nature genetics

The genetics of plant metabolism

Joost J B Keurentjes[1,2,8], Jingyuan Fu[3,8], C H Ric de Vos[4,5,8], Arjen Lommen[4–6], Robert D Hall[4,5], Raoul J Bino[2,4,5], Linus H W van der Plas[2], Ritsert C Jansen[3], Dick Vreugdenhil[2] & Maarten Koornneef[1,7]

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.

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[1]. Moreover, substantial quantitative and qualitative variation in metabolite composition is often observed within plant species[2].

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, which can affect metabolites[3], in segregating populations can identify loci explaining the observed variation[4]. 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[5–9]. 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[10,11]. QTL mapping will result in the localization of loci, and ultimately genes, causal for the observed variation and will allow the discovery of coregulated compounds. In this way, genome-wide genetic correlative metabolic analysis now becomes feasible, as we demonstrate here.

RESULTS

Metabolite variation is abundant and genetically controlled

To assess the natural variation in metabolite content present in *A. thaliana*, 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 *A. thaliana* (**Supplementary Table 1** online). We observed considerable quantitative and qualitative variation 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, we detected 964 mass peaks 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 (**Fig. 1a**). We found an average of 50 unique mass peaks per accession, 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 (**Fig. 1b**). This, together with the substantial

[1]Laboratory of Genetics, Wageningen University, Arboretumlaan 4, NL-6703 BD Wageningen, The Netherlands. [2]Laboratory of Plant Physiology, Wageningen University, Arboretumlaan 4, NL-6703 BD Wageningen, The Netherlands. [3]Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Kerklaan 30, NL-9751 NN Haren, The Netherlands. [4]Plant Research International, Droevendaalsesteeg 1, NL-6708 PB Wageningen, The Netherlands. [5]Centre for Biosystems Genomics, Droevendaalsesteeg 1, NL-6708 PB Wageningen, The Netherlands. [6]RIKILT–Institute of Food Safety, Bornsesteeg 45, NL-6700 AE Wageningen, The Netherlands. [7]Max Planck Institute for Plant Breeding Research, Carl von Linné weg 10, 50829 Cologne, Germany. [8]These authors contributed equally to this work. Correspondence should be addressed to M.K. (maarten.koornneef@wur.nl).

Received 10 March; accepted 5 May; published online 4 June 2006; doi:10.1038/ng1815

variation in metabolite composition observed within a single plant species promises great opportunities for metabolic engineering by classical breeding[12].

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)[13]. These were the two biochemically most distinct accessions for which such a mapping population was available (**Supplementary Methods** and **Supplementary Fig. 1** online). We found it striking that 853 of a total of 2,129 mass peaks identified in the RIL population were not detected in either parent (**Fig. 2a**). Although the number of lines analyzed in the RIL population (160 lines measured in duplicate) exceeded that of the number of parental lines (five 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. For 1,592 mass signals (74.8%), we detected at least one significant ($P < 0.0001$) QTL using a two-part parametric model[14]. This P threshold corresponded to a q value of 0.0002 in Storey's genome-wide false discovery rate (FDR) method[15]. On average, we found nearly 2.0 QTLs per analyzed mass, leading to a total of 4,213 QTLs (**Supplementary Fig. 2** online). Thus, after crossing these two distinct genotypes, variation in the presence and abundance of ~75% of the detected masses in their offspring could be at least partly explained by mappable genetic factors (**Fig. 2a**), consistent with the relatively high heritabilities found for many masses (**Supplementary Fig. 3** online). 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.

Analysis of the genomic distribution of the detected QTLs shows that these are not evenly distributed over the *A. thaliana* genome. Instead, we observed hot and cold spots for the regulation of metabolic content (**Fig. 2b,c**). 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[16]. We therefore computed empirical confidence levels by permutation tests (**Supplementary Methods** online) and found that in most cases, the frequency of QTLs occurring at hot spots was much higher than was expected by chance (**Fig. 2c**).

Map positions can uncover metabolic pathways

Colocation of QTLs coincides with clusters of highly correlated mass peaks, which are assumed to be enriched for masses regulated by the same genes. Coregulated metabolites may indicate that a specific biological function controls different components or that a specific step in a biochemical pathway is affected[17]. 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 *A. thaliana*[18]. The largest class comprises the aliphatic glucosinolates, which are all derived from methionine (**Fig. 3a**). Studies targeted to this class of metabolites have shown large quantitative and qualitative differences in accumulation of aliphatic glucosinolates between *A. thaliana* accessions[19]. In addition, QTL analysis of these glucosinolates in the Cvi × L*er* RIL population uncovered two major loci explaining the observed variation for most aliphatic glucosinolates[7]. The *MAM* locus at the top of chromosome 5 is responsible for the observed variation in chain length[20], whereas the *AOP* locus at the top of chromosome 4 is responsible for the observed variation in side chain modification[21]. 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 (**Supplementary Note** online).

By making use of the mass accuracy of the TOF-MS, we were able to identify most of the aliphatic glucosinolates reported for *A. thaliana*. Subsequent QTL analysis showed that all masses corresponding to an aliphatic glucosinolate indeed mapped to the *AOP* and/or *MAM* loci (**Fig. 3b**), thus confirming previous findings. Epistatic analysis of the two loci uncovered strong interactions for many of the detected glucosinolates (**Supplementary Methods** and **Supplementary Table 2** online).

The fact that we did not detect all glucosinolate QTLs found in another study[7] 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 × environment or genotype × developmental stage interaction, which can be expected also for metabolites. Furthermore, we mapped individual glucosinolates, whereas the other study[7] mapped total aliphatic glucosinolate content.

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) (**Supplementary Note** online). We observed strong

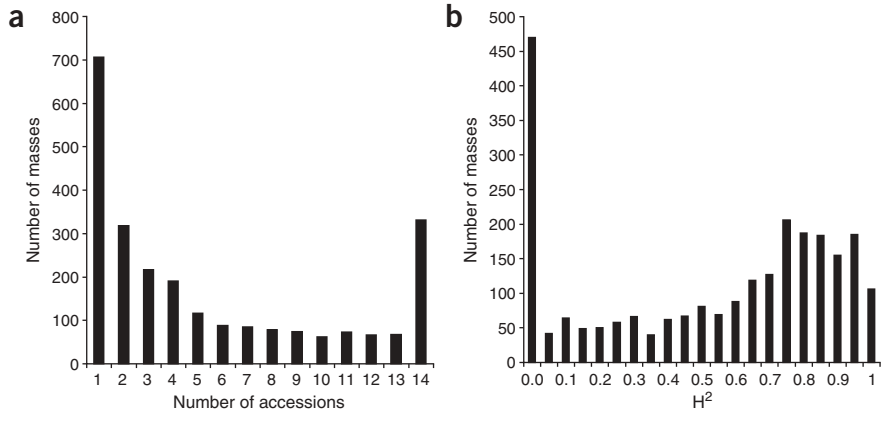

Figure 1 Natural variation in *A. thaliana* metabolite accumulation. (**a**) Frequency distribution of the number of different accessions each mass peak was detected in. (**b**) Frequency distribution of broad-sense heritability of each mass peak detected in the different accessions. Data are based on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

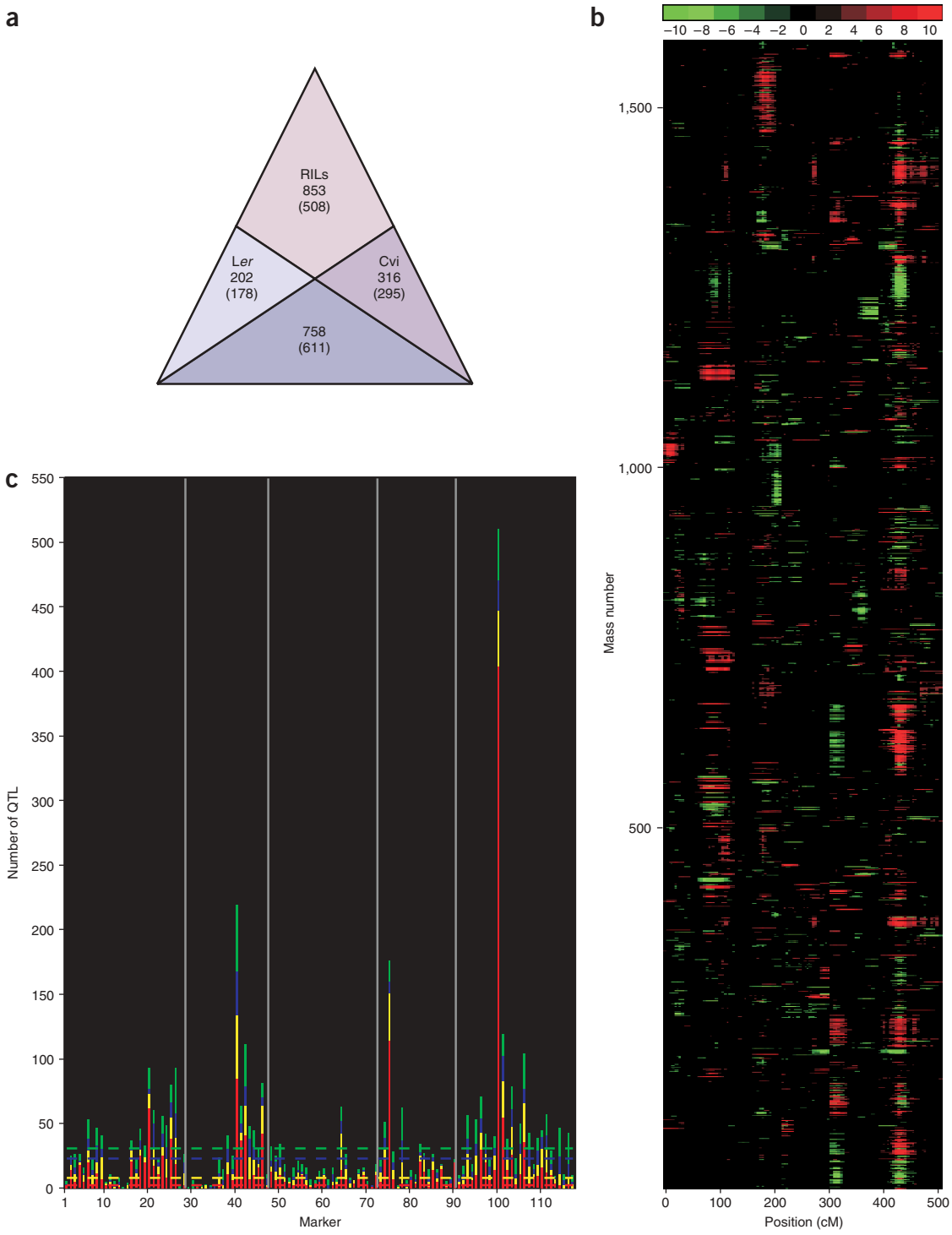

Figure 2 Genetic analysis of metabolite profiles in the *A. thaliana* Cvi × L*er* RIL population. (**a**) Number of masses detected in the RIL population and its parents. The triangle is subdivided into masses not detected in either parent (upper part), detected in one parent only (left and right) and 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. (**b**) Heat map of each mass in the RIL population for which at least one significant ($P < 0.0001$) QTL could be detected. Colors are according to their additive effects (red, Cvi; green, L*er*), and intensities represent significance of QTL likelihood ($-\log_{10}P$). (**c**) 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 are evenly spaced over the genome with an average distance of 5 cM between them. Chromosomal borders are indicated by vertical gray lines. The dashed 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} (green), 5×10^{-5} (blue), 10^{-5} (yellow) and 10^{-6} (red), respectively. Data represent two biological replicates per RIL and five biological replicates for each parent measured in two replicate extractions.

genetic correlations among aliphatic glucosinolates due to the colocation 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 covariance with any other pair[22]. We empirically estimated the significance threshold for the second-order correlations by permutation (**Supplementary Methods** online). Significant coefficients are shown in **Figure 3c** 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 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 (**Fig. 3b**) suggests that *AOP* acts downstream of *MAM*. Furthermore, we observed high levels of side chain–modified compounds in unexpected genotypic classes (**Supplementary Table 3** online). In

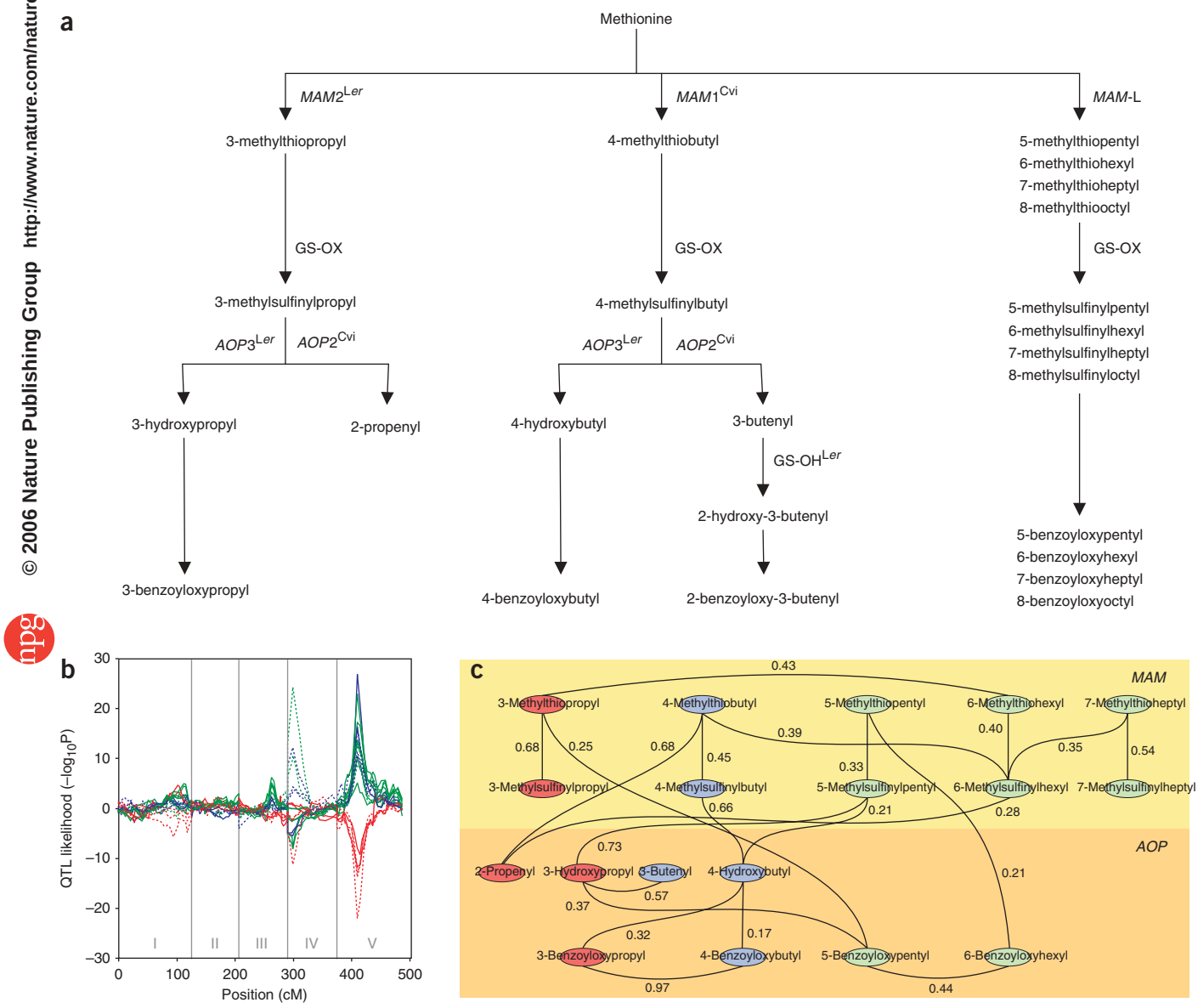

Figure 3 Genetic regulation of aliphatic glucosinolate accumulation in *A. thaliana*. (**a**) Scheme of aliphatic glucosinolate formation. Corresponding loci of enzymatic steps are shown in bold next to the arrows. (**b**) QTL likelihood profiles of aliphatic glucosinolates detected in the RIL population. 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 gray lines. (**c**) Second-order genetic correlations between aliphatic glucosinolates detected in the RIL population. Upper panel contains glucosinolates before side chain modification; 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. In **b** and **c**, colors represent different chain lengths (red, 3 C; blue, 4 C; green, >4 C).

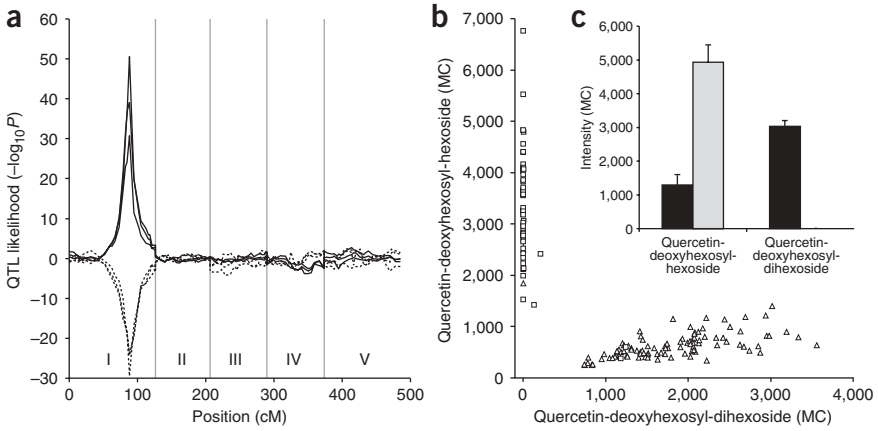

Figure 4 Genetic variation in flavonol-glycoside accumulation in *A. thaliana*. (**a**) QTL likelihood profiles of putatively identified flavonol glycosides in the RIL population. The sign of the value is related to the additive effect at each marker position (+, Cvi; –, L*er*). Dotted and solid lines represent flavonols with and without dihexosyl residues, respectively. Chromosomal borders are indicated by vertical gray 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 (gray). 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.

contrast to previous findings[21], 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 a previous study[21] also suggested alternative loci for 4-hydroxybutyl formation. These results validate our combined genetic and metabolomic approach to identify coregulated 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.

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 (**Supplementary Table 4** online). Most of the mass signals sharing this single QTL on chromosome 1 corresponded to different glycosylated flavonols (**Fig. 4a**). 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 1** and **Fig. 4b**,**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. 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 (that is, UGT79B10 and UGT79B11)[23]. UGT79B10 has been expressed as recombinant protein in *Escherichia coli*, but it showed no activity against quercetin glucosides in an *in vitro* analysis[24]. 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. Thus, the untargeted detection and subsequent mapping of metabolites enabled us to identify a number of putative flavonol-glycosides not previously reported in *A. thaliana*[25]. Colocation 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.

Table 1 Characteristics of putatively identified flavonols

Aglycone	Glycosylation	Significance (–log10*P*)	Effect (MC)	L*er* (MC ± s.e.m.)	Cvi (MC ± s.e.m.)
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	–1,326	1,334 ± 164	7 ± 0
Quercetin	Deoxyhexosyl-hexoside	50.7	2,659	1,293 ± 291	4,928 ± 517
Quercetin	Deoxyhexosyl-dihexoside	24.3	–1,721	3,031 ± 167	4 ± 0

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 –log10*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).

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 *A. thaliana* 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 coregulated 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 were 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[26,27]. 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[28]. 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 colocating QTLs can identify the genetic basis for these associations even when different experiments have been performed[29,30]. 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[31,32] can also readily be exploited, and this will aid further the construction of genetic regulatory networks[33]. 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[34]. 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 based 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.

METHODS

***A. thaliana* accessions and mapping population.** We analyzed 14 accessions of *A. thaliana* representing different regions of the global distribution of the species 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* (*Ler*) was used for QTL mapping of metabolite content. The F10 generation has been extensively genotyped[13] and is available from the *Arabidopsis* Biological Resource Center. All lines were advanced to the F13 generation, and residual heterozygous regions, estimated to be 0.71% in the F10 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, individual plants 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 d 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 h light (30 W m^{-2}) at 20 °C, 8 h dark at 15 °C and 75% relative humidity. After 6 d, 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 h into the light period, and all lines were harvested in random order within 2 h. 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 (20,000*g*) 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), using a Genesis Workstation (Tecan Systems). Samples were automatically injected (5 µl) and separated using an Alliance 2795 HT system (Waters) equipped with a Luna C_{18}-reversed phase column (150 × 2.1 mm, 3 µm; Phenomenex). Separation was performed at 40 °C by applying a 20 min gradient from 5–75% acetonitrile 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 ms and an interscan delay of 100 ms. Desolvation temperature was 250 °C with a nitrogen gas flow of 500 l/h, capillary spray was 2.75 kV, source temperature was 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 s, 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 was used for unbiased and unsupervised comparison of all LC-MS datasets[10,11]. In short, the program performs automated peak centering, local noise calculation, baseline correction and extraction of all relevant mass signals (that is, 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 data set. 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[11] (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).

Linkage map construction. Genotype data for the L*er*/Cvi population individuals are available at the web address listed below. The genetic map was constructed from a subset of the markers available with a few new markers added. The computer program JOINMAP 3.0 (ref. 35) was used for the calculation of linkage groups and genetic distances. Recombination frequencies were converted to distances in cM 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[14].

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 i^{th} 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: $p\{z = 1 \mid g = Ler\} = p\{z = 1 \mid g = Cvi\}$ versus the alternative hypothesis H_1: $p\{z = 1 \mid g = Ler\} \neq p\{z = 1 \mid g = Cvi\}$, 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: $u\{g = Ler\} = u\{g = Cvi\}$ versus the alternative hypothesis H_1: $u\{g = Ler\} \neq u\{g = Cvi\}$. The P value of the two-part model was then determined by the multiple of the P values from the two separate analyses ($P1$ and $P2$, respectively).

To calculate significance thresholds, we performed a simulation study following ref. 14. Each individual had probability 40% (the median proportion of null phenotype 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[15].

We then calculated the proportion of QTL significance explained by the binominal part by $\log P1/(\log P1 + \log P2)$, where $P1$ and $P2$ are the P values from the two separate parts of the model, respectively (**Supplementary Fig. 4** online). The variance explained by QTLs was calculated for both parts separately (**Supplementary Fig. 5** online). 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 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)[36]. 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[22,37,38] or from QTL profiles[39]. We combined and modified the methods from refs. 37 and 39 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 $\max(-\log_{10}P) \pm 1.5$; 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) Pairwise correlation coefficients between any two masses were then calculated as

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

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

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

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.

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

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 (**Supplementary Methods** online). 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 α level for the number of all pairwise analyses ($17 \times 18/2$) to avoid overcorrection. At this threshold, on average 0.1 correlation coefficients are significant by chance.

URLs. METALIGN is available at http://www.metAlign.nl. Genotype data for the L*er*/Cvi population individuals are available at http://nasc.nott.ac.uk/. JOINMAP is available at http://www.kyazma.nl.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

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).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Wink, M. Plant breeding: importance of plant secondary metabolites for protection against pathogens and herbivores. *Theor. Appl. Genet.* **75**, 225–233 (1988).
2. Windsor, A.J. *et al.* Geographic and evolutionary diversification of glucosinolates among near relatives of *Arabidopsis thaliana* (Brassicaceae). *Phytochemistry* **66**, 1321–1333 (2005).

3. Jansen, R.C. & Nap, J.P. Genetical genomics: the added value from segregation. *Trends Genet.* **17**, 388–391 (2001).
4. Jansen, R.C. Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211 (1993).
5. Bentsink, L. *et al.* Genetic analysis of seed-soluble oligosaccharides in relation to seed storability of *Arabidopsis*. *Plant Physiol.* **124**, 1595–1604 (2000).
6. Hobbs, D.H., Flintham, J.E. & Hills, M.J. Genetic control of storage oil synthesis in seeds of *Arabidopsis*. *Plant Physiol.* **136**, 3341–3349 (2004).
7. Kliebenstein, D.J., Gershenzon, J. & Mitchell-Olds, T. Comparative quantitative trait loci mapping of aliphatic, indolic and benzylic glucosinolate production in *Arabidopsis thaliana* leaves and seeds. *Genetics* **159**, 359–370 (2001).
8. Loudet, O., Chaillou, S., Merigout, P., Talbotec, J. & Daniel-Vedele, F. Quantitative trait loci analysis of nitrogen use efficiency in *Arabidopsis*. *Plant Physiol.* **131**, 345–358 (2003).
9. Mita, S., Murano, N., Akaike, M. & Nakamura, K. 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 (1997).
10. Tikunov, Y. *et al.* A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiol.* **139**, 1125–1137 (2005).
11. Vorst, O. *et al.* A non-directed approach to the differential analysis of multiple LC-MS-derived metabolic profiles. *Metabolomics* **1**, 169–180 (2005).
12. Dixon, R.A. Engineering of plant natural product pathways. *Curr. Opin. Plant Biol.* **8**, 329–336 (2005).
13. Alonso-Blanco, C. *et al.* 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 (1998).
14. Broman, K.W. Mapping quantitative trait loci in the case of a spike in the phenotype distribution. *Genetics* **163**, 1169–1175 (2003).
15. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
16. de Koning, D.J. & Haley, C.S. Genetical genomics in humans and model organisms. *Trends Genet.* **21**, 377–381 (2005).
17. Mitchell-Olds, T. & Pedersen, D. The molecular basis of quantitative genetic variation in central and secondary metabolism in *Arabidopsis*. *Genetics* **149**, 739–747 (1998).
18. Reichelt, M. *et al.* Benzoic acid glucosinolate esters and other glucosinolates from *Arabidopsis thaliana*. *Phytochemistry* **59**, 663–671 (2002).
19. Kliebenstein, D.J. *et al.* Genetic control of natural variation in *Arabidopsis* glucosinolate accumulation. *Plant Physiol.* **126**, 811–825 (2001).
20. Kroymann, J. *et al.* A gene controlling variation in *Arabidopsis* glucosinolate composition is part of the methionine chain elongation pathway. *Plant Physiol.* **127**, 1077–1088 (2001).
21. Kliebenstein, D.J., Lambrix, V.M., Reichelt, M., Gershenzon, J. & Mitchell-Olds, T. Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in *Arabidopsis*. *Plant Cell* **13**, 681–693 (2001).
22. de la Fuente, A., Bing, N., Hoeschele, I. & Mendes, P. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20**, 3565–3574 (2004).
23. Li, Y., Baldauf, S., Lim, E. & Bowles, D.J. Phylogenetic analysis of the UDP-glycosyltransferase multigene family of *Arabidopsis thaliana*. *J. Biol. Chem.* **276**, 4338–4343 (2001).
24. Lim, E., Ashford, D.A., Hou, B., Jackson, G. & Bowles, D.J. *Arabidopsis* glycosyltransferases as biocatalysts in fermentation for regioselective synthesis of diverse quercetin glucosides. *Biotechnol. Bioeng.* **87**, 623–631 (2004).
25. D'Auria, J.C. & Gershenzon, J. The secondary metabolism of *Arabidopsis thaliana*: growing like a weed. *Curr. Opin. Plant Biol.* **8**, 308–316 (2005).
26. Alba, R. *et al.* Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. *Plant Cell* **17**, 2954–2965 (2005).
27. Lumba, S. & McCourt, P. Preventing leaf identity theft with hormones. *Curr. Opin. Plant Biol.* **8**, 501–505 (2005).
28. Oksman-Caldentey, K.M. & Saito, K. Integrating genomics and metabolomics for engineering plant metabolic pathways. *Curr. Opin. Biotechnol.* **16**, 174–179 (2005).
29. Lall, S., Nettleton, D., DeCook, R., Che, P. & Howell, S.H. Quantitative trait loci associated with adventitious shoot formation in tissue culture and the program of shoot development in *Arabidopsis*. *Genetics* **167**, 1883–1892 (2004).
30. DeCook, R., Lall, S., Nettleton, D. & Howell, S.H. Genetic regulation of gene expression during shoot development in *Arabidopsis*. *Genetics* **172**, 1155–1164 (2006).
31. Brem, R.B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
32. Schadt, E.E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
33. Jansen, R.C. Studying complex biological systems using multifactorial perturbation. *Nat. Rev. Genet.* **4**, 145–151 (2003).
34. Koornneef, M., Alonso-Blanco, C. & Vreugdenhil, D. Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annu. Rev. Plant Biol.* **55**, 141–172 (2004).
35. Stam, P. Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. *Plant J.* **3**, 739–744 (1993).
36. McCullagh, P. & Nelder, J.A. *Generalized Linear Models* (Chapman & Hall, New York, 1989).
37. Bing, N. & Hoeschele, I. Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* **170**, 533–542 (2005).
38. Schadt, E.E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37**, 710–717 (2005).
39. Zhu, J. *et al.* An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.* **105**, 363–374 (2004).