapleton, D.S., Mata, C.M., Mui, E.T., Flowers, M.T., Schueler, K.L., Manly, K.F. *et al.* (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genet* **2,** e6.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I. *et al.* (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298,** 799-804.

Li, H., Lu, L., Manly, K.F., Chesler, E.J., Bao, L., Wang, J., Zhou, M., Williams, R.W. and Cui, Y. (2005). Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Hum Mol Genet* **14,** 1119-1125.

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.

Maloof, J.N. (2003). Genomic approaches to analyzing natural variation in *Arabidopsis thaliana*. *Curr Opin Genet Dev* **13,** 576-582.

Meinke, D.W., Meinke, L.K., Showalter, T.C., Schissel, A.M., Mueller, L.A. and Tzafrir, I. (2003). A sequence-based map of Arabidopsis genes with mutant phenotypes. *Plant Physiol* **131,** 409-418.

Mitchell-Olds, T. and Schmitt, J. (2006). Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature* **441,** 947-952.

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

Nordborg, M., Borevitz, J.O., Bergelson, J., Berry, C.C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J.N., Noyes, T., Oefner, P.J. ***et al.*** (2002). The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat Genet* **30,** 190-193.

Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R. ***et al.*** (2005). The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol* **3,** e196.

Palaniswamy, S.K., James, S., Sun, H., Lamb, R.S., Davuluri, R.V. and Grotewold, E. (2006). AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol* **140,** 818-829.

Paran, I. and Zamir, D. (2003). Quantitative traits in plants: beyond the QTL. *Trends Genet* **19,** 303-306.

Peck, S.C. (2005). Update on proteomics in Arabidopsis. Where do we go from here? *Plant Physiol* **138,** 591-599.

Price, A.H. (2006). Believe it or not, QTLs are accurate! *Trends Plant Sci* **11,** 213-216.

Rae, A.M., Howell, E.C. and Kearsey, M.J. (1999). More QTL for flowering time revealed by substitution lines in *Brassica oleracea*. *Heredity* **83 (Pt 5),** 586-596.

Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., Doebley, J., Kresovich, S., Goodman, M.M. and Buckler, E.S.t. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci U S A* **98,** 11479-11484.

Rockman, M.V. and Kruglyak, L. (2006). Genetics of global gene expression. *Nat Rev Genet* **7,** 862-872.

Ronald, J., Brem, R.B., Whittle, J. and Kruglyak, L. (2005). Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet* **1,** e25.

Ross-Ibarra, J. (2005). Quantitative trait loci and the study of plant domestication. *Genetica* **123,** 197-204.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G. ***et al.*** (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422,** 297-302.

Schauer, N., Steinhauser, D., Strelkov, S., Schomburg, D., Allison, G., Moritz, T., Lundgren, K., Roessner-Tunali, U., Forbes, M.G., Willmitzer, L. ***et al.*** (2005). GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett* **579,** 1332-1337.

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

Schmid, K.J., Sorensen, T.R., Stracke, R., Torjek, O., Altmann, T., Mitchell-Olds, T. and Weisshaar, B. (2003). Large-Scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res* **13,** 1250-1257.

Schmid, K.J., Torjek, O., Meyer, R., Schmuths, H., Hoffmann, M.H. and Altmann, T. (2006). Evidence for a large-scale population structure of *Arabidopsis thaliana* from genome-wide single nucleotide polymorphism markers. *Theor Appl Genet* **112,** 1104-1114.

Segal, E., Yelensky, R. and Koller, D. (2003). Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* **19 Suppl 1,** i273-282.

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.

Slate, J. (2005). Quantitative trait locus mapping in natural populations: progress, caveats and future directions. *Mol Ecol* **14,** 363-379.

Somerville, C. and Koornneef, M. (2002). Timeline: A fortunate choice: the history of Arabidopsis as a model plant. *Nat Rev Genet* **3,** 883-889.

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

Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302,** 249-255.

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.

Teng, S., Keurentjes, J.J.B., Bentsink, L., Koornneef, M. and Smeekens, S. (2005). Sucrose-specific induction of anthocyanin biosynthesis in Arabidopsis requires the MYB75/PAP1 gene. *Plant Physiol* **139,** 1840-1852.

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

Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M. *et al.* (2004). Global mapping of the yeast genetic interaction network. *Science* **303,** 808-813.

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.

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.

Vigouroux, Y., Mitchell, S., Matsuoka, Y., Hamblin, M., Kresovich, S., Smith, J.S., Jaqueth, J., Smith, O.S. and Doebley, J. (2005). An analysis of genetic diversity across the maize genome using microsatellites. *Genetics* **169,** 1617-1630.

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

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

Ward, J.L., Harris, C., Lewis, J. and Beale, M.H. (2003). Assessment of 1H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsis thaliana*. *Phytochemistry* **62,** 949-957.

Ward, J.L., Baker, J.M. and Beale, M.H. (2007). Recent applications of NMR spectroscopy in plant metabolomics. *Febs J* **274,** 1126-1131.

Weigel, D. and Nordborg, M. (2005). Natural variation in Arabidopsis. How do we find the causal genes? *Plant Physiol* **138,** 567-568.

West, M.A., van Leeuwen, H., Kozik, A., Kliebenstein, D.J., Doerge, R.W., St Clair, D.A. and Michelmore, R.W. (2006). High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis. *Genome Res* **16,** 787-795.

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

Winnacker, E.L. (2003). Interdisciplinary sciences in the 21st century. *Curr Opin Biotechnol* **14,** 328-331.

Yoon, D.B., Kang, K.H., Kim, H.J., Ju, H.G., Kwon, S.J., Suh, J.P., Jeong, O.Y. and Ahn, S.N. (2006). Mapping quantitative trait loci for yield components and morphological traits in an advanced backcross population between *Oryza grandiglumis* and the *O. sativa japonica* cultivar Hwaseongbyeo. *Theor Appl Genet* **112,** 1052-1062.

Zhang, X., Richards, E.J. and Borevitz, J.O. (2007). Genetic and epigenetic dissection of cis regulatory variation. *Curr Opin Plant Biol* **10,** 142-148.

Zhao, J., Becker, H.C., Zhang, D., Zhang, Y. and Ecke, W. (2006). Conditional QTL mapping of oil content in rapeseed with respect to protein content and traits related to plant development and grain yield. *Theor Appl Genet* **113,** 33-38.

Zhu, J., Lum, P.Y., Lamb, J., GuhaThakurta, D., Edwards, S.W., Thieringer, R., Berger, J.P., Wu, M.S., Thompson, J., Sachs, A.B. ***et al.*** (2004). An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet Genome Res* **105,** 363-374.

Chapter 2

Development of a Near-Isogenic Line population of *Arabidopsis thaliana* and comparison of mapping power with a Recombinant Inbred Line population

Joost J. B. Keurentjes, Leónie Bentsink, Carlos Alonso-Blanco, Corrie J. Hanhart, Hetty Blankestijn-De Vries, Sigi Effgen, Dick Vreugdenhil and Maarten Koornneef

Published in Genetics (2007) **175**, 891-905.

ABSTRACT

In Arabidopsis Recombinant Inbred Line (RIL) populations are widely used for Quantitative Trait Locus (QTL) analyses. However, mapping analyses with this type of populations can be limited because of masking effects of major QTLs and epistatic interactions of multiple QTLs. An alternative type of immortal experimental population commonly used in plant species are sets of introgression lines. Here we introduce the development of a genome-wide coverage Near Isogenic Line (NIL) population of *Arabidopsis thaliana*, by introgressing genomic regions from the Cape Verde Islands (Cvi) accession into the Landsberg *erecta* (L*er*) genetic background. We have empirically compared the QTL mapping power of this new population with an already existing RIL population derived from the same parents. For that, we analyzed and mapped QTLs affecting six developmental traits with different heritability. Overall, in the NIL population smaller-effect QTLs than in the RIL population could be detected although the localization resolution was lower. Furthermore, we estimated the effect of population size and of the number of replicates on the detection power of QTLs affecting the developmental traits. In general, population size is more important than the number of replicates to increase the mapping power of RILs, whereas for NILs, several replicates are absolutely required. These analyses are expected to facilitate experimental design for QTL mapping using these two common types of segregating populations.

INTRODUCTION

Quantitative traits are characterized by continuous variation. The establishment of the genetic basis of quantitative traits is commonly referred to as Quantitative Trait Locus (QTL) mapping, and has been hampered due to their multigenic inheritance and the often strong interaction with the environment. The principle of QTL mapping in segregating populations is based on the genotyping of progeny derived from a cross of distinct genotypes for the trait under study. Phenotypic values for the quantitative trait are then compared with the molecular marker genotypes of the progeny to search for particular genomic regions showing statistical significant associations with the trait variation, which are then called QTLs (Broman, 2001; Slate, 2005). Over the past few decades, the field has benefited enormously from the progress made in molecular marker technology. The ease by which such markers can be developed has enabled the generation of dense genetic maps and the performance of QTL mapping studies of the most complex traits (Borevitz and Nordborg, 2003).

QTL analyses make use of the natural variation present within species (Alonso-Blanco and Koornneef, 2000; Maloof, 2003) and have been successfully applied to various types of segregating populations. In plants, the use of 'immortal' mapping populations consisting of homozygous individuals is preferred because it allows performing replications and multiple analyses of the same population. Homozygous populations can be obtained by repeated selfing, like for Recombinant Inbred Lines (RILs), but also by induced chromosomal doubling of haploids, such as for Doubled Haploids (DHs) (Han *et al.*, 1997; Rae *et al.*, 1999; von Korff *et al.*, 2004). Depending on the species one can in principle also obtain immortality by vegetative propagation, although this is often more laborious. RILs are advantageous over DHs because of their higher recombination frequency in the population, resulting from multiple meiotic events occurred during repeated selfing (Jansen, 2003).

Another type of immortal population consists of Introgression Lines (ILs) (Eshed and Zamir, 1995), which are obtained through repeated backcrossing and extensive genotyping. These are also referred to as Near Isogenic Lines (NILs) (Monforte and Tanksley, 2000) or Backcross Inbred Lines (BILs) (Jeuken and Lindhout, 2004; Blanco *et al.*, 2006). Such populations consist of lines containing a single or a small number of genomic introgression fragments from a donor parent into an otherwise homogeneous genetic background. Although no essential differences exist between these populations, we use the term Near Isogenic Lines for the materials described here. A special case of ILs are Chromosomal

Substitution Strains (CSSs) (Nadeau *et al.*, 2000; Koumproglou *et al.*, 2002), where the introgressions span complete chromosomes. All immortal populations except those which can only be propagated vegetatively, share the advantage that they can easily be maintained through seeds, which allows the analysis of different environmental influences and the study of multiple, even invasive or destructive, traits. Statistical power of such analyses is increased because replicate measurements of genetically identical individuals can be done.

In plants, RILs and NILs are the most common types of experimental populations used for the analysis of quantitative traits. In both cases the accuracy of QTL localization, referred to as mapping resolution, depends on population size. For RILs, recombination frequency within existing lines is fixed and can therefore only be increased within the population by adding more lines (*i.e.* more independent recombination events). Alternatively, recombination frequency can be increased by intercrossing lines before fixation as homozygous lines by inbreeding (Zou *et al.*, 2005). In NIL populations resolution can be improved by minimizing the introgression size of each NIL. Consequently, to maintain genome-wide coverage a larger number of lines are needed. Despite the similarities between these two types of mapping populations, large differences exist in the genetic makeup of the respective individuals and the resulting mapping approach. In general, recombination frequency in RIL populations is higher than in equally sized NIL populations, which allows the analysis of less individuals. Each RIL contains several introgression fragments and, on average, each genomic region is represented by an equal number of both parental genotypes in the population. Therefore, replication of individual lines is often not necessary because the effect of each genomic region on phenotypic traits is tested by comparing the two genotypic RIL classes (each comprising approximately half the number of lines in the population). In addition, the multiple introgressions per RIL allow detection of genetic interactions between loci (epistasis). However, epistasis together with unequal recombination frequencies throughout the genome and segregation distortions caused by lethality or reduced fitness of particular genotypes may bias the power to detect QTLs. Furthermore, the wide variation of morphological and developmental traits present in most RIL populations may hamper the analysis of traits requiring the same growth and developmental stage of the individual lines. When many traits segregate simultaneously, this often affects the expression of other traits due to genetic interactions. Moreover, large-effect QTLs may mask the detection of QTLs with a small additive effect.

In contrast to RILs, NILs preferably contain only a single introgression per line, which increases the power to detect small-effect QTLs. However, the presence of a single introgression segment does not allow testing for genetic interactions and thereby the detection of QTLs expressed in specific genetic backgrounds (epistasis). In addition, because most of the genetic background is identical for all lines, NILs show more limited developmental and growth variation, increasing the homogeneity of growth stage within experiments. Nevertheless, lethality and sterility might sometimes hinder the obtaining of specific single introgression lines.

The choice of one mapping population over another depends on the plant species and the specific parents of interest. In cases where different cultivars or wild accessions are studied preference is often given to RILs. However, when different species or when wild and cultivated germplasm are combined NILs are preferred (Eshed and Zamir, 1995; Jeuken and Lindhout, 2004; von Korff *et al.*, 2004; Blair *et al.*, 2006; Yoon *et al.*, 2006). For instance, in tomato the high sterility in the offspring of crosses between cultivated and wild species made the use of NIL populations preferable because genome-wide coverage cannot be obtained with RIL populations due to sterility etc. (Eshed and Zamir, 1995). Furthermore, the analysis of agronomical important traits (such as fruit characters) cannot be performed when many genes conferring reduced fertility segregate. In Arabidopsis, the easiness to generate fertile RIL populations with complete genome coverage, due to its fast generation time, has led to their extensive use in mapping quantitative traits.

NILs have been developed in various studies using Arabidopsis to confirm and fine map QTLs previously identified in RILs (Alonso-Blanco *et al.*, 1998a, 2003; Swarup *et al.*, 1999; Bentsink *et al.*, 2003; Edwards *et al.*, 2005; Juenger *et al.*, 2005a; Teng *et al.*, 2005) for which also Heterogeneous Inbred Families (HIFs) (Tuinstra *et al.*, 1997) have been used (Loudet *et al.*, 2005; Reymond *et al.*, 2006). A set of chromosomal substitutions of the Landsberg *erecta* (L*er*) accession into Columbia (Col) has been developed to serve as starting material for making smaller introgressions (Koumproglou *et al.*, 2002). In mice CSSs are widely used for mapping purposes and have proven to be a valuable complement to other population types (Stylianou *et al.*, 2006). However, no genome-wide set of NIls that allows mapping to subparts of the chromosome has been described in Arabidopsis and, to our knowledge, no empirical comparative study has been performed between the two population types within a single species.

In this study we aim to compare a RIL population with a NIL population in terms of QTL detection power and localization resolution. For that, we generated a new genome-wide population of NILs using the same L*er* and Cvi parental accessions as used earlier to generate a RIL population (Alonso-Blanco *et al.*, 1998b). The two experimental populations were grown simultaneously in the same experimental setup, including multiple replicates. QTL mapping analyses were performed on six different traits and the results of these analyses were compared in both populations.

RESULTS

Construction of a genome-wide Near Isogenic Line population

We constructed a population of 92 introgression lines carrying between one and four Cvi introgression fragments in a L*er* genetic background. Lines were genotyped using 349 AFLP and 95 PCR markers to determine the number, position and size of the introgressions (see Materials and Methods). This set of lines was selected to provide together an almost complete genome-wide coverage (Figure 1). Forty lines contained a single introgression while 52 lines carried several Cvi fragments. From those, 32, 19, and 1 line bore two, three and four introgressions respectively. The genetic length of the introgression fragments was estimated using the map positions of the introgressed markers in the genetic map constructed from the existing RIL population derived from the same L*er* and Cvi parental accessions (Alonso-Blanco *et al.*, 1998b). The average genetic size of the main, second, third, and fourth introgression fragment was 31.7, 11.1, 6.7, and 5.2 cM respectively. Thus, lines with multiple Cvi fragments carried a main large introgression and several much smaller Cvi fragments. Additionally, we selected a core set of 25 lines that together covered more then 90% of the genome (supplemental Table 1 at http://www.genetics.org/supplemental/).

Genetic analyses of developmental traits

Six traits were measured and analyzed in the RIL and NIL populations (Table 1). Although plants were grown in four replicated blocks, block effects were negligible and was therefore not used as a factor in subsequent analyses. In both populations, among-genotype variance was highly significant (P < 0.0001) for all traits. In the RIL population, broad sense heritability estimates ranged from 0.34 (basal branch number) to 0.92 (total plant length) (Table 1). Statistical parameters of most traits were similar to those described by Alonso-Blanco *et al.* (1998a, 1999) and Juenger *et al.* (2005b). However, Ungerer *et al.* (2002) reported much lower average values for plant height and branch number although time to flower was similar. Moreover, among-genotype variance estimates were lower and within-genotype variance estimates higher resulting in lower heritability values compared to our analyses.

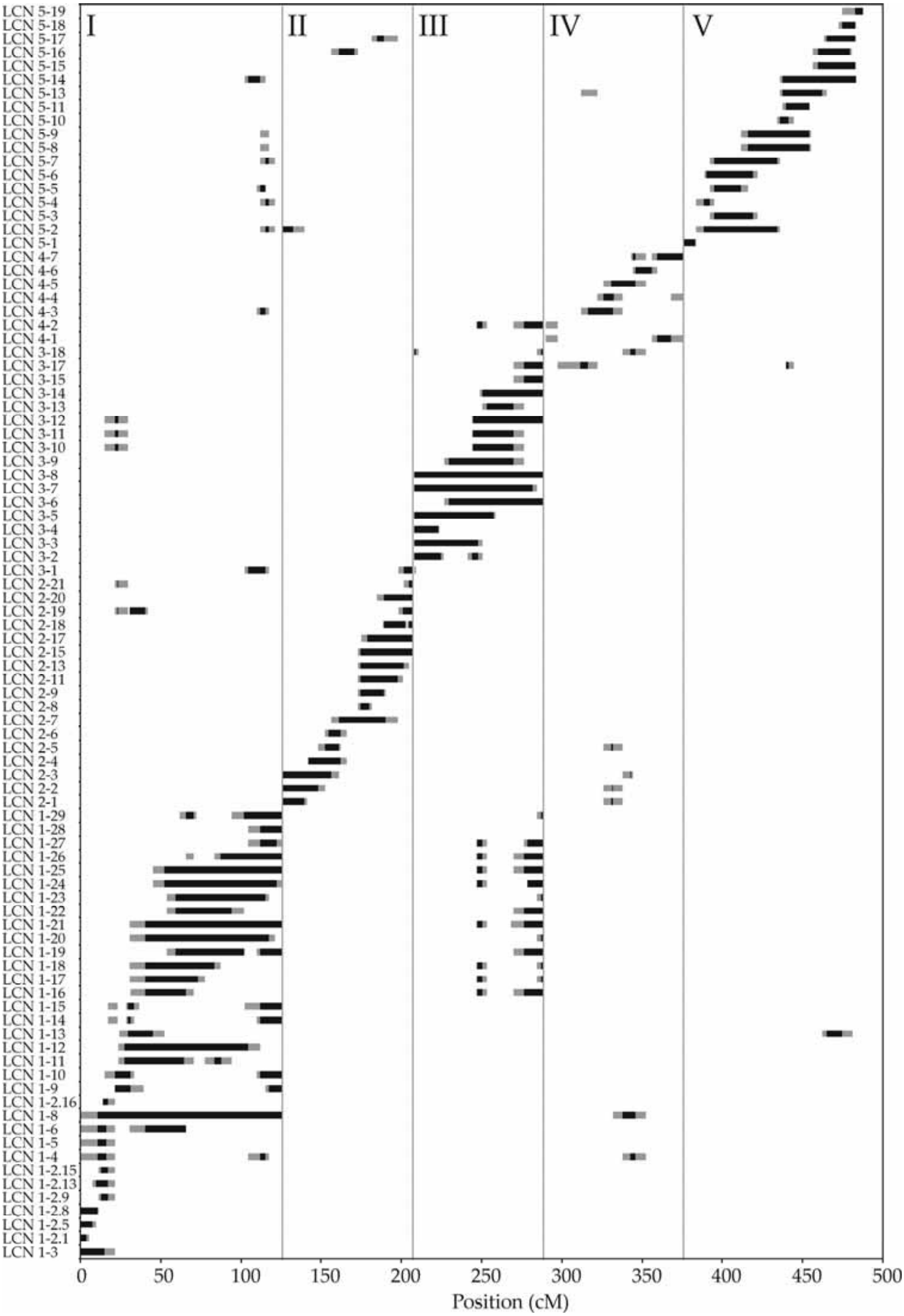

LCN 5-19
LCN 5-18
LCN 5-17
LCN 5-16
LCN 5-15
LCN 5-14
LCN 5-13
LCN 5-11
LCN 5-10
LCN 5-9
LCN 5-8
LCN 5-7
LCN 5-6
LCN 5-5
LCN 5-4
LCN 5-3
LCN 5-2
LCN 5-1
LCN 4-7
LCN 4-6
LCN 4-5
LCN 4-4
LCN 4-3
LCN 4-2
LCN 4-1
LCN 3-18
LCN 3-17
LCN 3-15
LCN 3-14
LCN 3-13
LCN 3-12
LCN 3-11
LCN 3-10
LCN 3-9
LCN 3-8
LCN 3-7
LCN 3-6
LCN 3-5
LCN 3-4
LCN 3-3
LCN 3-2
LCN 3-1
LCN 2-21
LCN 2-20
LCN 2-19
LCN 2-18
LCN 2-17
LCN 2-15
LCN 2-13
LCN 2-11
LCN 2-9
LCN 2-8
LCN 2-7
LCN 2-6
LCN 2-5
LCN 2-4
LCN 2-3
LCN 2-2
LCN 2-1
LCN 1-29
LCN 1-28
LCN 1-27
LCN 1-26
LCN 1-25
LCN 1-24
LCN 1-23
LCN 1-22
LCN 1-21
LCN 1-20
LCN 1-19
LCN 1-18
LCN 1-17
LCN 1-16
LCN 1-15
LCN 1-14
LCN 1-13
LCN 1-12
LCN 1-11
LCN 1-10
LCN 1-9
LCN 1-2.16
LCN 1-8
LCN 1-6
LCN 1-5
LCN 1-4
LCN 1-2.15
LCN 1-2.13
LCN 1-2.9
LCN 1-2.8
LCN 1-2.5
LCN 1-2.1
LCN 1-3
I
II
III
IV
V
0
50
100
150
200
250
300
350
400
450
500
Position (cM)

Figure 1: Graphical genotype of the L*er* x Cvi NIL population.
Bars represent introgressions. Solid bars represent the genetic position of Cvi introgressions in individual NILs. Shaded bars represent crossover regions between markers used for the genotyping of the lines. Numbers at the top indicate the five linkage groups.

Table 1: Descriptive statistics for six developmental traits analyzed in two mapping populations and their parents.

Trait	$\overline{X} \pm (SD)$	[VG]^{a}	[VE]^{b}	[H^{2}]^{c}	[CVG]^{d}
		Parents			
FT (days)	24.30 (1.03)^{e} 30.21 (2.47)^{f}	8.74	3.57	0.71	10.85
SL (cm)	9.58 (0.98)^{e} 13.21 (2.30)^{f}	3.27	3.14	0.51	15.87
TL (cm)	23.59 (1.92)^{e} 33.95 (4.17)^{f}	26.81	10.53	0.72	17.99
IB	2.21 (0.46)^{e} 2.49 (0.67)^{f}	0.02	0.33	0.05	5.53
BB	1.54 (0.68)^{e} 1.48 (0.91)^{f}	0.00	0.65	0.00	0.00
TB	3.75 (0.77)^{e} 3.97 (1.02)^{f}	0.01	0.82	0.01	1.88
		RIL population			
FT (days)	26.06 (6.03)	32.59	3.82	0.90	21.91
SL (cm)	9.89 (3.39)	9.70	1.80	0.83	31.49
TL (cm)	26.13 (9.22)	78.53	6.52	0.92	33.91
IB	2.34 (1.22)	0.99	0.50	0.67	42.66
BB	1.43 (0.93)	0.30	0.57	0.34	37.98
TB	3.77 (1.27)	0.78	0.84	0.48	23.36
		NIL population			
FT (days)	23.68 (3.60)	10.78	2.21	0.83	13.87
SL (cm)	9.81 (2.18)	3.17	1.58	0.65	18.15
TL (cm)	24.50 (5.95)	31.24	4.10	0.87	22.82
IB	2.26 (0.88)	0.51	0.27	0.65	31.42
BB	1.56 (0.84)	0.18	0.53	0.24	26.92
TB	3.82 (1.06)	0.48	0.64	0.42	18.25

FT, flowering time; SL, length at first silique; TL, total plant length; IB, main inflorescence branch number; BB, basal branch number; TB, total branch number. [a] Among-genotype variance component from ANOVA; tests whether genetic differences exist among genotypes for specified traits ($P < 0.0001$). [b] Residual variance component from ANOVA. [c] Measure of total phenotypic variance attributable to genetic differences among genotypes (broad sense heritability) calculated as $V_G/(V_G+V_E)$. [d] Coefficient of genetic variation calculated as $\left(100 \times \sqrt{V_G}\right)/\overline{X}$. [e] Landsberg *erecta* parent. [f] Cape Verde Islands parent.

For the NIL population, mean trait values were closer to those measured for L*er* due to the genetic structure of the population, consisting of lines carrying only small Cvi introgressions in a L*er* background. Furthermore, variance components from ANOVA were lower in the NIL population but heritability estimates differed only slightly compared to the RIL population (Table 1).

Strong and similar genetic correlations were observed between traits in the two L*er* x Cvi populations indicating partial genetic co-regulation (Table 2). Flowering time shows the highest correlation with the number of main inflorescence branches but is negatively correlated with basal branch number. Flowering time is also, but to a lesser degree, correlated with plant height. Correlations were also found between plant height and branching, with again positive values with the number of main inflorescence branches and negative correlations with basal branch number. These results contrasted with those from Ungerer *et al.* (2002), who found negative correlations between flowering time, plant height and branching in all pair-wise comparisons, which is probably due to the different environmental set up in the two laboratories.

Table 2: Genetic correlations among developmental traits analyzed in two mapping populations.

Trait	FT	SL	TL	IB	BB	TB
FT		0.63*	0.38*	0.97*	-0.49*	0.80*
SL	0.39*		0.90*	0.52*	-0.39*	0.35*
TL	0.21*	0.88*		0.18*	-0.32*	0.00
IB	0.91*	0.31*	0.09*		-0.54*	0.95*
BB	-0.26*	-0.28*	-0.26*	-0.35*		0.12*
TB	0.77*	0.15*	-0.07	0.85*	0.31*	

The top right and the bottom left halves of the table represent values calculated for the RIL and the NIL populations respectively. FT, flowering time; SL, length at first silique; TL, total plant length; IB, main inflorescence branch number; BB, basal branch number; TB, total branch number. * Significant at $P <$ 0.001.

Mapping quantitative traits in the L*er* x Cvi RIL population

Each trait was subjected to QTL analysis and three to eight QTLs were detected for each trait (Figure 2, Table 3). Major QTLs for flowering time, plant height and branching were in concordance with previously reported studies (Alonso-Blanco *et al.*, 1998a, 1999; Ungerer *et al.*, 2002, 2003; Juenger *et al.*, 2005b), although slight differences for minor QTLs were also found. Total explained variance for each trait ranged from 38.5% for basal branch number to 86.3% for total plant height. LOD scores for the largest-effect QTL ranged from 5.7 for basal branch number up to 60.7 for total plant height with corresponding explained variances of 11.0 and 64.0% respectively. The average genetic length of 2-LOD support intervals was 11.6 cM, ranging from 2.3 (length at first silique) to 33.3 cM (total branch number).

Table 3: QTLs detected in the RIL population.

Trait	Chr[a]	LOD score	support interval[b] (cM)	Explained Variance[c] (%)	Effect[d]	Total Explained Variance[e] (%)	Interaction[f] (%)
FT	1	11.9	1.5-9.8*	13.0	-3.9	68.4	9.6
	5	18.9	388.4-394.5*	22.2	5.7		
	5	11.9	408.2-413.7*	13.0	4.4		
SL	1	9.3	0.0-9.3	6.3	-1.7	79.5	15.0
	1	4.8	103.1-126.0	3.1	-1.3		
	2	39.7	173.2-175.5	43.2	4.5		
	3	2.9	234.2-253.6	1.9	1.0		
	3	5.0	281.5-287.8	3.2	-1.2		
	5	15.7	387.9-392.4*	11.8	2.9		
	5	10.2	403.6-409.7*	7.2	2.0		
TL	1	6.5	0.0-9.8*	2.8	-3.1	86.3	11.5
	1	5.0	73.9-84.6	2.1	-2.7		
	1	3.3	116.3-126.0	1.2	-2.3		
	2	60.7	173.2-176.0*	64.0	14.8		
	3	6.0	207.3-225.7*	2.6	-3.0		
	4	5.2	287.8-307.5*	2.2	-2.7		
	5	7.8	383.1-392.5*	3.6	4.1		
	5	5.1	403.6-411.7	2.2	3.0		
IB	1	5.0	0.0-13.5*	5.3	-0.4	65.0	20.5
	2	2.7	154.9-171.0*	2.8	-0.3		
	5	15.3	387.0-391.9*	19.7	0.9		
	5	10.4	398.8-411.7*	12.3	0.7		
	5	3.1	472.2-485.3	3.2	-0.3		
BB	1	5.7	72.4-91.0*	11.0	0.4	38.5	3.1
	2	3.2	167.0-200.2*	6.2	-0.3		
	4	4.6	360.7-373.5*	9.1	0.4		
	5	5.5	385.6-406.1*	11.3	-0.5		
TB	1	15.5	5.3-12.4*	16.1	-0.8	71.1	16.2
	1	4.9	81.7-93.8*	4.6	0.4		
	2	9.5	169.0-180.0*	9.1	-0.6		
	5	9.7	386.5-392.4*	9.4	0.6		
	5	10.9	403.3-412.2*	10.8	0.7		
	5	5.2	472.2-485.3	4.7	-0.4		

FT, flowering time; SL, length at first silique; TL, total plant length; IB, main inflorescence branch number; BB, basal branch number; TB, total branch number. [a] Chromosome number. [b] 2-LOD support interval. [c] Percentage of total variation explained by individual QTLs. [d] Effect of QTLs calculated as μ_B-μ_A, where A and B are RILs carrying L*er* and Cvi genotypes at the QTL positions, respectively. μ_B and μ_A were estimated by MapQTL®. Effects are given in days (flowering time), centimeters (length at first silique and total length) or numbers (elongated axils, basal branch number and total branch number). [e] Percentage of total variance explained by genetic factors estimated by MapQTL®. [f] Percentage of total variation explained by interaction between individual QTLs. * QTLs showing significant epistatic interactions ($P < 0.05$) and used to estimate the percentage of explained variance by genetic interactions.

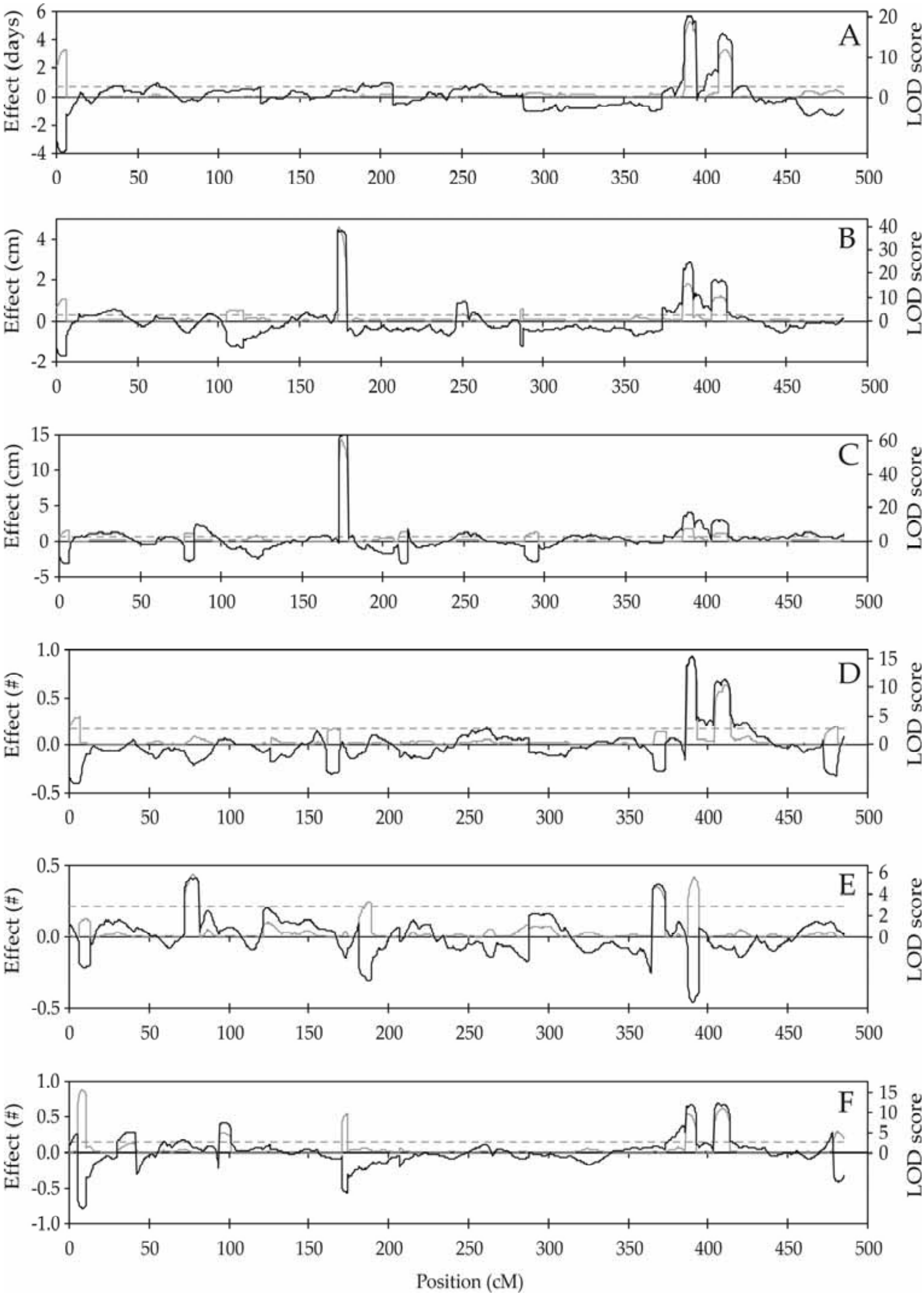

A
Effect (days)
LOD score
B
Effect (cm)
C
Effect (cm)
D
Effect (#)
E
Effect (#)
F
Effect (#)
Position (cM)
0
50
100
150
200
250
300
350
400
450
500

Figure 2: Genome-wide QTL profiles of traits analyzed in the RIL population.
(A) Flowering time, (B) Length at first silique, (C) Total plant length, (D) Number of main inflorescence branches, (E) Basal branch number and (F) Total branch number. Solid lines represent the QTL effect calculated as described in Materials and Methods. Shaded lines represent LOD scores. Shaded dashed lines represent genome-wide significance threshold levels for LOD scores determined by permutation testing.

Opposing effect QTLs were found for all traits, explaining the observed transgressive segregation within the population (data not shown). Genetic interaction among the detected QTLs was also tested. The proportion of variance explained by epistatic interactions ranged from 3.1 (basal branch number) to 20.5% (number of main inflorescence branches) and involved two to five of the detected QTLs (Table 3). Using a complete pairwise search of all markers (Chase *et al.*, 1997), a number of additional interactions were detected between loci not co-locating with major QTL positions (supplemental Figure 1 at http://www.genetics.org/supplemental/).

The smallest significant absolute effect detected was 4.4 days for flowering time, 1.0 and 2.3 cm for length at first silique and total plant length, respectively, and 0.3, 0.3, and 0.4 for the number of main inflorescence branches, basal branch number and total branch number, respectively. Relative effects, expressed as the fold difference between genotypes, calculated as $(|\mu_B-\mu_A|+\mu_A)/\mu_A$, then equaled 1.15-, 1.09-, 1.09-, 1.13-, 1.59-, and 1.10-fold, respectively (Tables 3 and 5). As expected, the total explained variance of a trait correlated positively with the smallest significantly detectable effect for that particular trait. In general, smaller effects could be detected with increasing total explained variance. When the chromosome-wide threshold for significance was used instead of the genome-wide threshold, one additional suggestive QTL was detected for main inflorescence branch number and total branch number and two for length at first silique.

Mapping quantitative traits in the L*er* x Cvi NIL population

To search for QTLs in the NIL population, we divided the Arabidopsis genetic map in adjacent genomic fragments that were individually tested. The complete genome was subdivided into 97 regions, defined by the position of the recombination events of the main introgressions of the 92 NILs (supplemental Table 2 at http://www.genetics.org/supplemental/). These regions are referred to as bins and each NIL was then assigned to those adjacent bins spanned by its Cvi introgression fragment. Thus, each bin contains a unique subset of lines with overlapping Cvi introgressions in that particular region, which were used to test the phenotypic effects of that bin. The average genetic length of the bins was 5.0 cM, ranging from 0.1 to 26.3 cM. The number of NILs per bin ranged from 0 to 13 with an average of

5.1 NILs. Because NILs were only assigned to bins when the complete bin was covered by the introgression, three bins remained empty [*viz.* bins 66 (26.3 cM), 73 (3.3 cM) and 77 (5.4 cM)]. On average each NIL was assigned to 5.4 adjacent bins. One NIL (LCN4-2) was not assigned to any bin because its introgression included only a single marker. Two NILs corresponded to complete chromosomal substitutions: line LCN3-8 (chromosome 3) and line LCN1-8 (chromosome 1), the latter carrying the largest introgression assigned to 27 adjacent bins.

To map QTLs in the NIL population, all bins were tested individually by comparing the phenotypes of the NILs assigned to each bin with that of L*er*. As shown in Figure 3 and Table 4, one to nine QTLs were detected for each trait. The total explained variance for each trait ranged from 26.7% for basal branch number up to 87.7% for total plant height. Explained variances for the largest-effect QTL for each trait ranged from 19.3% for basal branch number to 91.9% for total plant height as calculated from a restricted ANOVA using only lines from the most significant bin and L*er*. To show the relative effect of Mendelizing QTLs with respect to the total population variance we calculated the explained variances also when all lines of the population were subjected to ANOVA analysis using the most significant bin as fixed factor (Table 4). Relative effects of QTLs were much lower in this unrestricted analysis because all other QTLs in the population increase residual variation which is not corrected for, as is done in MQM mapping in the RIL population. Moreover, lines partly overlapping the QTL bin are not assigned to that bin but can still contain the QTL Cvi allele, further increasing the residual variation in the population.

The smallest significant QTL effect detected was 0.7 days for flowering time, 1.1 and 2.1 cm for length at first silique and total plant length, respectively, and 3.8, 0.5, and 0.4 for the number of main inflorescence branches, basal branch number and total branch number, respectively. Relative effects, expressed as the fold difference between genotypes, calculated as $(|\mu_B-\mu_A|+\mu_A)/\mu_A$, then equaled 1.03-, 1.11-, 1.09-, 2.71-, 1.30-, and 1.11-fold, respectively (Tables 4 and 5).

For a number of traits several QTLs were found that could not be significantly detected in the RIL population. In total 12 of such small-effect QTLs were detected for flowering time (3), length at first silique (5), total plant length (2), and basal branch number (2). None of those met the lower chromosome-wide significance threshold for suggestive QTLs in the RIL population. Although two were close to this threshold, ten of them did not reach LOD scores >1.0 in the RIL population (supplemental Table 3 at http://www.genetics.org/supplemental/).

Table 4: QTLs detected in the NIL population.

		Support	Support	Explained Variance (%)			Total Explained
Trait	Chr[a]	interval[b]	bin (cM)[c]	Restricted[d]	Unrestricted[e]	Effect[f]	Variance[g] (%)
FT	1	0.0 - 21.6	3.9 - 7.8	70.3	3.2	-3.2	83.2
	1	31.4 - 40.6	33.4 - 40.7	18.0	0.5	-1.0	
	1	73.3 - 122.0	83.6 – 87.0	7.1	0.7	-0.7	
	2	174.4 - 204.7	200.9 - 201.8	22.3	0.6	1.5	
	5	388.4 - 434.2	392.3 - 395.0	52.1	42.8	15.7	
SL	1	10.8 - 27.4	17.3 - 21.7	64.0	4.8	-3.1	66.1
	1	31.4 - 40.6	33.4 - 40.7	17.1	0.6	-1.1	
	1	73.3 - 125.9	122.1 - 126.0	34.9	2.8	-1.7	
	2	160.8 - 207.2	162.0 - 174.5	73.4	5.3	4.9	
	3	270.1 - 288.4	287.1 - 288.4	37.1	1.6	-1.7	
	4	359.5 - 375.7	368.2 - 375.7	32.2	1.7	-1.6	
	5	388.3 - 418.9	392.3 – 395.0	32.2	0.7	2.7	
	5	434.2 - 436.0	434.3 - 436.1	29.6	3.8	-1.4	
	5	441.4 - 459.3	454.3 - 459.4	28.2	1.1	-1.1	
TL	1	0.0 - 33.3	17.3 - 21.7	66.2	1.7	-6.3	87.7
	1	64.7 - 125.9	122.1 – 126.0	48.8	3.8	-3.8	
	2	160.8 - 207.2	174.5 - 178.8	91.9	10.5	18.5	
	3	287.0 - 288.4	287.1 - 288.4	19.0	0.4	-2.1	
	5	389.9 - 416.1	411.7 - 416.2	34.1	1.7	3.7	
	5	434.2 - 454.3	434.3 - 436.1	45.0	1.4	-3.9	
IB	5	388.3 - 434.2	392.3 – 395.0	46.3	37.7	3.8	66.1
BB	1	0.0 - 15.1	3.9 - 7.8	17.7	1.8	-0.6	26.7
	1	40.6 - 125.9	94.5 - 101.6	17.9	9.0	0.8	
	2	174.4 - 189.1	179.7 - 189.2	11.4	2.4	-0.5	
	5	388.3 - 434.2	392.3 – 395.0	14.4	1.7	-0.7	
	5	483.2 - 487.8	483.2 - 487.8	19.3	1.1	-0.8	
TB	1	0.0 - 15.9	7.8 - 9.9	24.1	2.2	-0.8	44.1
	1	40.6 - 125.9	94.5 - 101.6	14.0	4.1	0.8	
	2	174.4 - 189.1	179.7 - 189.2	7.6	1.5	-0.4	
	5	388.3 - 434.2	392.3 – 395.0	43.2	17.4	3.1	

FT, flowering time; SL, length at first silique; TL, total plant length; IB, main inflorescence branch number; BB, basal branch number; TB, total branch number. [a] Chromosome number. [b] The region spanned by consecutive bins, significantly ($P < 0.001$) differing from L*er* and sharing the same direction of effect, was taken as support interval. [c] Position of the bin within the QTL support interval showing the largest effect. [d] Within the QTL support interval, the bin showing the largest effect was compared to L*er* in an ANOVA analysis. The among-genotype component of ANOVA was taken as an estimator of explained variance. [e] All lines in the population were subjected to ANOVA using the bin described in footnote [d] as fixed factor. The among-genotype component of ANOVA was taken as an estimator of explained variance. [f] Effect of QTLs calculated as μ_B-μ_A, where μ_A is the mean value of all L*er* lines and μ_B is the mean value of all lines in the bin described in footnote [d]. Effects are given in days (flowering time), centimeters (length at first silique and total length) or numbers (main inflorescence branch number, basal branch number and total branch number). [g] All bins together with L*er* were analyzed by ANOVA and the among-genotype component was taken as a measure of totally explained variance.

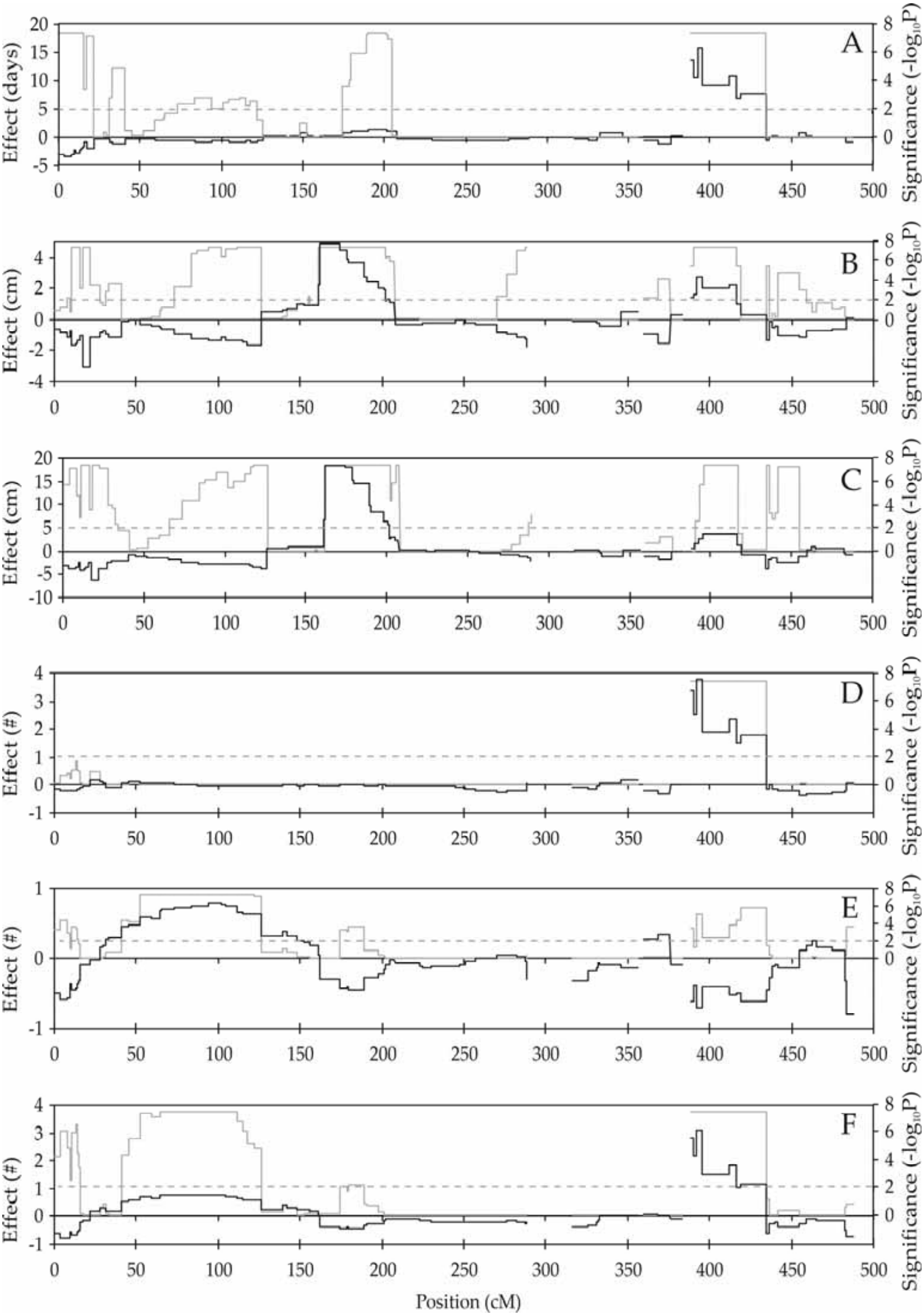

A
B
C
D
E
F
Effect (days)
Effect (cm)
Effect (cm)
Effect (#)
Effect (#)
Effect (#)
Significance (-log10P)
Position (cM)
0
50
100
150
200
250
300
350
400
450
500

Figure 3: QTL profiles of traits analyzed in the NIL population.
(A) Flowering time, (B) Length at first silique, (C) Total plant length, (D) Number of main inflorescence branches, (E) Basal branch number and (F) Total branch number. Solid lines represent the QTL effect calculated as described in Materials and Methods. Shaded lines represent significance scores. Shaded dashed lines represent significance threshold levels applied in this study.

Table 5: Comparative summary of QTL mapping parameters in the L*er* x Cvi RIL and NIL populations.

Trait	Population.[a]	QTLs[b] (no.)	Support[c] (cM)	Explained Variance[d] (%)	Total explained Variance (%)	Effect[e]	Relative effect[f]
FT	RIL	3	6.6	16.1	68.4	4.7	1.15
	NIL	5	35.5 (3.6)	34.0	83.2	4.4	1.03
SL	RIL	7	10.1	11.0	79.5	2.1	1.09
	NIL	9	23.3 (5.2)	38.7	66.1	2.1	1.11
TL	RIL	8	11.1	10.1	86.3	4.5	1.09
	NIL	6	31.4 (3.4)	50.8	87.7	6.4	1.09
IB	RIL	5	12.1	8.7	65.0	0.5	1.13
	NIL	1	45.9 (2.7)	46.3	66.1	3.8	2.71
BB	RIL	4	21.3	9.4	38.5	0.4	1.59
	NIL	5	33.1 (5.6)	16.1	26.7	0.7	1.30
TB	RIL	6	9.7	9.1	71.1	0.6	1.10
	NIL	4	40.5 (5.4)	22.2	44.1	1.3	1.11

FT, flowering time; SL, length at first silique; TL, total plant length; IB, main inflorescence branch number; BB, basal branch number; TB, total branch number. [a] Population type. [b] Number of QTLs detected. [c] Average length of support interval. In parentheses: average length of largest-effect bin. [d] Average explained variance for each QTL. [e] Average absolute effect for each QTL. Effects are given in days (flowering time), centimeters (length at first silique and total length) or numbers (elongated axils, basal branch number and total branch number). [f] Smallest relative effect significantly detected, expressed as fold difference compared to L*er*, calculated as $(|\mu_B-\mu_A|+\mu_A)/\mu_A$.

We defined the support interval in the NIL mapping population as the region spanned by consecutive bins, significantly differing from L*er* ($P < 0.001$) and sharing the same direction of effect. The length of support intervals estimated in this way ranged from 1.4 (total plant length) to 85.3 cM (basal branch number) with an average of 30.9 cM. Alternatively, we also searched for QTLs in the NIL population by comparing the phenotype of each NIL individually against L*er* (supplemental Figures 2-7 at http://www.genetics.org/supplemental/). In this case, support intervals can be estimated as the length of the overlapping regions between the Cvi introgression fragments of NILs significantly differing from L*er* in a particular genomic region. This second method increases the QTL localization resolution, but reduces statistical power. For each bin on average 116 plants could be tested against L*er* whereas only 24 plants were available for analysis of individual NILs. Moreover, individual lines may contain multiple opposing-effect QTLs, resulting in nonsignificant differences compared to L*er*. Therefore, lines spanning the bin support interval were occasionally not significantly different from L*er*. Likewise, lines bearing introgressions outside the bin support intervals sometimes differed significantly from L*er*, probably due to multiple additive small-effect QTLs. Together, the loss of power and the complexity of the traits under study hindered a confident estimation of a NIL support interval. Nevertheless, all QTLs detected in the bin analysis could also be detected by analyzing individual NILs. As a compromise between the two methods of support interval estimation we recorded the position of the largest-effect bin within the bin support interval (Table 4). However, it must be noted that bin support intervals may contain multiple QTLs of similar direction. The average size of these largest-effect bins was 4.6 cM. Within those bins, at least one individual NIL significantly differing from L*er,* was always found.

Power in RIL *vs.* NIL QTL mapping

The power to detect a QTL at a specific locus basically depends on the difference in mean trait values between A and B genotypes for that particular locus. Although other parameters like trait heritability, genetic interactions, and genetic map quality should not be ignored. Because power increases when variance for mean values decreases, QTL analyses can benefit greatly from multiple measurements. In a RIL population this can be achieved in two ways. First, because segregation of both alleles occurs randomly and each locus is represented equally by the A and B genotype, provided there is no segregation distortion (Doerge, 2002), increasing the number of RILs to be analyzed will increase the number of observations of each genotype at a given genomic position. A further advantage of increasing the RIL population size is that the number of recombination events increases, which can improve resolution. However, when the number of lines is fixed, more accurate trait values of lines can be achieved by measuring replicate individuals of the same line. In addition accurate trait values based on replicate measurements improve the possibility of detecting smaller-effect QTLs.

To test the effect of replicated measurements and population size on the QTL detection power of the two L*er* x Cvi populations we analyzed the phenotypic data obtained in these populations by varying both parameters. For the RIL population we performed QTL analyses on different numbers of RILs (population size) and used mean line values obtained with different number of replicates (replicate size). The total explained variance in the population, LOD score of the largest-effect QTL, and the number of detected QTLs were then recorded for each trait (Figure 4). When the population size was kept constant (161 lines), the recorded statistics increased when increasing the replicate number from one to four but this increase leveled off rapidly when measuring more replicates (Figure 4, A-C).

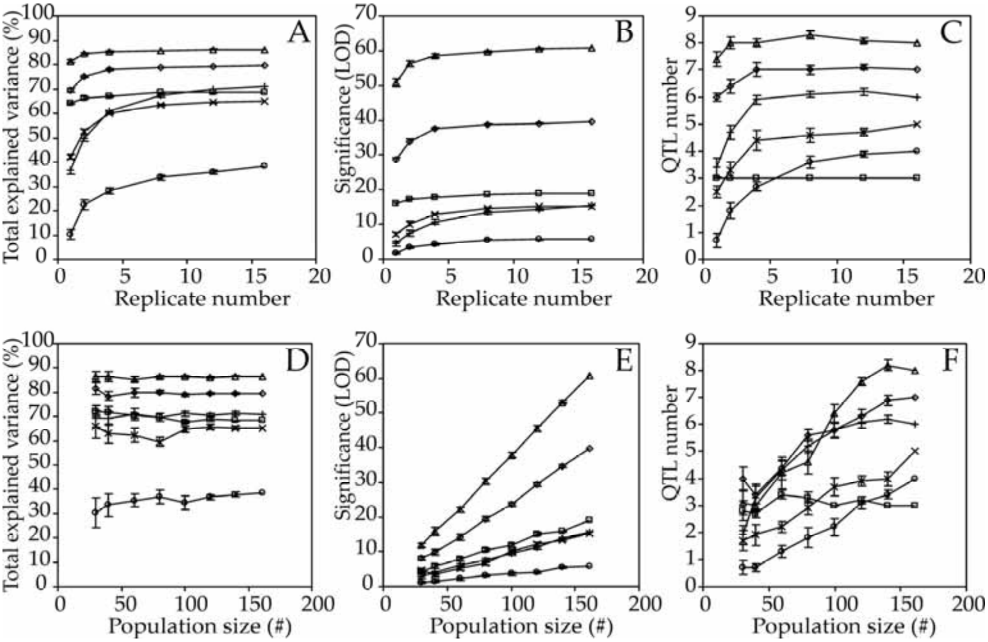


Figure 4: QTL detection power analysis of the L*er* x Cvi RIL population.
(A) Effect of replicate number on total explained variance. (B) Effect of replicate number on LOD score of the largest-effect QTL. (C) Effect of replicate number on the number of detected QTLs. (D) Effect of population size on total explained variance. (E) Effect of population size on LOD score of the largest-effect QTL. (F) Effect of population size on the number of detected QTLs. □, Flowering time; ◊, Length at first silique; Δ, Total plant length; x, Main inflorescence branch number; ○, Basal branch number; and +, Total branch number. Error bars represent SEM of ten independent analyses.

In contrast, when the number of replicates was kept constant (16 replicated measurements per RIL) and population size was increased, the QTL detection power improved more drastically. However, the total explained variance remained more or less constant over all population sizes (Figure 4D). This phenomenon is commonly known as the Beavis effect and is due to the fact that estimated explained variances of detected QTLs are sampled from a truncated distribution because QTLs are only taken into account when the test statistics reach a predetermined critical value (Xu, 2003). As a result, the expectations of detected QTL effects are biased upward. A second effect of increasing population size is the nearly linear increase of LOD scores, observed for all analyzed QTLs (Figure 4E). Significance thresholds determined by permutation tests for each population size, were steady around 2.7 LOD for population sizes >30 RILs and increased slightly with smaller population sizes (data not shown). The largest-effect QTL could be significantly detected at all population sizes for all traits except for basal branch number, whose largest-effect QTL could not be significantly detected in population sizes <80 RILs.

To evaluate the NIL population, we studied the effect of increasing the number of replicates per line by estimating the relative difference between line mean values that could still be significantly detected with different replicate numbers (see Materials and Methods). As shown in Figure 5A the power to detect significant phenotypic differences greatly increases when increasing the number of replicate individuals of NILs measured. Furthermore, the lower the heritability of the trait the larger the increase of detection power achieved by increasing the number of replicates per NIL. When a bin analysis was carried out using increasing replicate numbers a similar increase in the number of detected QTLs was observed (Figure 5B). Overall, the results presented in Figures 4 and 5 show that the number of replicates used in our analyses (16 individuals for each RIL and 24 individuals for each NIL) approximated the maximum QTL detection power of both L*er* x Cvi populations.

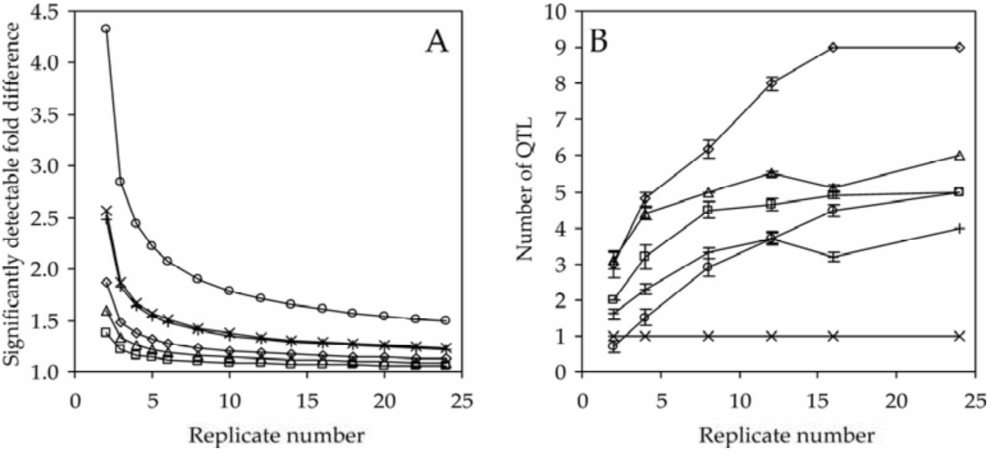


Figure 5: QTL detection power analysis of the L*er* x Cvi NIL population.

(A) Effect of replicate number on significantly detectable relative differences, expressed as fold difference between two lines. (B) Effect of replicate number on the number of detected QTLs. □, Flowering time; ◊, Length at first silique; Δ, Total plant length; x, Main inflorescence branch number; ○, Basal branch number; and +, Total branch number. Error bars represent SEM of ten independent analyses.

DISCUSSION

Experimental mapping populations are a basic resource to elucidate the genetic basis of quantitative multigenic traits. In this work, we have developed the first genome-wide population of NILs of *Arabidopsis thaliana* consisting of 92 lines carrying genomic introgression fragments from the parental accession Cvi into the genetic background of the common laboratory accession Landsberg *erecta*. In addition we have empirically compared the mapping power of this population with an existing population of recombinant inbred lines, derived from the same parental accessions. RIL and NIL populations have been used extensively in genetic studies (Eshed and Zamir, 1995; Rae *et al.*, 1999; Monforte and Tanksley, 2000; Koumproglou *et al.*, 2002; Han *et al.*, 2004; Koornneef *et al.*, 2004; Singer *et al.*, 2004; von Korff *et al.*, 2004) due to the advantages derived from their homozygosity and immortality: they can be used indefinitely; various traits can be analyzed in different experiments and environmental settings; and replicates of the individual lines can be analyzed, enabling a more accurate estimate of the line's phenotypic mean value. However, the main difference between the two populations lies in the nature of their genetic makeup. In a RIL population multiple genomic regions differ between most pairs of RILs and several segregating QTLs contribute to phenotypic differences between pairs of lines, making it impossible to assign the observed variation between pairs of lines to a specific genomic region. Therefore, to detect QTLs one must perform the simultaneous analysis of a large number of lines. In contrast, in a NIL population, the phenotypic variation observed between pairs of lines can be assigned directly to the distinct genomic regions introgressed in an otherwise similar genetic background. Depending on the desired resolution one can minimize the number of lines by analyzing lines carrying large introgressions or even chromosome substitution strains (Nadeau *et al.*, 2000).

A summary of the differences observed between the RIL and NIL populations derived from L*er* and Cvi is shown in Table 5 and in supplemental Figure 8 at http://www.genetics.org/supplemental/. The total number of QTLs detected did not differ much between the two populations. However, different loci were detected in both types of populations, showing their complementary properties. For both populations the detection of QTLs was highly dependent on the trait under consideration and its genetic architecture (*e.g.* effect and position of QTL, epistasis). The power of the new NIL population to detect the large-effect loci was close to that of the existing RIL population since most large-effect loci were detected in both populations. However, a few relatively large-effect loci showing significant epistatic interactions could only be detected in the RIL population, but not in the NILs (supplemental Table 3 at http://www.genetics.org/supplemental/). Moreover, localization resolution was higher in the RIL population compared to the bin analysis of the NIL population, allowing separation of linked QTLs. This was best illustrated by the two major QTLs for flowering time detected in the RIL population on the top of chromosome five, which not only are linked but also showed strong epistatic interaction. Consequently, these two QTLs could not be separated in the NIL population. Nevertheless, the QTL resolution in the NIL population can be increased when analyzing individual lines, although this will be at the cost of mapping power. In total, 14 QTLs detected in the RIL population could not be detected in the NIL population, of which 10 showed significant epistatic interactions with other QTLs and all others were closely linked to another significant QTL.

In contrast, the average explained variance of single QTLs was higher in the NIL population, increasing the power to detect small-effect QTLs. This difference can be attributed to the level of transgression, which is stronger in the RIL population, thereby increasing total phenotypic variance. As a result, 13 small-effect QTLs could be detected in the NIL population, which were not detected in the RIL population. Nevertheless, some of the small-effect QTLs detected in the NILs were close to the significance threshold in the RIL population when using the lower chromosomal LOD thresholds (supplemental Table 3 at http://www.genetics.org/supplemental/). Expectedly, the power to detect small-effect QTLs in the NIL population was higher for the more heritable traits (flowering time and plant height) than for those traits with low heritability (branching traits). The different power to detect small-effect QTLs of the two populations is due to the effect of the segregation of multiple QTLs in the RIL population, which increases the residual variance at each QTL under study.

The analyses of the RIL and NIL populations performed in this work were probably close to the maximum statistical power for the given population sizes since the number of detected QTLs leveled off at higher replication sizes (Figures 4 and 5). The power analyses presented here could guide the decision-making on the number of plants to be analyzed when experiments are costly, laborious, or time consuming and therefore may require the analysis of fewer plants. Overall, for RILs, the effect of population size on mapping power was larger than the effect of replicated measurements of individual lines. Therefore, to reduce the number of plants to be analyzed, it is preferable to first reduce the number of replicates per line, and only thereafter, if required, the number of lines. In our analyses major-effect QTLs for most traits could still be significantly detected when only 50 lines were analyzed without replicates (data not shown). However, due to the Beavis effect (Xu, 2003) the explained variances obtained with small population sizes were strongly overestimated. In the NIL population, the number of replicated measurements has a larger impact on mapping power and at least five replicated plants should be analyzed to obtain enough statistical power (Figure 5). However, fewer lines can be analyzed as long as genome-wide coverage is maintained. In this NIL population this can be achieved using a core set of 25 lines, although localization resolution was diminished. Nevertheless, most QTLs detected in the full set could still be detected in the core set (supplemental Figure 9 at http://www.genetics.org/supplemental/). Once a QTL has been identified in a particular region, one can zoom in with a minimal set of lines carrying smaller introgressions defined by crossovers in the support interval of the QTL of interest (Fridman *et al.*, 2004).

The L*er* x Cvi NIL population developed in this work provides a useful resource that will facilitate the genetic dissection of quantitative traits in Arabidopsis in various aspects. First, as shown here, it can be analyzed as an alternative segregating population to perform genome-wide QTL mapping, with the particular advantage of detecting small-effect QTLs. Second, this population can be used to confirm previously detected QTLs in the L*er* x Cvi RIL population. Third, individual lines of this population can serve as starting point for the rapid Mendelization of particular QTLs and for their fine mapping and cloning (Paran and Zamir, 2003). Finally, the single introgression lines of this population may also strongly facilitate the fine mapping of artificially induced mutant alleles in the common laboratory L*er* genetic background (or transferred to this accession). The fine mapping of mutant loci affecting quantitative adaptive traits is often hampered by the confounding effects of QTLs segregating in the mapping populations derived from crosses between the mutant and another Arabidopsis wild accession. Knowing the approximate genetic location of the mutant locus within a chromosomal arm, specific lines of this NIL population can be selected as carrying a single introgression spanning the map position of the locus of interest. These lines can then be used to derive the required monogenic mapping population, as has been illustrated with the flowering-time locus *FVE* (Ausin *et al.*, 2004). In conclusion, the elucidation of quantitative traits can benefit from the parallel analysis of both populations.

MATERIALS AND METHODS

Mapping populations

Two types of mapping populations were used to analyze six developmental traits. The first population consists of a set of 161 recombinant inbred lines (RILs) derived from a cross between the accessions Cape Verde Islands (Cvi) and Landsberg *erecta* (L*er*). The F_{10} generation has been extensively genotyped (Alonso-Blanco *et al.*, 1998b) and is available from the Arabidopsis Biological Resource Center. All lines were advanced to the F_{13} generation and residual heterozygous regions, estimated at 0.71% in the F_{10} generation, were genotyped again with molecular PCR markers to confirm that they were practically 100% homozygous.

The second population consists of a set of 92 near isogenic lines (NILs). NILs were generated by selecting appropriate L*er* x Cvi RILs and repeated backcrossing with L*er* as recurrent female parent. A number of these lines have been described previously (Alonso-Blanco *et al.*, 1998a, 2003; Swarup *et al.*, 1999; Bentsink *et al.*, 2003; Edwards *et al.*, 2005; Juenger *et al.*, 2005a; Teng *et al.*, 2005). The progeny of backcrosses was genotyped with PCR markers and lines containing a homozygous Cvi introgression into an otherwise L*er* background were selected. The set of selected lines was then extensively genotyped by AFLP analysis using the same restriction enzymes and primer combinations as those used for the genotyping of the RILs (Alonso-Blanco *et al.*, 1998b). The NILs will be made available through the Arabidopsis stock centers.

In both populations each line is almost completely homozygous and therefore individuals of the same line are genetically identical, which allows the pooling of replicated individuals and repeated measurements to obtain a more precise estimate of phenotypic values. For the RIL and NIL population 16 and 24 genetically identical plants were grown per line, respectively. Additionally, 96 replicates were grown for each parental accession L*er* and Cvi. All plants were grown in a single experiment with four completely randomized blocks containing 4, 6, and 24 replicates per RIL, NIL, and parent, respectively.

Plant growing conditions

Seeds were sown in petri dishes on water-soaked filter paper and incubated for five days in a cold room at 4°C in the dark to promote uniform germination. Subsequently, petri dishes were transferred to a climate chamber (24°C, 16 hr light per day) for two days before planting. Germinated seedlings were transferred to clay pots, placed in peat, containing a sandy soil mixture. A single plant per pot was grown under long-day light conditions in an air-conditioned green house from July until October. Plants were fertilized every two weeks using a liquid fertilizer.

Quantitative traits

A total number of six developmental traits, which were known to vary within the populations for the number of QTLs and heritability, were measured on all individuals. We quantified flowering time (FT); main inflorescence length at first silique (SL); total length of the main inflorescence (TL); basal branch number (BB), which is the number of side shoots growing out from the rosette; main inflorescence branch number (IB), which is the number of elongated axillary (secondary) inflorescences along the main inflorescence; and total number of side shoots (TB; basal plus main inflorescence). Flowering time was recorded as the number of days from the date of planting until the opening of the first flower. All other traits were measured at maturity.

Quantitative genetic analyses

For both populations and for each trait, total phenotypic variance was partitioned into sources attributable to genotype (V_G; *i.e.* the line effect) and error (V_E) using a random-effects analysis of variance (ANOVA, SPSS version 11.0) according to the model $Y = \mu + G + E$. Variance components were used to estimate broad sense heritability according to the formula $H^2 = V_G/(V_G + V_E)$, where V_G is the among-genotype variance component and V_E is the residual (error) variance component. Genetic correlations (r_G) were estimated as $r_G = \text{cov}_{1,2} / \sqrt{V_{G1} \times V_{G2}}$, where $\text{cov}_{1,2}$ is the covariance of trait means and V_{G1} and V_{G2} are the among-genotype variance components for those traits. The coefficient of genetic variation (CV_G) was estimated for each trait as $CV_G = \left(100 \times \sqrt{V_G}\right) / \overline{X}$, where V_G is the among-genotype variance component and $\overline{X}$ is the trait mean of the genotypes.

QTL analyses in the RIL population

To map QTLs using the RIL population, a set of 144 markers equally spaced over the Arabidopsis genetic map was selected from the RIL L*er* x Cvi map (Alonso-Blanco *et al.*, 1998b). These markers spanned 485 cM, with an average distance between consecutive markers of 3.5 cM and the largest genetic distance being 11 cM. The phenotypic values recorded, except basal branch number, were transformed ($\log_{10}(x+1)$) to improve the normality of the distributions and the values of 16 plants per RIL were used to calculate the means of each line for all traits. These means were used to perform the QTL analyses unless otherwise stated. The computer program MapQTL version 5.0 (Van Ooijen, 2004) was used to identify and locate QTLs linked to the molecular markers using both interval mapping and multiple QTL mapping (MQM). In a first step, putative QTLs were identified using interval mapping. Thereafter, a marker closely linked to each putative QTL was selected as a cofactor and the selected markers were used as genetic background controls in the approximate MQM of MapQTL. LOD statistics were calculated at 0.5 cM intervals. Tests of 1000 permutations were used to obtain an estimate of the number of type 1 errors (false positives). The genome-wide LOD score, which 95% of the permutations did not exceed, ranged from 2.6 to 2.8 and chromosome-wide LOD thresholds varied between 1.8 and 2.1 depending on trait and linkage group. The genome-wide LOD score was then used as the significance threshold to declare the presence of a QTL in MQM mapping, while the chromosome-wide thresholds were used to detect putative small-effect QTLs. In the final MQM model the genetic effect (μ_B-μ_A) and percentage of explained variance was estimated for each QTL and 2-LOD support intervals were established as an ~95% confidence level (Van Ooijen, 1992), using restricted MQM mapping.

Epistatic interactions between QTLs were estimated using factorial analysis of variance. For each trait, the mean phenotypic values were used as dependent variable and cofactors, corresponding to the detected QTLs, were used as fixed factors. The general linear model module of the statistical package SPSS version 11.0 was used to perform a full factorial analysis of variance or analysis of main effects only. Differences in R^2-values, calculated from the Type III sum of squares, were assigned to epistatic interaction effects of detected QTLs. Additionally we performed a complete pairwise search ($P < 0.001$, determined by Monte Carlo simulations) for conditional and coadaptive epistatic interactions for each trait using the computer program EPISTAT (Chase *et al.*, 1997).

The effect of replication on statistical power was analyzed by performing MQM mapping on means of trait values from 1, 2, 4, 8, 12, and 16 replicate plants, respectively. Analyses were performed on ten independent, stochastically sampled, data sets for each replication size and trait using automated cofactor selection ($P < 0.02$). Total explained variance, LOD score of the largest-effect QTL, and number of significant QTLs were recorded for each analysis.

The effect of population size on statistical power was analyzed by performing MQM mapping on increasing population sizes. Analyses were performed on ten independent, stochastically sampled, data sets for each population size. Subpopulations of increasing size, with a step size of 20 lines, were analyzed for each trait using automated cofactor selection ($P < 0.02$). Total explained variance, LOD score of the largest-effect QTL, and number of significant QTLs were recorded for each analysis.

Statistical analyses NILs

Differences in mean trait values of L*er* and NILs were analyzed by univariate analysis of variance, using the general linear model module of the statistical package SPSS version 11.0. Dunnett's pairwise multiple comparison *t*-test was used as a *post hoc* test to determine significant differences. For each analysis, trait values were used as dependent variable and NILs were used as fixed factor. Tests were performed 2-sided with a Bonferroni-corrected significance threshold level of 0.05 and L*er* as control category. In order to increase statistical power, similar analyses were conducted for bins (see results section). For this, trait values of all introgression lines assigned to a certain bin were pooled and compared to values of the L*er* parental line. Because each NIL can be a member of more than one bin the significance threshold was lowered to 0.001 to correct for multiple testing. The genetic effect of Cvi bins significantly differing from L*er* was calculated as μ_B-μ_A, where μ_A and μ_B are the mean trait values of L*er* and the Cvi bin, respectively. Explained variance was estimated from the partial η^2 of the univariate analysis of variance, where η^2 is the proportion of total variance attributable to factors in the analysis. The total percentage of explained variance was then estimated by using trait values as dependent variable and NILs as fixed factor, where all NILs where included as subjects. The percentage of explained variance of individual QTLs was estimated as a fraction of the total variation in the population (including all lines), using a single bin as fixed factor and as a fraction of the total variation in a comparison of a single bin with L*er* only.

To determine the effect of replicated measurements we calculated the power of detecting significant differences between L*er* and NILs using various replicate numbers. For each trait we calculated the minimal relative difference in mean trait values that could still be significantly detected. Calculations were performed using a normal distribution two-sample equal variance power calculator from the UCLA department of statistics (http://calculators.stat.ucla.edu/). We first calculated for each trait the mean phenotypic value of 96 L*er* replicate plants (μ_A) and for each line the standard deviation of 24 replicate plants. The mean line standard deviation of each trait was taken as a measure of variation (σ) in all subsequent calculations. The significance level, the probability of falsely rejecting the null hypothesis (H_0:μ_A=μ_B) when it is true, was set to 0.05 and power, the probability of correctly rejecting the null hypothesis when the alternative (H_1:$\mu_A \neq \mu_B$) is true, was set to 0.95. The sample size of L*er* (N_A) was always identical to the sample size of NILs (N_B) and ranged from 2 to 24 individuals. For each trait and sample size the mean trait value (μ_B) for NILs was then calculated as the minimum value to meet the alternative hypothesis (H_1:$\mu_A \neq \mu_B$) in a two-sided test. These minimum values were then converted in a fold-difference value compared to the L*er* value, calculated as $(|\mu_B-\mu_A|+\mu_A)/\mu_A$, to obtain a relative estimate independent of trait measurement units.

The effect of replication on statistical power was also analyzed by performing bin mapping using 2, 4, 8, 12, and 16 replicate plants, respectively. Analyses were performed on ten independent, stochastically sampled, data sets for each replication size and trait and the number of significant QTLs was recorded for each analysis.

Acknowledgements

We thank Kieron Edwards for sharing NILs, Johan van Ooijen for helpful assistance in the QTL mapping, and Piet Stam for critical reading of the manuscript. This work was supported by a grant from The Netherlands Organization for Scientific Research, Program Genomics (050-10-029).

REFERENCES

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

Alonso-Blanco, C., Blankestijn-de Vries, H., Hanhart, C.J. and Koornneef, M. (1999). Natural allelic variation at seed size loci in relation to other life history traits of *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A* **96,** 4710-4717.

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., Bentsink, L., Hanhart, C.J., Blankestijn-de Vries, H. and Koornneef, M. (2003). Analysis of natural allelic variation at seed dormancy loci of *Arabidopsis thaliana*. *Genetics* **164,** 711-729.

Ausin, I., Alonso-Blanco, C., Jarillo, J.A., Ruiz-Garcia, L. and Martinez-Zapater, J.M. (2004). Regulation of flowering time by FVE, a retinoblastoma-associated protein. *Nat Genet* **36,** 162-166.

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.

Blair, M.W., Iriarte, G. and Beebe, S. (2006). QTL analysis of yield traits in an advanced backcross population derived from a cultivated Andean x wild common bean (*Phaseolus vulgaris* L.) cross. *Theor Appl Genet* **112,** 1149-1163.

Blanco, A., Simeone, R. and Gadaleta, A. (2006). Detection of QTLs for grain protein content in durum wheat. *Theor Appl Genet* **112,** 1195-1204.

Borevitz, J.O. and Nordborg, M. (2003). The impact of genomics on the study of natural variation in Arabidopsis. *Plant Physiol* **132,** 718-725.

Broman, K.W. (2001). Review of statistical methods for QTL mapping in experimental crosses. *Lab Anim (NY)* **30,** 44-52.

Chase, K., Adler, F.R. and Lark, K.G. (1997). Epistat: a computer program for identifying and testing interactions between pairs of quantitative trait loci. *Theor Appl Genet* **94,** 724-730.

Doerge, R.W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet* **3,** 43-52.

Edwards, K.D., Lynn, J.R., Gyula, P., Nagy, F. and Millar, A.J. (2005). Natural allelic variation in the temperature-compensation mechanisms of the *Arabidopsis thaliana* circadian clock. *Genetics* **170,** 387-400.

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.

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.

Han, F., Ullrich, S.E., Kleinhofs, A., Jones, B.L., Hayes, P.M. and Wesenberg, D.M. (1997). Fine structure mapping of the barley chromosome-1 centromere region containing malting-quality QTLs. *Theor Appl Genet* **95,** 903-910.

Han, F., Clancy, J.A., Jones, B.L., Wesenberg, D.M., Kleinhofs, A. and Ullrich, S.E. (2004). Dissection of a malting quality QTL region on chromosome 1 (7H) of barley. *Mol Breed* **14,** 339-347.

Jansen, R.C. (2003). Quantitative trait loci in inbred lines. In Handbook of Statistical Genetics, D.J. Balding, M. Bishop and C. Cannings, eds (Chichester, UK: John Wiley & Sons), pp. 445-476.

Jeuken, M.J. and Lindhout, P. (2004). The development of lettuce backcross inbred lines (BILs) for exploitation of the *Lactuca saligna* (wild lettuce) germplasm. *Theor Appl Genet* **109,** 394-401.

Juenger, T.E., McKay, J.K., Hausmann, N., Keurentjes, J.J.B., Sen, S., Stowe, K.A., Dawson, T.E., Simms, E.L. and Richards, J.H. (2005a). Identification and characterization of QTL underlying whole-plant physiology in *Arabidopsis thaliana*: delta13C, stomatal conductance and transpiration efficiency. *Plant Cell Environ* **28,** 697-708.

Juenger, T.E., Sen, S., Stowe, K.A. and Simms, E.L. (2005b). Epistasis and genotype-environment interaction for quantitative trait loci affecting flowering time in *Arabidopsis thaliana*. *Genetica* **123,** 87-105.

Koornneef, M., Alonso-Blanco, C. and Vreugdenhil, D. (2004). Naturally occurring genetic variation in *Arabidopsis Thaliana*. *Annu Rev Plant Physiol Plant Mol Biol* **55,** 141-172.

Koumproglou, R., Wilkes, T.M., Townson, P., Wang, X.Y., Beynon, J., Pooni, H.S., Newbury, H.J. and Kearsey, M.J. (2002). STAIRS: a new genetic resource for functional genomic studies of Arabidopsis. *Plant J* **31,** 355-364.

Loudet, O., Gaudon, V., Trubuil, A. and Daniel-Vedele, F. (2005). Quantitative trait loci controlling root growth and architecture in *Arabidopsis thaliana* confirmed by heterogeneous inbred family. *Theor Appl Genet* **110,** 742-753.

Maloof, J.N. (2003). Genomic approaches to analyzing natural variation in *Arabidopsis thaliana*. *Curr Opin Genet Dev* **13,** 576-582.

Monforte, A.J. and Tanksley, S.D. (2000). Development of a set of near isogenic and backcross recombinant inbred lines containing most of the *Lycopersicon hirsutum* genome in a *L. esculentum* genetic background: a tool for gene mapping and gene discovery. *Genome* **43,** 803-813.

Nadeau, J.H., Singer, J.B., Matin, A. and Lander, E.S. (2000). Analysing complex genetic traits with chromosome substitution strains. *Nat Genet* **24,** 221-225.

Paran, I. and Zamir, D. (2003). Quantitative traits in plants: beyond the QTL. *Trends Genet* **19,** 303-306.

Rae, A.M., Howell, E.C. and Kearsey, M.J. (1999). More QTL for flowering time revealed by substitution lines in *brassica oleracea*. *Heredity* **83 (Pt 5),** 586-596.

Reymond, M., Svistoonoff, S., Loudet, O., Nussaume, L. and Desnos, T. (2006). Identification of QTL controlling root growth response to phosphate starvation in *Arabidopsis thaliana*. *Plant Cell Environ* **29,** 115-125.

Singer, J.B., Hill, A.E., Burrage, L.C., Olszens, K.R., Song, J., Justice, M., O'Brien, W.E., Conti, D.V., Witte, J.S., Lander, E.S. *et al.* (2004). Genetic dissection of complex traits with chromosome substitution strains of mice. *Science* **304,** 445-448.

Slate, J. (2005). Quantitative trait locus mapping in natural populations: progress, caveats and future directions. *Mol Ecol* **14,** 363-379.

Stylianou, I.M., Tsaih, S.W., DiPetrillo, K., Ishimori, N., Li, R., Paigen, B. and Churchill, G. (2006). Complex genetic architecture revealed by analysis of high-density lipoprotein cholesterol in chromosome substitution strains and F2 crosses. *Genetics* **174,** 999-1007.

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.

Teng, S., Keurentjes, J.J.B., Bentsink, L., Koornneef, M. and Smeekens, S. (2005). Sucrose-specific induction of anthocyanin biosynthesis in Arabidopsis requires the MYB75/PAP1 gene. *Plant Physiol* **139,** 1840-1852.

Tuinstra, M.R., Ejeta, G. and Goldsbrough, P.B. (1997). Heterogeneous inbred family (HIF) analysis: a method for developing near-isogenic lines that differ at quantitative trait loci. *Theor Appl Genet* **95,** 1005-1011.

Ungerer, M.C., Halldorsdottir, S.S., Modliszewski, J.L., Mackay, T.F. and Purugganan, M.D. (2002). Quantitative trait loci for inflorescence development in *Arabidopsis thaliana*. *Genetics* **160,** 1133-1151.

Ungerer, M.C., Halldorsdottir, S.S., Purugganan, M.D. and Mackay, T.F. (2003). Genotype-environment interactions at quantitative trait loci affecting inflorescence development in *Arabidopsis thaliana*. *Genetics* **165,** 353-365.

Van Ooijen, J.W. (1992). Accuracy of mapping quantitative trait loci in autogamous species. *Theor Appl Genet* **84,** 803-811.

Van Ooijen, J.W. (2004). MapQTL 5, Software for the mapping of quantitative trait loci in experimental populations (Wageningen, The Netherlands: Kyazma B.V.).

von Korff, M., Wang, H., Leon, J. and Pillen, K. (2004). Development of candidate introgression lines using an exotic barley accession (*Hordeum vulgare ssp. spontaneum*) as donor. *Theor Appl Genet* **109,** 1736-1745.

Xu, S. (2003). Theoretical basis of the Beavis effect. *Genetics* **165,** 2259-2268.

Yoon, D.B., Kang, K.H., Kim, H.J., Ju, H.G., Kwon, S.J., Suh, J.P., Jeong, O.Y. and Ahn, S.N. (2006). Mapping quantitative trait loci for yield components and morphological traits in an advanced backcross population between *Oryza grandiglumis* and the *O. sativa japonica* cultivar Hwaseongbyeo. *Theor Appl Genet* **112,** 1052-1062.

Zou, F., Gelfond, J.A., Airey, D.C., Lu, L., Manly, K.F., Williams, R.W. and Threadgill, D.W. (2005). Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations. *Genetics* **170,** 1299-1311.

Chapter 3

Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci

Joost J. B. Keurentjes*, Jingyuan Fu*, Inez R. Terpstra*, Juan M. Garcia, Guido van den Ackerveken, L. Basten Snoek, Anton J. M. Peeters, Dick Vreugdenhil, Maarten Koornneef and Ritsert C. Jansen

Published in Proceedings of the National Academy of Sciences USA (2007) **104**, 1708-1713.

* Equal contribution.

ABSTRACT

Accessions of a plant species can show considerable genetic differences that are effectively analyzed using Recombinant Inbred Line (RIL) populations. Here we describe the results of genome wide expression variation analysis in an RIL population of *Arabidopsis thaliana*. For many genes, variation in expression could be explained by expression Quantitative Trait Loci (eQTLs). The nature and consequences of this variation are discussed based on additional genetic parameters, such as heritability and transgression and by examining the genomic position of eQTLs versus gene position, polymorphism frequency, and gene ontology. Furthermore, we developed a novel approach for genetic regulatory network construction by combining eQTL mapping and regulator candidate gene selection. The power of our method was shown in a case study of genes associated with flowering time, a well studied regulatory network in Arabidopsis. Results that revealed clusters of co-regulated genes and their most likely regulators were in agreement with published data, and unknown relationships could be predicted.

INTRODUCTION

Analogous to classical traits, quantitative genetic variation is often observed for transcript levels of genes. Jansen and Nap (2001), therefore, introduced the concept of genetical genomics, in which Quantitative Trait Locus (QTL) analysis is applied to levels of transcript abundance and identifies genomic loci controlling the observed variation in expression (eQTLs). One of the best studied organisms with regard to gene expression regulation nowadays is yeast (Brem *et al.*, 2002, 2005; Yvert *et al.*, 2003; Bing and Hoeschele, 2005; Brem and Kruglyak, 2005; Ronald *et al.*, 2005; Storey *et al.*, 2005). However, in recent years several studies have demonstrated the feasibility of this approach in different organisms and diverse types of populations (Brem *et al.*, 2002; Schadt *et al.*, 2003; Morley *et al.*, 2004; Bystrykh *et al.*, 2005; Hubner *et al.*, 2005; DeCook *et al.*, 2006).

A logical next step would be the construction of genetic regulatory networks (Kendziorski and Wang, 2006), which only a few studies have addressed up to now (Bing and Hoeschele, 2005; Kliebenstein *et al.*, 2006). Although many studies on higher eukaryotes suffered from small populations or only analyzed a subset of genes present on the genome of the organism under study, the main reason holding back the identification of gene-by-gene regulation has been the lack of a reliable identification of candidate regulators. Although powerful in detecting loci controlling the observed variation for trait values, support intervals of QTLs are still of considerable width, often covering hundreds of genes. Consequently, the molecular dissection of quantitative trait regulation is still in its infancy and would greatly benefit from approaches reducing the number of candidate genes in a QTL support interval.

Promising results have been obtained by combining QTL analyses of physiological and gene expression traits, based on co-localization of (e)QTLs (Wayne and McIntyre, 2002; Hubner *et al.*, 2005; DeCook *et al.*, 2006). However, when expression differences in genes are caused by differences in expression of their regulator, it is likely that they show correlation in expression (Bing and Hoeschele, 2005). Moreover, multiple functionally related genes with co-inciding eQTLs, which might be members of a common pathway, are likely to have one and the same regulator. We therefore developed a novel approach for the assignment of maximum-likelihood regulators by combining QTL analysis of gene expression profiling and iterative Group Analysis (iGA) (Breitling *et al.*, 2004) of functionally related genes with co-inciding eQTLs.

To apply the concept of genetical genomics to higher plants we analyzed genome-wide gene expression variation in a large, well-studied Recombinant

Inbred Line (RIL) population of *Arabidopsis thaliana*. We show that for many genes the variation in transcript level can be explained by genetic factors. By integrating current knowledge of the genetics of a specific trait, we demonstrate the construction of genetic regulatory networks, which can serve to form hypotheses about as-yet-unknown regulatory steps.

RESULTS

Genetic control of gene expression in plants is highly complex

To determine the effect of genetic factors involved in the regulation of expression, we analyzed genome-wide gene expression in the parents and an RIL population of a cross between the distinct accessions Landsberg *erecta* (L*er*) and Cape Verde Islands (Cvi), consisting of 160 lines (Alonso-Blanco *et al.*, 1998b). Transcript levels of 24,065 genes were analyzed by DNA microarrays, of which 922 showed significant differential expression between the parents [$P < 2.5 \times 10^{-3}$; false-discovery rate (FDR) = 0.05]. Subsequent mapping resulted in 4,523 eQTLs detected for 4,066 genes ($P < 5.29 \times 10^{-5}$; FDR = 0.05, corresponding to a q value of 0.01) (Storey and Tibshirani, 2003).

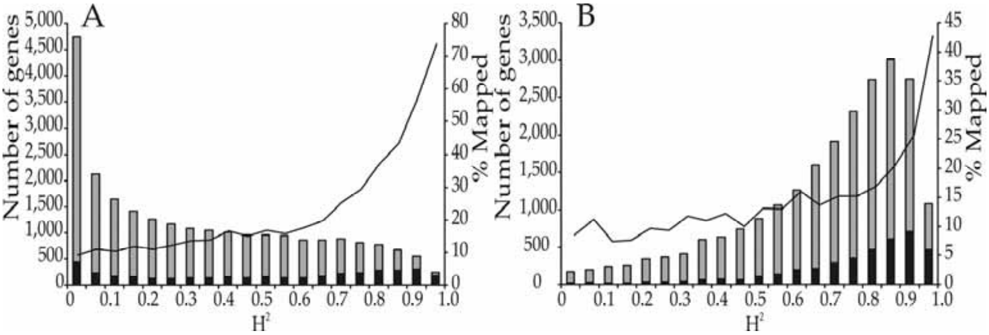


Figure 1: Frequency distributions of heritability values of gene expression.
(A) Data from a microarray comparison of the parents. (B) Data from a microarray analysis of the L*er* x Cvi RIL population. Solid and shaded bars represent the number of genes that could and could not be mapped, respectively. The solid line depicts the number of mapped genes as a proportion of the total number of genes for a given heritability class.

Because the microarray probe set was designed on the sequenced accession Columbia (Col), we performed hybridizations of genomic DNA of the parental lines and found relatively few hybridization differences (supplemental Table 1 at www.pnas.org/cgi/content/full/0610429104/DC1). However, the low power to detect differences, due to the small number of replicates, might have led to an under estimation, as indicated by other studies (Borevitz, 2006).

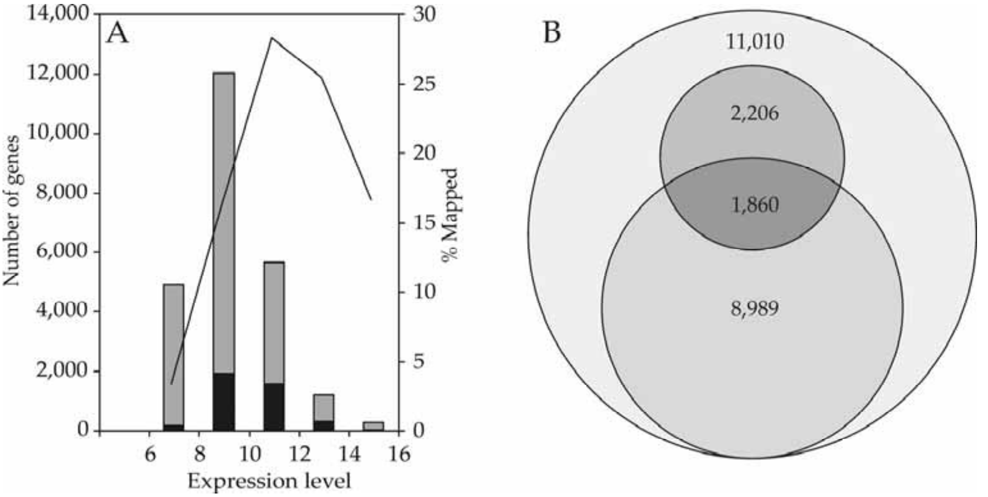


Figure 2: Effect of expression level and transgression on eQTL detection.
(A) Frequency distribution of the mean expression level of analyzed genes in the RIL population. Solid and shaded bars represent the number of genes that could and could not be mapped, respectively. The solid line depicts the number of mapped genes as a proportion of the total number of genes for a given class. (B) Diagram of the number of genes showing linkage and transgression. Circles are proportional to the number of genes. Increasing shading represents, respectively, the total number of genes analyzed (24,065), the number of genes whose expression showed significant linkage (4,066) and the number of genes whose expression showed transgressive segregation (10,849).

Heritability values calculated from the parental data and the RIL population reached a median value of 28.6 and 74.7%, respectively (Figure 1), which is in agreement with the discrepancy between the number of differentially expressed and mapped genes (*i.e.* genes for which an eQTL was found). Although the fraction of mapped genes increased with higher heritability values, for many genes showing high heritability, no eQTL could be significantly detected. These findings suggest that the regulation of expression of many genes is controlled by multiple eQTLs, of which many might not have passed the significance test because of their small effect. Likewise, only 65.6% of the genes differentially expressed between the parents could be mapped. However, for 15.0% of the genes for which the parents did not show a significant difference in expression levels, eQTLs could be detected. These observations and the much lower heritabilities calculated from the parental data, compared with those from the RIL population, indicate that eQTLs for a given gene might exert opposite additive effects, leading to a balanced expression in the parents but a transgressive expression pattern among the segregants of the population. To test this hypothesis, we tested each gene for significant transgression and found significant transgression of expression for 10,849 genes (45.1%). No relationship was found between the number of mapped genes and transgression (Figure 2B). These data indicate that the regulation of gene expression in plants is largely under genetic control but is highly complex because of the involvement of multiple genes.

Distribution of eQTLs identifies regulatory hot spots

To characterize in more detail the genes whose expression showed significant linkage, we determined several features. We first analyzed the distribution of eQTLs along the genome of Arabidopsis and found a number of genomic regions containing numbers of eQTLs significantly deviating from what can be expected by chance, as determined by permutation tests (Figure 3). These hot spots may reflect local gene-dense regions, in contrast to cold spots, which may reflect low-gene-density regions such as centromers. Alternatively, hot spots may contain master regulators: genes controlling the expression of many other genes. The large number of genes mapping to the *ERECTA* gene, which was included as a phenotypic marker, illustrate this finding. An empirical threshold for assessing a hot spot, providing a 0.05 genome-wide error rate was generated using a permutation procedure, which defined a hot spot as any marker associated with 43 or more genes. Because 176 genes mapped to the *ERECTA* marker, this locus was considered to be an eQTL hot spot. Polymorphisms in *ERECTA*, a receptor protein kinase (Torii *et al.*, 1996), are well known for their pleiotropic effect on many traits, including morphological differences (Koornneef *et al.*, 2004).

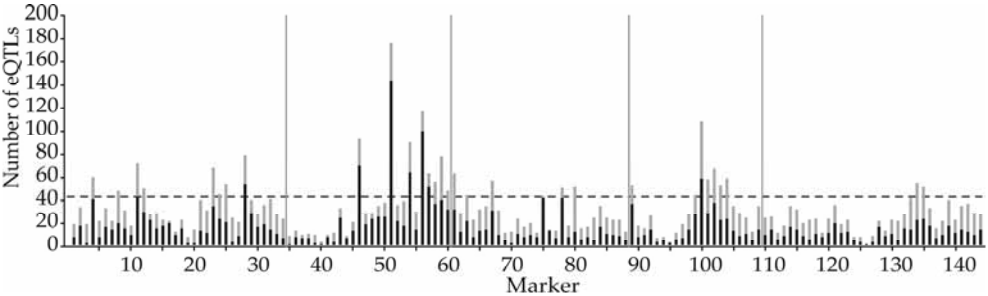


Figure 3: Genomic distribution of eQTLs.
Bars represent the number of distant (solid) and local (shaded) eQTLs detected at each marker position. Each eQTL was positioned at its best controlling marker. The dashed horizontal line represents the significance threshold value for defining a hot spot. Shaded vertical lines depict chromosomal borders.

Distant gene expression regulation occurs more frequently but local regulation is stronger

Genomic differences responsible for eQTLs either occur in regulatory genes affecting the transcript level of other genes (*trans*-regulation) or in the genes encoding the mRNA for which the eQTL was found (*cis*-regulation) (Rockman and Kruglyak, 2006). To compare the position of genes and their eQTLs, we anchored the genetic map to the physical map and found an almost linear genome-wide relation of 4.1 cM per Mbp (supplemental Figure 8 at www.pnas.org/cgi/content/full/0610429104/DC1). When the position of each eQTL was plotted against the position of the gene for which that eQTL was found, a strong enrichment along the diagonal of the graph was observed (Figure 4). This enrichment indicates that many genes, of which the majority is expected to be *cis*-regulated (Ronald *et al.*, 2005), map to their own physical position.

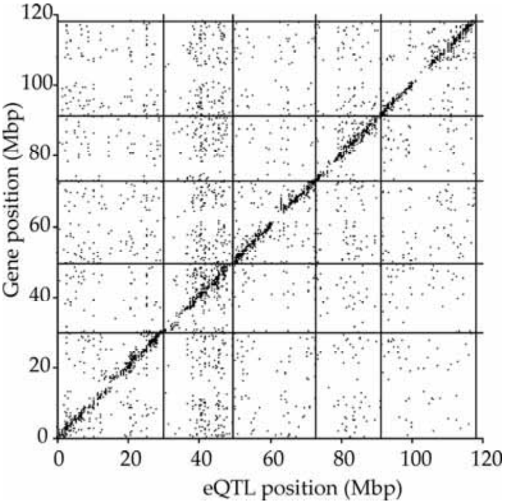


Figure 4: Distribution of mapped genes versus the position of their accompanying eQTL. Positions of detected eQTLs are plotted against the position of the gene for which that eQTL was found. Chromosomal borders are depicted as horizontal and vertical lines. Mbp, megabase pairs.

To quantify this result, we defined local/distant regulation in terms of the positional coincidence of genes and their accompanying eQTL(s). Of 4,066 mapped genes, 1,875 (46.1%) co-located with the support interval of one of their eQTLs, corresponding to a region consistent with max{-Log$_{10}$*P*} - 1.5 (where *P* expresses the significance of association (Keurentjes *et al.*, 2006)), and were therefore classified as locally regulated. Genes outside such intervals (1,958; 48.1%) were classified as distantly regulated. A minor number of 198 genes (4.9%) with multiple eQTLs showed both local and distant regulation, whereas the physical position of 35 genes (0.9%) was unknown (Table 1).

Table 1: The number of genes showing linkage, classified according to the position of eQTLs relative to the gene. Shown are the number of genes with a single or multiple eQTL(s) for different significance thresholds (P) and eQTL support intervals (max{-Log$_{10}P$} - x, where x = 1.5 and 2.0 respectively).

Position	Single eQTL	Multiple eQTLs
$P < 5.29 \times 10^{-5}$; max{-Log$_{10}P$} - 1.5		
Local	1875	
Distant	1752	206
Local + distant		198
Unknown	31	4
$P < 6.50 \times 10^{-4}$; max{-Log$_{10}P$} - 1.5		
Local	2167	
Distant	3671	916
Local + distant		794
Unknown	45	11
$P < 5.29 \times 10^{-5}$; max{-Log$_{10}P$} - 2.0		
Local	2007	
Distant	1676	156
Local + distant		192
Unknown	31	4

Because *cis*-regulation is often much stronger than *trans*-regulation (Bing and Hoeschele, 2005), as also indicated by the median –Log$_{10}P$ values of 7.1 and 5.3 and the median explained variance of 30.3 and 22.6% for local and distant eQTLs, respectively, the ratio of detected local versus distant eQTLs depends on the applied significance threshold (Schadt *et al.*, 2003; Morley *et al.*, 2004; Hubner *et al.*, 2005). The stringent threshold applied here, corrected for multiple testing, might therefore have underestimated distant regulation. When the threshold was decreased from 5.29×10^{-5} to 6.5×10^{-4} (FDR = 0.25, q = 0.05), 7,604 transcripts showed at least one linkage, with 2,167 (28.5%) being locally regulated, 4,587 (60.3%) being distantly regulated, and 794 (10.4%) being both locally and distantly regulated. Based on their P-value distributions (Storey and Tibshirani, 2003), the overall proportion of locally and distantly regulated genes were estimated at 40.5 and 15.3% respectively.

A second parameter affecting the assignment of locally versus distantly regulated transcripts is the setting of the eQTL support interval. However, when a wider interval of max{–Log$_{10}P$} - 2.0 was used at $P < 5.29 \times 10^{-5}$, results were similar with 2,007 (49.4%), 1,832 (45.1%), and 192 (4.7%) genes classified as locally, distantly, and both locally and distantly regulated, respectively.

Local regulation correlates with SNP frequency and is less frequent in regulatory genes

To determine whether a relationship exists between SNP or gene density and the number of mapped genes, we performed a sliding-window regression analysis. A strong correlation was observed between gene density and the number of locally and distantly regulated genes ($r^2 = 0.88$, $P < 0.0001$ and $r^2 = 0.91$, $P < 0.0001$, respectively) (Figure 5A).

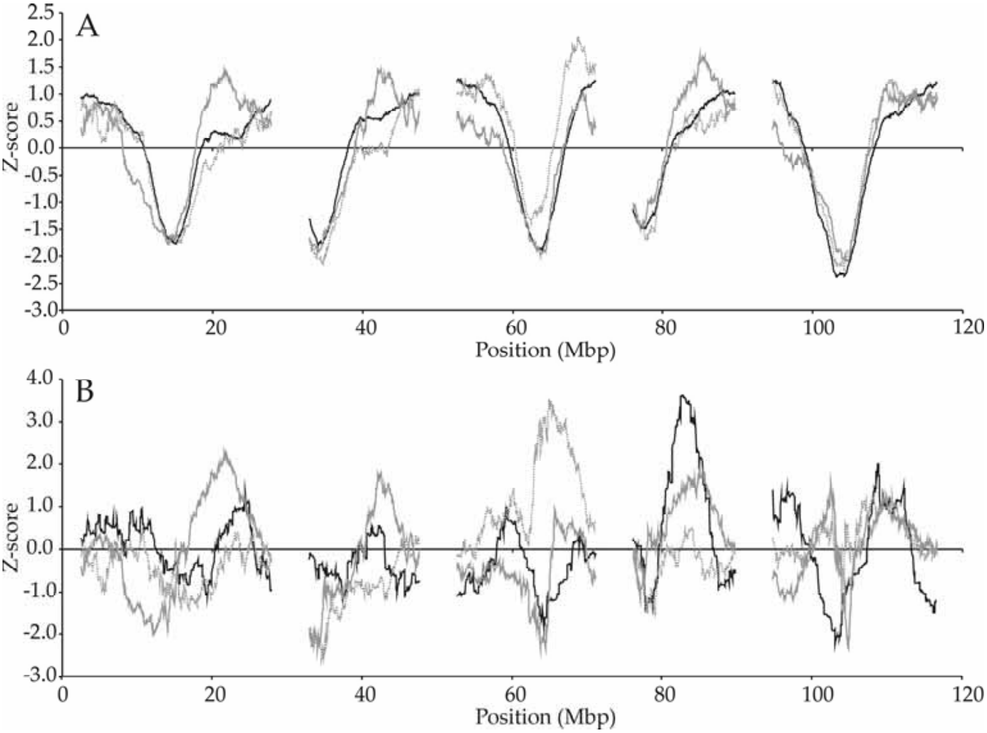


Figure 5: Relationship between gene and SNP frequency and the number of mapped genes.
(A) Relationship between gene frequency (solid lines) and the number of mapped genes, divided in locally (shaded lines) and distantly (dotted shaded lines) regulated genes. (B) Relationship between SNP frequency (solid lines) and the number of mapped genes, divided in locally (shaded lines) and distantly (dotted shaded lines) regulated genes, corrected for gene density. Gaps represent chromosomal borders. Mbp, megabase pairs.

A weaker but significant correlation was also found between gene and SNP frequency ($r^2 = 0.34$, $P < 0.0001$). Even when the number of mapped genes in a window was corrected for gene density, a significant correlation was still found between SNP frequency and the number of locally regulated genes ($r^2 = 0.32$, $P <$

0.0001), although incidental differences in hybridization efficiency might have contributed to an overestimation. Such a relationship was not found for distantly regulated genes ($r^2 = -0.003$, $P = 0.89$) (Figure 5B).

To assess whether there was a functional enrichment for genes whose variation in expression could be genetically explained, we computed the proportion of these genes for each Gene Ontology biological process and molecular function category (The Arabidopsis Information Resource; www.arabidopsis.org) (Figure 6). Genes involved in regulatory processes showed significantly less genetically explainable variation in expression (Al-Shahrour *et al.*, 2004) (supplemental Table 3 at www.pnas.org/cgi/content/full/0610429104/DC1). However, small changes in expression level, which may be more frequent in regulatory genes, are more difficult to detect but can nevertheless be very relevant biologically, because they may result in large changes in expression of target genes. Furthermore, many regulatory genes often display pleiotropic effects. A change in expression of such key regulators can affect the expression of many more target genes, which may skew the distribution of differently expressed genes in favor of classes containing predominantly target genes.

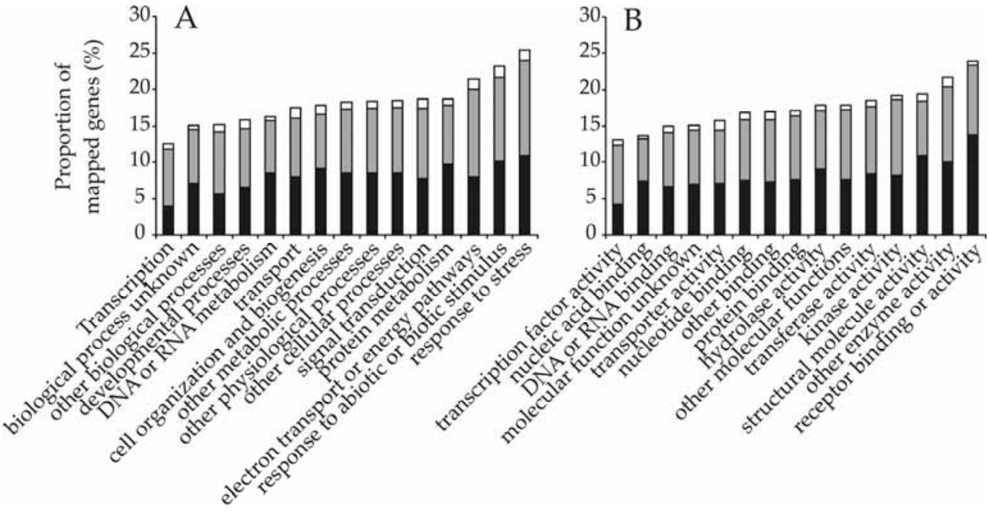


Figure 6: Frequency distribution of the proportion of mapped genes versus function.
(A) Proportion of genes that could be mapped in different Gene Ontology categories of biological processes. (B) Proportion of genes that could be mapped in different Gene Ontology categories of molecular functions. Solid, shaded, and white bars represent local, distant, and both local and distant regulation.

Interestingly, when these analyses were performed separately for locally and distantly regulated genes, regulatory categories showed a comparable proportion of distantly regulated genes with other classes but a much smaller proportion of locally regulated genes (Figure 6). Comparing locally to distantly regulated genes (Al-Shahrour *et al.*, 2004) resulted in significant overrepresentation of distantly regulated genes in ten Gene Ontology biological process categories, all involved in regulation. Only one category was detected in which locally regulated genes were overrepresented (Table 2). This finding agrees with the general assumption that regulatory genes are much more strongly conserved than other genes because of their often pleiotropic effects.

Table 2: Gene ontology categories with significantly different proportions of locally versus distantly regulated genes. The second and fourth column represent, for each category respectively, how many genes of the test set were locally and distantly regulated. The third and fifth column represent, for each category respectively, the proportion of the total number of annotated genes in the test set that were locally and distantly regulated. The sixth column represents, for each category, the *P*-value of observed differences between locally and distantly regulated genes. n.s., not significant.

	Local		Distant		
Gene Ontology category	Genes	%	Genes	%	*P*-value
Biological process					
regulation of cellular process	69	7.6	135	14.4	1.68E-03
regulation of cellular metabolism	58	7.2	120	14.2	1.68E-03
regulation of nucleic acid metabolism	57	8.2	118	16.1	1.68E-03
regulation of transcription	56	10.4	117	20.7	1.68E-03
regulation of metabolism	59	7.0	121	13.4	2.02E-03
regulation of cellular physiological process	69	8.2	135	15.0	2.02E-03
transcription	63	9.1	123	16.8	2.52E-03
regulation of physiological process	74	8.2	136	14.5	2.99E-03
RNA processing	31	5.7	9	1.6	3.86E-02
transcription, DNA-dependent	29	5.4	65	11.5	3.86E-02
regulation of transcription, DNA-dependent	29	8.6	64	17.9	3.86E-02
Molecular function					
transcription factor activity	64	6.2	125	11.5	1.60E-02
Cellular component					
n.s.					

A dual approach for the construction of regulatory networks reveals novel regulatory steps for flowering time

Genetic regulatory networks consist of a collection of genes, which are interconnected because one gene regulates the transcription of another directly or indirectly. The analysis of gene expression in a mapping population can greatly enhance the construction of such networks. If an eQTL results from differences in expression of a regulator, this regulator is likely to show correlation in expression levels with the gene that mapped to its position (Bing and Hoeschele, 2005). Multiple genes involved in the same biological process mapping to the same position indicates that many of them might be under the control of the same gene. We reasoned that the best candidate within an eQTL interval is the gene whose expression best correlates with multiple genes mapping to the position of that gene. We therefore combined expression trait profiling with eQTL mapping, gene annotation, and extended iterative Group Analysis (iGA) (Breitling *et al.*, 2004) to sort candidate regulators based on their PC-value (Possibility of Change), which tells, for a given regulator, how likely it is to observe a strong correlation with multiple members of a selected group of genes. This novel approach enabled us to drastically narrow down the number of candidate genes in an eQTL interval and select the best candidate for the construction of genetic regulatory networks.

To verify our approach, we focused on one of the best studied and most complete genetic regulatory networks available in plants: the regulation of flowering in Arabidopsis. Flowering time is highly variable between accessions of Arabidopsis (Koornneef *et al.*, 2004). Variation in flowering time also exists between L*er* and Cvi, and several studies have reported QTLs for this trait (Alonso-Blanco *et al.*, 1998a; Ungerer *et al.*, 2002; Juenger *et al.*, 2005). Although flowering starts much later, the expression of genes that indicate commitment to flowering are already apparent at very early stage and find their transcription peak in the seedling stage (Kobayashi *et al.*, 1999; Zimmermann *et al.*, 2004). We selected a set of 192 genes known to be involved in the control of flowering from recent literature (see supplemental Table 5 at www.pnas.org/cgi/content/full/0610429104/DC1 for a full list) and keyword searching in The Arabidopsis Information Resource database; 175 of these genes were analyzed in our study. Analysis of their expression level in the parental accessions assigned eight of them as being differentially expressed. However, 83 genes showed at least one eQTL at a genome-wide threshold of 2.23 x 10^{-3}. We calculated PC-values for correlation in expression profiles, using the group of 83 mapped flower genes and all candidate genes within their eQTL support intervals. We then selected the genes within the eQTL support interval of a given flower gene with significant PC-values (FDR = 0.05) as candidates for this eQTL (supplemental Table 5 at www.pnas.org/cgi/content/full/0610429104/DC1). Regulators were predicted for 51 genes, whereas for 32 genes no significant PC-value was obtained

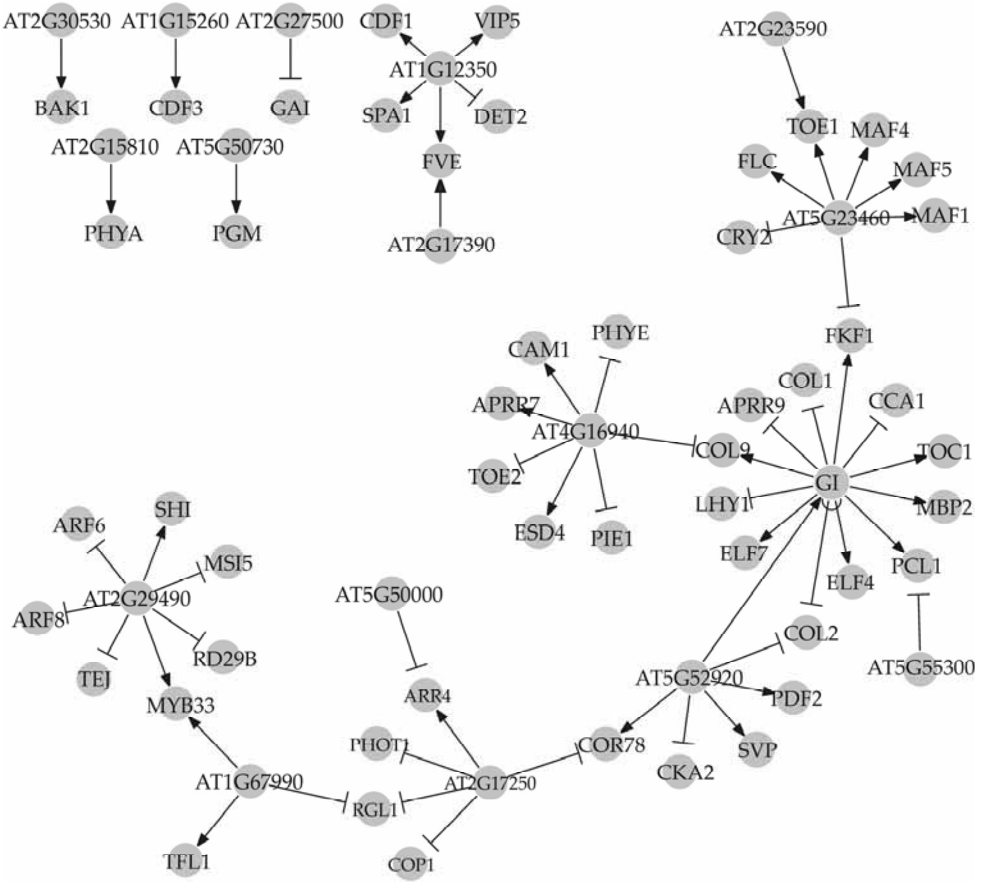


Figure 7: Regulatory network of genes involved in the transition to flowering.
Flower genes are connected to their most likely regulator by directional edges. Arrows and bars represent stimulative and repressive regulation, respectively.

Figure 7 shows a network of flower genes and their most likely regulators. The most significant regulator detected was *GIGANTEA* (*GI*) with a PC-value of 1.01 x 10^{-12}. Thirteen genes mapped to *GI*, including *GI* itself, and all of them contributed to the lowest PC-value. *GI* is the first member of an output pathway of the circadian clock that controls flowering time and has been shown to regulate circadian rhythms in Arabidopsis (Mizoguchi *et al.*, 2005). At the position of GI, a minor flowering-time QTL (Alonso-Blanco *et al.*, 1998a) and a circadian period length QTL (Swarup *et al.*, 1999; Michael *et al.*, 2003) were identified, which indicates the physiological consequences of this complex pattern of gene expression variation. Indeed many of the genes, like *CCA1* (see supplemental Table

5 at www.pnas.org/cgi/content/full/0610429104/DC1 for details), *LHY1, ELF4*, and *TOC1*, for which *GI* was identified as their most likely regulator, belong to the core circadian oscillator (Boss *et al.*, 2004). Others are involved in the regulation of the circadian clock, such as *PCL1*, *APRR9*, and *FKF1* (Michael *et al.*, 2003; Onai and Ishiura, 2005), or play a role in floral transition, such as *ELF7* and the *CONSTANS-LIKE* family *COL1*, *COL2*, and *COL9* (Ledger *et al.*, 2001; He *et al.*, 2004; Cheng and Wang, 2005). A second cluster of co-regulated genes is involved in floral repression and mapped to *FLG*, another major QTL for flowering time. Where the floral repressors *FLC*, *MAF1*, *MAF4*, *MAF5*, and *TOE1* (Boss *et al.*, 2004) are up-regulated, the floral promoter *CRY2* (Boss *et al.*, 2004) is down-regulated by this locus, in agreement with findings that *FLC* expression negatively correlates with *CRY2* (El-Din El-Assal *et al.*, 2003). In addition to *FLG*, *CRY2* and *FLC* are major-effect QTLs for flowering time in the L*er* x Cvi population, and significant epistasis has been found between *CRY2* and *FLC* (El-Din El-Assal *et al.*, 2003) and between the *FLC* region and the *FLG* locus (Alonso-Blanco *et al.*, 1998a). Although *HUA2* was previously suggested as a candidate for the *FLG* locus (Doyle *et al.*, 2005), we did not identify it as such and found a gene with unknown function (At5g23460) to be the most likely candidate. Other clusters are predominantly involved in hormonal pathways (*MYB33*, *ARF6*, *ARF8*, *RD29B*, and *SHI*) (Mouradov *et al.*, 2002; Nagpal *et al.*, 2005) and the photoperiod pathway (*PIE1*, *CAM1*, *PHYE*, and *ESD4*) (Levy and Dean, 1998; Boss *et al.*, 2004) of flowering.

To identify other possible target genes of the most significant regulator (*GI*), we calculated the correlation coefficient between the genes of the *GI* regulatory cluster and all other genes. Strong correlation was observed for 280 transcripts at an empirical correlation coefficient cutoff of 0.55, corresponding to a FDR of 9.5 x 10^{-5} (supplemental Table 6 at www.pnas.org/cgi/content/full/0610429104/DC1). Many of these genes showed no significant linkage at the position of *GI* but several displayed a suggestive eQTL. Although correlation can be a result of linked genetic effect, only 32 locally regulated genes were located within 2.5 Mbp of *GI*. The highest correlation coefficient (0.75) was found for a *CONSTANS-LIKE PROTEIN* encoding gene (At1g07050). The long day integrator *CONSTANS* (*CO*) has been shown to be a direct target of *GI* (Mizoguchi *et al.*, 2005), although it was not identified as such in our study. Two other genes associated with circadian rhythms, *APRR5* and *WNK1*, were detected, and both showed a suggestive eQTL at the position of *GI*. *APRR* genes are paralogs of *TOC1* and have been shown to be regulated by the protein kinase *WNK1* (Nakamichi *et al.*, 2002). These results suggest that the feedback regulation of the circadian clock by *GI* acts, at least partly, through *WNK1* and *APRR5*.

DISCUSSION

Genetic variation in gene expression is abundant and complex
We determined differences in gene expression between two distinct accessions of Arabidopsis and within an RIL population derived from these accessions.

Our data suggest that variation in gene expression among genetically different plants of the same species is for a large part genetically controlled and highly complex. Although eQTLs were detected for >4,000 genes, only 922 were differentially expressed between the parents, which suggests that the expression of many genes is controlled by multiple loci with opposing effects, avoiding large differences between natural accessions but generating strong transgression in a segregating population. This suggestion is supported by the differences in heritability, as calculated from the parental and population expression analyses. This difference between the two heritability estimates might have several reasons. First, statistical issues might bias the outcome of the analyses. False negatives might bias the number of genes differentially expressed between the parents downwards, because statistical power was limited to ten replicate measurements of each parent. On the other hand, false positives due to low signal-to-noise ratios for low-expressed genes might bias the number of mapped genes upwards. However, most mapped genes had medium-to-high expression levels (Figure 2A).

A second and more likely reason why mapped genes were not significantly differently expressed between the parents, given the extent of the difference in number, might be the complex genetic inheritance of gene expression. Illustrating this finding is that although the median heritability of mapped genes was 82.4%, only a median 28.4% of the variation observed for mapped genes could be explained by significant eQTLs. Furthermore, although the proportion of mapped genes increased with higher heritability values, many genes with a high heritability could not be mapped significantly. Together with the strong transgression observed for many genes, these data imply that regulation of expression often occurs through the added effect of numerous small-effect loci, each of which fail to pass the significance threshold.

Because two color arrays were used in this study, a dye effect can be expected in subsequent analyses. In our experiment, dye effect was controlled and corrected at two levels. At the level of the experimental design, we balanced the dye effect between two alleles by optimizing for the number of L*er*/Cvi and Cvi/L*er* comparisons at each marker position (Fu and Jansen, 2006). At the analysis level we included the gene-specific differential effect between the two dyes in the QTL analysis model (Dobbin *et al.*, 2005).

Molecular background of expression variation

Many factors, ranging from abiotic external influences to direct active control of transcriptional activity, influence the level of transcript abundance of a given gene. Here, we focused on genetic factors contributing to whole-genome transcript levels. Our data showed that genes whose transcript variation could be mapped are not equally distributed over the Arabidopsis genome. Although a strong correlation between the total number of genes per unit of chromosome and those that could be mapped was observed, other explanations, such as differences in chromatin structure or SNP frequency, cannot be excluded. Illustrative for this was the correlation observed between SNP frequency and the proportion of mapped genes.

Anchoring of the genetic map enabled us to define local versus distant regulation. Although, in general, local regulation seems stronger, distant regulation occurs more frequently. These findings were demonstrated by decreasing the significance threshold; only a minor number of additional locally regulated genes were detected, whereas the number of distantly regulated genes increased more than two-fold. Because the vast majority of genes showing local linkage are expected to be *cis*-regulated (Ronald *et al.*, 2005), this difference in increase can be explained by the direct influence of *cis*-polymorphisms on expression, whereas *trans*-polymorphisms exert their effect indirectly through a change in expression or coding sequence of a second gene. Taking together the strong transgression observed for many genes and the number of distantly versus locally regulated genes, it is conceivable that many *cis*-regulated genes exert pleiotropic effects on the expression of other genes and are causal for many of the eQTLs acting *in trans*.

Regulatory networks

For many biological processes, the genes contributing to a certain phenotype are often well known. However, in many cases, little is known about the regulation and interaction of these genes. We combined expression information with eQTL mapping, gene annotation, and iterative Group Analysis to identify likely regulators. This approach enabled, for the first time, the construction of maximum-likelihood genetic regulatory networks from a genome-wide genetical genomics experiment. A case study that used genes involved in the well-known process of transition from a vegetative state to a flowering state confirmed many of the interactions identified previously. Moreover, numerous novel interactions that can serve to form hypothesis for future studies were predicted. It must be noted, however, that analyses were performed on data from a single time point. It is not unlikely that regulation occurs differently at other developmental stages or diurnal phase or even organ, specifically. Especially for pathways influenced by the circadian clock, such as flowering time, expression differences at one time point can be caused by differences in circadian phase (Michael *et al.*, 2003; Darrah *et al.*, 2006). Accuracy and reliability would therefore benefit from gene expression analysis at multiple developmental stages and time points. Nevertheless, confidence in the followed approach was gained from the fact that many functionally related genes grouped together, indicating common and simultaneous regulation. We assigned the gene with the lowest PC-value as the most likely candidate responsible for this regulation although other genes with significant PC-values can not be ruled out *a priori*. Moreover, due to coincidental genetic linkage of regulators, independently regulated genes may show a high correlation in expression. This potential source of false candidate assignment is especially prone to hot spots of locally regulated genes. Subsequent in-depth analysis should be performed to unambiguously identify genes underlying eQTLs, but the number of candidate genes decreased substantially with the described method.

MATERIALS AND METHODS

Plant material and tissue collection

Aerial parts of seedlings from the accessions L*er* and Cvi and a population of 160 recombinant inbred lines derived from a cross between these parents (Alonso-Blanco *et al.*, 1998b; Keurentjes *et al.*, 2006) were grown and collected as described previously (Keurentjes *et al.*, 2006). In brief, seeds of lines were sown in petri dishes on 1/2MS agar and placed in a cold room for seven days. Petri dishes were then transferred to a climate chamber and seedlings were collected after seven days. Plant material was stored at -80°C until further processing.

Linkage map construction and anchoring to the physical map

The genetic map was constructed from a subset of the markers available, at http:/nasc.nott.ac.uk/, with a few new markers added. The computer program JoinMap 4 (van Ooijen, 2006) was used for the calculation of linkage groups and genetic distances. In total, 144 markers were used, with an average spacing of 3.5 cM. The largest distance between two markers was 10.8 cM.

To anchor the genetic map to the physical map of Arabidopsis, the total set of 291 available markers was analyzed. First, a genetic map that comprised all 291 markers was constructed. Physical positions of molecular PCR markers were obtained from The Arabidopsis Information Resource, release 6.0 (www. arabidopsis.org). Sequences of amplified fragment length polymorphism markers were obtained by *in silico* amplification of Col markers that were polymorphic between L*er* and Cvi (Peters *et al.*, 2001) or by sequencing fragments polymorphic between L*er* and Cvi but absent in Col. The retrieved marker sequences were then blasted against the completely sequenced Col genome, and center positions of positive hits were taken as the physical position. Physical positions could be established for 179 markers; positions of remaining markers were inferred from interpolation by using the closest nearby markers for which a physical position was known. The largest gap between two markers with confirmed physical position comprised 3.5 Mbp, which corresponded to a genetic distance of approximately 15 cM.

Sample preparation

Total RNA of each line was isolated from two biological replicates by using phenol-chloroform extraction (Jones *et al.*, 1985). Extracts were then combined and purified with RNeasy (Qiagen, Valencia, CA), amplified with the MessageAmp aRNA kit (Ambion, Austin, TX) incorporating 5-(3-aminoallyl)-UTP, and labeled with Cy3 or Cy5 mono-reactive dye (Amersham, Piscataway, NJ.). All RNA products were purified by using the Rneasy kit (Qiagen). Labeled RNA was fragmented for 15 minutes before hybridization (fragmentation reagent obtained from Ambion).

Microarray analyses

Arabidopsis DNA microarrays were provided by the Galbraith laboratory (University of Arizona, Tucson, AZ) and were produced from a set of 70-mer oligonucleotides, representing 24,065 unique genes (Array-Ready Oligo Set, version 1.0, Qiagen-Operon).

DNA probe immobilization and hybridization was performed according to instructions from the Galbraith laboratory. Arrays were scanned by using a ScanArray Express HT (PerkinElmer, Wellesley, MA) and quantified by using Imagene 6.0 (BioDiscovery, El Segundo, CA).

Experimental design

Genome-wide gene expression analysis was carried out for L*er* and Cvi and an RIL population derived from a cross between these two accessions. Ten replicates of the parental lines were compared in direct hybridizations by using a dye swap design. The 160 RILs were analyzed by direct hybridization of two genetically distant lines on each array, leading to a total of 80 slides. A novel distant pair design, which was proposed specifically for genetic studies on gene expression (Fu and Jansen, 2006) was used. An optimal design was obtained through simulated annealing, in which pairs of genetically distant lines were hybridized to maximize the direct comparisons between two different alleles at each marker. The numbers of Ler/Cvi and Cvi/Ler comparisons at each marker were optimized for equal ratio to balance dye effects, and their total number was optimized for minimal extra variation across other markers. The observed signal intensities on the arrays were subjected to general normalization procedures (Yang *et al.*, 2002; Smyth, 2004). Resulting log signal intensities and log ratios between co-hybridized RILs were used for further analyses.

Statistical analyses

Differential expression of genes between the two parents was tested for significance. For each gene, the *P*-value of a *t*-test and the corresponding *q*-values (Storey and Tibshirani, 2003) were computed (Smyth, 2004). The *P*-value significance threshold was 2.5 x 10^{-3} at a *q*-value cutoff of 0.05.

Log signal intensities of gene expression were used to test for genetic variance of expression traits. Spot effects were removed by treating it as a random effect in a linear mixed model.

Heritability of expression in the parental accessions was calculated as follows (Hegmann and Possidente, 1981):

$$H_P^2 = \frac{0.5 \times Vg}{0.5 \times Vg + Ve}$$

where *Vg* and *Ve* represent the components of variance among and within accessions, respectively. The factor 0.5 was applied to adjust for the 2-fold overestimation of additive genetic variance among inbred strains.

Heritability of expression within the RIL population was calculated by using the pooled variance of the parents as an estimate of the within line variance:

$$H_{RIL}^2 = \frac{V_{RIL} - Ve}{V_{RIL}}$$

where V_{RIL} and *Ve* are the variance among adjusted expression intensities in the segregants and the pooled variance within parental measurements, respectively. To prevent overestimation, we removed outliers more than three standard deviations away from the mean values. We discarded 1,470 (6.1%) negative heritability values.

Transgressive segregation was determined in terms of the pooled standard deviation of the parents (Brem and Kruglyak, 2005). We calculated the number of RILs, *n*, whose expression level lay beyond the region $\mu \pm 2$ x *SD*; where μ and *SD* are the mean and the standard deviation of parental phenotypic values, respectively. To determine significance, phenotype values of parents and segregants were reassigned at random to null parents and segregants for each transcript. The number of transgressive individuals, n_0, was then recorded. The total number of transcripts with n_0 greater than a given threshold *m* represented the genome-wide false-positive count at *m*. The FDR was computed as the ratio between estimated false-positive count at *m* and the number of non-permuted transcripts with $n > m$. Results were averaged over 20 permutations. The FDR = 0.05 cutoff corresponded to $m = 33$.

Multiple QTL analysis

Gene expression in the mapping population was analyzed for significant eQTLs. For each gene the log-ratios of signal intensities were subjected to multiple QTL mapping (MQM). Cofactors were selected by using a backward elimination process (Jansen, 1993) (see supplemental information at www.pnas.org/cgi/

content/full/0610429104/DC1). For every marker-by-gene combination, the MQM model can be given as:

$$y = \mu + b_k x_k + \sum_{i=1}^{m_k} b_i x_i$$

where y is the expression ratio of a transcript, μ is the gene-specific differential effect between Cy3 and Cy5 dyes (characterized as consistent across samples) (Dobbin *et al.*, 2005), x denotes the genotype comparison and takes the following values: 1 for Ler/Cvi, -1 for Cvi/Ler and 0 for Ler/Ler and Cvi/Cvi; b is the substitution effect; k is the k^{th} marker under study; and i denotes the cofactors from 1 to m_k, outside a 30-cM interval of the k^{th} marker. The P-value from a t-test that tested the hypothesis that $b_k = 0$ was used as a measure of significance of the association.

A genome-wide P-value threshold of 2.23 x 10^{-3} at $\alpha = 0.05$ for a single trait was estimated by a 10,000 permutation test (Churchill and Doerge, 1994). But for a study with 24,065 gene transcripts, we controlled the false discovery rate (FDR) based on the pool of P-values for all markers and all transcripts. Because the P-values are correlated when markers are linked, the FDR increases depending on the number of markers on a chromosome (Benjamini and Yekutieli, 2001). In our experiment, the maximum number of markers reached 35 (chromosome 5), and a simulation analysis (data not shown) that used Storey's algorithm to control the FDR (Storey, 2002) at a desired level indeed showed a 4.4-fold increase of the actual FDR. To account for this increase, we corrected the FDR by a factor of 5 and calculated the genome wide P-value threshold at Storey's FDR of 0.01 for all gene-marker P-values, to make sure that the real FDR rate is <0.05 (corrected FDR = 0.05). The estimated P-value threshold then corresponded to 5.29 x 10^{-5}, and this threshold was used as a significance threshold for the detection of eQTLs.

Explained variance of detected eQTLs was estimated by fitting expression ratios of all detected eQTLs and their interactions in a linear model. We used ANOVA to estimate the fraction of variance explained by each eQTL and eQTL interactions.

Local and distant regulation

We determined the physical position of each eQTL by anchoring the genetic map of the L*er* x Cvi population to the physical map of the sequenced accession Col. Support intervals were then calculated by setting left and right border positions associated with $max\{-Log_{10}P\} - 1.5$, where P represents the significance value for linkage (Keurentjes *et al.*, 2006).

The physical positions of genes (The Arabidopsis Information Resource, version 2005.12.8) showing significant linkage of expression values were then compared with the positions of their respective eQTL(s); a gene was classified as locally regulated when its position coincided with the support interval and as distantly regulated when it did not.

Distribution of hot spots

eQTL hot spots are shown by the frequency distribution of the number of significant eQTLs detected. Each eQTL is presented by the marker showing the most significant linkage. The frequency distribution of eQTLs by chance was empirically estimated by 250 permutations (de Koning and Haley, 2005). The 95th percentile, corresponding to 43 eQTLs, was used as a confidence threshold for the occurrence of a hot spot.

Sliding-window analyses

All 24,065 genes analyzed were positioned on the Arabidopsis physical map, and the ATG start codon was used as the start of each gene. Each gene was classified as locally regulated, distantly regulated, or non-regulated. The frequency of the total number of genes and the number of locally and distantly regulated genes along each chromosome was determined in a 5-Mbp sliding window by using a 50-Kbp step size.

Polymorphisms between L*er* and Cvi in 875 sequenced loci (Nordborg *et al.*, 2005) were downloaded from the MSQT website (http://msqt.weigelworld.org) and filtered for unique positions. INDELs were recorded as a single polymorphism by using the physical position of the first nucleotide difference. A total number of 4,032 polymorphisms were subjected to further analysis. A sliding-window analysis for SNP frequency was then carried out as described above.

Observed gene and SNP frequencies per window were standardized by using the genome-wide average and standard deviation, and resulting *z*-scores were plotted at the physical position of the center of each window.

Genetic network construction

A group of 83 functionally related genes and their potential regulators were used for the construction of a genetic regulatory network. All of the genes that were physically located in an eQTL interval were assigned as a regulator candidate for the gene for which that eQTL was detected. The candidates were sorted by using iterative Group Analysis (iGA) (Breitling *et al.*, 2004). We postulated that, among all possible regulators, the best candidates are those that correlate particularly well to a large number of their potential target genes.

To test that postulation, we calculated all pair-wise Spearman rank correlations on expression profiles (80 log-ratios of co-hybridized RILs) between each of the 83 functionally related genes and all potential regulators in their eQTL intervals. These values were then rank-ordered so that the strongly correlated gene-candidate pairs were at the top of the list. For each given candidate, we determined the iGA possibility of change value (PC-value, supplemental Table 5 at www.pnas.org/cgi/content/full/0610429104/DC1). The PC-value threshold was Bonferroni-adjusted as 0.05/*m*, where *m* is the total number of candidate genes. Any candidate with a significant PC-value is a putative regulator, and all genes contributing to this value are putative target genes. We defined the regulatory relation in terms of the sign of the correlation coefficient. If the correlation coefficient is negative, regulation is repressive; otherwise it is stimulative.

Potential target genes outside the initial group of functionally related genes were identified by using expression trait correlations (Lan *et al.*, 2006), for which we used the regulators and target genes obtained from the iGA study as seed transcripts. We then split the log-ratio gene-expression profile matrix (a x b) into two parts: a_1 x b and a_2 x b, where a is the total number of gene transcripts (a = 24,065 in our case); a_1 is the number of seed transcripts; a_2 is the number of other genes ($a_1 + a_2 = a$) and b is the number of arrays (b = 80 in our case). We then computed the Spearman correlation coefficient and its corresponding *P*-value between each a_1 seed gene and a_2 transcript. A 95 percentile empirical threshold (r = 0.55) and its corresponding FDR (Storey and Tibshirani, 2003) (FDR = 9.5 x 10^{-5}) were estimated by performing 1,000 permutations.

Acknowledgements

We thank David Galbraith for providing microarrays and protocols, Janny Peters for providing data of *In silico* AFLP analysis and Linus van der Plas for critical reading of the manuscript. This work was supported by a grant from the Netherlands Organization for Scientific Research, Program Genomics (050-10-029).

REFERENCES

Al-Shahrour, F., Diaz-Uriarte, R. and Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20,** 578-580.

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., Peeters, A.J., Koornneef, M., Lister, C., Dean, C., van den Bosch, N., Pot, J. and Kuiper, M.T. (1998b). 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.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29,** 1165-1188.

Bing, N. and Hoeschele, I. (2005). Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* **170,** 533-542.

Borevitz, J. (2006). Genotyping and mapping with high-density oligonucleotide arrays. *Methods Mol Biol* **323,** 137-145.

Boss, P.K., Bastow, R.M., Mylne, J.S. and Dean, C. (2004). Multiple pathways in the decision to flower: enabling, promoting, and resetting. *Plant Cell* **16 Suppl,** S18-31.

Breitling, R., Amtmann, A. and Herzyk, P. (2004). Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics* **5,** 34.

Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* **296,** 752-755.

Brem, R.B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A* **102,** 1572-1577.

Brem, R.B., Storey, J.D., Whittle, J. and Kruglyak, L. (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436,** 701-703.

Bystrykh, L., Weersing, E., Dontje, B., Sutton, S., Pletcher, M.T., Wiltshire, T., Su, A.I., Vellenga, E., Wang, J., Manly, K.F. *et al.* (2005). Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* **37,** 225-232.

Cheng, X.F. and Wang, Z.Y. (2005). Overexpression of COL9, a CONSTANS-LIKE gene, delays flowering by reducing expression of CO and FT in Arabidopsis thaliana. *Plant J* **43,** 758-768.

Churchill, G.A. and Doerge, R.W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138,** 963-971.

Darrah, C., Taylor, B.L., Edwards, K.D., Brown, P.E., Hall, A. and McWatters, H.G. (2006). Analysis of phase of LUCIFERASE expression reveals novel circadian quantitative trait loci in Arabidopsis. *Plant Physiol* **140,** 1464-1474.

de Koning, D.J. and Haley, C.S. (2005). Genetical genomics in humans and model organisms. *Trends Genet* **21,** 377-381.

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

Dobbin, K.K., Kawasaki, E.S., Petersen, D.W. and Simon, R.M. (2005). Characterizing dye bias in microarray experiments. *Bioinformatics* **21,** 2430-2437.

Doyle, M.R., Bizzell, C.M., Keller, M.R., Michaels, S.D., Song, J., Noh, Y.S. and Amasino, R.M. (2005). HUA2 is required for the expression of floral repressors in Arabidopsis thaliana. *Plant J* **41,** 376-385.

El-Din El-Assal, S., Alonso-Blanco, C., Peeters, A.J., Wagemaker, C., Weller, J.L. and Koornneef, M. (2003). The role of cryptochrome 2 in flowering in Arabidopsis. *Plant Physiol* **133,** 1504-1516.

Fu, J. and Jansen, R.C. (2006). Optimal design and analysis of genetic studies on gene expression. *Genetics* **172,** 1993-1999.

He, Y., Doyle, M.R. and Amasino, R.M. (2004). PAF1-complex-mediated histone methylation of FLOWERING LOCUS C chromatin is required for the vernalization-responsive, winter-annual habit in Arabidopsis. *Genes Dev* **18,** 2774-2784.

Hegmann, J.P. and Possidente, B. (1981). Estimating genetic correlations from inbred strains. *Behav Genet* **11,** 103-114.

Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V. *et al.* (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet* **37,** 243-253.

Jansen, R.C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135,** 205-211.

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

Jones, J.D., Dunsmuir, P. and Bedbrook, J. (1985). High level expression of introduced chimaeric genes in regenerated transformed plants. *Embo J* **4,** 2411-2418.

Juenger, T.E., Sen, S., Stowe, K.A. and Simms, E.L. (2005). Epistasis and genotype-environment interaction for quantitative trait loci affecting flowering time in Arabidopsis thaliana. *Genetica* **123,** 87-105.

Kendziorski, C. and Wang, P. (2006). A review of statistical methods for expression quantitative trait loci mapping. *Mamm Genome* **17,** 509-517.

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

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

Kobayashi, Y., Kaya, H., Goto, K., Iwabuchi, M. and Araki, T. (1999). A pair of related genes with antagonistic roles in mediating flowering signals. *Science* **286,** 1960-1962.

Koornneef, M., Alonso-Blanco, C. and Vreugdenhil, D. (2004). Naturally occurring genetic variation in Arabidopsis thaliana. *Annu Rev Plant Biol* **55,** 141-172.

Lan, H., Chen, M., Flowers, J.B., Yandell, B.S., Stapleton, D.S., Mata, C.M., Mui, E.T., Flowers, M.T., Schueler, K.L., Manly, K.F. *et al.* (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genet* **2,** e6.

Ledger, S., Strayer, C., Ashton, F., Kay, S.A. and Putterill, J. (2001). Analysis of the function of two circadian-regulated CONSTANS-LIKE genes. *Plant J* **26,** 15-22.

Levy, Y.Y. and Dean, C. (1998). The transition to flowering. *Plant Cell* **10,** 1973-1990.

Michael, T.P., Salome, P.A., Yu, H.J., Spencer, T.R., Sharp, E.L., McPeek, M.A., Alonso, J.M., Ecker, J.R. and McClung, C.R. (2003). Enhanced fitness conferred by naturally occurring variation in the circadian clock. *Science* **302,** 1049-1053.

Mizoguchi, T., Wright, L., Fujiwara, S., Cremer, F., Lee, K., Onouchi, H., Mouradov, A., Fowler, S., Kamada, H., Putterill, J. *et al.* (2005). Distinct roles of GIGANTEA in promoting flowering and regulating circadian rhythms in Arabidopsis. *Plant Cell* **17,** 2255-2270.

Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* **430,** 743-747.

Mouradov, A., Cremer, F. and Coupland, G. (2002). Control of flowering time: interacting pathways as a basis for diversity. *Plant Cell* **14 Suppl,** S111-130.

Nagpal, P., Ellis, C.M., Weber, H., Ploense, S.E., Barkawi, L.S., Guilfoyle, T.J., Hagen, G., Alonso, J.M., Cohen, J.D., Farmer, E.E. ***et al.*** (2005). Auxin response factors ARF6 and ARF8 promote jasmonic acid production and flower maturation. *Development* **132,** 4107-4118.

Nakamichi, N., Murakami-Kojima, M., Sato, E., Kishi, Y., Yamashino, T. and Mizuno, T. (2002). Compilation and characterization of a novel WNK family of protein kinases in Arabiodpsis thaliana with reference to circadian rhythms. *Biosci Biotechnol Biochem* **66,** 2429-2436.

Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R. ***et al.*** (2005). The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol* **3,** e196.

Onai, K. and Ishiura, M. (2005). PHYTOCLOCK 1 encoding a novel GARP protein essential for the Arabidopsis circadian clock. *Genes Cells* **10,** 963-972.

Peters, J.L., Constandt, H., Neyt, P., Cnops, G., Zethof, J., Zabeau, M. and Gerats, T. (2001). A physical amplified fragment-length polymorphism map of Arabidopsis. *Plant Physiol* **127,** 1579-1589.

Rockman, M.V. and Kruglyak, L. (2006). Genetics of global gene expression. *Nat Rev Genet* **7,** 862-872.

Ronald, J., Brem, R.B., Whittle, J. and Kruglyak, L. (2005). Local regulatory variation in Saccharomyces cerevisiae. *PLoS Genet* **1,** e25.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G. ***et al.*** (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422,** 297-302.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3,** Article3.

Storey, J.D. (2002). A direct approach to false discovery rates *J. R. Statist. Soc. B* **64,** 479-498.

Storey, J.D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100,** 9440-9445.

Storey, J.D., Akey, J.M. and Kruglyak, L. (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol* **3,** e267.

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.

Torii, K.U., Mitsukawa, N., Oosumi, T., Matsuura, Y., Yokoyama, R., Whittier, R.F. and Komeda, Y. (1996). The Arabidopsis ERECTA gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats. *Plant Cell* **8,** 735-746.

Ungerer, M.C., Halldorsdottir, S.S., Modliszewski, J.L., Mackay, T.F. and Purugganan, M.D. (2002). Quantitative trait loci for inflorescence development in Arabidopsis thaliana. *Genetics* **160,** 1133-1151.

van Ooijen, J.W. (2006). Joinmap 4, Software for the calculation of genetic linkage maps in experimental populations. In JoinMap (Wageningen, The Netherlands: Kyazma B.V.).

Wayne, M.L. and McIntyre, L.M. (2002). Combining mapping and arraying: An approach to candidate gene identification. *Proc Natl Acad Sci U S A* **99,** 14903-14906.

Yang, Y.H., Buckley, M.J., Dudoit, S. and Speed, T.P. (2002). Comparison of methods for image analysis on cDNA microarray data. *J Comput Graph Stat* **11,** 108-136.

Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R. and Kruglyak, L. (2003). Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat Genet* **35,** 57-64.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L. and Gruissem, W. (2004). GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol* **136,** 2621-263.

Chapter 4

The genetics of plant metabolism

Joost J. B. Keurentjes*, Jingyuan Fu*, C. H. Ric de Vos*, Arjen Lommen, Robert D. Hall, Raoul J. Bino, Linus H. W. van der Plas, Ritsert C. Jansen, Dick Vreugdenhil and Maarten Koornneef

Published in Nature Genetics (2006) **38**, 842-849.

* Equal contribution.

ABSTRACT

Variation for metabolite composition and content is often observed in plants. However, it is poorly understood to what extent this variation has a genetic basis. Here, we describe the genetic analysis of natural variation in the metabolite composition in *Arabidopsis thaliana*. Instead of focusing on specific metabolites, we have applied empirical untargeted metabolomics using Liquid Chromatography–Time of Flight Mass Spectrometry (LC-QTOF MS). This uncovered many qualitative and quantitative differences in metabolite accumulation between *A. thaliana* accessions. Only 13.4% of the mass peaks were detected in all 14 accessions analyzed. Quantitative Trait Locus (QTL) analysis of more than 2,000 mass peaks, detected in a Recombinant Inbred Line (RIL) population derived from the two most divergent accessions, enabled the identification of QTLs for about 75% of the mass signals. More than one-third of the signals were not detected in either parent, indicating the large potential for modification of metabolic composition through classical breeding. Combining partial interpretation of mass signals and QTL profiles allowed us to confirm biochemical pathways known from the literature and also to elucidate novel biosynthesis steps. This can lead to the identification of the underlying genes and the construction of biochemical networks in relation to other phenotypic traits.

INTRODUCTION

Metabolites are critical in biology, and plants are especially rich in diverse biochemical compounds. It has been estimated that over 100,000 metabolites can be found in plants, and each species may contain its own chemotypic expression pattern (Wink, 1988). Moreover, substantial quantitative and qualitative variation in metabolite composition is often observed within plant species (Windsor *et al.*, 2005).

Although knowledge on the regulation of metabolite formation is increasing, for thousands of metabolites, their function in the plant, their biosynthetic pathway and the regulation thereof is still unknown. QTL analysis of natural variation present in segregating populations, which can also concern metabolites (Jansen and Nap, 2001), can identify loci explaining the observed variation (Jansen, 1993). In recent years, a few studies have focused on identifying QTLs regulating a specific group of known metabolites using detection methods directed toward specific metabolite groups (Mita *et al.*, 1997; Bentsink *et al.*, 2000; Kliebenstein *et al.*, 2001a; Loudet *et al.*, 2003; Hobbs *et al.*, 2004). However, recent advances in mass spectrometry-based metabolomics and data processing techniques should now allow large-scale QTL analyses of untargeted metabolic profiles, which may uncover previously unknown regulatory functions of loci in metabolic pathways. Using dedicated alignment software, it is now possible to perform an unbiased comparison of large numbers of metabolite-derived masses detectable in large numbers of samples, arising from inherently large sets of genotypes (which are required for accurate mapping of QTLs) in an RIL population (Tikunov *et al.*, 2005; Vorst *et al.*, 2005). QTL mapping will result in the localization of loci, and ultimately genes, causal for the observed variation and will allow the discovery of co-regulated compounds. In this way, genome-wide genetic correlative metabolic analysis now becomes feasible, as we demonstrate here.

Relationships between biological traits are often inferred from correlation analysis within a given data set. However, many of these correlations need not to be causal or a result from pleiotropic effects of a common set of regulators. In studies focusing on a small number of traits this is usually not a problem because additional experiments can easily address this. In large data sets, such as those from gene expression analysis and metabolomics, more sophisticated approaches are needed to reduce the number of 'false positive' correlations (Kose *et al.*, 2001; Stuart *et al.*, 2003). Such methods are powerful in detecting relevant relationships and can be applied to any given data set even when data were acquired in different experiments. However, no information can be obtained about the underlying genetic regulation responsible for the observed correlation. The use of a mapping population to create comprehensive data sets, on the other hand, allows the identification of common regulators causal for the observed correlation between traits. Yvert *et al.* (2003) combined both approaches by first clustering traits based on segregation variation and subsequent mapping of the mean cluster values. Although this reduced multiple testing of traits and markers, noise may be introduced when multiple QTLs segregate in the cluster therewith reducing mapping power. Moreover, to exclude chance correlation from true coordinate regulation, stringent thresholds need to be applied to define clusters, thus individual outliers and less tightly linked genes are not included.

We chose to map each mass peak separately and determine relationships by correlation analysis of QTL profiles. This enables a pair wise genetic correlation analysis of each individual mass peak identifying related masses on the basis of co-regulation. Although this rules out experimental error and other non-genetical variation, we can not exclude developmental control of metabolite formation as the cause for the observed correlation when developmental traits segregate in the population.

RESULTS

Metabolite variation is abundant and genetically controlled

To assess the natural variation in metabolite content present in Arabidopsis, we performed HPLC-QTOF MS-based untargeted metabolic fingerprinting of acidified aqueous methanol extracts from seedlings of 14 different accessions originating from various parts of the global distribution range of Arabidopsis (supplemental Table 1 at http://www.nature.com/naturegenetics).

Considerable quantitative and qualitative variation was observed in the mass profiles of the different accessions. Although a metabolite may be represented by one to several mass signals in these analyses, depending on its chemical structure and abundance, each mass signal was treated as a separate element in subsequent analyses. On average, 964 mass peaks were detected per accession, with a minimum of 826 (Col) and a maximum of 1,337 (Cvi). We detected a total of 2,475 different mass peaks; 706 were unique to single accessions, and only 331 were present in all 14 accessions (Figure 1A). On average, 50 mass peaks per accession were found to be unique, with a minimum of 14 (Bay-0) and a maximum of 235 (Cvi). Although there might be a slight bias toward an overestimation of the number of accession specific mass peaks owing to low-abundance peaks detected around the threshold level, the observed frequency distribution pattern was similar when the threshold level was increased from six to ten times local noise. It can therefore be assumed that many of the differences observed between accessions are due to qualitative differences. For most masses, a large part of the observed variation can be assigned to genetic factors, as concluded from their often high broad-sense heritabilities (Figure 1B). This, together with the substantial variation in metabolite composition observed within a single plant species promises great opportunities for metabolic engineering by classical breeding (Dixon, 2005).

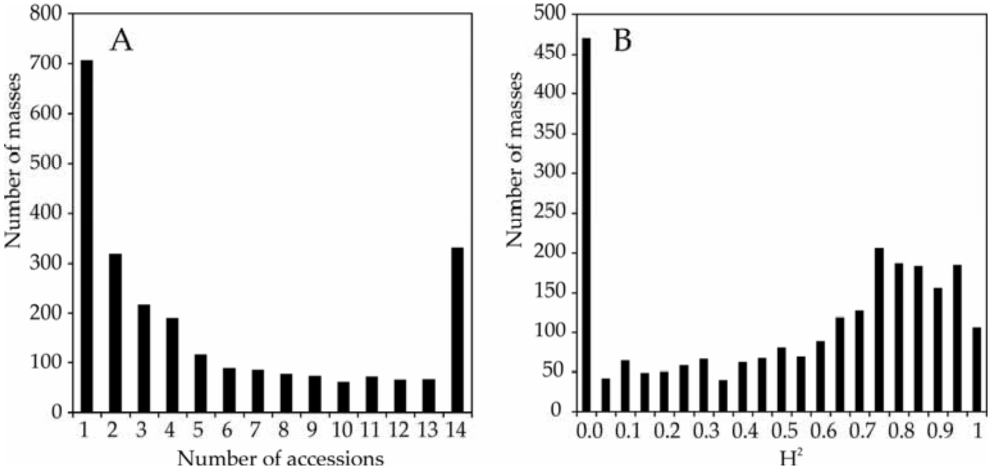


Figure 1: Natural variation in *Arabidopsis* metabolite accumulation.
(A) Frequency distribution of the number of different accessions each mass peak was detected in. (B) Frequency distribution of broad sense heritability of each mass peak detected in the different accessions. Data are based on at least two biological replicates per accession.

Most of the metabolic variation can be mapped

To uncover loci controlling the observed variation in metabolic profiles, we subsequently analyzed an RIL population derived from a cross between Landsberg *erecta* (L*er*) and Cape Verde Islands (Cvi) (Alonso-Blanco *et al.*, 1998). These were the two biochemically most distinct accessions for which such a mapping population was available (Figure 2).

Strikingly, 853 of a total of 2,129 mass peaks identified in the RIL population were not detected in either parent (Figure 3). Although the number of lines analyzed in the RIL population (160 lines measured in duplicate) exceeded that of the number of parental lines (5 replicates of each parent measured in duplicate), making the chance of detecting mass peak intensities around the threshold level higher, the observed ratio did not differ much when the threshold was increased modestly (data not shown). This suggests that many metabolites not present in either parent are produced as a result of the recombination of the genomes of the two parents.

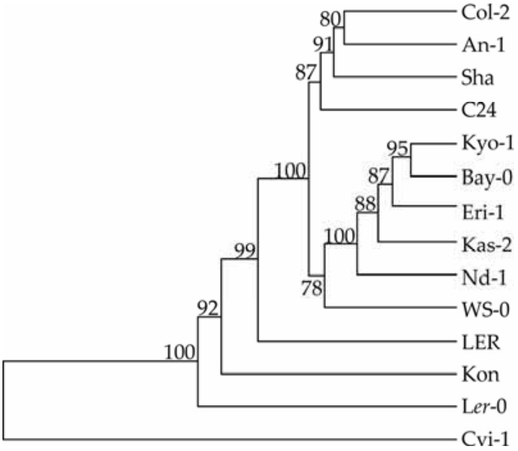


Figure 2: Hierarchical clustering of accessions for metabolite content.
The dendrogram depicts euclidean distance between groups after transformation of the data. Numbers represent confidence percentages after bootstrap analysis. Clustering on metabolite content for the different accessions shows the clear separation of Cvi from L*er* indicating large genetic differences for metabolite content.

For 1,592 mass signals (74.8%), at least one significant ($P < 0.0001$) QTL was detected using a two-part parametric model (Broman, 2003). This P-threshold corresponded to a q value of 0.0002 in Storey's genome-wide false discovery rate (FDR) method (Storey and Tibshirani, 2003). On average, we found nearly 2.0 QTLs per analyzed mass, leading to a total of 4,213 QTLs (supplemental Figure 2 at http://www.nature.com/naturegenetics). Thus, after crossing these two distinct genotypes, variation in the presence and abundance of ~75% of the detected masses in their offspring could at least partly be explained by mappable genetic factors (Figure 3), consistent with the relatively high heritabilities found for many masses (supplemental Figure 3 at http://www.nature.com/naturegenetics). At more stringent P-value thresholds of 5.0×10^{-5}, 1×10^{-5}, and 1×10^{-6}, corresponding to q values of 1×10^{-4}, 2.9×10^{-5}, and 4.1×10^{-6}, respectively, 1,500 (70.5%), 1,306 (61.3%), and 1,068 (50.2%) mass signals showed at least one significant linkage.

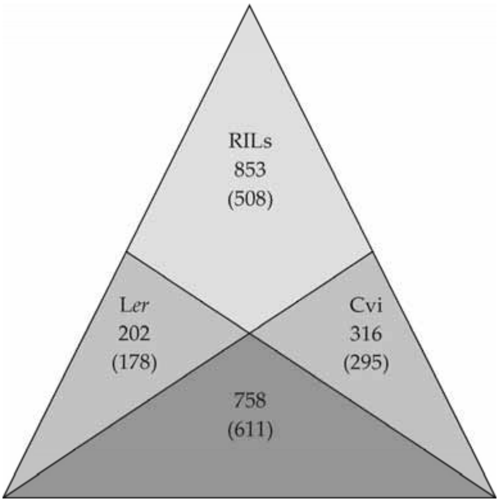


Figure 3: Number of masses detected in the RIL population and its parents.
The triangle is subdivided into masses not detected in either parent (upper part), detected in one parent only (left and right) and detected in both parents (lower part). The number of masses for which at least one significant ($P < 0.0001$) QTL was detected is shown in parentheses. Data represent two biological replicates per RIL and 5 biological replicates for each parent measured in 2 replicate extractions.

Analysis of the genomic distribution of the detected QTLs shows that these are not evenly distributed over the Arabidopsis genome. Instead, hot and cold spots for the regulation of metabolic content were observed (Figure 4). This unequal distribution of QTLs may occur for a number of reasons. Many of the metabolites detected by the approach chosen may be biochemically related and therefore have similar genetic control. In addition, genetic factors such as degree of genetic differentiation and effects of differential recombination rates might contribute to this heterogeneity. Finally, hot spots may reflect false-positive QTLs of traits highly correlated owing to technical or environmental factors (de Koning and Haley, 2005). We therefore computed empirical confidence levels by permutation tests (supplemental methods at http://www.nature.com/naturegenetics**)** and found that in most cases, the frequency of QTLs occurring at hot spots was much higher than was expected by chance (Figure 4).

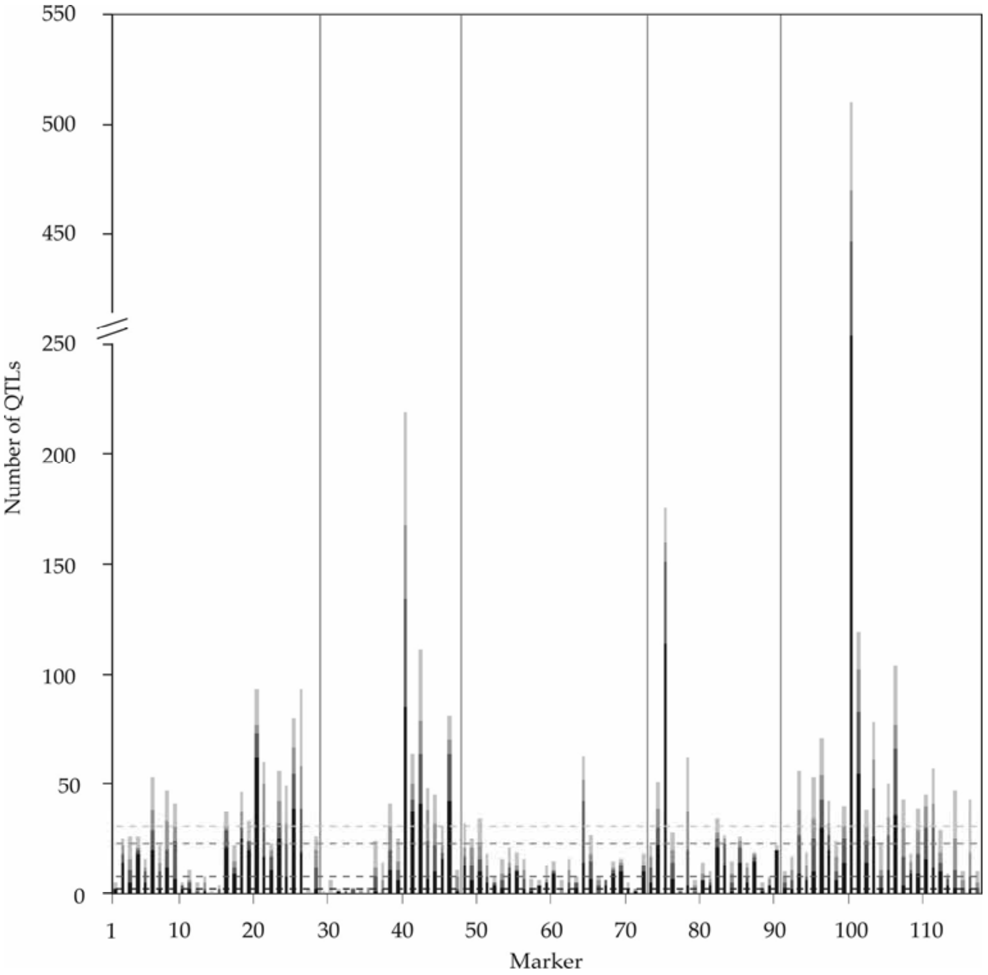


Figure 4: Frequency distribution of the number of significant QTLs detected at each marker position at four significance levels.

When, for a certain mass signal, consecutive markers showed significant linkage, only the most significant marker was counted. Markers were evenly spaced over the genome with an average distance of 5 cM between them. Chromosomal borders are indicated by vertical shaded lines. The dashed horizontal lines represent the 95% genome-wide frequency confidence thresholds for regulation hotspots obtained from 1,000 permutations. The corresponding values are 31, 23, 8, and 2 QTLs per marker expected by chance for significance levels of 10^{-4}, 5 x 10^{-5}, 10^{-5}, and 10^{-6} in increasing intensity, respectively. Data represent two biological replicates per RIL.

Map positions can reveal metabolic pathways

Co-location of QTLs coincides with clusters of highly correlated mass peaks, which are assumed to be enriched for masses regulated by the same genes. Co-regulated metabolites may indicate that a specific biological function controls different components or that a specific step in a biochemical pathway is affected (Mitchell-Olds and Pedersen, 1998). To demonstrate the latter possibility, we first focused on the mass signals corresponding to glucosinolates, for which over 30 different structures have already been identified in Arabidopsis (Reichelt *et al.*, 2002). The largest class comprises the aliphatic glucosinolates, which are all derived from methionine (Figure 5).

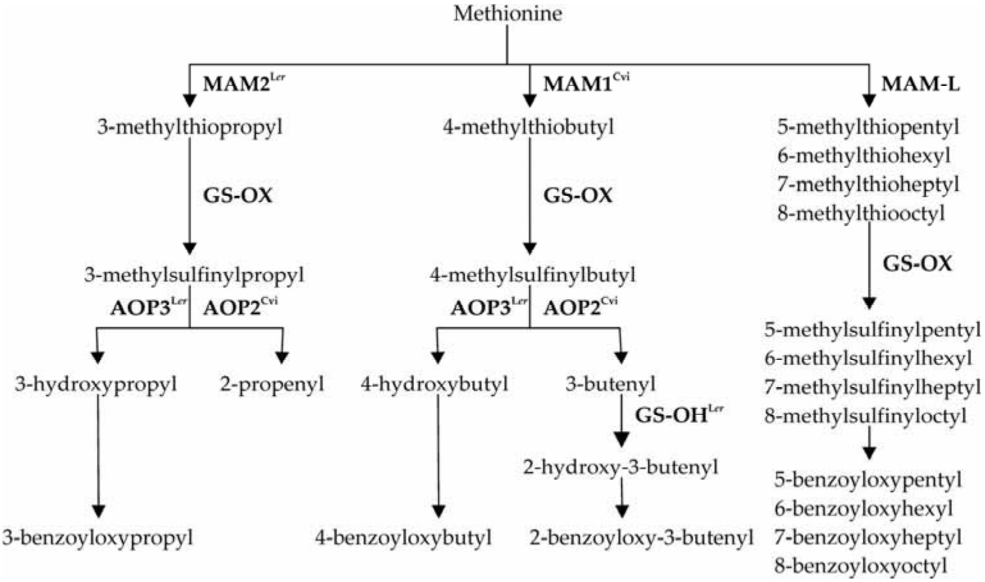


Figure 5: Genetic regulation of aliphatic glucosinolate accumulation in Arabidopsis. Corresponding loci of enzymatic steps are shown in bold next to the arrows.

Previous studies, targeted towards this class of metabolites, have shown large quantitative and qualitative differences in accumulation of aliphatic glucosinolates between Arabidopsis accessions (Kliebenstein *et al.*, 2001b). In addition, QTL analysis of these glucosinolates in the L*er* x Cvi RIL population uncovered two major loci explaining the observed variation for most aliphatic glucosinolates (Kliebenstein *et al.*, 2001a). The *MAM* locus at the top of chromosome 5 is responsible for the observed variation in chain length (Kroymann *et al.*, 2001), whereas the *AOP* locus at the top of chromosome 4 is responsible for the observed variation in side chain modification (Kliebenstein *et al.*, 2001c). Moreover, both loci, which contain multiple copies of genes having different biochemical functions, seem to control the quantitative variation in glucosinolate accumulation, with substantial interaction between the two loci. The MAM locus harbors a family of methylthioalkylmalate synthase (MAM) genes. In addition to a *MAM-L* (*MAM*-like) gene, the locus may harbor two further genes, *MAM1* and *MAM2* (Figure 5). Synthesis of C4 glucosinolates is completely dependent on the presence of a functional *MAM1* gene. Without this gene, C3 glucosinolates are synthesized. The occurrence of a *MAM-L* gene is responsible for the formation of glucosinolates with longer chain lengths. Both L*er* and Cvi contain a functional *MAM-L* gene whereas Cvi contains two *MAM1* genes arranged in tandem and L*er* contains a functional *MAM2* gene in addition to a truncated, non-functional *MAM1* gene (Kroymann *et al.*, 2001). The *AOP* locus is also a complex region containing genes encoding 2-oxoglutarate-dependent dioxygenases. At least three paralogs have been identified. The function of *AOP1* is still unknown but *AOP2* and *AOP3* functions have been described (Figure 5). All three *AOP* genes are present in both L*er* and Cvi but where *AOP1* is expressed at similar levels, *AOP2* is only expressed in Cvi and *AOP3* is only expressed in L*er* (Kliebenstein *et al.*, 2001c). Because the specific genes of the two loci, which are phylogenetic paralogs, are physically placed at the same genomic position, they segregate as alleles of each other.

By making use of the mass accuracy of the TOF-MS, we were able to identify most of the aliphatic glucosinolates reported for Arabidopsis. Subsequent QTL analysis showed that all masses corresponding to an aliphatic glucosinolate indeed mapped to the *AOP* and/or *MAM* loci (Figure 6), thus confirming previous findings. Epistatic analysis of the two loci revealed strong interactions for many of the detected glucosinolates (supplemental methods and supplemental Table 2 at http://www.nature.com/naturegenetics).

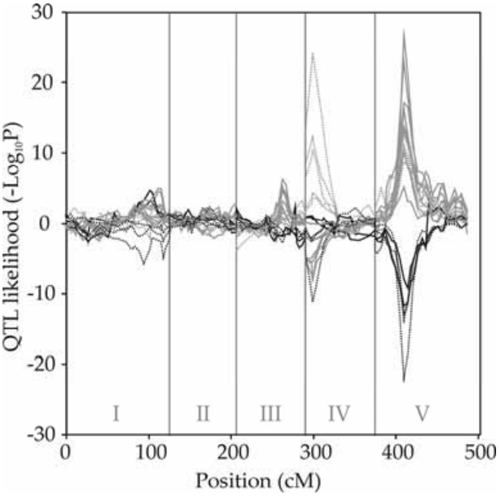


Figure 6: QTL likelihood profiles of aliphatic glucosinolates detected in the RIL population. The first QTL, at 303.3 cM, is at the *AOP* locus, the second, at 409.4 cM, is at the *MAM* locus. The sign of the value is related to the additive effect at each marker position (+, Cvi; -, L*er*). Solid lines represent glucosinolates before side chain modification and dotted lines glucosinolates after side chain modification. Chromosomal borders are indicated by vertical shaded lines. Colors represent different chain lengths (black, 3C; shaded, ≥4C**).**

The fact that we did not detect all glucosinolate QTLs found in another study (Kliebenstein *et al.*, 2001a) is most likely explained by the use of a different stage of plant development and differences in growing conditions. This is supported by the fact that they found different QTLs in seeds versus leaves. The observation that our *MAM* QTL was much stronger than in their study provides another example of such a genotype x environment or genotype x developmental stage interaction, which can be expected also for metabolites. Furthermore, we mapped individual glucosinolates whereas Kliebenstein *et al.* (2001a) showed the mapping of total aliphatic glucosinolate content.

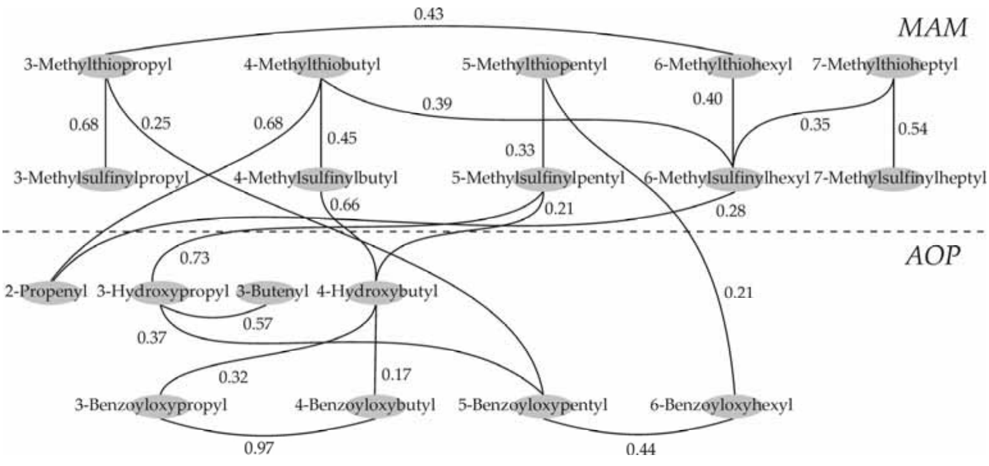


Figure 7: Second-order genetic correlations between aliphatic glucosinolates detected in the RIL population.
The upper panel contains glucosinolates before side chain modification; the lower panel contains glucosinolates after side chain modification. All edges depicted are significant at $\alpha = 0.05$, as determined by permutation. Corresponding correlation values are placed next to edges.

To assess the extent of genetic overlap between any two masses, we computed the correlation coefficients between QTL profiles (vectors of *P*-values associated with markers along the genome for each mass). Strong genetic correlations among aliphatic glucosinolates were observed due to the co-location of QTLs (data not shown). To extract the most relevant relationships between different glucosinolates, we also calculated second-order correlations defined by correlation between two glucosinolates independent of co-variance with any other pair (de la Fuente *et al.*, 2004). The significance threshold for the second-order correlations was empirically estimated by permutation (supplemental methods at http://www.nature.com/naturegenetics). Significant coefficients are shown in

Figure 7 as edges between metabolites; 0.1 false positive edges are expected by chance. The resulting network is essentially a reconstruction of a known pathway for glucosinolate formation (Figure 5) and groups glucosinolates according to their specific biosynthesis steps. The fact that the reconstructed network has similarities to the known pathway validates our methods, and the dissimilarities suggest possible previously unknown steps in the formation of glucosinolates.

Even if no prior information had been available, our mapping data alone suggest that at least two loci contribute to the observed variation in aliphatic glucosinolate formation. The fact that most *MAM*-regulated compounds do not show a QTL at the *AOP* locus and all *AOP*-regulated compounds also show a QTL at the *MAM* locus (Figure 6) suggests that *AOP* acts downstream of *MAM*. Furthermore, we observed high levels of side chain-modified compounds in unexpected genotypic classes (Table 1). In contrast to previous findings (Kliebenstein *et al.*, 2001c), this suggests, that both *AOP*2 and *AOP*3 are expressed in seedlings, indicating that regulation of glucosinolate formation is dependent on developmental stage. The reverse additive effect of the *AOP* locus for 4-hydroxybutyl, 2-propenyl and 4-benzoyloxybutyl formation shows that regulation can be completely different for different growth stages, although Kliebenstein *et al.* (2001c) also suggested alternative loci for 4-hydroxybutyl formation. These results validate our combined genetic and metabolomic approach to identify co-regulated masses and provide an independent line of evidence to validate or modify current knowledge. An untargeted approach should therefore facilitate the annotation of metabolites to existing or even to as-yet-unknown pathways.

Table 1: Phenotypic and mapping data of aliphatic glucosinolates. For each glucosinolate, significance (-log$_{10}$P) and additive effect of detected QTLs are given for the *MAM* and *AOP* locus, and relative abundance in the parental lines and RILs is given as mass signal intensities (MC, counts at maximum peak height).

	Locus				Parents		Genotype RILs[a]			
	MAM		AOP		L*er* (n=5)	Cvi (n=5)	AA (n=43)	AB (n=49)	BA (n=27)	BB (n=38)
Glucosinolate	Sign. (-logP)	Add. (MC)	Sign. (-logP)	Add. (MC)	Mean ± SD (MC)	Mean ± SD (MC)	Mean ± SD (MC)	Mean ± SD (MC)	Mean ± SD (MC)	Mean ± SD (MC)
3-Methylthiopropyl	11.5	-418	1.2	1086	555 (471)	4 (0)	296 (1162)	1255 (3748)	137 (584)	128 (757)
4-Methylthiobutyl	27.1	20416	5.3	-2419	269 (125)	22435 (9432)	856 (4981)	1076 (5309)	29646 (9469)	15984 (11843)
5-Methylthiopentyl	5.3	53	7.7	-69	7 (4)	38 (40)	49 (187)	9 (9)	157 (106)	35 (28)
6-Methylthiohexyl	17.4	53	2.2	-48	12 (10)	776 (995)	6 (7)	4 (2)	83 (97)	31 (45)
7-Methylthioheptyl	13.5	9238	0.2	610	18937 (3514)	29816 (4470)	8327 (6517)	7712 (5761)	16202 (7728)	17973 (7671)
3-Methylsulfinylpropyl	9.1	-291	0.8	77	663 (418)	15 (6)	264 (305)	447 (392)	109 (223)	52 (184)
4-Methylsulfinylbutyl	16.5	4998	4.4	-2607	42 (32)	801 (309)	348 (1124)	200 (692)	9183 (5125)	2484 (2043)
5-Methylsulfinylpentyl	10.4	110	8	-96	125 (23)	223 (68)	63 (80)	21 (16)	258 (131)	73 (51)
6-Methylsulfinylhexyl	23.1	424	2.3	-136	422 (132)	1362 (499)	267 (141)	188 (114)	792 (313)	547 (249)
7-Methylsulfinylheptyl	14	4842	0.1	250	1116 (194)	7843 (4787)	3851 (2647)	3682 (2326)	8396 (5699)	8751 (3621)
3-Hydroxypropyl	13.5	-6178	11.1	-5568	4371 (171)	261 (159)	11458 (5414)	2969 (1670)	1246 (2520)	413 (506)
4-Hydroxybutyl	16.3	122	4.5	54	4 (0)	104 (5)	6 (4)	26 (52)	84 (65)	179 (129)
2-Propenyl	21.9	-8	2.3	-3	24 (1)	10 (3)	20 (6)	17 (4)	11 (4)	9 (4)
3-Butenyl	11.5	8150	12.4	5242	13 (12)	17707 (7053)	8 (12)	313 (1160)	11 (18)	11212 (8989)
3-Benzoyloxypropyl	12.6	-5073	1.7	-1696	34807 (62)	31101 (1989)	32659 (4549)	32690 (1528)	29793 (2620)	26046 (5156)
4-Benzoyloxybutyl	10.8	1512	3.7	864	42 (11)	2654 (1269)	24 (54)	245 (1000)	701 (536)	2332 (2055)
5-Benzoyloxypentyl	12.8	255	10.8	156	16 (10)	1445 (460)	12 (37)	56 (154)	98 (65)	437 (317)
6-Benzoyloxyhexyl	8.9	92	24.5	57	4 (0)	470 (304)	4 (3)	21 (28)	4 (1)	110 (88)

[a] Genotype at the *MAM* and *AOP* locus, respectively; A=L*er*, B=Cvi.

Untargeted metabolomics uncovers new biosynthetic steps

To demonstrate the power of our untargeted metabolomics approach in uncovering previously unknown potential regulatory relationships between metabolites, we focused on a locus on chromosome 1 at 88.6 cM, where a number of mass signals could be mapped with high significance. We first determined the extent of QTL overlap, expressed as the correlation coefficient, of the mass with the most significant QTL with all other masses. Next, masses showing significant correlation were identified by calculating their accurate mass, interpreting their absorbance spectra (Photo Diode Array (PDA) signals) and using MS/MS fragmentation techniques (supplemental Table 4 at http://www.nature.com/naturegenetics). Most of the mass signals sharing this single QTL on chromosome 1 corresponded to different glycosylated flavonols (Figure 8A). The direction of the additive effect, however, suggests that genotypic variation at this locus exerts opposite effects on the glycosylation pattern. Lines carrying the L*er* allele(s) at this locus accumulate flavonols containing dihexosyl glycosides, whereas lines carrying the Cvi allele(s) at this position do not. L*er* genotypes, however, are able to synthesize all flavonols detected in Cvi genotypes (Table 2 and Figure 8, B and C). The present findings suggest that a specific not-previously-identified glycosyl transferase, catalyzing the production of flavonol-dihexosides, is active in L*er* but not in Cvi, thus affecting total flavonol composition.

Table 2: Characteristics of putatively identified flavonols. Each flavonol is presented as its aglycone with its distinguishing glycosylation pattern. Significance of the detected QTL on chromosome 1 at 88.6 cM for each flavonol is shown as –Log$_{10}P$ values and additive effect and relative abundance of each flavonol in the parental lines is given as mass signal intensities (MC, counts at maximum peak height).

Aglycone	Glycosylation	Sign. (-Log$_{10}P$)	Effect (MC)	L*er* (MC ± SE)	Cvi (MC ± SE)
Isorhamnetin	Deoxyhexosyl-hexoside	30.7	199	247 ± 54	212 ± 10
Isorhamnetin	Deoxyhexosyl-dihexoside	24.0	-123	258 ± 18	4 ± 0
Kaempferol	Dideoxyhexosyl-hexoside	39.1	197	13 ± 2	329 ± 40
Kaempferol	Deoxyhexosyl-dihexoside	29.5	-1326	1334 ± 164	7 ± 0
Quercetin	Deoxyhexosyl-hexoside	50.7	2659	1293 ± 291	4928 ± 517
Quercetin	Deoxyhexosyl-dihexoside	24.3	-1721	3031 ± 167	4 ± 0

Two genes putatively annotated as UDP-glucose:glycosyltransferases (UGTs) based on consensus sequence homology with Family 1 UGTs coincide with the support interval of the QTL (*viz.* UGT79B10 and UGT79B11) (Li *et al.*, 2001). UGT79B10 has been expressed as recombinant protein in *Escherichia coli*, but it showed no activity against quercetin glucosides in an *in vitro* analysis (Lim *et al.*, 2004). However, the coding sequence was obtained from the Columbia accession which might harbor allelic differences compared with L*er* or Cvi. No information about activity of UGT79B11 is currently available, but its sequence is highly homologous to UGT79B10, and the two genes probably arose from a duplication event. Therefore, both genes cannot be ruled out *a priori* as candidates for the observed QTL. Another possibility might be the presence of a gene in L*er* that is absent in Cvi and Col and therefore is not annotated in the Col sequence. Fine-mapping of this locus should demonstrate whether the QTL represents an encoding structural gene or a regulator thereof.

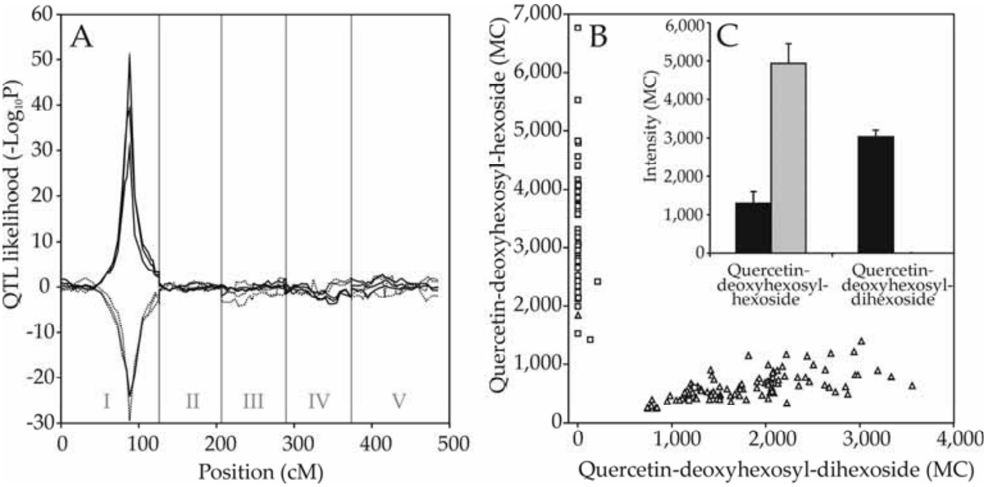


Figure 8: Genetic variation in flavonol-glycoside accumulation in Arabidopsis**.**
(A) QTL likelihood profiles of putatively identified flavonol glycosides in the RIL population. The sign of the value is related to the additive effect at each marker position (+, Cvi; -, L*er*). Dotted and solid lines represent flavonols with and without dihexosyl residues, respectively. Chromosomal borders are indicated by vertical shaded lines. (B) Typical example of relative levels of flavonol-dihexoside versus flavonol-monohexoside in the RIL population. Each symbol represents the average of two measurements per RIL. Squares and triangles represent lines carrying a Cvi or L*er* genotype at the QTL position, respectively. (C) Typical example of flavonol dihexoside and flavonol monohexoside accumulation in the parental lines L*er* (black) and Cvi (shaded). Data represent five biological replicates for each parent measured in two replicate extractions. In (B) and (C), values represent mass signal intensities (MC, counts at maximum peak height). Error bars represent s.e.m.

Thus, the untargeted detection and subsequent mapping of metabolites enabled us to identify a number of putative flavonol-glycosides not previously reported in Arabidopsis (D'Auria and Gershenzon, 2005). Co-location of QTLs suggests that variation in the accumulation of these flavonol species is attributable to a single locus affecting glycosylation of the basic flavonoid backbone.

DISCUSSION

The framework proposed here involves the untargeted detection of hundreds to potentially thousands of metabolites in a mapping population, thus enabling the mapping of QTLs for individual metabolites. This creates new opportunities for pathway elucidation and identification even when background knowledge is highly limited. We show that the biochemical variation in Arabidopsis is extensive but is nevertheless largely under genetic control, as concluded from the observation that genomic loci could be assigned for 75% of the LC-MS-detected mass peaks. The use of untargeted metabolomics is particularly useful in this context, because it allows the detection of previously unidentified metabolites. When such metabolites are co-regulated with known metabolites, this may facilitate the functional assignment of those unknown metabolites. Similarly, unexpected co-occurrence of well-known metabolites can also be discovered that would otherwise have been missed if detection was targeted to a specific subset of compounds. Genetic variation for metabolite composition might be important in adaptation to the specific environmental conditions in which the different accessions grow. In addition, they determine many aspects of the nutritional, sensory, and other aspects of crop plant quality.

Biological systems are often regulated at various molecular levels, including the influence of metabolites on plant development. A number of studies have indicated the influence of metabolites on whole plant morphology during early stages of development (Alba *et al.*, 2005; Lumba and McCourt, 2005). Thus, our understanding of biological function would benefit greatly from quantitative measurements of different classes of compounds (such as proteins and metabolites) and various processes (such as gene expression) carried out in parallel, preferably combined with other classical phenotypic analyses (Oksman-Caldentey and Saito, 2005). The implementation of different technologies then enables association analyses based on similar genetic control, as shown by similar QTL positions. In particular, the use of a perpetual mapping population such as an RIL population will have added value because co-locating QTLs can identify the genetic basis for these associations even when different experiments have been performed (Lall *et al.*, 2004; DeCook *et al.*, 2006). Our study can therefore easily be extended by using different extraction and analysis methods or by examining contrasting plant developmental stages. Moreover, the recent progress made in genetic analyses of gene expression (Brem *et al.*, 2002; Schadt *et al.*, 2003) can also readily be exploited, and this will aid further the construction of genetic regulatory networks (Jansen, 2003).

In the past, numerous studies have shown the usefulness of natural biodiversity for the elucidation of agronomically important traits, and pleiotropic loci have been identified controlling different traits simultaneously (Koornneef *et al.*, 2004). The parallel genetic analysis of physiological, transcriptional, and biochemical profiling can greatly enhance our understanding of metabolic regulatory circuitry and its relationship with phenotypic traits that segregate in the same population. The definitive identification of the most interesting chemical compounds represented by the various mass peaks would require additional chemical analysis. However, setting priorities for these analyses can now be performed effectively on the identified map positions of QTLs controlling such phenotypic traits.

Understanding the mechanisms that explain natural variation in metabolite profiles and how this correlates with phenotype is a primary challenge for evolutionary research and research geared to defining natural biodiversity and maximizing its use through directed plant breeding approaches. The strategy described here has universal application and can be used for any set of metabolites analyzed in mapping populations of any organism.

MATERIALS AND METHODS

Arabidopsis accessions and mapping population

Fourteen accessions of *A. thaliana* representing different regions of the global distribution of the species were analyzed for quantitative genetic variation in metabolite content. A population of 160 recombinant inbred lines derived from a cross between the accessions Cape Verde Islands (Cvi) and Landsberg *erecta* (L*er*) was used for QTL mapping of metabolite content. The F_{10} generation has been extensively genotyped (Alonso-Blanco *et al.*, 1998) and is available from the Arabidopsis Biological Resource Center. All lines were advanced to the F_{13} generation, and residual heterozygous regions, estimated to be 0.71% in the F_{10} generation, were genotyped again using molecular PCR markers. In addition, all lines were genotyped with a few extra markers to improve the quality of the genetic map. Because each line is almost completely homozygous, individuals of the same line are genetically identical, which allows the pooling of replicate individuals and repeated measurements to obtain a more precise estimate of phenotype values and broad sense heritabilities.

Germination, growth conditions and harvesting

Seeds of accessions and RILs were sown on 10 ml twice-diluted Murashigi and Skoog medium containing 2% agar in 6-cm Petri dishes. For each line, five replicate dishes were sown on five consecutive days with a density of a few hundred seeds per Petri dish. Petri dishes were placed in a cold room at 4°C for 7 days in the dark to promote uniform germination. Subsequently, dishes were randomly placed in five blocks in a climate chamber where each block contained one replicate dish of each line. Growing conditions were 16 hr light (30 $W.m^{-2}$) at 20°C, 8 hr dark at 15°C and 75% relative humidity. After 6 days the lids of the Petri dishes were removed to ensure seedlings were free of condensed water on the day of harvesting. On day 7, seedlings were harvested by submerging the complete Petri dish briefly in liquid nitrogen and scraping off the aerial parts with a razor blade. Harvesting started 7 hours into the light period and all lines were harvested in random order within 2 hours. Plant material was stored at -80°C until further processing.

Extract preparation and LC-MS analysis

For each line, plant material from two dishes was harvested to make one replicate sample and material from the other three dishes was harvested for the second sample. Samples were ground in liquid nitrogen, and 100 mg of each sample was weighed in 2.2 ml Eppendorf tubes. Aqueous-methanol extracts were prepared by adding 400 µl of ice-cold 92% methanol acidified with 0.1% (vol/vol) formic acid to the plant sample (final methanol concentration 75%, assuming 90% water in tissues). After sonication for 15 min and centrifugation (20,000g) for 10 min, the extracts were transferred to 96-well protein filtration plates (Captiva 0.45 µm, Ansys Technologies), vacuum filtrated and collected in 700-µl glass inserts in 96-well autosampler plates (Waters Corporation), using a Genesis Workstation (Tecan Systems Inc.). Samples were automatically injected (5 µl) and separated using an Alliance 2795 HT system (Waters Corporation) equipped with a Luna C_{18}-reversed phase column (150 x 2.1 mm, 3 µm; Phenomenex, CA). Separation was performed at 40°C by applying a 20 min gradient from 5-75% acetonitril in water, acidified with 0.1% formic acid, at a flow rate of 0.2 ml/min. Compounds eluting from the column were detected online, first by a Waters 996 photodiode array detector at 200-600 nm and then by a Q-TOF Ultima MS (Waters) with an Electron Spray Ionization (ESI) source. Ions were detected in negative mode in the range of m/z 100 to 1,500, using a scan time of 900 msec and an interscan delay of 100 msec. Desolvation temperature was 250°C with a nitrogen gas flow of 500 l/h, capillary spray was 2.75 kV, source temperature 120°C, cone voltage was 35 V with 50 l/h nitrogen gas flow and collision energy was 10 eV. The mass spectrometer was calibrated using 0.05% phosphoric acid in 50% acetonitrile and leucine enkaphalin (Sigma), detected online through a separate ESI interface every 10 sec, was used as a lock mass for exact mass measurements. MassLynx software version 4.0 (Waters) was used to control all instruments and for calculation of accurate masses.

Data pre-processing

The dedicated software program METALIGN (http://www.metAlign.nl) was used for unbiased and unsupervised comparison of all LC-MS datasets (Tikunov *et al.*, 2005; Vorst *et al.*, 2005). In short, the program performs automated peak centering, local noise calculation, baseline correction and extraction of all relevant mass signals (*i.e.* signal-to-noise ratio of 3 or higher) from all LC-MS datasets, and it subsequently uses landmark-dependent alignment algorithms to correct for local chromatographic drifts and obtain an ordered data matrix ('aligned mass peaks' versus samples). Mass peak signals generated are calculated as mass intensities (ion counts) at maximum peak height.

Quality improvement by reduction of the dataset

For each sample, the number of detected masses was reduced to improve the quality of the data set. Only masses that were detected in the optimized gradient phase (Vorst *et al.*, 2005) (between 3 and 20 min retention time) and that had a signal intensity higher than six times local noise were selected for further data analysis. For the RIL population, masses that had a signal intensity higher than six times local noise but that were detected in fewer than ten lines were discarded as well.

Statistical analyses

Total phenotypic variance was partitioned into sources attributable to genotype and error. Components of variance were used to estimate broad-sense heritability according to the formula $H^2 = V_G/(V_G + V_e)$, where V_G is the among-genotype variance component, and V_e is the residual (error) variance component of the analysis of variance (ANOVA).

The distance between accessions, based on metabolic content, was calculated by hierarchical clustering. Data were first transformed as $(x_{ij} - u_i)/sd_i$, where x_{ij} is the peak intensity of the ith mass in the jth accession; u_i is the mean intensity of the ith mass over all accessions, and sd_i is the standard deviation of the mean intensity of the ith mass over all accessions. Distance was then calculated using euclidean methods and clusters were constructed using average linkage clustering. To verify the clustering, we performed 1,000 bootstrap runs by using approximately-unbiased multistep-multiscale bootstrap resampling (Shimodaira, 2004). The *P*-values computed indicate how strongly each cluster was supported by the data.

Linkage map construction

Genotype data for the L*er* x Cvi population individuals are available at http:/nasc.nott.ac.uk/. The genetic map was constructed from a subset of the markers available with a few new markers added. The computer program JOINMAP 3.0 (Stam, 1993) (http://www.kyazma.com) was used for the calculation of linkage groups and genetic distances. Recombination frequencies were converted to centiMorgan distances using the Kosambi mapping function.

QTL analysis

For many masses, a spike in the phenotype distribution was observed, causing a departure from the assumption of normal distribution. The spike was caused by the absence of a mass peak in a considerable number of RILs, consequently leading to signal intensities equal to the detection threshold value (four times local noise). Because distributions were normal if only RILs were taken into account when signal intensities were above the detection threshold, we carried out a single-marker analysis using a two-part parametric model (Broman, 2003).

The first part describes a binominal model that tests for association of markers with presence or absence of mass peaks. For each mass peak, let y_i denote the mass intensity for the ith RIL. Let $z_i = 0$ if $y_i = 4$, and $z_i =1$ if $y_i > 4$. We then tested each marker for significant differences between the two genotypes for the probability of presence of the mass peak: H_0: $P\{z = 1 | g = \mathrm{L}er\} = P\{z = 1 | g = \mathrm{Cvi}\}$ versus the alternative hypothesis H_1: $P\{z = 1 | g = \mathrm{L}er\} \neq P\{z = 1 | g = \mathrm{Cvi}\}$, where g is the genotype (L*er* or Cvi) of a marker under analysis.

The second part describes a parametric model that tests for association of markers with intensity of the mass signal for those lines where $y_i > 4$. Under the assumption of normal distribution, we tested each marker for significant differences in the mean values between two genotypes: H_0: $\mu\{g = \mathrm{L}er\} = \mu\{g = \mathrm{Cvi}\}$ versus the alternative hypothesis H_1: $\mu\{g = \mathrm{L}er\} \neq \mu\{g = \mathrm{Cvi}\}$. The *P*-value of the two-part model was then determined by the multiple of the *P*-values from the two separate analyses (*P*1 and *P*2, respectively).

To calculate significance thresholds, we performed a simulation study following Broman (2003). Each individual had probability 40% (the median proportion of null phenotypes observed in mass data) of having a null phenotype and probability 60% of having a phenotype drawn from a normal distribution with mean 13 (the median value of mass phenotype data) and standard deviation 1. For each of 10,000 replicates, we simulated such data under the null hypothesis of no QTL, applied the two-part model and stored the genome-wide minimum *P*-value. The 98th percentile of the *P*-values corresponded to 0.0001. With the real data, the *q*-values corresponding to *P*-values were estimated using Storey's genome-wide false discovery rate (FDR) method (Storey and Tibshirani, 2003).

We next calculated the proportion of QTL significance explained by the binominal part by $\log P1/(\log P1 + \log P2)$, where *P*1 and *P*2 are the *P*-values from the two separate parts of the model respectively, (supplemental Figure 4 at http://www.nature.com/naturegenetics). The variance explained by QTLs was calculated for both parts separately (supplemental Figure 5 at http://www.nature.com/naturegenetics). In the quantitative model (part II), we used ANOVA to estimate the total sum of squares (SS_{total}) and the sum of squares between QTL genotypes (SS_{QTL}). The proportion of variance explained by the QTL was then calculated as SS_{QTL}/SS_{total}. For the binominal model (part I), we used the deviance instead of the sum of squares. We fitted the binominal data into a generalized linear (probit) model to estimate the deviances (dev) (McCullagh and Nelder, 1989). The proportion of variance explained by the QTL in the binominal model was then calculated as dev_{QTL}/dev_{total}.

Calculation of genetic correlations

Various methods have been developed and applied to uncover gene regulatory networks from expression profiles (de la Fuente *et al.*, 2004; Bing and Hoeschele, 2005; Schadt *et al.*, 2005) or from QTL profiles (Zhu *et al.*, 2004). We combined and modified the methods of Bing and Hoeschele (2005) and Zhu *et al.* (2004) and calculated the second-order partial correlation on QTL profiles between any pair of masses to assess the strength of their genetic relationship.

The calculation took three steps: (i) for each QTL significant at $P < 0.0001$, the QTL support interval was determined by setting left and right border positions associated with max{$-\log_{10}P$} ± 1.5; that is, the 1.5-LOD drop-off interval. Subsequently $-\log_{10}P$ values for positions outside the support intervals were set to zero. (ii) Pair wise correlation coefficients between any two masses were then calculated as:

$$r_{xy} = \frac{2\sum_{i=1}^{n} x_i \times y_i}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2}$$

where r_{xy} is the correlation coefficient between mass x and y, and i ($i = 1\ldots n$) is a marker. x_i and y_i represent $-\log_{10}P$ values for marker i. (iii) Finally, second-order partial correlations were calculated. The first-order correlation between variable x and y conditional on a single variable z is given by:

$$r_{xy|z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right)\left(1 - r_{yz}^2\right)}}$$

where r_{xy}, r_{xz} and r_{yz} are correlation coefficients on mass expression profiles between x and y, x and z, and y and z, respectively. The second-order partial correlation between x and y, conditional on a pair of variables z and k, is a function of first-order coefficients:

$$r_{xy|zk} = \frac{r_{xy|z} - r_{xk|z} r_{yk|z}}{\sqrt{\left(1 - r_{xk|z}^2\right)\left(1 - r_{yk|z}^2\right)}}$$

For each pair x and y, the second-order partial correlations were calculated conditional on each pair z and k, and the minimal value was stored. Having calculated these minimal values for all pairs x and y for aliphatic glucosinolates, the empirical threshold was obtained by permutation (supplemental methods at http://www.nature.com/naturegenetics). The second-order partial correlation coefficients between QTL profiles were computed in each of 20,000 permutations and sorted to derive the threshold of 0.14 at $\alpha = 0.05$, Bonferroni-adjusted for 17, the number of correlation tests for each glucosinolate. We did not correct the α

level for the number of all pair-wise analyses (17 x 18/2) to avoid over-correction. At this threshold, on average 0.1 correlation coefficients are significant by chance.

Acknowledgements
This work was supported by grants from the Netherlands Organization for Scientific Research, Program Genomics (050-10-029) and the Centre for Biosystems Genomics (CBSG, Netherlands Genomics Initiative).

REFERENCES

Alba, R., Payton, P., Fei, Z., McQuinn, R., Debbie, P., Martin, G.B., Tanksley, S.D. and Giovannoni, J.J. (2005). Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. *Plant Cell* **17,** 2954-2965.

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

Bentsink, L., Alonso-Blanco, C., Vreugdenhil, D., Tesnier, K., Groot, S.P. and Koornneef, M. (2000). Genetic analysis of seed-soluble oligosaccharides in relation to seed storability of Arabidopsis. *Plant Physiol* **124,** 1595-1604.

Bing, N. and Hoeschele, I. (2005). Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* **170,** 533-542.

Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* **296,** 752-755.

Broman, K.W. (2003). Mapping quantitative trait loci in the case of a spike in the phenotype distribution. *Genetics* **163,** 1169-1175.

D'Auria, J.C. and Gershenzon, J. (2005). The secondary metabolism of *Arabidopsis thaliana*: growing like a weed. *Curr Opin Plant Biol* **8,** 308-316.

de Koning, D.J. and Haley, C.S. (2005). Genetical genomics in humans and model organisms. *Trends Genet* **21,** 377-381.

de la Fuente, A., Bing, N., Hoeschele, I. and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20,** 3565-3574.

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

Dixon, R.A. (2005). Engineering of plant natural product pathways. *Curr Opin Plant Biol* **8,** 329-336.

Hobbs, D.H., Flintham, J.E. and Hills, M.J. (2004). Genetic control of storage oil synthesis in seeds of Arabidopsis. *Plant Physiol* **136,** 3341-3349.

Jansen, R.C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135,** 205-211.

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

Jansen, R.C. (2003). Studying complex biological systems using multifactorial perturbation. *Nat Rev Genet* **4,** 145-151.

Kliebenstein, D.J., Gershenzon, J. and Mitchell-Olds, T. (2001a). Comparative quantitative trait loci mapping of aliphatic, indolic and benzylic glucosinolate production in *Arabidopsis thaliana* leaves and seeds. *Genetics* **159,** 359-370.

Kliebenstein, D.J., Kroymann, J., Brown, P., Figuth, A., Pedersen, D., Gershenzon, J. and Mitchell-Olds, T. (2001b). Genetic control of natural variation in Arabidopsis glucosinolate accumulation. *Plant Physiol* **126,** 811-825.

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

Koornneef, M., Alonso-Blanco, C. and Vreugdenhil, D. (2004). Naturally occurring genetic variation in *Arabidopsis Thaliana*. *Annu Rev Plant Physiol Plant Mol Biol* **55,** 141-172.

Kose, F., Weckwerth, W., Linke, T. and Fiehn, O. (2001). Visualizing plant metabolomic correlation networks using clique- metabolite matrices. *Bioinformatics* **17,** 1198-1208.

Kroymann, J., Textor, S., Tokuhisa, J.G., Falk, K.L., Bartram, S., Gershenzon, J. and Mitchell-Olds, T. (2001). A gene controlling variation in Arabidopsis glucosinolate composition is part of the methionine chain elongation pathway. *Plant Physiol* **127,** 1077-1088.

Lall, S., Nettleton, D., DeCook, R., Che, P. and Howell, S.H. (2004). Quantitative trait loci associated with adventitious shoot formation in tissue culture and the program of shoot development in Arabidopsis. *Genetics* **167,** 1883-1892.

Li, Y., Baldauf, S., Lim, E.K. and Bowles, D.J. (2001). Phylogenetic analysis of the UDP-glycosyltransferase multigene family of *Arabidopsis thaliana*. *J Biol Chem* **276,** 4338-4343.

Lim, E.K., Ashford, D.A., Hou, B., Jackson, R.G. and Bowles, D.J. (2004). Arabidopsis glycosyltransferases as biocatalysts in fermentation for regioselective synthesis of diverse quercetin glucosides. *Biotechnol Bioeng* **87,** 623-631.

Loudet, O., Chaillou, S., Merigout, P., Talbotec, J. and Daniel-Vedele, F. (2003). Quantitative trait loci analysis of nitrogen use efficiency in Arabidopsis. *Plant Physiol* **131,** 345-358.

Lumba, S. and McCourt, P. (2005). Preventing leaf identity theft with hormones. *Curr Opin Plant Biol* **8,** 501-505.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models. (New York: Chapman & Hall).

Mita, S., Murano, N., Akaike, M. and Nakamura, K. (1997). Mutants of *Arabidopsis thaliana* with pleiotropic effects on the expression of the gene for beta-amylase and on the accumulation of anthocyanin that are inducible by sugars. *Plant J* **11,** 841-851.

Mitchell-Olds, T. and Pedersen, D. (1998). The molecular basis of quantitative genetic variation in central and secondary metabolism in Arabidopsis. *Genetics* **149,** 739-747.

Oksman-Caldentey, K.M. and Saito, K. (2005). Integrating genomics and metabolomics for engineering plant metabolic pathways. *Curr Opin Biotechnol* **16,** 174-179.

Reichelt, M., Brown, P.D., Schneider, B., Oldham, N.J., Stauber, E., Tokuhisa, J., Kliebenstein, D.J., Mitchell-Olds, T. and Gershenzon, J. (2002). Benzoic acid glucosinolate esters and other glucosinolates from *Arabidopsis thaliana*. *Phytochemistry* **59,** 663-671.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G. *et al.* (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422,** 297-302.

Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C. *et al.* (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* **37,** 710-717.

Shimodaira, H. (2004). Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Ann Statist* **32,** 2616-2641.

Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. *Plant J* **3,** 739-744.

Storey, J.D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100,** 9440-9445.

Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302,** 249-255.

Tikunov, Y., Lommen, A., de Vos, C.H., Verhoeven, H.A., Bino, R.J., Hall, R.D. and Bovy, A.G. (2005). A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiol* **139,** 1125-1137.

Vorst, O., de Vos, C.H.R., Lommen, A., Staps, R.V., Visser, R.G.F., Bino, R.J. and Hall, R.D. (2005). A non-directed approach to the differential analysis of multiple LC-MS-derived metabolic profiles. *Metabolomics* **1,** 169-180.

Windsor, A.J., Reichelt, M., Figuth, A., Svatos, A., Kroymann, J., Kliebenstein, D.J., Gershenzon, J. and Mitchell-Olds, T. (2005). Geographic and evolutionary diversification of glucosinolates among near relatives of *Arabidopsis thaliana* (*Brassicaceae*). *Phytochemistry* **66,** 1321-1333.

Wink, M. (1988). Plant breeding: importance of plant secondary metabolites for protection against pathogens and herbivores. *Theor Appl Genet* **75,** 225-233.

Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R. and Kruglyak, L. (2003). Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat Genet* **35,** 57-64.

Zhu, J., Lum, P.Y., Lamb, J., GuhaThakurta, D., Edwards, S.W., Thieringer, R., Berger, J.P., Wu, M.S., Thompson, J., Sachs, A.B. *et al.* (2004). An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet Genome Res* **105,** 363-374.

Chapter 5

Integrative analyses of genetic variation in enzyme activities of primary carbohydrate metabolism reveal distinct modes of regulation in *Arabidopsis thaliana*

Joost J. B. Keurentjes, Ronan Sulpice, Yves Gibon, Jingyuan Fu, Maarten Koornneef, Mark Stitt and Dick Vreugdenhil

Abstract

Plant primary carbohydrate metabolism is complex and flexible, and is regulated at many levels. Changes of transcript levels do not always lead to changes in enzyme activities, and these may not always affect metabolite levels and fluxes. To analyze interactions between these three levels of function, we have performed parallel genetic analyses of 15 enzymatic activities involved in primary carbohydrate metabolism, the transcript levels for their encoding structural genes, and their substrate and product metabolites, as well as a number of other related metabolites. Quantitative analyses of each trait were performed in the Arabidopsis L*er* x Cvi recombinant inbred line (RIL) population and subjected to correlation and quantitative trait locus (QTL) analysis. Specific regulation was often accompanied with correlations between traits, possibly due to developmental control affecting several genes, enzymes, or metabolites. For a number of enzymes, activity QTLs co-localized with expression QTLs (eQTLs) of their structural genes, or metabolite accumulation QTLs of their substrate and product. However, regulation often occurred through multiple loci, both due to posttranscriptional and *cis*- and *trans*-acting transcriptional control of structural genes, as well as independent of the structural genes. Although many of the regulatory processes in primary carbohydrate metabolism remain to be resolved, it is clear that such studies will benefit from the integrative genetic analysis of gene transcription, enzyme activity, and metabolite content. The multiparallel QTL analyses of the various interconnected transducers of biological information flow, described here for the first time, can assist in determining the cause and consequences of genetic regulation at different levels of complex biological systems.

INTRODUCTION

Carbon is probably the most prevalent and important element in any life form. Unlike most other organisms, which are dependent on uptake of organic forms of carbon, plants fix inorganic carbon through photosynthesis. Upon fixation, most of the inorganic carbon is converted into sucrose, which then acts as the major source of organic carbon for further metabolism. Some of the fixed carbon is temporarily stored as starch, and remobilized at night to support respiration and continued synthesis and export to other tissues. To meet the various demands of a growing plant for specific purposes, carbohydrates need to be allocated within the plant, and converted into a plethora of compounds (Koch, 2004). Carbohydrate metabolism in plants is more complex than in most other organisms; for example, there are alternative routes for the mobilization and metabolization of diverse components (Carrari *et al.*, 2003). Furthermore, depending on the tissue, part or all of the glycolytic pathway is present in the plastid as well as the cytosol (Lunn, 2007). Moreover, most enzymes in plant central metabolism are encoded by small gene families (The Arabidopsis Genome Initiative, 2000; Martienssen, 2000). As a result, a given substrate may be converted into different products, and products can be formed from different substrates. This versatility of enzymatic reactions in combination with substrate competition enables different metabolic routes and creates a dense metabolic network with short pathway lengths. Perturbations in sub parts of the network can therefore have strong consequences for other parts and even affect plant growth and development (Sturm and Tang, 1999; Roessner *et al.*, 2001; Fernie *et al.*, 2002). The complexity of the metabolic network may allow the plant to compensate for disturbance in one route, by enhancing the flux through an alternative route (Rontein *et al.*, 2002). To ensure a balanced carbon allocation through a plant's lifecycle, a strong and tight regulation is therefore essential. At the same time, this complexity means that there may be considerable redundancy, at least under standardized growth conditions. There are several reports where major changes in the expression of individual enzymes lead to little change in metabolism (*e.g.* (Neuhaus *et al.*, 1989).

Given the huge diversity in plant species, with large differences in their energy metabolism, growth and storage of reserves, it can be expected that there will be considerable variation in primary carbohydrate metabolism between species, and most likely also within species. For a thorough understanding of the role of natural variation in plant metabolism and development it is of pivotal importance to identify the genetic basis of variation in metabolic pathways and processes within species. The identification of genes affecting metabolic processes might also increase our knowledge about the regulatory control of pathways in general. The genetic control of primary carbohydrate metabolism is highly complex because many biochemical steps are involved, together with environmental and developmental factors. The finding, in Arabidopsis, that large differences in many enzyme activities and metabolite contents exist between accessions (Mitchell-Olds and Pedersen, 1998; Cross *et al.*, 2006), growing conditions (Gibon *et al.*, 2006; Morcuende *et al.*, 2007; Osuna *et al.*, 2007), developmental stages (Meyer *et al.*, 2007), time of day (Gibon *et al.*, 2004b), and tissues (Sergeeva *et al.*, 2004, 2006) illustrates this complexity. Cross *et al.* (2006) analyzed 24 Arabidopsis accessions for biomass production, metabolite content, and enzyme activity. Positive correlations were observed between biomass, enzyme activities, and carbohydrates. Further evidence for developmental control of plant metabolism is derived from a study by Meyer *et al.* (2007). The authors used GC-MS metabolic profiling of the Col x C24 RIL population in parallel with biomass determinations. No strong correlations between individual metabolites and biomass production could be observed but a strong canonical correlation was observed when all metabolites were taken into account. Among the metabolites contributing most to the observed correlation were intermediates of the hexose phosphate pool: fructose-6-phosphate, glucose-6-phosphate, and glucose-1-phosphate. Both positive and negative correlations between biomass and metabolites were observed although the large majority of metabolites, including sucrose, hexose phosphates and members of the TCA cycle, showed negative correlations. This, and the results of Cross *et al.* (2006), indicates that high rates of biomass production deplete pools of metabolites resulting in higher enzyme activities, as was also concluded from the relationship between tomato fruit size and metabolite content (Schauer *et al.*, 2006). Natural variation in, and spatial and temporal control of primary carbohydrate metabolism, therefore, suggest a tight relationship with plant development, although it is difficult to assess cause and consequence and regulation is highly complex.

Natural variation can be effectively analyzed in mapping populations, offering the possibility of locating genetic factors causal for the observed variation (Koornneef *et al.*, 2004). Although genetics has been successfully used to analyze quantitative variation in plant metabolism (Causse *et al.*, 1995; Mitchell-Olds and Pedersen, 1998; Prioul *et al.*, 1999; Hirel *et al.*, 2001; Rauh *et al.*, 2002; Loudet *et al.*, 2003; Fridman *et al.*, 2004; Harrison *et al.*, 2004; Sergeeva *et al.*, 2004, 2006; Calenge *et al.*, 2006; Keurentjes *et al.*, 2006; Schauer *et al.*, 2006), most studies addressed only a limited number of enzymes or metabolites, and did not integrate this with information about changes in transcript levels. Given the strong interdependency of enzyme activities and metabolites, genetic studies can benefit enormously from multidisciplinary approaches (Fiehn *et al.*, 2001; Winnacker, 2003). To gain insight into connectivity in metabolic networks it is therefore recommendable to analyze as many enzymes and metabolites involved in such a network as possible. The parallel analysis of gene expression would further enhance our understanding of genetic regulation (Urbanczyk-Wochniak *et al.*, 2003; Gachon *et al.*, 2005; Hirai *et al.*, 2005; Gibon *et al.*, 2006).

In the present study, we analyzed the activity of 15 different enzymes involved in primary carbohydrate metabolism as well as the transcript levels for their structural genes, in parallel with quantification of the most important carbohydrates and related metabolites in the Landsberg *erecta* (L*er*) x Cape verde islands (Cvi) recombinant inbred line (RIL) population of *Arabidopsis thaliana* (Alonso-Blanco *et al.*, 1998). RIL populations offer unique possibilities for such integrative studies because different types of experiments can be performed in replicates on the same genotypes. Furthermore a large number of genetic perturbations segregate in populations derived from crosses of distinct accessions. A relatively large set of lines can then be analyzed for correlations between traits as well as for quantitative trait loci (QTLs) controlling variation observed for these traits. The advantage of Arabidopsis is that the genome has been sequenced (The Arabidopsis Genome Initiative, 2000) and genes have been (putatively) annotated for nearly all enzymes in primary metabolism (The Arabidopsis Information Resource at http://www.arabidopsis.org/), allowing analysis of transcriptional regulation of these genes.

We show that genetically controlled variation exists for the activity of many enzymes as well as for transcript levels of their structural genes and for the metabolites they interconvert. By comparing the localization and responses of structural genes encoding the enzymes, eQTLs for their transcript levels, and QTLs for enzyme activities and metabolite contents, we demonstrate that genetically controlled regulation occurs through different modes of action and at multiple levels.

RESULTS

Natural variation in primary carbohydrate metabolism

To determine the extent of natural variation in primary carbohydrate metabolism in Arabidopsis we analyzed a Recombinant Inbred Line (RIL) population of a cross between the two distinct accessions Landsberg *erecta* (L*er*) and Cape Verde Islands (Cvi) (Alonso-Blanco *et al.,* 1998). Metabolic conversion rates attributable to enzyme activity were established for 15 specific enzymatic reactions in parallel with determinations of pools of metabolic carbon sources (Table 1, Figure 1).

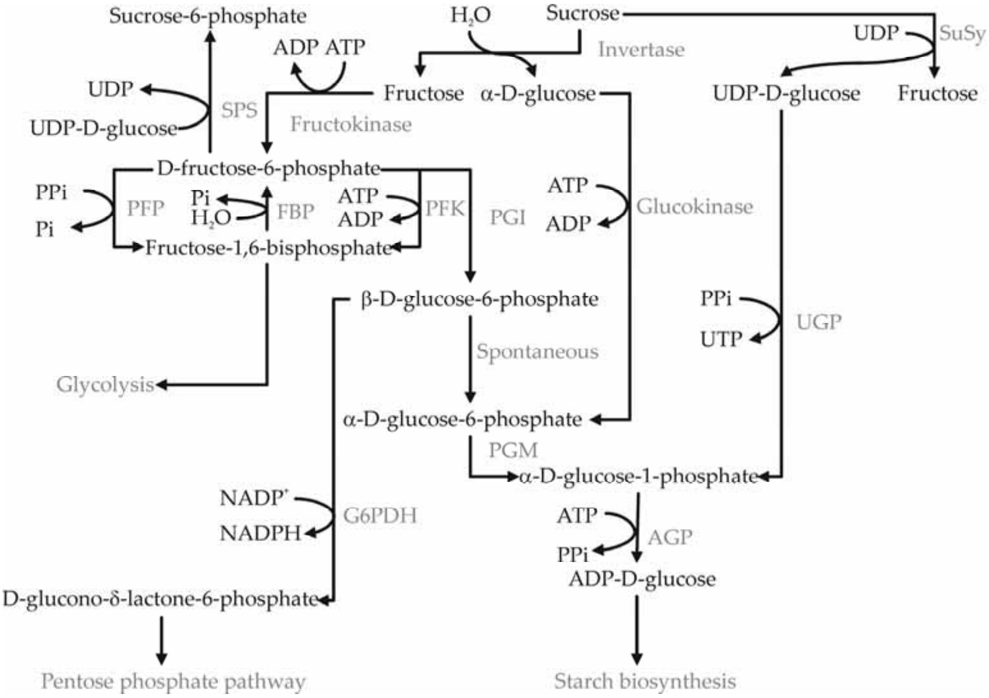


Figure 1: Enzymatic conversions in primary carbohydrate metabolism.
Reactions are given in the biologically most relevant direction, although several enzymes can catalyze reversible reactions. Metabolites are depicted in solid typeface and converting enzymes are depicted in shaded typeface on the right side of arrows.

Table 1: Summation of enzymes and metabolites analyzed and the abbreviations used. Reactions are given in the direction as they were assayed although several enzymes can also catalyze the reversible reactions.

Trait	Full name	Reaction
Inv	Acid soluble invertase, vacuolar	*Sucrose + H_2O → α-D-glucose + fructose*
AGP	ADP-glucose pyrophosphorylase	*ADP-D-glucose + PPi → α-D-glucose-1-phosphate + ATP*
FBP	Fructose-1,6-bisphosphate phosphatase, cytosolic isoform	*Fructose-1,6-bisphosphate + H_2O → D-fructose-6-phosphate + Pi*
G6PDH	Glucose-6-phosphate 1-dehydrogenase	*β-D-glucose-6-phosphate + $NADP^+$ → D-glucono-δ-lactone-6-phosphate + NADPH*
PFK	ATP dependent phosphofructokinase	*D-fructose-6-phosphate + ATP → fructose-1,6-bisphosphate + ADP*
PFP	Pyrophosphate: fructose-6-phosphate 1-phosphotransferase	*D-fructose-6-phosphate + PPi → fructose-1,6-bisphosphate + Pi*
PGM	Phosphoglucomutase	*α-D-glucose-1-phosphate → α-D-glucose-6-phosphate*
PGI	Phosphoglucose isomerase, cytosolic and plastidial isoforms	*D-fructose-6-phosphate → β-D-glucose-6-phosphate*
SPS	Sucrose phosphate synthase	*D-fructose-6-phosphate + UDP-D-glucose → sucrose-6-phosphate + UDP*
SuSy	Sucrose synthase	*Sucrose + UDP → UDP-D-glucose + fructose*
GK	Glucokinase	*α-D-glucose + ATP → α-D-glucose-6-phosphate + ADP*
FK	Fructokinase	*Fructose + ATP → D-fructose-6-phosphate + ADP*
UGP	UDP-glucose pyrophosphorylase	*UDP-D-glucose + PPi → α-D-glucose-1-phosphate + UTP*
Rubisco	Ribulose bisphosphate carboxylase/ oxygenase, initial and upon max activation	*H_2O + CO_2 + D-ribulose-1,5-bisphosphate → 2 3-phosphoglycerate + 2 H^+*
ChlA	Chlorophyl A	
ChlB	Chlorophyl B	
AA	Total Amino Acids	
Protein	Total Protein content	
Starch	Starch	
Suc	Sucrose	
Glu	Glucose	
Fru	Fructose	
G1P	α-D-glucose-1-phosphate	
G6P	α-D-glucose-6-phosphate	
UDPG	UDP-D-glucose	

Considerable variation was observed within the population for most of the analyzed traits, with coefficients of variation (CV) ranging from 13.7 (ChlA) to 54.2% (GK) (Table 2). In general CV values were higher for enzyme activity measurements than for contents of metabolites. A substantial part of the observed variation could be attributed to genetic factors, as concluded from QTL analyses. Significant QTLs were detected for ten of the enzyme activity traits and for nine metabolite traits (Table 2, Figure 2). In a number of cases, multiple QTLs were detected, sometimes with opposite effects, explaining the large variation and transgression that was observed, although in general the overall effect of QTLs was in concordance with the phenotypic differences observed between the parents. Very few co-locating QTLs were detected for the different enzyme activities, where co-location is defined as an overlap in 2 Mbp support intervals, even though several of them are from the same or related pathways (Table 2, Figure 3). Co-location of QTLs was more often the case for metabolic content due to the higher number of QTLs detected for the metabolic traits.

Despite this evidence for strong independent regulation, suggested by the detection of trait specific QTLs, when the values are compared across all the RILs, a positive correlation was observed between activity levels of all the enzymes analyzed (Figure 4). There was also a positive correlation between many enzyme activities and the structural metabolites protein and chlorophyll. A weaker positive correlation was observed between many enzyme activities and sucrose, amino acids, and starch, and a weak negative correlation with reducing sugars. This group of metabolites represents the end products of photosynthesis, and the primary compounds resulting from nitrogen incorporation. They are exported to the remainder of the plant or, in the case of starch, temporarily stored in the leaf and remobilized for export in the night. Stronger negative correlations were observed with intermediates of metabolic pathways, such as glucose-1-phosphate, glucose-6-phosphate, and UDP-glucose. These findings suggest that higher enzyme activities may allow higher fluxes, while lowering the levels of the intermediary substrates in the pathways.

Table 2: Genetic analyses of analyzed traits. The second to eighth column represent, respectively, the coefficient of variation for trait values within the RIL population, the chromosome number on which a QTL was detected, the position of the QTL on the chromosome in Mbp, the LOD score, percentage explained variance and direction of effect (+, L*er* > Cvi; -, L*er* < Cvi) of the QTL and the ²Log ratio of trait values for the parental accessions. PC1-8, principal components.

Trait	CV	Chr.	Mb	LOD	%Expl. Var	Effect	²Log Ler/Cvi
Inv	29.1	1	4.1	5.3	13.7	-	-0.13
AGP	21.3	4	12.4	3.1	8.0	+	-0.02
FBP	34.7	5	14.0	3.5	9.6	-	-1.00
G6PDH	39.0						-0.94
PFK	32.2						-0.38
PFP	26.0						0.36
PGM	37.8	1	26.9	16.0	17.5	+	-0.37
		5	20.9	36.4	56.3	-	
PGI(cyt)	22.8	1	16.8	3.1	6.8	+	0.35
		2	11.2	5.4	12.7	+	
		5	17.2	4.0	8.9	+	
PGI(pla)	22.9	5	16.7	3.1	8.4	-	0.33
PGI(tot)	15.5	1	14.9	3.2	8.8	+	0.34
SPS	20.6	5	7.0	6.4	18.0	+	0.36
SuSy	29.8						0.07
GK	54.2						ND
FK	47.8	5	16.6	3.6	9.4	-	ND
UGP	21.8	3	0.8	17.1	37.8	-	0.12
		5	5.2	5.1	9.3	+	
Rubisco(ini)	24.9						0.16
Rubisco(max)	20.9	3	20.5	3.1	9.0	+	0.21
Rubisco(ratio)	33.2						-0.50
ChlA	13.7	2	11.2	3.7	7.4	+	0.43
		3	0.3	3.4	6.8	+	
		4	10.6	3.4	6.7	+	
		5	1.7	3.8	7.6	+	
ChlB	14.0						0.32
AA	15.0	2	8.5	5.3	8.9	-	-0.53
		2	16.2	3.9	6.2	-	
		3	0.3	4.7	7.5	+	
		4	13.9	5.1	8.6	-	
		5	14.0	4.1	6.6	-	
Protein	14.2	2	12.9	3.2	7.6	+	0.35
		3	7.4	3.2	7.6	+	
Starch	17.8						-0.04
Suc	15.2	3	15.6	3.4	8.5	-	0.39
		3	23.3	5.8	15.1	+	
Glu	20.4	1	4.9	8.5	19.2	-	0.10
		2	11.2	4.4	9.1	-	
		3	13.0	5.8	13.8	-	

Table 2: Continued.

Trait	CV	Chr.	Mb	LOD	%Expl. Var	Effect	^{2}Log Ler/Cvi
Fru	19.4	1	5.4	5.0	10.9	-	0.03
		3	7.9	11.7	27.5	+	
		3	13.0	6.2	15.3	-	
G1P	32.7	3	0.3	4.5	12.1	-	-0.56
		5	7.2	3.3	8.8	+	
G6P	35.8	3	1.3	4.0	13.0	-	-0.38
UDPG	24.7	3	0.8	35.9	64.9	-	-0.71
PC1		2	11.2	4.7	11.6	+	
PC2		3	0.3	28.2	54.6	-	
PC3		1	4.4	4.7	13.0	-	
PC4							
PC5		5	8.6	4.1	11.9	-	
PC6		3	7.0	7.1	19.0	+	
PC7		5	18.2	10.8	28.5	-	
PC8		5	1.3	4.2	11.9	+	

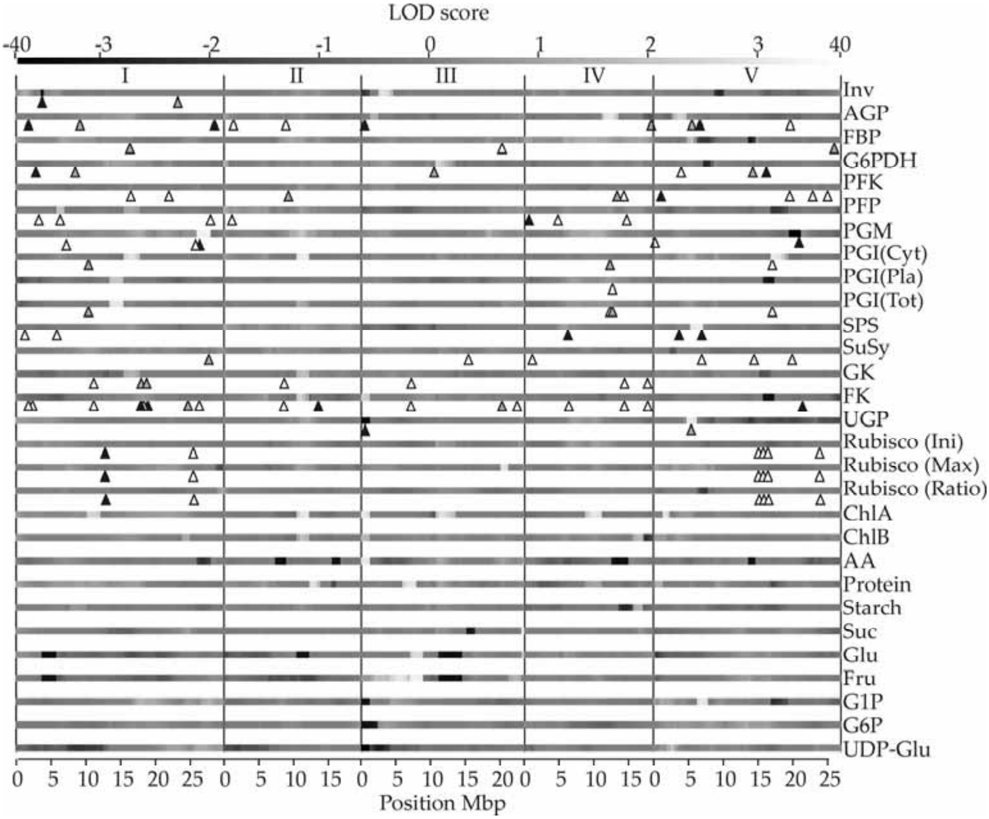


Figure 2: Heatmap of QTL profiles of each analyzed trait.

Shading intensities represent LOD scores. Positive effect loci are projected in decreasing intensity and negative effect loci in increasing intensity. Chromosomal borders are indicated by vertical shaded lines and the position of structural genes for the enzyme by triangles. Transcriptional regulation of structural genes is indicated by shading intensity of the triangles; Solid, local eQTL; shaded, distant eQTL; open, no eQTLs detected or gene not analyzed.

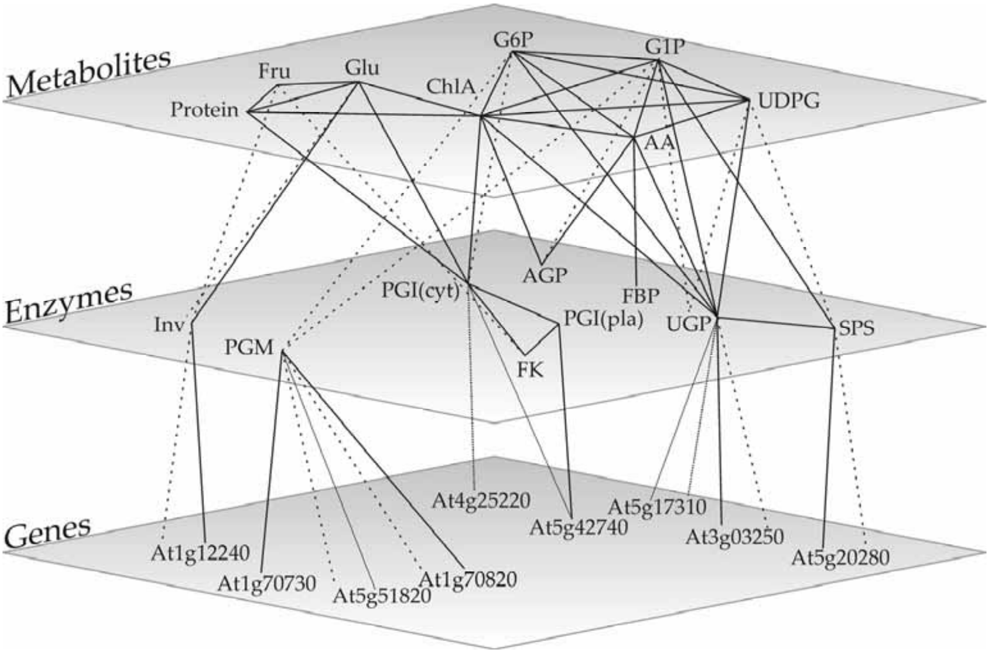
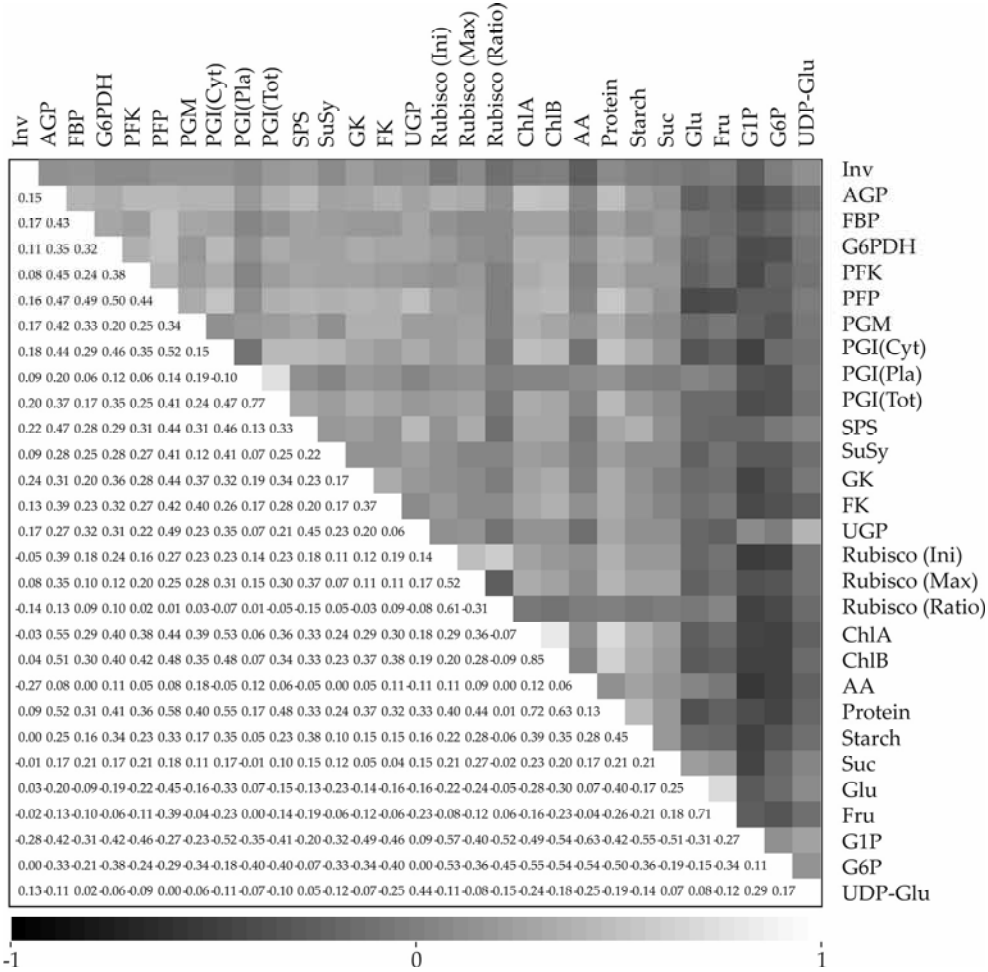


Figure 3: QTL co-location network of analyzed genes, enzymes and metabolites.
Edges between planes represent, respectively: between genes and enzymes: solid, position of structural gene co-locating with enzyme activity QTL; dashed, *cis*-eQTL co-locating with enzyme activity QTL; dotted, *trans*-eQTL co-locating with enzyme activity QTL; between enzymes and metabolites: solid, enzyme activity QTL co-locating with metabolite content QTL; dashed, enzymes connected to their substrate and/or product metabolites. Solid edges within planes connect traits with co-locating QTLs. Co-location was defined as an overlap in QTL support intervals.

To determine whether we could identify a common factor explaining the observed correlations we performed a principal component analysis (PCA) on all traits analyzed. For most traits a large part of the variation could be extracted in eight principal components (PC), explaining together 68% of the observed variation (Table 3). By far the best representative of all traits was PC1, which explained over 28% of the variance. Interestingly, in PC1 positive values were obtained for the enzyme activity traits and some end products, while negative values were obtained for the hexose pools, which is in line with the observed correlations between these traits. When the corresponding PC values for the individual RILs were subjected to QTL analysis a strong QTL for PC1 was observed at 11.2 Mbp on chromosome 2, which corresponds to the position of *ERECTA* (Table 2). This locus was also identified as a QTL for cytosolic phosphoglucose isomerase activity, chlorophyll A and glucose content. The *ERECTA* gene is polymorphic between the population's parental accessions L*er* and Cvi (Alonso-Blanco *et al.*, 1998) and causal for many of the morphological and developmental differences observed between these accessions (Torii *et al.*, 1996; Juenger *et al.*, 2005; Masle *et al.*, 2005). Moreover, *ERECTA* has been shown to exert pleiotropic effects on many growth related and metabolic traits (El-Lithy *et al.*, 2004; Keurentjes *et al.*, 2006, 2007a). It is therefore conceivable that *ERECTA* is responsible for a subtle simultaneous regulation of primary carbon metabolism, in parallel with its effects on development. It has been suggested earlier that there may be such links, but without any specific suggestions as to which genes might be involved (Cross *et al.*, 2006; Meyer *et al.*, 2007). Other PCs merely explain variation in a specific subset of traits, e.g. PC2 best explains most of the variation observed for UDP-glucose pyrophosphorylase, glucose-1-phosphate, glucose-6-phosphate and UDP-glucose. All of these traits show a QTL at the same position at the top of chromosome three, where a QTL for PC2 was also detected (Table 2) (see below for further discussion).

Table 3: Principal component analysis. Columns represent respectively the proportion of variance that could be explained by all components and by each component separately for the different traits analyzed. The last row represents the percentage of explained variance of all traits by all components and by each component separately.

	Extraction	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Inv	0.44	0.22	0.27	0.41	-0.24	0.06	0.26	-0.17	-0.01
AGP	0.64	0.78	0.06	0.10	-0.05	0.09	-0.01	0.03	0.05
FBP	0.53	0.48	0.21	0.15	-0.04	0.17	0.22	0.31	-0.24
G6PDH	0.59	0.70	-0.11	0.09	-0.06	0.23	0.09	-0.14	-0.10
PFK	0.42	0.56	0.02	-0.04	0.04	0.00	0.28	-0.15	-0.08
PFP	0.82	0.83	0.21	-0.06	-0.11	0.03	-0.05	0.13	-0.21
PGM	0.65	0.54	0.04	0.15	-0.19	0.09	-0.02	0.49	0.20
PGI(cyt)	0.70	0.76	0.12	-0.09	0.01	-0.08	0.13	-0.25	-0.10
PGI(pla)	0.84	0.33	-0.11	0.30	-0.54	-0.19	-0.54	-0.06	0.01
PGI(tot)	0.89	0.71	-0.04	0.16	-0.38	-0.22	-0.34	-0.23	-0.01
SPS	0.58	0.65	0.30	0.11	0.07	-0.15	-0.13	-0.04	-0.06
SuSy	0.35	0.45	0.14	-0.01	-0.12	-0.01	0.15	-0.05	-0.30
GK	0.51	0.60	0.02	0.02	-0.31	0.04	0.17	-0.11	0.06
FK	0.54	0.49	-0.23	0.06	-0.28	0.15	0.17	0.31	0.13
UGP	0.72	0.51	0.57	0.16	0.25	0.05	-0.18	0.08	-0.04
Rubisco(ini)	0.91	0.51	-0.20	0.10	0.33	0.53	-0.41	-0.16	0.13
Rubisco(max)	0.73	0.54	0.01	0.07	0.40	-0.29	-0.37	-0.10	0.21
Rubisco(ratio)	0.93	0.09	-0.24	0.05	0.02	0.91	-0.10	-0.10	-0.03
chlA	0.83	0.73	-0.24	-0.14	0.20	-0.15	0.25	-0.02	0.32
chlB	0.78	0.68	-0.19	-0.17	0.11	-0.05	0.36	-0.06	0.33
AA	0.70	0.13	-0.52	-0.01	0.13	-0.08	-0.25	0.51	-0.26
Protein	0.74	0.80	-0.13	-0.10	0.14	-0.10	0.06	0.02	0.18
Starch	0.59	0.55	-0.25	0.02	0.31	-0.18	-0.10	0.15	-0.25
Suc	0.70	0.24	-0.23	0.50	0.48	-0.11	0.19	0.02	-0.24
Glu	0.86	-0.39	-0.30	0.78	0.01	-0.11	0.06	-0.03	0.00
Fru	0.79	-0.27	-0.39	0.68	-0.01	-0.03	0.22	-0.07	0.23
G1P	0.69	-0.14	0.57	0.13	-0.01	0.04	0.00	0.39	0.43
G6P	0.48	-0.16	0.54	0.22	0.09	0.02	0.23	0.01	-0.23
UDPG	0.70	-0.13	0.69	0.26	0.27	0.09	-0.19	-0.07	0.15
% of Variance	67.82	28.25	9.08	6.64	5.47	5.36	5.25	4.00	3.77



	Inv	AGP	FBP	G6PDH	PFK	PFP	PGM	PGI(Cyt)	PGI(Pla)	PGI(Tot)	SPS	SuSy	GK	FK	UGP	Rubisco (Ini)	Rubisco (Max)	Rubisco (Ratio)	ChlA	ChlB	AA	Protein	Starch	Suc	Glu	Fru	G1P	G6P	UDP-Glu
Inv																													
AGP	0.15																												
FBP	0.17	0.43																											
G6PDH	0.11	0.35	0.32																										
PFK	0.08	0.45	0.24	0.38																									
PFP	0.16	0.47	0.49	0.50	0.44																								
PGM	0.17	0.42	0.33	0.20	0.25	0.34																							
PGI(Cyt)	0.18	0.44	0.29	0.46	0.35	0.52	0.15																						
PGI(Pla)	0.09	0.20	0.06	0.12	0.06	0.14	0.19	-0.10																					
PGI(Tot)	0.20	0.37	0.17	0.35	0.25	0.41	0.24	0.47	0.77																				
SPS	0.22	0.47	0.28	0.29	0.31	0.44	0.31	0.46	0.13	0.33																			
SuSy	0.09	0.28	0.25	0.28	0.27	0.41	0.12	0.41	0.07	0.25	0.22																		
GK	0.24	0.31	0.20	0.36	0.28	0.44	0.37	0.32	0.19	0.34	0.23	0.17																	
FK	0.13	0.39	0.23	0.32	0.27	0.42	0.40	0.26	0.17	0.28	0.20	0.17	0.37																
UGP	0.17	0.27	0.32	0.31	0.22	0.49	0.23	0.35	0.07	0.21	0.45	0.23	0.20	0.06															
Rubisco (Ini)	-0.05	0.39	0.18	0.24	0.16	0.27	0.23	0.23	0.14	0.23	0.18	0.11	0.12	0.19	0.14														
Rubisco (Max)	0.08	0.35	0.10	0.12	0.20	0.25	0.28	0.31	0.15	0.30	0.37	0.07	0.11	0.11	0.17	0.52													
Rubisco (Ratio)	-0.14	0.13	0.09	0.10	0.02	0.01	0.03	-0.07	0.01	-0.05	-0.15	0.05	-0.03	0.09	-0.08	0.61	-0.31												
ChlA	-0.03	0.55	0.29	0.40	0.38	0.44	0.39	0.53	0.06	0.36	0.33	0.24	0.29	0.30	0.18	0.29	0.36	-0.07											
ChlB	0.04	0.51	0.30	0.40	0.42	0.48	0.35	0.48	0.07	0.34	0.33	0.23	0.37	0.38	0.19	0.20	0.28	-0.09	0.85										
AA	-0.27	0.08	0.00	0.11	0.05	0.08	0.18	-0.05	0.12	0.06	-0.05	0.00	0.05	0.11	-0.11	0.11	0.09	0.00	0.12	0.06									
Protein	0.09	0.52	0.31	0.41	0.36	0.58	0.40	0.55	0.17	0.48	0.33	0.24	0.37	0.32	0.33	0.40	0.44	0.01	0.72	0.63	0.13								
Starch	0.00	0.25	0.16	0.34	0.23	0.33	0.17	0.35	0.05	0.23	0.38	0.10	0.15	0.15	0.16	0.22	0.28	-0.06	0.39	0.35	0.28	0.45							
Suc	-0.01	0.17	0.21	0.17	0.21	0.18	0.11	0.17	-0.01	0.10	0.15	0.12	0.05	0.04	0.15	0.21	0.27	-0.02	0.23	0.20	0.17	0.21	0.21						
Glu	0.03	-0.20	-0.09	-0.19	-0.22	-0.45	-0.16	-0.33	0.07	-0.15	-0.13	-0.23	-0.14	-0.16	-0.16	-0.22	-0.24	-0.05	-0.28	-0.30	0.07	-0.40	-0.17	0.25					
Fru	-0.02	-0.13	-0.10	-0.06	-0.11	-0.39	-0.04	-0.23	0.00	-0.14	-0.19	-0.06	-0.12	-0.06	-0.23	-0.08	-0.12	0.06	-0.16	-0.23	-0.04	-0.26	-0.21	0.18	0.71				
G1P	-0.28	-0.42	-0.31	-0.42	-0.46	-0.27	-0.23	-0.52	-0.35	-0.41	-0.20	-0.32	-0.49	-0.46	0.09	-0.57	-0.40	-0.52	-0.49	-0.54	-0.63	-0.42	-0.55	-0.51	-0.31	-0.27			
G6P	0.00	-0.33	-0.21	-0.38	-0.24	-0.29	-0.34	-0.18	-0.40	-0.40	-0.07	-0.33	-0.34	-0.40	0.00	-0.53	-0.36	-0.45	-0.55	-0.54	-0.54	-0.50	-0.36	-0.19	-0.15	-0.34	0.11		
UDP-Glu	0.13	-0.11	0.02	-0.06	-0.09	0.00	-0.06	-0.11	-0.07	-0.10	0.05	-0.12	-0.07	-0.25	0.44	-0.11	-0.08	-0.15	-0.24	-0.18	-0.25	-0.19	-0.14	0.07	0.08	-0.12	0.29	0.17	

Figure 4: Correlation matrix of analyzed enzymes and metabolites.
Values and shading intensities represent spearman rank correlation coefficients between two traits.

Relationship between structural gene expression and enzyme activity

The structural genes encoding enzymes capable of specific conversions are known for most steps in the metabolic pathways of primary carbohydrate metabolism in Arabidopsis. As noted in the introduction, in most cases multiple genes have been annotated. This redundancy in structural genes possibly results from a number of genome duplications during the evolutionary history of Arabidopsis (The Arabidopsis Genome Initiative, 2000). Empirical evidence for biological activity exists only for a limited number of genes, although for many, two or more genes may be needed as a minimum to encode the enzymes in different tissues and subcellular compartments. Many of the annotations are based on homology with genes with known biological activity, but functional analyses have not been performed. Furthermore homologous and paralogous genes might have lost or modified functions or their expression pattern might have changed.

Several cases were found where the position of structural genes co-locates with QTLs for activity of their encoded enzymes (Figure 2, Table 4), including invertase, phosphoglucomutase, phosphoglucose isomerase, sucrose phosphate synthase, and UDP-glucose pyrophosphorylase. In these cases, the variation observed in enzyme activity is most likely to be due to polymorphisms in the encoding structural genes. Such polymorphisms may occur (i) in the coding region of genes leading to an alteration of the specific activity or stability, or (ii) in promoter regions that affect transcription efficiency and subsequently protein levels. In the former case the changes of activity should be independent of changes of the transcript levels, whereas in the latter case they will be accompanied by qualitatively similar changes of transcript levels. To distinguish between these possibilities, we analyzed transcript levels for all of the putative structural genes, in parallel with the aforementioned enzyme activity assays. Samples were analyzed on full genome arrays (Keurentjes *et al.*, 2007b); signal intensities for each RIL were used to calculate the correlation coefficient between individual transcript levels and enzyme activities, and signal ratios of pairs of RILs on the same slide were used for QTL analyses.

Only a weak to medium correlation between enzyme activities and the transcript levels of the putative structural genes was observed (Table 4); (see later for a discussion of possible reasons). However, in some cases significant correlations were found. The strongest correlations were observed for structural genes co-locating with enzyme activity QTLs, indicating that part of the variation observed in enzyme activity can be explained by differential expression of structural genes. This is further supported by the fact that nearly all correlations of transcript levels of these genes with enzyme activities were positive. The only exception was a small non-significant negative correlation of a phosphoglucose mutase gene (At1g70820). Negative correlations possibly result from phase shifts in transcription and translation; although other explanations are also possible (see discussion).

Table 4: Statistics of structural genes. Columns represent respectively the encoded enzymes, the AGI gene codes of structural genes, the position of the structural gene on the chromosome indicated in the AGI code, the spearman rank correlation coefficient between enzyme activity and gene transcript levels, the *P*-value of the correlation coefficient, the chromosome number and, in parentheses, the position in Mbp, the LOD score, and the direction of effect (+, L*er* > Cvi; -, L*er* < Cvi) of detected eQTLs. Genes and eQTL positions in boldface co-locate with QTLs detected for enzyme activity. When more then one eQTL was detected, positions, LOD scores, and effects of the different eQTLs are separated by a semicolon. NA, Not Analyzed; NS, Not Significant.

Enzyme	Gene	Mb	R	P	eQTL	LOD	Effect
Inv	**at1g12240**	4.15	0.19	1.8E-02	**1(4.1)**	6.4	-
	at1g62660	23.20	-0.06	4.8E-01	1(7.9); 3(20.0)	3.7; 3.3	-; -
AGP	at1g27680	9.63	-0.08	3.2E-01	1(10.2)	3.0	-
	at1g05610	1.67	-0.04	6.1E-01	1(28.8); 3(20.5)	4.0; 3.5	-; +
	at1g74910	28.14	0.24	3.2E-03	1(22.3); 1(26.4)	3.7; 4.0	+; +
	at2g04650	1.62	0.03	7.4E-01	NS		
	at2g21590	9.25	-0.02	8.5E-01	NS		
	at3g03250	0.75	-0.26	1.1E-03	1(12.5); 3(1.4); 3(20.5)	3.2; 15.1; 3.2	-; -; -
	at4g39210	18.26	-0.20	1.4E-02	3(18.6)	3.5	-
	at5g17310	5.70	-0.23	3.9E-03	3(4.1)	9.6	-
	at5g19220	6.46	0.23	3.4E-03	5(8.1)	6.2	+
	at5g48300	19.59	0.15	6.2E-02	NS		
FBP	at1g43670	16.47	0.01	9.1E-01	1(12.2)	3.1	-
	at3g54050	20.03	-0.15	7.2E-02	NS		
	at5g64380	25.76	0.03	7.5E-01	5(22.4)	4.3	-
G6PDH	at1g09420	3.04	-0.06	4.5E-01	1(3.1)	4.8	-
	at1g24280	8.61	0.31	8.4E-05	2(6.9)	3.1	+
	at3g27300	10.08	0.17	3.7E-02	4(0.3)	3.5	-
	at5g13110	4.16	-0.02	7.6E-01	NS		
	at5g35790	13.97	0.12	1.3E-01	4(0.3); 4(13.9); 5(16.7)	3.1; 3.2; 4.7	-; -; -
	at5g40760	16.33	0.06	4.9E-01	5(16.7)	8.9	+
PFK	at1g43766	16.55	NA				
	at1g59810	22.01	NA				
	at2g22480	9.55	0.13	1.1E-01	1(18.0); 2(18.3); 5(2.5)	3.7; 4.9; 3.6	+; +; -
	at4g26270	13.30	0.25	1.4E-03	2(10.0); 2(11.2)	3.1; 3.5	+; +
	at4g29220	14.40	-0.08	3.4E-01	NS		
	at5g03300	0.80	0.10	2.0E-01	5(0.8)	21.7	+
	at5g47810	19.37	0.04	6.4E-01	NS		
	at5g56630	22.94	-0.01	9.0E-01	NS		
	at5g61580	24.78	0.09	2.8E-01	NS		

Table 4: Continued.

Enzyme	Gene	Mb	R	P	eQTL	LOD	Effect
PFP	at1g12000	4.05	0.01	9.3E-01	NS		
	at1g20950	7.30	0.01	9.3E-01	NS		
	at1g76550	28.73	0.35	9.5E-06	NS		
	at2g05150	1.86	NA				
	at4g04040	1.94	-0.15	6.3E-02	1(3.8)	3.5	+
	at4g08876	5.68	NA				
	at4g32840	15.84	0.27	7.7E-04	NS		
PGM	at1g23190	8.22	0.11	1.7E-01	NS		
	at1g70730	26.67	0.04	6.3E-01	NS		
	at1g70820	26.71	-0.13	1.1E-01	**1(28.0)**	5.6	-
	at5g17530	0.58	NA				
	at5g51820	21.08	0.69	4.4E-23	5(1.7); **5(21.0)**	7.4; 36.6	+; -
PGI(Cyt)	at1g30560	10.82	-0.02	8.4E-01	4(6.6); 4(10.6)	3.4; 3.4	+; +
	at4g25220	12.92	0.32	4.6E-05	**2(11.2)**	3.1	+
	at5g42740	17.15	0.19	1.6E-02	NS		
PGI(Pla)	at4g24620	12.71	-0.17	3.0E-02	NS		
PGI(Tot)	at1g30560	10.82	0.01	9.3E-01	4(6.6); 4(10.6)	3.4; 3.4	+; +
	at4g24620	12.71	-0.23	3.5E-03	NS		
	at4g25220	12.92	0.15	5.8E-02	2(11.2)	3.1	+
	at5g42740	17.15	0.17	3.1E-02	NS		
SPS	at1g04920	1.39	0.10	2.2E-01	NS		
	at1g16570	5.67	0.12	1.3E-01	NS		
	at4g10120	6.31	-0.08	3.1E-01	4(6.2)	7.0	+
	at5g11110	3.54	0.13	1.2E-01	5(3.7)	4.5	+
	at5g20280	6.84	0.23	4.2E-03	**5(7.2)**	9.2	+
SuSy	at1g73370	27.59	0.16	4.0E-02	5(14.0)	8.2	-
	at3g43190	15.19	-0.07	4.0E-01	NS		
	at4g02280	0.99	0.07	3.8E-01	NS		
	at5g20830	7.05	0.10	2.1E-01	NS		
	at5g37180	14.74	0.14	7.7E-02	NS		
	at5g49190	19.96	0.27	5.5E-04	NS		

Table 4: Continued.

Enzyme	Gene	Mb	R	P	eQTL	LOD	Effect
GK	at1g30660	10.88	0.18	2.5E-02	NS		
	at1g47840	17.62	0.04	6.5E-01	1(16.8)	3.8	-
	at1g50460	18.70	0.02	7.9E-01	1(18.0)	17.0	-
	at2g19860	8.58	0.10	2.3E-01	NS		
	at3g20040	6.99	0.22	6.3E-03	NS		
	at4g29130	14.35	0.07	4.0E-01	NS		
	at4g37840	17.79	0.19	2.0E-02	NS		
FK	at1g06020	1.82	-0.09	2.7E-01	NS		
	at1g06030	1.83	0.01	8.8E-01	NS		
	at1g30660	10.88	0.17	3.1E-02	NS		
	at1g47840	17.62	-0.11	1.6E-01	1(16.8)	3.8	-
	at1g50390	18.67	NA				
	at1g50460	18.70	-0.09	2.9E-01	1(18.0)	17.0	-
	at1g66430	24.78	-0.11	1.6E-01	1(28.8); 2(16.8)	3.4; 3.5	-; -
	at1g69200	26.02	-0.07	4.0E-01	NS		
	at2g19860	8.58	-0.03	7.2E-01	NS		
	at2g31390	13.39	-0.15	7.2E-02	2(12.5)	5.0	-
	at3g20040	6.99	0.22	5.3E-03	NS		
	at3g54090	20.04	0.26	1.2E-03	3(11.0)	3.3	+
	at3g59480	21.99	0.05	5.6E-01	NS		
	at4g10260	6.37	NA				
	at4g29130	14.35	0.18	2.5E-02	NS		
	at4g37840	17.79	0.19	1.7E-02	NS		
	at5g51830	21.09	0.27	5.8E-04	5(21.0)	26.4	-
UGP	**at3g03250**	0.75	0.41	1.3E-07	1(12.5); **3(1.4)**	4.5; 43.9	-; -
	at5g17310	5.70	0.42	7.1E-08	1(12.5); **3(1.9)**	4.6; 30.3	-; -
Rubisco	at1g34630	12.69	0.19	1.9E-02	1(13.4)	3.0	-
	at1g67090	25.05	-0.03	7.3E-01	NS		
	at5g38410	15.39	0.02	8.4E-01	NS		
	at5g38420	15.40	NA				
	at5g38430	15.40	0.05	5.2E-01	NS		
	at5g58240	23.58	0.20	1.3E-02	NS		

We next subjected the observed transcript levels of the structural genes to QTL analyses. For each encoded enzyme; we found significant QTLs for at least one of the encoding structural genes (eQTLs) (Table 4). Both local and distant regulation was observed, as judged from the position of genes and their respective eQTLs; locally observed eQTLs indicate that regulation occurs *in cis* whereas distant eQTLs suggests regulation to occur *in trans* (Rockman and Kruglyak, 2006).

Examples of strong local regulation include UDP-glucose pyrophosphorylase (At3g03250), phosphoglucomutase (At5g51820), phosphofructokinase (At5g03300), and hexokinase (At1g50460). As noted above, enzyme activity correlated with the transcript level for several of these genes. Moreover, strong local regulation of structural genes co-locating with a QTL for activity of their encoded enzyme was observed [*e.g.* invertase (At1g12240), phosphoglucomutase (At1g70820 and At5g51820), sucrose phosphate synthase (At5g20280), and UDP-glucose pyrophosphorylase (At3g03250)]. The only exception was a structural gene for UDP-glucose pyrophosphorylase (At5g17310), which showed strong distant regulation. These findings again suggest that *cis*-regulatory variation in expression of structural genes is at least partly responsible for observed variation in enzyme activity.

In other cases, both locally and distantly acting significant eQTLs for structural genes, that did not co-locate with QTLs for enzyme activity, were found, even though significant correlation was sometimes observed between transcript levels of these genes and enzyme activity. In the case of cytosolic phosphoglucose isomerase a *trans*-acting eQTL for a structural gene (At4g25220) co-locates with a QTL for enzyme activity (Table 4). Moreover, of all genes annotated as a phosphoglucose isomerase, the transcript levels of this gene showed the highest correlation with enzyme activity. This indicates that also *trans*-acting regulatory variation in structural gene transcription can explain variation observed in enzyme activity. For two structural genes, co-locating with their encoding enzyme activity QTLs (*viz.* At1g70730, phosphoglucomutase and At5g42740, cytosolic phosphoglucose isomerase), no significant eQTL was observed.

Finally, both locally and distantly acting significant eQTLs for structural genes were detected without coinciding positions of genes and activity QTLs or co-locating (e)QTLs and for which no significant correlation between transcript level and enzyme activity was found. These findings suggest that not all annotated genes actually contribute to the observed activity of the putatively encoded enzyme and might serve other functions independently regulated from their current annotation. However, our results do not exclude other explanations, such as spatial and temporal control, post-transcriptional and (post)-translational regulation, additive effects of multiple genes, and temporal shifts between transcription and translation (see also discussion).

Different modes of action in the genetic control of enzymatic activity

Although variation in activity was observed for many of the analyzed enzymes, the strongest genetically controlled variation was found for phosphoglucomutase (PGM) and UDP-glucose pyrophosphorylase (UGP). For these two enzymes, we also investigated substrate and product levels. When combined with the parallel analysis of transcript levels of the structural genes, this offers the opportunity of gaining deeper insight into the mechanisms of genetic regulation of these traits.

For PGM-activity two highly significant QTLs were detected with opposite effects (Figure 5A). One strong activity QTL for PGM was detected at the lower arm of chromosome five, with activity being strongly decreased in L*er* genotypes for this locus, compared to Cvi genotypes. This activity QTL co-located with a structural gene for the plastidic PGM (At5g51820 (*PGM1*) (Kofler *et al.*, 2000; Periappuram *et al.*, 2000). QTL analysis of transcript levels of this gene revealed an equally significant eQTL at the identical position of this structural gene and the enzyme activity QTL. Since the direction of the additive effect of both QTLs is also identical, this suggests that *cis*-regulatory variation in the expression of a structural gene is causal for the observed variation in enzyme activity. The second activity QTL for PGM is located on the lower arm of chromosome 1 and coincides with two putatively annotated structural genes for cytosolic isoforms of PGM (At1g70730 and At1g70820 (The Arabidopsis Information Resource). No eQTL could be detected explaining variation in transcript levels of At1g70730, but a minor eQTL was detected explaining transcript level variation of At1g70820. This minor eQTL was located at a similar position as the QTL for PGM-activity, although with an opposite additive effect. There are several alternative explanations why an eQTL and activity QTL have different signs. One is that a polymorphism in the structural gene is leading to increased activity or protein stability, which results in changes of metabolites that weakly repress the transcription of the structural gene (negative feedback). Another is that there are actually two *cis* polymorphisms, one affecting transcription and one affecting protein function, which interact to regulate the eventual level of enzyme activity. The L*er* allele, compared to the Cvi allele, then leads to lower transcript levels but the encoded enzyme shows higher activity for the conversion of G1P into G6P. For At1g70730, functional polymorphisms in the coding sequence alone could explain the observed variation in enzyme activity since no genetically regulated variation in transcript levels was observed for this gene.

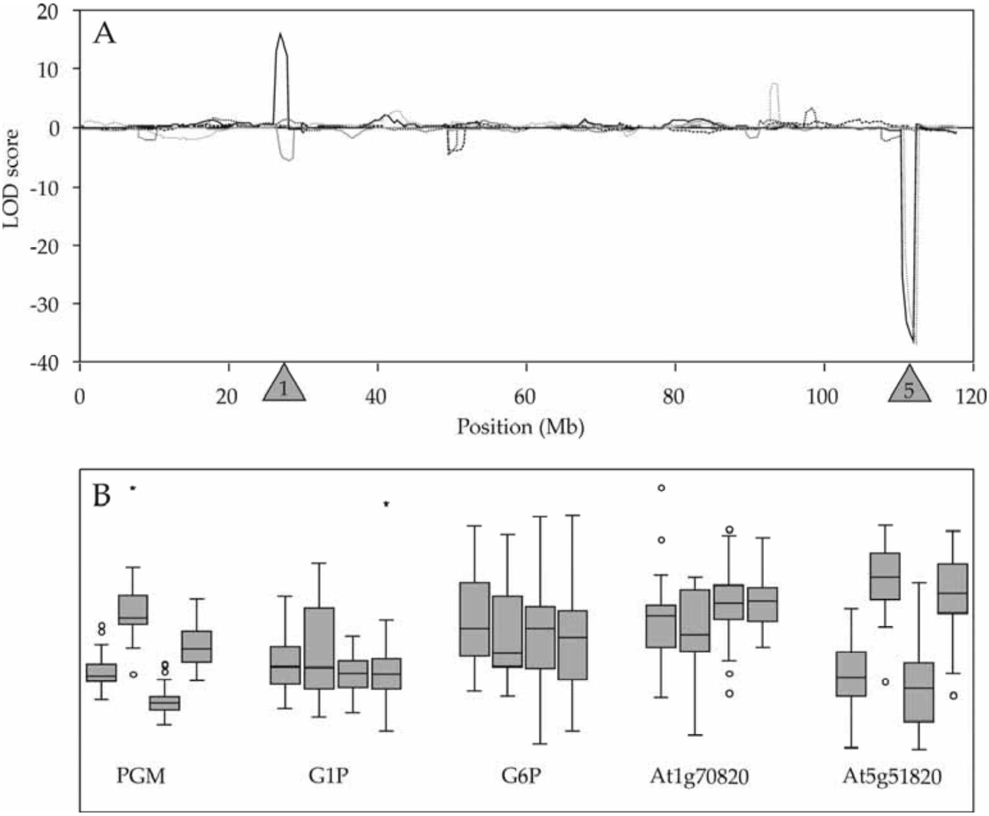


Figure 5: QTL profiles and boxplots of PGM related traits.
(A) LOD scores plotted against genomic position, the sign of the LOD score is determined by the direction of effect (+, L*er* > Cvi; -, L*er* < Cvi). Solid line, PGM activity; dotted line, G1P content; dashed line, G6P content; shaded solid line At1g70820 expression level; shaded dotted line, At5g51820 expression level. Shaded triangles indicate positions of structural genes: 1, At1g70820; 5, At5g51820. (B) Boxplots for four genotypic classes. Each class represents genotypic identical individuals for the two QTLs at chromosome one and five (from left to right: A_1A_5, A_1B_5, B_1A_5, B_1B_5; A = L*er*, B = Cvi). Boxplots show the median, interquartile range, outliers (o) and extreme cases (*) of individual variables. All traits are plotted in arbitrary units.

The levels of substrate and product of PGM were not affected by PGM-activity QTLs (Figure 5B). Although minor QTLs were detected for glucose-1-phosphate (G1P) and glucose-6-phosphate (G6P) content, these did not co-locate with QTLs for PGM-activity, suggesting that the size of the hexose phosphate pool is not determined by flux rates, as catalyzed by PGM, but regulated independently. Note that G1P and G6P are present in the plastid and the cytosol, with larger pools in the cytosol. As the strong PGM-activity QTL is likely to be caused by the plastidic PGM, then quite large changes in the pools of the plastid might not have been seen in the overall measurements.

In contrast, strong co-regulation was observed for the activity of UGP and its metabolite substrate UDP-glucose (UDPG) and to a lesser degree its product G1P. Two QTLs with opposite effect were detected for UGP-activity, each of them co-locating with a putatively annotated structural gene (Figure 6A). The UGP-activity QTL at the top of chromosome three co-locates with the position of the structural gene At3g03250, for which an eQTL with the same direction of effect was detected at the identical position. This suggests that variation in UGP-activity can be explained by cis-regulated differences in transcript levels of At3g03250. The second QTL for UGP-activity maps to the upper arm of chromosome five, and co-locates with the structural gene At5g17310. When the At5g17310 transcript levels were subjected to QTL analysis, a highly significant *trans*-acting eQTL was detected at the same position as the chromosome three UGP-activity QTL and the At3g03250 eQTL, and with the same direction of effect. This implies that the UGP-activity QTL at chromosome five cannot be explained by transcription differences of At5g17310, but might result from *cis* polymorphisms in the coding sequence. Instead, transcript level differences of At5g17310 might contribute to the chromosome three UGP-activity QTL. Although the encoded enzyme of the Cvi allele of At5g17310, compared to the L*er* allele, might have a lower specific activity it is much stronger transcribed in lines carrying the Cvi genotype at the chromosome three locus (figure 6B). Given the strong homology in sequence and function between At3g03250 and At5g17310, and the fact that for both genes a highly significant eQTL was detected at an identical position, it is likely that they are co-regulated by the same genetic factor. This could imply that At3g03250 is not *cis*-regulated, as suggested earlier, but, like At5g17310, is regulated *in trans* by a tightly linked locus.

Interestingly, a QTL for both UDPG- and G1P-content was detected at the chromosome three locus (Figure 6A), each with the same direction of effect as the (e)QTLs for UGP-activity and gene transcript levels. The direction of effect and the position of the G1P QTL can be explained by product accumulation (G1P) upon higher conversion rates of UGP. However, the direction of the highly significant QTL for the substrate UDPG is against expectations since increasing conversion rates are incompatible with accumulation of substrate (UDPG). It is therefore unlikely that UDPG content is controlled by the activity level of UGP. Instead, we hypothesize that accumulation of UDPG triggers upregulation of the expression of UGP encoding genes leading to higher enzyme activity and accumulation of G1P.

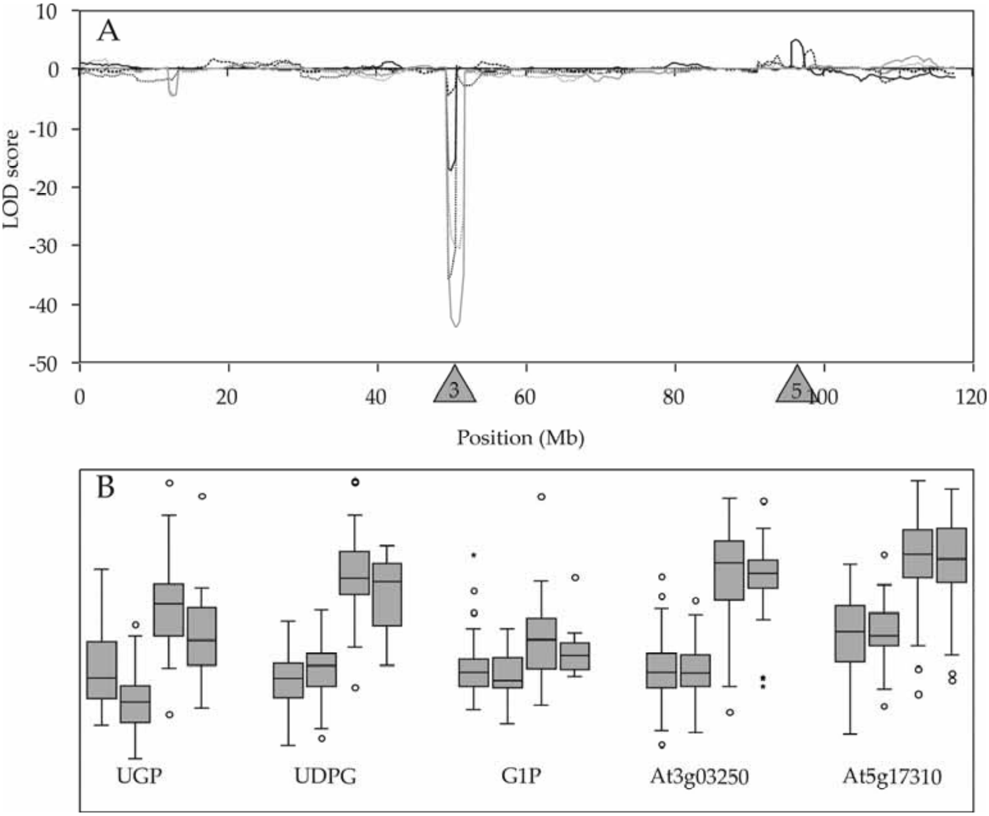


Figure 6: QTL profiles and boxplots of UGP related traits.
(A) LOD scores plotted against genomic position, the sign of the LOD score is determined by the direction of effect (+, L*er* > Cvi; -, L*er* < Cvi). Solid line, UGP activity; dotted line, UDPG content; dashed line, G1P content; shaded solid line At3g03250 expression level; shaded dotted line, At5g17310 expression level. Shaded triangles indicate positions of structural genes: 3, At3g03250; 5, At5g17310. (B) Boxplots for four genotypic classes. Each class represents genotypic identical individuals for the two QTLs at chromosome three and five (from left to right: A_3A_5, A_3B_5, B_3A_5, B_3B_5; A = L*er*, B = Cvi). Boxplots show the median, interquartile range, outliers (o) and extreme cases (*) of individual variables. All traits are plotted in arbitrary units.

DISCUSSION

Natural variation in primary carbohydrate metabolism

Natural diversity provides a rich source of genetic perturbations which has been effectively analyzed for carbohydrate metabolism by quantitative genetics in a number of studies and a variety of plant species (Causse *et al.*, 1995; Eshed and Zamir, 1995; Mitchell-Olds and Pedersen, 1998; Prioul *et al.*, 1999; Chen *et al.*, 2001; Fridman *et al.*, 2004; Sergeeva *et al.*, 2004, 2006; Li *et al.*, 2005; Cross *et al.*, 2006; Schauer *et al.*, 2006). However, most of these studies did not incorporate transcription analysis of relevant genes or even combined enzyme activity and metabolite content measurements. Here we present, for the first time, a comprehensive genetic analysis of all intermediate entities of the path from genotype-to-phenotype, including gene transcription, enzyme activity, and metabolite content. We have shown that natural variation in primary carbohydrate metabolism is extensive in Arabidopsis. A substantial part of this variation was attributable to genetic regulation, resulting in many QTLs detected for the analyzed traits; including 15 QTLs for the 15 enzyme activities and 23 QTLs for the 11 metabolites analysed in this study. Many of those QTLs could be explained by genetic variation in structural genes.

Several other studies in Arabidopsis have reported QTL analyses of carbohydrate metabolism traits in RIL populations. Mitchell-Olds and Pedersen (1998) analyzed activities of ten enzymes among which phosphoglucose isomerase (PGI), phosphoglucomutase (PGM) and fructose-1,6-bisphosphate phosphatase (FBP) in the Col x L*er* RIL population. No QTL was found for FBP, in contrast to our findings. For PGI two QTLs were found at other positions than the three loci identified in our study. The single QTL for PGM on chromosome five however co-located with one of the QTLs identified in our study. PGM activity was also analyzed in the L*er* x Cvi population by Sergeeva *et al.* (2004) who reported at least three QTLs, of which two co-located with the two QTLs found in our study. In another study by Sergeeva *et al.* (2006), soluble acid invertase (Inv) activity was analyzed in the *Ler* x Cvi population revealing several QTLs, among which the one that was confirmed in our analyses.

With respect to metabolite QTLs, amino acid content was analyzed in the Bay-0 x Sha population by Loudet *et al.* (2003). Similar to our results a high number of QTLs were detected of which a few co-located. However, no co-location was observed between the most significant QTLs in both studies. The extracts used in the study of Loudet *et al.* (2003) were also analyzed for starch, glucose, fructose, and sucrose content (Calenge *et al.*, 2006). Multiple QTLs were detected for each analyzed trait under the two different environmental conditions that were tested. QTLs for starch content were not detected in our study, possibly due to differences in sampling time point and growth stage. For glucose, fructose, and sucrose multiple QTLs were also detected in our study. However, co-location with QTLs detected by Calenge *et al.* (2006) was only observed for the strongest QTL for glucose content on chromosome 1 and for a minor QTL for fructose content on chromosome 3. The evident dissimilarities between the different studies might reflect genotypic differences between populations or differences in developmental stage, timing of sampling, or environmental growth conditions. Loudet *et al.* (2003) and Calenge *et al.* (2006) showed large differences in regulation of carbohydrate content when plants were grown under different nitrogen supply regimes. Moreover, Sergeeva *et al.* (2004, 2006) showed organ specific regulation of enzyme activity. These results illustrate that genetic regulation of primary carbohydrate metabolism is under spatial and temporal control involving a multitude of loci, which can be revealed depending on genotype, environment, development stages, and their mutual interactions.

Nevertheless, residual fractions of variance could often not be explained by detected QTLs due to minor environmental and developmental differences between samples, and sampling and analytical variation. When high fractions of unexplained residual variation are observed this might also reflect the complex regulation of primary carbohydrate metabolism due to the genetic regulation by many QTLs, each with a relatively small effect. Such minor QTLs may fail to pass the QTL significance threshold. Segregation of these small-effect QTLs, however, in addition to possible epistatic interactions, may contribute to transgression and to the large genetic variation that is observed. Another indication of the complex regulation of primary carbohydrate metabolism was the finding that specific QTLs were detected for most analyzed traits. When co-location of QTLs for different traits was observed this might often be due to the direct inter-dependence of the traits. For instance, UDP-glucose pyrophosphorylase converts UDP-glucose into glucose-1-phosphate and all three traits map to a similar position on the genome.

Despite the seemingly specific independent regulation of many traits, indicated by the position of the identified QTLs, there was a striking correlation pattern between many traits. Positive correlations were observed between different enzyme activity levels and between enzyme activities and the structural components, such as chlorophyll and proteins, and weaker correlations with some end products, such as sucrose, starch, and amino acids. Negative correlations, however, were observed between enzyme activities and dynamic (phosphorylated) intermediates of carbohydrate metabolic pathways. These results suggest that in addition to the often specific independent regulation of metabolic pathways a more general level of regulation is acting on carbohydrate metabolism in plants, which could be related to the growth and developmental status of the plant. Subsequent data analysis suggested developmental differences to be causal for the observed correlations. The principal component best explaining the variation observed for all traits mapped to the position of *ERECTA* (AT2G26330), a gene well known for its involvement in developmental control of Arabidopsis. The entwinement of plant growth with carbohydrate metabolism was also reported in other studies for enzyme activities (Cross *et al.*, 2006) and metabolite content (Cross *et al.*, 2006; Schauer *et al.*, 2006; Meyer *et al.*, 2007).

Relationship between structural gene expression and enzyme activity

Many metabolic conversions in plants are catalyzed by enzymes and variation in enzymatic activity can have a high impact on metabolic fluxes and metabolite content. It is conceivable that natural variation in enzyme activity is inflicted by genomic variation in the structural genes encoding these enzymes.

We found strong evidence that natural variation for enzyme activity levels is at least partially regulated by variation in structural genes or regulatory loci controlling the transcription of these genes. First, co-location of structural genes and enzyme activity QTLs suggests natural variation for these genes to be causal for the observed variation in enzyme activity. When *cis*-acting eQTLs were detected for these genes, regulation is likely to occur on the transcriptional level, otherwise regulation might act post-transcriptionally, possibly due to altered specific activity or protein stability. Secondly, co-location of *trans*-acting eQTLs for structural genes and enzyme activity QTLs suggests *trans*-regulatory variation of these genes to be causal for the observed variation in enzyme activity. Such regulation is likely to occur through transcriptional regulation of the structural gene due to variation for a distant regulator. Both *cis*- and *trans*-acting transcriptional as well as *cis*-acting post-transcriptional regulation of structural genes were identified as potential causes for observed variation in enzyme activity. However, for many enzymes, QTLs were also detected which did not co-locate with structural genes or their eQTLs, suggesting that regulation occurs at multiple levels, partly independent of variation in (transcript levels of) structural genes. Likewise, for many structural genes eQTLs were detected which did not co-locate with QTLs for enzyme activity, which often explained the low correlation observed between transcript levels and activity. Apparently variation observed in the transcript levels of these genes does not contribute to the variation observed in enzyme activity, suggesting that their encoded proteins might serve other functions than their current annotation, or that variation at the transcriptional level is ‘overruled’ by other regulating mechanisms or by temporal differences between gene expression and subsequent processes. Finally, for a number of structural genes no significant eQTL could be detected which can be the result of low (variation in) transcript levels that could not be detected in the microarray experiment

Often only a weak to medium correlation exists between levels of enzyme activity and transcript levels of structural genes. This can be partly explained by the redundancy in structural genes when different genes each contribute only partially to the eventual level of enzyme. However, different genes of a gene family might have different specific activities for the metabolic conversions under study, for which also natural variation might be present between accessions. In a segregating population this diversity of genetic variants and possible epistatic interactions between them can severely complicate correlation analyses. On the other hand, correlations might be difficult to establish when relationships between transcript levels and protein levels are not linear. Deviations from perfect correlations and linearity can be caused due to delays in protein formation and/or activation upon transcription. Moreover, regulation of enzyme activity can occur post-transcriptionally through mRNA- and protein-stability, protein-folding, activation by or dependency on co-factors, (de)-phosphorylation, etc. Finally, lack of correlation can be simply a result of non-functionality at the sampled developmental stage or due to a dilution effect when genes are only transcribed in specific cells or tissues. Negative correlations might be the result from negative feedback due to high transcription levels of redundant genes, or phase shifts in diurnal rhythms of transcription and translation (Gibon *et al.*, 2004b, 2006; Blasing *et al.*, 2005).

Different modes of action in the genetic control of enzymatic activity

For many enzymes natural variation was observed in their level of activity. In many cases enzyme activity was related to metabolite content, among which substrates and products of the analyzed enzymes. In several cases QTLs for enzyme activity co-located with structural genes encoding these enzymes or eQTLs for those genes. Differences in correlation pattern and QTL profiles between gene expression, enzyme activity and metabolite content indicate, however, that genetic regulation causal for observed variation is not similar for all analyzed traits. Instead, various modes of genetic control, using different mechanisms, seem to act in the regulation of carbohydrate metabolism.

For phosphoglucomutase, one of the enzymes for which the highest variation in activity was observed, it was shown that most of this variation could be explained by genetic factors. Parallel analysis of enzyme activity and structural gene expression suggested *cis*-regulatory variation in transcription of one of the structural genes (At5g51820) to be causal for the major PGM activity QTL. However, differences in the variation in transcription of structural genes and enzyme activity also indicated polymorphisms in coding regions of structural genes at a second locus to account for the observed variation in enzyme activity. Furthermore, although significant negative correlations were observed between PGM activity and its substrate and product G1P and G6P, these correlations are not caused by any of the detected QTLs. This suggests that other levels of regulation are also active for which no genomic variation could be detected within the analyzed population.

In contrast, the combined analysis of variation in the activity of UGP, its substrate and product and transcription of its encoding structural genes suggested *trans*-regulated transcription differences to be the major cause for variation in enzyme activity. In this case the strong positive correlation between UDPG and UGP suggests UDPG levels to be the driving force for this *trans*-acting regulation. This would mean that plants are able to sense and respond to changes in UDPG accumulation, which has been suggested and shown also for other sugars (Rolland *et al.*, 2002; Halford *et al.*, 2003; Avonce *et al.*, 2005; Gonzali *et al.*, 2006). Although it remains speculative to assign which genetic factor(s) determine(s) the variation observed in UDPG accumulation it is interesting to note that the inorganic phosphate status in Arabidopsis affects the transcription of UGP-encoding genes (Ciereszko *et al.*, 2001, 2005). Moreover, natural variation for phosphate and phytate, the major source of inorganic phosphate in plants, has been observed in the L*er* x Cvi population and a common QTL explaining most of the variation co-locates with the QTL for UDPG-content and UGP-activity (Bentsink *et al.*, 2003). Furthermore, a QTL for the accumulation of the phosphorylated hexoses G1P and G6P was detected at this position, which might indicate that high levels of inorganic phosphate results in elevated levels of phosphorylated sugars. In conclusion, variation in phosphorus levels would then regulate the accumulation of UDPG, which in turn triggers the expression of UGP-encoding structural genes, leading to higher activity of UGP.

MATERIALS AND METHODS

Plant material and tissue collection

Aerial parts of seedlings from the accessions L*er* and Cvi and a population of 160 recombinant inbred lines derived from a cross between these parents (Alonso-Blanco *et al.*, 1998; Keurentjes *et al.*, 2006) were grown and collected as described previously (Keurentjes *et al.*, 2006). In brief, seeds of lines were sown in petri dishes on 1/2MS agar and placed in a cold room for seven days. Petri dishes were then transferred to a climate chamber and seedlings were collected after seven days. Plant material was stored at -80°C until further processing.

Linkage map construction and anchoring to the physical map

The genetic map was constructed from a subset of the markers available, at http:/nasc.nott.ac.uk/, as described in Keurentjes *et al.* (2007b). In total, 144 markers were used, with an average spacing of 3.5 cM. The largest distance between two markers was 10.8 cM. The genetic map was anchored to the physical map as described in Keurentjes *et al.* (2007b), with an almost linear genome-wide relation of 4.1 cM per Mbp.

Metabolite and enzyme measurements

Metabolites were extracted and analyzed as described previously; ChlA, ChlB, AA, protein, sucrose, glucose, and fructose (Cross *et al.*, 2006); starch, G1P, and G6P (Gibon *et al.*, 2002); UDPG (Keurentjes *et al.*, 2006). Enzymes were extracted as described in Gibon *et al.* (2004a) and analyzed as described previously; Inv, AGP, FBP, G6PDH, PFK, PFP, SPS, SuSy, GK, FK, and UGP (Gibon *et al.*, 2004a); PGI (Cross *et al.*, 2006); PGM (Manjunath *et al.*, 1998); Rubisco (Sulpice *et al.*, 2007). Samples were randomized during extraction and analysis, and two biological replicates were analyzed for each trait.

Microarray analyses

Transcript levels of genes were analyzed on two-color DNA-microarrays as described previously (Keurentjes *et al.*, 2007b). Resulting ^{2}log signal intensities were used for correlation analyses and ^{2}log ratios between co-hybridized RILs were used for QTL analyses.

Statistical analyses

Spearman rank correlations were determined in Excel (Microsoft) for mean trait values as follows:

$$R_{jk} = \frac{n\left(n^2 - 1\right) - 6\sum_{i=1}^{n}\left(y_{ij} - y_{ik}\right)^2 - \frac{1}{2}\left(T_j + T_k\right)}{\sqrt{\left[n\left(n^2 - 1\right) - T_j\right]\left[n\left(n^2 - 1\right) - T_k\right]}}, \quad j, k = 1,2,\ldots,m;$$

where n is the number of observations, y is the rank of observations for variables j to m, and $T_j = \sum t_j\left(t_j^2 - 1\right)$, t_j being the number of ties of a particular value of variable j, and the summation being over all tied values of variable j (Siegel, 1956).

QTL analyses for gene transcript levels were performed as described in (Keurentjes *et al.*, 2007b). For QTL analyses of metabolite and enzyme traits the computer program MapQTL version 5.0 (Van Ooijen, 2004) was used to identify and locate QTLs linked to the molecular markers using multiple QTL mapping (MQM). LOD statistics were calculated at 0.5 cM intervals. Tests of 1000 permutations were used to obtain an estimate of the number of type 1 errors (false positives). The genome-wide LOD score, which 95% of the permutations did not exceed, ranged from 2.4 to 2.7. A LOD score of 3.0, to correct for multiple testing, was then used as the significance threshold to declare the presence of a QTL. In the MQM model the genetic effect (μ_B-μ_A) and percentage of explained variance was estimated for each QTL, and 2 Mbp-support intervals were established as an ~95% confidence level (Van Ooijen, 1992). Co-location of (e)QTLs was defined as an overlap in the 2Mbp-support intervals.

Genomic positions of genes were inferred from the Arabidopsis information resource (The Arabidopsis Genome Initiative, 2000). When physical positions of genes fell in the 2 Mbp-support interval of (e)QTLs this was considered as co-location.

Principal component and box plot analyses were performed in SPSS (version 12.0).

Acknowledgements

We thank Linus van der Plas for critical reading of the manuscript. This work was supported by grants from the Netherlands Organization for Scientific Research, Program Genomics (050-10-029), the Centre for Biosystems Genomics (CBSG, Netherlands Genomics Initiative) and a Genomics Fellowship from The Netherlands Genomics Initiative (050-72-412).

REFERENCES

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

Avonce, N., Leyman, B., Thevelein, J. and Iturriaga, G. (2005). Trehalose metabolism and glucose sensing in plants. *Biochem Soc Trans* **33,** 276-279.

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.

Blasing, O.E., Gibon, Y., Gunther, M., Hohne, M., Morcuende, R., Osuna, D., Thimm, O., Usadel, B., Scheible, W.R. and Stitt, M. (2005). Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in Arabidopsis. *Plant Cell* **17,** 3257-3281.

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.

Carrari, F., Urbanczyk-Wochniak, E., Willmitzer, L. and Fernie, A.R. (2003). Engineering central metabolism in crop species: learning the system. *Metab Eng* **5,** 191-200.

Causse, M., Rocher, J.P., Henry, A.M., Charcosset, A., Prioul, J.L. and De Vienne, D. (1995). Genetic dissection of the relationship between carbon metabolism and early growth in maize, with emphasis on key enzyme loci. *Mol Breed* **1,** 259-272.

Chen, X., Salamini, F. and Gebhardt, C. (2001). A potato molecular-function map for carbohydrate metabolism and transport. *Theor Appl Genet* **102,** 284-295.

Ciereszko, I., Johansson, H., Hurry, V. and Kleczkowski, L.A. (2001). Phosphate status affects the gene expression, protein content and enzymatic activity of UDP-glucose pyrophosphorylase in wild-type and pho mutants of Arabidopsis. *Planta* **212,** 598-605.

Ciereszko, I., Johansson, H. and Kleczkowski, L.A. (2005). Interactive effects of phosphate deficiency, sucrose and light/dark conditions on gene expression of UDP-glucose pyrophosphorylase in Arabidopsis. *J Plant Physiol* **162,** 343-353.

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.

El-Lithy, M.E., Clerkx, E.J., 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.

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.

Fiehn, O., Kloska, S. and Altmann, T. (2001). Integrated studies on plant biology using multiparallel techniques. *Curr Opin Biotechnol* **12,** 82-86.

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.

Gachon, C.M., Langlois-Meurinne, M., Henry, Y. and Saindrenan, P. (2005). Transcriptional co-regulation of secondary metabolism enzymes in Arabidopsis: functional and evolutionary implications. *Plant Mol Biol* **58,** 229-245.

Gibon, Y., Vigeolas, H., Tiessen, A., Geigenberger, P. and Stitt, M. (2002). Sensitive and high throughput metabolite assays for inorganic pyrophosphate, ADPGlc, nucleotide phosphates, and glycolytic intermediates based on a novel enzymic cycling system. *Plant J* **30,** 221-235.

Gibon, Y., Blaesing, O.E., Hannemann, J., Carillo, P., Hohne, M., Hendriks, J.H., Palacios, N., Cross, J., Selbig, J. and Stitt, M. (2004a). 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.

Gibon, Y., Blasing, O.E., Palacios-Rojas, N., Pankovic, D., Hendriks, J.H., Fisahn, J., Hohne, M., Gunther, M. and Stitt, M. (2004b). Adjustment of diurnal starch turnover to short days: depletion of sugar during the night leads to a temporary inhibition of carbohydrate utilization, accumulation of sugars and post-translational activation of ADP-glucose pyrophosphorylase in the following light period. *Plant J* **39,** 847-862.

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

Gonzali, S., Loreti, E., Solfanelli, C., Novi, G., Alpi, A. and Perata, P. (2006). Identification of sugar-modulated genes and evidence for in vivo sugar sensing in Arabidopsis. *J Plant Res* **119,** 115-123.

Halford, N.G., Hey, S., Jhurreea, D., Laurie, S., McKibbin, R.S., Paul, M. and Zhang, Y. (2003). Metabolic signalling and carbon partitioning: role of Snf1-related (SnRK1) protein kinase. *J Exp Bot* **54,** 467-475.

Harrison, J., Hirel, B. and Limami, A.M. (2004). Variation in nitrate uptake and assimilation between two ecotypes of *Lotus japonicus* and their recombinant inbred lines. *Physiol Plant* **120,** 124-131.

Hirai, M.Y., Klein, M., Fujikawa, Y., Yano, M., Goodenowe, D.B., Yamazaki, Y., Kanaya, S., Nakamura, Y., Kitayama, M., Suzuki, H. *et al.* (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.

Hirel, B., Bertin, P., Quillere, I., Bourdoncle, W., Attagnant, C., Dellay, C., Gouy, A., Cadiou, S., Retailliau, C., Falque, M. *et al.* (2001). Towards a better understanding of the genetic and physiological basis for nitrogen use efficiency in maize. *Plant Physiol* **125,** 1258-1270.

Juenger, T.E., McKay, J.K., Hausmann, N., Keurentjes, J.J.B., Sen, S., Stowe, K.A., Dawson, T.E., Simms, E.L. and Richards, J.H. (2005). Identification and characterization of QTL underlying whole-plant physiology in *Arabidopsis thaliana*: delta13C, stomatal conductance and transpiration efficiency. *Plant Cell Environ.* **28,** 697-708.

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

Keurentjes, J.J.B., Bentsink, L., Alonso-Blanco, C., Hanhart, C.J., Blankestijn-De Vries, H., 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., Terpstra, I.R., Garcia, J.M., van den Ackerveken, G., Snoek, L.B., Peeters, A.J., 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.

Koch, K. (2004). Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development. *Curr Opin Plant Biol* **7,** 235-246.

Kofler, H., Hausler, R.E., Schulz, B., Groner, F., Flugge, U.I. and Weber, A. (2000). Molecular characterisation of a new mutant allele of the plastid phosphoglucomutase in Arabidopsis, and complementation of the mutant with the wild-type cDNA. *Mol Gen Genet* **263,** 978-986.

Koornneef, M., Alonso-Blanco, C. and Vreugdenhil, D. (2004). Naturally occurring genetic variation in *Arabidopsis Thaliana*. *Annu Rev Plant Physiol Plant Mol Biol* **55,** 141-172.

Li, L., Strahwald, J., Hofferbert, H.R., Lubeck, J., Tacke, E., Junghans, H., Wunder, J. and Gebhardt, C. (2005). DNA variation at the invertase locus invGE/GF is associated with tuber quality traits in populations of potato breeding clones. *Genetics* **170,** 813-821.

Loudet, O., Chaillou, S., Merigout, P., Talbotec, J. and Daniel-Vedele, F. (2003). Quantitative trait loci analysis of nitrogen use efficiency in Arabidopsis. *Plant Physiol* **131,** 345-358.

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

Manjunath, S., Lee, C.H., VanWinkle, P. and Bailey-Serres, J. (1998). Molecular and biochemical characterization of cytosolic phosphoglucomutase in maize. Expression during development and in response to oxygen deprivation. *Plant Physiol* **117,** 997-1006.

Martienssen, R.A. (2000). Weeding out the genes: the Arabidopsis genome project. *Funct Integr Genomics* **1,** 2-11.

Masle, J., Gilmore, S.R. and Farquhar, G.D. (2005). The ERECTA gene regulates plant transpiration efficiency in Arabidopsis. *Nature* **436,** 866-870.

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

Mitchell-Olds, T. and Pedersen, D. (1998). The molecular basis of quantitative genetic variation in central and secondary metabolism in Arabidopsis. *Genetics* **149,** 739-747.

Morcuende, R., Bari, R., Gibon, Y., Zheng, W., Pant, B.D., Blasing, O., Usadel, B., Czechowski, T., Udvardi, M.K., Stitt, M. *et al.* (2007). Genome-wide reprogramming of metabolism and regulatory networks of Arabidopsis in response to phosphorus. *Plant Cell Environ* **30,** 85-112.

Neuhaus, H.E., Kruckeberg, A.L., Feil, R., Gottlieb, L. and Stitt, M. (1989). Dosage mutants of phosphoglucose isomerase in the cytosol and chloroplasts of *Clarkia xantiana*. II. Study of the mechanisms which regulate photosynthate partitioning. *Planta* **178,** 110-122.

Osuna, D., Usadel, B., Morcuende, R., Gibon, Y., Blasing, O.E., Hohne, M., Gunter, M., Kamlage, B., Trethewey, R., Scheible, W.R. *et al.* (2007). Temporal responses of transcripts, enzyme activities and metabolites after adding sucrose to carbon-deprived Arabidopsis seedlings. *Plant J* **49,** 463-491.

Periappuram, C., Steinhauer, L., Barton, D.L., Taylor, D.C., Chatson, B. and Zou, J. (2000). The plastidic phosphoglucomutase from Arabidopsis. A reversible enzyme reaction with an important role in metabolic control. *Plant Physiol* **122,** 1193-1199.

Prioul, J.L., Pelleschi, S., Sene, M., Thevenot, C., Causse, M., de Vienne, D. and Leonardi, A. (1999). From QTLs for enzyme activity to candidate genes in maize. *J Exp Bot* **50,** 1281-1288.

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

Rockman, M.V. and Kruglyak, L. (2006). Genetics of global gene expression. *Nat Rev Genet* **7,** 862-872.

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

Rolland, F., Moore, B. and Sheen, J. (2002). Sugar sensing and signaling in plants. *Plant Cell* **14 Suppl,** S185-205.

Rontein, D., Dieuaide-Noubhani, M., Dufourc, E.J., Raymond, P. and Rolin, D. (2002). The metabolic architecture of plant cells. Stability of central metabolism and flexibility of anabolic pathways during the growth cycle of tomato cells. *J Biol Chem* **277,** 43948-43960.

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

Sergeeva, L.I., Vonk, J., Keurentjes, J.J.B., van der Plas, L.H., Koornneef, M. and Vreugdenhil, D. (2004). Histochemical analysis reveals organ-specific quantitative trait loci for enzyme activities in Arabidopsis. *Plant Physiol* **134,** 237-245.

Sergeeva, L.I., Keurentjes, J.J.B., Bentsink, L., Vonk, J., van der Plas, L.H., Koornneef, M. and Vreugdenhil, D. (2006). Vacuolar invertase regulates elongation of *Arabidopsis thaliana* roots as revealed by QTL and mutant analysis. *Proc Natl Acad Sci U S A* **103,** 2994-2999.

Siegel, S. (1956). Non-parametric statistics for the behavioral sciences. (New York: McGraw-Hill).

Sturm, A. and Tang, G.Q. (1999). The sucrose-cleaving enzymes of plants are crucial for development, growth and carbon partitioning. *Trends Plant Sci* **4,** 401-407.

Sulpice, R., Tschoep, H., von Korff, M., Bussis, D., Usadel, B., Hoehne, M., Witucka-Wall, H., Altmann, T., Stitt, M. and Gibon, Y. (2007). Description and applications of a rapid and sensitive non-radioactive microplate-based assay for maximum and initial activity of ribulose-1,5-bisphosphate carboxylase. *Plant Cell Environ* **In press,** doi: 10.1111/j.1365-3040.2007.01679.x.

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

Torii, K.U., Mitsukawa, N., Oosumi, T., Matsuura, Y., Yokoyama, R., Whittier, R.F. and Komeda, Y. (1996). The Arabidopsis ERECTA gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats. *Plant Cell* **8,** 735-746.

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.

Van Ooijen, J.W. (1992). Accuracy of mapping quantitative trait loci in autogamous species. *Theor Appl Genet* **84,** 803-811.

Van Ooijen, J.W. (2004). MapQTL 5, Software for the mapping of quantitative trait loci in experimental populations (Wageningen, The Netherlands: Kyazma B.V.).

Winnacker, E.L. (2003). Interdisciplinary sciences in the 21st century. *Curr Opin Biotechnol* **14,** 328-331.

Chapter 6

General discussion

One of the intriguing observations in nature is the enormous diversity in characteristic properties of various species. However, natural variation can also be observed within species (Alonso-Blanco and Koornneef, 2000), which is supposed to be one of the driving forces of species formation. Farmers and breeders have used natural occurring genetic variation for centuries to improve crop species (Koornneef and Stam, 2001; Zamir, 2001). The identification of the genetic factors controlling natural variation would therefore improve our understanding of genetic regulatory processes and give insight into the evolutionary significance of variation (Mitchell-Olds and Schmitt, 2006). For many traits quantitative variation is observed, suggesting that it is controlled by multiple genes. Possible interactions between genes and between genes and the environment further add to the complexity of quantitative traits, making the genetic dissection of such traits difficult.

A classical first step in the genetic analysis of traits is the determination of inheritance patterns in the progeny of a cross between distinct varieties. Quantitative traits do not segregate in distinct classes but instead display a more continuous variation in trait values as a result from the segregation of multiple independent loci. To relate these quantitative trait loci (QTLs) to genomic positions it is pivotal to be able to determine the genotype of segregants. The development of molecular markers greatly enhanced the ease in which mapping populations can be genotyped. Molecular markers represent genomic polymorphisms between genotypically different lines. By crossing distinct accessions, numerous polymorphisms will segregate in a progeny, enabling the construction of genetic maps. Because quantitative traits may also segregate in the same offspring population, QTLs for these traits can then be mapped by analyzing co-segregation of trait values with molecular markers used for the construction of the genetic map (Broman, 2001; Doerge, 2002; Jansen, 2003).

Depending on the species and the ease by which mapping populations can be generated several approaches have been applied. A relatively fast approach, requiring a minimal number of generations, is the generation of an F_2 or back cross (BC) population. However, such populations still contain a high level of heterozygosity which may compromise the construction of genetic maps, especially when dominant markers are being used. Moreover, such populations can not be propagated sexually without further segregation of the heterozygous regions, making additional genotyping in later generations necessary. Populations consisting of homozygous lines on the other hand, need several rounds of selfing or backcrossing to reach full homozygosity. Alternatively, homozygosity can be obtained by generating double haploids. When fully homozygous, lines can be propagated without introducing further genotypic changes in their progeny. At this stage the population has become immortal and a single round of genotyping is sufficient to generate a genetic map for any further experimentation. Homozygous lines offer the advantage of replicated measurements at genotypically identical individuals and also allow comparing different experiments in time and environment.

In Arabidopsis, recombinant inbred lines (RILs) have become the mapping population of choice because of its selfing nature and short generation times (Somerville and Koornneef, 2002). In other species however, near isogenic lines (NILs) are sometimes favorable because of sterility problems and intolerance towards inbreeding (Eshed and Zamir, 1995). Both types of population can be used for mapping purposes although they differ markedly in their genetic make-up due to differences in the crossing scheme. RILs are generated from an F_1 without backcrossing and therefore contain on average equal contributions of both parental genomes. NILs on the other hand, are generated from an F_1 through repeated backcrossing with a recurrent parent and contain only a limited amount of the donor genome. As a result RILs can contain multiple introgressions whereas NILs preferably contain only a single introgression into an otherwise isogenic background. Because of these differences, different mapping strategies are required for the two types of population.

Chapter two of this thesis described the development of the first genome-wide coverage NIL population in Arabidopsis allowing the comparison of mapping purposes with an already existing RIL population derived from the same parental accessions (Alonso-Blanco *et al.*, 1998). These comparisons revealed a higher mapping power for small effect loci but lower mapping resolution in the NIL population compared to the RIL population. However, results were greatly depending on the genetic architecture of traits and population size and structure. For RIL populations both mapping power and resolution can be increased by increasing population size, which would also diminish the need for replicated measurements. For NIL populations resolution can also be improved by adding more lines but, depending on the size of introgressions and the amount of overlap, power has to be increased by replicated measurements. Because of the much higher recombination frequency, RIL populations are often favorable over NIL populations when mapping experiments are limited by the number of plants that can be analyzed. The segregation of multiple loci in RIL populations might mask small-effect QTLs, but allows the detection of genetic interactions, which is not possible in NIL populations. Nils have been shown to be very useful for the confirmation of QTLs and as starting material for the fine-mapping of so-called Mendelized QTL.

Although natural phenotypic variation can be observed for many quantitative traits in Arabidopsis, which can be effectively analyzed in RIL populations (Alonso-Blanco and Koornneef, 2000; Koornneef *et al.*, 2004), QTL analysis often reveals only a limited number of steps in the complex regulatory pathways of quantitative traits. The path from genotype to phenotype often involves multiple intermediate steps and it is therefore difficult to determine whether QTLs regulate traits directly or indirectly. Moreover, regulation can occur at different levels, ranging from variation in presence or expression of genes to variation in protein function. Until the cloning of a QTL and the identification of the causal polymorphism(s) it therefore remains uncertain at which point in a pathway traits are regulated. To fully understand the complex regulation of quantitative traits it is therefore recommendable to genetically analyze different levels and intermediates at which genetic control might act (Fiehn *et al.*, 2001; Winnacker, 2003).

The recent advance in analytical technologies (transcriptomics, proteomics and metabolomics) now enables the large-scale genetic analysis of different entities in the circuitry from gene to phenotype. The expression of genes often determines the onset of pathways resulting in a particular phenotype. Therefore, phenotypic variation might be inflicted by variation in gene expression. At the other end of the information flow from DNA sequence to gene function, metabolites, as products from the encoded proteins, stand closest to the eventual phenotype. It is conceivable that genetically controlled variation in metabolite composition and accumulation determines, at least partly, the observed phenotypic variation. In chapter three and four high throughput ‘omic’ technologies were used for the analysis of natural variation in gene expression (transcriptomics) and metabolite composition and content (metabolomics).

Analogous to ‘classical’ quantitative phenotypic traits, natural variation can also be observed for gene expression, when that expression is under genetic control (Borevitz *et al.*, 2003; Kliebenstein *et al.*, 2006). Chapter three describes the genetic analysis of genome-wide gene expression variation in Arabidopsis (genetical genomics) (Jansen and Nap, 2001). These analyses revealed extensive genetic control of gene expression, judged from the fact that for more than 4,000 genes expression QTLs (eQTLs) could be detected. However, many more genes showed high heritability values, even though no eQTL could be detected. This suggests that their expression is regulated by multiple eQTLs, of which many might not have passed the stringent significance threshold due to their small effect. Both local and distant regulation (Rockman and Kruglyak, 2006) was observed although local regulation was often much stronger. Local regulation might be a result from polymorphisms in *cis*-regulatory elements which directly affects the expression of the gene under study. Interestingly, local regulation correlated with polymorphism frequency, further supporting the suggestion that expression variation can be a result from local sequence differences. Moreover, regulatory genes showed much less local regulation, which can be explained by much stronger conservation due to their pleiotropic effects.

Distant regulation on the other hand, is most likely a result from polymorphisms in a regulator, affecting expression in *trans*, possibly through multiple intermediates. Since regulators may exert pleiotropic effects on numerous genes, directly or indirectly, multiple eQTLs would map to the position of this regulator. Indeed several eQTL hot spots were identified, possibly containing such master regulators. The detection of distant eQTLs indicates that the expression of the gene under study is controlled by genetic factors in *trans*. When the causal gene underpinning the *trans* eQTL can be identified this allows the possibility of establishing gene regulation networks (Jansen, 2003). However, QTL support intervals often contain hundreds of genes, each of which can be a candidate regulator. Positive confirmation of a candidate can only be obtained upon cloning of the eQTL, which is practically difficult to achieve for genome-wide expression studies. The assignment of candidate genes therefore relies on additional information, such as co-expression and gene ontology. The power of such a computational approach was demonstrated by the reconstruction of a regulatory network for genes involved in the regulation of flowering time. Nonetheless, variation in expression of genes can not always be explained by expression differences of their regulator, especially when expression of the regulator is not *cis*-regulated. When the causal polymorphism(s) reside(s) in the coding sequence of the regulator this might alter protein function or stability and the expression of target genes then depends on the allelic form of the regulator rather than on its expression level. The identification of such relationships from additional information such as genome sequences, e.g. binding site data, and experimentation, *e.g.* protein-DNA interaction data, can further improve the assignment of candidate regulators and ultimately the construction of regulatory networks. The recognition of genetically controlled expression of genes and the reduction in the number of candidate regulators through QTL analysis should therefore further guide the detailed analysis of gene-by-gene regulation.

Unlike microarrays for the genome-wide analysis of the transcriptome, no single platform exists for the simultaneous analysis of the complete metabolome. In contrast to mRNA, which has an identical chemical structure for all genes, metabolites represent a plethora of different chemical classes and no universal analytical tools are available yet. However, advances in mass spectrometry have made the detection and quantification of hundreds of compounds of specific chemical classes possible (Fiehn *et al.*, 2000). Plants are especially rich in the number of secondary metabolites, which is possibly a consequence of their sessile nature. Since plants are unable to migrate they need to adapt to local environments for their survival. The wide range of habitats makes it amenable that natural variation in secondary metabolite composition and accumulation plays an important role in the diversification of plants (Fiehn, 2002).

Chapter four describes the genetic analysis of metabolite composition in Arabidopsis using large-scale untargeted liquid chromatography-time of flight mass spectrometry (LC-QTOF MS). Although untargeted, LC-MS predominantly detects semi-polar secondary metabolites, which are amongst the most variable compounds in nature. When applied to accessions originating from various parts of the global distribution range of *A. thaliana*, considerable quantitative and qualitative differences were observed. The majority of compounds could only be detected in a limited number of accessions and each accession analyzed contained unique compounds not found in any other accession. However, a substantial number of compounds, presumably representing essential metabolites, could be detected in all accessions analyzed. The extensive natural variation in metabolite content together with the often observed high heritabilities indicates that metabolite composition is largely under genetic control. Indeed, when quantification in a RIL population was subjected to QTL analysis, for 75% of the detected compounds significant QTL(s) could be assigned. The impact of genetic factors on the dynamic range of metabolite content was further demonstrated by the fact that a high number of compounds which were not found in either one of the parents could be detected in RILs. This suggests that metabolic pathways in the parents are blocked at different steps which can be overcome by complementation due to recombination of their genomes. Natural variation therefore offers a large potential for metabolic engineering of crop species through classical breeding.

Similar to the distribution of eQTLs along the genome, hot- and cold-spots could be observed for metabolite accumulation QTLs. Interestingly, for both analyses a hotspot was observed at the position of the *ERECTA* gene, a receptor protein kinase well known for its pleiotropic effects (Torii *et al.*, 1996). The *ERECTA* gene is polymorphic for the parental accessions of the RIL population and causal for many of the morphological differences observed between the parents. Co-location of QTLs implies that accumulation of the metabolites mapping to the same position might be controlled by a common regulator. Although no information can be inferred whether this control acts directly or indirectly through downstream effects of a regulatory step, co-regulated metabolites are likely to be part of a common pathway or involved in the same biological process. When those metabolites can be identified, information can be obtained about the mechanism of regulation, and the number and order of metabolites in a metabolic pathway. These features were demonstrated by the reconstruction of the aliphatic glucosinolate formation pathway and the discovery of variation in glycosyl transferase activity affecting flavonol composition. However, untargeted metabolomic approaches detect anonymous compounds and the identification of these compounds is still in its infancy (Schauer *et al.*, 2005; Moco *et al.*, 2006). The unraveling of metabolic pathways would therefore benefit much from the development of mass identification libraries.

The large-scale genetic analyses of gene expression and metabolite content clearly have shown their usefulness in constructing genetic regulatory networks. Yet, none of these approaches can fully explain the complex regulation of phenotypic quantitative traits. Moreover, interactions and cross-talk between components of the various regulatory levels are probably eminent and it is not always possible to distinguish cause and consequence of natural variation without further experimentation. However most of the tools, including the genome sequence, are now available to study biological systems as a genetic system in its entirety. The integration of data collected from multiparallel analyses of the various interconnected transducers of biological information flow will therefore thrive our understanding of complex biological systems.

Chapter five describes the integrative analysis of genetic variation in enzyme activities of primary carbohydrate metabolism. Carbohydrates are essential for many biological processes ranging from growth to energy metabolism and plants contain a multitude of enzymes for the allocation and conversion of the necessary compounds. Perturbations affecting the functionality of these enzymes can therefore have large effects on plant growth. The activity of 15 enzymes involved in carbohydrate metabolism were analyzed in the Landsberg *erecta* x Cape Verde Islands RIL population and subjected to QTL analyses. In addition, the expression of the structural genes encoding those enzymes and a number of carbohydrate metabolites, as substrate and products of the enzymes, were analyzed in parallel.

The natural variation observed for a large number of enzymes and metabolites could partly be explained by detected QTLs. Moreover, both positive and negative correlations were observed between enzyme activities and metabolite contents, although only few co-locating QTLs were detected. These findings suggested that genetic control of primary carbohydrate metabolism acts at different levels: a direct independent regulation of individual components and a more general simultaneous regulation of all components. Principal component analyses further suggested that such simultaneous regulation of carbohydrate metabolism might be under developmental control.

The parallel analysis of structural gene expression and enzyme activity also revealed distinct modes of regulation. From the position of structural genes, their eQTLs and enzyme activity QTLs, together with correlation analyses of gene expression and enzyme activity levels the involvement of structural gene variation could be evaluated. In a number of cases *cis*-regulated expression variation of structural genes was suggested to be causal for observed variation in enzyme activity. However, *trans*-regulated expression variation was also observed and might have contributed to the observed variation in activity for some enzymes. Furthermore the lack of expression variation in some instances indicated altered protein function to affect specific activity of enzymes. Finally, the detection of enzyme activity QTLs not co-locating with structural genes encoding the enzyme under study suggests other regulatory mechanisms, independent of structural genes, to be active (*e.g.* post-translational control). The different regulatory mechanisms, including the role of metabolites, were demonstrated by the detailed analysis of phosphoglucomutase and UDP-glucose pyrophosphorylase activity, their structural genes and their respective substrates and products.

The work described in this thesis has shown the extensive variation in quantitative traits in Arabidopsis including variation in gene expression and metabolite content. The use of natural variation in combination with genetic approaches such as QTL analyses has further shown the power in elucidating the often complex genetic regulation of traits. Moreover, the application of high throughput 'omic' technologies enabled the construction of regulatory networks which were unlikely to be uncovered from targeted small scale approaches. However QTL analyses are limited by the amount of natural variation segregating in the mapping population and the genetic make-up of the employed RIL population only consists of two genotypes. The analysis of traits in multiple populations, generated from different accessions, is likely to reveal additional regulatory steps. Alternatively, multiple distinct accessions could be intercrossed in a single mapping population, thereby increasing the amount of segregating natural variation. The ultimate mapping population however, consists of the worldwide collection of accessions and the recent advances in linkage disequilibrium mapping have only just begun to make the exploration of this comprehensive reservoir of natural variation possible (Nordborg *et al.*, 2002).

Another impediment obstructing the comprehensive elucidation of genetic regulation is the often observed spatial and temporal control of quantitative traits. Due to cost and time considerations analyses are often limited in the number of developmental stages and tissues that can be sampled. However, to get a full understanding of the complex regulatory mechanisms of traits it is recommendable to analyze traits in multiple developmental stages and tissues. Likewise, for many traits genetic interactions with the environment are observed and analyzing traits in different circumstances might therefore reveal specific regulatory steps.

Although powerful in mapping genomic regions, the resolution of QTL analysis is often not high enough to identify the causal gene (QTG), and ultimately the causal changes at the nucleotide level (QTN), affecting the trait of interest. Due to the often small effects of QTLs and the complex regulation of traits, it is not always easy to obtain definitive proof for the identification of a QTG or QTN. Initial QTL mapping is usually followed by confirmation in NILs, which can also be used for fine-mapping. Once a select set of candidate genes has been defined, several lines of experimentation can be followed to provide evidence for the identification of a QTG or QTN. Such lines include natural variation surveys within and between species, comparative sequence analyses, gene expression analyses, functional (*in vitro*) gene analyses, knockout or mutational analyses, and (transgenic) complementation tests (Borevitz and Nordborg, 2003; Weigel and Nordborg, 2005). Although some of these analyses will provide stronger evidence than others, usually several tests are needed to demonstrate a causal link between allelic variation and a particular phenotype.

Finally, here Arabidopsis was chosen as a model plant for the genetic analyses of quantitative variation. The availability of the complete genome sequence, commercially available genome-wide microarrays and the publicly available high quality mapping populations together with the numerous tools and techniques developed for this species make Arabidopsis the perfect choice for these kinds of analyses (Alonso and Ecker, 2006). However, due to the developments in sequence technologies and comparative genomics many of the findings in Arabidopsis can be readily 'translated' to other species (Gale and Devos, 1998; Hall *et al.*, 2002). Moreover, the rapid progress in genomic technologies and the increasing number of mapping populations for other (crop) species should no longer restrict the analyses described in this thesis to model species. Many of the tools and techniques developed in Arabidopsis can readily be applied to other species and now the time has come that applied sciences will benefit from the groundbreaking work in model species such as *Arabidopsis thaliana.*

REFERENCES

Alonso, J.M. and Ecker, J.R. (2006). Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in Arabidopsis. *Nat Rev Genet* **7,** 524-536.

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

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

Borevitz, J.O., Liang, D., Plouffe, D., Chang, H.S., Zhu, T., Weigel, D., Berry, C.C., Winzeler, E. and Chory, J. (2003). Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res* **13,** 513-523.

Borevitz, J.O. and Nordborg, M. (2003). The impact of genomics on the study of natural variation in Arabidopsis. *Plant Physiol* **132,** 718-725.

Broman, K.W. (2001). Review of statistical methods for QTL mapping in experimental crosses. *Lab Anim (NY)* **30,** 44-52.

Doerge, R.W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet* **3,** 43-52.

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.

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

Fiehn, O., Kloska, S. and Altmann, T. (2001). Integrated studies on plant biology using multiparallel techniques. *Curr Opin Biotechnol* **12,** 82-86.

Fiehn, O. (2002). Metabolomics--the link between genotypes and phenotypes. *Plant Mol Biol* **48,** 155-171.

Gale, M.D. and Devos, K.M. (1998). Plant comparative genetics after 10 years. *Science* **282,** 656-659.

Hall, A.E., Fiebig, A. and Preuss, D. (2002). Beyond the Arabidopsis genome: opportunities for comparative genomics. *Plant Physiol* **129,** 1439-1447.

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

Jansen, R.C. (2003). Studying complex biological systems using multifactorial perturbation. *Nat Rev Genet* **4,** 145-151.

Kliebenstein, D.J., West, M.A., van Leeuwen, H., Kim, K., Doerge, R.W., Michelmore, R.W. and St Clair, D.A. (2006). Genomic survey of gene expression diversity in *Arabidopsis thaliana*. *Genetics* **172,** 1179-1189.

Koornneef, M. and Stam, P. (2001). Changing paradigms in plant breeding. *Plant Physiol* **125,** 156-159.

Koornneef, M., Alonso-Blanco, C. and Vreugdenhil, D. (2004). Naturally occurring genetic variation in *Arabidopsis Thaliana*. *Annu Rev Plant Physiol Plant Mol Biol* **55,** 141-172.

Mitchell-Olds, T. and Schmitt, J. (2006). Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature* **441,** 947-952.

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

Nordborg, M., Borevitz, J.O., Bergelson, J., Berry, C.C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J.N., Noyes, T., Oefner, P.J. *et al.* (2002). The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat Genet* **30,** 190-193.

Rockman, M.V. and Kruglyak, L. (2006). Genetics of global gene expression. *Nat Rev Genet* **7,** 862-872.

Schauer, N., Steinhauser, D., Strelkov, S., Schomburg, D., Allison, G., Moritz, T., Lundgren, K., Roessner-Tunali, U., Forbes, M.G., Willmitzer, L. *et al.* (2005). GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett* **579,** 1332-1337.

Somerville, C. and Koornneef, M. (2002). Timeline: A fortunate choice: the history of Arabidopsis as a model plant. *Nat Rev Genet* **3,** 883-889.

Torii, K.U., Mitsukawa, N., Oosumi, T., Matsuura, Y., Yokoyama, R., Whittier, R.F. and Komeda, Y. (1996). The Arabidopsis ERECTA gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats. *Plant Cell* **8,** 735-746.

Weigel, D. and Nordborg, M. (2005). Natural variation in Arabidopsis. How do we find the causal genes? *Plant Physiol* **138,** 567-568.

Winnacker, E.L. (2003). Interdisciplinary sciences in the 21st century. *Curr Opin Biotechnol* **14,** 328-331.

Zamir, D. (2001). Improving plant breeding with exotic genetic libraries. *Nat Rev Genet* **2,** 983-989.

Summary

Plants show considerable genetic differences between accessions of the same species, which are reflected in phenotypic variation. This natural variation is often displayed as a continuous distribution of trait values and is therefore called quantitative variation. Quantitative variation is the result of the interplay of multiple genes and environmental factors. Because the contribution of each gene to the eventual phenotype can be quite small, sophisticated statistical methods are needed to associate genomic regions with the trait of interest. Such an approach is known as quantitative trait locus (QTL) analysis. In QTL analysis, the trait of interest is quantified in a genotyped mapping population, derived from a cross between distinct genotypes.

Arabidopsis thaliana is the leading model species in modern plant sciences. Its short generation time, small and fully sequenced genome, and wide global distribution range make Arabidopsis especially suited for the analysis of quantitative traits. Because of its natural self-pollination, recombinant inbred lines (RILs) have become the mapping population of choice in Arabidopsis. However, in other species near-isogenic lines (NILs) are favorable due to intolerance toward inbreeding and fertility issues. NILs are also useful for the confirmation and fine-mapping of QTLs identified in RIL populations. Chapter two describes the development of a genome-wide coverage NIL population from a cross between the distinct accessions Landsberg *erecta* (L*er*) and Cape Verde Islands (Cvi), for which a RIL population was developed previously. The genetic make-up of these two types of populations differs in the number of introgressions that segregate in the population. RILs contain multiple introgressions, whereas NILs preferably contain only a single introgression per line. As a consequence, in contrast to NIL populations, epistatic interactions between loci can be detected in RIL populations. Furthermore, owing to the higher recombination frequency, fewer replications per line need to be analyzed in RIL populations. However, the simultaneous segregation of multiple QTLs diminishes the power to detect small-effect QTLs in RIL populations compared to NIL populations.

The segregation of phenotypic variation can be observed for numerous traits, including quantitative ones, in populations derived from crosses between Arabidopis accessions. However, quantitative traits are often the resultant of many intermediary steps from genotype to phenotype. To fully understand the complex regulatory circuitry of quantitative traits it is therefore recommendable to analyze genetic regulation at different levels of the biological information flow. The recent advances in 'omic' technologies now make the large scale analysis of gene

expression (transcriptomics), protein content (proteomics), and metabolite content (metabolomics) feasible.

The expression of genes often determines the onset of biological pathways and it is conceivable that variation in expression levels is reflected in phenotypic variation. The genetic analysis of genome-wide gene expression variation in the L*er* x Cvi RIL population in chapter three revealed high heritabilities for many genes, indicating that their expression is under genetic control. Indeed for a substantial number of genes expression QTLs (eQTLs) could be detected. In depth analysis uncovered both *cis*-regulated expression, resulting from polymorphisms in the gene itself, and *trans*-regulated expression, resulting from genetic differences in distant regulators. Identifying *trans*-regulators offers the possibility to determine gene-to-gene regulation and ultimately the construction of regulatory networks. For a number of genomic regions, unexpectedly high numbers of *trans*-eQTLs were detected. Such hot spots are possibly caused by pleiotropic effects of regulators (*e.g.* transcription factors). When multiple genes, involved in the same biological process, map to the same position this indicates that many of them might be regulated by the same gene. This information was successfully used to demonstrate the construction of a regulatory network for flowering time.

On the other end of the information chain, metabolites stand closest to physiological phenotypes. It is therefore likely that genetic variation, leading to physiological differences, is also causal for differences in metabolite content. The untargeted metabolic analyses described in chapter four uncovered extensive natural variation in metabolite composition in 14 different accessions of Arabidopsis. QTL analysis of more than 2,000 high quality mass peaks, detected in the L*er* x Cvi RIL population, enabled the identification of QTLs for about 75% of the mass signals. The finding that more than one-third of the mass signals, detected in the RILs, were not detected in either parent suggests that many metabolites are formed due to the recombination of the parental genomes. The identification of anonymous mass peaks, that appear to be co-regulated as based on the positions of QTLs, enabled the (re)construction of metabolic pathways and uncovered novel biosynthetic steps and compounds in Arabidopsis. These results indicate the large potential for modification of metabolic composition through classical breeding.

Although each of the different entities in the path from genotype to phenotype can be effectively analyzed, a thorough understanding of the interaction between these different levels can only be obtained from the integrated study of multi-parallel analyses. In chapter five, the complex regulation of primary carbohydrate metabolism was analyzed in a case study. The activities of 15 enzymes involved in carbohydrate metabolism, in parallel with the expression of their structural genes, and contents of their metabolic substrates and products,

were genetically analyzed in the L*er* x Cvi RIL population. For many enzymes QTLs explaining variation observed in their activity were detected. A number of these QTLs co-located with the position of structural genes, indicating that natural variation in structural genes can be causal for variation in enzyme activity. From the expression analyses of these structural genes it was concluded that both expression variation and variation in protein function determine the differences in observed enzyme activity. However, not all enzyme activity QTLs co-located with structural genes, suggesting that regulation occurs at multiple levels. To further complicate the regulation of carbohydrate metabolism, significant correlations between enzyme activities and metabolite contents were observed, although this was not always accompanied by co-locating QTLs. Further analysis suggested a relationship between the regulation of carbohydrate metabolism and plant development.

The results of this thesis demonstrate the power of combining genetic approaches with large-scale high-throughput technologies for the construction of genetic regulatory networks and metabolic pathways. The integration of multi-parallel analyses will further enhance our understanding of the complex circuitry of genetic regulation of quantitative traits.

Samenvatting

Genetische verschillen tussen accessies van planten openbaren zich vaak als fenotypische variatie. Deze natuurlijke variatie vertoont in veel gevallen een continue verdeling en wordt daarom ook wel kwantitatieve variatie genoemd. Kwantitatieve variatie is het gevolg van het samenspel van meerdere genen en de invloed van omgevingsfactoren. Omdat de bijdrage van ieder gen aan het uiteindelijke fenotype erg klein kan zijn, zijn geavanceerde statistische methodes nodig om een associatie van genomische regio's met een bepaalde eigenschap aan te tonen. Een dergelijke aanpak staat bekend als quantitative trait locus (QTL) analyse. In QTL analyses wordt de gewenste eigenschap gekwantificeerd in een genetische kartering populatie, welke verkregen is door een kruising van verschillende genotypes.

Arabidopsis thaliana (Zandraket) is de meest gebruikte modelplant in moderne plantwetenschappen. Door de combinatie van een korte levenscyclus, een klein en volledig opgehelderd genoom en een wijde verspreiding over de wereld is Arabidopsis bij uitstek geschikt voor de genetische analyse van kwantitatieve eigenschappen. Omdat het een zelfbevruchter is, zijn recombinante inteelt lijnen (Recombinant Inbred Lines; RILs) het meest gangbaar als genetische kartering populatie in Arabidopsis. In andere soorten zijn bijna-isogene lijnen (Near-Isogenic Lines; NILs) echter beter bruikbaar door inteelt en vruchtbaarheidsproblemen in RILs. NILs zijn ook erg nuttig voor het bevestigen en de precieze positionering van QTLs die in RIL populaties gevonden zijn. Hoofdstuk twee beschrijft de ontwikkeling van een volledig genoomdekkende NIL populatie verkregen uit een kruising tussen de verschillende accessies Landsberg *erecta* (L*er*) en Cape Verde Islands (Cvi). Uit deze kruising was eerder al een RIL populatie ontwikkeld. De genetische opmaak van deze twee types populatie verschilt in het aantal introgressies. RILs bevatten meerdere introgressies terwijl NILs bij voorkeur slechts een enkele introgressie per lijn bevatten. Hierdoor kunnen in RIL populaties, in tegenstelling tot NIL populaties, epistatische interacties aangetoond worden. Bovendien kunnen er minder herhalingen per lijn geanalyseerd worden in RIL populaties omdat de recombinatiefrequentie hoger is dan in NIL populaties. Echter, de kans op het detecteren van QTLs met een klein effect is kleiner in RIL populaties, vergeleken met NIL populaties, omdat meerdere QTLs tegelijkertijd uitsplitsen.

Voor vele eigenschappen, inclusief kwantitatieve, kan uitsplitsing van fenotypische variatie worden waargenomen in populaties verkregen uit kruisingen met verschillende Arabidopsis accessies. Kwantitatieve eigenschappen zijn echter

vaak het gevolg van vele tussenliggende stappen op het traject van genotype naar fenotype. Om een volledig beeld te krijgen van de complexe reguleringscircuits van kwantitatieve eigenschappen is het daarom aan te bevelen om de genetische regulatie op verschillende niveaus van de biologische informatiestroom te analyseren. De recente voortgang in zogenaamde 'omic' technologieën maakt het momenteel mogelijk om de expressie van genen (transcription; transcriptomics) en de aanwezigheid van eiwitten (proteins; proteomics) en inhoudstoffen (metabolites; metabolomics) op grote schaal te analyseren.

De expressie van genen bepaalt vaak het begin van biologische routes en het is aannemelijk dat variatie in expressie niveaus zijn weerslag heeft op fenotypische variatie. De genetische analyse van genexpressie variatie van het complete genoom in de L*er* x Cvi RIL populatie in hoofdstuk drie toonde aan dat voor vele genen de gevonden variatie erfelijk is. Dit wijst er op dat de expressie van deze genen genetisch gereguleerd is. Voor een groot aantal genen werden inderdaad expressie QTLs (eQTLs) gevonden. Gedetailleerde analyses toonden zowel *cis*-gereguleerde expressie, als gevolg van polymorfismen in het gen zelf, als ook *trans*-gereguleerde expressie, als gevolg van genetische verschillen in regulatoren elders op het genoom, aan. De identificatie van *trans*-regulatoren biedt de mogelijkheid om gen-tot-gen regulatie aan te tonen en uiteindelijk om regulatienetwerken te construeren. Voor een aantal genomische regio's werden onverwacht hoge aantallen *trans*-eQTLs gevonden. Deze hot spots worden mogelijk veroorzaakt door pleiotrope effecten van regulatoren (b.v. transcriptie factoren). Als voor meerdere genen, die ieder bij hetzelfde biologische proces betrokken zijn, op dezelfde positie een eQTL gevonden wordt dan wijst dit er op dat vele wellicht door hetzelfde gen gereguleerd worden. Deze informatie werd succesvol aangewend om de constructie van een regulatie netwerk voor bloeitijd te demonstreren.

Aan het andere eind van de informatieketen staan metabolieten het dichtst bij het uiteindelijke fysiologische fenotype. Het is daarom waarschijnlijk dat genetische variatie, leidend tot fysiologische verschillen, ook de oorzaak is van verschillen in metaboliet niveaus. De ongerichte metaboliet analyses zoals beschreven in hoofdstuk vier toonden de uitgebreide natuurlijke variatie in metaboliet samenstelling in 14 verschillende accessies van Arabidopsis aan. QTL analyses van meer dan 2.000 kwalitatief betrouwbare massapieken, gedetecteerd in de L*er* x Cvi RIL populatie, resulteerde in de identificatie van QTLs voor ongeveer 75% van de massapieken. Meer dan een-derde van de massapieken die in de RILs konden worden gedetecteerd werden in geen van de ouders aangetroffen. Dit suggereert dat vele metabolieten werden gevormd door de recombinatie van de genomen van de ouders. De identificatie van massapieken, die op basis van QTL posities identiek gereguleerd lijken te zijn, maakte het mogelijk om metabole routes te (re)construeren en om nieuwe biosynthetische stappen en metabolieten in Arabidopsis aan te tonen. Deze resultaten geven de hoge potentie voor het modificeren van metaboliet samenstelling door klassieke veredeling weer.

Hoewel ieder van de verschillende stappen in het traject van genotype naar fenotype effectief geanalyseerd kan worden, kan een volledig begrip van de interactie tussen verschillende niveaus alleen verkregen worden door geïntegreerde studies van multi-parallelle analyses. In hoofdstuk vijf werd de complexe regulatie van het primaire koolhydraatmetabolisme geanalyseerd in een modelstudie. De activiteiten van 15 enzymen die allen een rol spelen in dit metabolisme werden parallel met de expressie van hun structurele genen en accumulatie van hun metabole substraten en producten genetisch geanalyseerd in de L*er* x Cvi RIL populatie. Voor veel enzymen konden QTLs worden gevonden die de variatie in activiteit verklaarden. Een deel van deze QTLs werd gevonden op posities van structurele genen. Dit wijst er op dat natuurlijke variatie in structurele genen de oorzaak kan zijn van de variatie in enzymactiviteit. Na expressieanalyses van deze structurele genen kon geconcludeerd worden dat zowel expressievariatie als variatie in enzymfunctie de verschillen in enzymactiviteit bepalen. Niet alle enzymactiviteit QTLs werden echter op posities van structurele genen gevonden, wat suggereert dat regulatie op meerdere niveaus plaats vindt. De complexe regulatie van het hydraatmetabolisme werd verder geïllustreerd door de significante correlaties tussen enzymactiviteiten en metaboliet accumulaties zonder dat er sprake was van QTLs op identieke posities. Statistische analyses suggereerden een relatie tussen koolhydraatmetabolisme en plantontwikkeling als een mogelijke oorzaak van correlaties.

De resultaten van dit proefschrift tonen de kracht aan van het combineren van genetische analyses met grootschalige ‘omics` technologieën om genetische regulatienetwerken en metabole routes te construeren. De integratie van multi-parallelle analyses zal ons begrip van de complexe circuits van genetische regulatie van kwantitatieve eigenschappen verder vergroten.

Publications

Publications from this thesis

Keurentjes, J.J.B., J. Fu, I.R. Terpstra, J.M. Garcia, G. van den Ackerveken, L.B. Snoek, A.J.M. Peeters, D. Vreugdenhil, M. Koornneef and R.C. Jansen. Regulatory network construction in Arabidopsis using genome-wide gene expression QTLs. Proc Natl Acad Sci U S A. 104: 1708-13. 2007.

Keurentjes, J.J.B., L. Bentsink, C. Alonso-Blanco, C.J. Hanhart, H. Blankestijn-De Vries, S. Effgen, D. Vreugdenhil and M. Koornneef. 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. 2007.

Keurentjes, J.J.B., J. Fu, C.H. de Vos, A. Lommen, R.D. Hall, R.J. Bino, L.H.W. van der Plas, R.C. Jansen, D. Vreugdenhil and M. Koornneef. The genetics of plant metabolism. Nat Genet. 38: 842-9. 2006.

Fu, J., M.A. Swertz, **J.J.B. Keurentjes** and R.C. Jansen. MetaNetwork: a computational protocol for the genetic study of metabolic networks. Nat protoc. 2: 685-94. 2007.

De Vos, C.H., S. Moco, A. Lommen, **J.J.B. Keurentjes**, R.J. Bino and R.D. Hall. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. Nat protoc. 2: 778-91. 2007.

Keurentjes, J.J.B., M. Koornneef and D. Vreugdenhil. Genome studies and molecular genetics. In preparation for Curr Opin Plant Biol.

Keurentjes, J.J.B., R. Sulpice, Y. Gibon, J. Fu, M. Koornneef, M. Stitt and D. Vreugdenhil. Integrative analyses of genetic variation in enzyme activities of primary carbohydrate metabolism reveal distinct modes of regulation in Arabidopsis thaliana. In preparation for Genome Biol.

Fu J., M. Dijkstra, **J.J.B. Keurentjes**, M. Koornneef, R. Breitling and R.C. Jansen. Systems biology through system genetics. In preparation.

Publications related to this thesis

Sergeeva, L.I., **J.J.B. Keurentjes**, L. Bentsink, J. Vonk, L.H.W. van der Plas, M. Koornneef and D. Vreugdenhil. Vacuolar invertase regulates elongation of Arabidopsis thaliana roots as revealed by QTL and mutant analysis. Proc Natl Acad Sci U S A. 103: 2994-9. 2006.

Teng, S., **J.J.B. Keurentjes**, L. Bentsink, M. Koornneef and S. Smeekens. Sucrose-specific induction of anthocyanin biosynthesis in Arabidopsis requires the MYB75/PAP1 gene. Plant Physiol. 139: 1840-52. 2005.

Juenger, T.E., J.K. mcKay, N. Hausmann, **J.J.B. Keurentjes**, S. Sen, K.A. Stowe, T.E. Dawson, E.L. Simms and J.H. Richards. Identification and characterization of QTL underlying whole-plant physiology in Arabidopsis thaliana: δ13C, stomatal conductance and transpiration efficiency. Plant Cell Env. 28: 697-702. 2005.

Sergeeva, L.I, J. Vonk, **J.J.B. Keurentjes**, L.H.W. van der Plas, M. Koornneef and D. Vreugdenhil. Histochemical analysis reveals organ-specific quantitative trait loci for enzyme activities in Arabidopsis. Plant Physiol. 134: 237-45. 2004.

Sicard O., O. Loudet, **J.J.B. Keurentjes**, T. Candresse, O. Le Gall, F. Revers and V. Decroocq. Identification of QTLs controlling symptom development during viral infection in Arabidopsis thaliana. Submitted to Plant Physiol.

Tessadori F., M. van Zanten, P. Pavlova, B.L. Snoek, F.F. Millenaar, R.K. Schulkes, **J.J.B. Keurentjes**, R. van Driel, L.A.C.J. Voesenek, P. Fransz and A.J.M. Peeters. Natural variation in heterochromatin content among Arabidopsis accessions is controlled by light-intensity. In preparation for Plant cell.

Other publications

Léon-Kloosterziel, K.M., B.W.M. Verhagen, **J.J.B. Keurentjes**, J.A. Van Pelt, M. Rep, L.C. Van Loon and C.M.J. Pieterse. Colonization of the Arabidopsis rhizosphere by fluorescent Pseudomonas spp. activates a root-specific, ethylene-responsive PR-5 gene in the vascular bundle. Plant Mol. Biol. 57: 731-48. 2005.

De Boer, M., P. Bom, F. Kindt, **J.J.B. Keurentjes**, I. Van der Sluis, L.C. Van Loon, and P.A.H.M. Bakker. Control of Fusarium wilt of radish by combining Pseudomonas putida strains that have different disease-suppressive mechanisms. Phytopathology. 93: 626-32. 2003.

Léon-Kloosterziel, K.M., B.W.M. Verhagen, **J.J.B. Keurentjes**, L.C. Van Loon and C.M.J. Pieterse. Identification of genes involved in rhizobacteria-mediated induced systemic resistance in Arabidopsis. In: Induced Resistance in Plants Against Insects and Diseases (A. Schmitt and B. Mauch-Mani, eds), IOBC/wprs Bulletin. 25: 71-4. 2002.

Pieterse, C.M.J., J.A. van Pelt, S.C.M. van Wees, J. Ton, K.M. Léon-Kloosterziel, **J.J.B. Keurentjes**, B.W.M. Verhagen, M. Knoester, I. Van der Sluis, P.A.H.M. Bakker and L.C. van Loon. Rhizobacteria-mediated induced systemic resistance: triggering, signalling, and expression. Eur. J. Plant Pathol. 107: 51-61. 2001.

De Boer, M., I. van der Sluis, **J.J.B. Keurentjes**, L.C. van Loon and P.A.H.M. Bakker. Modes of action of suppression of fusarium wilt of radish by the combination of Pseudomonas putida RE8 and P. fluorescens RS111. In: Proceedings of the Fifth International Plant-Growth Promoting Rhizobacteria Workshop. Cordoba, Argentina. 2000.

De Boer, M., I. Van der Sluis, **J.J.B. Keurentjes**, L.C. van Loon and P.A.H.M. Bakker. Verbetering van biologische beheersing van Fusarium-verwelkingsziekte in radijs door het gebruik van combinaties van Pseudomonas-stammen. Gewasbescherming. 31: 56-7. 2000.

Pieterse, C.M.J., S.C.M. van Wees, J. Ton, K.M. Léon-Kloosterziel, J.A. van Pelt, **J.J.B. Keurentjes**, M. Knoester and L.C. van Loon. Rhizobacteria-mediated induced systemic resistance (ISR) in Arabidopsis: involvement of jasmonate and ethylene. In: Biology of Plant-Microbe Interactions, Volume 2 (P.J.G.M. De Wit, T. Bisseling and W.J. Stiekema, eds), The International Society for Molecular Plant-Microbe Interactions, St. Paul, MN., 291-6, 1999.

De Boer, M., P. Bom, F. Kindt, I. Van der Sluis, **J.J.B. Keurentjes**, L.C. van Loon and P.A.H.M Bakker. Het gebruik van de combinatie van Pseudomonas putida stammen RE8 en WCS358 kan biologische bestrijding van Fusarium verwelkingsziekte in radijs verbeteren. Gewasbescherming. 30: 85-6. 1999.

Falque, M., **J.J.B. Keurentjes**, J.M.T. Bakx-Schotman and P.J. van Dijk. Development and characterisation of microsatellite markers in the sexual-apomictic complex Taraxacum officinale (dandelion). Theor. Appl. Genet. 97: 283-92. 1998.

Gorissen A., J.H. van Ginkel, **J.J.B. Keurentjes** and J.A. van Veen. Grass root decomposition is retarded when grass has been grown under elevated CO2. Soil Biol. Biochem. 27: 117-20. 1995.

Bonnier, F.J.M., **J.J.B. Keurentjes** and J.M. van Tuyl. Ion leakage as a criterion for viability of lily bulb scales after storage at -2°C for 0.5, 1.5 and 2.5 years. Hort Science. 29: 1332-4. 1994.

Honours, Awards and Fellowships

CBSG Special Achievement Award. 2005.

ZonMw, National Genomics Initiative. International research fellowship. 2006.

Curriculum vitae

Joost Keurentjes werd geboren op 25 december 1968 te Doetinchem. Na het behalen van het HAVO diploma aan het ISALA college te Silvolde in 1988 voltooide hij het MLO te Arnhem in 1992. Aansluitend was hij twee jaar werkzaam als onderzoeksmedewerker aan het AB-DLO alvorens in 1994 de studie plantenbiotechnologie aan de IAHL te Velp te beginnen. In 1997 werd deze studie afgerond en volgden er enkele dienstverbanden als onderzoeksmedewerker bij het IPO-DLO, de universiteit Utrecht (Fytopathologie) en als senior onderzoeker bij Hercules b.v.. In 2002 begon hij aan het promotieonderzoek beschreven in dit proefschrift bij de leerstoelgroepen Erfelijkheidsleer en Plantenfysiologie aan de Wageningen Universiteit. Met ingang van 1 mei 2007 is hij in dienst als post-doc onderzoeker bij voornoemde leerstoelgroepen.

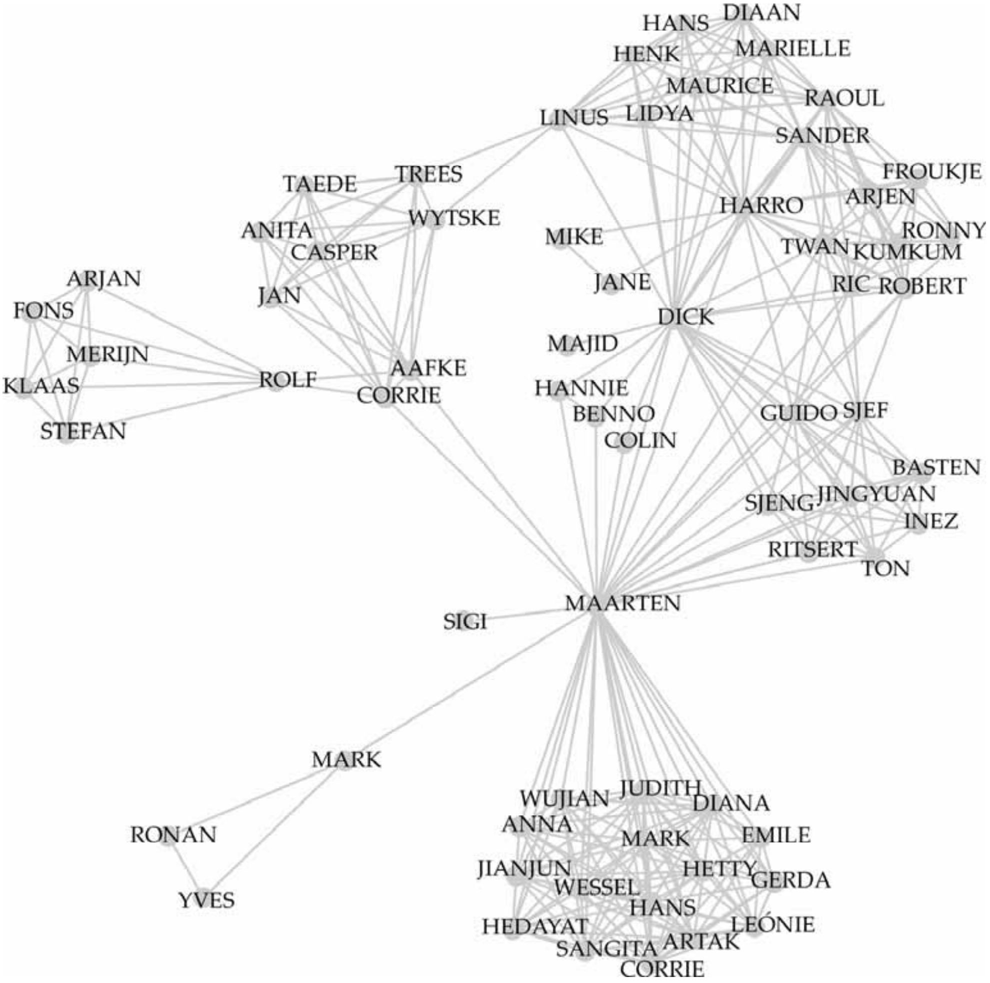
Nawoord

Op het moment dat u dit leest hoop ik dat u ook de moeite heeft genomen, of nog zult nemen, om de voorgaande hoofdstukken door te nemen. De inhoud van dit proefschrift is namelijk met zorg en toewijding samengesteld en gelukkig niet alleen door mij. Het vermelden van een ieder die heeft bijgedragen aan de totstandkoming ervan zou slechts leiden tot een droge opsomming. Het staat echter buiten kijf dat dit boekje er heel anders uit had gezien zonder de hulp van velen. Hoewel de inbreng van de één misschien omvangrijker of zinvoller is geweest dan van de ander wil ik toch geen onderscheid maken in waardering. Ik prijs mij gelukkig om deel uit te hebben mogen maken van een omvangrijk netwerk van specialisten van wier expertise ik dankbaar gebruik heb gemaakt. Ik heb geprobeerd de complexiteit van dit netwerk weer te geven in de figuur op de volgende pagina. Kenners zullen onmiddellijk opmerken dat het een topologisch robuust, modulair en schaalvrij hierarchisch netwerk is met een hoge graad van connectiviteit. In de praktijk staat dit synonym voor een kwalitatief hoogwaardig samenwerkingsverband met korte lijnen tussen de deelnemers, het zogenaamde 'kleine-wereld effect' (iedereen kent wel iemand die iemand anders kent).

Toch wil ik er graag een aantal personen uitlichten die in de achterliggende jaren bijzonder veel voor mij betekend hebben. In de eerste plaats mijn promotoren Maarten en Linus en co-promotor Dick. Deze synergistische drie-eenheid heeft mij vrijwel probleemloos door mijn promotietraject geloodst. Een speciaal woord van dank ook aan mijn twee paranimfen. Jingyuan, zonder wier hulp en tomeloze inzet een groot deel van dit proefschrift niet tot stand was gekomen. Judith, met wie, als mede-AIO en kamergenoot, ik meer dan vier jaar opgetrokken ben. We hebben veel lief en leed gedeeld en gelukkig meer lief dan leed.

Rest mij nog te vermelden dat ik het allemaal met veel plezier volbracht heb. Het in mij gestelde vertrouwen om er nog eens vier jaar aan vast te plakken verheugt mij dan ook zeer.

Joost


HANS
DIAAN
HENK
MARIELLE
MAURICE
RAOUL
LINUS
LIDYA
SANDER
TAEDE
TREES
FROUKJE
HARRO
ARJEN
WYTSKE
ANITA
CASPER
MIKE
RONNY
TWAN
KUMKUM
ARJAN
JANE
RIC
ROBERT
FONS
JAN
DICK
MERIJN
MAJID
KLAAS
ROLF
AAFKE
CORRIE
HANNIE
BENNO
GUIDO
SJEF
STEFAN
COLIN
BASTEN
SJENG
JINGYUAN
INEZ
RITSERT
TON
SIGI
MAARTEN
MARK
WUJIAN
JUDITH
DIANA
ANNA
MARK
EMILE
RONAN
JIANJUN
WESSEL
HETTY
GERDA
YVES
HANS
HEDAYAT
LEÓNIE
SANGITA
ARTAK
CORRIE

Education Statement of the Graduate School Experimental Plant Sciences

The Graduate School EXPERIMENTAL PLANT SCIENCES

Issued to: **Joost J. B. Keurentjes**
Date: **7 September 2007**
Group: **Laboratories of Plant Physiology and Genetics, Wageningen University**

1) Start-up phase	*date*
► **First presentation of your project**	
Using natural variation for dissecting pathways of plant performance traits	Apr 07, 2003
► **Writing or rewriting a project proposal**	
► **Writing a review or book chapter**	
► **MSc courses**	
► **Laboratory use of isotopes**	
Subtotal Start-up Phase	*1.5 credits*[*]

2) Scientific Exposure	*date*
► **EPS PhD student days**	
EPS PhD student day, Utrecht University	Mar 27, 2003
EPS PhD student day, Vrije Universiteit Amsterdam	Jun 03, 2004
EPS PhD student day, Radboud University Nijmegen	Jun 02, 2005
EPS PhD student day, Wageningen University	Sep 19, 2006
► **EPS theme symposia**	
EPS Theme 3 symposium 'Metabolism and Adaptation', Wageningen university	Mar 23, 2003
EPS Theme 3 symposium 'Metabolism and Adaptation', Wageningen university	Oct 25, 2004
EPS Theme 3 symposium 'Metabolism and Adaptation', Utrecht university	Nov 24, 2005
► **NWO Lunteren days and other National Platforms**	
ALW meeting Experimental Plant Sciences, Lunteren	Apr 07-08, 2003
ALW meeting Experimental Plant Sciences, Lunteren	Apr 05-06, 2004
ALW meeting Experimental Plant Sciences, Lunteren	Apr 04-05, 2005
ALW meeting Experimental Plant Sciences, Lunteren	Apr 02-03, 2007
► **Seminars (series), workshops and symposia**	
Frontiers in Plant Science, Wageningen university	2003
Flying seminars, Wageningen university	2003-2007
CBSG Cluster meeting Arabidopsis, Wageningen (3x)	2004-2005
CBSG Summit, Wageningen (2x)	2005 & 2007
15th symposium ALW-Discussion Group "Secondary Metabolism in Plant and Plant Cell", Zeist	May 20, 2005

16th symposium ALW-Discussion Group “Secondary Metabolism in Plant and Plant Cell”, Leiden	Oct 6, 2006
Netherlands BioInformatics Centre Workshop Bioinformatics for Metabolomics, Wageningen	Nov 29, 2006
► **Seminar plus**	
► **International symposia and congresses**	
7th International Congress of Plant Molecular Biology (ISPMB), Barcelona, Spain	Jun 23-28, 2003
Keystone symposia, Biological discovery using diverse high-throughput data, Steamboat Springs, USA	Mar 30-Apr 4, 2004
16th International Conference on Arabidopsis Research, Madison, USA	Jun 15-19, 2005
4th Plant Genomics European Meetings, Amsterdam, The Netherlands	Sep 20-23, 2005
15th Crucifer Genetics Workshop: Brassica 2006, Wageningen, The Netherlands	Sep 30–Oct 4, 2006
5th Plant Genomics European Meetings, Venice, Italy	Oct 11-14, 2006
18th International Conference on Arabidopsis Research, Beijing, China	Jun 20-23, 2007
► **Oral presentations**	
7th International Congress of Plant Molecular Biology (ISPMB), Barcelona, Spain	Jun 24, 2003
Heidelberg Institute of Plant Science, Heidelberg University, Heidelberg, Germany	Sep, 2003
CBSG Cluster meeting Arabidopsis, Wageningen (3x)	2004-2005
EPS Theme 3 symposium Metabolism and Adaptation, Wageningen university	Oct 25, 2004
Plant Research International, Bioscience, Plant Development Systems, Wageningen	Nov, 2004
Plant Research International, Bioscience, Metabolic Regulation, Wageningen	Dec, 2004
ALW meeting Experimental Plant Sciences, Lunteren	Apr 5, 2005
EPS/VLAG/CBSG Workshop Metabolomics, Wageningen University	May 3, 2005
16th International Conference on Arabidopsis Research, Madison, USA	Jun 17, 2005
4th Plant Genomics European Meetings, Amsterdam, The Netherlands	Sep 23, 2005
EPS Theme 3 symposium Metabolism and Adaptation, Utrecht university	Nov 24, 2005
Max-Planck-Institute of Molecular Plant Physiology, Golm, Germany	May 5, 2006
De Ruiter Seeds, Bergschenhoek	Jul 25, 2006
15th Crucifer Genetics Workshop: Brassica 2006, Wageningen, The Netherlands	Oct 4, 2006
16th symposium ALW-Discussion Group “Secondary Metabolism in Plant and Plant Cell”, Leiden	Oct 6, 2006
5th Plant Genomics European Meetings, Venice, Italy	Oct 14, 2006
Netherlands BioInformatics Centre, Workshop Bioinformatics for Metabolomics, Wageningen	Nov 29, 2006
CBSG Summit, Wageningen	Feb 6, 2007
ALW meeting Experimental Plant Sciences, Lunteren	Apr 2-3, 2007

Utrecht Genetic Seminar Series, Hubrecht laboratory, Utrecht	Apr 12, 2007
Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China	Jun 20, 2007
18th International Conference on Arabidopsis Research, Beijing, China (2x)	Jun 21, 2007
PhD summerschool; Environmental signaling: Arabidopsis as a model, Utrecht University	Aug 27, 2007
► **IAB interview**	May, 2005
► **Excursions**	
Subtotal Scientific Exposure	37.2 credits*
3) In-Depth Studies	date
► **EPS courses or other PhD courses**	
International Summerschool, The analysis of natural variation within crop and model plants, Wageningen	Apr 22-25, 2003
EPS Summerschool, Functional Genomics: theory and hands-on data analysis, Utrecht university	Aug 25-28, 2003
EPS/VLAG/CBSG Workshop Metabolomics, Wageningen University	May 2-4, 2005
ABIES/PE&RC/EPS/SdV Workshop Mathematics in Plant Biology, Paris, France	Jun 30-Jul 1, 2005
PhD summerschool; Environmental signaling: Arabidopsis as a model, Utrecht University	Aug 27-29, 2007
► **Journal club**	
member of literature discussion group	2002-2003
► **Individual research training**	
Netherlands Genomics Initiative fellowship, MPI for Molecular Plant Physiology, Golm Germany	feb 1-may 5, 2006
Subtotal In-Depth Studies	9.6 credits*
4) Personal development	date
► **Skill training courses**	
► **Organisation of PhD students day, course or conference**	
Wageningen International, Training programme on the conservation, management and use of plant genetic resources in agriculture; Biotechnology for genetic resources conservation and crop improvement	May 9-Jul 1, 2005
► **Membership of Board, Committee or PhD council**	
Member of PhD council	2003-2006
Subtotal Personal Development	2.9 credits*
TOTAL NUMBER OF CREDIT POINTS*	**51.2**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises a minimum total of 30 credits.

* *A credit represents a normative study load of 28 hours of study*

The research described in this thesis was performed in the project 'QTL express: identification of plant performance traits in Arabidopsis by combining high-throughput mapping and expression profiling', financially supported by a grant from the Netherlands Organization for Scientific Research, Program Genomics (050-10-029).

Printed by: Ponsen & Looijen BV, Wageningen, The Netherlands.