Metabolomics-assisted breeding: a viable option for crop improvement?

Alisdair R. Fernie¹ and Nicolas Schauer²

¹Max-Planck-Institute für Molekulare Pflanzenphysiologie, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany
²De Ruiter Seeds, Leeuwenhoekweg 52, 2661 CZ Bergschenhoek, the Netherlands

Metabolomics approaches enable the parallel assessment of the levels of a broad range of metabolites and have been documented to have great value in both phenotyping and diagnostic analyses in plants. These tools have recently been turned to evaluation of the natural variance apparent in metabolite composition. Here, we describe exciting progress made in the identification of the genetic determinants of plant chemical composition, focussing on the application of metabolomics strategies and their integration with other high-throughput technologies. Metabolomics represents an important addition to the tools currently employed in genomics-assisted selection for crop improvement.

Breeding crop compositional quality

Although the improvement of crop species has been a fundamental human pursuit since cultivation began some ten thousand years ago, we have only recently developed the capability to select for more than a handful of traits. For this reason, both early domestication and modern breeding activities imposed genetic bottlenecks; consequently, cultivated varieties of plants contain only a small fraction of the variation present in the gene pool. The wild ancestors of most plant species can still be found in natural habitats and germplasm centres have been set up worldwide to conserve these valuable resources in the form of seed banks [1], providing a source of genetic variation for crop improvement. This approach has been much exploited as a source of monogenic traits (for reviews, see Refs [2–4]), however, arguably it has been under-exploited in the study of quantitative traits. The utility of these seed banks was greatly enhanced by the widespread development of molecular-marker techniques in the early 1980s, which not only revolutionized plant breeding but also greatly assisted basic research by facilitating the introgression of defined genes or genomic regions from wild species or landraces.

Recent years have seen a dramatic increase in interest in understanding natural variance in plants and a growing number of research groups are using the introgression approach to study complex traits influenced by quantitative trait loci (QTL). Many of these studies identified QTL underlying yield (for example, see Refs [5,6]) or biotic and abiotic stress resistance (for example, see Refs [7,8] and, for recent reviews, see Refs [9–11]). Moreover, the spectacular technical advances of the post-genomic era have brought about a wealth of data, which enable us to elucidate associations between natural genetic and phenotypic variations in plants. Although many such studies have focussed on the model species Arabidopsis thaliana, they are increasingly being adopted in investigations in crop species. The nutritional status of crop plants is ultimately dependent on their metabolic composition and recent studies have highlighted the importance of compositional quality of crops for human health [12]. In this review, we focus on how the combination of genetic and metabolic approaches has been used to improve crop nutritional quality and evaluate the wider potential of this strategy. Although high costs (estimated at between 15 and 400€ per sample, depending on the technique) currently limit the use of metabolomic tools [13], they should be regarded as an additional, rather than an alternative, route towards crop improvement. Indeed, the costs for many post-genomic profiling methods, including metabolomics (see Glossary), are rapidly decreasing. Metabolomics is now an order of magnitude cheaper than transcript profiling [14] and is not reliant on having a pre-available genome sequence [15]. Although our knowledge of the chemical composition traits in plants usually lags behind that of yield and biotic and abiotic resistance traits, recent research in protein [16], oil [16,17] and provitamin A content in maize [18], starch content in potato and rice,

Glossary

**Dominant inheritance**: the situation wherein the allele inherited from one parent exerts its influence irrespective of the allele inherited from the other parent.

**Epistasis**: the interaction between genes. Epistasis occurs when the action of one gene is modified by one or several other genes, which are sometimes called modifier genes. The gene from which the phenotype is expressed is said to be epistatic, whereas the phenotype that is altered or suppressed is said to be hypostatic.

**Flux profiling**: evaluation of the rate of exchange of a labelled atom or atoms through multiple biochemical pathways. An important complement to metabolite profiling.

**Metabolite profiling**: the measurement of the broad range of metabolites within a single extract.

**Metabolomics**: the measurement of the small molecular metabolite complement of the cell.

**Overdominant inheritance**: or best-parent heterosis - the situation in which the offspring displays higher (or lower) levels of a trait than either of its parents.

**Primary metabolism**: encompasses essential reactions involving those compounds that are formed as a part of the normal anabolic and catabolic processes, which result in assimilation, respiration, transport and differentiation processes that take place in most, if not all, cells of an organism.

**Secondary metabolism**: a compound is classified as a secondary metabolite if it does not seem to directly function in the processes of growth and development. Even though secondary compounds are a normal part of the metabolism of an organism, they are often produced in specialized cells and tend to be more complex than primary compounds.
Box 1. Metabolite profiling technologies

Two techniques dominate metabolite profiling strategies: (i) mass spectrometry (MS); and (ii) nuclear magnetic resonance (NMR). Metabolomics, or the more modestly termed metabolite profiling, has been carried out since the mid 1970s [78], but only became a standard laboratory technique in the past decade [79]. Here, we focus on providing short definitions of the techniques and their relative advantages and disadvantages.

Gas-chromatography-mass-spectrometry (GC-MS), gas-chromatography-time-of-flight-mass-spectrometry (GC-TOF-MS) and liquid-chromatography-mass-spectrometry (LC-MS) are currently the standard mass-spectrometry methods for metabolite analyses. GC-MS technologies enable the identification and robust quantification of a few hundred primary metabolites within a single extract [80,81]. The main advantage of this instrument stems from the fact that it has long been used for metabolite profiling and, therefore, there are stable protocols for machine set-up, maintenance and usage. GC-TOF-MS offers several advantages, most notably, fast scan times, which give rise to either improved peak deconvolution (the ability to resolve partially co-eluting peaks) or higher sample throughput. Compared with GC-MS technologies, LC-MS offers several distinct advantages, chiefly its adaptability to measure a far broader range of metabolites encompassing both primary and secondary metabolites [28,77]. However, LC-MS usually uses electrospray ionization, which is prone to ion suppression (i.e. the competition of co-eluting entities for ionization energy) making it important to validate novel applications of this type of instrumentation. In addition to these machines, use of capillary-electrophoresis-mass-spectrometry (CE-MS) and Fourier-transform-ion-cyclotron-resonance-mass-spectrometry (FT-ICR-MS) systems have been demonstrated (for a review, see Ref. [82]). The first of these, CE-MS, is a highly sensitive methodology that can detect low-abundance metabolites and that provides good analyte separation, whereas the second, FT-ICR-MS, relies solely on very high resolution mass analysis, which potentially enables the measurement of the empirical formula for thousands of metabolites, however, it is somewhat limited by the lack of chromatographic separation.

NMR approaches, which rely on the detection of magnetic nuclei of atoms after application of a constant magnetic field, are the main alternative to MS-based approaches for metabolite profiling [79]. These are well-developed and well-validated methods and the computer software associated with NMR instrumentation is, consequently, also advanced. Furthermore, despite limitations in its sensitivity and, therefore, in metabolite coverage, it retains an advantage over MS-based approaches for certain biological questions. For example, it can be used non-invasively (i.e. on living cells) because the pH of the vacuole is different from that found elsewhere in the cell. NMR can provide subcellular information and it is easier to derive atomic information for flux modelling from NMR than from MS-based approaches.

and carotenoid content in tomato (for a review, see Ref. [19]), has advanced the understanding of these traits. In the past few years, rapid development of high-throughput tools for metabolic profiling (the parallel detection of the levels of multiple metabolites in a single extract; see Box 1 for details and Table 1 for an overview of technologies) has facilitated the analysis of a broad range of metabolites. Given that metabolic engineering in plants using targeted reverse genetic approaches often has unanticipated consequences, either on plant yield or on the levels of other cellular metabolites, the ability to screen a wide range of metabolites at once is very useful. Not only does this enable the detection of unwanted traits but it also facilitates a greater understanding of the metabolic network and how this interacts with developmental phenotypes. This is already true from the datasets acquired to date; however, because most metabolomic approaches are unbiased, the profiles they produce contain many unannotated peaks, representing unknown metabolites. Therefore, it seems likely that the power of metabolomics as a platform for the selection of breeding material can only improve. Owing to the increasing availability of immortalised plant populations, the acceleration in mapping and sequencing techniques and the decreasing unit cost of metabolomics-based phenotyping, a compelling argument can be made for the adoption of metabolomics as an integral component in plant breeding programs (see Figure 1 for a typical example).

Emerging data from a range of model and crop species are facilitating a better understanding of plant metabolic networks and are starting to uncover mechanisms of interaction between metabolism and development. Although metabolomics is a new scientific field (Box 1 and Table 1), a large amount of data have already been published on its application to widely divergent genetic populations. These data include assessments of the relative contribution of genotype and environment on metabolite composition, analyses of metabolite heritability and the integration of metabolite data with morphological phenotyping. Perhaps most excitingly, these recent studies demonstrate that, by using hybrid material, the contents of certain metabolites can be enhanced by a mechanism that does not invoke a yield penalty. Together with the recent advances in sequencing and transcript profiling (Boxes 2 and 3), the integration of data from several different genomics platforms is becoming economically feasible within a single project. Our focus is the potential of metabolomics in genomics-assisted breeding. We begin by selecting recent and historic success stories in which single chemical composition traits have been successfully bred.

Improving crop composition one metabolite at a time

Owing to technical limitations, researchers traditionally focused on a single or, at most, a handful of metabolic traits that were of greatest importance either for industrial or nutritional value. Prime examples of these targeted approaches include carotenoid content of tomato, protein content of maize and starch content of potato and rice (see Refs [16,19,20]). Researchers also focussed on simple metabolic processes, such as cold-sweetening in potato [21]. Perhaps the best example for a long-term program at improvement of crop compositional quality is the Illinois long-term selection experiment for protein and oil content in maize (http://www.ideals.uiuc.edu/handle/2142/3524), which began in 1896. Indeed, this experiment is arguably the longest continuous genetic experiment, comprising >100 cycles of selection and producing nine related maize populations with phenotypic extremes for grain composition [16]. These populations contain the known phenotypic extremes for maize kernel composition (i.e. individuals displaying the lowest and highest levels of either protein or oil) and are still used in current breeding programs as a favourable source of alleles associated with oil, protein and starch content. More recently, a combination of QTL map-based cloning, transgenesis and association mapping has been used to reveal the amino acid of the enzyme acyl-CoA:diacylglycerol acyltransferase...
Box 2. The utility of ‘next generation’ sequencing technology

The past few years have seen several advances in sequencing technology, including the development of massively parallel sequencing [83]. Although it took >10 years to sequence the human genome, complete genome sequencing can now be performed in a few months [17]. Traditional Sanger-based sequencing relies on the cloning and amplification of the DNA. The future promises faster and more sensitive whole genome sequencing technologies, the so-called ‘next generation’ sequencing, including single-molecule sequencing, sequencing by synthesis, and the even more futuristic method of nanopore sequencing [84]. Nanopore sequencing uses a single DNA molecule without the need of amplification and cloning. Although this technology is promising, it will take a few more years until it is used more widely by researchers. Sequencing costs are considerable, although it cost ~$3 billion to sequence the first human genome, the sequencing of James Watson’s genome cost only $1 million and latest estimates for a human genome sequence are $60K with a six week completion time (http://press.appliedbiosystems.com/corpcomm/applerapress.nsf/ABIDisplayPress/F426CD6F553255C2882574090082573E?OpenDocument&doctype=abi). The era of $1000 whole genome sequencing seems to be upon us –200 base pair, instead of single base pair, detection will probably rapidly accelerate sequencing and, thus, enable us to access the genetic basis of metabolomics associated traits much more rapidly than currently. It is perhaps the parallel development of both technologies that renders the incorporation of metabolomics within genome-assisted breeding strategies feasible.

In plant breeding, marker-assisted selection (MAS) employs restriction fragment length polymorphism (RFLP), cleaved amplified polymorphic sequences (CAPS), amplified fragment length polymorphism ( AFLP) or single sequence repeat (SSR) markers to track traits of interests. For the differentiation between two different alleles, single nucleotide polymorphism (SNP) markers are highly informative and easy to develop once the polymorphic region has been identified. SNP detection is somewhat limited in sample throughput. The use of PCR and proprietary systems such as SNPVawel (Keygene BV; http://www.keygene.com/keygene-products) can allow multiplex assays. However, advances in sequencing technologies enable the detection of thousand of SNPs in a single short run. Recent ‘proof-of-concept’ studies used 454 sequencing to discover genome wide transcriptomic SNPs in maize [85] and eucalyptus [86]. These studies revealed that the advances in sequence technologies can greatly enhance marker-assisted selection, although the costs are currently prohibitively high. However, if the expense is overcome, breeding strategies will almost certainly shift from single molecular marker analyses to sequencing-assisted breeding (SAB) to maximize control of trait segregation and hybrid purity. Thus, it seems highly likely that the association of metabolic trait properties to their underlying genetic basis will be dramatically accelerated by the combination of this approach with the application of metabolomics strategies.

Box 3. Transcriptomic approaches

The investigation of the total transcript content of a biological sample, known as transcriptomics, enables the detection of changes in transcript levels between different conditions and can, thus, be used in an attempt to identify mechanisms underlying quantitative variation in traits. The recent employment of microarray technology to identify genomic regions in whole-genome-covering RIL populations facilitates the identification of expression QTLs (eQTLs) controlling the transcript levels for individual genes. These loci can reside very close to the gene (e.g. in the promoter region) or near a transcription factor on another chromosome. In combination with phenotypic or metabolic studies, this integrated approach can facilitate the identification of genomic factors responsible for metabolic, yield, stress or disease resistance QTL. For example, Rowe et al. [87] integrated metabolic QTL analysis with eQTL studies in an Arabidopsis RIL population to identify a new regulatory myb factor subfamily for glucosinolate biosynthesis. A recent study of two barley varieties revealed >2000 polymorphic regions and extending the study to a 136 line double haploid population genome wide eQTL analysis exposed >23K eQTL affecting 16K genes [88]. A similar study to explore genes underlying resistance to wheat stem rust in barley by integration of disease resistance data revealed six major loci [89]. Two of these loci were already known to be determinants for stem rust resistance, but one of the four novel loci provided a very strong candidate gene encoding a histidine kinase which, therefore, represents a good target for crop improvement.

Tiling arrays that cover the whole genome can detect changes even in untranslated regions of the genome. Zeller et al. [90] have used this approach to detect polymorphic regions in a comparative proof-of-concept study of twenty Arabidopsis accessions. Whereas Zhang et al. [91] have used tiling arrays to assess genetic, epigenetic and transcriptional polymorphism in Arabidopsis. There are many potential applications of tiling arrays but, for crop breeding, whole genome polymorphism discovery is by far the most interesting.

Microarrays can be used to detect polymorphic regions in the transcriptome, even in moderately sized genomes such as Arabidopsis. Marker-assisted selection is of crucial importance in modern-day breeding. The increase in SNP-based markers is leading to bottlenecks in throughput and costs of genotyping. Recent studies have shown the applicability of microarrays for mapping a large number of SNPs [92,93].

An expanding catalogue of metabolite QTL

In the past few years, researchers have begun to use pathway-based approaches to identify the genetic determinants of crop compositional quality in several plant species. These approaches have led to a detailed dissection and an increase in our understanding of glucosinolate biosynthesis [24], seed oil synthesis [25] and oligosaccharide metabolism [26] in Arabidopsis and flavonoid biosynthesis in Arabidopsis [27,28], tomato [29] and Populus [30]. Furthermore, in the past two years, several studies have been carried out at the metabolomic level in Arabidopsis, tomato, wheat, rice, sesame, broccoli and mustard [31–41], which have lead to a far richer description of the natural variation of chemical composition in these species facilitating the identification of importance sources of allelic provitamin A levels in maize. This finding is particularly pertinent given the severe health disorders that result from vitamin A deficiency. Two of these strategies were at least partially reliant on association mapping, whereas as yet, no metabolomics studies have been published that have adopted this approach, the genetic determinants of many traits have nevertheless been detected using conventional map-based strategies.
variance for metabolic engineering (for a relevant overview, see Table 2).

The studies on Arabidopsis were based on three independent recombinant inbred line populations and demonstrated wide natural variation in both primary [32–34] and secondary [31] metabolism. Keurentjes et al. [31] focussed on the analysis of a Landsberg erecta (Ler) × Cape Verdi Islands (Cvi) recombinant inbred line (RIL) population and examined parental lines of a further 12 accessions. By profiling leaf material from these samples using an untargeted liquid chromatography mass spectrometry (LC-MS) method, they revealed a large quantitative variation in

<table>
<thead>
<tr>
<th>Technology</th>
<th>Application</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC-MS</td>
<td>Analyses of polar or lipophilic compounds (e.g. sugars, organic acids, tocopherols, vitamins).</td>
<td>Accuracy: &lt;50 ppm Mass range: &lt;350 Da</td>
</tr>
<tr>
<td>GC × GC-MS</td>
<td>Similar to GC-MS, but with better separation of co-eluting compounds and increased sensitivity owing to GC × GC.</td>
<td>Accuracy: &lt;50 ppm Mass range: &lt;350 Da</td>
</tr>
<tr>
<td>SPME GC-MS</td>
<td>Analyses of volatile compounds (e.g. aroma components, repellents).</td>
<td>Accuracy: &lt;50 ppm Mass range: &lt;350 Da</td>
</tr>
<tr>
<td>CE-MS</td>
<td>Analyses of polar compounds (e.g. amino acids, CoA-Derivates, sugars, organic acids, tocopherols, vitamins).</td>
<td>Accuracy: &lt;50 ppm Mass range: &lt;1000 Da</td>
</tr>
<tr>
<td>LC-MS</td>
<td>Analyses of mainly secondary metabolites (e.g. carotenoids, flavonoids, glucosinolates, vitamins).</td>
<td>Accuracy: 50–100 ppm Mass range: &lt;1500 Da</td>
</tr>
<tr>
<td>FT-ICR-MS</td>
<td>High-resolution MS in combination with LC is highly powerful. Enables the identification of unknown metabolites by m/z mass to charge ratio.</td>
<td>Accuracy: &lt;1 ppm Mass range: &lt;1500 Da</td>
</tr>
<tr>
<td>NMR</td>
<td>Non destructive analyses of abundant metabolites in a sample.</td>
<td>Mass range: &lt;~50 kDa</td>
</tr>
<tr>
<td>Direct-injection-MS</td>
<td>Non separative technique giving a fingerprint of the metabolic content in a biological sample.</td>
<td>Accuracy: 50–100 ppm Mass range: &lt;1500 Da</td>
</tr>
<tr>
<td>FAIMS-MS</td>
<td>Next generation hyphenation technology to MS. Enables selection of specific ions, reducing ion suppression and matrix effects. FAIMS enables the separation of isobaric compounds in combination with selective MS.</td>
<td>Accuracy: 50–100 ppm Mass range: &lt;1500 Da</td>
</tr>
</tbody>
</table>

Abbreviations: Da, Dalton; FT-ICR, fourier transform ion-cyclotron resonance; FAIMS, field asymmetric waveform ion mobility spectrometry; ppm, parts per million; SPME, solid phase micro extraction.

Figure 1. Profiling large populations to define novel metabolic QTL. Combining metabolomics, transcriptomics analysis and extensive phenotyping of large, genetically diverse populations (e.g. tomatoes) with an integrated bioinformatics platform will facilitate the identification of novel mQTL and the underlying genetics of the trait of interest. This schema serves to display how multiparallel metabolite and transcript profiling will probably inform future breeding strategies.
metabolism and showed that there were also qualitative differences in the range of metabolites present in the accessions. In addition, this study not only enabled an evaluation of the genetic architecture of aliphatic glucosinolate accumulation in Arabidopsis but also enabled inference of the structure of the underlying pathways. This work produced a very nice complement to early work in Arabidopsis in the groups of Richard Mithen and Jonathon Gershey (for example, see Refs [42] and [43]) and to subsequent work in broccoli and mustard [41]. These studies should, thus, aid in the selection of breeding lines that could potentiate the development of plants containing compounds that inhibit carcinogenesis.

By contrast, Meyer et al. [32] used gas chromatography mass-spectrometry (GC-MS) to study the primary metabolism of Columbia (Col) × C24 RIL population. Although no single primary metabolite displayed a strong correlation with plant biomass, Meyer et al. [32] identified a metabolic signature composed of contributions from various metabolites. Further studies on the QTL in the RIL population and in an introgression line (IL) population derived from the same parental accession led to the identification of six biomass QTL and 157 metabolic QTL. Two of the biomass QTL coincide with significantly more metabolic QTL (mQTL) than statistically expected, supporting the notion that the metabolic profile and biomass accumulation of a plant are linked. Furthermore, three of the six biomass QTL could be mathematically predicted based purely on their metabolite composition. More recently, a similar study was published on the RIL population resulting from a Bayreuth-0 (Bay) × Shahdara (Sha) cross [34]. This study, which was based on two independent experiments and enabled evaluation of the hereditability of mQTL in comparison to those of eQTL determined for the same samples (Box 2), found that the mQTL tended to be less heritable than the eQTL.

Moreover, statistical analyses of the data revealed that numerous mQTL displaying a moderate phenotypic effect frequently had most of their variation controlled by epistatic interactions, thereby enabling the generation and evaluation of network models that might help elucidate poorly defined metabolic pathways, such as those involved in the synthesis of important plant volatiles and hormones.

### Identifying metabolite QTL – moving from model species to crops

Not surprisingly, the most extensive studies on metabolomic natural variation have been conducted in Arabidopsis. However, increasingly, crop species have become the focus of metabolomic approaches. Astonishingly, many of these crop studies have been carried out on material from a single harvest [39–41,44], which makes it impossible to discriminate the effects of genotype from those of environment. However, these first ‘proof-of-concept’ investigations have provided important information about the natural diversity of metabolism.

#### Single harvest studies

Studies on rice, the staple food of almost half the world’s population, which furthermore provides three-quarters of the calorific intake of inhabitants of Asia [45], are particularly pertinent for world agriculture. In 2007, Kusano et al. [39] profiled 70 rice cultivars (including 68 of the rice world core collection; http://www.shigen.nig.ac.jp/rice/oryzabase/wild/coreCollection.jsp) using a combination of 1D and 2D GC coupled to MS, yielding a highly accurate inventory of the nutritional value of these cultivars.

In a similar, albeit smaller-scale study, Laurentin and co-workers used a combination of high-performance liquid chromatography (HPLC) and amplified fragment length polymorphism (AFLP) to determine the relationship between genetic and metabolic diversity in sesame [40]. Intriguingly, this study demonstrated that there was a large difference in the patterns of diversity at the genomic and metabolic levels, indicating that they were not tightly associated to one another. On the one hand, this observation, like that of the low heritability of the metabolome, argues against metabolomics as a means of selection. On the other hand, given that yield traits with a heritability of ~10% have been successfully incorporated into breeding programs, the fact that metabolite heritabilities of 25–35% are commonly estimated bodes well for the addition of this technique in future breeding strategies.

In tomato, screening of carotenoid metabolites by matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF-MS) was recently demonstrated to be useful for the screening of large populations. For this purpose, selected lines from two tomato populations (S. pennellii introgression lines and saturated mutants) were profiled to identify germplasm that is likely to be of high utility in the breeding of fruit containing high levels of these important nutriceuticals [44]. In addition to the health-promoting properties of certain anti-oxidant isoprenoids, such as carotenoids and vitamin E, the value in identifying and quantifying isoprenoids is also

### Table 2. Overview of crop studies employing metabolite profiling

<table>
<thead>
<tr>
<th>Crop</th>
<th>Main findings</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barley</td>
<td>P-deficiency in barley leads to shifts in carbohydrate metabolism, a reduction in organic acids and P-containing metabolites.</td>
<td>[75]</td>
</tr>
<tr>
<td>Corn</td>
<td>Targeted metabolite profiling revealed gene versus environmental effects in a set of corn hybrids and the influence of water stress on metabolite content.</td>
<td>[38,46]</td>
</tr>
<tr>
<td>Cucumis in cucumber</td>
<td>Combined transcript and metabolite profiling elucidated QTL involved in spider-mite-induced volatile biosynthesis</td>
<td>[76]</td>
</tr>
<tr>
<td>Potato</td>
<td>Genetic modification or environmental perturbations of potato plants result large effects on potato tubers composition.</td>
<td>[81]</td>
</tr>
<tr>
<td>Rice</td>
<td>Application of 2D GC-MS for the identification of natural variation on the metabolic level in 70 rice varieties revealed large metabolic differences between cultivars.</td>
<td>[39]</td>
</tr>
<tr>
<td>Tomato</td>
<td>Comprehensive metabolite profiling of a tomato introgression line library enables the identification of &gt;880 mQTL and the mode of inheritance of those QTL.</td>
<td>[35,36]</td>
</tr>
</tbody>
</table>
illustrated by the fact that they are an important target site for bleaching herbicides.

**Multiple harvest studies**
The metabolomic approach has also been performed in material from multi-harvest crops. A wide range of compositional traits including protein and oil contents, fatty acid, amino acid and organic acid content were analysed in three maize hybrids grown at three separate locations [46]. A broad profiling of tomato volatiles, which are extremely important flavour components, in a population of 74 *Solanum lycopersicum* × *S. pennellii* ILs yielded 100 QTL that were conserved across harvests [47]. Physiological studies on two of these volatiles – 2-phenylethanol and 2-phenylacetaldelyde – used a combination of metabolic and flux profiling alongside reverse genetic studies to confirm the biological pathway of these important aromatic compounds in tomato [48]. Similar, albeit not so extensive, studies have been carried out using intraspecific crosses of *S. lycopersicum* [49], documenting the levels of a subset of the most important volatile components of the fruit and defining a range of QTL for them. Studies in our laboratory on the same *S. pennellii* ILs described, using an established GC-MS method [36] over two independent harvests, resulted in the identification of 889 QTL governing the accumulation of 74 metabolites, including important primary metabolites, such as sugars, organic acids, essential amino acids, intermediate metabolites and vitamins. Although in many cases the metabolite content was increased, this was often associated with a yield penalty. To find out whether these traits were heritable, we grew the *S. pennellii* ILs for a third harvest, alongside lines that were heterozygous for the introgression (ILHs), enabling the evaluation of hereditability and the QTL mode of inheritance [35]. These studies revealed that the mean hereditability of the metabolite QTL was of a range that would be regarded as intermediate (i.e. between 0.20 and 0.35 – as was also found in *Arabidopsis* [34]). However, a handful of the traits were nevertheless highly correlated and displayed reasonable heritability (a mean r of between 0.3 and 0.69). A similar finding was observed in the maize study, which revealed a greater influence of environment on the metabolite profiles of three genotypes studied [46]. The comparative study of the tomato IL and ILHs, however, revealed that most of the metabolic QTL were dominantly inherited with a considerable number displaying an additive or recessive mode of action and only a negligible amount displaying the characteristics of overdominant inheritance. Interestingly, the mode of inheritance was quantitatively different between diverse classes of compounds with, for example, sugars and acids displaying significantly different patterns of inheritance. Moreover, several metabolite pairs belonging to the same pathway displayed a similar mode of inheritance at the same chromosomal loci, indicating that the variation in both metabolites is probably mediated by enzymes involved in their interconversion. However, the association between morphological and metabolic traits was far less prominent in the ILHs than in the ILs, which has wide implications for breeding strategies. The possibility of uncoupling enhanced metabolite content from any penalties with respect to plant performance and fecundity and redevelopment of hybrid genetic material could prove an important advance in the use of genomics-driven breeding approaches.

**Integration with other profiling data**
Integrating results from metabolic and morphological profiling proves to be a powerful strategy for crop improvement. Several recent studies have illustrated the utility of combining data from metabolomics with that from other genomics platforms to provide new insights on both gene annotation [50–53] and regulation in complex biological systems [54–56]. These approaches have resulted in the identification of numerous candidate genes including several in which expression correlates strongly with the levels of metabolites with important nutritional or organoleptic properties. To date, use of this approach on populations of wide genetic diversity has been restricted to *Arabidopsis* concentrating on the Bay × Sha Sha and Ler × Cvi RILs described earlier. Both of these populations were analysed by a combination of metabolomic and expression profiling [57,58] (Box 2) and the Ler × Cvi RILs were also analysed by enzymatic profiling [59]. These analyses revealed the full complexity of interaction across the various levels of cellular organization and, thus, the full scale of the challenge of engineering plants by targeted methods.

Evaluation of the Bay × Sha data was focussed on the aliphatic and indole pathways of glucosinolate biosynthesis and revealed that all loci controlling expression variation also affected the accumulation of the resulting metabolites and that epistasis was more apparent for the metabolic traits than the expression traits. Furthermore, the analysis indicated that, although natural variation in transcripts can significantly impact phenotypic variation, the natural variation in metabolites or the enzymatic loci that correspond to them can feedback to affect the transcripts [60]. Similar conclusions were made following the analysis of the integrated data relating to the central primary metabolism of the Col × Cvi RILs. The additional data provided at the enzymatic level revealed many examples of the complex circuitry governing metabolism [59]. Similarly to the Bay × Sha results, the natural variation in plant primary metabolism could be attributed to allelic differences in structural genes of catalytic enzymes such as those involved in glucosinolate biosynthesis, by the identification of regulatory loci or via metabolic signalling. The increasing availability and interest in cross-laboratory phenotyping of immortalised populations of both model and crop species [22,61–63] promises to be of great help in defining both the genetic and physiological mechanisms underlying trait variance, thereby rendering emergent QTL database resources [64,65] essential if we are to maximise the opportunities afforded us by these rich datasets. However, mining data for correlations only enables us to conclude that the variance in two traits is associated; we need to clone the QTL to understand the mechanisms by which these changes are brought. Most of the QTLs already cloned displayed major (dominant) effects and were identified in wide crosses (see Ref. [66] for a recent review). Recent developments in genetic and molecular biological
Combining metabolomics and association mapping
Association mapping has only recently been adopted in plant genetic research (for a review, see Ref. [67]) and it has been used for a few traits relevant to chemical composition research [18,68–71]. However, given the potential of this approach, particularly now that sequencing costs are rapidly decreasing (Box 2), it certainly should be considered within the wider context discussed in this article. Such mapping approaches have recently pinpointed associations between genomic regions of maize and kernel composition as well as starch content in potatoes, pigment content in tomato and provitamin A content in maize [68–70,72,73]. However, as yet, the number of cultivars or accessions that have been examined at high-throughput within a single study is limited. Several prototype studies assessing the combination of association mapping at the metabolomic level are currently underway worldwide. By and large, these approaches all adopt the same strategies as the studies already described, but they are carried out on a far greater number of genetically variant individuals. The success of the targeted metabolite approaches indicates that metabolomics studies could greatly benefit from the advantages afforded by a multiparallel approach because this would probably encompass the use of a higher mapping resolution, a greater allele number and a reduced time span to establish association as opposed to linkage analysis [67]. It seems likely to be only a matter of time before the efficacy of such strategies can be effectively assessed.

Concluding remarks and future perspectives
We have highlighted the current status of metabolomics in the assessment of broad genetic variance and focussed on platforms (Box 4) should greatly accelerate this cloning process.
Box 4. RNAi and miRNA approaches to breeding

Recent advances in our understanding of native gene silencing have facilitated the adoption of more rapid reverse genetic strategies, such as those afforded by functional testing of alleles. Both small interference RNAs (siRNAs) and microRNAs (miRNAs) have a pivotal role in gene silencing [94], with miRNAs being able to inactivate either specific genes or entire gene families. When brought into a plant, artificial miRNAs function as dominant suppressors of gene activity and these approaches have recently become a focus of crop researchers and commercial agricultural companies. For example, Warthmann and co-workers have recently designed artificial miRNAs (amiRNAs) to study agricultural important genes in rice [95]. The authors targeted a phytoene desaturase, which causes an albino phenotype, a GA20 oxidase, which results in dwarfism, and a gene encoding a phytochrome P450 monoxygenase, which results in an elongated upper internode. For each gene, two amiRNA constructs were designed to elucidate the importance of sequence properties to effectively silence gene expression. RNAi has also been used to silence the first step of flavonoid biosynthesis, which resulted in parthenocarpic tomato fruits [96]. Parthenocarpic leads to seedless fruits and is, thus, a highly desirable trait in crop plants for the consumer and for the seed provider. Recently, Monsanto and colleagues have developed a transgenic system based on RNAi to control colopteran insects, such as root worms. This technology is highly likely to be implemented in breeding programs in the near-future.

its potential role in informing breeding strategies. Although the cost and the extent of heritability need to be taken into account, the vast amount of knowledge accrued over a few years argues that this approach should be continued and extended. The shift from single metabolite measurements to platforms that can provide information on hundreds of metabolites has led to the development of better models to describe the links both within metabolism itself and between metabolism and yield-associated traits. The use of hybrids makes it possible to engineer plants that produce high levels of metabolites without accruing a yield penalty. The ongoing efforts to elucidate the metabolic response to biotic and abiotic stresses indicate that metabolomics-assisted breeding might also be useful in the development of crops that are more resistant to these stresses. The application of post-genomics tools should accelerate the selection process (Figure 2) and the combined use of metabolomics, genome sequencing and high-throughput reverse genetics (Box 4) will probably considerably shorten the time required for the production of elite lines. For this reason, we strongly believe that metabolomics-assisted breeding [74] can be applied to crop species in a similar manner to that which has already proven successful in breeding programs to increase disease resistance and herbicide or salinity tolerance [2,3,10] and which is certainly a viable option for crop improvement.

Acknowledgements

The Max-Planck-Gesellschaft, the Deutsche Forschungsgemeinschaft and the Bundesministerium für Bildung und Forschung is acknowledged for its support to the Fernie laboratory.

References

23 Kliebenstein, D.J. et al. (2001) Comparative quantitative trait loci mapping of aliphatic, indolic and benzylc glucosinolate production in Arabidopsis thaliana leaves and seeds. Genetics 159, 359–370


Laurentin, H. et al. (2008) Relationship between metabolic and genomic diversity in sesame (Sesamum indicum L.). BMC Genomics 9, 250


Causse, M. et al. (2002) QTL analysis of fruit quality in fresh market tomato: a few chromosome regions control the variation of sensory and instrumental traits. J. Exp. Bot. 53, 2089–2098

Fridman, E. et al. (2005) Metabolic, genomic, and biochemical analyses of glandular trichomes from the wild tomato species Lycopersicon hirsutum identify a key enzyme in the biosynthesis of methylketones. Plant Cell 17, 1252–1267


Alba, R. et al. (2005) Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. Plant Cell 17, 2954–2965


Okawa, A. et al. (2008) Rice metabolomics. Rice 1, 63–71


Barbazuk, W.B. et al. (2007) SNP discovery via 454 transcriptome sequencing. Plant J. 51, 910–918

Novaes, E. et al. (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics 9, 312
92 Pavy, N. et al. (2008) Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. BMC Genomics 9, 21
96 Schijlen, E.G. et al. (2007) RNAi silencing of chalcone synthase, the first step in the flavonoid biosynthesis pathway, leads to parthenocarpic tomato fruits. Plant Physiol. 144, 1520–1530