Review

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

Alisdair R. Fernie[1] and Nicolas Schauer[2]

[1] Max-Planck-Institute für Molekulare Pfanzenphysiologie, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany
[2] De 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 their 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 underexploited 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 a 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.

Corresponding author: Fernie, A.R. (fernie@mpimp-golm.mpg.de).

0168-9525/$ – see front matter doi:10.1016/j.tig.2008.10.010 Available online 21 November 2008

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.

NNR 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. 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, by ligation 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 ~$3billion to sequence the first human genome, the sequencing of James Watson's genome cost only $1million 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?OpenDocumentandtype=abi). The era of $1000 whole genome sequencing seems to be upon us and techniques relying on 5–200 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 SNPWave™ (Kegene 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 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].

responsible for determining oil content and composition in maize [17]. In a similar approach, albeit one that did not rely on association mapping, screening of a tomato introgression line population harbouring introgression of the wild species *Solanum pennellii* resulted in the identification of multiple QTL for total soluble solid content. One of these introgression lines (*Brix*9-2-5), was delimited to a single base-pair change in *LIN5*, an apoplastic invertase coding sequence and the line containing the allele from the wild species had a greater ability to bind sucrose and, hence, an increased sugar yield [22,23]. The tomato hybrid AB2 harbours a QTL from *S. pennellii* and is currently a leading processing variety. Another interesting example of is the recent identification, by association mapping, of lycopene ε cyclase as a key determinant of 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.

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 oligosaccharine 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

Table 1. Common and envisaged technologies in metabolite profiling[a]

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

[a]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.

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

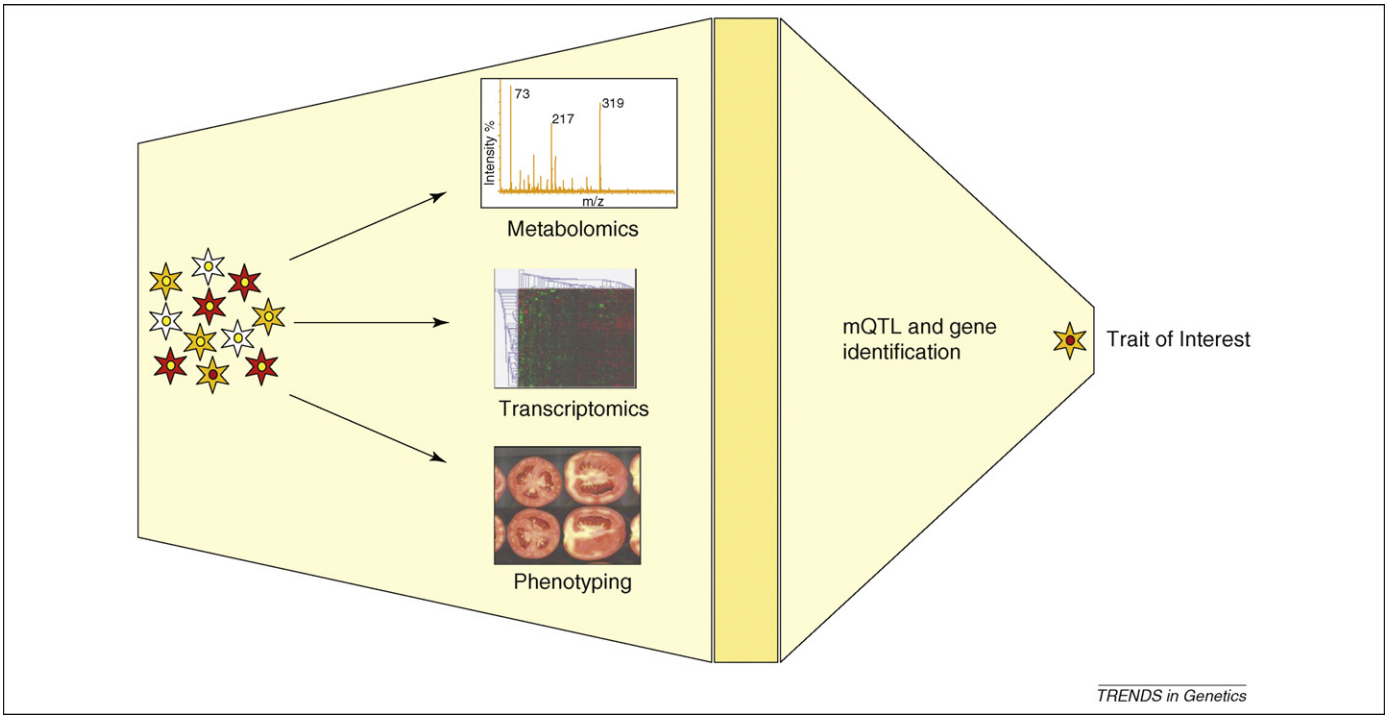

Figure 1. Profiling large populations to define novel metabolic QTL. Combining metabolomics, transcriptomics analysis and extensive phenotyping of large, genetically diverse populations (e.g. tomatoes) with an integrated bioinformatics platform will facilitate the identification of novel mQTL and the underlying genetics of the trait of interest. This schema serves to display how multiparallel metabolite and transcript profiling will probably inform future breeding strategies.

Table 2. Overview of crop studies employing metabolite profiling

Crop	Main findings	Refs
Barley	P-deficiency in barley leads to shifts in carbohydrate metabolism, a reduction in organic acids and P-containing metabolites. Shifting carbohydrates into amino and organic acid metabolism could lead to more efficient use of carbon under P-stress.	[75]
Corn	Targeted metabolite profiling revealed gene versus environmental effects in a set of corn hybrids and the influence of water stress on metabolite content.	[38,46]
Cucumis sp.	Combined transcript and metabolite profiling elucidated QTL involved in spider-mite-induced volatile biosynthesis in cucumber.	[76]
Potato	Genetic modification or environmental perturbations of potato plants result large effects on potato tubers composition.	[81]
Rice	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.	[39]
Tomato	Comprehensive metabolite profiling of a tomato introgression line library enables the identification of >880 mQTL and the mode of inheritance of those QTL.	[35,36]

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 Gershezhon (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 OTL. 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

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-phenylacetaldehyde – 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. penellii* 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 great 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

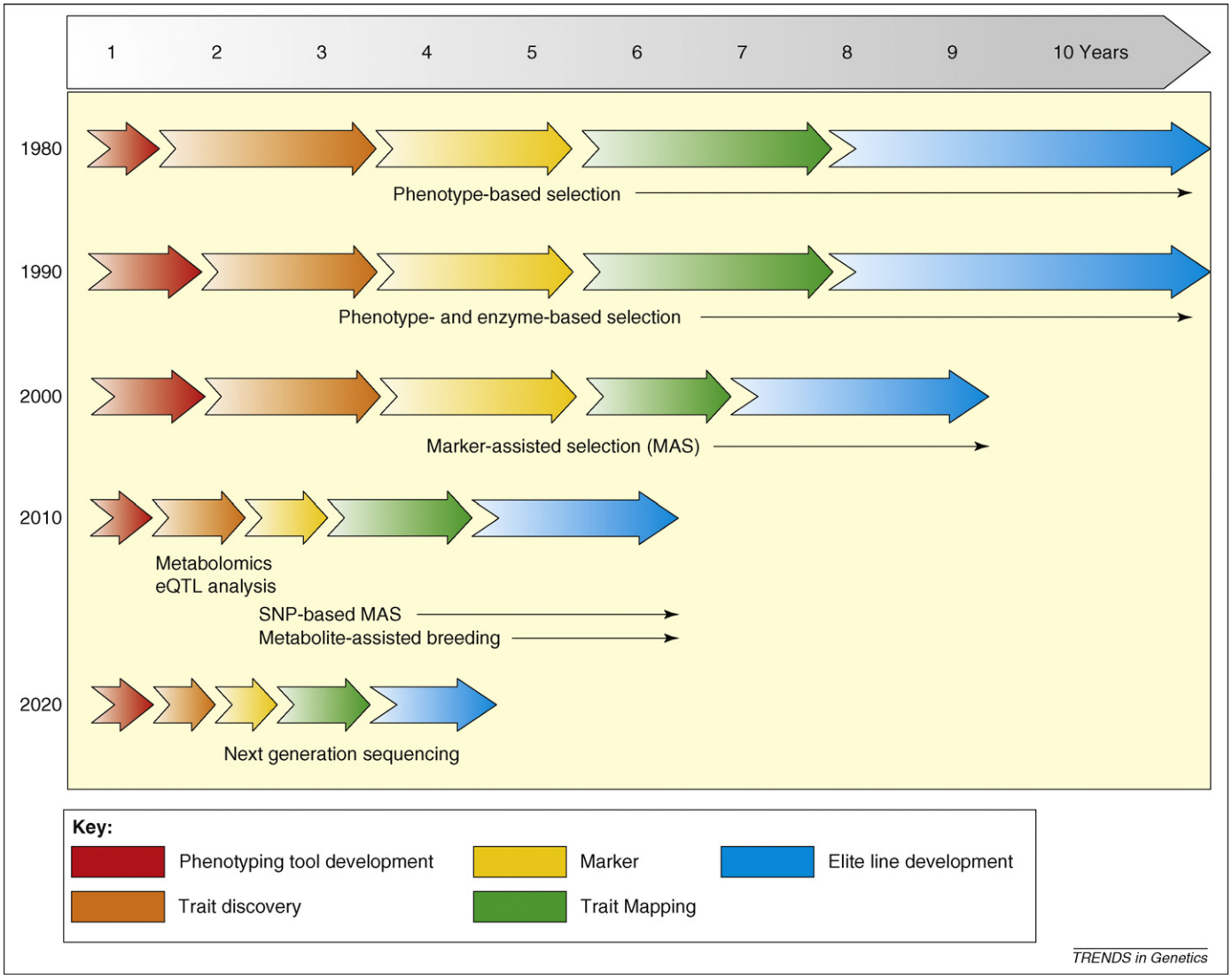

Figure 2. Breeding technology pipeline from past to present to future. The breeding pipeline from 1980 to that envisaged in 2020. In the past, trait discovery was mainly based on phenotypic observations, whereas marker development was restricted to phenotypic or enzymatic or protein markers. Thus, trait mapping and elite line development was a laborious task. The technological advances of molecular biology in the 1980s and 1990s enabled the application of molecular markers and improved the speed of trait mapping and commercial material development. Today, the application of marker-assisted selection in combination with new -omics approaches, such as metabolomics or transcriptomics (e.g. eQTL studies) enabled rapid discovery of new traits and allelic variation and, thus, improves the time to market by several years. In the future, the progress in trait discovery tools, plus simultaneous whole genome sequencing for marker development and trait mapping should shorten the market introduction of new varieties to ~4–5 years. Abbreviations: SNP, single nucleotide polymorphism.

platforms (Box 4) should greatly accelerate this cloning process.

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

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 monooxygenase, 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]. Parthenocarpy 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 insects [97]. In this study, the authors used RNAi as an enabling technology 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

1 Tanksley, S.D. and McCouch, S.R. (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* 277, 1063–1066
2 McCouch, S. (2004) Diversifying selection in plant breeding. *PLoS Biol.* 2, e347
3 Zamir, D. (2001) Improving plant breeding with exotic genetic libraries. *Nat. Rev. Genet.* 2, 983–989
4 Moose, S.P. and Mumm, R.H. (2008) Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiol.* 147, 969–977
5 Gur, A. and Zamir, D. (2004) Unused natural variation can lift yield barriers in plant breeding. *PLoS Biol.* 2, e245
6 Cong, B. *et al.* (2008) Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. *Nat. Genet.* 40, 800–804
7 Nelson, D.E. *et al.* (2007) Plant nuclear factor Y (NF-Y) B subunits confer drought tolerance and lead to improved corn yields on water-limited acres. *Proc. Natl. Acad. Sci. U. S. A.* 104, 16450–16455
8 Takano, J. *et al.* (2002) *Arabidopsis* boron transporter for xylem loading. *Nature* 420, 337–340
9 Tanksley, S.D. (2004) The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. *Plant Cell* 16 (Suppl), S181–S189
10 Takeda, S. and Matsuoka, M. (2008) Genetic approaches to crop improvement: responding to environmental and population changes. *Nat. Rev. Genet.* 9, 444–457
11 Varshney, R.K. *et al.* (2005) Genomics-assisted breeding for crop improvement. *Trends Plant Sci.* 10, 621–630
12 Demmig-Adams, B. and Adams, W.W. (2002) Antioxidants in photosynthesis and human nutrition. *Science* 298, 2149–2153
13 Borrás, L. and Slafer, G.A. (2008) Agronomy and plant breeding are key to combating food crisis. *Nature* 453, 1177
14 Kopka, J. *et al.* (2004) Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.* 5, 109
15 Stitt, M. and Fernie, A.R. (2003) From measurements of metabolites to metabolomics: an 'on the fly' perspective illustrated by recent studies of carbon-nitrogen interactions. *Curr. Opin. Biotechnol.* 14, 136–144
16 Moose, S.P. *et al.* (2004) Maize selection passes the century mark: a unique resource for 21st century genomics. *Trends Plant Sci.* 9, 358–364
17 Zheng, P. *et al.* (2008) A phenylalanine in DGAT is a key determinant of oil content and composition in maize. *Nat. Genet.* 40, 367–372
18 Harjes, C.E. *et al.* (2008) Natural genetic variation in lycopene ε cyclase tapped for maize biofortification. *Science* 319, 330–333
19 Fernie, A.R. *et al.* (2006) Natural genetic variation for improving crop quality. *Curr. Opin. Plant Biol.* 9, 196–202
20 Gebhardt, C. *et al.* (2005) Plant genome analysis: the state of the art. *Int. Rev. Cytol.* 247, 223–284
21 Menendez, C.M. *et al.* (2002) Cold sweetening in diploid potato: mapping quantitative trait loci and candidate genes. *Genetics* 162, 1423–1434
22 Fridman, E. *et al.* (2004) Zooming in on a quantitative trait for tomato yield using interspecific introgressions. *Science* 305, 1786–1789
23 Zamir, D. (2008) Plant breeders go back to nature. *Nat. Genet.* 40, 269–270
24 Kliebenstein, D.J. *et al.* (2001) Comparative quantitative trait loci mapping of aliphatic, indolic and benzylic glucosinolate production in *Arabidopsis thaliana* leaves and seeds. *Genetics* 159, 359–370
25 Hobbs, D.H. *et al.* (2004) Genetic control of storage oil synthesis in seeds of *Arabidopsis*. *Plant Physiol.* 136, 3341–3349
26 Bentsink, L. *et al.* (2000) Genetic analysis of seed-soluble oligosaccharides in relation to seed storability of *Arabidopsis*. *Plant Physiol.* 124, 1595–1604
27 Yonekura-Sakakibara, K. *et al.* (2007) Identification of a flavonol 7-o-rhamnosyltransferase gene determining flavonoid pattern in *Arabidopsis* by transcriptome coexpression analysis and reverse genetics. *J. Biol. Chem.* 282, 14932–14941
28 Tohge, T. *et al.* (2005) Functional genomics by integrated analysis of metabolome and transcriptome of *Arabidopsis* plants over-expressing an MYB transcription factor. *Plant J.* 42, 218–235
29 Spencer, J.P. *et al.* (2005) The genotypic variation of the antioxidant potential of different tomato varieties. *Free Radic. Res.* 39, 1005–1016
30 Morreel, K. *et al.* (2006) Genetical metabolomics of flavonoid biosynthesis in *Populus*: a case study. *Plant J.* 47, 224–237
31 Keurentjes, J.J. *et al.* (2006) The genetics of plant metabolism. *Nat. Genet.* 38, 842–849

32 Meyer, R.C. *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
33 Lisec, J. *et al.* (2008) Identification of metabolic and biomass QTL in *Arabidopsis thaliana* in a parallel analysis of RIL and IL populations. *Plant J.* 53, 960–972
34 Rowe, H.C. *et al.* (2008) Biochemical networks and epistasis shape the *Arabidopsis thaliana* metabolome. *Plant Cell* 20, 1199–1216
35 Schauer, N. *et al.* (2008) Mode of inheritance of primary metabolic traits in tomato. *Plant Cell* 20, 509–523
36 Schauer, N. *et al.* (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nat. Biotechnol.* 24, 447–454
37 Schauer, N. *et al.* (2005) Metabolic profiling of leaves and fruit of wild species tomato: a survey of the *Solanum lycopersicum* complex. *J. Exp. Bot.* 56, 297–307
38 Harrigan, G.G. *et al.* (2007) Metabolite analyses of grain from maize hybrids grown in the United States under drought and watered conditions during the 2002 field season. *J. Agric. Food Chem.* 55, 6169–6176
39 Kusano, M. *et al.* (2007) Application of a metabolomic method combining one-dimensional and two-dimensional gas chromatography-time-of-flight/mass spectrometry to metabolic phenotyping of natural variants in rice. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 855, 71–79
40 Laurentin, H. *et al.* (2008) Relationship between metabolic and genomic diversity in sesame (*Sesamum indicum* L.). *BMC Genomics* 9, 250
41 Rochfort, S.J. *et al.* (2008) Class targeted metabolomics: ESI ion trap screening methods for glucosinolates based on MSn fragmentation. *Phytochemistry* 69, 1671–1679
42 Magrath, R. *et al.* (1993) The inheritance of aliphatic glucosinolates in *Brassica napus*. *Plant Breed.* 111, 55–72
43 Kliebenstein, D.J. *et al.* (2001) Genetic control of natural variation in *Arabidopsis* glucosinolate accumulation. *Plant Physiol.* 126, 811–825
44 Fraser, P.D. *et al.* (2007) Metabolite profiling of plant carotenoids using the matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Plant J.* 49, 552–564
45 Hall, R.D. *et al.* (2008) Plant metabolomics and its potential application for human nutrition. *Physiol. Plant.* 132, 162–175
46 Harrigan, G.G. *et al.* (2007) Impact of genetics and environment on nutritional and metabolite components of maize grain. *J. Agric. Food Chem.* 55, 6177–6185
47 Tieman, D.M. *et al.* (2006) Identification of loci affecting flavour volatile emissions in tomato fruits. *J. Exp. Bot.* 57, 887–896
48 Tieman, D. *et al.* (2006) Tomato aromatic amino acid decarboxylases participate in synthesis of the flavor volatiles 2-phenylethanol and 2-phenylacetaldehyde. *Proc. Natl. Acad. Sci. U. S. A.* 103, 8287–8292
49 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
50 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
51 Goossens, A. *et al.* (2003) A functional genomics approach toward the understanding of secondary metabolism in plant cells. *Proc. Natl. Acad. Sci. U. S. A.* 100, 8595–8600
52 Achnine, L. *et al.* (2005) Genomics-based selection and functional characterization of triterpene glycosyltransferases from the model legume *Medicago truncatula*. *Plant J.* 41, 875–887
53 Hagel, J.M. *et al.* (2008) Quantitative 1H nuclear magnetic resonance metabolite profiling as a functional genomics platform to investigate alkaloid biosynthesis in opium poppy. *Plant Physiol.* 147, 1805–1821
54 Hirai, M.Y. *et al.* (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 101, 10205–10210
55 Urbanczyk-Wochniak, E. *et al.* (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.* 4, 989–993
56 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
57 Keurentjes, J.J. *et al.* (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
58 West, M.A. *et al.* (2007) Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* 175, 1441–1450
59 Keurentjes, J.J. *et al.* (2008) Integrative analyses of genetic variation in enzyme activities of primary carbohydrate metabolism reveal distinct modes of regulation in *Arabidopsis thaliana*. *Genome Biol.* 9, R129
60 Wentzell, A.M. *et al.* (2007) Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. *PLoS Genet.* 3, 1687–1701
61 Yu, J. *et al.* (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics* 178, 539–551
62 Pillen, K. *et al.* (2003) Advanced backcross QTL analysis in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* 107, 340–352
63 Ashikari, M. *et al.* (2005) Cytokinin oxidase regulates rice grain production. *Science* 309, 741–745
64 Zeng, H. *et al.* (2007) PlantQTL-GE: a database system for identifying candidate genes in rice and *Arabidopsis* by gene expression and QTL information. *Nucleic Acids Res.* 35, D879–D882
65 Gur, A. *et al.* (2004) Real time QTL of complex phenotypes in tomato interspecific introgression lines. *Trends Plant Sci.* 9, 107–109
66 Salvi, S. and Tuberosa, R. (2005) To clone or not to clone plant QTLs: present and future challenges. *Trends Plant Sci.* 10, 297–304
67 Yu, J. and Buckler, E.S. (2006) Genetic association mapping and genome organization of maize. *Curr. Opin. Biotechnol.* 17, 155–160
68 Wilson, L.M. *et al.* (2004) Dissection of maize kernel composition and starch production by candidate gene association. *Plant Cell* 16, 2719–2733
69 Palaisa, K.A. *et al.* (2003) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. *Plant Cell* 15, 1795–1806
70 Li, F. *et al.* (2008) Identification of the wax ester synthase/Acyl-CoA:diacylglycerol acyltransferase WSD1 required for stem wax ester biosynthesis in *Arabidopsis thaliana*. *Plant Physiol.* 148, 97–107
71 Buntjer, J.B. *et al.* (2005) Haplotype diversity: the link between statistical and biological association. *Trends Plant Sci.* 10, 466–471
72 Stich, B. *et al.* (2008) Multi-trait association mapping in sugar beet (*Beta vulgaris* L.). *Theor. Appl. Genet.* 117, 947–954
73 Zhao, J. *et al.* (2007) Association mapping of leaf traits, flowering time, and phytate content in *Brassica rapa*. *Genome* 50, 963–973
74 Morandini, P. and Salamini, F. (2003) Plant biotechnology and breeding: allied for years to come. *Trends Plant Sci.* 8, 70–75
75 Huang, C.Y. *et al.* (2008) Metabolite profiling reveals distinct changes in carbon and nitrogen metabolism in phosphate-deficient barley plants (*Hordeum vulgare* L.). *Plant Cell Physiol.* 49, 691–703
76 Mercke, P. *et al.* (2004) Combined transcript and metabolite analysis reveals genes involved in spider mite induced volatile formation in cucumber plants. *Plant Physiol.* 135, 2012–2024
77 Huhman, D.V. *et al.* (2005) Quantification of saponins in aerial and subterranean tissues of *Medicago truncatula*. *J. Agric. Food Chem.* 53, 1914–1920
78 Jellum, E. (1977) Profiling of human body fluids in healthy and diseased states using gas chromatography and mass spectrometry, with special reference to organic acids. *J. Chromatogr.* 143, 427–462
79 Fernie, A.R. *et al.* (2004) Innovation - Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.* 5, 763–769
80 Fiehn, O. *et al.* (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* 18, 1157–1161
81 Roessner, U. *et al.* (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* 13, 11–29
82 Oikawa, A. *et al.* (2008) Rice metabolomics. *Rice* 1, 63–71
83 Margulies, M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380
84 Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141
85 Barbazuk, W.B. *et al.* (2007) SNP discovery via 454 transcriptome sequencing. *Plant J.* 51, 910–918
86 Novaes, E. *et al.* (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 9, 312

87 Sonderby, I.E. *et al.* (2007) A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates. *PLoS One* 2, e1322
88 Potokina, E. *et al.* (2008) Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. *Plant J.* 53, 90–101
89 Druka, A. *et al.* (2008) Exploiting regulatory variation to identify genes underlying quantitative resistance to the wheat stem rust pathogen Puccinia graminis f. sp. tritici in barley. *Theor. Appl. Genet.* 117, 261–272
90 Zeller, G. *et al.* (2008) Detecting polymorphic regions in *Arabidopsis thaliana* with resequencing microarrays. *Genome Res.* 18, 918–929
91 Zhang, X. *et al.* (2008) Global analysis of genetic, epigenetic and transcriptional polymorphisms in *Arabidopsis thaliana* using whole genome tiling arrays. *PLoS Genet.* 4, e1000032
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
93 Edwards, J.D. *et al.* (2008) Development and evaluation of a high-throughput, low-cost genotyping platform based on oligonucleotide microarrays in rice. *Plant Methods* 4, 13
94 Ossowski, S. *et al.* (2008) Gene silencing in plants using artificial microRNAs and other small RNAs. *Plant J.* 53, 674–690
95 Warthmann, N. *et al.* (2008) Highly specific gene silencing by artificial miRNAs in rice. *PLoS ONE* 3, e1829
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
97 Baum, J.A. *et al.* (2007) Control of coleopteran insect pests through RNA interference. *Nat. Biotechnol.* 25, 1322–1326

Natural Variation in Arabidopsis: From Molecular Genetics to Ecological Genomics[1][W][OA]

Detlef Weigel*

Max Planck Institute for Developmental Biology, 72076 Tuebingen, Germany

One of the most remarkable biological insights in the past 30 years has been that many genetic programs for complex traits, such as flower or limb development, are shared across broad groups of organisms. These conserved pathways in turn can be tuned to produce tremendous phenotypic differences, not only between, but also within species. Intraspecific variation is often quantitative, one example being the onset of flowering, although there is also qualitative variation, such as in the ability to resist pathogens.

While many tools for quantitative genetics were developed by breeders, the model plant Arabidopsis (*Arabidopsis thaliana*) was adopted for studying the genetic architecture of quantitative traits soon after molecular markers for mapping became available (Chang et al., 1988; Nam et al., 1989). The species belongs to a small genus with nine members. Different from most of its congeners, Arabidopsis is self-compatible, and its life cycle can be as short as 6 weeks, both properties that greatly facilitate genetic studies. Its native range is considered to be continental Eurasia and North Africa (Al-Shehbaz and O'Kane, 2002), but it has been introduced throughout much of the rest of the world, especially around the northern hemisphere.

The potential of genetic variation to inform many different areas of Arabidopsis biology was most strongly advocated by Maarten Koornneef and his students. From the mid-1990s, they published both an impressive number of original research articles on this subject and a series of influential review articles that advertised the impact that the study of natural genetic variation could have on questions of both development and physiology (Alonso-Blanco and Koornneef, 2000; Koornneef et al., 2004).

Today, the study of natural variation in Arabidopsis continues to reveal new biology. In addition, the entire genus is increasingly being used to address fundamental questions of evolution (Mitchell-Olds and Schmitt, 2006; Bergelson and Roux, 2010). Some of the problems studied are: How, and how frequently, do new variants arise? Why do some variants rise to high frequency, while others are eliminated? And why are certain combinations of new variants incompatible with each other? Here, I will first give an overview of the tools and resources available for the study of natural variation in Arabidopsis. Next, I will present a few examples of how our knowledge of important biological processes has been improved through insights obtained from varieties other than the common laboratory accessions. Where similar or contrasting findings have been made in other species of the Brassicaceae, to which Arabidopsis belongs, I will mention these. The article concludes with a discussion of recent work that aims to integrate evolutionary and ecological studies with functional tests.

A final introductory note: Natural accessions of Arabidopsis have in the past often been referred to as "ecotypes." This term implies that a line has a unique ecology and is adapted to specific environments, as opposed to differing only in genotype from other varieties (Turesson, 1922b). Preferable is the neutral term accession, which merely means that a unique identifier in a collection has been assigned (Alonso-Blanco and Koornneef, 2000).

GENETIC TOOL KIT FOR THE STUDY OF NATURAL VARIATION

Experimental Populations for Genetic Mapping

Accessions of Arabidopsis vary in a number of traits (Fig. 1; Table I). The most general way to identify genes is by crossing two accessions, which may or may not have a different phenotype, but produce nonuniform F2 progeny. In the F2 or later generations, specific phenotypes are then associated with segregating genetic markers that distinguish the contributions from the parental genomes. When phenotypic classes are not discrete, this is done using the methods of quantitative trait locus (QTL) mapping (Falconer and Mackay, 1996).

Because marker analysis used to be very tedious and expensive, substantial efforts were invested early on into producing recombinant inbred lines (RILs), which constitute immortal populations in which recombinant chromosomes have been fixed through inbreed-

[1] This work was supported in part by Framework Programme 7 Collaborative Project AENEAS (contract Knowledge Based BioEconomy-2009–226477), by TRANSNET of the Bundesministerium für Bildung und Forschung program PLANT-Knowledge Based BioEconomy, by Schwerpunktprogramm 1529 "Adaptomics" and Schwerpunktprogramm 1530 "Flowering Time Control" of the Deutsche Forschungsgemeinschaft, by a Gottfried Wilhelm Leibniz Award of the Deutsche Forschungsgemeinschaft, and by the Max Planck Society.

* E-mail weigel@weigelworld.org.

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.111.189845

Figure 1. Gross morphological variation in Arabidopsis and relatives. A, Variation between Arabidopsis accessions. On top, vegetative rosettes of accessions grown for 4 weeks in long days are shown. They vary in rosette diameter and compactness, leaf shape, and tissue necrosis or onset of senescence. Similarly, variation in size and shape of individual leaves, in this case the sixth in the rosette, is apparent in the 10 examples shown on the bottom left. Finally, differences in overall architecture are illustrated with five plants. On the left is an early flowering accession with few rosette leaves. The next two flower later, but the second one from the left has reduced apical dominance. Finally, the two accessions on the right have similarly tall main inflorescences but differ in the number of secondary inflorescences. The appearance on the far right is common among wild-grown plants. B, Some characters, such as flower size and fruit shape, vary relatively little within Arabidopsis, but more dramatic variation is found in comparison with closely related taxa, such as *Capsella rubella* (left) and *A. lyrata* (right). Images courtesy of Eunyoung Chae, Sang-Tae Kim, and George Wang.

ing (Reiter et al., 1992; Lister and Dean, 1993; Fig. 2). RILs, which were first developed in mice (Bailey, 1971), have the advantage that they need to be genotyped only once but can be phenotyped repeatedly for many different traits and under different environmental conditions. An advantage of Arabidopsis is its self-compatibility, so that inbred lines can be easily generated by selfing and single-seed descent. Around 60 RIL populations are available from the stock centers as of the time this article is written (end of 2011; http://www.inra.fr/internet/Produits/vast/RILs.htm, http://www.arabidopsis.org/ and http://www.arabidopsis.info/). Importantly, the lengthy inbreeding process can now be bypassed through a revolutionary technology introduced by the laboratory of Simon Chan. This method allows the facile production of doubled haploid plants from recombinant populations (Ravi and Chan, 2010).

Even after five or six generations of inbreeding, which is customary for RILs, a small percent of the genome remains heterozygous. This turns out to have its own benefits. In such a heterogeneous inbred family (HIF), only a small portion of the genome segregates for the two parental alleles (Tuinstra et al., 1997). Additional recombinants that further reduce an interval of interest are easily derived from heterozygous HIF individuals, as are near isogenic lines (NILs) that are homozygous for either parental allele at this locus. A disadvantage of HIF-derived NILs is that each HIF has a unique genome composition and that one can therefore not easily place several QTL in a common genetic background.

NILs that carry only a small genomic region from one parent in a background that is otherwise composed of the genome of the other parent can also be generated directly by repeated backcrosses (Fig. 2). Such NILs, pioneered in crops where they are also called introgression lines (Seevers et al., 1971; Rhodes et al., 1989; Eshed and Zamir, 1995), are powerful for systematic analyses of interactions between genes from different genomes, although epistatic interactions among alleles from the introgressed genome are mostly lost. The properties of NIL sets are in many ways complementary to those of RILs, and they are particularly useful when introgression is performed in two directions (Törjék et al., 2008). NILs can identify QTL of smaller effect but with lower resolution than RIL populations (Falconer and Mackay, 1996; Keurentjes et al., 2007).

Although the genomes of RILs already contain more recombination events than F2 populations and therefore afford higher mapping resolution, this can be further increased with advanced intercross RILs, in which individuals from the F2 and later generations are intermated before inbred lines are derived (Darvasi and Soller, 1995; Balasubramanian et al., 2009). Other approaches involve the use of multiple parents, as in the MAGIC (for multiple advanced generation intercross) and AMPRIL (for Arabidopsis multiparent RIL) populations (Fig. 2; Kover et al., 2009; Huang et al., 2011). The MAGIC design is more elaborate and generates more recombination events per line than the AMPRIL strategy, but the founder genomes are less evenly represented in the final lines. Mapping in either population is more complex than with RILs, but with a sufficiently high density of intermediate frequency markers, one can infer the most likely local founder genotype. Even more so than simple F2 or RIL populations, AMPRILs and MAGIC lines are likely to contain genotypic combination not found in the wild.

QTL mapping accuracy increases with the MAGIC and AMPRIL populations, but not all possible QTL that can be found in pairwise crosses between some of the parents are detected. An alternative would be to combine the most informative subsets of RIL populations and to perform a joint QTL analysis. Especially when genotyped with common markers, a joint analysis can confirm common QTL (Bentsink et al., 2010; Salomé et al., 2011b).

Table I. *Traits studied by natural variation in Arabidopsis*
For references, see Supplemental Table S1.

Trait	Gene(s) Cloned?[a]
Aluminum content	N
Autonomous endosperm development	N
Auxin response	N
Carbohydrate availability and content	N
Cell wall composition	N
Chiasma frequency	N
Chromatin compaction	N
Circadian clock	C
Copper tolerance	Y
Crowding response	C
Disease resistance	Y
Drought response	N
Editing and processing of mitochondrial transcripts	Y
Elemental composition	Y/N
Flowering time	Y
Freezing tolerance	Y
Fruit number	C
Genetic robustness	N
Glucosinolate content	Y
Inflorescence replacement (mimicking grazing)	N
Jasmonate response	N
Leaf senescence	N
Leaf, inflorescence, and flower morphology	Y
Lethality in interploidy crosses	N
Life history traits other than flowering and growth	N
Light response	Y
Molybdenum content	Y
Nitrogen availability response	N
Oil content	N
Osmotic and salt stress tolerance	N
Phosphate content	N
Phytate content	N
Recruitment of bacterial rhizosphere communities	N
Root hydraulics	N
Root system size	N
Salicylic acid response	N
Salinity tolerance	N
Seed dormancy	Y
Seed germination, longevity	N
Seed lipids	N
Seed mucilage composition	Y
Sinapoylmalate biosynthesis	Y
Sodium accumulation	Y
Stomata density	N
Submergence tolerance	N
Sulfate content	Y
Terpene biosynthesis	Y
Thermal dissipation	N
Trichome density	Y
Zinc response	Y

[a]Y, Yes; N, no; C, likely candidates.

Some of the advantages of using RIL-type populations will continue to apply in the future. Trait values, especially those with low heritability, can be estimated more precisely due to replication (Soller and Beckmann, 1990; Mackay, 2001). Perhaps most importantly, one can study correlations between different traits, which can reveal fitness trade-offs, and reaction norms, the response of a specific genotype to different environments. However, not every geographic region where Arabidopsis is found is fairly represented in the available RIL populations because geographic sampling of Arabidopsis has so far been rather uneven (Fig. 3). Thus, forward genetics in additional material, even if composed mostly of F2 populations, will likely be informative. Fortunately, with reduced representation approaches such as restriction-associated DNA sequencing (RAD-seq) or genotyping-by-sequencing (Baird et al., 2008; Elshire et al., 2011) and multiplexing of genomic DNA from many individuals (currently, at least 96), costs for interrogating thousands of markers have dropped to a few U.S. dollars.

Finally, a general caveat when performing conventional genetic mapping is that chiasma frequencies differ between accessions (Sanchez-Moran et al., 2002). Data from F2 populations also support the conclusion that recombination rates vary depending on the cross (Salomé et al., 2011a). Thus, the ease with which loci are mapped will differ from cross to cross, even more so if structural variants interfere with recombination near the loci of interest.

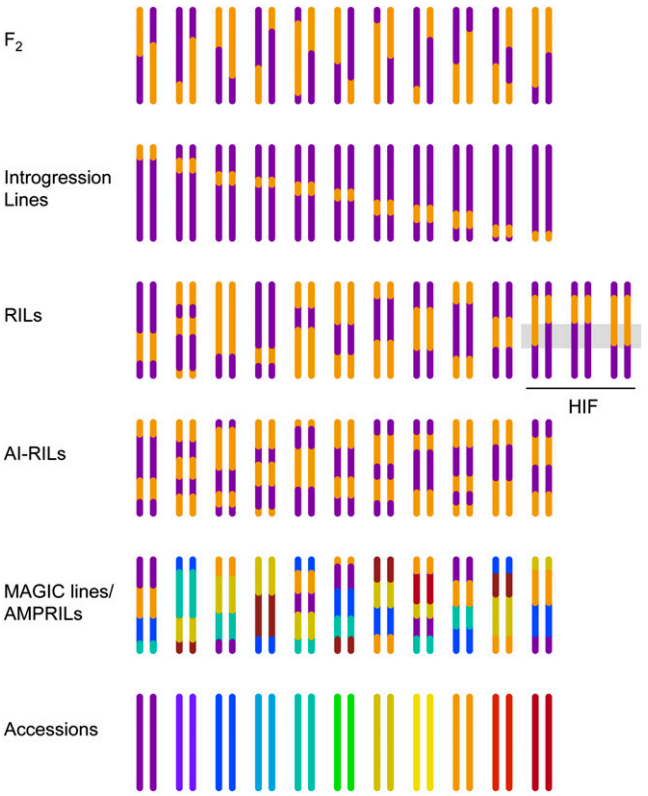

Figure 2. Populations for mapping genes causing trait variation. Colors indicate contribution from different parental accessions. Only one chromosome pair is shown for each individual. HIF individuals are derived from RILs, in which a small portion of the genome is still heterozygous.

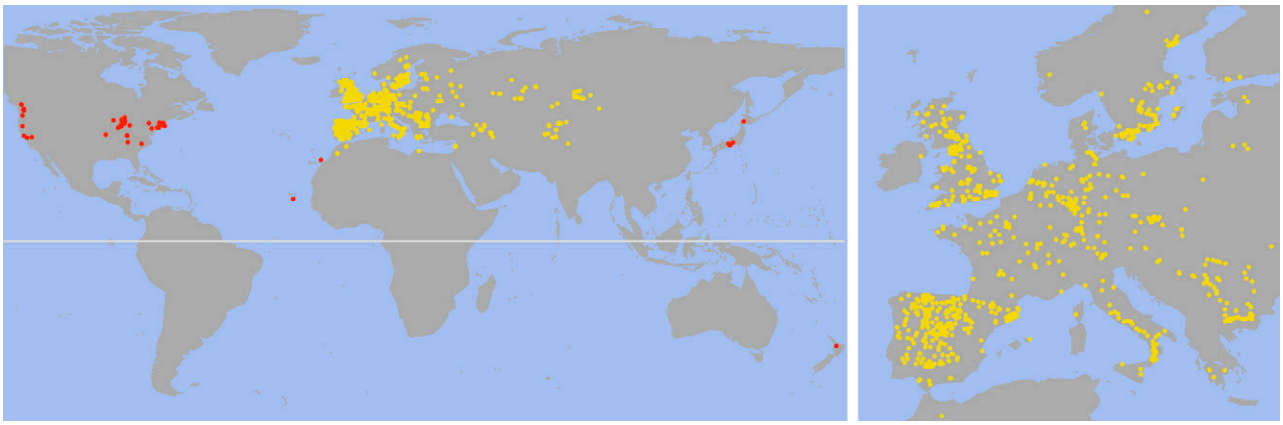

Figure 3. Distribution of over 7,000 Arabidopsis accessions collected from the wild and available in the stock center or soon-to-be-released collections. Western and southern Europe, including Great Britain, is heavily overrepresented, although sampling is not even. Accessions from the presumed native range are in yellow and likely introductions in red. Whether the distribution across China to Japan is continuous with the native range is unclear. Arabidopsis has been reported in additional locales, such as South Korea, and several African countries (Alonso-Blanco and Koornneef, 2000). Maps courtesy of George Wang.

Identification and Validation of Causal Genes and Polymorphisms

After a genomic interval underlying phenotypic differences has been identified, there are various options to track down the responsible gene, assuming that only a single gene is causal. Different from induced mutations, simply resequencing a region with dozens or more genes is on its own generally not informative because of the high number of polymorphisms that distinguish an arbitrary pair of accessions, about 1 in every 200 bp. Fortunately, compared to other multicellular organism in which natural variation is studied, Arabidopsis has the enormous advantage that almost all accessions are quite easily transformed by dipping flowering plants into a suspension of *Agrobacterium tumefaciens* containing a T-DNA vector with the transgene of interest (Clough and Bent, 1998).

If the final mapping interval does not contain a gene previously implicated in the trait of interest, one of the first steps will often be to investigate whether null alleles affect this trait. For the vast majority of genes, T-DNA insertion lines in the reference Columbia-0 (Col-0) background are available from the stock centers (http://arabidopsis.org, http://arabidopsis.info; for review, see Alonso and Ecker, 2006). The most straightforward approach to investigate the activity of individual genes in other genetic backgrounds is gene silencing, and collections of vectors for knocking down a large fraction of genes present in the reference genome are available, both for conventional hairpin RNA interference and artificial microRNAs (amiRNAs; for review, see Ossowski et al., 2008b). Gene silencing is a convenient tool to test the relative activity of alleles, an approach that we have called quantitative knockdown (Schwartz et al., 2009). It is conceptually related to quantitative complementation, where different alleles are examined in the hemizygous state, by crossing a homozygous strain to a tester that carries a knockout allele of the gene of interest (Mackay, 2001; Fig. 4).

As an alternative, one can introduce genomic fragments spanning the region of interest to identify the gene(s) affecting the trait under investigation. Transgenic complementation also allows the examination of chimeric genes in different backgrounds to pinpoint the causal region, or even nucleotide, within an allele. A possible complication arises from the fact that the addition of an extra wild-type copy of an independent gene in the same pathway can quantitatively affect the phenotype and thus confound the interpretation of the observed phenotypes. An attractive feature of amiRNAs is that one can engineer transgenes that do not change the encoded protein but do not respond to silencing by a specific amiRNA anymore (Palatnik et al., 2003). One can thus use an amiRNA to knock out the endogenous gene and at the same time introduce a variant copy of the gene that is not affected by the amiRNA. This allows in essence the functional replacement of one allele with another.

A final word of caution: Spontaneous mutations are not as rare as one might think, with direct measurements indicating about one new single base pair mutation per haploid genome and generation (Ossowski et al., 2010). Thus, not every genetic variant that distinguishes accessions must be a natural variant in the sense that it was present in nature. Indeed, there are now several reports of mutations with large phenotypic effects that were segregating in an accession and may only have arisen after the accession was collected. Two of these cases affect parents of commonly used RIL populations, Landsberg *erecta*-0 and Bayreuth-0 (Doyle et al., 2005; Loudet et al., 2008; Laitinen et al., 2010). Thus, even if misidentification of an accession has been ruled out, which is not uncommon (Anastasio et al., 2011; Simon et al., 2011), there can be true genetic

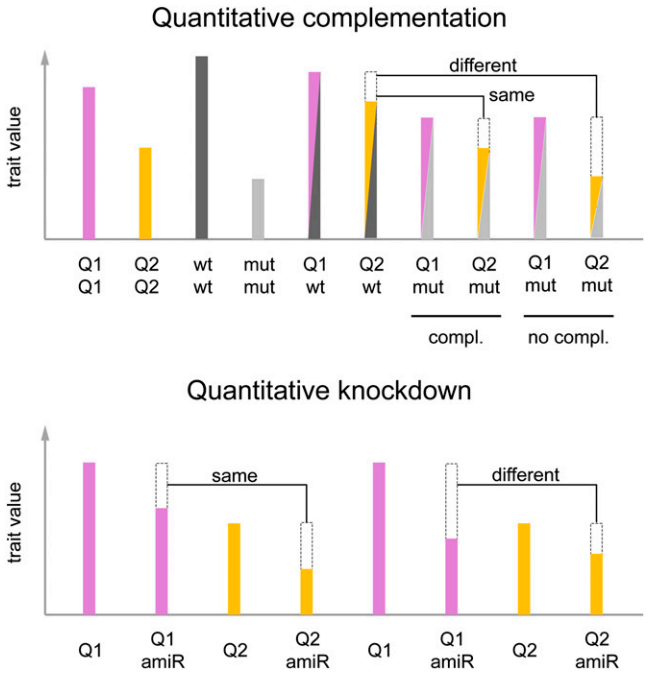

Figure 4. Quantitative complementation and knockdown to determine whether QTL are allelic to a candidate gene. Both tests rely on quantitative comparisons between genotypes; the dashed boxes indicate phenotypic differences to the genotype to the left. In a quantitative complementation test, one determines whether the two QTL alleles, Q1 and Q2, are differentially affected when heterozygous with the wild-type (wt) or mutant (mut) allele of a candidate gene (Mackay, 2001). If the QTL alleles respond differently, i.e. if in this example only Q1 complements the mutant phenotype, the candidate gene and the QTL are probably allelic. Similarly, in a quantitative knockdown experiment, a differential effect of an amiRNA (amiR) against the candidate gene indicates that the Q1 allele has lower activity than Q2 and that the candidate gene is likely responsible for the QTL.

and phenotypic differences between accessions that share recent common ancestry.

WHOLE-GENOME RESOURCES FOR THE STUDY OF NATURAL VARIATION

Enabling Genome-Wide Association Studies

Genetic mapping in crosses is greatly facilitated when genome-wide polymorphisms, or better yet the entire genome sequences, of the investigated accessions are known. If a sufficient number of genome sequences is available, one can even dispense with experimental crosses and exploit shared ancestry to directly identify common alleles that are responsible for phenotypic variation in the entire population. This approach was first proposed for human, already before the first finished human genome sequence was in sight (Lander, 1996; Risch and Merikangas, 1996). Because obtaining complete genome sequences for many individuals of the same species was out of question at the time, it was proposed to rely on linkage disequilibrium (LD). LD refers to the fact that in most species there has not been enough historic recombination to produce all possible combinations of physically adjacent polymorphisms, but rather that sequence variants are normally found in haplotype blocks of various lengths. Thus, a causal polymorphism can in principle be identified indirectly through its association with any of the other sequence variants in its haplotype block (Kruglyak, 1999; Jorde, 2000). The term that is normally used today for this experimental strategy is genome-wide association study (GWAS). A shortcut that reduces the required genotyping effort has been to make use of prior information and to first focus on genes already shown to affect a trait of interest (Long et al., 1998; Caicedo et al., 2004; Olsen et al., 2004; Balasubramanian et al., 2006; Ehrenreich et al., 2009), but this has become largely obsolete today.

While the principles of GWAS are easy to understand, important limitations arise from population structure, that is, not all investigated individuals being equally distantly related to each other. Powerful methods have been developed to correct for population structure, but how to reliably detect alleles that are largely fixed between populations remains a challenge. Other issues are allelic heterogeneity, that is, alleles at a single locus with similar effects on gene function having arisen repeatedly; or complex genetic architecture, where many different genes affect the same trait. A recent article by Myles et al. (2009) provides an excellent primer of the challenges for GWAS.

As with RIL analyses, the selfing nature of Arabidopsis is a boon for GWAS, since each accession needs to be genotyped or sequenced only once but can be phenotyped many times. Magnus Nordborg almost single-handedly convinced the Arabidopsis community of the feasibility and usefulness of GWAS approaches, even before high-density genotype information was available (Aranzana et al., 2005; Zhao et al., 2007). While initial estimates of LD in Arabidopsis were too high (Nordborg et al., 2002, 2005), it finally turned out that LD in the global population extends over not more than about 5 to 10 kb, or one to two genes, which is very convenient for GWAS (Kim et al., 2007). It is thought that the relatively low LD reflects a history of frequent outcrossing together with rapid dispersal enabled by the selfing mode of reproduction.

The first enterprise with the goal of finding a large fraction of sequence variants used high-density custom arrays with almost one billion unique oligonucleotides to interrogate the genomes of 20 accessions, including the Col-0 reference accession (Clark et al., 2007). This set was chosen to be maximally diverse based on a previous analysis of 96 accessions, from which about 1,000 short fragments distributed throughout the genome had been dideoxy sequenced (Nordborg et al., 2005). The most important information to come from the array-based resequencing study was a collection of hundreds of thousands of nonsingleton single nucleotide polymorphisms (SNPs) that could be used for

GWAS (Kim et al., 2007). About 216,000 SNPs, or one every 0.5 kb, have been subsequently typed in over 1,000 accessions (Horton et al., 2012), chosen from a larger panel of more than 5,000 accessions for which information from 149 intermediate frequency markers was available (Platt et al., 2010). The high density of SNPs meant that a typical haplotype block was tagged with several SNPs, which made GWAS in Arabidopsis right away more powerful than in humans. Despite similar LD characteristics, GWAS in human initially used only about 1 SNP per 6 kb (Wellcome Trust Case Control Consortium, 2007).

Prospects of GWAS in Arabidopsis

Several proof-of-concept examples have now been published, indicating that GWAS will often be successful in Arabidopsis. In the first comprehensive study, over 100 different morphological, physiological, and molecular traits were analyzed in 96 to 192 accessions (Atwell et al., 2010). In several cases, known genes were rediscovered, and in many others, plausible candidates were identified with high precision. The most impressive results, in agreement with previous pilot studies (Aranzana et al., 2005), were obtained for disease resistance, which is often controlled by single genes with very large effects. This is in contrast with humans, where effect sizes of QTL detected by GWAS are often small (McCarthy et al., 2008; Manolio et al., 2009).

The utility of GWAS can be increased by making use of prior information, such as functional data from mutant studies, gene annotation, or membership of genes in specific regulatory networks to prioritize GWAS candidates (Aranzana et al., 2005; Schadt et al., 2005; Atwell et al., 2010; Chan et al., 2011). Similarly, QTL mapping in experimental populations can greatly reduce the portion of the genome that one has to consider for the location of GWAS QTL (Brachi et al., 2010; Nemri et al., 2010). This approach becomes particularly powerful when both strategies are directly integrated using experimental populations with several parents, so that alleles pinpointed by GWAS are represented in multiple founder backgrounds. The term nested association mapping has been coined for this approach, which was pioneered in maize (*Zea mays*; Yu et al., 2008; McMullen et al., 2009). Arabidopsis populations, such as the MAGIC lines and AMPRILs, serve a similar purpose (Kover et al., 2009; Huang et al., 2011). An alternative will be to examine several independent RIL populations. An advantage of using RIL sets over F2 individuals in this case is that for each set of founders, the lines can be chosen to be maximally informative in terms of contribution of the founder genomes, thus greatly reducing phenotyping efforts (Xu et al., 2005; Simon et al., 2008).

Because of the plasticity of plant development and physiology, the influence of genes on the phenotype is very often dependent on the environment, often codified as gene-by-environment or GxE interaction. Similarly, the effects of individual genes are often modified by other genes in the genome because genes do not act on their own but form more or less complex functional networks. When genes have nonadditive effects, this is called GxG or more commonly an epistatic interaction. While the identification of epistatic QTL is standard fare for mapping in experimental populations (Mackay, 2001), this continues to be a major challenge for GWAS. This has been suggested to be computationally and statistically feasible several years ago (Marchini et al., 2005), and several computational strategies have been developed since (Mitchell-Olds, 1995; Cordell, 2009; Kam-Thong et al., 2011). However, I am not aware of an example where all variants were used in a GWAS to detect epistatic loci. Here again, mapping in experimental populations, perhaps in combination with network reconstruction (Rowe et al., 2008; Jiménez-Gómez et al., 2010; Kerwin et al., 2011), should help to reduce the search space for GWAS of epistatic loci.

A Proliferation of Genome Sequences

In addition to the anonymous SNPs for the first generation of GWAS in Arabidopsis, array-based resequencing revealed tens of thousands of amino acid replacements along with hundreds of more drastic mutations that are likely to eliminate the function of many genes in various accessions. In addition, a large percentage of the reference genome was found to be missing in each accession (Borevitz et al., 2007; Clark et al., 2007; Zeller et al., 2008; Plantegenet et al., 2009). This implied that, conversely, the reference accession Col-0 likely lacked a substantial portion of genes present in other accessions. The analysis of individual loci had already shown that some gene families could differ greatly between accessions. Foremost are the disease resistance genes of the nucleotide-binding site-Leu-rich repeat (NB-LRR) class, with both presence/absence polymorphisms and highly divergent alleles in different accessions (Grant et al., 1995; Caicedo et al., 1999; Noël et al., 1999; Stahl et al., 1999; Rose et al., 2004). A logical next step was therefore to scrutinize the genomes of accessions for sequences not represented in the reference genome. With the advent of new sequencing technologies, this goal became attainable at a reasonable cost. Even before these methods were exploited to the same end for human genomes, it was shown that they not only gave an accurate account of small-scale polymorphisms in Arabidopsis genomes but that they could also be used to detect copy number variants and to assemble sequences absent from the reference (Ossowski et al., 2008a).

The 1001 Genomes Project for Arabidopsis was announced in 2007 (Nordborg and Weigel, 2008; Weigel and Mott, 2009). The initial proposal was to pursue a two-pronged hierarchical strategy for defining the pangenome of Arabidopsis. The first hierarchical aspect was a sampling of accessions throughout the range of Arabidopsis such that diversity could be

analyzed at global, regional, and local scales. Thus, rather than equidistant distribution of samples, it was envisioned that the project would include regional populations separated by distances measured in kilometers as well as individuals from within local stands spaced only meters apart. The second hierarchical aspect was to produce genome sequences at different levels of accuracy and completeness such that a relatively small number of highly accurate and complete genomes would inform the analysis of a much larger number of genomes that had not been completely assembled. The rationale behind this proposal was that mere lists of sequence variants that result from simple resequencing approaches, in which sequence reads are only aligned to a target genome, can be misleading. Specifically, because of false-negative problems, trying to reconstruct contiguous sequences by superimposing known isolated polymorphisms on the reference genome information can be problematic. To overcome these limitations, two groups have introduced reference-guided assembly approaches (Gan et al., 2011; Schneeberger et al., 2011), in which the Col-0 reference genome (Arabidopsis Genome Initiative, 2000) is first used to identify portions of the genome that are conserved in other accessions. Gaps are then filled in by assembling sequence reads and anchoring them to the known bits. As expected, multiple out-of-phase insertions or deletions in coding sequences can combine to restore open reading frames (Schneeberger et al., 2011). Similarly, additional mutations can make up for defects in splice acceptor or donor sites, as can be inferred from transcriptome analysis by RNA sequencing (Gan et al., 2011). The error rates of these reference-guided assemblies in single-copy regions were close to what was deemed as the lower acceptable bound in the initial reference genome sequencing project, about 1 in 10,000 bp (although final error rates in the reference genome were probably only about one-fifth; Ossowski et al., 2008a).

As expected from previous resequencing studies, up to 2% of reference positions were judged to be absent from the new assemblies. Conversely, up to 0.6% of the new assemblies represented sequences not found in the reference genome (Gan et al., 2011; Schneeberger et al., 2011). Because the new sequencing technologies generate more error-prone and shorter reads, and the insert sizes for paired-end sequencing libraries are generally smaller as well (Metzker, 2010), there are limits to closing gaps between regions that are well conserved relative to the reference genome. That bases present in the reference, but missing from a nonreference accession, outnumber the opposite class several-fold indicates the shortcomings of the reference-guided assemblies, since it should be equally likely that insertions and deletions occur on either lineage. We are thus currently faced with a paradox: >90% of the euchromatic portion of an accession's genome can be sequenced for a few hundred dollars, but the remainder can only be recovered when investing many hundred or thousand times that amount. This is particularly relevant because some of the most interesting genes in the genome, such as many disease resistance genes, reside in highly variable gene clusters with often nearly identical tandem repeats that are even challenging for assembly from dideoxy sequenced bacterial artificial chromosomes or fosmid clones (Noël et al., 1999).

While the most common approach for the identification and annotation of variants has been comparison against the reference, a multiple alignment consensus benefits the evaluation of complex alleles (Gan et al., 2011). However, with the rapid increase in the number of genome sequences, simple all-against-all comparisons will soon not be feasible anymore because of the time required to perform them. It has therefore been proposed to represent the pangenome, that is, the collection of all possible sequence variants along each chromosome, in a single data structure as a graph, which would both facilitate the identification of polymorphisms in newly sequenced genomes and their classification as shared or unique (Schneeberger et al., 2009).

Insights from Comparing Genome Sequences

Apart from supporting forward genetic studies in Arabidopsis, genome sequences have increased our understanding of the evolutionary history of the species. Array-based comparison of 20 accessions revealed only a single large region in the genome that was shared by the majority of accessions, indicative of this region having experienced recent and strong selection in many different populations (Clark et al., 2007). Remarkably, the much more fine-grained information from short-read sequencing of 80 lines did not substantially change this picture of strong selective sweeps being rare, even though population differentiation along the genome is not uniform (Cao et al., 2011).

In addition to local polymorphism patterns that are shaped by selection and demography, there are consistent chromosomal-scale differences that are probably caused by molecular and genetic factors, such as mutation, recombination, and biased gene conversion. One of these is an excess of polymorphisms in regions adjacent to the centromeres (Borevitz et al., 2007; Clark et al., 2007), which has also been reported in *Medicago truncatula* and rice (*Oryza sativa*), but not in maize (Gore et al., 2009; Huang et al., 2010c; Branca et al., 2011). The interpretation of polymorphism patterns in Arabidopsis has also benefited from the high-quality reference sequence available now for the close relative *Arabidopsis lyrata* (Hu et al., 2011). In agreement with lack of conservation between the two species reflecting either that sequences are dispensable or subject to species-specific positive selection, regions found only in Arabidopsis are more polymorphic than shared regions (Cao et al., 2011).

Finally, Arabidopsis accessions harbor extensive variation in mitochondrial genomes (Forner et al.,

2005; Arrieta-Montiel et al., 2009), in subtelomeric regions (Kuo et al., 2006; Wang et al., 2010), and in heterochromatic repeats, including retrotransposons and rDNA (Fransz et al., 2000; Davison et al., 2007; Ito et al., 2007). Structural differences between mitochondrial genomes can be revealed relatively easily by new sequencing methods (Davila et al., 2011). Furthermore, although read lengths and insert sizes are insufficient for long-range reconstruction of highly repetitive regions of the genome, read coverage and sequence variation in individual reads can be exploited to determine differences in genome size and repeat content (James et al., 2009; Tenaillon et al., 2011).

Utility of Genome Sequences

As of the time that this article was written (end of 2011), over 100 genome sequences for Arabidopsis had been published. In addition, sequence data for over 300 additional accessions were already publicly available. In aggregate, commitments for over 700 accessions had been made, indicating that the initial goal of 1,001 genome sequences would be reached well before the end of 2012 (http://1001genomes.org).

Several of the Arabidopsis genome sequences were immediately useful. For example, the Landsberg *erecta* accession is commonly used for mutant screens, and its genome sequence is facilitating the mapping and analysis of induced mutations. Similarly, several of the accessions are parents of RIL populations (Ossowski et al., 2008a; Schneeberger et al., 2009, 2011; Gan et al., 2011), and their genome sequences are aiding the identification of polymorphisms responsible for QTL. Genome sequences also provide an inventory of potential knockout mutations, which is informative given that a considerable fraction of natural genetic variation is due to loss-of-function alleles. Examples are new alleles of *PHYTOCHROME D* (*PHYD*) and *FRIGIDA LIKE1* (*FRL1*), for which before only single alleles were known (Aukerman et al., 1997; Schläppi, 2006; Cao et al., 2011).

In addition, the 1001 Genomes Project is advancing GWAS. As discussed above, the first phase of GWAS in Arabidopsis has been based on a set of 216k tag SNPs, which were estimated to predict >90% of all common variants (Kim et al., 2007; Horton et al., 2012). It is simple to call the same SNPs in any of the accessions of the 1001 Genomes Project and to include any line that has not been array genotyped into GWAS projects that makes use of the 216k tag SNP array data. Furthermore, it is possible to accurately impute common variants identified by whole-genome sequencing in array genotyped accessions and GWAS with imputed data detects additional polymorphisms linked to traits under consideration (Cao et al., 2011).

Apart from increasing the chances that sequence differences directly responsible for trait variation are found by GWAS, a major advantage of complete genome sequences is that they support the prediction of activity differences between potentially causal alleles. For example, in coding regions, mutations that disrupt the open reading frame or affect splicing are more likely to affect gene function than codon or silent changes. And among amino acid substitutions, one can estimate how probable it is that a mutation has deleterious effects based on conservation of that amino acid in other species (Ng and Henikoff, 2006).

Complete genome sequences will thus help to tackle one of the major challenges of GWAS, allelic heterogeneity, where several different alleles have similar effects on the trait of question. That independent alleles at the same locus can have the same phenotypic consequences has been known for a quarter of a century, since the first genes responsible for genetic disorders or cancer in humans were cloned (Royer-Pokora et al., 1986; Clark et al., 1989; Botstein and Risch, 2003). In Arabidopsis, the flowering regulators *FRIGIDA* (*FRI*) and *FLOWERING LOCUS C* (*FLC*) are often partially or completely inactivated, with many of these alleles being found only in single accessions (Johanson et al., 2000; Le Corre et al., 2002; Gazzani et al., 2003; Michaels et al., 2003; Lempe et al., 2005; Shindo et al., 2005; Méndez-Vigo et al., 2011). Drastic mutations that prematurely terminate or partially delete the same open reading frame are found more often than expected by chance in the genomes of different accessions (Cao et al., 2011; Fig. 5). This might be the outcome of positive selection, as is the case for *FRI* and *FLC* (Toomajian et al., 2006), or purifying selection being weak or absent. In either case, the presence of multiple alleles with similar effects on a particular phenotype makes the detection of such loci in GWAS analyses difficult since each polymorphism is considered separately (Myles et al., 2009). If, instead, all alleles with similar predicted activity differences were combined or, better yet, if alleles were considered according to their relative degree of activity, this hurdle could be overcome.

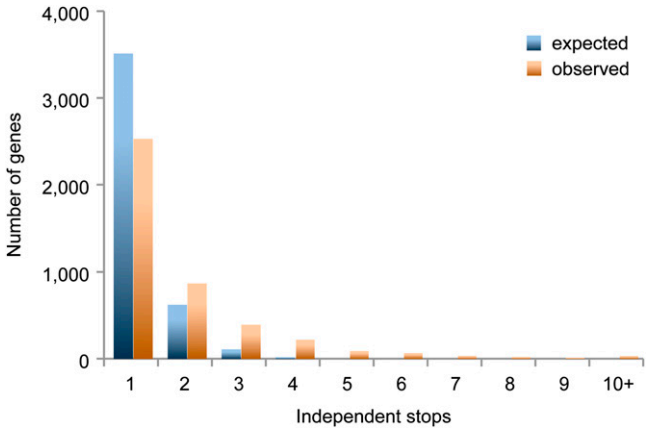

Figure 5. Comparison of expected and observed occurrences of 8,133 independent premature stops in 4,263 protein coding genes, considering all genes with >90% coverage in 75 out of 80 accessions. Data are from Cao et al. (2011).

The methods discussed in the preceding paragraph would be a considerable improvement over the strategy that is gaining popularity in humans: the search for an excess of rare variants in candidate genes. In rare-variant-burden methods, rare variants are combined for the purposes of contrasting phenotypically distinct classes of individuals, but functional effects of alleles are ignored, and these methods are not integrated into standard GWAS (Asimit and Zeggini, 2010).

Epigenomic Variation

GWAS in humans, where it is not unusual that tens of thousands of individuals are analyzed, has been successful in detecting many alleles, even with very small effects, but the fraction of the total variation explained by these variants is often only small. This also has been the case for traits such as height that are known to be highly heritable from family studies. Some possibilities are that genetic architecture may be more complex, with many interacting loci, or that rare alleles are more important than anticipated (see above). An alternative explanation, which is en vogue in many circles, is that epigenetic variation unlinked to sequence variants and, hence, not detectable by conventional GWAS is responsible for many phenotypic differences (McCarthy et al., 2008; Manolio et al., 2009).

Epigenetic differences can have obvious consequences in plants. In several species, including Arabidopsis, spontaneously occurring epialleles with overt phenotypes have been described (Jacobsen and Meyerowitz, 1997; Cubas et al., 1999; Hollick et al., 2000; Soppe et al., 2000; Stam et al., 2002; Manning et al., 2006; Martin et al., 2009). The epialleles often show increased cytosine methylation of the promoter and strongly reduced RNA expression. In several cases, the epialleles are associated with structural changes, such as the *g* mutation in melon, which is apparently caused by the insertion of a transposon and spread of DNA methylation into adjacent sequences.

Tiling array analyses comparing two different pairs of Arabidopsis accessions have shown that these differ in the extent of methylation at individual cytosines. That there are fewer differences in transposable element than genic methylation between natural accessions (Vaughn et al., 2007) agrees with transposable element methylation being more stable in inbred lines (Becker et al., 2011; Schmitz et al., 2011). Methylation differences seem to be largely stable in F1 hybrids (Woo and Richards, 2008; Zhang et al., 2008; Groszmann et al., 2011), but methylation patterns can change at relatively high rates, around 1% or more, in subsequent generations (Vaughn et al., 2007). The fluidity of the genomic methylation landscape after crosses is consistent with RNA-dependent DNA methylation mediated by short interfering RNAs being able to target other loci in trans, as long as these harbor sufficient levels of sequence similarity (Melquist and Bender, 2003). This is substantiated by nonadditive expression levels of short interfering RNAs and correlated effects on DNA methylation in F1 hybrids (Groszmann et al., 2011).

Importantly, although epialleles with phenotypic effects are largely stable and can be inherited over many generations, most revert occasionally to the wild-type form (Jacobsen and Meyerowitz, 1997; Cubas et al., 1999; Hollick et al., 2000; Soppe et al., 2000; Stam et al., 2002; Manning et al., 2006; Martin et al., 2009). The stability of DNA methylation in inbred Arabidopsis lines has recently been examined directly (Becker et al., 2011; Schmitz et al., 2011). While loss and gain of methylation at individual sites occurred much more often than mutations in the nucleotide sequence (Ossowski et al., 2010), changes in larger methylated regions similar to the ones that distinguish epialleles identified by forward genetics were rare. However, both types of methylation changes were distinguished from DNA mutations in that the same positions were affected in independent lines much more often than expected by chance and that there was an appreciable rate of reversions.

Crosses of wild-type lines to mutant strains with largely demethylated genomes have also revealed a wide range in the stability of epialleles after the causal mutations had been segregated away (Reinders et al., 2009; Teixeira et al., 2009). Consistent with the more labile nature of epialleles, heritability estimates in such lines are considerably lower than they are in natural accessions for the same traits (Roux et al., 2011). Thus, while the large majority of DNA methylation differences is sufficiently stable to account for inheritance within a limited number of generations, it remains unclear how often epialleles can become subject to Darwinian selection and thus make a contribution to long-term evolution. If reversion rates exceed the selective advantage conferred by an epiallele, its frequency in the population will be largely determined by the equilibrium of forward and reverse epimutation rates (Slatkin, 2009; Johannes and Colomé-Tatché, 2011).

In summary, although natural epialleles are often due to nearby structural variation, crosses between divergent accessions can induce new epialleles in trans. While the first class does not pose a problem for conventional GWAS, as such alleles should be tagged by linked sequence polymorphisms, the second class would only be revealed if GWAS would be extended to directly include information on DNA methylation profiles. A different question is whether epialleles are equally, more, or less likely than DNA alleles to reflect adaptation to the local environment.

LEARNING NEW BIOLOGY FROM THE STUDY OF NATURAL VARIATION

While knowledge about the origin and phenotypic effects of sequence polymorphisms is central to un-

derstanding how species adapt to their natural environment, most studies of genetic variation in Arabidopsis have probably been motivated by the desire to identify regulatory and other genes that are not present in the common laboratory accessions. An especially original use of natural variation has been the search for second site modifiers of *ABA insensitive3* and *leafy cotyledon1* mutant phenotypes. Both mutants suffer from impaired seed maturation, and seed viability declines much more rapidly than in wild-type plants. Introgression of the mutant alleles into other accessions identified natural modifiers that can partially suppress the mutant phenotypes, possibly pointing to new regulators of seed maturation (Sugliani et al., 2009). In a similar manner, the *CAULIFLOWER* (*CAL*) gene was discovered serendipitously as an enhancer of the *apetala1* (*ap1*) mutant phenotype. *CAL* and *AP1* turned out to be paralogs with an asymmetrical relationship: While *AP1* can compensate for loss of *CAL* activity, the reverse is not true. Thus, in contrast with induced *ap1* mutations, natural loss-of-function alleles of *CAL* have no overt phenotype on their own and are only noticed if *AP1* is inactive as well (Bowman et al., 1993; Kempin et al., 1995).

Arabidopsis was used early on to identify genes that control seed dormancy (van Der Schaar et al., 1997). For ease of cultivation, common laboratory accessions had been selected to be early flowering (more below) and to have little dormancy, meaning that seeds would germinate relatively quickly after harvest. The *DELAY OF GERMINATION1* (*DOG1*) locus, the first dormancy QTL cloned, encodes the prototype of a small gene family of unknown molecular function. There is extensive variation in *DOG1* expression levels between accessions, suggesting the presence of many functionally distinct alleles of *DOG1* (Bentsink et al., 2006). Arabidopsis accessions also remain an important resource for functional and evolutionary analyses of large-effect resistance genes (Staskawicz et al., 1995). This is a large area for which there are several recent in-depth reviews (Nishimura and Dangl, 2010).

Below, I will discuss three naturally variable traits in some more detail: trichome density, which provides a paradigm for how information from multiple genome sequences can be used to pinpoint causal polymorphisms; glucosinolate content, which has an underlying biochemical pathway with variation at almost every step; and the onset of flowering, a developmental trait with a well-understood molecular basis.

Trichome Density

Early studies by Rodney Mauricio and Mark Rausher came to the conclusion that both physical defenses in the form of leaf hairs (trichomes) and chemicals (glucosinolates) reduce herbivore damage to Arabidopsis in the field but that these are not without costs (Mauricio and Rausher, 1997; Mauricio, 1998). Several genes have been identified as affecting trichome density of natural Arabidopsis accessions. The most dramatic effects are seen in accessions that are glabrous, that is, lack trichomes completely, and at least two different loss-of-function mutations at *GLABRA1* (*GL1*) have been found. Whether a fitness tradeoff, as suggested for other defense traits, underpins the *GL1* polymorphisms is unknown. Balancing selection, however, which is often taken as a sign of trade-offs, does not appear to be responsible for maintaining different *GL1* alleles (Hauser et al., 2001). Glabrousness caused by inactivating mutations in *GL1* also segregates in *A. lyrata* and *Arabidopsis halleri* populations (Hauser et al., 2001; Kärkkäinen and Ågren, 2002; Kivimäki et al., 2007; Kawagoe et al., 2011).

A less extreme phenotype of reduced trichome density is caused in some Arabidopsis accessions by a nonsynonymous substitution in *MYC1* (Symonds et al., 2011). As another warning to population geneticists, one of the exons was found to exhibit a strong signal of divergent selection, with many amino acid substitutions. However, this signal was not correlated with trichome density.

Other accessions have increased trichome number relative to the Col-0 reference accession, and *ENHANCER OF TRY AND CPC2* (*ETC2*) has been identified as the causal gene (Hilscher et al., 2009). *ETC2*, *MYC1*, and *GL1* all encode transcription factors, with GL1 promoting and ETC2 repressing trichome formation by competing for interaction with common partners, a group of basic helix-loop-helix proteins that includes GL3 and MYC1 (Ishida et al., 2008). In contrast with *MYC1*, the high- and low-activity variants of *ETC2* segregate at intermediate frequencies, indicating that *ETC2* is a major determinant of natural variation in trichome number. *ETC2* very likely corresponds to one of the first QTL that was mapped in Arabidopsis, *REDUCED TRICHOME NUMBER* (Larkin et al., 1996), and consistent with alleles of different activity being common, *ETC2* can also be detected by GWAS (Atwell et al., 2010). Notably, it had initially been suggested that *ETC2* has only a minor role in trichome formation, a conclusion that came from studies done with common accessions that have an *ETC2* allele without obvious disruptions but with nevertheless low activity.

The work on *ETC2* is noteworthy because of how the causal polymorphism was first pinpointed using a strategy that should be broadly applicable. To triangulate the causal region in the final mapping interval, accessions with either very high or very low trichome densities were selected, and the extent of haplotype sharing in each group was compared, which identified a small region with only two candidate polymorphisms (Hilscher et al., 2009). Transformation with chimeric transgenes provided conclusive support that one of the variants, a nonsynonymous mutation, was reducing the activity of *ETC2*. With the resources of the 1001 Genomes Project, these types of local association studies should become a common strategy for the endgame in identifying QTL after conventional mapping in F2 or similar populations.

Glucosinolate Content

In addition to the gene-for-gene resistance loci that are effective against individual pathogen strains (for review, see Nishimura and Dangl, 2010), Arabidopsis accessions also show quantitative variation in resistance, in particular against herbivorous insects. As with trichomes, chemical defenses in the form of a Brassicaceae-specific class of secondary metabolites, the glucosinolates, can reduce herbivore damage (Blau et al., 1978). There are considerable inter- and intraspecific differences in the repertoire of glucosinolates, which are hydrolyzed by the enzyme myrosinase into the active defense compounds (Kliebenstein et al., 2005). In Arabidopsis, *METHYLTHIOALKYLMALATE SYNTHASE* (*MAM*) and *AOP* are the two major loci responsible for variation in glucosinolate biosynthesis, with additional contributions from the *GSL-OH* locus (Kliebenstein et al., 2001; Kroymann et al., 2001, 2003). Hydrolysis of the glucosinolates is further affected by the polymorphic *EPITHIOSPECIFIER PROTEIN* and *EPITHIOSPECIFIER MODIFIER1* loci (Lambrix et al., 2001; Zhang et al., 2006). In other Brassicaceae, several of the same genes are responsible for intraspecific variation in glucosinolate content, including *A. lyrata* (Li and Quiros, 2003; Heidel et al., 2006).

Notably, both the *MAM* and *AOP* loci are complex, with several tandem arrayed genes that vary in presence, enzyme activity, or expression level between accessions, giving rise to more than two alternative allelic states, processes that are apparently driven by positive selection (Kliebenstein et al., 2001; Kroymann et al., 2001, 2003). At least *MAM* shows a similar pattern of diversity created by gene duplication and neofunctionalization between other members of the Arabidopsis genus as well as closely related genera (Benderoth et al., 2006).

The detailed understanding of the control of glucosinolate accumulation in turn supports research into broader questions of genetic variation, such as the importance of stochastic variation, which was found to be genetically encoded as well (Jimenez-Gomez et al., 2011).

Flowering Time

Seed production is one of the most important components of fitness, and to optimize seed set, plants need to flower at the right time of year. In agreement with Arabidopsis is found in places with very different growing seasons, natural accessions differ greatly in their flowering behavior. Beginning with Laibach (1943, 1951), several investigators reported flowering variation not only in inbred accessions, but also in individuals collected from the wild (Napp-Zinn, 1957; Cetl et al., 1968; Jones, 1971; Westerman, 1971). That this trait is under selection has also been inferred from population genomics analyses (Flowers et al., 2009) and from the finding of latitudinal and altitudinal clines, likely due to covariation of flowering time with climatic factors (Caicedo et al., 2004; Stinchcombe et al., 2004; Lempe et al., 2005).

The first natural allele to be mapped with molecular markers in Arabidopsis was at the *FRI* locus, which segregates in a Mendelian manner in crosses between late- and early-flowering accessions (Lee et al., 1993; Clarke and Dean, 1994). The first QTL mapped in Arabidopsis were also ones controlling flowering (Kowalski et al., 1994; Clarke et al., 1995), followed by many additional QTL studies (for review, see Koornneef et al., 2004; Shindo et al., 2007). Mapping in crosses and GWAS have shown that flowering time variation can be explained by relatively few large-effect QTL (Atwell et al., 2010; Brachi et al., 2010; Li et al., 2010; Salomé et al., 2011b; Strange et al., 2011), which is very different from maize (Buckler et al., 2009).

FRI and the epistatically acting *FLC* gene are responsible for a large fraction of flowering time variation in Arabidopsis accessions when these are not exposed to a winter-like vernalization treatment. *FRI* promotes expression of the FLC transcription factor, which directly represses genes with positive roles in flowering (Li et al., 2008; Deng et al., 2011). Allelic variation at *FLC* likely accounts for flowering time differences in other Brassicaceae as well, including *Capsella bursa-pastoris* and some, but not all, *Brassica* species (Long et al., 2007; Razi et al., 2008; Slotte et al., 2009; Zhao et al., 2010). A role for *FRI* in flowering time variation in *A. lyrata* and *Brassica napus* has been inferred from association studies (Kuittinen et al., 2008; Wang et al., 2011).

Strikingly, there are many alleles at both *FRI* and *FLC* (Michaels and Amasino, 1999; Johanson et al., 2000; Le Corre et al., 2002; Gazzani et al., 2003; Lempe et al., 2005; Shindo et al., 2005; Méndez-Vigo et al., 2011). Because of the convenience of early flowering, commonly used laboratory accessions have a loss-of-function allele at one or both loci. However, while low-activity *FRI* alleles typically have disrupted open reading frames, *FLC* alleles are predominantly characterized by noncoding structural variation. During vernalization, *FLC* becomes epigenetically silenced, and natural alleles differ in the duration of vernalization needed for stably switching off *FLC* expression (Shindo et al., 2006). In addition to its repressive effects on flowering, high-activity alleles of *FLC* promote germination in the cold, which in turn allows plants to experience the longer cold period required for flowering when *FLC* is active (Chiang et al., 2009). The *FRI* homologs *FRL1* and *FRL2* along with the *FLC* homologs *FLM/MAF1* and *MAF2* provide additional routes to flowering time variation (Werner et al., 2005; Schläppi, 2006; Caicedo et al., 2009; Rosloski et al., 2010).

Flowering time control is one of the most intensively investigated developmental processes in Arabidopsis, and well over 100 genes are known to affect flowering, with many having substantial pleiotropic effects on plant growth (Srikanth and Schmid, 2011). Remark-

ably, only one gene with very few nonflowering phenotypes, the central flowering activator *FT*, has been shown to contribute extensively to flowering time variation between Arabidopsis accessions (Schwartz et al., 2009; Li et al., 2010; Huang et al., 2011; Salomé et al., 2011b; Strange et al., 2011). QTL studies have implicated *FT* as being the cause of flowering time differences also in *B. napus* (Long et al., 2007).

Several other genes responsible for flowering time variation in Arabidopsis have multiple functions during plant development, including the photoreceptor encoding genes *CRYPTOCHROME2*, *PHYC*, and *PHYD* (Aukerman et al., 1997; El-Din El-Assal et al., 2001; Balasubramanian et al., 2006; Méndez-Vigo et al., 2011). In addition, there is functional allelic variation at *PHYA* and *PHYB*. Both regulate flowering (Srikanth and Schmid, 2011), although the effects of the natural alleles on flowering have not been studied (Maloof et al., 2001; Filiault et al., 2008). Two other pleiotropically acting, naturally variable flowering regulators are *FY* (Adams et al., 2009) and *HUA2*. In addition to affecting flowering time, a natural *HUA2* change-of-function allele has a dramatic effect on plant architecture that had not been anticipated from mutant studies (Alonso-Blanco et al., 1998a; Wang et al., 2007; Huang et al., 2011; Strange et al., 2011). Finally, additional loci responsible for flowering time regulation have been identified by growing plants under variable conditions (Weinig et al., 2002; Brachi et al., 2010; Li et al., 2010).

TOWARD AN UNDERSTANDING OF THE FORCES SHAPING GENETIC VARIATION

Apart from extending our knowledge of biological mechanisms and pathways in Arabidopsis, a major motivation for studying genetic variation is to understand how a species adapts to different local environments, which traces adaptation leaves in the genome, and how this leads to the formation of new species. In this section, I describe how genome analyses have provided insights into the history of the species, what is being learned about epistatic interactions between alleles from different genomes, and how evidence for local adaptation is emerging.

Geographic Distribution of Population Diversity

Until a decade ago, the vast majority of the few hundred Arabidopsis accessions available from the stock centers came from western Europe. In the past years, collections have been substantially expanded, with more than 2,000 genotypically distinct accessions having been described (Schmuths et al., 2006; Beck et al., 2008; Picó et al., 2008; Montesinos et al., 2009; Bomblies et al., 2010; Lewandowska-Sabat et al., 2010; Platt et al., 2010; Cao et al., 2011; Méndez-Vigo et al., 2011). With whole-genome data, the pattern of isolation-by-distance that had been deduced from more sparse data before came into even sharper focus. In addition, it was found that geographic regions differ greatly both with respect to the total number of polymorphisms distinguishing accessions within a region from each other and from other regions and the relative frequency of variants that are shared with other regions.

There is an overall gradient from west to east: The greatest diversity is found at the western end of the native range, in the Iberian Peninsula, including North Africa, while the most uniform regions are in Central Asia. This is consistent with the view that Arabidopsis populations in the west are the oldest, with later expansion into the eastern end of its native distribution, along with recently colonized regions, such as the Alps, in the center of the range (Sharbel et al., 2000; Nordborg et al., 2005; Schmid et al., 2005; Ostrowski et al., 2006; Beck et al., 2008; Picó et al., 2008; Platt et al., 2010; Cao et al., 2011). In addition, there is also altitudinal stratification within regions, with populations from high altitude being overall less diverse than those from lower altitude (Montesinos et al., 2009; Lewandowska-Sabat et al., 2010; Gomaa et al., 2011). It has also been suggested that there is evidence for migration from east to west, accompanying the spread of agriculture (François et al., 2008); however, knowing that the Iberian Peninsula is the most diverse region, it is unclear what to make from this. The regional differences have certainly important implications for the design of GWAS, since LD extends further in less diverse regions (Cao et al., 2011).

In continental Eurasia, identical multilocus genotypes are almost exclusively found only in the same local patches of Arabidopsis individuals (Picó et al., 2008; Bomblies et al., 2010; Lewandowska-Sabat et al., 2010; Platt et al., 2010). Exceptions are the British Isles and North America. In both regions, one specific genotype is found in many different places. For North America, recent and widespread, but uneven, introduction by European settlers has been suggested as the most likely cause; this scenario is compatible with the absence of genetic isolation by distance in North America (Platt et al., 2010).

Epistatic Interactions between Genomes

Despite its selfing nature, and contrary to what early analyses had suggested, stands of Arabidopsis plants can include several different multilocus genotypes. Moreover, outcrossing rates of Arabidopsis in nature can be several percent, and heterozygous individuals are thus not that rare (Stenøien et al., 2005; Bakker et al., 2006; Jorgensen and Emerson, 2008; Bomblies et al., 2010; Platt et al., 2010).

Superior performance in heterozygous F1 hybrids is known as heterosis or hybrid vigor. Heterosis in Arabidopsis is generally not as dramatic as in other species, but heterotic QTL for biomass and metabolites have been identified by backcrossing RILs derived from two inbred accessions to the founders (Syed and Chen, 2005; Kusterer et al., 2007; Lisec et al., 2009;

Meyer et al., 2010). There is also extensive evidence for nonadditive, or epistatic, effects on gene expression in intra- and interspecific hybrids (Wang et al., 2006; Zhang and Borevitz, 2009; Zhang et al., 2011). In both stable allotetraploids and F1 hybrids of Arabidopsis × *arenosa*, circadian gene expression programs are altered, and a similar trend is apparent in F1 hybrids between two Arabidopsis accessions that exhibit hybrid vigor. The heterotic effects are mediated by central regulators of the circadian clock (Ni et al., 2009), although the proximate causes that alter activity of these regulators, and their relationship to the heterosis QTL identified in the same cross before, remain unknown.

Inferior performance of F1 hybrids is known as hybrid weakness or incompatibility, with extreme cases presenting as hybrid sterility or lethality. In addition, a decline in fitness of later generations is called hybrid breakdown or inbreeding depression (Hochholdinger and Hoecker, 2007; Charlesworth and Willis, 2009; Bomblies, 2010) A commonly observed incompatibility phenomenon is cytoplasmic male sterility (CMS), due to a mismatch between nuclear genes that encode proteins active in mitochondria and the mitochondrial genome (Fujii and Toriyama, 2008). Despite well over 1,000 different interaccession crosses having been examined (Bomblies et al., 2007), CMS has not yet been reported in Arabidopsis, even though weak CMS has been observed in *A. lyrata* (Leppälä and Savolainen, 2011). The most common obvious defect in F1 hybrids of Arabidopsis appears to be an autoimmune syndrome, hybrid necrosis, that is also known from many other plants.

Hybrid necrosis can often be explained by one or two epistatically interacting loci (Bomblies et al., 2007; Bomblies and Weigel, 2007). At least one of the genes causal for hybrid necrosis in Arabidopsis encodes an immune receptor of the NB-LRR class (Bomblies et al., 2007), consistent with the identification of immune genes underlying hybrid necrosis in other species (Krüger et al., 2002; Jeuken et al., 2009; Yamamoto et al., 2010). The NB-LRR family is the most variable gene family in plants, with genes often being found in clusters that have a complex history of gene duplication, deletion, and gene conversion. NB-LRR genes are engaged in recognition of diverse proteins (Nishimura and Dangl, 2010), providing an intuitive explanation for why hybrid necrosis is so common. In a broader context, hybrid necrosis is a manifestation of the costs of disease resistance (Tian et al., 2003).

In some instances, hybrid necrosis becomes only expressed in the F2 generation (Alcázar et al., 2009). In one such case, one of the causal genes encodes a receptor kinase homolog, with evidence of positive selection for disease resistance having increased the frequency of this allele in Central Asia (Alcázar et al., 2010). A receptor-kinase-like gene of a different class is responsible for an incompatibility that primarily causes growth defects. This specific case involves an interaction between alleles at a single locus with similar properties as many NB-LRR loci, namely being composed of a highly variable tandem array of genes (Smith et al., 2011). Notably, not every highly variable gene family appears to cause problems in hybrids. Cytochrome P450s, which are important for plant insect defense and are produced by one of the most highly variable gene families (Clark et al., 2007; Cao et al., 2011), have so far not been tied to hybrid weakness, perhaps because they are not designed to interact with a diverse set of other proteins.

Most F2 incompatibilities were not discovered because of overt phenotypic effects but were deduced from segregation distortion, that is, the absence of certain genotypic combinations, in F2 or RIL populations (Lister and Dean, 1993; Mitchell-Olds, 1995; Alonso-Blanco et al., 1998b; Loudet et al., 2002; Werner et al., 2005; Törjék et al., 2006; Simon et al., 2008; Balasubramanian et al., 2009; Salomé et al., 2011a). For RILs, this can be due to inadvertent selection, e.g. because late-germinating lines are eliminated, but several cases are associated with lethality of specific segregants. One example involves a pair of paralogs that arose from a very recent ectopic duplication event and that independently sustained inactivating mutations in different lineages (Bikard et al., 2009). About three-quarters of accessions carry inactive copies of one or the other paralog, suggesting that increased dosage is disfavored. A similar situation of reciprocally mutated paralogs explains an epistatic interaction affecting shoot growth (Vlad et al., 2010). Both cases differ from other examples of complex duplication and mutation events, where the paralogs have become neofunctionalized and have now distinct activities (Kliebenstein et al., 2001; Kroymann et al., 2003; Huang et al., 2010a).

Experimental Ecology and Ecological Genomics

The worldwide distribution of Arabidopsis can be well described by climatic range boundaries; these indicate that laboratory conditions commonly used for growth of Arabidopsis are at the extreme end of its normal habitats, which are normally much cooler and drier (Hoffmann, 2002). This has important implications for interpreting phenotypic differences observed in the greenhouse. For example, strains with differential activity of the key flowering regulators *FRI* and *FLC*, known to vary in many accessions, only differ strongly in their flowering behavior outdoors when germinated at specific times of the year, with a critical period in early fall having a disproportionately large effect on flowering time, namely, whether plants overwinter (Wilczek et al., 2009). Such knowledge is essential if one wants to predict responses to a changing climate (Wilczek et al., 2010). Furthermore, by culturing plants in seminatural settings, in which either variable light and temperature conditions are reproduced in climate chambers or plants are germinated in the greenhouse, then transplanted outdoors, one can detect QTL that are not found when plants are grown

in a uniform environment. Whether either type of QTL is more relevant is unclear and can only be addressed by phenotyping truly naturally growing individuals. Nevertheless, analysis in seminatural conditions provides insights into the genetic basis of traits considered to be indicative of fitness, such as germination, survival, fruit and seed number, or competitiveness (Weinig et al., 2002, 2003a, 2003b; Stinchcombe et al., 2004; Donohue et al., 2005; Li et al., 2006; Brachi et al., 2010; Huang et al., 2010b; Li et al., 2010; Fournier-Level et al., 2011).

Different experimental approaches are beginning to reveal local adaptation in Arabidopsis. When 74 accessions were monitored in the greenhouse under different temperatures, it was found that accessions from cold regions respond in their growth more strongly to elevated temperatures than accessions from warm regions, which are only moderately inhibited by colder temperatures (Hoffmann et al., 2005). Systematic correlation of phenotypes with environmental gradients can indicate adaptation (Endler, 1977), and there are also latitudinal clines in light sensitivity and altitudinal clines in flowering-related traits (Maloof et al., 2001; Méndez-Vigo et al., 2011). It has been similarly proposed that populations of Arabidopsis near oceans or saline soils are more likely to carry an allele at the *HKT1* locus that increases sodium accumulation in leaves (Baxter et al., 2010). However, the accessions investigated were unevenly sampled, information about soil salinity at the places of origin was not available, and the relationship between compromised activity of *HKT1* and salt tolerance is complex (Mäser et al., 2002; Berthomieu et al., 2003). Thus, the conclusions about adaptation to salinity should be taken with the proverbial grain of salt.

Reciprocal transplantation experiments have produced evidence for local adaptation in *A. lyrata* (Leinonen et al., 2009, 2011). Somewhat surprisingly, this approach, a gold standard in ecology (Turesson, 1922a), has so far only been sparingly applied in Arabidopsis. This has recently been remedied, with an impressive study in which hundreds of accessions were grown at several different places in the native range of the species (Fournier-Level et al., 2011). Alleles associated with superior fitness at each site were most likely to be found in accessions originating near that site. GWAS identified several candidates for survival and fruit number, although only one, the photoreceptor gene *PHYB*, which affects light response, can be easily connected to local adaptation based on prior knowledge. Additional evidence for local adaptation comes from GWAS for climate variables at the place of origin combined with fitness tests at a single site (Hancock et al., 2011). Both of these studies were carried out predominantly with accessions from the western European and Scandinavian part of the native range, and it will be interesting to repeat these experiments with a broader spectrum of accessions and test locales.

OUTLOOK

Our knowledge of natural variation in Arabidopsis has advanced tremendously in the past decade, with an impressive set of genetic and genomic approaches and resources that are now available (Fig. 6). In the near future, the simultaneous application of different strategies will lead to genetic variation increasingly informing basic plant biology. Combined analyses of global transcript and metabolite levels and biomass across accessions and RIL populations is supporting the reconstruction of functional networks (Wentzell et al., 2007; Lisec et al., 2008; Rowe et al., 2008; Sulpice et al., 2009, 2010). Integration of QTL data with such information has shown that in addition to biosynthetic and metabolic enzymes, upstream transcription factors of the MYB class contribute to diversity in glucosinolate content (Sønderby et al., 2007) and that the clock gene *ELF3* has a role in shade avoidance (Jiménez-Gómez et al., 2010). Another instructive example of how natural variation can help to discover a new regulatory pathway comes from the study of xylem expansion (Sibout et al., 2008). The authors noted that the xylem expansion loci colocalized with flowering time QTL, which led them to hypothesize that the onset of flowering causes xylem expansion in both the shoot and the root. They subsequently confirmed such a model by transiently inducing the activity of a central floral regulator. There is similarly great promise in GWAS with the same material to identify cases of pleiotropic action of natural sequence variants.

I have also highlighted the many opportunities Arabidopsis offers for the study of interactions between divergent genomes, which may both promote or reduce outcrossing, and thereby affect the partitioning

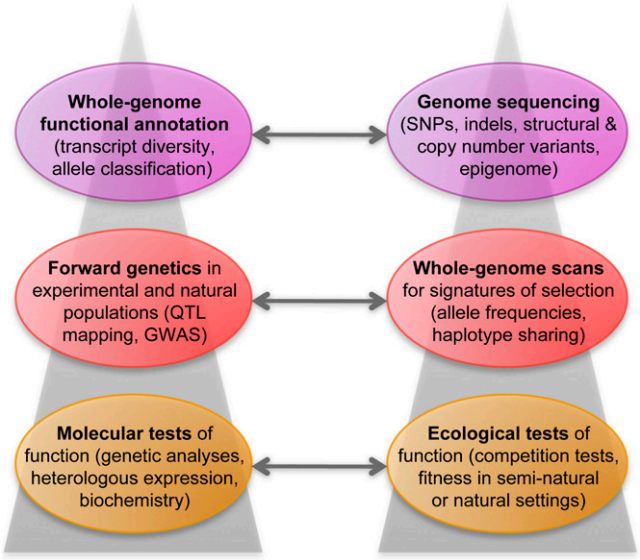

Figure 6. Relationship between approaches to the study of genetic variation.

of genetic diversity into different lineages (and ultimately into different species). So far, the parents for the investigated crosses have largely been chosen randomly. With increasing information about the genome-wide and population-specific distribution of sequence polymorphisms, more judicious and systematic choices of genotype combinations should accelerate the pace with which we can obtain insights into the fascinating questions of hybrid performance.

Another important direction will be to phenotype naturally growing plants in situ over several years (Montesinos et al., 2009). Genotyping of very large numbers of wild plants has become very affordable with next-generation sequencing methods, which will facilitate linking genotype and phenotype even on an individual basis (Baird et al., 2008; Elshire et al., 2011). An example for such strategies is a study that monitored over 4 years the load of five different viruses that had been known before to infect wild Brassicaceae (Pagán et al., 2010). Such experiments are required to test claims about fitness trade-offs between disease resistance and growth (Tian et al., 2003; Todesco et al., 2010). Finally, selection experiments are a tool that should not be underestimated for their potential to provide insights into favorable allele combinations (Ungerer et al., 2003; Ungerer and Rieseberg, 2003; Scarcelli and Kover, 2009; Fakheran et al., 2010).

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Table S1. Full references for Table 1.

ACKNOWLEDGMENTS

I thank Eunyoung Chae, Sang-Tae Kim, and George Wang for plant images; Joy Bergelson, Carlos Alonso-Blanco, Jun Cao, Karl Schmid, and George Wang for help in producing the map of Arabidopsis accessions; and Annie Schmitt and Joy Bergelson for preprints. I am especially grateful to three anonymous reviewers, who provided insightful comments and helped to correct several oversights in the original manuscript.

Received October 24, 2011; accepted December 5, 2011; published December 6, 2011.

LITERATURE CITED

Adams S, Allen T, Whitelam GC (2009) Interaction between the light quality and flowering time pathways in Arabidopsis. Plant J **60:** 257–267

Alcázar R, García AV, Kronholm I, de Meaux J, Koornneef M, Parker JE, Reymond M (2010) Natural variation at Strubbelig Receptor Kinase 3 drives immune-triggered incompatibilities between *Arabidopsis thaliana* accessions. Nat Genet **42:** 1135–1139

Alcázar R, García AV, Parker JE, Reymond M (2009) Incremental steps toward incompatibility revealed by *Arabidopsis* epistatic interactions modulating salicylic acid pathway activation. Proc Natl Acad Sci USA **106:** 334–339

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

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

Alonso-Blanco C, Peeters AJ, Koornneef M, Lister C, Dean C, van den Bosch N, Pot J, Kuiper MT (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

Al-Shehbaz I, O'Kane S Jr (2002) Taxonomy and phylogeny of *Arabidopsis* (Brassicaceae). The Arabidopsis Book **1:** 1–22, doi/10.1199/tab.001

Anastasio AE, Platt A, Horton M, Grotewold E, Scholl R, Borevitz JO, Nordborg M, Bergelson J (2011) Source verification of mis-identified *Arabidopsis thaliana* accessions. Plant J **67:** 554–566

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

Aranzana MJ, 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

Arrieta-Montiel MP, Shedge V, Davila J, Christensen AC, Mackenzie SA (2009) Diversity of the *Arabidopsis* mitochondrial genome occurs via nuclear-controlled recombination activity. Genetics **183:** 1261–1268

Asimit J, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annu Rev Genet **44:** 293–308

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, et al (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature **465:** 627–631

Aukerman MJ, Hirschfeld M, Wester L, Weaver M, Clack T, Amasino RM, Sharrock RA (1997) A deletion in the *PHYD* gene of the *Arabidopsis* Wassilewskija ecotype defines a role for phytochrome D in red/far-red light sensing. Plant Cell **9:** 1317–1326

Bailey DW (1971) Recombinant-inbred strains. An aid to finding identity, linkage, and function of histocompatibility and other genes. Transplantation **11:** 325–327

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE **3:** e3376

Bakker EG, Stahl EA, Toomajian C, Nordborg M, Kreitman M, Bergelson J (2006) Distribution of genetic variation within and among local populations of *Arabidopsis thaliana* over its species range. Mol Ecol **15:** 1405–1418

Balasubramanian S, Schwartz C, Singh A, Warthmann N, Kim MC, Maloof JN, Loudet O, Trainer GT, Dabi T, Borevitz JO, et al (2009) QTL mapping in new *Arabidopsis thaliana* advanced intercross-recombinant inbred lines. PLoS ONE **4:** e4318

Balasubramanian S, Sureshkumar S, Agrawal M, Michael TP, Wessinger C, Maloof JN, Clark R, Warthmann N, Chory J, Weigel D (2006) The *PHYTOCHROME C* photoreceptor gene mediates natural variation in flowering and growth responses of *Arabidopsis thaliana*. Nat Genet **38:** 711–715

Baxter I, Brazelton JN, Yu D, Huang YS, Lahner B, Yakubova E, Li Y, Bergelson J, Borevitz JO, Nordborg M, et al (2010) A coastal cline in sodium accumulation in *Arabidopsis thaliana* is driven by natural variation of the sodium transporter AtHKT1;1. PLoS Genet **6:** e1001193

Beck JB, Schmuths H, Schaal BA (2008) Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. Mol Ecol **17:** 902–915

Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, Weigel D (September 20, 2011) Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature **480:** 245–249

Benderoth M, Textor S, Windsor AJ, Mitchell-Olds T, Gershenzon J, Kroymann J (2006) Positive selection driving diversification in plant secondary metabolism. Proc Natl Acad Sci USA **103:** 9118–9123

Bentsink L, Hanson J, Hanhart CJ, Blankestijn-de Vries H, Coltrane C, Keizer P, El-Lithy M, Alonso-Blanco C, de Andrés MT, Reymond M, et al (2010) Natural variation for seed dormancy in *Arabidopsis* is regulated by additive genetic and molecular pathways. Proc Natl Acad Sci USA **107:** 4264–4269

Bentsink L, Jowett J, Hanhart CJ, Koornneef M (2006) Cloning of *DOG1*, a

quantitative trait locus controlling seed dormancy in *Arabidopsis*. Proc Natl Acad Sci USA **103:** 17042–17047

Bergelson J, Roux F (2010) Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. Nat Rev Genet **11:** 867–879

Berthomieu P, Conéjéro G, Nublat A, Brackenbury WJ, Lambert C, Savio C, Uozumi N, Oiki S, Yamada K, Cellier F, et al (2003) Functional analysis of *AtHKT1* in *Arabidopsis* shows that Na^+ recirculation by the phloem is crucial for salt tolerance. EMBO J **22:** 2004–2014

Bikard D, Patel D, Le Metté C, Giorgi V, Camilleri C, Bennett MJ, Loudet O (2009) Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. Science **323:** 623–626

Blau PA, Feeny P, Contardo L, Robson DS (1978) Allylglucosinolate and herbivorous caterpillars: a contrast in toxicity and tolerance. Science **200:** 1296–1298

Bomblies K (2010) Doomed lovers: mechanisms of isolation and incompatibility in plants. Annu Rev Plant Biol **61:** 109–124

Bomblies K, Lempe J, Epple P, Warthmann N, Lanz C, Dangl JL, Weigel D (2007) Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. PLoS Biol **5:** e236

Bomblies K, Weigel D (2007) Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. Nat Rev Genet **8:** 382–393

Bomblies K, Yant L, Laitinen RA, Kim ST, Hollister JD, Warthmann N, Fitz J, Weigel D (2010) Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. PLoS Genet **6:** e1000890

Borevitz JO, Hazen SP, Michael TP, Morris GP, Baxter IR, Hu TT, Chen H, Werner JD, Nordborg M, Salt DE, et al (2007) Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*. Proc Natl Acad Sci USA **104:** 12057–12062

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet (Suppl) **33:** 228–237

Bowman JL, Alvarez J, Weigel D, Meyerowitz EM, Smyth DR (1993) Control of flower development in *Arabidopsis thaliana* by *APETALA1* and interacting genes. Development **119:** 721–743

Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M, Bergelson J, Cuguen J, Roux F (2010) Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. PLoS Genet **6:** e1000940

Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J, Bharti AK, Woodward JE, May GD, Gentzbittel L, et al (2011) Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. Proc Natl Acad Sci USA **108:** E864–E870

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, et al (2009) The genetic architecture of maize flowering time. Science **325:** 714–718

Caicedo AL, Richards C, Ehrenreich IM, Purugganan MD (2009) Complex rearrangements lead to novel chimeric gene fusion polymorphisms at the *Arabidopsis thaliana MAF2-5* flowering time gene cluster. Mol Biol Evol **26:** 699–711

Caicedo AL, Schaal BA, Kunkel BN (1999) Diversity and molecular evolution of the *RPS2* resistance gene in *Arabidopsis thaliana*. Proc Natl Acad Sci USA **96:** 302–306

Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD (2004) Epistatic interaction between *Arabidopsis FRI* and *FLC* flowering time genes generates a latitudinal cline in a life history trait. Proc Natl Acad Sci USA **101:** 15670–15675

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet **43:** 956–963

Cetl I, Dobrovolná J, Effmertova E (1968) The developmental character of natural populations of *Arabidopsis thaliana* (L.) Heynh in relation to the geographical-climatic conditions of localities. Folia Fac Sci Nat Univ Purk Brun (Biol.) **18:** 37–49

Chan EK, Rowe HC, Corwin JA, Joseph B, Kliebenstein DJ (2011) Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*. PLoS Biol **9:** e1001125

Chang C, Bowman JL, DeJohn AW, Lander ES, Meyerowitz EM (1988) Restriction fragment length polymorphism linkage map for *Arabidopsis thaliana*. Proc Natl Acad Sci USA **85:** 6856–6860

Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. Nat Rev Genet **10:** 783–796

Chiang GC, Barua D, Kramer EM, Amasino RM, Donohue K (2009) Major flowering time gene, flowering locus C, regulates seed germination in *Arabidopsis thaliana*. Proc Natl Acad Sci USA **106:** 11661–11666

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, et al (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science **317:** 338–342

Clark SS, Crist WM, Witte ON (1989) Molecular pathogenesis of Ph-positive leukemias. Annu Rev Med **40:** 113–122

Clarke JH, Dean C (1994) Mapping *FRI*, a locus controlling flowering time and vernalization response in *Arabidopsis thaliana*. Mol Gen Genet **242:** 81–89

Clarke JH, Mithen R, Brown JKM, Dean C (1995) QTL analysis of flowering time in *Arabidopsis thaliana*. Mol Gen Genet **248:** 278–286

Clough SJ, Bent AF (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. Plant J **16:** 735–743

Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet **10:** 392–404

Cubas P, Vincent C, Coen E (1999) An epigenetic mutation responsible for natural variation in floral symmetry. Nature **401:** 157–161

Darvasi A, Soller M (1995) Advanced intercross lines, an experimental population for fine genetic mapping. Genetics **141:** 1199–1207

Davila JI, Arrieta-Montiel MP, Wamboldt Y, Cao J, Hagmann J, Shedge V, Xu YZ, Weigel D, Mackenzie SA (2011) Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. BMC Biol **9:** 64

Davison J, Tyagi A, Comai L (2007) Large-scale polymorphism of heterochromatic repeats in the DNA of *Arabidopsis thaliana*. BMC Plant Biol **7:** 44

Deng W, Ying H, Helliwell CA, Taylor JM, Peacock WJ, Dennis ES (2011) FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of *Arabidopsis*. Proc Natl Acad Sci USA **108:** 6680–6685

Donohue K, Dorn L, Griffith C, Kim E, Aguilera A, Polisetty CR, Schmitt J (2005) Environmental and genetic influences on the germination of *Arabidopsis thaliana* in the field. Evolution **59:** 740–757

Doyle MR, Bizzell CM, Keller MR, Michaels SD, Song J, Noh YS, Amasino RM (2005) *HUA2* is required for the expression of floral repressors in *Arabidopsis thaliana*. Plant J **41:** 376–385

Ehrenreich IM, Hanzawa Y, Chou L, Roe JL, Kover PX, Purugganan MD (2009) Candidate gene association mapping of *Arabidopsis* flowering time. Genetics **183:** 325–335

El-Din El-Assal S, Alonso-Blanco C, Peeters AJ, Raz V, Koornneef M (2001) A QTL for flowering time in *Arabidopsis* reveals a novel allele of *CRY2*. Nat Genet **29:** 435–440

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE **6:** e19379

Endler JA (1977) Geographic Variation, Speciation, and the Clines. Princeton University Press, Princeton, NJ

Eshed Y, 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

Fakheran S, Paul-Victor C, Heichinger C, Schmid B, Grossniklaus U, Turnbull LA (2010) Adaptation and extinction in experimentally fragmented landscapes. Proc Natl Acad Sci USA **107:** 19120–19125

Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics, Ed 4. Addison Wesley Longman, Harlow, Essex, UK

Filiault DL, Wessinger CA, Dinneny JR, Lutes J, Borevitz JO, Weigel D, Chory J, Maloof JN (2008) Amino acid polymorphisms in *Arabidopsis* phytochrome B cause differential responses to light. Proc Natl Acad Sci USA **105:** 3157–3162

Flowers JM, Hanzawa Y, Hall MC, Moore RC, Purugganan MD (2009) Population genomics of the *Arabidopsis thaliana* flowering time gene network. Mol Biol Evol **26:** 2475–2486

Forner J, Weber B, Wiethölter C, Meyer RC, Binder S (2005) Distant sequences determine 5′ end formation of cox3 transcripts in *Arabidopsis thaliana* ecotype C24. Nucleic Acids Res **33:** 4673–4682

Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM (2011) A map of local adaptation in *Arabidopsis thaliana*. Science **334:** 86–89

François O, Blum MG, Jakobsson M, Rosenberg NA (2008) Demographic

history of european populations of *Arabidopsis thaliana*. PLoS Genet **4:** e1000075

Fransz PF, Armstrong S, de Jong JH, Parnell LD, van Drunen C, Dean C, Zabel P, Bisseling T, Jones GH (2000) Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: structural organization of heterochromatic knob and centromere region. Cell **100:** 367–376

Fujii S, Toriyama K (2008) Genome barriers between nuclei and mitochondria exemplified by cytoplasmic male sterility. Plant Cell Physiol **49:** 1484–1494

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, et al (2011) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature **477:** 419–423

Gazzani S, Gendall AR, Lister C, Dean C (2003) Analysis of the molecular basis of flowering time variation in *Arabidopsis* accessions. Plant Physiol **132:** 1107–1114

Gomaa NH, Montesinos-Navarro A, Alonso-Blanco C, Picó FX (2011) Temporal variation in genetic diversity and effective population size of Mediterranean and subalpine *Arabidopsis thaliana* populations. Mol Ecol **20:** 3540–3554

Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, et al (2009) A first-generation haplotype map of maize. Science **326:** 1115–1117

Grant MR, Godiard L, Straube E, Ashfield T, Lewald J, Sattler A, Innes RW, Dangl JL (1995) Structure of the *Arabidopsis RPM1* gene enabling dual specificity disease resistance. Science **269:** 843–846

Groszmann M, Greaves IK, Albertyn ZI, Scofield GN, Peacock WJ, Dennis ES (2011) Changes in 24-nt siRNA levels in *Arabidopsis* hybrids suggest an epigenetic contribution to hybrid vigor. Proc Natl Acad Sci USA **108:** 2617–2622

Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperone FG, Toomajian C, Roux F, Bergelson J (2011) Adaptation to climate across the *Arabidopsis thaliana* genome. Science **334:** 83–86

Hauser MT, Harr B, Schlötterer C (2001) Trichome distribution in *Arabidopsis thaliana* and its close relative *Arabidopsis lyrata*: molecular analysis of the candidate gene *GLABROUS1*. Mol Biol Evol **18:** 1754–1763

Heidel AJ, Clauss MJ, Kroymann J, Savolainen O, Mitchell-Olds T (2006) Natural variation in *MAM* within and between populations of *Arabidopsis lyrata* determines glucosinolate phenotype. Genetics **173:** 1629–1636

Hilscher J, Schlötterer C, Hauser MT (2009) A single amino acid replacement in ETC2 shapes trichome patterning in natural *Arabidopsis* populations. Curr Biol **19:** 1747–1751

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

Hoffmann MH (2002) Biogeography of *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae). J Biogeogr **29:** 125–134

Hoffmann MH, Tomiuk J, Schmuths H, Koch C, Bachmann K (2005) Phenological and morphological responses to different temperature treatments differ among a world-wide sample of accessions of *Arabidopsis thaliana*. Acta Oecol **28:** 181–187

Hollick JB, Patterson GI, Asmundsson IM, Chandler VL (2000) Paramutation alters regulatory control of the maize *pl* locus. Genetics **154:** 1827–1838

Horton M, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, Muliyati W, Platt A, Sperone FG, Vilhkálmsson BJ, et al (2012) Genome-wide pattern of genetic variation in worldwide *Arabidopsis thaliana* accessions from the *RegMap* panel. Nat Genet (in press)

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet **43:** 476–481

Huang M, Abel C, Sohrabi R, Petri J, Haupt I, Cosimano J, Gershenzon J, Tholl D (2010a) Variation of herbivore-induced volatile terpenes among Arabidopsis ecotypes depends on allelic differences and subcellular targeting of two terpene synthases, TPS02 and TPS03. Plant Physiol **153:** 1293–1310

Huang X, Paulo MJ, Boer M, Effgen S, Keizer P, Koornneef M, van Eeuwijk FA (2011) Analysis of natural allelic variation in *Arabidopsis* using a multiparent recombinant inbred line population. Proc Natl Acad Sci USA **108:** 4488–4493

Huang X, Schmitt J, Dorn L, Griffith C, Effgen S, Takao S, Koornneef M, Donohue K (2010b) The earliest stages of adaptation in an experimental plant population: strong selection on QTLs for seed dormancy. Mol Ecol **19:** 1335–1351

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, et al (2010c) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet **42:** 961–967

Ishida T, Kurata T, Okada K, Wada T (2008) A genetic regulatory network in the development of trichomes and root hairs. Annu Rev Plant Biol **59:** 365–386

Ito H, Miura A, Takashima K, Kakutani T (2007) Ecotype-specific and chromosome-specific expansion of variant centromeric satellites in *Arabidopsis thaliana*. Mol Genet Genomics **277:** 23–30

Jacobsen SE, Meyerowitz EM (1997) Hypermethylated SUPERMAN epigenetic alleles in *Arabidopsis*. Science **277:** 1100–1103

James SA, O'Kelly MJ, Carter DM, Davey RP, van Oudenaarden A, Roberts IN (2009) Repetitive sequence variation and dynamics in the ribosomal DNA array of *Saccharomyces cerevisiae* as revealed by whole-genome resequencing. Genome Res **19:** 626–635

Jeuken MJ, Zhang NW, McHale LK, Pelgrom K, den Boer E, Lindhout P, Michelmore RW, Visser RG, Niks RE (2009) *Rin4* causes hybrid necrosis and race-specific resistance in an interspecific lettuce hybrid. Plant Cell **21:** 3368–3378

Jimenez-Gomez JM, Corwin JA, Joseph B, Maloof JN, Kliebenstein DJ (2011) Genomic analysis of QTLs and genes altering natural variation in stochastic noise. PLoS Genet **7:** e1002295

Jiménez-Gómez JM, Wallace AD, Maloof JN (2010) Network analysis identifies *ELF3* as a QTL for the shade avoidance response in Arabidopsis. PLoS Genet **6:** e1001100

Johannes F, Colomé-Tatché M (2011) Quantitative epigenetics through epigenomic perturbation of isogenic lines. Genetics **188:** 215–227

Johanson U, West J, Lister C, Michaels S, Amasino R, Dean C (2000) Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. Science **290:** 344–347

Jones ME (1971) The population genetics of *Arabidopsis thaliana*. III. The effect of vernalisation. Heredity **27:** 59–72

Jorde LB (2000) Linkage disequilibrium and the search for complex disease genes. Genome Res **10:** 1435–1444

Jorgensen TH, Emerson BC (2008) Functional variation in a disease resistance gene in populations of *Arabidopsis thaliana*. Mol Ecol **17:** 4912–4923

Kam-Thong T, Pütz B, Karbalai N, Müller-Myhsok B, Borgwardt K (2011) Epistasis detection on quantitative phenotypes by exhaustive enumeration using GPUs. Bioinformatics **27:** i214–i221

Kärkkäinen K, Ågren J (2002) Genetic basis of trichome production in *Arabidopsis lyrata*. Hereditas **136:** 219–226

Kawagoe T, Shimizu KK, Kakutani T, Kudoh H (2011) Coexistence of trichome variation in a natural plant population: a combined study using ecological and candidate gene approaches. PLoS ONE **6:** e22184

Kempin SA, Savidge B, Yanofsky MF (1995) Molecular basis of the *cauliflower* phenotype in *Arabidopsis*. Science **267:** 522–525

Kerwin RE, Jimenez-Gomez JM, Fulop D, Harmer SL, Maloof JN, Kliebenstein DJ (2011) Network quantitative trait loci mapping of circadian clock outputs identifies metabolic pathway-to-clock linkages in *Arabidopsis*. Plant Cell **23:** 471–485

Keurentjes JJ, Bentsink L, Alonso-Blanco C, Hanhart CJ, Blankestijn-De Vries H, Effgen S, Vreugdenhil D, Koornneef M (2007) 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

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet **39:** 1151–1155

Kivimäki M, Kärkkäinen K, Gaudeul M, Løe G, Ågren J (2007) Gene, phenotype and function: *GLABROUS1* and resistance to herbivory in natural populations of *Arabidopsis lyrata*. Mol Ecol **16:** 453–462

Kliebenstein DJ, Kroymann J, Mitchell-Olds T (2005) The glucosinolate-myrosinase system in an ecological and evolutionary context. Curr Opin Plant Biol **8:** 264–271

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

Koornneef M, Alonso-Blanco C, Vreugdenhil D (2004) Naturally occur-

ring genetic variation in *Arabidopsis thaliana*. Annu Rev Plant Biol **55:** 141–172

Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R (2009) A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. PLoS Genet **5:** e1000551

Kowalski SP, Lan TH, Feldmann KA, Paterson AH (1994) QTL mapping of naturally-occurring variation in flowering time of *Arabidopsis thaliana*. Mol Gen Genet **245:** 548–555

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

Kroymann J, Textor S, Tokuhisa JG, Falk KL, Bartram S, Gershenzon J, 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

Krüger J, Thomas CM, Golstein C, Dixon MS, Smoker M, Tang S, Mulder L, Jones JD (2002) A tomato cysteine protease required for Cf-2-dependent disease resistance and suppression of autonecrosis. Science **296:** 744–747

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet **22:** 139–144

Kuittinen H, Niittyvuopio A, Rinne P, Savolainen O (2008) Natural variation in *Arabidopsis lyrata* vernalization requirement conferred by a *FRIGIDA* indel polymorphism. Mol Biol Evol **25:** 319–329

Kuo HF, Olsen KM, Richards EJ (2006) Natural variation in a subtelomeric region of Arabidopsis: implications for the genomic dynamics of a chromosome end. Genetics **173:** 401–417

Kusterer B, Piepho HP, Utz HF, Schön CC, Muminovic J, Meyer RC, Altmann T, Melchinger AE (2007) Heterosis for biomass-related traits in Arabidopsis investigated by quantitative trait loci analysis of the triple testcross design with recombinant inbred lines. Genetics **177:** 1839–1850

Laibach F (1943) *Arabidopsis thaliana* (L.)Heynh. als Objekt für genetische und entwicklungsphysiologische Untersuchungen. Bot Arch **4:** 439–445

Laibach F (1951) Über sommer- und winterannuelle Rassen von *Arabidopsis thaliana* (L.) Heynh. Ein Beitrag zur Ätiologie der Blütenbildung. Beitr Biol Pflanzen **28:** 173–210

Laitinen RA, Schneeberger K, Jelly NS, Ossowski S, Weigel D (2010) Identification of a spontaneous frame shift mutation in a nonreference *Arabidopsis* accession using whole genome sequencing. Plant Physiol **153:** 652–654

Lambrix V, Reichelt M, Mitchell-Olds T, Kliebenstein DJ, Gershenzon J (2001) The *Arabidopsis* epithiospecifier protein promotes the hydrolysis of glucosinolates to nitriles and influences *Trichoplusia ni* herbivory. Plant Cell **13:** 2793–2807

Lander ES (1996) The new genomics: global views of biology. Science **274:** 536–539

Larkin JC, Young N, Prigge M, Marks MD (1996) The control of trichome spacing and number in *Arabidopsis*. Development **122:** 997–1005

Le Corre V, Roux F, Reboud X (2002) DNA polymorphism at the *FRIGIDA* gene in *Arabidopsis thaliana*: Extensive nonsynonymous variation is consistent with local selection for flowering time. Mol Biol Evol **19:** 1261–1271

Lee I, Bleecker A, Amasino R (1993) Analysis of naturally occurring late flowering in *Arabidopsis thaliana*. Mol Gen Genet **237:** 171–176

Leinonen PH, Remington DL, Savolainen O (2011) Local adaptation, phenotypic differentiation, and hybrid fitness in diverged natural populations of *Arabidopsis lyrata*. Evolution **65:** 90–107

Leinonen PH, Sandring S, Quilot B, Clauss MJ, Mitchell-Olds T, Ågren J, Savolainen O (2009) Local adaptation in European populations of *Arabidopsis lyrata* (Brassicaceae). Am J Bot **96:** 1129–1137

Lempe J, Balasubramanian S, Sureshkumar S, Singh A, Schmid M, Weigel D (2005) Diversity of flowering responses in wild *Arabidopsis thaliana* strains. PLoS Genet **1:** 109–118

Leppälä J, Savolainen O (2011) Nuclear-cytoplasmic interactions reduce male fertility in hybrids of *Arabidopsis lyrata* subspecies. Evolution **65:** 2959–2972

Lewandowska-Sabat AM, Fjellheim S, Rognli OA (2010) Extremely low genetic variability and highly structured local populations of *Arabidopsis thaliana* at higher latitudes. Mol Ecol **19:** 4753–4764

Li D, Liu C, Shen L, Wu Y, Chen H, Robertson M, Helliwell CA, Ito T, Meyerowitz E, Yu H (2008) A repressor complex governs the integration of flowering signals in *Arabidopsis*. Dev Cell **15:** 110–120

Li G, Quiros CF (2003) In planta side-chain glucosinolate modification in Arabidopsis by introduction of dioxygenase Brassica homolog BoGSL-ALK. Theor Appl Genet **106:** 1116–1121

Li Y, Huang Y, Bergelson J, Nordborg M, Borevitz JO (2010) Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. Proc Natl Acad Sci USA **107:** 21199–21204

Li Y, Roycewicz P, Smith E, Borevitz JO (2006) Genetics of local adaptation in the laboratory: flowering time quantitative trait loci under geographic and seasonal conditions in *Arabidopsis*. PLoS ONE **1:** e105

Lisec J, Meyer RC, Steinfath M, Redestig H, Becher M, Witucka-Wall H, Fiehn O, Törjék O, Selbig J, Altmann T, et al (2008) Identification of metabolic and biomass QTL in *Arabidopsis thaliana* in a parallel analysis of RIL and IL populations. Plant J **53:** 960–972

Lisec J, Steinfath M, Meyer RC, Selbig J, Melchinger AE, Willmitzer L, Altmann T (2009) Identification of heterotic metabolite QTL in *Arabidopsis thaliana* RIL and IL populations. Plant J **59:** 777–788

Lister C, Dean C (1993) Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. Plant J **4:** 745–750

Long AD, Lyman RF, Langley CH, Mackay TF (1998) Two sites in the Delta gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. Genetics **149:** 999–1017

Long Y, Shi J, Qiu D, Li R, Zhang C, Wang J, Hou J, Zhao J, Shi L, Park BS, et al (2007) Flowering time quantitative trait loci analysis of oilseed brassica in multiple environments and genomewide alignment with Arabidopsis. Genetics **177:** 2433–2444

Loudet O, Chaillou S, Camilleri C, Bouchez D, Daniel-Vedele F (2002) Bay-0 x Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in *Arabidopsis*. Theor Appl Genet **104:** 1173–1184

Loudet O, Michael TP, Burger BT, Le Metté C, Mockler TC, Weigel D, Chory J (2008) A zinc knuckle protein that negatively controls morning-specific growth in *Arabidopsis thaliana*. Proc Natl Acad Sci USA **105:** 17193–17198

Mackay TF (2001) The genetic architecture of quantitative traits. Annu Rev Genet **35:** 303–339

Maloof JN, Borevitz JO, Dabi T, Lutes J, Nehring RB, Redfern JL, Trainer GT, Wilson JM, Asami T, Berry CC, et al. (2001) Natural variation in light sensitivity of *Arabidopsis*. Nat Genet **29:** 441–446

Manning K, Tör M, Poole M, Hong Y, Thompson AJ, King GJ, Giovannoni JJ, Seymour GB (2006) A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. Nat Genet **38:** 948–952

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al (2009) Finding the missing heritability of complex diseases. Nature **461:** 747–753

Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet **37:** 413–417

Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, Pitrat M, Dogimont C, Bendahmane A (2009) A transposon-induced epigenetic change leads to sex determination in melon. Nature **461:** 1135–1138

Mäser P, Eckelman B, Vaidyanathan R, Horie T, Fairbairn DJ, Kubo M, Yamagami M, Yamaguchi K, Nishimura M, Uozumi N, et al (2002) Altered shoot/root Na^+ distribution and bifurcating salt sensitivity in *Arabidopsis* by genetic disruption of the Na^+ transporter *AtHKT1*. FEBS Lett **531:** 157–161

Mauricio R (1998) Costs of resistance to natural enemies in field populations of the annual plant *Arabidopsis thaliana*. Am Nat **151:** 20–28

Mauricio R, Rausher MD (1997) Experimental manipulation of putative selective agents provides evidence for the role of natural enemies in the evolution of plant defense. Evolution **51:** 1435–1444

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet **9:** 356–369

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, et al (2009) Genetic properties of the maize nested association mapping population. Science **325:** 737–740

Melquist S, Bender J (2003) Transcription from an upstream promoter

controls methylation signaling from an inverted repeat of endogenous genes in *Arabidopsis*. Genes Dev **17:** 2036–2047

Méndez-Vigo B, Picó FX, Ramiro M, Martínez-Zapater JM, Alonso-Blanco C (2011) Altitudinal and climatic adaptation is mediated by flowering traits and *FRI*, *FLC*, and *PHYC* genes in Arabidopsis. Plant Physiol **157:** 1942–1955

Metzker ML (2010) Sequencing technologies: the next generation. Nat Rev Genet **11:** 31–46

Meyer RC, Kusterer B, Lisec J, Steinfath M, Becher M, Scharr H, Melchinger AE, Selbig J, Schurr U, Willmitzer L, et al (2010) QTL analysis of early stage heterosis for biomass in Arabidopsis. Theor Appl Genet **120:** 227–237

Michaels SD, Amasino RM (1999) *FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell **11:** 949–956

Michaels SD, He Y, Scortecci KC, Amasino RM (2003) Attenuation of *FLOWERING LOCUS C* activity as a mechanism for the evolution of summer-annual flowering behavior in *Arabidopsis*. Proc Natl Acad Sci USA **100:** 10102–10107

Mitchell-Olds T (1995) Interval mapping of viability loci causing heterosis in Arabidopsis. Genetics **140:** 1105–1109

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

Montesinos A, Tonsor SJ, Alonso-Blanco C, Picó FX (2009) Demographic and genetic patterns of variation among populations of *Arabidopsis thaliana* from contrasting native environments. PLoS ONE **4:** e7213

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell **21:** 2194–2202

Nam HG, Giraudat J, Den Boer B, Moonan F, Loos WD, Hauge BM, Goodman HM (1989) Restriction fragment length polymorphism linkage map of *Arabidopsis thaliana*. Plant Cell **1:** 699–705

Napp-Zinn K (1957) Untersuchungen über das Vernalisationsverhalten einer winterannuellen Rasse von *Arabidopsis thaliana* (L.) Heynh. Planta **50:** 177–210

Nemri A, Atwell S, Tarone AM, Huang YS, Zhao K, Studholme DJ, Nordborg M, Jones JD (2010) Genome-wide survey of *Arabidopsis* natural variation in downy mildew resistance using combined association and linkage mapping. Proc Natl Acad Sci USA **107:** 10302–10307

Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet **7:** 61–80

Ni Z, Kim ED, Ha M, Lackey E, Liu J, Zhang Y, Sun Q, Chen ZJ (2009) Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. Nature **457:** 327–331

Nishimura MT, Dangl JL (2010) Arabidopsis and the plant immune system. Plant J **61:** 1053–1066

Noël L, Moores TL, van Der Biezen EA, Parniske M, Daniels MJ, Parker JE, Jones JD (1999) Pronounced intraspecific haplotype divergence at the *RPP5* complex disease resistance locus of *Arabidopsis*. Plant Cell **11:** 2099–2112

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, et al (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet **30:** 190–193

Nordborg M, Hu TT, 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

Nordborg M, Weigel D (2008) Next-generation genetics in plants. Nature **456:** 720–723

Olsen KM, Halldorsdottir SS, Stinchcombe JR, Weinig C, Schmitt J, Purugganan MD (2004) Linkage disequilibrium mapping of *Arabidopsis CRY2* flowering time alleles. Genetics **167:** 1361–1369

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D (2008a) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res **18:** 2024–2033

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science **327:** 92–94

Ossowski S, Schwab R, Weigel D (2008b) Gene silencing in plants using artificial microRNAs and other small RNAs. Plant J **53:** 674–690

Ostrowski MF, David J, Santoni S, McKhann H, Reboud X, Le Corre V, Camilleri C, Brunel D, Bouchez D, Faure B, et al. (2006) Evidence for a large-scale population structure among accessions of *Arabidopsis thaliana*: possible causes and consequences for the distribution of linkage disequilibrium. Mol Ecol **15:** 1507–1517

Pagán I, Fraile A, Fernandez-Fueyo E, Montes N, Alonso-Blanco C, García-Arenal F (2010) *Arabidopsis thaliana* as a model for the study of plant-virus co-evolution. Philos Trans R Soc Lond B Biol Sci **365:** 1983–1995

Palatnik JF, Allen E, Wu X, Schommer C, Schwab R, Carrington JC, Weigel D (2003) Control of leaf morphogenesis by microRNAs. Nature **425:** 257–263

Picó FX, Méndez-Vigo B, Martínez-Zapater JM, Alonso-Blanco C (2008) Natural genetic variation of *Arabidopsis thaliana* is geographically structured in the Iberian peninsula. Genetics **180:** 1009–1021

Plantegenet S, Weber J, Goldstein DR, Zeller G, Nussbaumer C, Thomas J, Weigel D, Harshman K, Hardtke CS (2009) Comprehensive analysis of *Arabidopsis* expression level polymorphisms with simple inheritance. Mol Syst Biol **5:** 242

Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, Agren J, Bossdorf O, Byers D, Donohue K, et al (2010) The scale of population structure in *Arabidopsis thaliana*. PLoS Genet **6:** e1000843

Ravi M, Chan SW (2010) Haploid plants produced by centromere-mediated genome elimination. Nature **464:** 615–618

Razi H, Howell EC, Newbury HJ, Kearsey MJ (2008) Does sequence polymorphism of *FLC* paralogues underlie flowering time QTL in *Brassica oleracea*? Theor Appl Genet **116:** 179–192

Reinders J, Wulff BB, Mirouze M, Marí-Ordóñez A, Dapp M, Rozhon W, Bucher E, Theiler G, Paszkowski J (2009) Compromised stability of DNA methylation and transposon immobilization in mosaic *Arabidopsis* epigenomes. Genes Dev **23:** 939–950

Reiter RS, Williams JG, Feldmann KA, Rafalski JA, Tingey SV, Scolnik PA (1992) Global and local genome mapping in *Arabidopsis thaliana* by using recombinant inbred lines and random amplified polymorphic DNAs. Proc Natl Acad Sci USA **89:** 1477–1481

Rhodes D, Rich PJ, Brunk DG, Ju GC, Rhodes JC, Pauly MH, Hansen LA (1989) Development of two isogenic sweet corn hybrids differing for glycinebetaine content. Plant Physiol **91:** 1112–1121

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science **273:** 1516–1517

Rose LE, Bittner-Eddy PD, Langley CH, Holub EB, Michelmore RW, Beynon JL (2004) The maintenance of extreme amino acid diversity at the disease resistance gene, *RPP13*, in *Arabidopsis thaliana*. Genetics **166:** 1517–1527

Rosloski SM, Jali SS, Balasubramanian S, Weigel D, Grbic V (2010) Natural diversity in flowering responses of *Arabidopsis thaliana* caused by variation in a tandem gene array. Genetics **186:** 263–276

Roux F, Colomé-Tatché M, Edelist C, Wardenaar R, Guerche P, Hospital F, Colot V, Jansen RC, Johannes F (2011) Genome-wide epigenetic perturbation jump-starts patterns of heritable variation found in nature. Genetics **188:** 1015–1017

Rowe HC, Hansen BG, Halkier BA, Kliebenstein DJ (2008) Biochemical networks and epistasis shape the *Arabidopsis thaliana* metabolome. Plant Cell **20:** 1199–1216

Royer-Pokora B, Kunkel LM, Monaco AP, Goff SC, Newburger PE, Baehner RL, Cole FS, Curnutte JT, Orkin SH (1986) Cloning the gene for an inherited human disorder—chronic granulomatous disease—on the basis of its chromosomal location. Nature **322:** 32–38

Salomé PA, Bomblies K, Fitz J, Laitinen RAE, Warthmann N, Yant L, Weigel D (November 9, 2011a) The recombination landscape in *Arabidopsis thaliana* F_2 populations. Heredity http://dx.doi.org/10.1038.hdy.2011.95

Salomé PA, Bomblies K, Laitinen RA, Yant L, Mott R, Weigel D (2011b) Genetic architecture of flowering-time variation in *Arabidopsis thaliana*. Genetics **188:** 421–433

Sanchez-Moran E, Armstrong SJ, Santos JL, Franklin FC, Jones GH (2002) Variation in chiasma frequency among eight accessions of *Arabidopsis thaliana*. Genetics **162:** 1415–1422

Scarcelli N, Kover PX (2009) Standing genetic variation in *FRIGIDA* mediates experimental evolution of flowering time in *Arabidopsis*. Mol Ecol **18:** 2039–2049

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, 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

Schläppi MR (2006) FRIGIDA LIKE 2 is a functional allele in Landsberg

erecta and compensates for a nonsense allele of FRIGIDA LIKE 1. Plant Physiol **142:** 1728–1738

Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T (2005) A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics **169:** 1601–1615

Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ, Ecker JR (2011) Transgenerational epigenetic instability is a source of novel methylation variants. Science **334:** 369–373

Schmuths H, Bachmann K, Weber WE, Horres R, Hoffmann MH (2006) Effects of preconditioning and temperature during germination of 73 natural accessions of *Arabidopsis thaliana*. Ann Bot (Lond) **97:** 623–634

Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O, Weigel D (2009) Simultaneous alignment of short reads against multiple genomes. Genome Biol **10:** R98

Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, Smith LM, Cao J, Fitz J, Warthmann N, et al (2011) Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. Proc Natl Acad Sci USA **108:** 10249–10254

Schwartz C, Balasubramanian S, Warthmann N, Michael TP, Lempe J, Sureshkumar S, Kobayashi Y, Maloof JN, Borevitz JO, Chory J, et al (2009) Cis-regulatory changes at *FLOWERING LOCUS T* mediate natural variation in flowering responses of *Arabidopsis thaliana*. Genetics **183:** 723–732, 721SI–727SI

Seevers PM, Daly JM, Catedral FF (1971) The role of peroxidase isozymes in resistance to wheat stem rust disease. Plant Physiol **48:** 353–360

Sharbel TF, Haubold B, Mitchell-Olds T (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. Mol Ecol **9:** 2109–2118

Shindo C, Aranzana MJ, Lister C, Baxter C, Nicholls C, Nordborg M, Dean C (2005) Role of *FRIGIDA* and *FLOWERING LOCUS C* in determining variation in flowering time of Arabidopsis. Plant Physiol **138:** 1163–1173

Shindo C, Bernasconi G, Hardtke CS (2007) Natural genetic variation in Arabidopsis: tools, traits and prospects for evolutionary ecology. Ann Bot (Lond) **99:** 1043–1054

Shindo C, Lister C, Crevillen P, Nordborg M, Dean C (2006) Variation in the epigenetic silencing of *FLC* contributes to natural variation in *Arabidopsis* vernalization response. Genes Dev **20:** 3079–3083

Sibout R, Plantegenet S, Hardtke CS (2008) Flowering as a condition for xylem expansion in *Arabidopsis* hypocotyl and root. Curr Biol **18:** 458–463

Simon M, Loudet O, Durand S, Bérard A, Brunel D, Sennesal FX, Durand-Tardif M, Pelletier G, Camilleri C (2008) Quantitative trait loci mapping in five new large recombinant inbred line populations of *Arabidopsis thaliana* genotyped with consensus single-nucleotide polymorphism markers. Genetics **178:** 2253–2264

Simon M, Simon A, Martins F, Botran L, Tisné S, Granier F, Loudet O, Camilleri C (November 11, 2011) DNA fingerprinting and new tools for fine-scale discrimination of *Arabidopsis thaliana* accessions. Plant J http://dx.doi.org/10.1111/j.1365-1313X.2011.04852.x

Slatkin M (2009) Epigenetic inheritance and the missing heritability problem. Genetics **182:** 845–850

Slotte T, Huang HR, Holm K, Ceplitis A, Onge KS, Chen J, Lagercrantz U, Lascoux M (2009) Splicing variation at a *FLOWERING LOCUS C* homeolog is associated with flowering time variation in the tetraploid *Capsella bursa-pastoris*. Genetics **183:** 337–345

Smith LM, Bomblies K, Weigel D (2011) Complex evolutionary events at a tandem cluster of *Arabidopsis thaliana* genes resulting in a single-locus genetic incompatibility. PLoS Genet **7:** e1002164

Soller M, Beckmann JS (1990) Marker-based mapping of quantitative trait loci using replicated progenies. Theor Appl Genet **80:** 205–208

Sønderby IE, Hansen BG, Bjarnholt N, Ticconi C, Halkier BA, Kliebenstein DJ (2007) A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates. PLoS ONE **2:** e1322

Soppe WJ, Jacobsen SE, Alonso-Blanco C, Jackson JP, Kakutani T, Koornneef M, Peeters AJ (2000) The late flowering phenotype of *fwa* mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. Mol Cell **6:** 791–802

Srikanth A, Schmid M (2011) Regulation of flowering time: all roads lead to Rome. Cell Mol Life Sci **68:** 2013–2037

Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J (1999) Dynamics of disease resistance polymorphism at the *Rpm1* locus of *Arabidopsis*. Nature **400:** 667–671

Stam M, Belele C, Dorweiler JE, Chandler VL (2002) Differential chromatin structure within a tandem array 100 kb upstream of the maize *b1* locus is associated with paramutation. Genes Dev **16:** 1906–1918

Staskawicz BJ, Ausubel FM, Baker BJ, Ellis JG, Jones JD (1995) Molecular genetics of plant disease resistance. Science **268:** 661–667

Stenøien HK, Fenster CB, Tonteri A, Savolainen O (2005) Genetic variability in natural populations of *Arabidopsis thaliana* in northern Europe. Mol Ecol **14:** 137–148

Stinchcombe JR, Weinig C, Ungerer M, Olsen KM, Mays C, Halldorsdottir SS, Purugganan MD, Schmitt J (2004) A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene *FRIGIDA*. Proc Natl Acad Sci USA **101:** 4712–4717

Strange A, Li P, Lister C, Anderson J, Warthmann N, Shindo C, Irwin J, Nordborg M, Dean C (2011) Major-effect alleles at relatively few loci underlie distinct vernalization and flowering variation in *Arabidopsis* accessions. PLoS ONE **6:** e19949

Sugliani M, Rajjou L, Clerkx EJ, Koornneef M, Soppe WJ (2009) Natural modifiers of seed longevity in the Arabidopsis mutants *abscisic acid insensitive3-5* (*abi3-5*) and *leafy cotyledon1-3* (*lec1-3*). New Phytol **184:** 898–908

Sulpice R, Pyl ET, Ishihara H, Trenkamp S, Steinfath M, Witucka-Wall H, Gibon Y, Usadel B, Poree F, Piques MC, et al (2009) Starch as a major integrator in the regulation of plant growth. Proc Natl Acad Sci USA **106:** 10348–10353

Sulpice R, Trenkamp S, Steinfath M, Usadel B, Gibon Y, Witucka-Wall H, Pyl ET, Tschoep H, Steinhauser MC, Guenther M, et al (2010) Network analysis of enzyme activities and metabolite levels and their relationship to biomass in a large panel of *Arabidopsis* accessions. Plant Cell **22:** 2872–2893

Syed NH, Chen ZJ (2005) Molecular marker genotypes, heterozygosity and genetic interactions explain heterosis in *Arabidopsis thaliana*. Heredity **94:** 295–304

Symonds VV, Hatlestad G, Lloyd AM (2011) Natural allelic variation defines a role for *ATMYC1*: trichome cell fate determination. PLoS Genet **7:** e1002069

Teixeira FK, Heredia F, Sarazin A, Roudier F, Boccara M, Ciaudo C, Cruaud C, Poulain J, Berdasco M, Fraga MF, et al (2009) A role for RNAi in the selective correction of DNA methylation defects. Science **323:** 1600–1604

Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J (2011) Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. Genome Biol Evol **3:** 219–229

Tian D, Traw MB, Chen JQ, Kreitman M, Bergelson J (2003) Fitness costs of *R*-gene-mediated resistance in *Arabidopsis thaliana*. Nature **423:** 74–77

Todesco M, Balasubramanian S, Hu TT, Traw MB, Horton M, Epple P, Kuhns C, Sureshkumar S, Schwartz C, Lanz C, et al (2010) Natural allelic variation underlying a major fitness trade-off in *Arabidopsis thaliana*. Nature **465:** 632–636

Toomajian C, Hu TT, Aranzana MJ, Lister C, Tang C, Zheng H, Zhao K, Calabrese P, Dean C, Nordborg M (2006) A nonparametric test reveals selection for rapid flowering in the *Arabidopsis* genome. PLoS Biol **4:** e137

Törjék O, Meyer RC, Zehnsdorf M, Teltow M, Strompen G, Witucka-Wall H, Blacha A, Altmann T (2008) Construction and analysis of 2 reciprocal Arabidopsis introgression line populations. J Hered **99:** 396–406

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

Tuinstra MR, Ejeta G, Goldsbrough PB (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

Turesson G (1922a) The genotypical response of the plant species to the habitat. Hereditas **3:** 211–350

Turesson G (1922b) The species and the variety as ecological units. Hereditas **3:** 100–113

Ungerer MC, Linder CR, Rieseberg LH (2003) Effects of genetic background on response to selection in experimental populations of *Arabidopsis thaliana*. Genetics **163:** 277–286

Ungerer MC, Rieseberg LH (2003) Genetic architecture of a selection response in *Arabidopsis thaliana*. Evolution **57:** 2531–2539

van Der Schaar W, Alonso-Blanco C, Léon-Kloosterziel KM, Jansen RC, van Ooijen JW, Koornneef M (1997) QTL analysis of seed dormancy in *Arabidopsis* using recombinant inbred lines and MQM mapping. Heredity **79:** 190–200

Vaughn MW, Tanurdzić M, Lippman Z, Jiang H, Carrasquillo R, Rabinowicz PD, Dedhia N, McCombie WR, Agier N, Bulski A, et al (2007) Epigenetic natural variation in *Arabidopsis thaliana.* PLoS Biol **5:** e174

Vlad D, Rappaport F, Simon M, Loudet O (2010) Gene transposition causing natural variation for growth in *Arabidopsis thaliana*. PLoS Genet **6:** e1000945

Wang CT, Ho CH, Hseu MJ, Chen CM (2010) The subtelomeric region of the *Arabidopsis thaliana* chromosome IIIR contains potential genes and duplicated fragments from other chromosomes. Plant Mol Biol **74:** 155–166

Wang J, Tian L, Lee HS, Wei NE, Jiang H, Watson B, Madlung A, Osborn TC, Doerge RW, Comai L, et al (2006) Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. Genetics **172:** 507–517

Wang N, Qian W, Suppanz I, Wei L, Mao B, Long Y, Meng J, Muller AE, Jung C (2011) Flowering time variation in oilseed rape (*Brassica napus* L.) is associated with allelic variation in the *FRIGIDA* homologue *BnaA.FRI. a*. J Exp Bot **62:** 5641–5658

Wang Q, Sajja U, Rosloski S, Humphrey T, Kim MC, Bomblies K, Weigel D, Grbic V (2007) *HUA2* caused natural variation in shoot morphology of *A. thaliana*. Curr Biol **17:** 1513–1519

Weigel D, Mott R (2009) The 1001 genomes project for *Arabidopsis thaliana*. Genome Biol **10:** 107

Weinig C, Dorn LA, Kane NC, German ZM, Halldorsdottir SS, Ungerer MC, Toyonaga Y, Mackay TF, Purugganan MD, Schmitt J (2003a) Heterogeneous selection at specific loci in natural environments in *Arabidopsis thaliana*. Genetics **165:** 321–329

Weinig C, Stinchcombe JR, Schmitt J (2003b) QTL architecture of resistance and tolerance traits in *Arabidopsis thaliana* in natural environments. Mol Ecol **12:** 1153–1163

Weinig C, Ungerer MC, Dorn LA, Kane NC, Toyonaga Y, Halldorsdottir SS, Mackay TF, Purugganan MD, Schmitt J (2002) Novel loci control variation in reproductive timing in *Arabidopsis thaliana* in natural environments. Genetics **162:** 1875–1884

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature **447:** 661–678

Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ (2007) Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. PLoS Genet **3:** 1687–1701

Werner JD, Borevitz JO, Warthmann N, Trainer GT, Ecker JR, Chory J, Weigel D (2005) Quantitative trait locus mapping and DNA array hybridization identify an *FLM* deletion as a cause for natural flowering-time variation. Proc Natl Acad Sci USA **102:** 2460–2465

Westerman JM (1971) Genotype-environment interaction and developmental regulation in *Arabidopsis thaliana*. IV. Wild material; analysis. Heredity **26:** 383–395

Wilczek AM, Burghardt LT, Cobb AR, Cooper MD, Welch SM, Schmitt J (2010) Genetic and physiological bases for phenological responses to current and predicted climates. Philos Trans R Soc Lond B Biol Sci **365:** 3129–3147

Wilczek AM, Roe JL, Knapp MC, Cooper MD, Lopez-Gallego C, Martin LJ, Muir CD, Sim S, Walker A, Anderson J, et al (2009) Effects of genetic perturbation on seasonal life history plasticity. Science **323:** 930–934

Woo HR, Richards EJ (2008) Natural variation in DNA methylation in ribosomal RNA genes of *Arabidopsis thaliana*. BMC Plant Biol **8:** 92

Xu Z, Zou F, Vision TJ (2005) Improving quantitative trait loci mapping resolution in experimental crosses by the use of genotypically selected samples. Genetics **170:** 401–408

Yamamoto E, Takashi T, Morinaka Y, Lin S, Wu J, Matsumoto T, Kitano H, Matsuoka M, Ashikari M (2010) Gain of deleterious function causes an autoimmune response and Bateson-Dobzhansky-Muller incompatibility in rice. Mol Genet Genomics **283:** 305–315

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics **178:** 539–551

Zeller G, Clark RM, Schneeberger K, Bohlen A, Weigel D, Rätsch G (2008) Detecting polymorphic regions in *Arabidopsis thaliana* with resequencing microarrays. Genome Res **18:** 918–929

Zhang X, Borevitz JO (2009) Global analysis of allele-specific expression in *Arabidopsis thaliana*. Genetics **182:** 943–954

Zhang X, Cal AJ, Borevitz JO (2011) Genetic architecture of regulatory variation in *Arabidopsis thaliana*. Genome Res **21:** 725–733

Zhang X, Shiu SH, Cal A, Borevitz JO (2008) Global analysis of genetic, epigenetic and transcriptional polymorphisms in *Arabidopsis thaliana* using whole genome tiling arrays. PLoS Genet **4:** e1000032

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

Zhao J, Kulkarni V, Liu N, Del Carpio DP, Bucher J, Bonnema G (2010) *BrFLC2* (*FLOWERING LOCUS C*) as a candidate gene for a vernalization response QTL in *Brassica rapa*. J Exp Bot **61:** 1817–1825

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007) An *Arabidopsis* example of association mapping in structured samples. PLoS Genet **3:** e4

Review

Cell PRESS

Feature Review

Molecular mechanisms of polyploidy and hybrid vigor

Z. Jeffrey Chen

Section of Molecular Cell and Developmental Biology, Section of Integrative Biology, Center for Computational Biology and Bioinformatics, and Institute for Cellular and Molecular Biology, The University of Texas at Austin, One University Station A4800, Austin, TX 78712, USA

Hybrids such as maize (*Zea mays*) or domestic dog (*Canis lupus familiaris*) grow bigger and stronger than their parents. This is also true for allopolyploids such as wheat (*Triticum* spp.) or frog (i.e. *Xenopus* and *Silurana*) that contain two or more sets of chromosomes from different species. The phenomenon, known as hybrid vigor or heterosis, was systematically characterized by Charles Darwin (1876). The rediscovery of heterosis in maize a century ago has revolutionized plant and animal breeding and production. Although genetic models for heterosis have been rigorously tested, the molecular bases remain elusive. Recent studies have determined the roles of nonadditive gene expression, small RNAs, and epigenetic regulation, including circadian-mediated metabolic pathways, in hybrid vigor, which could lead to better use and exploitation of the increased biomass and yield in hybrids and allopolyploids for food, feed, and biofuels.

Polyploidy and hybrid vigor – a general view

Hybrids and polyploids (whole genome duplication) occur in many plants and some animals. Hybrids are formed by hybridizing different strains, varieties, or species. Although heterosis or hybrid vigor is evolutionarily defined as that the heterozygotes have higher fitness in a population than the homozygotes, heterosis generally refers to superior levels of biomass, stature, growth rate, and/or fertility in the hybrid offspring than in the parents. The latter definition is adopted in this review. Polyploidy refers to an organism or cell that contains two or more sets of basic chromosomes. An autopolyploid is formed by duplicating a genome within the same species, such as potato (*Solanum tuberosum*), alfalfa (*Medicago sativa*), and sugarcane (*Saccharum*), whereas an allopolyploid is derived from hybridization between different species followed by chromosome doubling or from fusion of unreduced gametes between species. An allopolyploid is a 'doubled interspecific hybrid', leading to permanent fixation of heterozygosity and hybrid vigor. Many crops, including maize (*Zea mays*) and sorghum (Sorghum bicolor), are grown mainly as hybrids, and many other crops, such as bread wheat (*Triticum aestivum*), upland cotton (*Gossypium hirsutum*), and oilseed rape (*Brassica napus*, also known as canola), are grown as allopolyploids. Despite the evolutionary significance of polyploidy and agricultural importance of hybrid vigor, the mechanisms of polyploidy and hybrid vigor are poorly understood. In this review, I outlined a historical perspective of hybrids, allopolyploids, and hybrid vigor and reevaluated genetic models for heterosis in relation to the recent findings for the roles of nonadditive gene expression, allelic expression variation (see Glossary),

Glossary

Allelic expression variation: the expression pattern or level of the alleles in the hybrids is different from that in the parents. This can also refer to the expression of homoeologous loci in interspecific hybrids.
Allopolyploid: an organism or individual that contains two or more sets of genetically distinct chromosomes, usually by hybridization between different species.
Amphidiploid: synonymous to allopolyploid. Contains a diploid set of chromosomes derived from each parent. Strictly speaking, only bivalents are formed in an amphidiploid, whereas multivalents are formed in an allopolyploid.
Aneuploid: an individual in which the chromosome number is not an exact multiple of the typical haploid set for that species.
Apomixis: only one parent (usually female) contributes genes to the offspring.
Autopolyploid: a polyploid created by the multiplication of one basic set of chromosomes (in one species).
Epigenetics: non-Mendelian inheritance, heritable changes in gene expression without changes in primarily DNA sequences.
Gametic imprinting: the expression of a gene is dependent on its parental origin in the offspring.
Genomic shock: the release of genome-wide chromatin constraints of gene expression, including activation of transposons in response to environmental changes and genomic hybridization.
Heterosis: the greater vigor of growth, survival, and fertility in hybrids than in the parents.
Homoeologous: chromosomes or genes in related species that are derived from the same ancestor and coexist in an allopolyploid.
Homologous: genes or structures that share a common evolutionary ancestor.
Homoploid hybrids: hybrids formed between different species, in some cases resulting in a derivative hybrid species without a change in chromosome number.
Imprinting or genomic imprinting: unequal expression of maternal and paternal alleles in the offspring.
Nonadditive gene expression: the expression level of a gene in an allotetraploid is not equal to the sum of two parental loci ($1 + 1 \neq 2$), leading to activation (>2), repression (<2), dominance, or overdominance.
Orthologous: chromosomes or genes in different species that have evolved from the same ancestor.
Paralogous: two or more genes in the same species that share a single ancestral origin.
Paramutation: heritable changes in gene expression induced by allelic interactions.
Ploidy: the number of basic chromosome sets.
Polyploid: an individual or cell that has two or more basic sets of chromosomes.
X-chromosome inactivation: during mammalian development, the repression of one of the two X-chromosomes in the somatic cells of females as a method of dosage compensation.

Corresponding author: Chen, Z.J. (zjchen@mail.utexas.edu).

1360-1385/$ – see front matter doi:10.1016/j.tplants.2009.12.003 Available online 18 January 2010

and small RNAs in hybrid vigor and incompatibility. The molecular mechanisms for single-locus heterosis were highlighted using empirical data on altered epigenetic regulation of master regulators such as circadian clock genes that control physiological and metabolic pathways, leading to increased growth vigor and biomass in hybrids and allopolyploids. A better understanding of the mechanisms of polyploidy and hybrid vigor will help us manipulate gene expression and heterosis in hybrid plants and polyploid crops that are directly relevant to the growing demand of plant materials for food, feed, and fuels.

Hybrids, allopolyploids, and hybrid vigor – a historical perspective

"I raised close together two large beds of self-fertilised and crossed seedlings from the same plant of *Linaria vulgaris*. To my surprise, the crossed plants when fully grown were plainly taller and more vigorous than the self-fertilised ones." – Charles Darwin (The Effects of Cross and Self Fertilisation in the Vegetable Kingdom, 1876).

In his book [1], Charles Darwin systematically documented the growth, development, and seed fertility of cross-pollinated plants compared with that plants for more than 60 different species of plants, including pea (*Pisum sativum*), tomato (*Solanum lycopersicum*), and tobacco (*Nicotiana tabacum*). The results led him to conclude that inbreeding was generally deleterious (later known as inbreeding depression), and cross-fertilization was generally beneficial. Thirty-two years later, George H. Shull published a landmark paper, entitled 'The composition of a field of maize' [2], which marked the rediscovery of hybrid vigor or heterosis and the beginning of applying heterosis in plant breeding. Shull indicated that selfing maize (corn; *Zea mays*) plants led to a reduction of overall growth vigor and yield. The notion was well supported from maize inbreeding experiments by Edward M. East [3]. East predicted that the low seed yield in the inbred lines would impede hybrid production. Shull then demonstrated that the hybrids had uniformly superior growth vigor and yield to the inbreeding parents. The low seed yield in the inbreds was improved by using double-cross (i.e. making the hybrids by crossing two hybrids derived from two pairs of inbred lines). Maize breeders continued to improve seed production in inbred lines until there were sufficient seeds to make the single-cross hybrids with a significant increase in yield [4]. Maize yield has steadily increased sixfold since the introduction of hybrids in the 1920s [5].

Hybrid rice was first studied in 1964 in China. A rice breeder, Yuan, Long Ping, initiated the research on hybrid rice and heterosis in China. The technology of hybrid rice seed production was developed in the 1970s. The most commonly used hybrids are produced between different varieties within a subspecies or between the subspecies *Oryza sativa* ssp. *indica* and *O. sativa* ssp. *japonica* [6]. Although the grain quality of intraspecific hybrids could be further improved, the yield from hybrid rice is $\geq$20% greater than that from conventional rice and accounts for 50% of the total rice area in many rice producing countries, including China, India, and Indonesia.

When US scientists produced hybrid maize a century ago, Russian scientists developed a new species named *Rhaphanobrassica* from the hybrids between two plant genera *Raphanus* and *Brassica* [7]. The cytologist G.D. Karpechenko hoped to produce plants that would have the roots of radish and the leaves of cabbage. The hybrids were made from artificial crosses between two vegetables, the radish (*Raphanus sativus*, $2n = 18$) and the cabbage (*Brassica oleracea*, $2n = 18$). However, the F_1 hybrids had the roots of cabbage and the leaves of radish, and were highly sterile, probably because of a failure in chromosome pairing. A few fertile plants were found to be spontaneous allotetraploids that contained 36 chromosomes, and these plants had vegetative growth vigor. Unfortunately, the new species was as short-lived as its creator, who was executed in 1941 for his association with N. Vavilov in an alleged 'anti-Soviet group'.

Numerous *Nicotiana* hybrids and allopolyploids have been produced. Some, such as *Nicotiana glutinosa* × *N. tabacum*, were not vigorous but rather dwarf [8], whereas others such as *N. glutinosa* × *Nicotiana tomentosa* had great vigor [9].

Triticale (× *Triticale* Tschermak) is a successful man-made interspecies hybrid or allopolyploid [10,11]. Triticale is derived from crossing two cereals, hexaploid bread wheat (*T. aestivum*) or tetraploid durum wheat (*Triticum turgidum*) and rye (*Secale cereale*). In 1875, A.S. Wilson reported the first hybrid between wheat and rye in Scotland (UK), and a decade later W. Rimpau produced the first doubled-fertile hybrid that showed little heterosis. In Russia during the crop season of 1918, thousands of natural hybrids between wheat and rye appeared in many wheat fields. For the next 16 years, G.K. Meister and his colleagues exploited these vigorous hybrids [11]. In 1935, M. Lindschau and E. Oehler named triticale after Tschermak, one of the rediscoverers of Mendelian Law. In theory, triticale combines the high yield potential and good grain quality of wheat with the disease and stress tolerance of rye. Triticale has vigor in vegetative growth, biomass, and tolerance to adverse conditions such as limited water and poor soil conditions. It is grown mainly for forage and animal feed because of poor baking quality and seed fertility, which need to be improved. Triticale is primarily grown in Poland, Australia, Germany, France, and China. The Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT) has a triticale program that is aimed at improving food production and nutrition in developing countries. Triticale can be considered an energy crop because of its increased levels of biomass heterosis.

Modern view of hybrids, allopolyploids, and hybrid vigor

Humans have simply replicated a few examples of these remarkable natural processes that have produced many hybrid and polyploid plants not recorded in literature. Estimates indicate that ~10% of animal and ~25% of plant species hybridize with at least one other species [12]. A recent study estimates that 15% of angiosperm and 31% of fern speciation events are accompanied by an increase in ploidy [13]. The proportion of polyploid flowering plants might be 70% or more [14], and the majority (~75%) are allopolyploids [15,16]. Many agricultural crops such as wheat, cotton, and oilseed rape are allopolyploids.

Allopolyploids are presumably formed spontaneously by crossing related species via unreduced gametes or via spontaneous chromosome doubling of the resulting interspecific hybrids. A large number of hybrids spontaneously form between wheat and rye in wheat fields, suggesting that hybridization between species (and genera) occurs frequently if growth and physiological conditions overcome hybridization barriers. Interspecific hybrids and allopolyploids have been formed in *Tragopogon* [17], *Spartina* [18], and *Senecio* [19] in recent centuries. Allopolyploid *Spartina townsendii* is derived from *Spartina alternifolia* and *Spartina stricta*. The allopolyploid is so vigorous that it has replaced the parental forms and spread all over southern England (UK) and to France [18]. *Senecio* species are native in France, and the allopolyploids have spread to England [19]. *Tragopogon* is native to Euroasia; allopolyploids were formed in the early nineteenth century in North America and become invasive in local environments [17]. Some allopolyploids such as *Tragopogon* [17] and *Brassica* [20] are formed through multiple origins and by reciprocal crosses (with different combinations of maternal cytoplasm and paternal nucleus), whereas others such as cotton [21], wheat [22], and *Arabidopsis* (*Arabidopsis thaliana*) [23] are formed by a single or a few hybridization events.

Durum wheat (*T. turgidum,* AABB, $2n = 4x = 28$) is an allotetraploid formed by crossing two extant diploid wild grasses, *Triticum monococcum* or *Triticum urartu* (AA, $2n = 14$) and a wild goatgrass such as *Triticum searsii* or *Triticum speltoides* (BB, $2n = 14$). The exact donor of the B genome is unknown. Approximately 8000–10 000 years ago, hexaploid wheat or bread wheat (*T. aestivum*, $2n = 6x = 42$, AABBDD) was formed in farmers' fields through hybridization between a domesticated tetraploid wheat and a wild diploid grass (*Triticum tauschii*, DD, $2n = 14$). The hexaploid bread wheat has been domesticated and cultivated since the history of human civilization [24].

Cotton belongs to the genus *Gossypium*, which includes about 45 species split across two ploidy levels, diploid ($2n = 2x = 26$) and tetraploid ($2n = 4x = 52$) [21]. A polyploidization event occurred ~1.5 million years ago between AA and DD extant diploid species, and the AADD allotetraploids diverged into five species that are distributed throughout the world [21]. Among them, upland or American cotton, *Gossypium hirsutum*, accounts for >95% of cotton produced worldwide. Pima or Egyptian cotton, *Gossypium barbadense*, accounts for <5% of the cotton produced. The AA progenitor species produce both lint (long) fibers, which are spinnable into yarn, and shorter fibers called fuzz. By contrast, the DD genome progenitor species produce few lint fibers, which are initiated pre-anthesis, but are much shorter in length than the lint fibers of the AA genome progenitor. Interestingly, the allotetraploids produce more abundant and higher quality fibers than the extant descendant species, suggesting strong selection on polyploid cotton for fiber traits.

The genus *Brassica* offers a textbook example of reciprocal hybrids and allopolyploids formed between three diploid species, which is known as U-triangle [20]. The three diploid species are *Brassica nigra* ($2n = 2x = 16$), *Brassica oleracea* ($2n = 2x = 18$), and *Brassica campestris* or *rapa* ($2n = 2x = 20$), and each allotetraploid species is formed between two diploid species. For example, *B. napus* ($2n = 4x = 38$) is an allotetraploid between *Brassica rapa* and *B. oleracea*, *Brassica juncea* ($2n = 4x = 34$) is formed between *B. nigra* and *B oleracea*, and *Brassica carinata* ($2n = 4x = 36$) is formed between *B. nigra* and *B. rapa*. *Brassica napus* (oilseed rape) has higher oil content and better oil composition than its parents, probably because of natural selection and human domestication for these traits in the interspecific hybrids or allotetraploids.

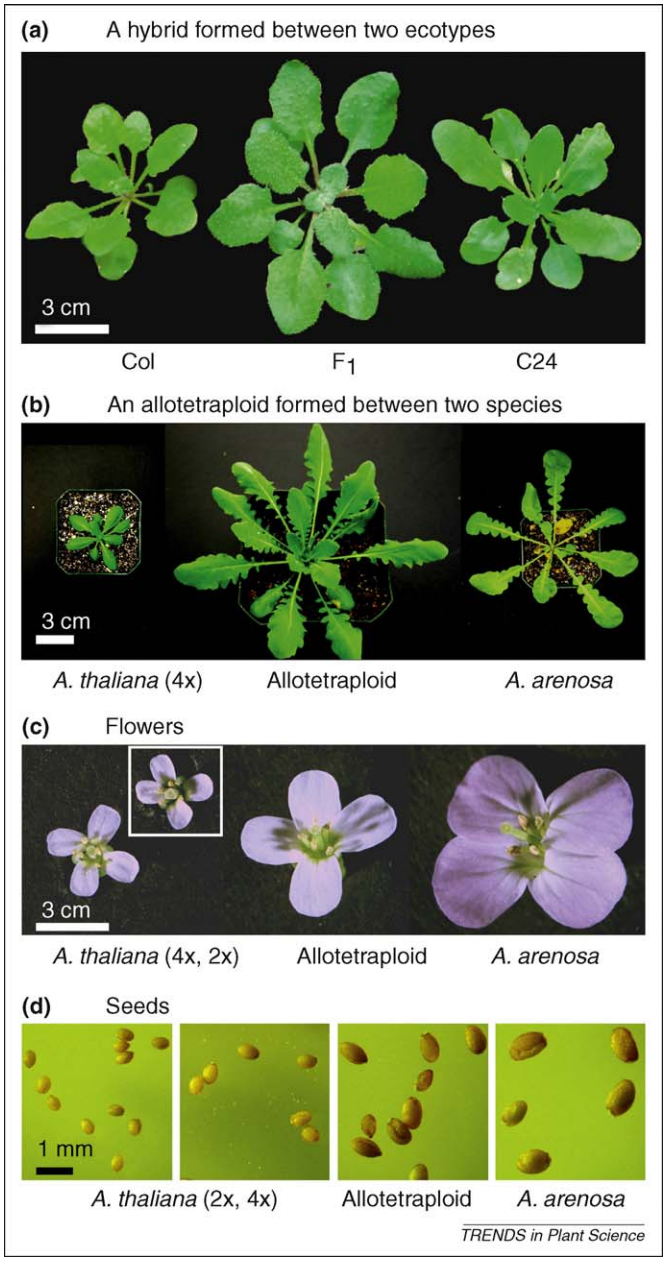

Figure 1. *Arabidopsis* hybrids and allotetraploids. **(a)** Seedlings of the F_1 hybrid produced by crossing *Arabidopsis thaliana* Columbia × *A. thaliana* C24. **(b)** A stable allotetraploid (in F_8 generation) was maintained by self-pollination. **(a)** and **(b)** were reproduced from [101] with permission. The F_1 interspecific hybrid or allotetraploid was produced by pollinating *A. thaliana Ler* autotetraploid with pollen from the outcrossing *Arabidopsis arenosa* tetraploid [48,132]. **(c)** Typical flowers of the allotetraploid and its progenitors, *A. thaliana* tetraploid (inset, diploid) and *A. arenosa*. **(d)** Seeds of the allotetraploid and its progenitors, *A. thaliana* Ler tetraploid and *A. arenosa*. Seeds of *A. thaliana Ler* diploid are also shown.

Hybrids and allopolyploids also occur in *Arabidopsis*, a member of the Brassicaeae family. Many hybrids formed between different ecotypes do not have obvious growth vigor. Only a handful of hybrid combinations give rise to growth vigor [25] and other traits such as cold tolerance [26] (Figure 1a). The available genetic resources such as recombinant inbred lines (RILs) have been used to dissect and study quantitative trait loci (QTL) that are associated with growth-related and life history traits [25,27,28].

Arabidopsis suecica ($2n = 4x = 26$) is a natural allotetraploid formed between extant *A. thaliana* and *Arabidopsis arenosa* 12,000–300,000 years ago [29]. New allotetraploids can be readily resynthesized by crossing these two species *A. thaliana* ($2n = 4x = 20$) and *A. arenosa* ($2n = 4x = 32$) (Figure 1b). During vegetative growth, the allotetraploids are 3–5 times larger than *A. thaliana* and twice as large as *A. arenosa*. Under long-day conditions (light/dark of 16/8 h), the allotetraploids flower slightly later than the late-flowering parent *A. arenosa*, and produce 18–25 rosette leaves, whereas *A. arenosa* has 10–12 leaves at flowering. The flowers of allotetraploids are intermediate between those of the two parents. The seeds are roughly twice the size of *A. thaliana* and slightly smaller than that of *A. arenosa*, a natural outcrossing autotetraploid. The seed germination rates are much higher in the stable allotetraploids (after 7–8 generations of selfing) than in *A. arenosa*. A large portion of *A. arenosa* seeds are not fully developed, probably resulting from failure of embryo and endosperm development as a consequence of being an autotetraploid [30,31].

By definition, most heterozygous animals, including humans, are hybrids that carry different alleles from female and male parents. Mating among siblings leads to accumulation of deleterious mutations and recessive alleles, a phenomenon known as inbreeding depression [32]. Although interspecific hybrids and polyploids are rarer in animals than in plants [33,34], interspecific hybrids do occur in mammals (e.g. a mule is a hybrid between a horse and a donkey). Mammalian interspecific hybrids are sterile, probably because of incompatibility and/or imbalance in imprinting and sex chromosome dosage, as proposed by H. Muller [33]. The number of homoploid hybrid-species in animals is growing rapidly [35]. They include a recent invasive sculpin, a hybrid fish (*Cottus gobio*) derived from *Cottus perifretum* and *Cottus rhenanum*, a cyrinid fish *Gila seminuda* that is formed between *Gila robusta* and *Gila elegans*, *Rhagoletis* fruitflies, and *Heliconius* butterflies [36,37]. Like plant hybrids, animal hybrids grow generally better than their parents. For example, mules are generally tougher than horses, and they endure heat better than horses. They have denser muscling from their donkey parents than horses and have fewer leg problems than horses, but they do not run as fast as horses, a trait probably inherited from their donkey parents.

Many interspecific hybrids have reduced viability and fertility. The Bateson–Dobzhansky–Muller model suggests that the hybrid incompatibilities are caused by interactions between genes that have functionally diverged in the respective hybridizing species [38,39]. These incompatibilities appear concurrently with speciation or consequently after species divergence. The incompatibility genes include hybrid lethality genes found in *Drosophila* [40,41], *Caenorhabditis elegans* [42], and *Arabidopsis* [43,44]. In *Drosophila*, the lethality in F_1 hybrid males is caused by the interaction between Lethal hybrid rescue (Lhr), which has functionally diverged in *Drosophila simulans* and Hybrid male rescue (Hmr), which has functionally diverged in *Drosophila melanogaster* [40]. In another study, hybrid lethality is caused by the nucleoporin 160 kDa (Nup160) gene of *D. simulans*, which is incompatible with one or more factors from the *D. melanogaster* X-chromosome [41]. In *C. elegans*, the interactions between two tightly linked but diverged alleles *zeel-1* and *peel-1* cause widespread genetic incompatibility [42]. Recent work in *Arabidopsis* supports functional divergence between duplicate genes that lead to hybrid incompatibilities between ecotypes [44] or hybrid necrosis in intraspecific hybrids [43]. In mammals, hybrid incompatibilities are related to abnormal expression patterns of imprinting genes in interspecific hybrids in *Peromyscus* [45] or epigenetic activation of retroelements in marsupial hybrids [46]. In plants, some imprinted genes were abnormally silenced in *Arabidopsis* interspecific hybrids [31,47], and many protein-coding genes are epigenetically regulated in allotetraploids [48,49].

For genetically viable hybrids, the degree of heterosis is proportional to the genetic differences in two parental strains [50]. In other words, the levels of heterosis increase as the genetic distances between the parents increase. After evaluating the phenotypic data from 37 genera, including *Zea*, *Solanum*, and *Nicotiana*, E.M. East (1936) [50] noted that interspecific hybrids generally show more heterosis than intraspecific hybrids, if the genetic difference between the species or genera does not prevent them from forming compatible hybrids. The hybrids formed between different subgenera show more heterosis than the hybrids formed between species within the same subgenera. If the hybrids are incompatible, they are dwarf and stunted, probably because dramatic differences in growth and reproductive development inherited from the divergent parents fail to be reconciled. Indeed, the hybrids formed between subgenera often have more heterosis as well as more dwarfs. For example, most intergenic or interspecific hybrids are abnormal, and yet the greatest amount of heterosis is found in the hybrids derived from *Raphanus* and *Brassica* [7]. In rice, the hybrids between two subspecies show more heterosis than the hybrids between varieties within a subspecies. However, the notion may not be generalized across all hybrids. In maize (*Z. mays*) and tobacco, although the varieties (inbred lines) are genetically similar, the hybrids formed between different combinations of varieties show dramatic levels of heterosis. This suggests that the interaction between a few genes or the combination of a few genes in a genetic cross plays an important role in heterosis, as observed in tomato [51]. Alternatively, large-scale recombination suppression accompanied by a high level of residual heterozygosity could be associated with inbreeding depression and heterosis in maize [52,53]. Notably, genetic mechanisms responsible for heterosis may be different between the species that are naturally self-pollinating and out-crossing.

Heterosis is more predominant in outcrossing than inbreeding species, and the inbreeding populations do not have obvious heterosis of fitness.

Notably, heterosis in interspecific hybrids is permanently fixed in the respective allopolyploids in which the chromosomes are doubled. This is facilitated by many allopolyploids that become self-pollinating irrespective of pollinating patterns in the parents. Thus, the heterosis is heritable and selected in the allopolyploid progeny. Although heterosis in interspecific hybrids and allopolyploids is generally high, the heterosis in autopolyploids is not obvious [50,54]. In *Arabidopsis*, diploids and autotetraploids often have similar morphology, leaf size, and plant stature. The autotetraploids have slightly larger flowers and seeds (Figure 1c and d), and flower later than the diploids, depending on the combination of genotypes. For example, the difference in flowering time between a diploid and an autotetraploid is greater in Columbia ecotype than in Landsberg *erecta* ecotype.

The degree of heterosis may shift during different stages of growth and development [51]. If growth vigor is shown in the early stages, it is often exhibited not only in seedlings, vegetative tissues, and organs such as rosettes, and overall biomass, but also in the late stages of reproduction such as in the flowers and fruits. In some plants, heterosis in vegetative growth is different from that in reproductive development because they are controlled by different sets of genes and regulatory pathways. It is notable that biomass heterosis in plants is largely dependent on flowering time. For example, late flowering and indefinite inflorescent plants often have greater biomass than the early flowering and definite inflorescent plants. The flowering time is controlled by a few loci in inbreeding *Arabidopsis* and rice [55–58]. The single-locus heterosis in tomato could be controlled by a flowering locus T (FT)-like locus that regulates the transition from definite to indefinite inflorescence [51]. In outcrossing maize, the flowering time is mediated by additive effects of numerous (two-dozen or more) QTLs, each with only a small effect on the trait [59]. Interestingly, in *B. napus* late flowering is heterotic, whereas in maize hybrids early flowering is heterotic, suggesting different effects of gene actions (repression or activation) on heterosis.

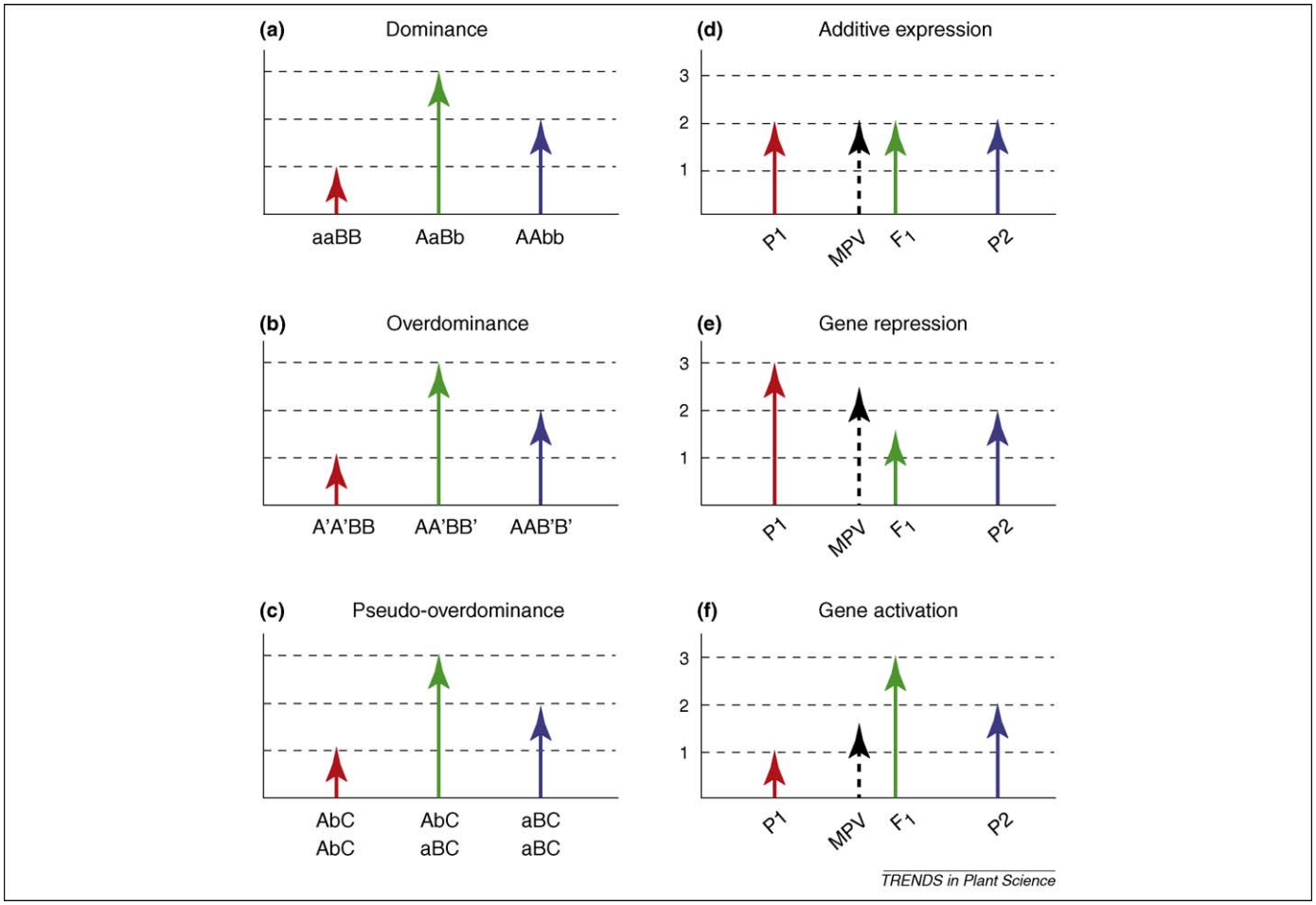

Figure 2. Genetic models and nonadditive gene expression for heterosis. **(a)** The dominance model. The F_1 with both dominant alleles (AaBb) of two loci is superior to the parents that contain only one pair of dominant alleles (aaBB and AAbb) because the superior or dominant allele complements the inferior or recessive allele. **(b)** The overdominance model. The interactions between heterozygous alleles in F_1 (AA'BB') causes superior phenotypes compared with the combinations of homozygous alleles in the parents (A'A'BB and AAB'B'). **(c)** The pseudo-overdominance model. The combination of dominant alleles (AaBb) in repulsion (AbC/aBC) in the F_1 acts as overdominance compared with homozygous parents (AAbbCC and aaBBCC). The presence of dominant alleles in F_1 complements the recessive alleles, leading to a better phenotype. **(d)** Additive expression. The expression level of a gene, genotype, or phenotype is additive. Abbreviations: MPV, mid-parent value (1/2P1 + 1/2P2); P1, parent 1; P2, parent 2. P1, P2, MPV, and F_1 represent the values of gene expression, genotype, or phenotype. **(e, f)** Nonadditive expression. **(e)** Gene repression. The expression of a gene, genotype, or phenotype is lower than the MPV. **(f)** Gene activation. The expression of a gene, genotype, or phenotype is higher than the MPV, which includes dominance, overdominance, and pseudo-overdominance models. Gene repression and activation also explain epistatic interactions. Relative expression levels (1, 2, and 3) are shown on the y-axis.

Genetic models for hybrid vigor

The genetic basis for hybrid vigor or heterosis has been debated for over a century, but little consensus has been reached. Several hypotheses including dominance, overdominance, and pseudo-overdominance are available to explain the phenomenon of hybrid vigor. According to the dominance model [60,61], inbred parents contain inferior or deleterious alleles in several loci that inhibit overall good performance, whereas in the hybrids these inferior alleles in one parent are complemented by the superior or dominant alleles from the other parent (Figure 2a). As a result, the hybrids have an overall better performance than the parents. The model is based on the dominance (wild type) and recessive (mutant) aspect of trait performance, and genetic complementation is likely to occur in the combination of alleles from respective parents. Moreover, one can apply statistical models to dissect additive and dominant components of genetic variation. In theory, the parent that contains homozygous superior or dominant alleles for all possible loci would perform better than the hybrids, but hybrid maize breeding practice has indicated otherwise. In spite of dramatic improvement of inbred parents by eliminating deleterious alleles, the heterotic (or allelomorphic) responses in the hybrids often exceed those in the parents [50]. Maize is naturally outcrossing and requires a certain amount of combinational dominant and recessive alleles in some genetic loci to avoid inferior performance or lethality from being completely inbred. In other words, the parent with recessive alleles in all genetic loci would be deleterious, as would the parent with dominant alleles in all genetic loci.

The overdominance model [2,50,62] suggests that novel allelic interactions within each of many genetic loci lead to superior function over the homozygous states in the inbred parents (Figure 2b). This model is favored because hybrids always outperform the parents that have been excessively inbred and selected and contain many superior or dominant genetic loci [50]. Moreover, it is the allelic combination in the hybrids that determines the levels of heterosis. The genetic composition of inbred parents does not necessarily predict the levels of hybrid vigor. A challenge for this model is to identify the best combination of a single genetic locus or a few loci that contribute to the overall heterosis, which seems to contradict the hybrid performance of many agronomic traits that are controlled by multiple genetic loci.

A recent study [63] has suggested an alternative model, pseudo-overdominance (Figure 2c). The overdominance is associated with the complementation of two or more linked dominant and recessive alleles in repulsion, in which the dominant and recessive alleles are located on opposite homologs of the two genes, acting as overdominance. The heterosis associated with pseudo-overdominance can dissipate in the selfing progeny because genetic recombination leads to the dissociation of the alleles from the repulsion state, which is exactly what is observed in a study with tomato hybrids [63]. This pseudo-overdominance can also arise from numerous alleles in recombination suppression regions where good and bad allele combinations are in repulsion [52,53].

These genetic models have limitations. For example, heterosis in rice has been found to be associated with three different models, namely, dominance [64], overdominance [65], and epistasis [66]. These different conclusions are probably related to the complexity of genetic bases and trait variability for heterosis. First, heterosis can result from epistatic interactions among the alleles in different loci, which cannot be easily explained by statistical models. Epistasis is involved in many QTLs associated with inbreeding depression and heterosis in maize [67] and rice [66,68]. Second, heterosis is affected by genetic backgrounds. For example, one of the two QTLs controlling differences in morphology and inflorescence architecture between maize and its ancestor (teosinte, *Zea mays* ssp. *parviglumis*) has strong phenotypic effects in the teosinte background but reduced effects in the maize genetic background [69]. When the two QTLs are combined into one genotype, both morphology and inflorescence architecture are altered. In an extensive analysis of heterosis for dry biomass in 63 *Arabidopsis* accessions that were crossed with three reference lines (Col-0, C24, and Nd), 29 out of 169 crosses had significant heterosis for shoot biomass, and the biomass heterosis was higher in some hybrids (e.g. Col-0 $\times$ C24) than in others [25]. This is consistent with the higher levels of growth vigor in interspecific hybrids than in the ecotype hybrids (Figure 1). Third, altered levels of heterosis observed in different genetic backgrounds also suggest a role for maternal and paternal effects of genetic loci in hybrid performance [70], although allelic expression variation is not commonly observed in reciprocal crosses [71,72]. However, a recent study suggested otherwise, and nearly 50% genes showed paternal dominant expression patterns in the seedlings of maize reciprocal hybrids [73], which is inconsistent with similar phenotypes observed in reciprocal hybrids [54,71,72]. It is likely that some of these changes in gene expression may dissipate over time. Fourth, heterosis is affected by many genetic loci. Statistical and genetic models cannot accurately estimate the relative contribution of individual loci to a particular pathway or trait. Some transcription factors and chromatin proteins may control the expression of many other genes in various biological pathways. Finally, these genetic models do not explain well the heterosis in polyploid plants because allelic and genomic dosage may play a more important role than the allelic complementation or interactions. Changes in dosage-dependent gene expression may be more profound than alteration in allelic interactions. In maize, the increased number of genes and the genome dosage appears to have a negative effect on growth vigor and increased levels of inbreeding depression [54].

Nonadditive gene expression in the hybrids and allotetraploids

At gene expression levels, the dominance model suggests that the expression of genes in the hybrids is a result of combined or additive expression of two alleles in the parents (e.g. $1 + 1 = 2$) (Figure 2d), whereas the overdominance model indicates that allelic interactions in the hybrids lead to nonadditive expression of the alleles derived from the parents ($1 + 1 \neq 2$) (Figure 2e and f).

If the interactions lead to positive effects or gene activation, the outcome is expected to be $1 + 1 > 2$. If the interactions result in negative effects or gene repression, the expected outcome would be $1 + 1 < 2$. The expression of some genes falls in the range between additive and nonadditive expression. Nonadditive expression explains positive as well as negative epistatic interactions.

Nonadditive expression of 30 selected genes was studied in maize diploid and triploid hybrids using RNA blots and normalized expression values with internal controls [74]. The expression values of 19–20 genes in reciprocal hybrids are different from the mid-parent values (MPV). The transcript levels in the diploid hybrids correlate negatively with the levels in diploid inbreds. Moreover, genome dosage affects transcript levels in diploid and triploid hybrids. The transcript levels are higher in triploids than in diploids. The transcript levels for nearly half of the genes tested are different in reciprocal triploid crosses, suggesting strong maternal effects of gene expression in triploid hybrids.

In a study using cDNA microarrays, ~10% of ESTs were expressed differently between the two inbred parents [75]. Among them, 78% (1062 of 1367) of ESTs were additively expressed in the hybrids relative to the MPV, and 22% were nonadditively expressed. The expression patterns include all possible modes of nonadditive expression: high- and low-parent dominance, underdominance, and overdominance. The data suggest that multiple molecular mechanisms, including overdominance, contribute to heterosis. In a similar study using microarrays, ~80% of the genes that were expressed differently between the two parents were additively expressed in the hybrids [72]. However, among 20% of nonadditively expressed genes, many were expressed at levels within the parental range. Few genes showed expression levels higher than the high parent or lower than the low parent. Further analysis of allele-specific expression patterns in the hybrid indicates that gene expression variation is largely associated with *cis*-regulatory variation. The data suggest that *cis*-regulatory variation between the alleles maintains allelic expression levels in the F_1 hybrid, leading to additive expression. Another study [76] suggested that hybrid yield and heterosis are associated positively with the proportion of additively expressed genes, negatively with the proportion of paternally expressed genes, and not correlated with over- or under-expression of some specific genes. These different conclusions related to the relative contribution of additive and nonadditive expression to the hybrid performance in similar studies using the same pair of inbred parents might be caused by developmental variation among different tissues examined, various normalization methods and/or different statistical tools used in microarray and RNA blot analyses. Moreover, it is not surprising to identify positive effects of additive expression on heterosis because the proportion of additively expressed genes is generally high (~80%).

Allelic expression variation varies from unequal expression of both alleles (biallelic) to expression of a single allele (monoallelic) in the hybrids, which is reminiscent of developmental reactivation of silenced rRNA genes in *Brassica* allotetraploids [77] and organ-specific reciprocal silencing in cotton allotetraploids [78], although they involve two homoeologous loci. In maize hybrids, the allelic expression variation can respond to planting density and drought stress [71]. For example, biallelic expression for seven of 15 genes examined is found in a genetically improved modern hybrid, whereas mono-allelic expression is observed in a less improved old hybrid. The two alleles of stress responsive genes in the hybrid are differentially expressed in response to density and drought stresses. Although maternal or paternal effects on allelic expression are not commonly observed in vegetative tissues and seedlings, expression of many genes (~8%) deviates from a 1:1 ratio, the expected ratio in the hybrids of reciprocal crosses, and 2:1, the expected ratio in three stages of endosperm development in the hybrids of reciprocal crosses [70]. These genes resemble maternally or paternally expressed genes, which is probably associated with genomic imprinting. The gene encoding a no-apical-meristem (NAM) related protein 1 (nrp1) is expressed only in the endosperm, in which the maternally transmitted alleles are expressed, whereas the paternally transmitted alleles are silenced throughout the three developmental stages.

Genome-wide nonadditive expression of homoeologous loci has been extensively studied in many interspecific hybrids and allopolyploids, including *Arabidopsis*, *Brassica*, cotton, *Drosophila*, *Senecio*, and wheat (see review and ref. [79]). Although the levels of gene expression detected vary from one experimental species to another, the trends are similar. The levels of differentially expressed genes between the related species are higher than those within species. Over 15–50% of genes are differentially expressed between the related species in plants or animals. The number of nonadditively expressed genes ranges from 5–38% in *Arabidopsis* allotetraploids to ~30% in cotton allotetraploids [80]. In *Senecio*, the number of differentially expressed genes between the natural and synthetic allopolyploids can be as high as ~60% [81], although some of this could be related to genotypic differences between synthetic and natural allopolyploids. In *Arabidopsis* allotetraploids, over 65% of the nonadditively expressed genes are repressed, and over 94% of the repressed genes in the allotetraploids are expressed at higher levels in *A. thaliana* than in *A. arenosa*, consistent with the silencing of *A. thaliana* rRNA genes subjected to nucleolar dominance [77] and with overall suppression of the *A. thaliana* phenotype in the synthetic allotetraploids and natural *A. suecica* [82]. The data suggest transcriptome and phenotypic dominance of *A. arenosa* over *A. thaliana* in the allotetraploids. In cotton, the A-genome ESTs are selectively enriched in the allotetraploid [83], a result consistent with the production of long lint fibers in A-genome species. However, in another study, the expression of homoeologous loci is shifted toward the D-genome species [80], which does not produce spinnable fibers. Moreover, ~20% of the genes show locus-specific expression patterns in different stages of fiber development. The data support the role of developmental regulation in the expression rRNA genes and protein-coding genes found in *Arabidopsis* and *Brassica* allotetraploids [77,82].

Genome-wide gene expression data collectively support the genetic models of dominance and overdominance at the

Box 1. Central role of the circadian clock in plant growth and development

Every organism under the sun lives by day and night with a constant cycle of ~24 h. Plants, in particular, during the day, convert sunlight, water, and carbon dioxide into carbohydrates and eventually biomass, and emit oxygen as a byproduct of photosynthesis. At night, plants store, transport, and use the carbohydrates, and release energy, carbon dioxide, and water as a byproduct of respiration. Moreover, the temperature and growth conditions change during day and night. These rhythmic cycles are known as the circadian clock, which is derived from the Latin words 'circa' (about) and 'dies' (day) [136]. The scientific literature on circadian rhythms began with the daily leaf movements of heliotrope plants even in continuous darkness [137], suggesting an internal circadian rhythm. Figure I (**a**) Internal time keepers or circadian clock regulators include CCA1, LHY, and TOC1 in a major negative feedback loop (Loop I) of the circadian oscillator in *Arabidopsis*, which produces a self-sustaining and constant periodicity of 24 h, even when plants are grown under constant light and temperature. CCA1 Hiking Expedition (CHE) has recently been shown to be a negative regulator of CCA1 [90]. In addition to CCA1, LHY, and TOC1, other regulatory loops include one (Loop III) consisting of PSEUDO-RESPONSE REGULATOR (PRR) 7 and 9, another (Loop II) of GI and unknown protein, and another (Loop IV) of ZEITLUPE (ZTL), GI, and PRR3. Figure I (**b**) Diagram of *CCA1* and *LHY* (red line) and *TOC1* (green line) expression rhythms in a 24-h clock with 16 h of light (open bar) and 8 h of darkness (filled bar). Zeitgeber (ZT) is German for time giver, and dawn is defined as ZT0. Period is the time for completing one cycle of rhythms and is shown from one peak to another (or form one trough to another). The expression amplitude of rhythm is defined as one-half the distance between the peak and trough. Many aspects of plant physiology, metabolism, and development are under circadian control, and a large proportion of transcriptome (from 15% up to ~90%) shows circadian regulation [96,98]. For further information, see the many excellent reviews in the field, including historical perspectives of circadian rhythms [94], how plants tell time [138], regulation of output from the circadian clock [139], and the most recent reviews of circadian systems in higher plants [95,140].

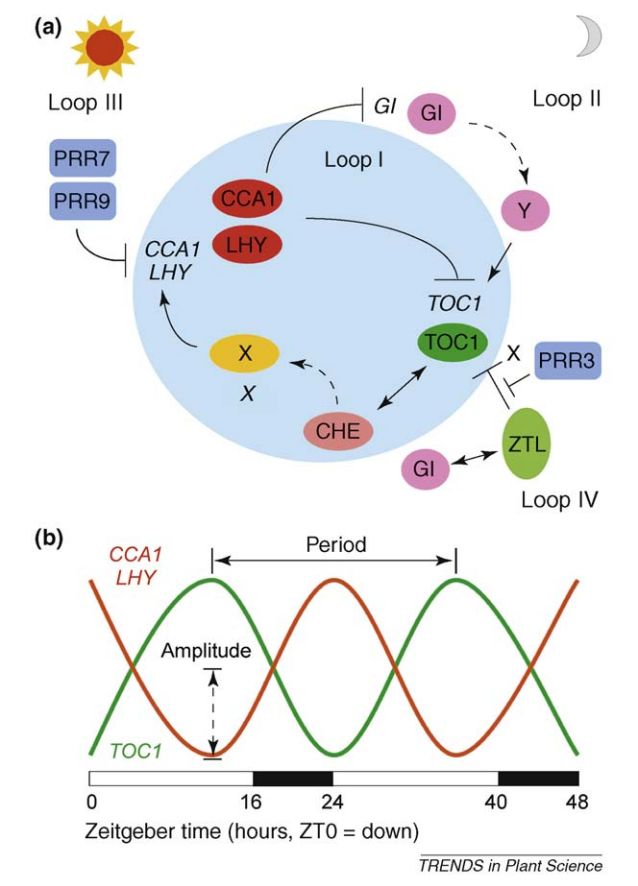

Figure I. Central oscillators of circadian clock and their diurnal expression patterns.

levels of individual genes but do not provide mechanistic insights into the molecular basis for heterosis.

A molecular clock model for growth vigor in hybrids and allopolyploids

At the molecular levels, both dominance and overdominance models suggest nonadditive expression of alleles in the hybrids relative to the parents. The dominant mode of gene expression represents one extreme: monoallelic expression in the hybrids, whereas overdominant mode of gene expression indicates another: biallelic expression in the hybrids at levels either higher than the high-parent value or lower than the low-parent value. Neither the dominance nor the overdominance model can explain the epistatic interactions among different genes and gene products that are involved in the same or different regulatory and/or biological pathways, leading to an altered trait or phenotype. Moreover, heterosis changes over time or during growth and development of plants and animals. For example, heterosis in biomass such as vigorous growth in seedlings, roots, and other vegetative tissues may not be directly translated into large fruits or seeds because different sets of genes in the biological pathways control vegetative growth and reproductive development, although some pathways are intricately related. Therefore, a molecular model for heterosis should define individual genes in specific regulatory pathways. One model is that epigenetic regulation induces nonadditive expression of one or more key regulator genes in the hybrids, which in turn mediates the expression of many other genes in the same regulatory networks associated with changes in developmental and physiological pathways, leading to heterosis in specific stages of growth and development. As a result, nonadditive expression of many genes collectively in various biological pathways gives rise to an overall vigor of vegetative growth and yield.

Circadian clocks affect many physiological and developmental processes, including various metabolic pathways and fitness traits in animals and plants, and photosynthesis and starch metabolism in plants (see Box 1) [84–87]. In *Arabidopsis*, the central oscillators of the circadian clock consist of negative regulators CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) and LATE ELONGATED HYPOCOTYL (LHY) [88,89] and reciprocal positive regulators TIMING OF CAB EXPRESSION 1 (TOC1), CCA1 Hiking Expedition (CHE) [90], and GIGANTEA (GI) [89,91,92]. CHE, a transcription factor belonging to the TCP class, represses *CCA1* expression [90]. CCA1 and LHY are partially redundant MYB-domain transcription factors with incompletely overlapping functions [88,89]. CCA1 and LHY negatively regulate *TOC1* and *GI* expression, whereas TOC1 binds to the CCA1 promoter and interacts with CHE, positively regulating CCA1 and LHY expression [89–91,93]. This circular feedback regulation affects the rhythms,

amplitude, and/or period of the circadian clock as well as its input and output pathways in *Arabidopsis* [94,95]. At least ~15% of genes, including those involved in photosynthesis and starch metabolism [96,97], and up to 90% of transcriptome [98] are affected by the circadian clock regulators. Moreover, day-length and circadian effects on transitory starch degradation and maltose metabolism correlate with the diurnal expression patterns of these metabolic genes [99]. Consequently, maintaining circadian regulation increases CO_2 fixation and growth, whereas disrupting circadian rhythms reduces fitness [87,100].

Analyzing genome-wide nonadditively expressed genes in *Arabidopsis* allotetraploids [82], the authors [101] found that among ~130 genes upregulated in the allotetraploids, two thirds of them in their upstream regions contain at least one (CCA1)-binding site (CBS; AAAAATCT) or evening element (AAAATATCT) [96]. One subset of the genes encodes protochlorophyllide (pchlide) oxidoreductases a and b (PORA and PORB) that mediate the light-requiring step in chlorophyll biosynthesis in higher plants [102]. Both *PORA* and *PORB* are upregulated in the allotetraploids. In *A. thaliana*, *PORA* and *PORB* are expressed at high levels in seedlings and young leaves, and overexpression of *PORA* and *PORB* increases chlorophyll a and b content [103]. The other subset of genes encodes all major enzymes in starch metabolism and sugar transport [104,105], many of which contain EE/CBS and are upregulated in the allotetraploids. As a result, the allotetraploids accumulate ~60% more starch than the low parent and ~30% more than the high parent, and ~70% more chlorophyll than the low parent. The starch amount in the allotetraploids is 3 to 5 times more than the low parent and 70% more than the high parent, and the sugar content is 50–100% more in the allotetraploids than in the parents.

The study further established a direct connection between epigenetic repression of *CCA1* and *LHY* and upregulation of the genes involved in the light-requiring processes of photosynthesis, starch metabolism, and sugar biosynthesis in the hybrids and allopolyploids [101]. This daytime-specific repression of clock genes has an epigenetic cause because it correlates with loss of histone modifications (e.g. H3K9 acetylation and H3K4 dimethylation) that are normally associated with active transcription from the *CCA1* and *LHY* genes. By contrast, upregulation of *TOC1* and *GI* correlates with increased levels of H3K9 acetylation and H3K4 dimethylation. Interestingly, similar repression of *CCA1* and *LHY* and upregulation of *TOC1* are also found in the F_1 hybrids made by crossing C24 and Colombia strains of *A. thaliana* without ploidy changes. However, the levels of changes in gene expression, chlorophyll, and starch content in the hybrids are lower than in the allotetraploids. This observation is consistent with a positive correlation between the levels of heterosis and genetic distances among the parents used in the hybrids. Similar expression changes of a *CCA1-like* gene were observed in maize hybrids to those observed in *Arabidopsis* hybrids [101] (unpublished data).

Altering the clock amplitude but maintaining the rhythmic phase increases growth vigor in the hybrids and allotetraploids (Figure 3). Expressing *TOC1::CCA1* and *TOC1::cca1(RNAi)* in the diploid transgenic plants mimics

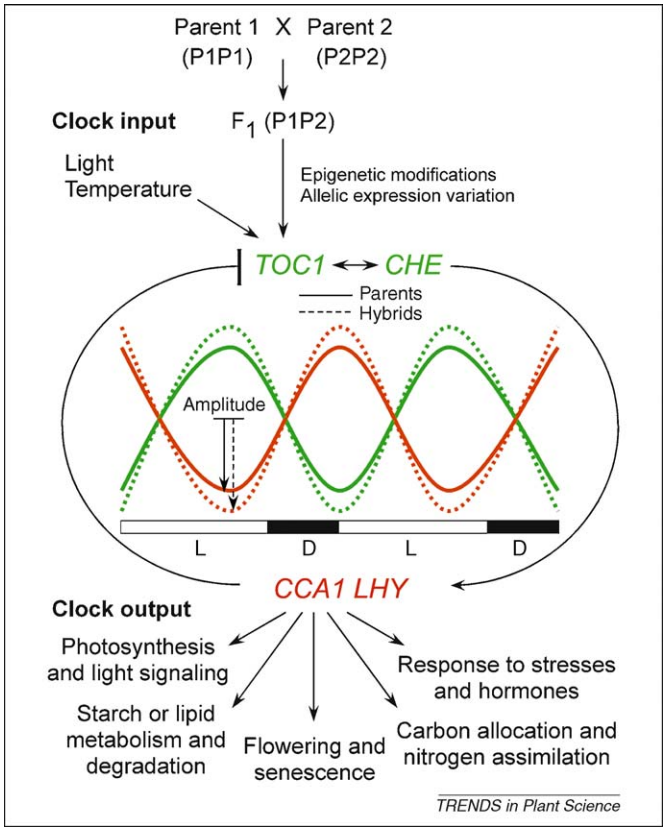

Figure 3. Growing around the clock: a molecular mechanism for hybrid vigor. A molecular clock model explains the basis of heterosis. The internal clocks of plants are controlled by multiple feedback loops, including a major loop that consists of two transcription repressors CCA1 and LHY with redundant but incompletely overlapping functions and feedback regulators TOC1 and CHE (see Box 1). The clock receives input signals such as lights and temperature and controls output traits and pathways, including photosynthesis and light signaling, flowering, starch biosynthesis and metabolism, responses to stresses and hormones, and carbon allocation and nitrogen assimilation, through the expression of evening element (EE) or CCA1 binding site (CBS)-associated genes. The expression amplitude and periodicity of circadian clock regulators can be changed or fine-tuned in response to input (external) signals such as light and temperature, as well as internal mechanisms such as allelic expression variation. L and D indicate the length of light (L) and darkness (D) in a circadian cycle. In the hybrids, the allelic interactions between parent 1 (P1) and parent 2 (P2) induce epigenetic repression of *CCA1* and *LHY* expression amplitudes (red dashed line) and upregulation of *TOC1* expression amplitudes (green dashed line) relative to the expression values in the parents (solid red and green lines, respectively), whereas the periodicity of the clock remains the same [101] because maintaining clock periodicity and rhythm is important for plant growth and fitness [84]. The reduced amount of CCA1 repressors in the hybrids during the day induces the expression of circadian-clock-associated genes (CCGs) in various output pathways, including chlorophyll biosynthesis, and starch metabolism and degradation. As a result, the hybrids produce more chlorophyll and starch than the parents, which promotes vegetative growth and morphological vigor. The *CCA1* expression amplitude is regulated by chromatin modifications, where the levels of active histone marks are reduced during the day and increased at night. The hybrid-induced changes in the *CCA1* expression amplitude are reminiscent of expression alterations in response to changes in input signals such as light (intensities) and temperature. The clock modulates auxin signaling and responses [141]. In addition, the output pathways also produce feedback regulation for the internal clocks. For example, circadian oscillator regulation requires organic nitrogen signals [142] and free cytosolic Ca^{2+} [143]. Allelic interactions in the hybrids induce superior performance of physiological pathways for chlorophyll biosynthesis and starch metabolism. The overdominant performance is caused by epigenetic repression (nonadditive expression) of a key regulator in the feedback loop of the clock oscillator, which mediates the downstream genes in chlorophyll biosynthesis and starch metabolism. Clock-mediated heterosis is probably universal because internal clocks mediate physiological and metabolic pathways in plants and animals. Moreover, this model can be extrapolated to explain superior traits of many other biological pathways.

alteration in the *CCA1* expression amplitude. Repressing or overexpressing *CCA1* under the *TOC1* promoter might also slightly affect rhythmic phase and have pleiotropic (but mild) effects on flowering time and plant growth [101], but these effects may be minimal. Completely knocking out clock-genes affects other aspects of plant growth and development, and the plants may lose their fitness and growth vigor. Although the results obtained in *cca1* and *lhy* mutants also show increased growth vigor in the early stages [101], over time the constant loss of rhythmic phase in the mutants induces many other changes, including flowering time and physiological syndromes, leading to low fitness and small plants in the late stages of development [84]. The mutant plants are likely to develop indirect effects independent of original *cca1 lhy* double mutations such as flowering time defects [106]. Moreover, the genetic interaction between CCA1 or LHY and TOC1 is complex. TOC1 mediates the floral transition in a CCA1 or LHY-dependent manner, whereas CCA1/LHY function upstream of TOC1 in regulating a photomorphogenic process [107]. In *Arabidopsis* C24 × Columbia F_1 hybrids, heterosis for biomass (leaf size and dry shoot mass) is 2–3 times higher at high light intensity than at low and intermediate light intensities [25]. The relative growth rates of the hybrids are high in the early developmental stages under low and intermediate light intensities and constantly high over the entire vegetative phase under high light intensity. The above data suggest other factors such as light intensities and light signaling pathways affect the degree and early onset of heterosis for biomass.

Do the changes in circadian clock genes affect other traits in hybrids? Many life history traits, including plant height and leaf length and number, were coincidently mapped in the locations of *CCA1* (bottom of chromosome 2) and *LHY* (top of chromosome 1) in the RILs derived from L*er* and Cvi [27] (unpublished data). Another locus *CRY2* in the vicinity of *LHY* was also a candidate for fruit length and ovule number but not for other traits [28], suggesting a role of epistatic interactions among *CCA1*, *LHY*, and *CRY2* in life history traits. As noted above, heterosis is manifested in many different forms during growth and development. Other key regulators and/or environmental factors such as light intensities, photoperiod, nutrients, and the conditions for optimal growth can also affect many other pathways and traits such as plant stature, flower size, seed fertility, and yield.

How is the allelic or locus-specific expression of CCA1 and other clock regulators established in the hybrids and allotetraploids? Although allelic expression variation of clock genes has not been studied in the *Arabidopsis* hybrids, the locus-specific expression was observed in two allotetraploid lines examined [101]. In respective parents, *A. thaliana* and *A. arenosa* loci were equally expressed. In the allotetraploids *A. thaliana CCA1* (*AtCCA1*) was repressed 2–3-times more than the *A. arenosa CCA1* (*AaCCA1*) whose expression was slightly reduced. Similarly, the repression of *AaLHY* was 1.5-times more than the *AtLHY* in the allotetraploids. Conversely, *AtTOC1* and *AtGI* loci were upregulated more than *AaTOC1* and *AaGI* in the allotetraploids. The data collectively indicate that *A. thaliana* genes are more sensitive to expression changes (repression or activation) than the homoeologous *A. arenosa* genes through epigenetic modifications in the allotetraploids [82,101,108]. Moreover, both *A. thaliana* C24 and Columbia alleles in the hybrids or both *A. thaliana* and *A. arenosa* loci in the allotetraploids are expressed but either upregulated or repressed relative to the MPV, suggesting a role for expression overdominance or repression in hybrid vigor.

Altering expression of a few genes in the circadian clock regulation to promote growth vigor is reminiscent of single locus heterosis, which has been documented for the *erecta* and *angustifolia* loci in *A. thaliana* [109]. These loci also show an overdominant mode of expression and encode regulatory proteins, namely, a receptor-like kinase [110] and a transcription factor [111], respectively. This offers a solution to clone QTLs that have been extensively studied in the hybrids of *Arabidopsis*, tomato, maize, and rice. For example, the genetic basis of heterosis in an elite rice hybrid is controlled by single-locus heterotic effects and dominance-by-dominance interactions [112].

A good example is the domestication of maize (*Z. mays* spp s. *mays*), which involves a transition of apical dominance (a collection of stem cells for the development of main stem and axillary branches) from its probable wild ancestor, teosinte (*Z. mays* ssp. *parviglumis*). The apical dominance is controlled by a major genetic locus named *teosinte branched 1* (*tb1*), which encodes a protein with homology to the cycloidea in snapdragon. *tb1* represses the growth of axillary organs and promotes the formation of female inflorescences. The maize allele of *tb1* is expressed at twice the level of the teosinte allele, suggesting that gene regulatory changes underlie the evolutionary divergence of maize from teosinte [113]. Another example is the domestication of tomato (Solanaceae). The wild type produces few-flowered inflorescences, but the mutants *compound inflorescence* (*s*) and *anantha* (*an*) are highly branched, and *s* produces hundreds of flowers [51]. The *S* and *AN* encode a homeobox transcription factor and an F-box protein, respectively. Apical dominance and branch formation are controlled by a few regulatory genes, suggesting a molecular basis for single-locus heterosis. However, the connection between the gene function and morphological variation in these studies has yet to be established, and also it is debatable whether the control of inflorescence architecture (e.g. from definite to indefinite) by promoting progression of an inflorescence meristem to floral organs is part of heterosis or developmental variation.

Allelic activation and repression through *cis*- and *trans*-acting effects in hybrids or allopolyploids is reminiscent of paramutation [114,115], X-inactivation [116,117], and repeat-associated gene silencing [118]. However, in hybrids and allopolyploids allelic- and locus-specific expression occurs on a genome-wide scale, which occurs on any chromosomes but does not occur at every locus in a specific chromosome or even in a small chromosomal segment [49]. In some cases, epigenetic regulation is stochastic and takes several generations to establish [48]. In contrast to random inactivation of paternal and maternal X-chromosomes in somatic cells, there is a dominance hierarchy for locus-specific gene expression in allopolyploids. The expression of homoeologous genes, including

rDNA loci, is dominant from one parent over the other in the interspecific hybrids or allopolyploids. The dominance phenomenon is similar to paramutagenic and paramutable alleles in paramutation, but the expression of two alleles and loci in the hybrids and allopolyploids is additive, whereas the paramutagenic allele exerts trans-generational effects on the expression of the paramutable allele. Compared with epigenetic silencing of endogenous repeat gene loci, the alleles or homoeologous loci examined in the hybrids and allopolyploids do not have obvious internal repeats. If epigenetic mechanisms are responsible for allelic- and locus-specific gene expression in hybrids and allopolyploids, they probably operate through *cis*- and *trans*-acting effects [119,120], chromatin modifications, and/or small RNAs that discriminate between homoeologous loci [108,121].

Roles for small RNAs in hybrid vigor and incompatibility in allotetraploids

The above models suggest that epigenetic and transcriptional regulation of key regulatory genes leads to heterosis. Nonadditive gene expression is also controlled by post-transcriptional mechanisms via RNA-mediated pathways [108,122]. Small RNAs, including microRNAs (miRNAs) [123], small interfering RNAs (siRNAs) [124], and *trans*-acting siRNAs (tasiRNAs) [125,126], mediate post-transcriptional regulation, RNA-directed DNA methylation, and chromatin remodeling. miRNAs are produced from genetic loci independent of their targets and serve as negative regulators of gene expression by targeting RNA degradation or translational repression [123]. tasiRNAs arise in plants from specific *TAS* loci that are transcribed into precursors, which are cleaved by miRNA-guided mechanisms. The resulting 21-nt tasiRNAs direct the degradation of target mRNAs [125,126]. miRNAs and tasiRNAs control the expression of genes that encode transcription factors and proteins that are important for growth and development. It is conceivable that different ecotypes and species might have developed specific growth and developmental patterns, which are partly mediated by miRNAs and tasiRNAs. Combination of miRNAs and their targets of different parental origins in the hybrids or new allopolyploid species may reprogram expression of miRNAs and tasiRNAs and their targets [127]. Indeed, many miRNA targets are nonadditively expressed in the allotetraploids [82], suggesting a role for miRNAs in buffering genetic clashes between species [127]. In a recent study using massive parallel sequencing of small RNAs and microarray analysis of miRNAs in resynthesized and natural *Arabidopsis* allotetraploids and their progenitors, the miRNAs and tasiRNAs but not the siRNAs were associated with nonadditive expression of target genes in the allotetraploids [122]. Although the sequences of many miRNAs are conserved, miRNA accumulation levels are nonadditive in the leaves or flowers of interspecific hybrids and allotetraploids relative to the parents. Nonadditive accumulation levels of miRNAs are associated positively with the expression levels of miRNA biogenesis genes such as *AGO1* and *DCL1* but negatively with many miRNA targets. The data suggest that expression variation of miRNAs and their targets in the hybrids and allotetraploids are controlled by epigenetic mechanisms at transcriptional and post-transcriptional levels. The genome merger in the allotetraploids induces epigenetic modifications [108], leading to nonadditive expression of some miRNA targets, miRNA primary transcripts, and miRNA biogenesis genes. At the post-transcriptional level, nonadditive expression of miRNA biogenesis genes can affect the processing efficiency of miRNA precursors, resulting in nonadditive accumulation of miRNAs. Moreover, differential expression of *A. thaliana* and *A. arenosa* miRNAs and their targets in the allotetraploids leads to biased target degradation, probably because the efficiency of target mRNA degradation is dependent on a threshold of miRNA concentration [128]. In addition, although the target loci of different parental origins are conserved, their secondary structures might have diverged, which affects the efficiency of miRNA-triggered degradation [129].

Repeat-associated siRNAs (rasiRNAs) are predominately derived from transposons and repeats and highly enriched in centromeres and heterochomatic regions [130], and diverge rapidly among closely related species. The rasiRNA population is relatively low in F_1, and many rasiRNAs absent in F_1 are restored in late and natural allotetraploids, indicating that it takes several generations to establish stable expression patterns of siRNAs of protein-coding genes [48]. Although the proportion of rasiRNAs is lower in F_1 than in *A. thaliana*, the number of miRNA reads is higher in F_1 than in *A. thaliana*, indicating rapid and dynamic changes of siRNAs and miRNAs in early stages of allopolyploid formation. A few transposons generated new siRNAs in F_1, F_7 allotetraploids, and *A. suecica*. This might be related to sequencing depth or activation of these elements in allopolyploids. Reduction of siRNAs in F_1 may activate some transposable elements in response to 'genomic shock' [131] in marsupial interspecific hybrids [46] and induce genome instability and infertility in *Arabidopsis* allotetraploids [48,132]. siRNA-directed DNA methylation and chromatin modifications are required for the establishment and maintenance of heterochromatin and centromeres [124,130], leading to genome stability. Consistent with the notion, siRNA accumulation is related to DNA hypermethylation of *A. thaliana* homoeologous centromeres in natural allotetraploid *A. suecica* [121]. During F_1 and early stages of allotetraploid formation, genomic shock causes meiotic disorders and genome instability [131], probably resulting from a temporary loss of siRNAs. Over time, genome stability is restored through regeneration of rasiRNAs in genetically stable allotetraploids.

Some rasiRNAs are associated with gene repression in diploids but weakly with gene expression changes between the related species or in allotetraploids. The correlation between siRNA-generating genes and the genes that are nonadditively expressed in the allotetraploids is insignificant, which is consistent with a few genes that are affected by DNA hypomethylation in *A. suecica* [121]. This is because siRNAs are tightly regulated for the maintenance of heterochromatin and genome stability. It is also likely that the majority of nonadditively expressed genes encode proteins, and siRNA-containing transposons and repeats are underrepresented in microarrays [82].

A probable model is that siRNAs are inherited maternally to silence transposons that are reactivated during gametogenesis. The repression of *A. thaliana* homoeologous loci [82] and accumulation of *A. thaliana* centromeric siRNAs [121] are similar to the repression of transposons through maternal transmission of endogenous siRNAs in *Drosophila* [133]. Indeed, interspecific hybrids and allotetraploids can only be produced using *A. thaliana* as the maternal parent [48,132], suggesting an important role of maternal inheritance in overcoming hybrid incompatibility. A recent study has shown that the expression of PolIV-dependent siRNAs (p4siRNAs) is initiated in the female gametophyte and persists during seed development [134], suggesting a role for maternally inherited siRNAs in maintaining genomic stability of the new hybrids and offspring. Unlike conventional imprinting genes, the inheritance of maternal p4siRNAs is independent of DNA methylation. It is proposed that activating factors related to the maternal expression of RNAi genes such as *NRPD1A*, *RDR2*, and *DCL3* are responsible for maternal p4siRNA production. Alternatively, repressive factors in the paternal genome can also be involved. The loss of p4siRNAs in the sperm cells is consistent with expression loss of chromatin remodeling factor DDM1, suggesting transcriptional repression of paternal p4siRNAs during male gamete formation, which persists after fertilization [135].

The rasiRNAs may be directly related to suppression of transposons and indirectly related to genomic stability and growth vigor in the hybrids. The maternal inheritance of p4siRNAs and paternal suppression of rasiRNAs occur only in the hybrids, which may lead to morphological and developmental changes in the hybrids but not in the parents. As a result, both increase in growth vigor and post-zygotic failures are frequently observed in hybrid plants, depending on the presence or absence of rasiRNAs that are required to maintain genome stability and fertility.

Future perspectives

Heterosis or hybrid vigor results from genome-wide changes and interactions between paternal and maternal alleles. Heterozygosity is a prerequisite to changes in gene expression and phenotypic variation in hybrids and allopolyploids. The heterotic effects on gene expression changes in the hybrids can be augmented in polyploids (e.g. diploid versus tetraploid hybrids). Expression alteration of the genes that encode transcription factors and chromatin proteins is expected to cause cascade effects on the expression of downstream genes and their biological processes. In that sense, heterosis can be explained by a single gene or a few genes in the biological pathways. Epigenetic regulation of circadian-mediated changes in chlorophyll biosynthesis and starch metabolism offers one of the direct links to growth vigor in plant hybrids and allopolyploids. Maternal inheritance and paternal suppression of rasiRNAs affect post-zygotic failures and seed fertility and development, whereas reprogramming of miRNAs and tasiRNAs in the hybrids leads to nonadditive phenotypes and growth vigor. Several questions remain to be answered. First, what causes the allelic expression variation in the hybrids and allopolyploids? For example, how and why does the genomic mixture turn down the expression amplitude of circadian clock genes without affecting the duration of internal clocks? Why are the rasiRNAs maternally inherited? How are allelic expression variation and genetic divergence established and maintained? Is heterosis caused by genome-wide chromatin modifications or modifications of a few regulatory genes? Second, how can heterosis be permanently fixed? Apomixis (seed production without paternal genetic material) has been extensively pursued as a means for fixation of hybrid vigor. Doubling chromosomes in hybrids, particularly in the intraspecific or interspecific hybrids, offers an alternative solution to the permanent fixation of hybrid vigor. Finally, many hybrids, particularly intraspecific and interspecific hybrids, cannot survive, probably because of speciation or lethality genes that existed before speciation or diverged after speciation, which cause hybrid incompatibilities. Piwi-piRNA and transposons are associated with germline defects in *Drosophila*, a phenomenon known as hybrid dysgenesis. Hybrid vigor and hybrid incompatibility are two-edges of a magic sword that is hidden in the parents but revealed in the hybrids and allopolyploids. A better understanding of the genes and regulatory mechanisms for polyploidy and hybrid vigor will help us effectively select the best combinations of parents for producing best-performing hybrids and polyploids, as well as genetically manipulate the expression of key regulatory genes in the hybrid and polyploid plants for the increased production of seeds, fruits, biomass, and metabolites, such as carbohydrates, celluloses, sugars, lipids, and oils, for the growing demand for these materials to produce food, feed, and biofuels.

Acknowledgements

I thank many former and current members, including but not limited to Hyeon-Se Lee, Jianlin Wang, Lu Tian, Zhongfu Ni, Meng Chen, Erika Lackey, Misook Ha, Eun-Deok Kim, Danny Ng, Changqing Zhang, Gyoungju Nah, Jie Lu, Marisa Miller, and Dae Kwan Ko, for their invaluable contributions to the research program. I am grateful to Edward Buckler and an anonymous reviewer for their insightful and constructive suggestions to improve the manuscript. I apologize for not citing additional relevant references owing to space limitations. The work was supported by the grants from the National Institutes of Health (GM067015) and the National Science Foundation (DBI0733857 and DBI0624077).

References

1 Darwin, C.R. (1876) *The Effects of Cross- and Self-fertilization in the Vegetable Kingdom*, John Murry
2 Shull, G.H. (1908) The composition of a field of maize. *Amer. Breeders Assoc. Rep.* 4, 296–301
3 East, E.M. (1908) *Inbreeding in corn. In Reports of the Connecticut Agricultural Experiment Station for Years 1907–1908*, Connecticut Agricultural Experiment Station, pp. 419–428
4 Duvick, D.N. (2001) Biotechnology in the 1930s: the development of hybrid maize. *Nat. Rev. Genet.* 2, 69–74
5 Crow, J.F. (1998) 90 years ago: the beginning of hybrid maize. *Genetics* 148, 923–928
6 Cheng, S.H. *et al.* (2007) Progress in research and development on hybrid rice: a super-domesticate in China. *Ann. Bot. (Lond.)* 100, 959–966
7 Karpechenko, G.D. (1927) Polyploid hybrids of Raphanus sativus L. × Brassica oleracea L. *Bull. Appl. Bot.* 17, 305–410
8 Clausen, R.E. and Goodspeed, T.H. (1925) Interspecific hybridization in Nicotiana. II. a tetraploid GLUTINOSA-TABACUM hybrid, an experimental verification of Winge's hypothesis. *Genetics* 10, 278–284

9 Goodspeed, T.H. (1933) Chromosome number and morphology in Nicotiana VI: Chromosome numbers of forty species. *Proc. Natl. Acad. Sci. U. S. A.* 19, 649–653
10 O'Mara, J.G. (1953) Cytogenetics of triticale. *Bot. Rev.* 19, 587–605
11 Guedes-Pinto, H. *et al.* (1996) *Triticale: Today and Tomorrow*, Springer
12 Mallet, J. (2004) Hybridization as an invasion of the genome. *Trends Ecol. Evol.* 20, 229–237
13 Wood, T.E. *et al.* (2009) The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. U. S. A.* 106, 13875–13879
14 Masterson, J. (1994) Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* 264, 421–424
15 Grant, V. (1981) *Plant Speciation*, Columbia University Press
16 Brochmann, C. *et al.* (2004) Polyploidy in arctic plants. *Biol. J. Linn. Soc.* 82, 521–536
17 Soltis, D.E. and Soltis, P.S. (1999) Polyploidy: recurrent formation and genome evolution. *Trends Ecol. Evolu.* 14, 348–352
18 Baumel, A. *et al.* (2001) Molecular investigations in populations of Spartina anglica C.E. Hubbard (Poaceae) invading coastal Brittany (France). *Mol. Ecol.* 10, 1689–1701
19 Abbott, R.J. and Lowe, A.J. (2004) Origins, establishment and evolution of new polyploid species: Senecio cambrensis and S. eboracensis in the British Isles. *Biol. J. Linn. Soc.* 82, 467–474
20 U, N. (1935) Genome analysis in Brassica with special references to the experimental formation of B. napus and peculiar mode of fertilization. *Jpn. J. Genet*. 7, 389–452.
21 Wendel, J.F. and Cronn, R.C. (2003) Polyploidy and the evolutionary history of cotton. *Adv. Agron.* 78, 139–186
22 Salamini, F. *et al.* (2002) Genetics and geography of wild cereal domestication in the near east. *Nat. Rev. Genet.* 3, 429–441
23 Sall, T. *et al.* (2003) Chloroplast DNA indicates a single origin of the allotetraploid Arabidopsis suecica. *J. Evol. Biol.* 16, 1019–1029
24 Dubcovsky, J. and Dvorak, J. (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316, 1862–1866
25 Meyer, R.C. *et al.* (2004) Heterosis of biomass production in Arabidopsis. Establishment during early development. *Plant Physiol.* 134, 1813–1823
26 Rohde, P. *et al.* (2004) Heterosis in the freezing tolerance of crosses between two Arabidopsis thaliana accessions (Columbia-0 and C24) that show differences in non-acclimated and acclimated freezing tolerance. *Plant J.* 38, 790–799
27 Alonso-Blanco, C. *et al.* (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
28 el-Assal, S.E. *et al.* (2004) Pleiotropic effects of the Arabidopsis cryptochrome 2 allelic variation underlie fruit trait-related QTL. *Plant Biol. (Stuttg.)* 6, 370–374
29 Jakobsson, M. *et al.* (2006) A unique recent origin of the allotetraploid species Arabidopsis suecica: Evidence from nuclear DNA markers. *Mol. Biol. Evol.* 23, 1217–1231
30 Comai, L. *et al.* (2003) FISH analysis of meiosis in Arabidopsis allopolyploids. *Chromosome Res.* 11, 217–226
31 Bushell, C. *et al.* (2003) The basis of natural and artificial postzygotic hybridization barriers in Arabidopsis species. *Plant Cell* 15, 1430–1442
32 Charlesworth, B. and Charlesworth, D. (1999) The genetic basis of inbreeding depression. *Genet. Res.* 74, 329–340
33 Muller, H.J. (1925) Why polyploidy is rarer in animals than in plants. *Amer. Nat.* 59, 346–353
34 Mable, B.K. (2004) 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. *Biol. J. Linn. Soc.* 82, 453–466
35 Dowling, T.E. and Secor, C.L. (1997) The role of hybridization and introgression in the diversification of animals. *Annu. Rev. Ecol. Syst.* 28, 593–619
36 Mavarez, J. *et al.* (2006) Speciation by hybridization in Heliconius butterflies. *Nature* 441, 868–871
37 Mallet, J. (2007) Hybrid speciation. *Nature* 446, 279–283
38 Muller, H.J. (1942) Isolating mechanisms, evolution and temperature. *Biol. Symp.* 6, 71–125
39 Dobzhansky, T. (1936) Studies on Hybrid Sterility. II. Localization of Sterility Factors in Drosophila Pseudoobscura Hybrids. *Genetics* 21, 113–135
40 Brideau, N.J. *et al.* (2006) Two Dobzhansky-Muller genes interact to cause hybrid lethality in Drosophila. *Science* 314, 1292–1295
41 Tang, S. and Presgraves, D.C. (2009) Evolution of the Drosophila nuclear pore complex results in multiple hybrid incompatibilities. *Science* 323, 779–782
42 Seidel, H.S. *et al.* (2008) Widespread genetic incompatibility in C. elegans maintained by balancing selection. *Science* 319, 589–594
43 Bomblies, K. *et al.* (2007) Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. *PLoS Biol.* 5, e236
44 Bikard, D. *et al.* (2009) Divergent evolution of duplicate genes leads to genetic incompatibilities within A. thaliana. *Science* 323, 623–626
45 Vrana, P.B. *et al.* (2000) Genetic and epigenetic incompatibilities underlie hybrid dysgenesis in Peromyscus. *Nat. Genet.* 25, 120–124
46 O'Neill, R.J. *et al.* (1998) Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* 393, 68–72
47 Josefsson, C. *et al.* (2006) Parent-dependent loss of gene silencing during interspecies hybridization. *Curr. Biol.* 16, 1322–1328
48 Wang, J. *et al.* (2004) Stochastic and epigenetic changes of gene expression in Arabidopsis polyploids. *Genetics* 167, 1961–1973
49 Lee, H.S. and Chen, Z.J. (2001) Protein-coding genes are epigenetically regulated in Arabidopsis polyploids. *Proc. Natl. Acad. Sci. U. S. A.* 98, 6753–6758
50 East, E.M. (1936) Heterosis. *Genetics* 21, 375–397
51 Lippman, Z.B. *et al.* (2008) The making of a compound inflorescence in tomato and related nightshades. *PLoS Biol.* 6, e288
52 Gore, M.A. *et al.* (2009) A first-generation haplotype map of maize. *Science* 326, 1115–1117
53 McMullen, M.D. *et al.* (2009) Genetic properties of the maize nested association mapping population. *Science* 325, 737–740
54 Birchler, J.A. *et al.* (2003) In search of the molecular basis of heterosis. *Plant Cell* 15, 2236–2239
55 Michaels, S.D. and Amasino, R.M. (1999) FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. *Plant Cell* 11, 949–956
56 Corbesier, L. *et al.* (2007) FT protein movement contributes to long-distance signaling in floral induction of Arabidopsis. *Science* 316, 1030–1033
57 Valverde, F. *et al.* (2004) Photoreceptor regulation of CONSTANS protein in photoperiodic flowering. *Science* 303, 1003–1006
58 Sasaki, A. *et al.* (2002) Green revolution: a mutant gibberellin-synthesis gene in rice. *Nature* 416, 701–702
59 Buckler, E.S. *et al.* (2009) The genetic architecture of maize flowering time. *Science* 325, 714–718
60 Jones, D.F. (1917) Dominance of linked factors as a means of accounting for heterosis. *Genetics* 2, 466–479
61 Bruce, A.B. (1910) The Mendelian theory of heredity and the augmentation of vigor. *Science* 32, 627–628
62 Crow, J.F. (1948) Alternative hypothesis of hybrid vigor. *Genetics* 33, 477–487
63 Semel, Y. *et al.* (2006) Overdominant quantitative trait loci for yield and fitness in tomato. *Proc. Natl. Acad. Sci. U. S. A.* 103, 12981–12986
64 Xiao, J. *et al.* (1995) Dominance is the major genetic basis of heterosis in rice as revealed by QTL analysis using molecular markers. *Genetics* 140, 745–754
65 Li, Z.K. *et al.* (2001) Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. I. Biomass and grain yield. *Genetics* 158, 1737–1753
66 Yu, S.B. *et al.* (1997) Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid. *Proc. Natl. Acad. Sci. U. S. A.* 94, 9226–9231
67 Stuber, C.W. *et al.* (1992) Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. *Genetics* 132, 823–839
68 Luo, L.J. *et al.* (2001) Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. II. Grain yield components. *Genetics* 158, 1755–1771
69 Doebley, J. *et al.* (1995) teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics* 141, 333–346

70 Guo, M. *et al.* (2003) Genome-wide mRNA profiling reveals heterochronic allelic variation and a new imprinted gene in hybrid maize endosperm. *Plant J.* 36, 30–44
71 Guo, M. *et al.* (2004) Allelic variation of gene expression in maize hybrids. *Plant Cell* 16, 1707–1716
72 Stupar, R.M. and Springer, N.M. (2006) Cis-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. *Genetics* 173, 2199–2210
73 Swanson-Wagner, R.A. *et al.* (2009) Paternal dominance of trans-eQTL influences gene expression patterns in maize hybrids. *Science* 326, 1118–1120
74 Auger, D.L. *et al.* (2004) A test for a metastable epigenetic component of heterosis using haploid induction in maize. *Theor. Appl. Genet.* 108, 1017–1023
75 Swanson-Wagner, R.A. *et al.* (2006) All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. *Proc. Natl. Acad. Sci. U. S. A.* 103, 6805–6810
76 Guo, M. *et al.* (2006) Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. *Theor. Appl. Genet.* 113, 831–845
77 Chen, Z.J. and Pikaard, C.S. (1997) Transcriptional analysis of nucleolar dominance in polyploid plants: biased expression/silencing of progenitor rRNA genes is developmentally regulated in Brassica. *Proc. Natl. Acad. Sci. U. S. A.* 94, 3442–3447
78 Adams, K.L. *et al.* (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. U. S. A.* 100, 4649–4654
79 Jackson, S.A. and Chen, Z.J. (2009) Genomic and expression plasicity of polyploidy. *Curr. Opin. Plant Biol.* (in press), doi:10.1016/j.pbi.2009.11.004
80 Hovav, R. *et al.* (2008) Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. *Proc. Natl. Acad. Sci. U. S. A.* 105, 6191–6195
81 Hegarty, M.J. *et al.* (2006) Transcriptome shock after interspecific hybridization in senecio is ameliorated by genome duplication. *Curr. Biol.* 16, 1652–1659
82 Wang, J. *et al.* (2006) Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. *Genetics* 172, 507–517
83 Yang, S.S. *et al.* (2006) Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton. *Plant J.* 47, 761–775
84 Dodd, A.N. *et al.* (2005) Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science* 309, 630–633
85 Wijnen, H. and Young, M.W. (2006) Interplay of circadian clocks and metabolic rhythms. *Annu. Rev. Genet.* 40, 409–448
86 Panda, S. *et al.* (2002) Circadian rhythms from flies to human. *Nature* 417, 329–335
87 Michael, T.P. *et al.* (2003) Enhanced fitness conferred by naturally occurring variation in the circadian clock. *Science* 302, 1049–1053
88 Mizoguchi, T. *et al.* (2002) LHY and CCA1 are partially redundant genes required to maintain circadian rhythms in Arabidopsis. *Dev. Cell* 2, 629–641
89 Alabadi, D. *et al.* (2001) Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. *Science* 293, 880–883
90 Pruneda-Paz, J.L. *et al.* (2009) A functional genomics approach reveals CHE as a component of the Arabidopsis circadian clock. *Science* 323, 1481–1485
91 Strayer, C. *et al.* (2000) Cloning of the Arabidopsis clock gene TOC1, an autoregulatory response regulator homolog. *Science* 289, 768–771
92 Park, D.H. *et al.* (1999) Control of circadian rhythms and photoperiodic flowering by the Arabidopsis GIGANTEA gene. *Science* 285, 1579–1582
93 Wang, Z.Y. and Tobin, E.M. (1998) Constitutive expression of the CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) gene disrupts circadian rhythms and suppresses its own expression. *Cell* 93, 1207–1217
94 McClung, C.R. (2006) Plant circadian rhythms. *Plant Cell* 18, 792–803
95 Harmer, S.L. (2009) The Circadian System in Higher Plants. *Annu. Rev. Plant. Biol.* 60, 357–377
96 Harmer, S.L. *et al.* (2000) Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. *Science* 290, 2110–2113
97 Smith, S.M. *et al.* (2004) Diurnal changes in the transcriptome encoding enzymes of starch metabolism provide evidence for both transcriptional and posttranscriptional regulation of starch metabolism in Arabidopsis leaves. *Plant Physiol.* 136, 2687–2699
98 Covington, M.F. *et al.* (2008) Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biol.* 9, R130
99 Lu, Y. *et al.* (2005) Daylength and circadian effects on starch degradation and maltose metabolism. *Plant Physiol.* 138, 2280–2291
100 Dodd, A.N. *et al.* (2005) The plant clock shows its metal: circadian regulation of cytosolic free Ca^{2+}. *Trends Plant Sci.* 10, 15–21
101 Ni, Z. *et al.* (2009) Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. *Nature* 457, 327–331
102 Reinbothe, S. *et al.* (1996) PORA and PORB, two light-dependent protochlorophyllide-reducing enzymes of angiosperm chlorophyll biosynthesis. *Plant Cell* 8, 763–769
103 Sperling, U. *et al.* (1997) Overexpression of light-dependent PORA or PORB in plants depleted of endogenous POR by far-red light enhances seedling survival in white light and protects against photooxidative damage. *Plant J.* 12, 649–658
104 Lloyd, J.R. *et al.* (2005) Leaf starch degradation comes out of the shadows. *Trends Plant Sci.* 10, 130–137
105 Smith, A.M. *et al.* (2005) Starch degradation. *Annu. Rev. Plant Biol.* 56, 73–98
106 Fujiwara, S. *et al.* (2008) Circadian clock proteins LHY and CCA1 regulate SVP protein accumulation to control flowering in Arabidopsis. *Plant Cell* 20, 2960–2971
107 Ding, Z. *et al.* (2007) A complex genetic interaction between Arabidopsis thaliana TOC1 and CCA1/LHY in driving the circadian clock and in output regulation. *Genetics* 176, 1501–1510
108 Chen, Z.J. (2007) Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu. Rev. Plant Biol.* 58, 377–406
109 Redei, G.P. (1962) Single locus heterosis. *Mol. Gen. Genet.* 93, 164–170
110 Shpak, E.D. *et al.* (2004) Synergistic interaction of three ERECTA-family receptor-like kinases controls Arabidopsis organ growth and flower development by promoting cell proliferation. *Development* 131, 1491–1501
111 Kim, G.T. *et al.* (2002) The ANGUSTIFOLIA gene of Arabidopsis, a plant CtBP gene, regulates leaf-cell expansion, the arrangement of cortical microtubules in leaf cells and expression of a gene involved in cell-wall formation. *EMBO J.* 21, 1267–1279
112 Hua, J. *et al.* (2003) Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. *Proc. Natl. Acad. Sci. U. S. A.* 100, 2574–2579
113 Doebley, J. *et al.* (1997) The evolution of apical dominance in maize. *Nature* 386, 485–488
114 Chandler, V.L. and Stam, M. (2004) Chromatin conversations: mechanisms and implications of paramutation. *Nat. Rev. Genet.* 5, 532–544
115 Mittelsten Scheid, O. *et al.* (2003) Formation of stable epialleles and their paramutation-like interaction in tetraploid Arabidopsis thaliana. *Nat. Genet.* 34, 450–454
116 Lee, J.T. and Jaenisch, R. (1997) The (epi)genetic control of mammalian X-chromosome inactivation. *Curr. Opin. Genet. Dev.* 7, 274–280
117 Lee, J.T. (2005) Regulation of X-chromosome counting by Tsix and Xite sequences. *Science* 309, 768–771
118 Bender, J. and Fink, G.R. (1995) Epigenetic control of an endogenous gene family is revealed by a novel blue fluorescent mutant of Arabidopsis. *Cell* 83, 725–734
119 Wittkopp, P.J. *et al.* (2004) Evolutionary changes in cis and trans gene regulation. *Nature* 430, 85–88
120 Wang, J. *et al.* (2006) Nonadditive regulation of FRI and FLC loci mediates flowering-time variation in Arabidopsis allopolyploids. *Genetics* 173, 965–974
121 Chen, M. *et al.* (2008) RNAi of met1 reduces DNA methylation and induces genome-specific changes in gene expression and centromeric small RNA accumulation in Arabidopsis allopolyploids. *Genetics* 178, 1845–1858

122 Ha, M. *et al.* (2009) Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allopolyploids. *Proc. Natl. Acad. Sci. U. S. A.* 106, 17835–17840
123 Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297
124 Baulcombe, D. (2004) RNA silencing in plants. *Nature* 431, 356–363
125 Vazquez, F. *et al.* (2004) Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. *Mol. Cell* 16, 69–79
126 Peragine, A. *et al.* (2004) SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. *Genes Dev.* 18, 2368–2379
127 Ha, M. *et al.* (2008) Interspecies regulation of microRNAs and their targets. *Biochim. Biophys. Acta* 1779, 735–742
128 Brown, B.D. *et al.* (2007) Endogenous microRNA can be broadly exploited to regulate transgene expression according to tissue, lineage and differentiation state. *Nat. Biotechnol.* 25, 1457–1467
129 Long, D. *et al.* (2007) Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol.* 14, 287–294
130 Lippman, Z. and Martienssen, R. (2004) The role of RNA interference in heterochromatic silencing. *Nature* 431, 364–370
131 McClintock, B. (1984) The significance of responses of the genome to challenge. *Science* 226, 792–801
132 Comai, L. *et al.* (2000) Phenotypic instability and rapid gene silencing in newly formed Arabidopsis allotetraploids. *Plant Cell* 12, 1551–1568
133 Brennecke, J. *et al.* (2008) An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* 322, 1387–1392
134 Mosher, R.A. *et al.* (2009) Uniparental expression of PolIV-dependent siRNAs in developing endosperm of Arabidopsis. *Nature* 460, 283–286
135 Slotkin, R.K. *et al.* (2009) Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* 136, 461–472
136 Halberg, F. *et al.* (1959) Phase relations of 24-hour periodicities in blood corticosterone, mitoses in cortical adrenal parenchyma, and total body activity. *Endocrinology* 64, 222–230
137 de Mairan, J. (1729) Observation botanique. *Hist. Acad. Roy. Sci.* 35–36
138 Gardner, M.J. *et al.* (2006) How plants tell the time. *Biochem. J.* 397, 15–24
139 Yakir, E. *et al.* (2007) Regulation of output from the plant circadian clock. *FEBS J.* 274, 335–345
140 McClung, C.R. (2008) Comes a time. *Curr. Opin. Plant Biol.* 11, 514–520
141 Covington, M.F. and Harmer, S.L. (2007) The circadian clock regulates auxin signaling and responses in Arabidopsis. *PLoS Biol.* 5, e222
142 Gutierrez, R.A. *et al.* (2008) Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1. *Proc. Natl. Acad. Sci. U. S. A.* 105, 4939–4944
143 Dodd, A.N. *et al.* (2007) The Arabidopsis circadian clock incorporates a cADPR-based feedback loop. *Science* 318, 1789–1792

Review

Mendel, 150 years on

T.H. Noel Ellis[1], Julie M.I. Hofer[1], Gail M. Timmerman-Vaughan[2], Clarice J. Coyne[3] and Roger P. Hellens[4]

[1] Institute of Biological, Environmental & Rural Sciences, Aberystwyth University, Gogerddan Campus, Aberystwyth, Ceredigion, SY23 3EB, UK
[2] The New Zealand Institute for Plant & Food Research Ltd, Christchurch 8140, New Zealand
[3] USDA-ARS Western Regional Plant Introduction Station, Washington State University, Pullman, Washington, USA
[4] The New Zealand Institute for Plant & Food Research Ltd, Auckland, New Zealand

Mendel's paper 'Versuche über Pflanzen-Hybriden' is the best known in a series of studies published in the late 18th and 19th centuries that built our understanding of the mechanism of inheritance. Mendel investigated the segregation of seven gene characters of pea (*Pisum sativum*), of which four have been identified. Here, we review what is known about the molecular nature of these genes, which encode enzymes (*R* and *Le*), a biochemical regulator (*I*) and a transcription factor (*A*). The mutations are: a transposon insertion (*r*), an amino acid insertion (*i*), a splice variant (*a*) and a missense mutation (*le-1*). The nature of the three remaining uncharacterized characters (green versus yellow pods, inflated versus constricted pods, and axial versus terminal flowers) is discussed.

Mendel's studies: species, traits and genes

Mendel's paper 'Versuche über Pflanzen-Hybriden' [1] is the best known in a series of studies published in the late 18th and 19th centuries [2–4] that built our understanding of the mechanism of inheritance [5]. The title of Mendel's paper is usually mistranslated in English as 'Experiments in Plant Hybridisation' rather than 'Experiments on Plant Hybrids', reflecting the impact of his work on the science of genetics rather than Mendel's own concern with the nature of hybrids and their implications for the 'Umwandlung einer Art in eine andere' - transformation of one species into another. There is also a misconception, as a result of R.A. Fisher's attack on Mendelism [6], that Mendel's results and experimentation were in some way suspect. These defamatory criticisms include imputations on the scope of his experimental work, his understanding of what he wrote and statistical interpretations of his results; although they have been roundly debunked [7,8], they remain embedded in common opinion.

In his paper, Mendel described eight single gene characters of pea, of which he investigated the segregation of seven. The eighth is the 'purple podded' character determined by the gene *Pur* on linkage group I. He also discussed the segregation of three traits (tall versus short, green versus yellow pods and inflated versus constricted pods) in common bean (*Phaseolus vulgaris*) that are likely orthologues of the corresponding characters he studied in pea. For both species Mendel used additional species names (such as *Phaseolus nanus* or *Pisum saccharatum*). These names are no longer used and we would consider these types as variants – Mendel commented that there is no 'sharp line between the hybrids of species and varieties as between species and varieties themselves'.

From a biological perspective Mendel's genes appear to be an unrelated set of genes that are uninformative about a single process; but they did elucidate the process of genetic inheritance itself. They are therefore important from an historical perspective and they illustrate a diversity of gene functions and types of mutation. Uncovering the molecular basis of these mutations solves a longstanding mystery in genetics.

This review focuses on the identification of four of Mendel's genes (*R/r*, round versus wrinkled seed; *I/i*, yellow versus green cotyledons; *A/a*, coloured versus unpigmented seed coats and flowers; and *Le/le*, long versus short internode length). In addition, the possible natures of three other characters studied by Mendel (*Gp/gp*, green versus yellow pods; *P/p* or *V/v*, inflated versus constricted pods; and *Fa/fa* or *Fas/fas*, axial versus terminal flowers) are discussed.

Linkage

A major conclusion from Mendel's work was that the factors determining individual traits segregated independently of one another. We now know that this is not always the case. The associated segregation of parental allelic combinations, known as genetic linkage, is well established. Fortunately Mendel studied segregation at multiple unlinked loci. This meant his results were not confounded by linkage, which would have been much more difficult to interpret. The issue of linkage is sometimes egregiously combined with the criticism of the quality of Mendel's data to imply falsely that he somehow suppressed inconvenient data [7]. Unfortunately these discussions suffered from confusion in the literature regarding chromosome numbers, linkage data and their combination [9]. Our current view of the position of the genetic loci Mendel studied is presented in Figure 1. As discussed below, there is some uncertainty about the identity of the genes for the fasciated (terminal) flowers (*Fa* or *Fas*) or the constricted pod phenotypes (*P* or *V*); therefore, the map locations of all are indicated. From this distribution of genetic loci it is clear that there are two possible cases where linkage could have confounded Mendel's results: these are *R–Gp* and *Le–V*.

The wrinkled seed character that Mendel studied was *R* versus *r* on linkage group V [10]. The character 'green versus

Corresponding author: Hellens, R.P. (roger.hellens@plantandfood.co.nz).

 1360-1385/$ – see front matter doi:10.1016/j.tplants.2011.06.006

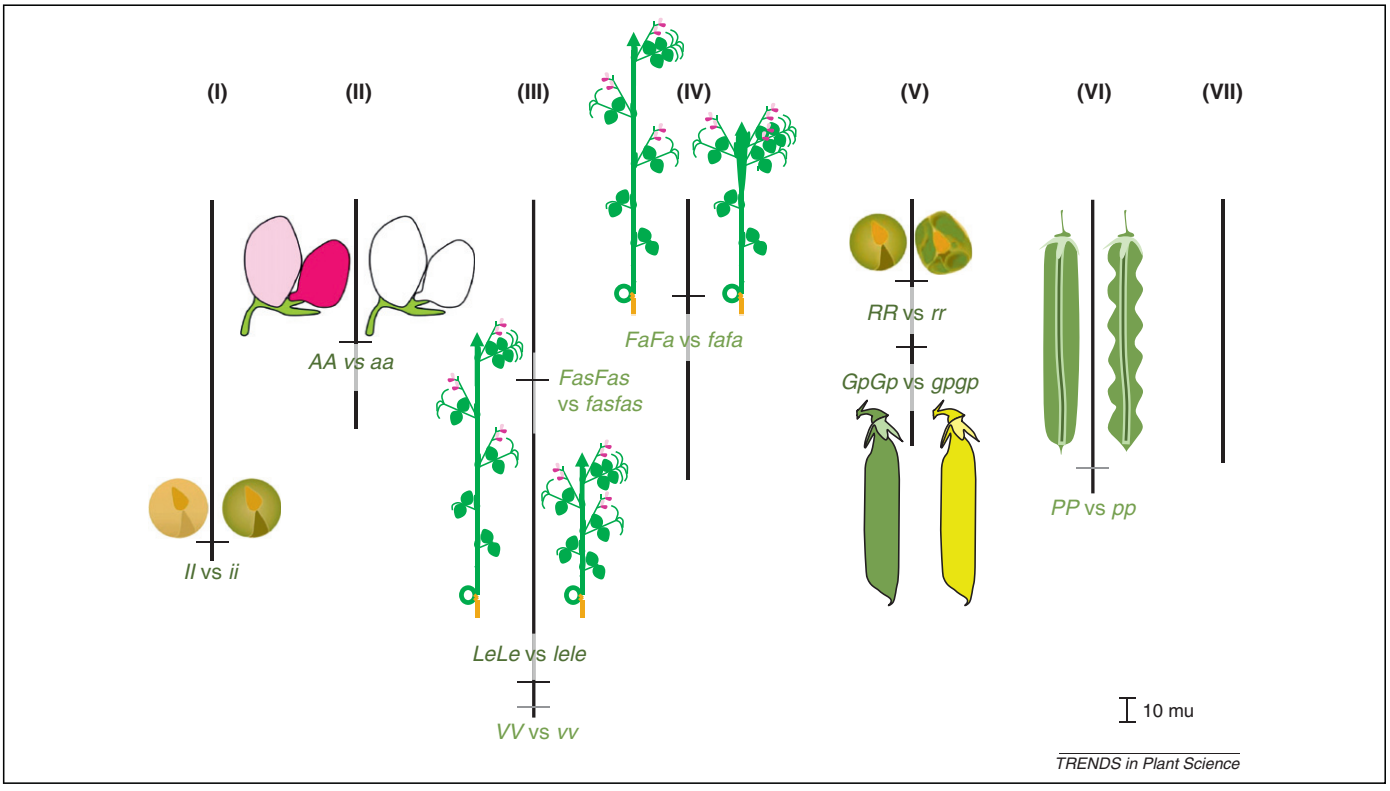

Figure 1. Genetic location of Mendel's seven characters on pea linkage groups. Yellow versus green cotyledons *II*/*ii* on linkage group **(I)**; seed coat (and flower) colour *AA*/*aa* on linkage group **(II)**; tall versus dwarf plants (*LeLe*/*lele*) on linkage group **(III)**; difference in the form of the ripe pods (*PP*/*pp* or *VV*/*vv*) on linkage groups **(III)** and **(VI)**, respectively; difference in the position of the flower (*FasFas*/*fasfas* or *FaFa*/*fafa*) on linkage groups **(III)** or **(IV)**, respectively; round versus wrinkled (*RR*/*rr*) on linkage group **(V)**; and colour of unripe pod (*GpGp*/*gpgp*) on linkage group **(V)**.

yellow pod' is unambiguously *Gp* versus *gp*, also on linkage group V. Linkage between these two loci can be detected [11]. In Mendel's study of two- and three-factor crosses he used approximately 600 F_2 individuals. He did not present data on the combination of *RR GpGp* crossed with *rr gpgp*, but some F_2 plants derived from this cross were probably grown as implied by the text, 'further experiments were made with a smaller number of experimental plants in which the remaining characters by twos and threes were united as hybrids' [1]. In one recombinant inbred population derived from the cross between the inbred John Innes Germplasm lines JI15 and JI399 [12], the recombination fraction between the *R* locus (genotyped using a molecular marker assay) and *Gp* is 36%, resulting in an expected segregation ratio of 9.6:2.4:2.4:1.6 rather than 9:3:3:1. Mendel would have needed about 200 plants in the 'smaller number' to have a 5% statistically significant deviation from independent assortment. Furthermore, linkage group V in pea, most likely corresponding to chromosome 3, behaves unusually in this cross because the number of chiasmata is never greater than one [12]; usually two or three occur. The recombination fraction calculated above is therefore the smallest that Mendel could have encountered, so it is unlikely that genetic linkage would have been discernable in any of the crosses that Mendel examined.

Genes and their mutant alleles

Round versus wrinkled (R *versus* r)

The wrinkled phenotype is striking because plants that appear completely normal bear seeds that are irregular in shape (Figure 1). The immature seeds do not appear unusual, but by maturity there are many differences between the wild-type and mutant seeds. These include diverse features such as subcellular arrangement of organelles, the ratio of the two major types of storage protein, the shape of starch granules, the amylose to amylopectin ratio of the starch polymers and sugar content [13]. There are several genes in pea that confer a wrinkled (rugosus) phenotype and all are lesions in enzymes involved in starch biosynthesis [14–17]. However, only the *r* mutant is known to have been available to Mendel [10].

A biochemical approach was taken to identify the gene encoded by *R* [10]. It was known that *rr* lines were distinguished from wild-type by their reaction to an antibody raised against the starch branching enzyme, so this antibody was used to identify cDNA clones. These cDNAs provided the route to isolating the structural gene encoding a starch branching enzyme (EC 2.4.1.18). Subsequent analysis showed that this gene co-segregated with the *R* locus and that wrinkled (*r*) mutants were disrupted in this gene by the insertion of a non-autonomous type II transposon (called *Ips-r*) related to the Ac/Ds family [10] (Figure 2). Thus the first of Mendel's mutants to be characterized corresponded to a mutation in a gene encoding a biosynthetic enzyme and it was potentially associated with an active transposon. No systematic search for other alleles at the *R* locus has been undertaken and the active and autonomous form of the transposon has not been identified.

Yellow versus green cotyledons (I *versus* i)

Ripe wild-type *II* seeds are yellow because the chlorophyll is lost as the seeds mature, whereas *ii* seeds remain green

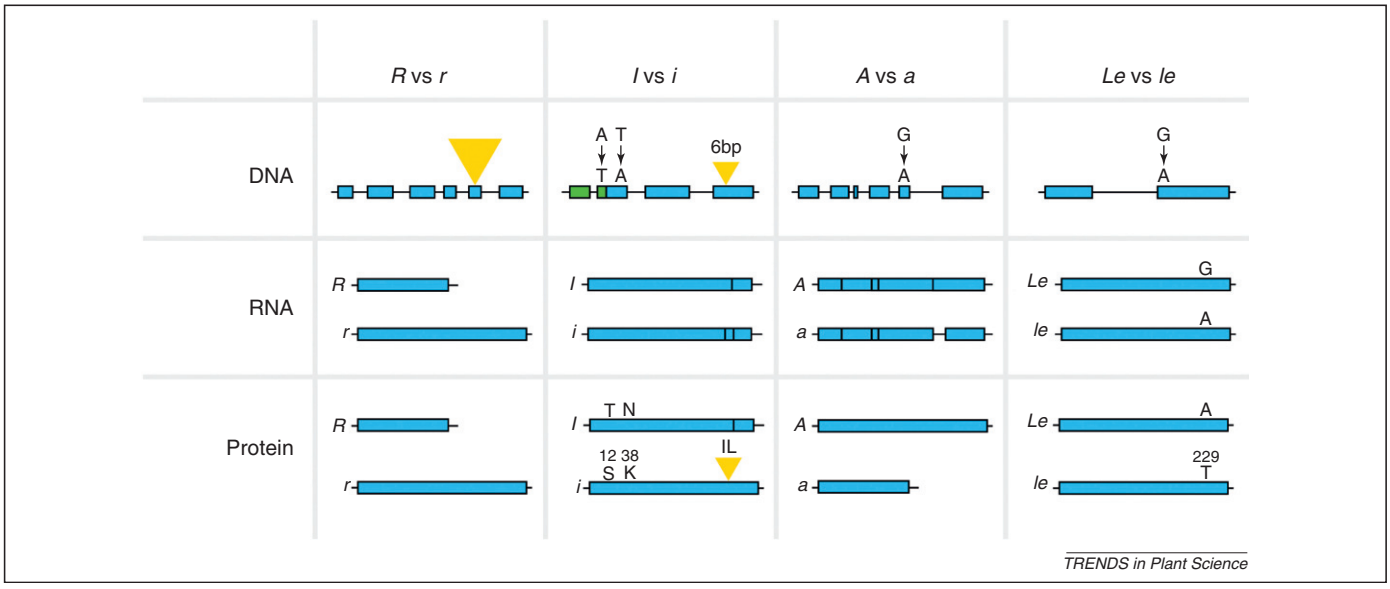

Figure 2. Mutations in Mendel's genes. Round versus wrinkled (*R* vs *r*): encoding starch branching enzyme I (SBEI). In the mutant allele, a transposon is inserted into the open reading frame (large triangle), disrupting both transcription (larger transcript) and translation in mutant lines. Yellow versus green cotyledons (*I* vs *i*): encoding a stay-green protein (SGR). In the mutant allele, a six nucleotide insertion in the coding sequence leads to a two amino acid insertion in the translated protein, disrupting gene function. Other amino acid changes in the signal peptide are not thought to disrupt function. Seed coat (and flower) colour (*A* vs *a*): encoding a basic helix–loop–helix transcription factor (bHLH). In the most common mutant allele, a single nucleotide change at an intron junction disrupts RNA processing leading to a transcript with an additional eight nucleotides and a truncated protein. Tall versus dwarf plants (*Le* vs *le*): encoding gibberellic acid 3-oxidase. A single nucleotide substitution in the coding sequence leads to an alanine (A) to threonine (T) substitution at position 229 that reduces the activity of the enzyme.

(Figure 1). This difference can be seen through the seed coat, but is clearest if the testa is removed. The phenotype is somewhat variable: wild-type seeds that dry out early sometimes retain green colour, whereas green *ii* seeds can sometimes bleach. Chlorophyll is a central component of the plant photosynthetic machinery and the compound responsible for the green colour in plants. A dynamic pathway of chlorophyll biosynthesis and degradation [18] maintains the amount of chlorophyll in photosynthetic tissues and reduces it in low light, during senescence, or other specific phases of plant development.

As green cotyledons are the recessive phenotype, a mutation in the chlorophyll degradation pathway best explains the molecular nature of this trait. Studies in species such as rye grass (*Festuca pratensis*) [19], rice (*Oryza sativa*) [20], maize (*Zea mays*) [21], pepper (*Capsicum annuum*) [22] and *Arabidopsis* (*Arabidopsis thaliana*) [23,24], have identified several genes with 'stay-green' phenotypes, such as: *PAO*, which encodes pheophorbide *a* oxygenase that converts pheophorbide *a* to red chlorophyll catabolite [21–24]; *CBR*, which encodes chlorophyll *b* reductase that converts chlorophyll *b* to chlorophyll *a* [20]; *Stay-Green* (*SGR*), which encodes a protein that is thought to aid the disassembly of light harvesting complex II, allowing chlorophyll to enter the degradation pathway [22,25]; and *PPH*, which encodes pheophytin pheophorbide hydrolase that converts pheophytin *a* to pheophorbide *a* [26].

The first indications that a mutation in a *SGR* gene might be responsible for the *i* mutation were the observations of genetic linkage between a pea orthologue of *SGR* from rice and the *I* locus together with a reduction in the accumulation of *SGR* transcripts in *ii* pea lines [27]. The molecular nature of this lesion was later described [28], and several sequence differences were observed. Two nucleotide differences in the region predicted to function as a signal peptide were initially considered as explanations for the stay-green phenotype because they lead to amino acid substitutions. However, when bombarded into onion (*Allium cepa*) epithelial cells, both the *I* and *i* sequences fused with green fluorescent protein (GFP) were able to target fluorescence into the plastid compartment, indicating that the function of the signal peptide was not compromised by these substitutions. A third sequence difference in *ii* lines consisted of a six-nucleotide (two amino acid) insertion (Figure 2). To assess the consequence of this insertion, a modified form of the rice *SGR2* gene containing the same insertion was transformed into the *sgr-2* mutant. This construct was unable to complement the *sgr-2* mutant rice line, whereas the rice *SGR2* gene was able to complement the mutant. Neither *SGR* allele from pea was able to complement the rice *sgr-2* mutation [28].

SGR appears to direct chlorophyll to the degradation pathway [25]. Although the mechanism by which the protein achieves this is unclear, it seems that a small modification of the protein sequence, as seen in the green cotyledon pea lines, might be sufficient to disrupt the function of the protein.

*Seed coat (and flower) colour (*A *versus* a*)*

The *a* mutation abolishes anthocyanin pigmentation throughout the plant. In pea, as in many other plants, the appearance of red, purple or blue pigments is due to the accumulation of anthocyanin compounds. The different shades of red, purple and blue pigmentation are due to the chemical makeup of the individual anthocyanin compounds, in particular the presence of hydroxyl groups and sugar moieties, together with the pH of the vacuole where they accumulate and other compounds that

complex with the anthocyanins [29–31]. Anthocyanin pigmentation is patterned in space and in response to environmental stimuli such as high light or cold temperatures. Mutants that affect the pattern of pigmentation (such as *Pur* q.v.) are well represented in *Pisum* germplasm but in *a* mutants there is no accumulation of anthocyanin in any part of the plant.

A gene that encodes a basic helix–loop–helix (bHLH) transcription factor was identified as a candidate gene for the *A* locus through comparative genomics [32]. The genetic map of pea was aligned to genomic sequences of *Medicago* (*Medicago truncatula*) using the sequences of cDNA probes known to flank the *A* locus. Annotated genes within about a 10 Mb region of the medicago genome were then scrutinized to identify candidate genes with predicted functions known to influence anthocyanin accumulation. No putative biosynthetic genes were identified in this region. Only one potential regulatory gene, a *bHLH* gene similar to *Arabidopsis TT8* was identified. Degenerate primers designed to the medicago gene were used to isolate the pea orthologue, which was then mapped to linkage group II and shown to co-segregate with the *A* locus (Figure 1). Gene models for this *bHLH* gene were derived from BAC DNA sequences from both coloured and white-flowered lines [32].

Of the 16 single nucleotide polymorphisms (SNPs) identified between the two gene models, the majority (13/16) were silent mutations. Two SNPs predicting amino acid changes were subsequently found in both coloured and white-flowered lines, excluding them as candidates for the causal mutation. The remaining SNP, a G-to-A transition at the splice donor site of intron six of the gene model, occurred only in white-flowered lines. This change interferes with RNA splicing such that eight nucleotides of intron sequence are retained in the processed mRNA, corresponding to a truncated peptide on translation. To confirm that this SNP determined the mutant phenotype, the white-petal phenotype was complemented by transient transformation using biolistics. BAC DNA from both coloured and white-flowered lines was shot into pea petals and coloured foci were observed after introduction of the wild-type but not the mutated gene. Finally, the *A* gene was sequenced from a range of pea germplasm. In this selection, all 60 pea lines with coloured flowers had an intact intron junction (Figure 2), and most but not all (78/88) white-flowered lines had the mutated intron junction. Of the ten remaining white-flowered lines, seven exotic lines carried a different mutation, a single nucleotide insertion in exon six that is predicted to introduce a frameshift and lead to truncation of the protein on translation. No significant deviation from the wild-type sequence has been found in the three other white-flowered lines; however, it is not certain that these three lines are *a* mutants, and the entirety of the gene has not been sequenced [32].

*Tall versus short (*Le *versus* le*)*

Many pea genes are now known to be involved in the synthesis of (*Lh*, *Ls*, *Na*, *Sln* and *Le*) [33–38], or in the response to (*LaI* and *Cry*), the plant hormone gibberellin (GA) [39]. On the basis of its phenotype and distribution among varieties the *Le* gene is considered to be the one studied by Mendel [33,40–42]. *LeLe* plants are tall, *lele* plants are dwarf (Figure 1); this difference is due to internode length rather than the number of nodes.

The *Le* gene product was implicated in GA biosynthesis in early experiments that showed that stem elongation in dwarf seedlings was stimulated by application of GA_3 [43,44]. The activity of the *Le* gene product was established because the conversion of GA_{20} to GA_1 (one of the active forms of GA) was much greater for *LeLe* than for *lele* plants [36], and GA_1 levels were higher in the shoots of *LeLe* versus *lele* plants, whereas GA_{20} amounts were elevated in *lele* plants [45,46]. As a consequence of these studies, it was hypothesized that *Le* encodes a GA 3-oxidase (GA 3β-hydroxylase). GA 3-oxidase activity was shown to be reduced in *lele* plants [45] and subsequent identification of the *Le* gene demonstrated that it encodes a GA 3-oxidase (EC 1.14.11.15) [40,41].

A partial *Le* sequence was obtained by screening a cDNA library at low stringency with *Arabidopsis GA 3-oxidase* (*AtGA4*) probe. This enabled the isolation of full length *le* and *Le* genomic sequences [40]. Sequence alignment revealed a G-to-A transition conferring an alanine-to-threonine substitution at position 229 in the *le-1* gene product (Figure 2). Although this residue is not invariant among plant 2-oxoglutarate-dependent dioxygenases, the class of enzymes to which GA 3-oxidase belongs, it nevertheless lies within a highly conserved region of the protein. Linkage analysis demonstrated co-segregation of the pea *GA3ox* sequence and *Le*. GA 3-oxidase enzymatic activity was demonstrated following recombinant expression of the cDNAs from *Le* and *le-1* plants in *Escherichia coli*. The GA_{20} substrate was converted to GA_1 by both cDNA expression products but the enzyme encoded by *le-1* showed approximately 5% of the activity of the wild-type. The identity of *Le* as *GA3ox* was further supported by the characterization of two additional induced alleles; *le-2* (formerly known as le^d) and *le-3* [41]. The *le-2* mutant was found to carry both the alanine-to-threonine substitution at position 229 found in *le-1*, and a second mutation; a single base deletion of G376, which was inferred to confer a frameshift and premature termination of translation. This mutation at a second site confirms that the *le-2* allele is derived from Mendel's *le-1* allele [47]. The *le-3* line contains a C-to-T transition resulting in a histidine-276 to tyrosine amino acid substitution. The gene products from *Le*, *le-1*, *le-2* and *le-3* GA 3-oxidase clones were assayed following expression in *E. coli*. The relative activities of the recombinant enzymes for two substrates, GA_4 (converted to GA_9) and GA_{20} (converted to GA_1), were $Le > le\text{-}1 \approx le\text{-}3 > le\text{-}2$ [41,47].

The *le-2* allele is likely to be a null allele because the recombinant protein exhibits no activity when GA_{20} is used as a substrate and GA_1 product is measured [47]. Until recently, this was difficult to equate with the *le-2* plant phenotype, which is not an extreme dwarf, and is capable of limited GA_{20} to GA_1 conversion [45]. A second pea *GA 3-oxidase* gene (*GA3ox2*) that is expressed primarily in roots but also in shoots might be responsible for the low level of GA 3-oxidase activity in *le-2* plants [39].

Uncharacterized genes

*Inflated versus constricted pods (*P *versus* p *or* V *versus* v*)*

The inflated versus constricted pod phenotype refers to the presence or absence of a layer of lignified cells (sclerenchyma) adjoining the epidermis of the pod wall and is referred to as parchment (Figure 1). Pods without 'that rough skinny membrane' are described in Gerard's 1597 *Herball* [48], and in general this cell layer is absent in vegetable pea types where the whole pod is eaten (mangetout). Absence of this cell layer leads to a pod that is constricted around the seeds at maturity. There are two sub-types of this class of cultivar, one with a thickened pod wall (*nn*) and no 'string' along the ventral suture (*sin2sin2*) called 'snap' or 'sugar snap' peas. The second type, sometimes called 'snow pea' has thin pod walls (*NN*) and usually has a stringy pod (*Sin2Sin2*). Although breeders commonly combine *NN* with a wrinkled seed character (*rr*) it is difficult to recover vigorous stringless plants with wrinkled seeds [49]. Mendel referred to peas with this pod characteristic as '*P. saccharatum*' suggesting that he used a 'sugar snap' type (probably *rr Sin2Sin2 NN* with either *vv* or *pp*).

It is difficult to be sure which locus Mendel was studying because homozygous individuals carrying mutations in either of the two genes *P* or *V* lack this cell layer [42]. However, we can make some deductions. Mendel studied the segregation of multiple factors in single crosses and although he did not report studies of the joint segregation of 'stem length' and 'pod form' he did report that small-scale experiments combining characters 'by twos and threes' were undertaken. The *V* and *Le* loci are linked, about 15 cM apart on linkage group III [50], so the combination of these two characters in a single cross would have deviated from his expectation for independent segregation. There are therefore two likely possibilities: (i) Mendel studied *vv* in small populations where the deviation from expectation due to linkage was not seen, or (ii) the character state he studied was determined by *pp*; the *P* locus is located on linkage group VI [51] and so would have segregated independently of *Le*.

The 'parchments' of *PP* and *VV* genotypes are secondary cell walls deposited after the cessation of cell growth and are composites of cellulose, hemicelluloses and lignins [52]. Secondary wall biosynthesis has been characterized biochemically and genetically, and studies on transcription factors in *Arabidopsis* indicate that a group of NAC domain proteins and their downstream targets act as regulators [53]. Three homologues of these transcription factor genes are located on chromosome 2 of medicago in regions syntenic with *P* and *V* and are under investigation as candidates for *P* or *V* in pea.

*Green versus yellow pods (*Gp *versus* gp*)*

The *gp* mutation conveys another striking phenotype. *GpGp* plants have green pods whereas *gpgp* plants have yellow pods (Figure 1). Young stems and buds at flowering are also noticeably yellow, whereas leaflets are green as normal. As with the cotyledon colour locus *I*, the green pod/yellow pod *Gp* locus appears as a difference in the accumulation of chlorophyll. In contrast to the *I* locus where the wild-type dominant form is yellow and the recessive mutant form is green, for the *Gp* locus the wild-type dominant form is green and the recessive mutant form is yellow. This suggests that the mutant form *i* represents a failure of chlorophyll degradation, whereas the mutant form *gp* fails to develop a normal chlorophyll complex in the pods [54,55]. Interestingly the region of the medicago genome syntenic to *Gp* contains a gene *Medtr7g080590* annotated as 'chloroplast lumenal protein related' which has similarity to the *Arabidopsis LCD1* gene, mutants of which have a pale phenotype under standard growth conditions and bleach in response to ozone [56]. This phenotype has some similarities to *gp* suggesting that it is a candidate worth investigating further.

*Axial versus terminal flowers (*Fa *versus* fa *or* Fas *versus* fas*)*

The position of flowers, and hence the seeds, on a crop plant is of great importance in agriculture. Several genes determine flower location in pea. Homozygous mutants carrying the *det* gene are determinate with a terminal inflorescence. This mutation has been characterized at the molecular level [57]; however, it is most unlikely that this was the gene studied by Mendel because he described the mutant form as having 'a false umbel', implying a fasciated type with a broadened stem and a 'crown' of many flowers. Alleles of fasciation genes have been widely used in pea breeding, particularly in conjunction with a mutation that confers synchronicity in flowering time [58].

In pea, mutations at several different loci are known to confer a fasciated phenotype; of these, the genes *Fa* (linkage group IV) and *Fas* (linkage group III) are two that are not also defective in nodulation [59]. Mendel would most likely have noticed the yellowness of plants defective in nodulation, thus *Fa* and *Fas* are contenders for the 'difference in the position of the flowers' character. The *Fa* locus has been conventionally assigned to Mendel's trait, but the evidence for this is not definitive. The distance between *Fas* and *Le* on linkage group III is sufficiently large that linkage would have been difficult to detect without intervening markers.

In general, stem fasciation is thought to result from failure of cellular organization within the shoot apical meristem. In *Arabidopsis*, several classes of genes are known to contribute to this organization and to show loss-of-function phenotypes that include shoot fasciation. These include small secreted peptides encoded by *CLAVATA3 (CLV3)/ENDOSPERM SURROUNDING REGION (ESR)*-related genes, which are known to act as ligands for transmembrane proteins such as CLV1 in the shoot apical meristem [60]. These interactions transmit a signal that keeps the stem cell population in check. A failure in the CLV signalling pathway leads to increased stem cell accumulation, as seen in the fasciated phenotypes of *clv1* and *clv3* mutants [61–63]. *CLV*-related genes are therefore obvious candidates for *Fa* and *Fas* in pea, as well as genes that affect the cell cycle. Although many cell cycle mutants are embryo-lethal, several have been characterized in *Arabidopsis* that are viable and have a fasciated shoot phenotype. Among these are the *atbrca2* mutants, which carry lesions in a homologue of the breast tumour susceptibility factor BRCA2 [64], and the *fasciata1* (*fas1*)

and *fas2* mutants [65], which are affected in the genes encoding the p150 and p60 subunits of chromatin assembly factor 1, respectively. A BLAST search of the region of the medicago genome syntenic with *Fa* shows that it contains a homologue of *CLV1* whereas the region syntenic with *Fas* contains a *CLV1* homologue and a *BRCA2* homologue.

Conclusion

As 150 years have elapsed since Mendel's experiments [6,32], it is difficult to state with certainty that the alleles he studied have been identified. In this respect, the diversity of mutant alleles can be informative: lines studied by Mendel must have carried spontaneous mutations. We should also bear in mind that multiple independent spontaneous mutations are unlikely because spontaneous mutation rates are very low with respect to the time since domestication. We do not know the number of different spontaneous alleles for *r* in pea germplasm, but for *Le* the *le-2* allele appears to be derived from *le-1* (and *le-3* is an induced mutation) so there has been a single introduction of this dwarf trait into cultivars. The diversity of mutant *a* alleles has been studied and one is predominant in cultivated lines. A second rare allele restricted to a small subset of landraces is known. This again suggests a single introduction of this character into modern cultivars, or cultivars available in Mendel's time. In contrast, several spontaneous *i* alleles exist, suggesting independent introductions of this trait, which seems remarkable. The types of lesion in Mendel's mutants are various: transposon insertion (*r*), missense mutation (*le-1*), splice variant (*a*) and amino acid insertion (*i*). The mutations affect diverse biological processes; two genes encode enzymes (*R*, *Le*), one is a regulator of a biochemical pathway (*I*) and the most recently described (*A*) is a transcription factor of a family first known for its role in cancer biology. So far, a range of different approaches has been used to identify four of Mendel's seven loci; new comparative genomic tools have identified candidates for the three remaining loci.

Acknowledgements

We thank Rebecca McGee, Andrew Allan and William Laing for helpful comments on this manuscript.

References

1 Mendel, G. (1866) Versuche über pflanzen-hybriden. *Verhandlungen der naturfoschung Vereins* 4, 3–47
2 Knight, T. (1799) An Account of Some Experiments on the Fecundation of Vegetables. In a Letter from Thomas Andrew Knight, Esq. to the Right Hon. Sir Joseph Banks, K.B. P. R. S. *Philos. Transact. R. Soc. Lond.* (1776–1886) 89, 195–204
3 Goss, J. (1824) On variation in the colour of peas, occasioned by cross impregnation. *Hort. Trans.* 5, 234–237
4 Sageret, A. (1826) Considérations sur la Production des Hybrides, des Variantes et des Variétés en General et sur celles de la Famille des Cucurbitacées en Particulier. *Annales des Sciences Naturelles*, *1st ser.* 8, 294–314
5 Olby, R. (1966) *Origins of Mendelism*, University of Chicago Press
6 Fisher, R.A. (1936) Has Mendel's work been rediscovered? *Ann. Sci.* 1, 115–137
7 Fairbanks, D.J. and Rytting, B. (2001) Mendelian controversies: a botanical and historical review. *Am. J. Bot.* 88, 737–752
8 Hartl, D.L. and Fairbanks, D.J. (2007) Mud sticks: On the alleged falsification of Mendel's data. *Genetics* 175, 975–979
9 Ellis, T.H.N. and Poyser, S.J. (2002) An integrated and comparative view of pea genetic and cytogenetic maps. *New Phytol.* 153, 17–25
10 Bhattacharyya, M.K. *et al.* (1990) The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. *Cell* 60, 115–122
11 Rozov, S.M. *et al.* (1993) A new version of pea linkage group 5. *Pisum Genet.* 25, 46–51
12 Hall, K.J. *et al.* (1997) The relationship between genetic and cytogenetic maps of pea. II. Physical maps of linkage mapping populations. *Genome* 40, 755–769
13 Wang, T.L. and Hedley, C.L. (1993) Genetic and developmental analysis of the seed in peas. In *Genetics, Molecular Biology and Biotechnology* (Casey, R. and Davies, D.R., eds), CAB International
14 Martin, C. and Smith, A.M. (1995) Starch Biosynthesis. *Plant Cell* 7, 971–985
15 Harrison, C.J. *et al.* (2000) The *rug3* locus of pea encodes plastidial phosphoglucomutase. *Plant Physiol.* 122, 1187–1192
16 Craig, J. *et al.* (1999) Mutations at the *rug4* locus alter the carbon and nitrogen metabolism of pea plants through an effect on sucrose synthase. *Plant J.* 17, 353–362
17 Bogracheva, T.Y. *et al.* (1999) The effect of mutant genes at the *r*, *rb*, *rug3*, *rug4*, *rug5* and *lam* loci on the granular structure and physicochemical properties of pea seed starch. *Carbohydr. Polymers* 39, 303–314
18 Eckhardt, U. *et al.* (2004) Recent advances in chlorophyll biosynthesis and breakdown in higher plants. *Plant Mol. Biol.* 56, 1–14
19 Moore, B.J. *et al.* (2005) Molecular tagging of a senescence gene by introgression mapping of a stay-green mutation from *Festuca pratensis*. *New Phytol.* 165, 801–806
20 Kusaba, M. *et al.* (2007) Rice *NON-YELLOW COLORING1* is involved in light-harvesting complex II and grana degradation during leaf senescence. *Plant Cell* 19, 1362–1375
21 Gray, J. *et al.* (1997) A novel suppressor of cell death in plants encoded by the *Lls1* gene of maize. *Cell* 89, 25–31
22 Borovsky, Y. and Paran, I. (2008) Chlorophyll breakdown during pepper fruit ripening in the chlorophyll retainer mutation is impaired at the homolog of the senescence-inducible stay-green gene. *Theor. Appl. Genet.* 117, 235–240
23 Pružinská, A. *et al.* (2003) Chlorophyll breakdown: pheophorbide a oxygenase is a Rieske-type iron-sulfur protein, encoded by the *ACCELERATED CELL DEATH 1* gene. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15259–15264
24 Tanaka, R. *et al.* (2003) The Arabidopsis-accelerated cell death gene *ACD1* is involved in oxygenation of pheophorbide a: Inhibition of the pheophorbide a oxygenase activity does not lead to the "stay-green" phenotype in Arabidopsis. *Plant Cell Physiol.* 44, 1266–1274
25 Aubry, S. *et al.* (2008) Stay-green protein, defective in Mendel's green cotyledon mutant, acts independent and upstream of pheophorbide a oxygenase in the chlorophyll catabolic pathway. *Plant Mol. Biol.* 67, 243–256
26 Schelbert, S. *et al.* (2009) Pheophytin pheophorbide hydrolase (pheophytinase) is involved in chlorophyll breakdown during leaf senescence in Arabidopsis. *Plant Cell* 21, 767–785
27 Armstead, I. *et al.* (2007) Cross-species identification of Mendel's *I* locus. *Science* 315, 73
28 Sato, Y. *et al.* (2007) Mendel's green cotyledon gene encodes a positive regulator of the chlorophyll-degrading pathway. *Proc. Natl. Acad. Sci. U.S.A.* 104, 14169–14174
29 Harborne, J.B. (1967) *Comparative Biochemistry of the Flavonoids*, Academic Press
30 Brouillard, R. and Dangles, O. (1993) Flavonoids and flower colour. In *The Flavonoids: Advances in Research since 1986* (Harborne, J.B., ed.), pp. 565–587, Chapman & Hall
31 Yoshida, K. *et al.* (2009) Blue flower color development by anthocyanins: from chemical structure to cell physiology. *Nat. Prod. Rep.* 26, 884–915
32 Hellens, R.P. *et al.* (2010) Identification of Mendel's white flower character. *PLoS ONE* 5, e13230
33 Blixt, S. (1972) Mutation genetics in *Pisum*. *Agri. Hort. Genet.* 30, 1–293
34 Davidson, S.E. *et al.* (2004) The pea gene *LH* encodes ent-kaurene oxidase. *Plant Physiol.* 134, 1123–1134
35 Davidson, S.E. *et al.* (2003) The pea gene *NA* encodes ent-kaurenoic acid oxidase. *Plant Physiol.* 131, 335–344

36 Ingram, T.J. and Reid, J.B. (1987) Internode length in *Pisum*. Gene *na* may block gibberellin synthesis between ent-7α-hydroxykaurenoic acid and gibberellin A_{12}-aldehyde. *Plant Physiol.* 83, 1048–1053

37 Martin, D.N. *et al.* (1999) The *SLENDER* gene of pea encodes a gibberellin 2-oxidase. *Plant Physiol.* 121, 775–781

38 Reid, J.B. and Potts, W.C. (2011) Internode length in *Pisum*. Two further mutants, *lh* and *ls*, with reduced gibberellin synthesis, and a gibberellin insensitive mutant, *lk*. *Physiol. Plant.* 66, 417–426

39 Weston, D.E. *et al.* (2008) The pea DELLA proteins LA and CRY are important regulators of gibberellin synthesis and root growth. *Plant Physiol.* 147, 199–205

40 Lester, D.R. *et al.* (1997) Mendel's stem length gene (*Le*) encodes a gibberellin 3β-hydroxylase. *Plant Cell* 9, 1435–1443

41 Martin, D.N. *et al.* (1997) Mendel's dwarfing gene: cDNAs from the *Le* alleles and function of the expressed proteins. *Proc. Natl. Acad. Sci. U.S.A.* 94, 8907–8911

42 White, O.E. (2011) The present state of knowledge of heredity and variation in peas. *Proc. Am. Phil. Soc.* 56, 487–588

43 Brian, P.W. (1957) The effects of some microbial metabolic products on plant growth. *Symp. Soc. Exp. Biol.* 11, 166–181

44 Brian, P.W. and Hemming, H.G. (1955) The effect of gibberellic acid on shoot growth of pea seedlings. *Physiol. Plant.* 8, 669–681

45 Ross, J.J. *et al.* (1989) Internode length in *Pisum*. Estimation of GA1 levels in genotypes *Le*, *le* and le^d. *Physiol. Plant.* 76, 173–176

46 Proebsting, W.M. *et al.* (1992) Gibberellin concentration and transport in genetic lines of pea: effects of grafting. *Plant Physiol.* 100, 1354–1360

47 Lester, D.R. *et al.* (1999) The influence of the null *le-2* mutation on gibberellin levels in developing pea seeds. *Plant Growth Regul.* 27, 83–89

48 Gerard, J. (1597) *Herball or Generall Historie of Plantes*, Norton

49 McGee, R.J. and Baggett, J.R. (1992) Inheritance of stringless pod in *Pisum sativum* L. *J. Am. Soc. Hort. Sci.* 117, 628–632

50 Lamprecht, H. and Mrkos, H. (1950) Die vererbung des vorblattes bei Pisumsowie die koppelung des gens br. *Agri. Hort. Genet.* 9, 153–162

51 Gritton, E.T. and Hagedorn, D.J. (1975) Linkage of genes *sbm* and *wlo* in peas. *Crop Sci.* 15, 447–448

52 Reiter, W-D. (2002) Biosynthesis and properties of the plant cell wall. *Curr. Opin. Plant Biol.* 5, 536–542

53 Zhong, R. *et al.* (2008) A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis. *Plant Cell* 20, 2763–2782

54 Price, D.N. *et al.* (1988) The effect of the *Gp* gene on fruit development in *Pisum sativum*-L.1. Structural and physical aspects. *New Phytol.* 110, 261–269

55 Price, D.N. and Hedley, C.L. (1988) The effect of the *gp* gene on fruit development in *Pisum sativum*-L.2. Photosynthetic implications. *New Phytol.* 110, 271–277

56 Barth, C. and Conklin, P.L. (2003) The lower cell density of leaf parenchyma in the *Arabidopsis thaliana* mutant *lcd1-1* is associated with increased sensitivity to ozone and virulent *Pseudomonas syringae*. *Plant J.* 35, 206–218

57 Foucher, F. *et al.* (2003) *DETERMINATE* and *LATE FLOWERING* are two *TERMINAL FLOWER1/CENTRORADIALIS* homologs that control two distinct phases of flowering initiation and development in pea. *Plant Cell* 15, 2742–2754

58 Weller, J.L. *et al.* (2009) Update on the genetic control of flowering in garden pea. *J. Exp. Bot.* 60, 2493–2499

59 Sinjushin, A.A. and Gostimskii, S.A. (2007) Relationship between different fasciated lines of pea. *Pisum Genet.* 39, 16–18

60 Miwa, H. *et al.* (2009) Plant meristems: CLAVATA3/ESR-related signaling in the shoot apical meristem and the root apical meristem. *J. Plant Res.* 122, 31–39

61 Clark, S.E. *et al.* (1995) CLAVATA3 is a specific regulator of shoot and floral meristem development affecting the same processes as CLAVATA1. *Development* 121, 2057–2067

62 Clark, S.E. *et al.* (1997) The *CLAVATA1* gene encodes a putative receptor kinase that controls shoot and floral meristem size in Arabidopsis. *Cell* 89, 575–585

63 Fletcher, J.C. *et al.* (1999) Signaling of cell fate decisions by CLAVATA3 in Arabidopsis shoot meristems. *Science* 283, 1911–1914

64 Abe, K. *et al.* (2009) Inefficient double-strand DNA break repair is associated with increased fasciation in Arabidopsis *BRCA2* mutants. *J. Exp. Bot.* 60, 2751–2761

65 Kaya, H. *et al.* (2001) *FASCIATA* genes for Chromatin Assembly Factor-1 in *Arabidopsis* maintain the cellular organization of apical meristems. *Cell* 104, 131–142

Science
AAAS

Breeding Technologies to Increase Crop Production in a Changing World

Mark Tester, *et al.*

Science **327**, 818 (2010);
DOI: 10.1126/science.1183700

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

22. Forum for Agricultural Research in Africa, *Framework for African Agricultural Productivity* (Forum for Agricultural Research in Africa, Accra, Ghana, 2006).
23. K. Anderson, Ed., *Distortions to Agricultural Incentives, a Global Perspective 1955-2007* (Palgrave Macmillan, London, 2009).
24. J. N. Pretty, A. S. Ball, T. Lang, J. I. L. Morison, *Food Policy* **30**, 1 (2005).
25. G. C. Nelson *et al*., *Climate Change: Impact on Agriculture and Costs of Adaptation* (International Food Policy Research Institute, Washington, DC, 2009).
26. N. Stern, *The Economics of Climate Change* (Cambridge Univ. Press, Cambridge, 2007).
27. J. N. Pretty *et al*., *Environ. Sci. Technol.* **40**, 1114 (2006).
28. P. Hazell, S. Wood, *Philos. Trans. R. Soc. London Ser. B Biol. Sci.* **363**, 495 (2008).
29. K. Deininger, G. Feder, *World Bank Res. Obs.* **24**, 233 (2009).
30. P. Collier, *Foreign Aff.* **87**, 67 (2008).
31. L. Cotula, S. Vermeulen, L. Leonard, J. Keeley, *Land Grab or Development Opportunity? Agricultural Investment and International Land Deals in Africa* [International Institute for Environment and Development (with FAO and International Fund for Agricultural Development), London, 2009].
32. A. Aksoy, J. C. Beghin, Eds., *Global Agricultural Trade and Developing Countries* (World Bank, Washington, DC, 2005).
33. R. A. Gilbert, J. M. Shine Jr., J. D. Miller, R. W. Rice, C. R. Rainbolt, *Field Crops Res.* **95**, 156 (2006).
34. IAASTD, *International Assessment of Agricultural Knowledge, Science and Technology for Development: Executive Summary of the Synthesis Report*, www.agassessment.org/index.cfm?Page=About_IAASTD&ItemID=2 (2008).
35. P. G. Lemaux, *Annu. Rev. Plant Biol.* **60**, 511 (2009).
36. D. Lea, *Ethical Theory Moral Pract.* **11**, 37 (2008).
37. Cabinent Office, *Food Matters: Towards a Strategy for the 21st Century* (Cabinet Office Stategy Unit, London, 2008).
38. Waste and Resources Action Programme (WRAP), *The Food We Waste* (WRAP, Banbury, UK, 2008).
39. T. Stuart, *Uncovering the Global Food Scandal* (Penguin, London, 2009).
40. FAO, www.fao.org/english/newsroom/factfile/IMG/FF9712-e.pdf (1997).
41. California Integrated Waste Management Board, www.ciwmb.ca.gov/FoodWaste/FAQ.htm#Discards (2007).
42. FAO, *World Agriculture Towards 2030/2050* (FAO, Rome, Italy, 2006).
43. FAO, *World Agriculture Towards 2030/2050* (FAO, Rome, Italy, 2003).
44. M. D. Smith *et al*., *Science* **327**, 784 (2010).
45. A. G. J. Tacon, M. Metian, *Aquaculture* **285**, 146 (2008).
46. D. Whitmarsh, N. G. Palmieri, in *Aquaculture in the Ecosystem*, M. Holmer, K. Black, C. M. Duarte, N. Marba, I. Karakassis, Eds. (Springer, Berlin, Germany, 2008).
47. P. R. Hobbs, K. Sayre, R. Gupta, *Philos. Trans. R. Soc. London Ser. B Biol. Sci.* **363**, 543 (2008).
48. W. Day, E. Audsley, A. R. Frost, *Philos. Trans. R. Soc. London Ser. B Biol. Sci.* **363**, 527 (2008).
49. J. Gressel, *Genetic Glass Ceilings* (Johns Hopkins Univ. Press, Baltimore, 2008).
50. FAO, Livestock's Long Shadow (FAO, Rome, Italy, 2006).
51. C. P. Reij, E. M. A. Smaling, *Land Use Policy* **25**, 410 (2008).
52. UNEP, *Africa: Atlas of Our Changing Environment* (UNEP, Nairobi, Kenya, 2008).
53. The authors are members of the U.K. Government Office for Science's Foresight Project on Global Food and Farming Futures. J.R.B. is also affiliated with Imperial College London. D.L. is a Board Member of Plastid AS (Norway) and owns shares in AstraZeneca Public Limited Company and Syngenta AG. We are grateful to J. Krebs and J. Ingrahm (Oxford), N. Nisbett and D. Flynn (Foresight), and colleagues in Defra and DfID for their helpful comments on earlier drafts of this manuscript. If not for his sad death in July 2009, professor Mike Gale (John Innes Institute, Norwich, UK) would also have been an author of this paper.

10.1126/science.1185383

REVIEW

Breeding Technologies to Increase Crop Production in a Changing World

Mark Tester* and Peter Langridge

To feed the several billion people living on this planet, the production of high-quality food must increase with reduced inputs, but this accomplishment will be particularly challenging in the face of global environmental change. Plant breeders need to focus on traits with the greatest potential to increase yield. Hence, new technologies must be developed to accelerate breeding through improving genotyping and phenotyping methods and by increasing the available genetic diversity in breeding germplasm. The most gain will come from delivering these technologies in developing countries, but the technologies will have to be economically accessible and readily disseminated. Crop improvement through breeding brings immense value relative to investment and offers an effective approach to improving food security.

Although more food is needed for the rapidly growing human population, food quality also needs to be improved, particularly for increased nutrient content. In addition, agricultural inputs must be reduced, especially those of nitrogenous fertilizers, if we are to reduce environmental degradation caused by emissions of CO_2 and nitrogenous compounds from agricultural processes. Furthermore, there are now concerns about our ability to increase or even sustain crop yield and quality in the face of dynamic environmental and biotic threats that will be particularly challenging in the face of rapid global environmental change. The current diversion of substantial quantities of food into the production of biofuels puts further pressure on world food supplies (*1*).

Breeding and agronomic improvements have, on average, achieved a linear increase in food production globally, at an average rate of 32 million metric tons per year (*2*) (Fig. 1). However, to meet the recent Declaration of the World Summit on Food Security (*3*) target of 70% more food by 2050, an average annual increase in production of 44 million metric tons per year is required (Fig. 1), representing a 38% increase over historical increases in production, to be sustained for 40 years. This scale of sustained increase in global food production is unprecedented and requires substantial changes in methods for agronomic processes and crop improvement. Achieving this increase in food production in a stable environment would be challenging, but is undoubtedly much more so given the additional pressures created by global environmental changes.

Global Environmental Change Alters Breeding Targets

Certain aspects of global environmental change are beneficial to agriculture. Rising CO_2 acts as a fertilizer for C3 crops and is estimated to account for approximately 0.3% of the observed 1% rise in global wheat production (*4*), although this benefit is likely to diminish, because rising temperatures will increase photorespiration and nighttime respiration. A benefit of rising temperatures is the alleviation of low-temperature inhibition of growth, which is a widespread limitation at higher latitudes and altitudes. Offsetting these benefits, however, are obvious deleterious changes, such as an increased frequency of damaging high-temperature events, new pest and disease pressures, and altered patterns of drought. Negative effects of other pollutants, notably ozone, will also reduce benefits to plant growth from rising CO_2 and temperature.

Particularly challenging for society will be changes in weather patterns that will require alterations in farming practices and infrastructure; for example, water storage and transport networks. Because one-third of the world's food is produced on irrigated land (*5*, *6*), the likely impacts on global food production are many. Along with agronomic- and management-based approaches to improving food production, improvements in a crop's ability to maintain yields with lower water supply and quality will be critical. Put simply, we need to increase the tolerance of crops to drought and salinity.

In the context of global environmental change, the efficiency of nitrogen use has also emerged as a key target. Human activity has already more than doubled the amount of atmospheric N_2 fixed

Australian Centre for Plant Functional Genomics, University of Adelaide, South Australia SA 5064, Australia.

*To whom correspondence should be addressed. E-mail: mark.tester@acpfg.com.au

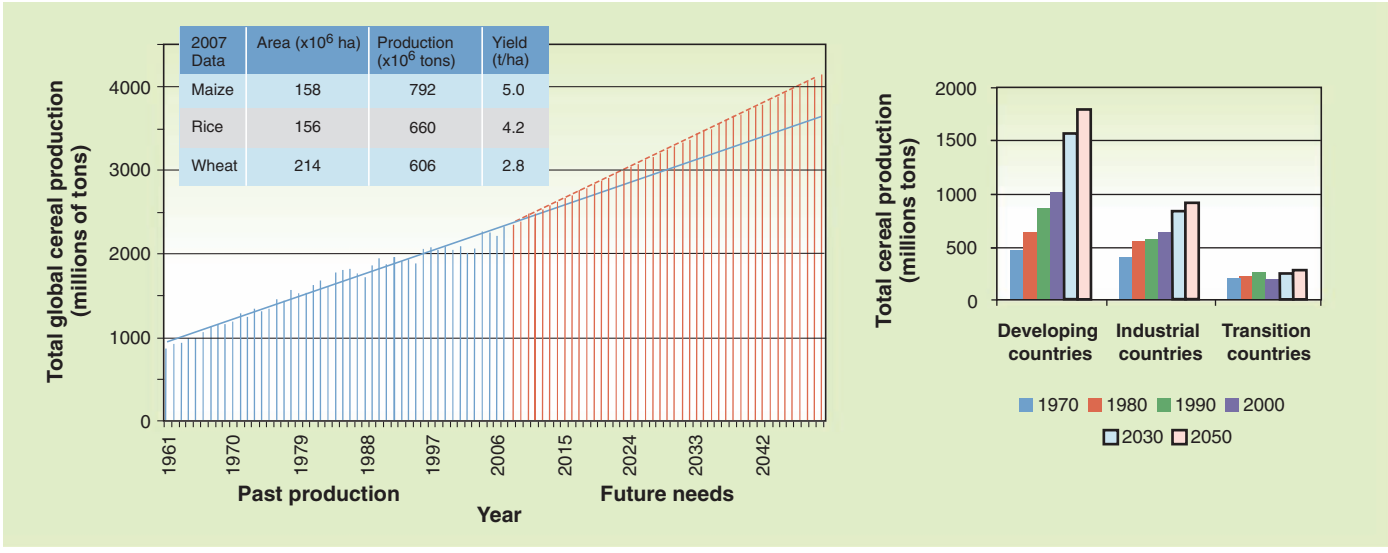

Fig. 1. Cereal production targets. (**Left**) Global cereal production has risen from 877 million metric tons in 1961 to 2351 million metric tons in 2007 (blue). However, to meet predicted demands (*3*), production will need to rise to over 4000 million metric tons by 2050 (red). The rate of yield increase must move from the blue trend line (32 million metric tons per year) to the red dotted line (44 million metric tons per year) to meet this demand, an increase of 37%. The inset table shows the 2007 data for the three major cereals. Data are from the FAO: http://faostat.fao.org/. (**Right**) The greatest demand for yield increases will be from countries in the developing world. [Based on FAO data (*26*)].

annually, which has led to environmental impacts, such as increased water pollution, and the emission of greenhouse gases, such as nitrous oxide. Nitrogen inputs are increasingly being managed by legislation that limits fertilizer use in agriculture. Furthermore, rising energy costs means that fertilizers are now commonly the highest input cost for farmers. New crop varieties will need to be more efficient in their use of reduced nitrogen than current varieties are (*7*). Therefore, it is important that breeding programs develop strategies to select for yield and quality with lower nitrogen inputs.

Current Approaches to Crop Improvement

Arguably, increased yield in conditions of abiotic stresses, such as drought and salinity, could be best achieved by selecting for increased yield under optimal production conditions: Plants with higher yields in good conditions are more likely to have higher yields in stressed conditions (*8*). Such an approach will also increase yield in high-yield environments. However, it is becoming increasingly apparent that specific selection strategies are needed to enhance yield in low-yield (stressed) environments. Given that average global yields of wheat are less than 3 metric tons/ha (Fig. 1) and given there are many areas with yields as high as 10 metric tons/ha, the majority of land cropped to wheat delivers yields below 3 metric tons/ha. Therefore, by virtue of the much larger areas of low-yielding land globally, low-yielding environments offer the greatest opportunity for substantial increases in global food production. Increasing yield by 1 metric ton/ha in a low-yielding area delivers a much higher relative increase than does the same increase in high-yielding environments. This increase can be achieved by tackling major limitations on yield in poor environments (termed yield stability); for example, by protecting plants and yield from factors such as salinity and heat or drought periods. The local social benefits of supporting farmers on low-yielding lands would also be great.

It is often thought that concentration on yield stability may come at the expense of high yields in good years; however, yield penalties in more favorable conditions do not necessarily accompany drought tolerance (Fig. 2). Yield stability is harder to select for than improved yield is, because selection in breeding programs requires many years and many sites for evaluation. However, there is evidence for a genetic basis for yield stability and, hence, an opportunity for gain (*9*). Transgenic approaches are also likely to improve yield stability (*10*). There are several clear examples where single genes have been able to substantially increase yield, notably to drive domestication (to control tiller number, branching, and seed number) and the green revolution (for dwarfing). Initial results suggest that a gene conferring increased drought tolerance may also have a widespread impact on yield (*10*).

This is not to say that efforts to maintain yield should be reduced. In particular, maintaining resistance to rapidly evolving pests and pathogens is an essential mainstay of breeding programs. Interactions between breeders, pathologists, and agronomists must be maintained to ensure that crops and cropping systems change coordinately. No-till farming, in which plowing of the soil is avoided, for example, has changed the spectrum of diseases and pests attacking crops, to the extent that a change in breeding targets was needed. The development of multiple cropping systems will also demand interactions between agronomists

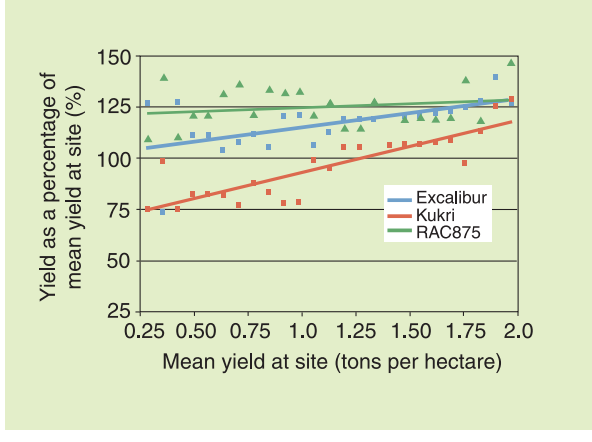

Fig. 2. Yield under severe drought stress. Shown are differences in maintenance of yield with lower water supply for three lines of Australian bread wheat. Low-yielding environments are water-limited fields in southern Australia. The yield for each of the three lines is plotted relative to the average yield for that site of at least 50 independent genotypes. The lines were evaluated in 25 environments (multiple sites for several years).

Box 1. New breeding technologies.

MAS uses a marker such as a specific phenotype, chromosomal banding, a particular DNA or RNA motif, or a chemical tag that associates with the desired trait. For example, a DNA marker closely linked to a disease resistance locus can be used to predict whether a plant is likely to be resistant to that disease.

• Gene pyramiding can usually only be accomplished by using MAS. For example, pyramiding is used to create durable disease resistances by selecting for two or more resistance genes against a pathogen. Multiple, partial, rust-resistance genes in wheat can be accumulated into elite varieties to provide strong and durable resistance. Single genes would give only weak resistance, and MAS offers the only effective method for accumulating multiple resistances (*22*).

• Marker-assisted recurrent selection (MARS) involves crossing in selected individuals at each cycle of crossing and selection. In this way, desirable alleles can be brought into the breeding scheme from many different sources. This technique has been applied to sunflower, soybean, and maize to bring desirable alleles at several target loci into single elite lines (*27*).

• Genome-wide or genomic selection also relies on MAS and is under evaluation for the feasibility of incorporating desirable alleles at many loci that have small genetic effect when used individually. In this approach, breeding values can be predicted for individual lines in a test population based on phenotyping and whole-genome marker screens. These values can then be applied to progeny in a breeding population based on marker data only, without the need for phenotypic evaluation. Modeling studies indicate that this method can lead to considerable increases in the rates of genetic gain by accelerating the breeding cycles (*20*). In the oil palm, for example, this approach could lead to the release of improved germplasm after only 6 years as compared with the current time of 19 years (*28*).

• Complex trait dissection uses high-throughput technologies to determine the phenotypic components of complex traits. For example, robotic greenhouse systems use nondestructive imaging to monitor growth rates, stem and leaf architecture, and root structure (for example, see www.lemnatec.com/). Similar systems can also be adapted for the detection of characteristics of chlorophyll fluorescence (which indicate aspects of plant responses to the environment) or fluorescent protein–labeled genotypes.

• The analysis of complex traits has recently been bolstered by developments in statistical and modeling methods for the analysis of phenotypic data obtained from field and controlled environment studies. For example, in assessing drought tolerance in wheat and sorghum, modeling can be used generate an "index of the climatic environment" to identify the stages of crop development where there is the strongest interaction between genotype and the environment and to identify aspects of the crop response that can be most readily enhanced by breeding and selection (*29*).

• Increasing genetic diversity requires an expansion of the germplasm base in breeding programs (*22*), but this is dependent on enhancing techniques for assessing the value of the program and using individual accessions from germplasm collections. Improvements in phenotyping and genotyping will help remove this limitation by facilitating the identification and characterization of key adaptive QTLs. For example, increased expression of a boron transporter in a barley landrace leads to high tolerance to soil boron in elite varieties when the high-expression allele is transferred. Screening for variation in expression levels for this gene in germplasm collections may identify new sources of tolerance (*30*).

• Introgression of novel alleles from landraces and wild relatives is often slow and tedious, but options are now being developed for accelerating introgression as we learn more about the recombinational behavior of plant genomes and develop new breeding methods.

• The wider deployment of GM approaches will be needed for the introduction of novel genes and alleles from diverse sources, and particularly for traits that are absent from plant genomes (for example, *Bacillus thuringiensis* toxin from soil bacteria) or where there is insufficient variation for practical utility (for example, vitamin A accumulation in rice endosperm).

• The constraints on regulatory and consumer acceptance of GM can be reduced by adopting alternative approaches for engineering plants. For example, consumer acceptance may be greater and regulatory approvals simpler for plants transformed with cis-genic vectors in which only host gene sequences are used in the transformation construct (www.cisgenics.com/). Similarly, the creation of marker-free plants, where only the DNA that has a biological effect remains in the plant, has been used to develop plants without antibiotic-resistance genes, which has caused much controversy (*31*).

• Heterosis (hybrid vigor) for inbreeding species (that is, species that usually self-pollinate, such as rice and wheat) can offer 20% to over 50% yield increases, and, for example, a 68% increase in yield has been achieved in foxtail millet (*32*). Strategies for using heterosis more widely to increase yields in inbreeding crops center on finding ways of reducing the cost and increasing the efficiency of producing hybrid seed. These include identifying new sources of male sterility for hybrid creation [such as thermosensitive genic male sterility in rice (*33*)] and using GM approaches to engineer sterility and restore fertility (such as the InVigor Canola from Bayer CropScience)]. Another possible mechanism for producing hybrid seed involves the use of apomixis, where plants produce seed without the need for fertilization. This allows hybrid vigor to be fixed. Creating apomictic crop plants may also be possible as we learn more about the genes controlling this process.

• Direct targeting of key heterotic loci may also be achievable as we learn more about the molecular basis of hybrid vigor (for example, in maize) (*34*).

Limitations

Of course, none of this will happen without suitably trained staff in plant breeding and molecular biology, so substantial increases in the education of plant breeders are essential. Most countries are struggling to maintain strong breeding capabilities. A vital adjunct is the free communication of resources and capabilities from technology developers to technology users. Resource and capacity building within breeding programs is essential to develop novel approaches, particularly in developing countries. Furthermore, developing countries critically need support for the development of crops, where there has been little interest from the developed world and, consequently, little investment. In many cases, these "orphan crops," such as cassava and plantain, are of critical importance for food security.

For many of the new breeding technologies, access to equipment, reagents, and skilled personnel is critical. Whereas service providers deliver this support to breeding programs in some parts of the world, they are often too expensive for poorly resourced breeding programs, and the logistics of sending plant tissue samples for analysis in a timely fashion can be prohibitive. Some organizations are attempting to address this limitation by establishing support services for breeding programs in the developing world (www.generationcp.org/).

and breeders. However, it is clear that more is required than can be provided by traditional breeding approaches.

Emerging Technologies for Crop Breeding

The production and evaluation of genetically modified (GM) crops is an active area of research, but the access of growers to this technology in many countries is currently restricted primarily because of political and bioethical issues (Box 1). Nevertheless, GM technologies permit the generation of novel variation beyond that which is available in naturally occurring (or even deliberately mutated) populations. Classic applications of GM include the use of proteinaceous toxins to control insect pests and "golden rice," which is biofortified with vitamin A (*11*). Crucial to the future deployment of GM crops are the discovery and characterization not only of genes but of promoters that provide accurate and stable spatial and temporal control of the expression of the genes (*12*). Development of cis-genic vectors and marker-free transgenic plants (Box 1) may help to ease some of the political concerns about GM technologies. Nevertheless, the widespread application of GM technologies will remain limited while regulatory demands impose high costs on releasing GM crops (Box 1). Although it is likely that most of the important contributions to crop improvement in the coming 5 to 10 years will continue to be from non-GM approaches, we consider that transgenic technologies will inevitably be deployed for most major crops in the future.

Methods of crop breeding have undergone major changes, and a range of technologies is improving the rate and success of crop improvement in some breeding programs, but these have yet to be widely adopted. Contributions are being made through new selection strategies that are informed by sophisticated genetics, the use of computers to track and manage field trials, and biometric methods for field-trial design and assessment of interactions between genotype, environment, and management (*13*).

Marker-assisted selection (MAS) techniques (Box 1) are free of the political issues that have plagued the application of GM technologies. MAS involves using variation at the DNA level to track and monitor specific regions of the genomes during crossing and selection (*14*). The greatest benefit of MAS occurs where the target traits are of low heritability, are recessive in nature, and involve difficult and costly phenotyping, and where pyramiding of genes is desired for results such as disease and pest resistance. In these cases, MAS is likely to be more reliable, more convenient, or cheaper than phenotype-based selection, and MAS currently provides the only viable method for gene pyramiding. Molecular markers are also important in analyzing the mode of inheritance of certain traits and assessing genetic diversity. In cases where desirable traits are closely linked and in repulsion, markers can be critical in selecting rare recombination events.

In many cases, MAS provides an important alternative to phenotypic selection. However, the success of markers depends on their reliability in predicting phenotype. Many key stresses associated with rapid environment changes, notably drought and salinity tolerance, are complex and highly variable. For these types of traits, it is necessary to dissect tolerance into component contributory traits and to identify genetic regions encoding the traits, rather than overall plant tolerance (*6*, *15*, *16*). However, this genetic approach requires high-throughput phenotyping (phenomics) (*17*) (Box 1). Phenomics also allows screening of populations for particular traits and will facilitate the introgression of novel variation from wild germplasm. Phenomics will enable tighter definition of the properties of molecular markers, allowing introgression of appropriate combinations of tolerance traits into commercial varieties for particular target environments.

The combination of reliable phenotyping and MAS has been particularly important in transferring desirable alleles by simple backcrossing into elite germplasm. Although MAS has been used to track multiple independent loci (*18*), conventional breeding schemes become quite complex as the number of target loci expands. To overcome the problems of dealing with multiple loci, in particular, multiple loci of small genetic effect, two relatively new methods involving MAS can be deployed: marker-assisted recurrent selection (MARS) and genome-wide or genomic selection (GWS) (*19*, *20*) (Box 1). MARS involves crossing selected individuals at each selection cycle so that desirable alleles at the target loci are introduced one at a time or through the merging of multiple crossing and selection streams. A problem with this approach is that it is most effective for genes or quantitative trait loci (QTLs) of major effect. In contrast, GWS does not require prior information on marker trait associations and can be used to select for multiple loci of small genetic effect. In this approach, populations are extensively genotyped to give full genome coverage and phenotyped. Subsequently, these data allow the prediction of phenotypic performance of an individual on the basis of whole-genome marker surveys.

These new breeding and selection strategies rely on the availability of cheap and reliable marker systems. A serious limitation in marker application for some species has been the paucity of useful markers. However, the new sequencing platforms have allowed large-scale discovery of single-nucleotide polymorphisms (SNPs) for species where few markers were previously available. The new marker systems combined with the new marker-based selection and screening strategies provide a base for a revolution in crop breeding and genetics.

Expanding the Germplasm Base for Plant Breeding

The success of plant breeding over the past century has been associated with a narrowing of the available genetic diversity within elite germplasm, particularly for some species such as peanut and soybean. New sources of variation include landraces and wild relatives of crop species, and although exploiting wild relatives as a source of novel alleles is challenging, it has provided notable successes in crop improvement. A particularly important example of the introgression of genetic information from a relative was the use of the short arm of rye chromosome 1R in wheat. In the early 1990s, this wheat-rye translocation was used in 45% of 505 bread wheat cultivars in 17 countries (*21*). Increasingly easy gene discovery, improved enabling technologies for genetics and breeding, and a better understanding of the factors limiting practical exploitation of exotic germplasm promise to transform existing, and to accelerate the development of new, strategies for efficient and directed germplasm use (Box 1).

Most crop geneticists agree that enrichment of the cultivated gene pool will be necessary to meet the challenges that lie ahead. However, to fully capitalize on the extensive reservoir of favorable alleles within wild germplasm, many advances are still needed. These include increasing our understanding of the molecular basis for key traits, expanding the phenotyping and genotyping of germplasm collections, improving our molecular understanding of recombination in order to enhance rates of introgression of alien chromosome regions, and developing new breeding strategies that permit introgression of multiple traits (*22*). Recent progress has shown that each of these challenges is tractable and within reach if some of the basic problems limiting the application of new technologies can be tackled.

Limitations in Applying the New Technologies

Several issues are likely to limit the application of these new methods, particularly for breeding programs in the public sector (Box 1). Regulatory complexity and high costs have prevented the widespread delivery of GM technologies (Box 1). Over the coming decade or so, however, it seems inevitable that GM technologies will become much more widely used—it is probably a case of "when," not "if." A consequence emerging for crops that are now dominated by GM varieties (such as cotton, soybean, and maize) is that breeding programs are now based around GM varieties, and consequently, breeding programs in non-GM jurisdictions have limited access to current advances. The key limitations for traditional breeding include lack of resources, training, and capabilities for most of the world's

crop improvement programs (*23, 24*) (Box 1). It is important, therefore, that we expand the scope of and access to new marker platforms to provide efficient, cost-effective screening services to the breeders. Communication and mechanisms for delivery of material to breeders must be developed. There is an urgent need to expand the capacity of breeding programs to adopt new strategies. The clearly documented high rate of return on such investments in the past should be kept in mind (*25*).

The concerns about food security and the likely impact of environmental change on food production have injected a new urgency into accelerating the rates of genetic gain in breeding programs. Further technological developments are essential, and a major challenge will be to also ensure that the technological advances already achieved are effectively deployed.

References and Notes

1. Organisation for Economic Cooperation and Development (OECD), *OECD-FAO Agricultural Outlook 2009–2018* (OECD, Paris, 2009).
2. J. M. Alston, J. M. Beddow, P. G. Pardey, *Science* **325**, 1209 (2009).
3. Food and Agriculture Organisation (FAO) of the United Nations, Declaration of the World Summit on Food Security, Rome, 16–18 November 2009 (www.fao.org/wsfs/world-summit/en/).
4. R. A. Fischer, G. O. Edmeades, *Crop Sci.* **50**, in press (2010).
5. FAO, FAO Land and Plant Nutrition Management Service (2008) (www.fao.org/nr/land/en/).
6. R. Munns, M. Tester, *Annu. Rev. Plant Biol.* **59**, 651 (2008).
7. M. B. Peoples, A. R. Mosier, J. R. Freney, in *Nitrogen Fertilization in the Environment*, P. E. Bacon, Ed. (Marcel Dekker, New York, 1995), pp. 505–602.
8. R. A. Richards, *Plant Soil* **146**, 89 (1992).
9. A. T. W. Kraakman, R. E. Niks, P. M. Van den Berg, P. Stam, F. A. Van Eeuwijk, *Genetics* **168**, 435 (2004).
10. D. E. Nelson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 16450 (2007).
11. J. E. Mayer, W. H. Pfeiffer, P. Beyer, *Curr. Opin. Plant Biol.* **11**, 166 (2008).
12. I. S. Møller *et al.*, *Plant Cell* **21**, 2163 (2009).
13. P. S. Baenzinger *et al.*, *Crop Sci.* **46**, 2230 (2006).
14. S. P. Moose, R. H. Mumm, *Plant Physiol.* **147**, 969 (2008).
15. G. H. Salekdeh, M. Reynolds, J. Bennett, J. Boyer, *Trends Plant Sci.* **14**, 488 (2009).
16. M. Reynolds, Y. Manes, A. Izanloo, P. Langridge, *Ann. Appl. Biol.* **155**, 309 (2009).
17. E. Finkel, *Science* **325**, 380 (2009).
18. B. C. Y. Collard, D. J. Mackill, *Philos. Trans. R. Soc. London Ser. B* **363**, 557 (2008).
19. R. Bernardo, A. Charcosset, *Crop Sci.* **46**, 614 (2006).
20. E. L. Heffner, M. E. Sorrells, J. L. Jannink, *Crop Sci.* **49**, 1 (2009).
21. S. V. Rabinovich, in *Wheat: Prospects for Global Improvement*, H. J. Braun *et al.*, Eds. (Kluwer Academic, Dordrecht, The Netherlands, 1998), pp. 401-418.
22. C. Feuillet, P. Langridge, R. Waugh, *Trends Genet.* **24**, 24 (2008).
23. The Global Partnership Initiative for Plant Breeding can be found at http://km.fao.org/gipb/.
24. H. C. J. Godfray *et al.*, *Science* **327**, 812 (2010).
25. J. M. Alston *et al.*, *A Meta-Analysis of Rates of Return to Agricultural R&D: Ex Pede Herculem?* (International Food Policy Research Institute, Washington, DC, 2000).
26. FAO, *World Agriculture: Toward 2030/2050*. Interim Report, Global Perspective Studies Unit (FAO, Rome, 2006).
27. S. R. Eathington, T. M. Crosbie, M. D. Edwards, R. S. Reiter, J. K. Bull, *Crop Sci.* **47** (suppl. 3), S154 (2007).
28. C. K. Wong, R. Bernardo, *Theor. Appl. Genet.* **116**, 815 (2008).
29. S. C. Chapman, *Euphytica* **161**, 195 (2008).
30. T. Sutton *et al.*, *Science* **318**, 1446 (2007).
31. B. Darbani, A. Eimanifar, C. N. Stewart Jr., W. N. Camargo, *Biotechnol. J.* **2**, 83 (2007).
32. M. M. Siles *et al.*, *Crop Sci.* **44**, 1960 (2004).
33. X. Wu, *Agron. J.* **101**, 688 (2009).
34. M. D. McMullen *et al.*, *Science* **325**, 737 (2009).

35. We thank C. Morris for her helpful comments on the manuscript. Support for our research programs from the Australian Research Council, Grains Research Development Corporation, South Australian State Government, and the University of Adelaide is gratefully acknowledged. P.L. is the chief executive officer of the Australian Centre for Plant Functional Genomics.

10.1126/science.1183700

PERSPECTIVE

Smart Investments in Sustainable Food Production: Revisiting Mixed Crop-Livestock Systems

M. Herrero,[1]* P. K. Thornton,[1] A. M. Notenbaert,[1] S. Wood,[2] S. Msangi,[2] H. A. Freeman,[3] D. Bossio,[4] J. Dixon,[5] M. Peters,[6] J. van de Steeg,[1] J. Lynam,[7] P. Parthasarathy Rao,[8] S. Macmillan,[1] B. Gerard,[9] J. McDermott,[1] C. Seré,[1] M. Rosegrant[2]

Farmers in mixed crop-livestock systems produce about half of the world's food. In small holdings around the world, livestock are reared mostly on grass, browse, and nonfood biomass from maize, millet, rice, and sorghum crops and in their turn supply manure and traction for future crops. Animals act as insurance against hard times, and supply farmers with a source of regular income from sales of milk, eggs, and other products. Thus, faced with population growth and climate change, small-holder farmers should be the first target for policies to intensify production by carefully managed inputs of fertilizer, water, and feed to minimize waste and environmental impact, supported by improved access to markets, new varieties, and technologies.

"Business as usual" investments in agriculture, although necessary (*1*, *2*), are unlikely to deliver sustainable solutions as the world rapidly changes (*3*, *4*). At the recent G8 summit in Italy, the leaders of the world's wealthiest countries promised to invest U.S.$20 billion to improve global food security. Most of that money is likely to flow to the developing world, where over the next few decades agricultural systems, already facing a variety of stresses, will be expected to accommodate a massive population surge. Even an investment of this magnitude could fail to generate food security if its deployment is not well planned and based on sound science.

The usual culprits, such as inefficient aid delivery, government corruption, and political unrest, are a barrier to progress but are not the most important problem. Rather, it involves a fundamental failure to appreciate the range of different agricultural systems that are expected to feed our planet in the coming decades and their policy needs. The diverse pressures that are acting on agricultural systems in various parts of the world include population increase, rising incomes and urbanization, a rapidly rising demand for animal products in many developing countries, and a fierce competition for land and water (*3*, *5*, *6*), all of which will have profound effects on food security (*1*). Croppers and livestock keepers the world over have steadily accumulated local experience and knowledge that will help them to adapt in the future, but the rapid rates of change seen in many agricultural systems in developing countries may simply outstrip their capacity. Yet, recent scientific assessments (*1*, *2*, *7–10*) and the technical and policy recommendations that flow from them have not fully captured the complex biological, social, and economic dynamics of the variety of chal-

[1]International Livestock Research Institute (ILRI), Post Office Box 30709, Nairobi, Kenya. [2]International Food Policy Research Institute (IFPRI), 2033 K Street NW, Washington, DC 20006, USA. [3]International Finance Corporation, The World Bank Group, Washington, DC 20433, USA. [4]International Water Management Institute (IWMI), Colombo, Sri Lanka. [5]Australian Centre for International Agricultural Research, Canberra, ACT, Australia. [6]Centro Internacional de Agricultural Tropical (CIAT), Cali, Colombia. [7]Independent consultant, Nairobi, Kenya. [8]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India. [9]CGIAR Systemwide Livestock Programme, Addis Ababa, Ethiopia.

*To whom correspondence should be addressed. E-mail: m.herrero@cgiar.org