ELSEVIER

Review

TRENDS in Plant Science Vol.10 No.12 December 2005

Full text provided by www.sciencedirect.com

SCIENCE DIRECT

Genomics-assisted breeding for crop improvement

Rajeev K. Varshney[1,2], Andreas Graner[1] and Mark E. Sorrells[3]

[1]Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, D-06466 Gatersleben, Germany
[2]Present address: International Crops Research Institute for Semi Arid Tropics (ICRISAT), Patancheru – 502 324, Andhra Pradesh, India
[3]Department of Plant Breeding, Cornell University, Ithaca, NY 14853, USA

Genomics research is generating new tools, such as functional molecular markers and informatics, as well as new knowledge about statistics and inheritance phenomena that could increase the efficiency and precision of crop improvement. In particular, the elucidation of the fundamental mechanisms of heterosis and epigenetics, and their manipulation, has great potential. Eventually, knowledge of the relative values of alleles at all loci segregating in a population could allow the breeder to design a genotype *in silico* and to practice whole genome selection. High costs currently limit the implementation of genomics-assisted crop improvement, particularly for inbreeding and/or minor crops. Nevertheless, marker-assisted breeding and selection will gradually evolve into 'genomics-assisted breeding' for crop improvement.

Potential of genomics research

In recent years, an impressive number of advances in genetics and genomics have greatly enhanced our understanding of structural and functional aspects of plant genomes and have integrated basic knowledge in ways that can enhance our ability to improve crop plants to our benefit (Box 1). The complete genome sequences of *Arabidopsis* and rice, as well as an enormous number of plant expressed sequence tags (ESTs) (see Glossary), have become available. Further sequencing projects to enhance our knowledge of major crops are under way, and combining the new knowledge from genomic research with traditional breeding methods is essential for enhancing crop improvement. Superior varieties can result from the discovery of novel genetic variation, improved selection techniques or the identification of genotypes with new or improved attributes caused by superior combinations of alleles at multiple loci. Advances in genomics can contribute to crop improvement in two general ways. First, a better understanding of the biological mechanisms can lead to new or improved screening methods for selecting superior genotypes more efficiently. Second, new knowledge can improve the decision-making process for more efficient breeding strategies. Here, we present an overview of the status and availability of genomic resources and genomics research in crop plant species, and discuss strategies and approaches for effectively exploiting genomics research for crop improvement (Box 1, Figure 1).

Strategies for the future

Functional molecular markers

During the past few years, functionally characterized genes, EST and genome sequencing projects have facilitated the development of molecular markers from the transcribed regions of the genome. Among the more important and popular molecular markers that can be developed from ESTs are single-nucleotide polymorphisms (SNPs) [1], simple sequence repeats (SSRs) [2] or conserved orthologous sets of markers (COSs) [3] (Table 1). Putative functions can be deduced for the markers derived from ESTs or genes using homology searches (BLASTX) with protein databases (e.g. NR-PEP and SWISSPROT). Therefore, molecular markers

Glossary

Association mapping: also known as linkage disequilibrium (LD) mapping or association analysis is a population-based survey used to identify trait–marker relationships based on linkage disequilibrium.

Biparental populations: the progeny derived after crossing two genotypes as male and female parents. Such populations include F_2 genotypes generated from F_1 progeny, lines generated after doubling the haploids (DHs, obtained from F_1 plants through anther, egg cell or ovule culture or distant hybridization), or recombinant inbred lines (RILs), which are derived by single seed descent for at least five or more generations by repeated selfing or sibling mating.

Candidate gene: a gene that has been identified as related to a particular trait (phenotype, disease or condition). Candidate genes in general can be divided into two categories: positional and functional. A positional candidate gene is one that might be associated with a trait, based on the location of a gene on a chromosome. A functional candidate gene is one whose function has something in common biologically with the trait under investigation. Positional candidate genes are identified through QTL- and map-based cloning approaches, whereas functional genomics approaches such as transcriptomics and expression genetics provide the set of functional candidate genes.

COS: conserved orthologous set of markers that are used for comparative mapping between closely related species. For a given group of species, a COS is formed by identifying genes from each species that are orthologous to genes of other species in the set.

Epistasis: a form of gene interaction whereby one gene interferes with the phenotypic expression of another nonallelic gene or genes. Gene A is said to be epistatic to gene B if an allele of gene A masks the encoded effects of gene B. In case of epistasis, the combined phenotypic effect of two or more genes is either less than (negative epistasis) or greater than (positive epistasis) the sum of the effects of individual genes.

Expressed sequence tags (ESTs): partial sequences obtained from 5′ or 3′ end of cDNAs.

Corresponding authors: Varshney, R.K. (r.k.varshney@cgiar.org), Varshney, R.K. (varshney.raj@gmail.com).

www.sciencedirect.com 1360-1385/$ - see front matter © 2006 Elsevier Ltd. All rights reserved. doi:10.1016/j.tplants.2005.10.004

generated from (gene) sequence data are known as ‘functional markers’ (FMs) [4].

FMs have been developed extensively for plant species in which ESTs or gene sequence data are available [5]. By screening the unigene consensus sequences (based on ESTs) from over 50 plant species, Stephen Rudd and colleagues [3] demonstrated the feasibility of predicting molecular markers (e.g. SSRs, SNPs and COSs) that can be used to develop FMs in large numbers for several species (PlantMarkers, http://markers.btk.fi/). As a community effort, the compilation of developed EST-derived SSR markers for Triticeae (cereal) species is in progress at the Triticeae EST–SSR Coordination’s website (http://wheat.pw.usda.gov/ITMI/EST-SSR/) for making them publicly accessible.

FMs have some advantages over RMs (random markers that are generated from an anonymous region of the genome) because they are completely linked to the desired trait allele. Such markers can be derived from the gene responsible for the trait of interest and target the functional polymorphism in the gene, thus allowing selection in different genetic backgrounds without revalidating the marker–quantitative-trait-locus (QTL) allele relationship. Thus, they have also been referred as ‘perfect markers’, even though different alleles with the same polymorphism (resulting from intragenic recombination, insertion, deletion or mutation) might produce different phenotypes. A perfect marker allows breeders to track specific alleles within pedigrees and populations, and to minimize linkage drag flanking the gene of interest. As markers become more abundant, breeders develop strategies for use that are compatible with financial resources and breeding goals. Increasingly, markers are being applied to the selection of parental materials and for the accelerated selection of loci controlling traits that are difficult to select phenotypically. Examples include the pyramiding of genes for disease resistance or quality and those that interact with the environment or are costly to evaluate. Linked deleterious alleles are a potential problem as the number of loci selected increase, particularly if the donor parent is a related wild species. Frédéric Hospital [6] examined the efficiency of marker-assisted selection for reducing the size of the flanking donor segment. He showed that the efficiency of selection for the reduction of linkage drag in backcross programs depends on the population sizes, the number of backcross generations and the distances between the flanking markers and the introgressed gene. Closely linked markers are most desirable for reducing linkage drag but this requires larger population sizes and more backcross generations.

Transcriptomics and functional genomics

The salient challenge of applied genetics and functional genomics is identification of the genes underlying a trait of interest so that they can be exploited in crop improvement programmes (see the Review by Willem Albert Rensink and Robin Buell in this issue of *Trends in Plant Science*). Macro- and microarrays have been successfully used in many plant species to understand the basic physiology, developmental processes and environmental stress

Gene space: long gene-rich regions that contain the vast majority of genes, separated by long gene-poor regions in a genome of given species. Occurrence of ‘gene space’ is a common feature of plant species, which have a large genome size owing to the abundance of repetitive DNA (transposons and retrotransposons) in their genome.

Linkage disequilibrium (LD): non-random association between two markers, genes or QTLs on the same chromosome in a population owing to their tendency to be co-inherited. When variants of two genetic loci are in LD, the variant seen at one locus predicts the variant found at the other.

Marker-assisted selection: a method that uses molecular markers for indirect selection of difficult traits at the seedling stage, speeding up the process of conventional plant breeding and facilitating the improvement of traits that cannot be easily selected using conventional methods.

Map-based cloning: involves the identification of a mutant phenotype for the trait of interest (obtained by mutagenesis or from natural variation) and genetic fine mapping using many progeny plants. This map is then used for chromosome walking or landing, with the help of large-insert DNA libraries or physical maps to isolate the gene.

Metabolomics: an extended discipline of biochemistry that involves the analysis (usually high throughput or broad scale) of small-molecule metabolites and polymers such as starch. It also involves descriptions of biological pathways and current metabolomic databases such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG).

Micro- or macroarray: cDNAs or oligonucleotides (representing the whole or partial genome of an organism) immobilized on a glass slide (called microarray) or other substrate such as nylon membrane (called macroarray) that are probed with labelled cDNAs from treated and control tissue for gene expression analysis.

Molecular markers: a set of DNA-based genetic markers that can detect DNA polymorphism both at the level of specific loci and at the whole genome level. There are many types of molecular markers: restriction fragment length polymorphisms (RFLPs) were the first to be developed; others include random amplification of polymorphic DNAs (RAPDs), cleaved amplified polymorphic sequence (CAPS), simple sequence repeats (SSRs) and amplified fragment length polymorphisms (AFLPs); the most recently developed markers are single nucleotide polymorphisms (SNPs) and single feature polymorphisms (SFPs).

Near isogenic lines (NILs): a set of lines (generally for a given variety) that are genetically similar except for a gene, marker or trait and surrounding DNA that is associated owing to linkage drag.

Phenomics: high-throughput analysis of phenotypes that includes detailed and systematic analysis of phenotypes in terms of data repository and a means of structured interrogation.

Proteomics: expanded area of protein biochemistry that encompasses database of protein sequences, databases of predicted protein structures and, more recently, databases of protein expression analysis.

Pyramiding of genes: the process of bringing together several disease resistance or agronomically important genes from different sources into one genetic background (genotype).

Quantitative trait loci (QTLs): genomic regions that are associated with a phenotypic trait exhibiting continuous variation.

Single feature polymorphisms (SFPs): identified in transcript profiling data by visualizing differences in hybridization signals in different cultivars. The polymorphisms present in DNA are transcribed into the messenger RNA and can potentially affect hybridization to the microarrays or GeneChip probes if present in a region complementary to the probe. Polymorphisms generated during mRNA processing, such as alternative splicing and polyadenylation could also affect hybridization of the target RNA.

Single nucleotide polymorphisms (SNPs): an alteration of one nucleotide in a DNA sequence, SNPs can be detected and used as markers. Their frequent occurrence provides a large source of widely distributed genetic markers that are likely to be found close to target genes of interest.

Simple sequence repeats (SSRs): commonly called microsatellites, SSRs consist of simple, tandemly repeated di- to pentanucleotide sequence motifs. Because they are abundant, hypervariable, multiallelic and evenly distributed throughout the nuclear genomes of most organisms, they provide a valuable source of polymorphism and are thus an important class of genetic markers. The exceptionally high levels of polymorphism detected by SSRs are due to the variability in the number of tandem repeats at a particular locus.

Targeting induced local lesions in genomes (TILLING): a reverse genetic method that combines random chemical mutagenesis with PCR-based screening of gene regions of interest. This provides a range of allele types, including mis-sense and knockout mutations. By comparing the phenotypes of isogenic genotypes differing in single sequence motifs, TILLING provides direct proof of function of both induced and natural polymorphisms without the use of transgenic modifications.

Transcriptomics: the application of micro- or macroarrays and sequence-based methods to conduct expression profiling to determine the level of gene expression at a global (genome wide) level.

Unigenes: a non-redundant set of genes that is defined after clustering (computational) analysis of sequences generated through an EST or a genome sequencing project.

Box 1. Genomic resources, technologies and bioinformatics

Resources for major crop species include detailed, high-density genetic maps, cytogenetic stocks, contig-based physical maps and deep coverage, large-insert libraries [60,61]. These tools have facilitated the isolation of genes via map-based cloning, the localization of quantitative trait loci (QTLs) and the sequencing and annotation of large genomic DNA fragments in several plant species [62].

Complete genome sequences of *Arabidopsis* [63] and two rice cultivars (representing both the *indica* and *japonica* subspecies [64–66]) have become available. Sequence comparisons of the rice subspecies have revealed many insertions and deletions in both genomes [66–68]. Whole genome or gene space sequencing is being carried out for several plant species such as maize (http://www.maizegenome.org/), sorghum [69], wheat (http://www.wheatgenome.org/), tomato (http://sgn.cornell.edu/help/about/tomato_sequencing.html), tobacco (http://www.intl-pag.org/13/abstracts/PAG13_P027.html), poplar (http://genome.jgi-psf.org/Poptr1/), *Medicago* (http://www.medicago.org/genome/) and lotus (http://www.kazusa.or.jp/lotus/). Just as valuable is the resequencing of large regions of the genome.

The widespread use of transcriptome sampling strategies is a complementary approach to genome sequencing, and results in a large collection of expressed sequence tags (ESTs) for almost all the important plant species (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). The plant EST database has recently passed the five million sequence landmark. More than 50 plant species, each with >5000 ESTs, are represented [3]. Comparative sequence analysis can be used in some cases to facilitate isolation of genes in species lacking ESTs. However, EST resources have some limitations, such as unidentified contaminants, chimeric sequences, multiple forms in polyploids (homoeoalleles) and putatively non-functional transcripts. Moreover, they lack untranscribed regulatory factors and under-represented genes.

Comparative genomics among the cereals has revealed extensive colinearity among molecular marker maps based on restriction fragment length polymorphism (RFLP) (e.g. [70,71]). Brandon Gaut [72] reanalyzed previously published comparative RFLP-based mapping studies among the cereals and concluded that the genomes were evolving more rapidly than previously thought. Recently, sequence-based maps have revealed extensive breakdown of colinearity between wheat and rice ([73] and see volume 168 of *Genetics*), maize and rice [74], and sorghum and rice [75]. However, comparative sequencing between maize inbreds has revealed striking differences in both coding and repetitive sequences, and the structural heterogeneity resulting in non-shared genes among maize inbreds has led to speculation that the complementarity of haplotypes could contribute to the heterosis [76–78]. The repetitive sequence environment might affect gene expression, and complementation of repetitive sequences has also been proposed to be the cause of heterosis [77]. Stephan Brunner and colleagues [78] hypothesized that heterosis might result from chromatin restructuring caused by the unshared flanking repetitive sequences. These studies of maize inbreds and rice subspecies support the hypothesis that grass genomes are evolving rapidly. This is advantageous to plant breeders because novel genetic variation is fundamental to breeding progress. One of the hallmarks of genomics research has been the discovery of new mechanisms contributing to genome evolution.

Bioinformatics facilitates both the analysis of genomic and post-genomic data, and the integration of data from the related fields of transcriptomics, proteomics, metabolomics and phenomics. Several bioinformatic tools and databases (Table 1) have been developed for DNA sequence analysis, marker discovery and querying and analyzing information. For instance, the GenBank metadatabase is the repository of choice for public DNA sequence data worldwide and contains more than 7.4 million plant DNA sequences. Similarly, there are genome databases such as RGP (http://rgp.dna.affrc.go.jp/), The Institute for Genomic Research (http://www.tigr.org/), Gramene (http://www.gramene.org/) and GrainGenes (http://wheat.pw.usda.gov/) that incorporate analytical, visualization and interrogation tools. Enhanced bioinformatic tools, genome databases and integration of information from different fields enable the identification of genes and gene products, and can elucidate the functional relationships between genotype and observed phenotype [79]. Probably the most important future prospect is the enhancement of visualization tools that extend beyond simple relationships and help us more clearly to interpret the complex multidimensional biological networks of genes and their relationships to phenotypes.

responses, and to identify and genotype mutations [7,8]. However, use of these technologies for applied aspects in plant breeding has been limited because, except for near-isogenic lines (NILs), differential gene expression is caused not only by the trait of interest but also by the variation present in the genetic background. Therefore, background effects must be eliminated to establish a functional association between the level of gene expression and a given trait. Elena Potokina and colleagues [8] used ten barley genotypes that were characterized for six malting quality parameters and a cDNA array with 1400 unigenes to identify candidate genes for each of the six malting parameters. Such studies suggest that a functional association analysis strategy can provide a useful link between functional genomics and plant breeding. However, there are severe technical limitations to this approach including: (i) false positive signals from genes, caused by hitchhiking effects and low heritability of gene expression patterns and gene interactions; (ii) limited population sizes, which is partly because of cost; and (iii) limited correlations with QTL studies because of their comparatively low resolution. Furthermore, genes encoding transcription factors (TFs), the master-control proteins in all living cells, can control or influence many biological processes and many TFs are themselves regulated at the level of transcription [9]. TFs are generally produced at low levels in plants, frequently in a cell-type- or tissue-specific manner and often only transiently during development [10]. Therefore, it is more likely that the transcripts of many TF genes will be difficult to detect and quantify with DNA array technologies. However, the reverse-transcription polymerase chain reaction (RT-PCR) is estimated to be at least 100 times more sensitive than DNA arrays at detecting transcripts [11]. As a result, Tomasz Czechowski and colleagues [10] recently developed a real-time RT-PCR-based resource for quantitative measurement of TF encoding genes. Thus, knowledge about where and when TF encoding genes are transcribed and how such transcription is affected by internal and external cues will be valuable in elucidating the specific biological roles of the cognate proteins, particularly in response to environmental stresses.

Microarray-based gene expression data between two genetically different lines can also be used to identify single feature polymorphisms (SFPs) for SNP detection in a highly parallel manner [12], and can be exploited to develop FMs. In a recent study using Affymetrix GeneChip expression data, >10 000 SFPs have been identified between two genotypes of barley, a species with a large and complex genome [13]. However, identification of SFPs involves the problem of sensitivity versus

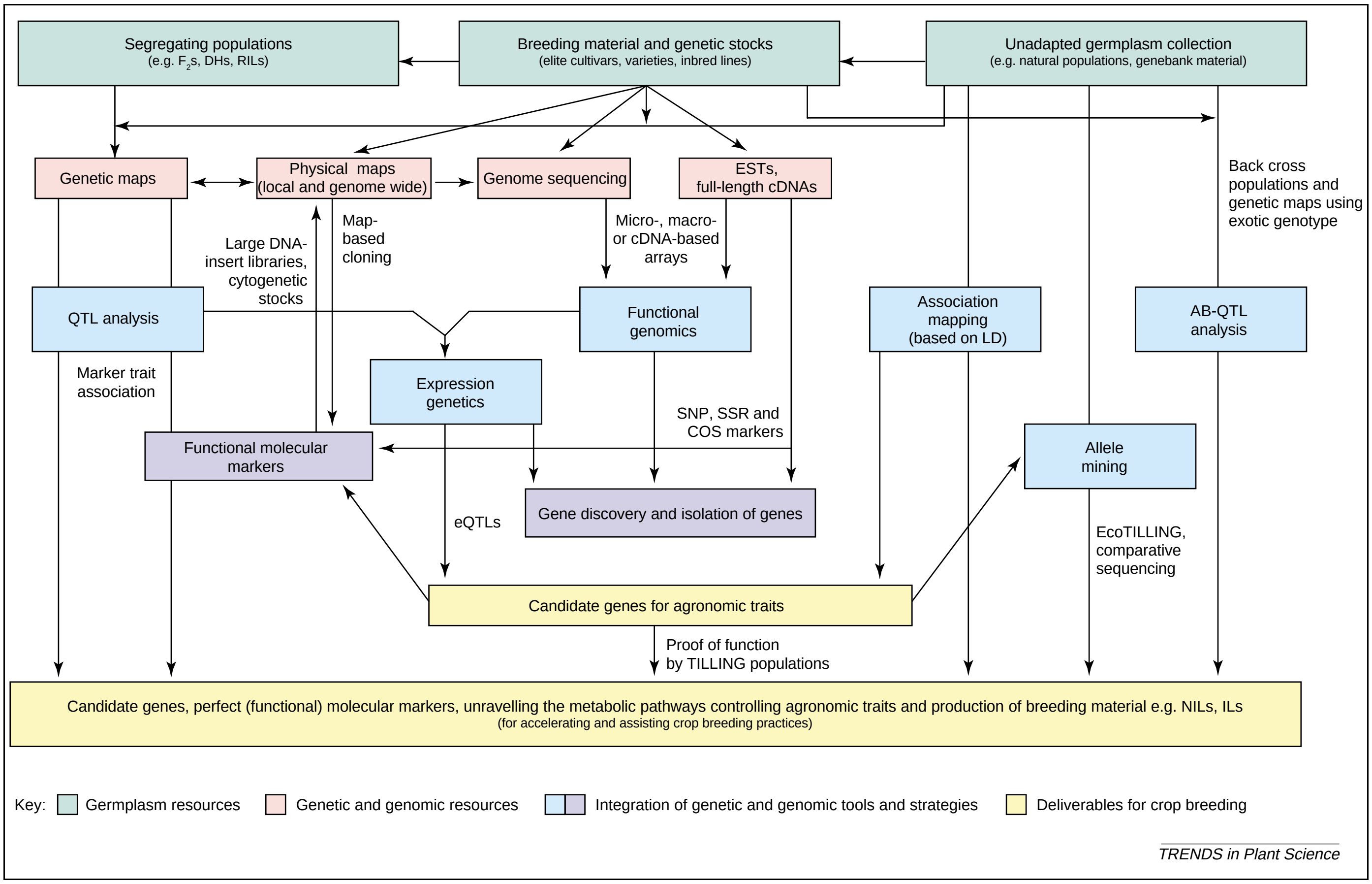

Figure 1. An integrated view of exploitation of genomic resources for crop improvement via different genetic and genomic strategies. Abbreviations: AB-QTL, advanced backcross QTL; COS, conserved orthologous set; DHs, doubled haploids; eQTLs, expression QTLs; ESTs, expressed sequence tags; ILs, introgression lines; LD, linkage disequilibrium; NILs, near isogenic lines; QTL, quantitative trait locus; RILs, recombinant inbred lines; SNP, single nucleotide polymorphism; SSR, simple sequence repeat or microsatellite; TILLING, targeted induced local lesions in genome.

selectivity (i.e. many putative SNPs could not be confirmed). Furthermore, the development of SNP markers in polyploid crop species such as wheat is complicated by the multiple genomes, resulting in the need to distinguish intergenome polymorphisms (between the A, B and D genomes) from intervarietal polymorphisms [14].

It is important to realize that some studies have shown that different microarray platforms (e.g. Affymetrix, Agilent, Amersham) with the same RNA sample or analysis of the same microarray gene expression data with different bioinformatic tools might not identify the same set of differently expressed genes for a given trait [15–17]. Such studies highlight the need for caution when analyzing and interpreting functional genomics studies for the purpose of extracting candidate gene lists.

Expression genetics and eQTLs

Ritsert Jansen and Jan-Peter Nap [18] proposed the use of gene expression data in QTL analysis. By analyzing the expression levels of genes or clusters of genes within a segregating population, it is possible to map the inheritance of that expression pattern. Expression QTLs (eQTLs) can be classified as *cis* or *trans* acting based on the location of the transcript compared with that of the eQTL influencing expression of that transcript [19]. Because of this feature, eQTL analysis makes it possible to identify factors influencing the level of mRNA expression. The regulatory factor (second order effect) is of specific interest because more than one QTL can be putatively connected to a *trans*-acting factor [20]. Thus, the mapping of eQTLs allows multifactorial dissection of the expression profile of a given mRNA, cDNA, protein or metabolite into its underlying genetic components, as well as localization of these components on the genetic map [18]. Subsequently, the eQTL analysis for each gene or gene product analyzed in the segregating population can identify the regions of the genome influencing its expression. Furthermore, for plant species in which the sequence of the whole genome is available, the annotation of those genomic regions will be helpful for the identification of the genes and the regulatory sequences involved in their expression.

The mapping of expression profiles has demonstrated its utility in understanding complex traits in humans [20, 21], fruit flies [22] and yeast [23]. After analyzing mRNA transcript abundances as quantitative traits for maize, Eric Schadt and colleagues [20] identified 18 805 genes that were differentially expressed (type I error = 0.05) in the ear leaf tissue from two different inbred lines. In a population of 76 F_2 individuals from the cross between these inbreds, expression patterns of 6481 genes were associated with at least one QTL (LOD $\geq$3.0). Most of the genes in their study had a single eQTL and 80%

Table 1. List of some important bioinformatics tools and databases for genomics research

Names	URL	Description and application
Tools		
MISA	http://pgrc.ipk-gatersleben.de/misa/	A Perl script-based module that allows the identification and localization of perfect microsatellites as well as compound microsatellites in sequences.
AutoSNP	http://www.cerealsdb.uk.net/discover.htm	A SNP detection program to identify putative polymorphisms between orthologous and parologous sequences from expressed sequence databases.
SNP2CAPS	http://pgrc.ipk-gatersleben.de/snp2caps/	For computational conversion of SNPs into CAPS markers.
MicroArray Software Catalogue	https://www.cs.tcd.ie/Nadia.Bolshakova/softwaretotal.html	Bioinformatic tools for microarray data analysis, datamining and data visualization software package.
TASSEL	http://www.maizegenetics.net/bioinformatics/tasselindex.htm	A software package to evaluate trait associations, evolutionary patterns and linkage disequilibrium.
Structure	http://pritch.bsd.uchicago.edu/software.html	A software package for using multi-locus genotype data to investigate population structure, such as inferring the presence of distinct populations, assigning individuals to populations, studying hybrid zones and identifying migrants and admixed individuals
Databases		
AceDB	http://www.acedb.org/	A genome database designed specifically for handling bioinformatic data flexibly; it includes tools designed to manipulate genomic data but is increasingly also used for non-biological data.
KEGG	http://www.genome.ad.jp/kegg/	Bioinformatic resources to enable computational prediction of higher level complexity of cellular processes and organism behaviours from genomic information.
NCBI	http://www.ncbi.nih.gov/	Public databases and software tools for storing, disseminating and analyzing genome data.
EMBL nucleotide sequence database	http://www.ebi.ac.uk/embl/	A public (European) nucleotide sequence database that allows user friendly downloading of sequence data.
SwissProt	http://us.expasy.org/sprot/	A curated protein sequence database that strives to provide a high level of annotation (e.g. the function of a protein and its domain structure, post-translational modifications and variants).
GRAMENE	http://www.gramene.org/	A curated, open-source, web-accessible data resource for comparative genome analysis in the grasses.
GrainGenes	http://wheat.pw.usda.gov/	A suite of services for the Triticeae and oat communities, including databases, documents, tools, data files, websites, announcements, curation and community assistance.
ArMet	http://www.armet.org/	A framework for the description of plant metabolomics experiments and their results.
MapMan	http://gabi.rzpd.de/projects/MapMan/	A user-driven tool that displays large datasets (e.g. gene expression data from Affymetrix arrays) onto diagrams of metabolic pathways or other processes.
PlantMarkers	http://markers.btk.fi/	A database of predicted plant molecular markers (e.g. SSR, SNP and COS markers).
HarvEST	http://harvest.ucr.edu/	EST database viewing software that emphasizes gene function and is oriented to comparative genomics and the design of oligonucleotides to support activities such as microarray content design, function annotation, physical and genetic mapping.
PEDANT	http://pedant.gsf.de/	A genome database that provides exhaustive automatic analysis of genomic sequences using a large range of bioinformatics tools.
Sources of multiple tools		
Tools for datamining	http://www.ncbi.nlm.nih.gov/Tools/	Common bioinformatic tools such as BLAST (for comparing gene and protein sequences against others in public databases), ORFfinder (for identification of all possible open reading frames in a DNA sequence) and e-PCR (to search DNA sequence for sequence tagged sites) for genome analysis.
Bioinformatic.Net	http://www.bioinformatics.vg/	A directory for bioinformatics, genomics, proteomics, biotechnology and molecular biology that lists databases and bioinformatic tools and analyses.
Genamics SoftwareSeek	http://genamics.com/software/	A repository and database of freely distributable and commercial tools for use in molecular biology and biochemistry.
Sequence Manipulation Suite	http://wire.ndsu.nodak.edu/DEALING/DMtools/SMS	A collection of web-based programs for analyzing and formatting DNA and protein sequences.
Molecular Biology Database Collection	http://www3.oup.co.uk/nar/database/c/	A compilation of nucleotide sequence databases, RNA sequence databases, protein sequence database, structure databases, metabolic and signaling pathways, microarray data and other gene expression databases and plant databases (including for *Arabidopsis*, rice and other plant databases such as BarleyBase, CR-EST, Mendel, PlantCARE, PlantGDB).

Abbreviations: CAPS, cleaved amplified polymorphic sequence; COS, conserved orthologous set of markers; EST, expressed sequence tag; SNP, single-nucleotide polymorphism; SSR, simple sequence repeat.

of those with a LOD score ≥7 were colocated with the gene when the gene location was known. Gene–gene interactions similar to epistasis were also reported, and the interacting eQTLs were sometimes found on different chromosomes. More recently, Matias Kirst and colleagues [24] used this approach to map expression profiles associated with xylem growth in eucalyptus. Using 91 lines from an interspecific backcross between *Eucalyptus grandis* and *Eucalyptus globules*, they identified many gene expression patterns correlated with differences in xylem growth. Many of the differentially expressed genes are known to be involved in the biosynthesis of lignin and lignin components, and they shared eQTL with a wood growth QTL (based on DNA polymorphism data and wood growth phenotype). Also, in a study of the effects of artificial selection on the maize genome (other than an expression genetics study), Stephen Wright and colleagues [25] have also shown the clustering of candidate genes with putative functions in plant growth near QTLs contributing to phenotypic differences between maize and its wild progenitor teosinte. Thus, the colocalization of candidate genes with QTLs controlling a particular phenotype supports the use of the candidate gene as a potential source for developing perfect markers for selecting the phenotype in marker-assisted breeding (Figure 1).

To reduce the number of eQTL tests, dimension reduction and correlation analyses can be used to select expression phenotypes of genes tentatively associated with the physiological phenotype. These eQTL locations are then compared with QTL locations for the phenotypes of interest using confidence intervals to identify the number and location of genes affecting trait-related gene expression. To analyse putative overlapping QTLs further, multiple-trait QTL analyses can be used, taking advantage of structured correlation of the data for a robust statistical test of pleiotropy versus linkage [26].

It is important to realize that, in a recent comparison of two eQTL studies in human cell lines [27,28], Dirk-Jan de Koning and Chris Haley [19] suggested that results of eQTL analyses should be interpreted with caution. They have shown how technical and environmental factors (that might not have been taken into account) can result in the detection of false ‘hot spots’ or hubs of *trans* acting eQTLs that affect the expression of many more genes than expected by chance.

Exploitation of natural variation in germplasm collections

Extensive germplasm collections are available for crop plant species but, to date, there have been relatively few comprehensive characterizations using molecular markers. There are several strategies for exploiting the variation in germplasm collections and they vary according to the objectives of the breeding program (Figure 1). For some traits, it might be necessary to use wild ancestors of crop plants and to introgress some of the diversity that was lost during domestication to improve agricultural yields under optimal as well as stress conditions. Most of the genetic variation present in wild species and unadapted germplasm available in gene banks has a negative effect on the adaptation of plants to agricultural environments; hence, the challenge is to identify and make use of the advantageous alleles in a breeding programme. This is particularly the case for quantitative traits because the value of a wild or exotic accession for contributing useful alleles cannot be determined *a priori* with certainty. The concern that breeding is reducing genetic diversity is controversial. For example, Elena Khlestkina and colleagues [29] examined the genetic diversity of cultivated wheat that was sampled over 50 years in Europe and Asia, and found no significant change, whereas Yong-Bi Fu and colleagues [30] reported a loss of 19% of SSR alleles over 100 years. The key questions are whether the marker alleles that are lost are associated with undesirable trait alleles and whether desirable linked trait alleles are lost with those that are eliminated. These are difficult questions that are likely to have different answers for different breeding programs.

Advanced backcross QTL analysis

Many useful traits have been transferred from wild relatives into crop species, most of which are controlled by single genes or gene clusters conferring resistance to various diseases [31]. For transferring the QTLs of agronomically important traits from a wild species into a crop variety, an approach named ‘advanced backcross QTL analysis’ (AB-QTL) was proposed by Steven Tanksley and Clare Nelson [32]. In this approach, a wild species is backcrossed to a superior cultivar with selection for domestication traits. Selection is imposed to retain individuals that exhibit domestication traits such as non-shattering. The segregating BC_2F_2 or BC_2F_3 population is then evaluated for traits of interest and genotyped with polymorphic molecular markers. These data are then used for QTL analysis, potentially resulting in the identification of QTLs while transferring these QTLs into adapted genetic backgrounds.

The AB-QTL approach has been evaluated in many crop plant species to determine whether genomic regions (QTLs) derived from wild or unadapted germplasm have the potential to improve yield [33–37]. However, the wild species’ chromosome segments mask the magnitude of some favorable effects that were identified for certain introgressed alleles [38]. Thus, the yield-promoting QTL did not make a substantial contribution to the phenotype and the best lines were inferior to commercial cultivars. However, in tomato, the pyramiding of independent yield promoting chromosome segments resulted in new varieties with increased productivity under normal and stress conditions [39]. One disadvantage is that the value of the wild accession for contributing useful QTL alleles is unknown before a major investment in mapping. Another major limitation to AB-QTL is difficulty in maintaining an adequate population size in selected backcross populations so that useful alleles are not lost and the QTLs can be accurately mapped.

Association mapping based on linkage disequilibrium

The primary goal of association mapping is to detect correlations between genotypes and phenotypes in a sample of individuals based on linkage disequilibrium (LD). Biparental populations such as doubled haploids (DHs), F_2 or recombinant inbred lines (RILs) have been widely used to construct molecular marker maps and to identify genes or QTLs for traits of interest. However, these mapping populations are the products of just one or a few cycles of meiotic recombination, limiting the resolution of genetic maps, and are often not representative of germplasm that is actively used in breeding programs. By contrast, the use of unrelated genotypes or natural populations in association mapping can provide greater resolution for identifying genes responsible for variation in a quantitative trait [40–42]. Details about linkage disequilibrium (LD), its measurement and decay, and factors affecting it have been reviewed in many articles [41–43] and are therefore not covered here.

For a study of marker–trait association based on LD, two methods have been proposed: (i) candidate gene sequencing; and (ii) whole genome scanning [44,45]. Whole genome scan and candidate gene methods are similar and differ primarily in the scale at which the analysis is performed. Understanding the level of LD across the genome in the sample population will facilitate the choice of appropriate method and germplasm for genetic association mapping. Where there is significant LD, of the order of several hundreds of kilobases or more, it might be feasible to identify genetic regions that are associated with a particular trait of interest by scanning the entire genome with closely linked markers. However, if the LD declines rapidly around or in the causative genes, they can only be evaluated (not necessarily identified) by comparing DNA sequences of candidate genes.

One of the primary limitations of LD-based association mapping in plant species has been the frequent occurrence of related subgroups in the sample, which results in a high probability of type I error. Jonathan Pritchard and colleagues [46] proposed a Bayesian approach for inferring population structure based on unlinked markers. The program Structure (Table 1) assigns individuals to subpopulations and uses that information to test marker–trait associations. This method was extended by Jeff Thornsberry and colleagues [47] for the analysis of quantitative traits by using the matrix of population assignments and the quantitative traits as predictors in a logistic regression model, in which the dependent variable is a binary genetic polymorphism. If the marker allele is unique in the population, marker–trait association is only expected when a QTL is tightly linked to the marker (unless the marker allele pre-exists in the breeding population) because the accumulated recombination events occurring since a common ancestor will reduce or eliminate the marker–trait association if the QTL is not tightly linked to a molecular marker.

In crop plant species, marker–trait associations have been demonstrated by exploiting candidate gene sequencing methodology [42,43,47,48]. Comprehensive genome-wide scans for polymorphism using current technologies are generally not practical for plant species with large genomes and limited genomic resources. Thus, the alternative approach of focusing on variation in candidate genes or DNA markers closely linked to previously identified QTLs is the most appropriate strategy. A high degree of LD facilitates association analysis of markers linked to a QTL but high LD hinders the identification of candidate genes [42]. In maize, the rapid decay of LD provides a means of identifying candidate genes with high precision and at the same time allows one to associate alleles with phenotypic values [47]. For those species with high LD, comparative mapping and transcript profiling are necessary to narrow the target region. A longer-term goal for crop plants is to develop resources such as haplotype maps for genome-wide association studies. SNPs are the most abundant form of DNA variation and hundreds of thousands of SNPs are required for whole genome coverage. A subset of common SNPs that is maximally informative must be selected for association mapping. A haplotype map is a useful resource for designing LD studies and association mapping because it consists of selected SNPs that belong to blocks of limited diversity and that describe a high proportion of the genotypes in various populations with a frequency of more than 5%. Such maps can be used to identify regions of the genome associated with traits of interest in populations with high LD as well as candidate genes in populations with low LD. Haplotype maps will be particularly useful for whole genome selection.

Allele mining or EcoTILLING

To devise plant breeding strategies for crop improvement, a breeder would ideally like to know the relative value of all alleles for genes of interest in the primary germplasm, an unlikely prospect. However, information can be gathered for all alleles of a fully characterized gene in a germplasm collection and the process is known as ‘allele mining’. In this context, a strategy based on targeting induced local lesions in genomes (TILLING), called EcoTILLING, was developed for detecting multiple types of polymorphisms in germplasm collections (e.g. natural population, breeding or gene bank materials) [49]. EcoTILLING allows natural alleles at a locus to be characterized across many germplasms, enabling both SNP discovery and haplotyping. This can be done at a fraction of the cost of SNP genotyping or haplotyping methods, which require large scale sequencing. Haplotypes generated after EcoTILLING across a range of germplasm can be binned (sorted into groups) and confirmatory sequencing done on only the unique haplotypes.

EcoTILLING is expected to provide a series of alleles for those genes that are involved in important processes of the plant even though the known variants for these genes have not been observed through genetic studies. Extensive information about the candidate genes in terms of structure and regulation or phenotypic expression is important for designing the primer pairs for EcoTILLING. The necessity of also screening regulatory regions, which are often distant from the effector genes, indicates that selecting the candidate sequences for EcoTILLING is not a trivial task. After identifying all alleles that are available, they must be evaluated for their relative value in adapted genotypes in the target environment. These analyses might help in designing synthetic alleles that are superior to those found in nature. This could be accomplished by recombining the coding regions of genes either randomly (e.g. by gene shuffling) or deliberately (e.g. by domain swapping).

Challenges in phenotyping

Successful exploitation of genomics tools and strategies in plant breeding programmes requires extensive and precise phenotyping of agronomic traits for breeding materials, mapping populations and natural populations or gene bank materials. Dissecting phenotypes into components can improve heritability and aid our understanding of biological systems causing the phenotype. Another strategy for linking a gene with phenotype is phenotypic characterization of large mutagenized populations (mutant plants) or TILLING populations ('phenomics').

Gene networks

There is a great plasticity in plant genomes, which makes it possible to produce various phenotypes from little genetic variation. Other complexities, reviewed and proposed by Michele Morgante and Francesco Salamini [50], should also be considered. One example is the role of epistasis in QTL variation. Simulation studies have shown the key role of epistasis in the long term evolution of adaptive traits and in the dynamics of population divergence [51]. Similarly, the epigenetic phenomenon and the relationships between gene silencing, DNA methylation, RNA interference and heterochromatic DNA have demonstrated the complexity of RNA regulation operating through small non-coding RNAs. For instance, after analyzing and comparing the genome sequence data of human, dog, chicken, mouse and rat, Benjamin Lewis and colleagues [52] reported that the nucleotide sequence of regulatory microRNAs has been conserved for at least 310 million years. MicroRNAs were found to have direct regulatory effects on more than 5300 human genes, comprising 30% of the genome. The presence of microRNAs and their role in development and morphogenesis in plant systems has been confirmed [53]. Michael Axtell and David Bartel [54] used a microarray designed to measure expression of microRNAs and found that, in tissues in which a given microRNA was highly expressed, the corresponding gene target was unlikely to show high expression. The regulatory variation of gene expression that frequently concerns gene or genomic regions (such as promoters, introns, silencers and other non-coding DNA sequences, away from transcriptional units) has been shown to be more variable than protein coding DNA sequences [54]. This regulatory variation is genetic and fully heritable in nature. Erich Grotewold [55] has proposed a model to explain how new metabolic pathways can rapidly evolve when regulatory genes are duplicated and diverge. Keiichi Mochida and colleagues [56] were able to measure differential gene expression in hexaploid wheat using SNPs to distinguish the expression profiles of homoeologous genes. As microarray technology evolves, gene networks and the regulatory factors controlling them will become a focal point for genomics-assisted breeding. At present, it is difficult to understand and to assign a measurable proportion of the phenotypic variation of a trait to regulatory mechanisms [51].

One of the least understood phenomena is epigenetics, a term that refers to a collection of stable changes in gene expression that are not caused by DNA base changes. Gene silencing is a type of epigenetic change in which gene expression is permanently lost, and includes DNA methylation, changes in the histone code and RNA interference [57]. Altering chromatin structure can cause large scale genomic effects, thus altering transcriptional activity. Andreas Madlung and Luca Comai [57] reviewed the genomic effects of stress caused by tissue culture, pathogen attack, abiotic factors and hybridization. They concluded that epigenetic regulation can be relaxed under stress conditions and that this might result in the activation of suppressed genes and secondary effects during the re-establishment of genomic order. Selection can then act on the resulting genetic and epigenetic changes in the population. The development of knowledge and tools that allow the controlled manipulation of epigenetic phenomena could lead to a new paradigm for crop improvement strategies.

A way to the future: genomics-assisted breeding

Considerable progress has been made building infrastructure for applying genomics approaches. These include one-dimensional genetic information (genome sequences), many ESTs and gene knockout populations in several plant species of biological and agronomic importance. New knowledge and new tools are changing the strategies used in crop plant research and will thus reduce the costs and increase the throughput of the assays. There is a continuing need to integrate disciplines such as structural genomics, transcriptomics, proteomics and metabolomics with plant physiology and plant breeding (Figure 1). Bioinformatics is providing the means for integration and structured interrogation of datasets that will facilitate the cross-fertilization of disciplines (Table 1).

Genomics research has successfully unraveled various metabolic pathways and provided molecular markers for agronomic traits. However, the mechanisms of epigenetic phenomena are only beginning to be understood and their potential role in crop improvement is unknown. Similarly, tantalizing bits of information concerning the possible basis of heterosis are gradually emerging. Eventual elucidation of the mechanism of heterosis might be one of the most important contributions of molecular genetics research to crop improvement.

Ultimately, the goal of the breeder will be to assay the genetic makeup of individual plants rapidly and to select desirable genotypes in breeding populations. The construction of 'graphical genotypes' of each plant or progeny row would allow the breeder to determine which chromosome sections are inherited from each parent to facilitate the selection process and perhaps to reduce the need for extensive field tests [58]. A logical extension of whole genome selection for the breeder would be to design the superior genotypes *in silico*, an approach described as ‘breeding by design’ [59]. Thus, in the post-genomics era, high-throughput approaches combined with automation, increasing amounts of sequence data in the public domain and enhanced bioinformatics techniques will contribute to genomics research for crop improvement. However, the costs of applying genomics strategies and tools are often more than is available in commercial or public breeding programmes, particularly for inbreeding crops or crops that are only of regional importance. Nevertheless, marker-assisted breeding or marker-assisted selection will gradually evolve into ‘genomics-assisted breeding’ for crop improvement. Newly developed genetic and genomics tools will enhance, but not replace, the conventional breeding and evaluation process. The ultimate test of the value of a genotype is its performance in the target environment and acceptance by farmers.

References

1 Rafalski, A. (2002) Applications of single nucleotide polymorphism in crop genetics. *Curr. Opin. Plant Biol.* 5, 94–100

2 Varshney, R.K. *et al*. (2005) Genic microsatellite markers in plants: features and applications. *Trends Biotechnol.* 23, 48–55

3 Rudd, S. *et al*. (2005) PlantMarkers – a database of predicted molecular markers from plants. *Nucleic Acids Res.* 33, D628–D632

4 Andersen, J.R. and Lubberstedt, T. (2003) Functional markers in plants. *Trends Plant Sci.* 8, 554–560

5 Gupta, P.K. and Rustgi, S. (2004) Molecular markers derived from expressed/transcribed portion of the genome in higher plants. *Funct. Integr. Genomics* 4, 139–162

6 Hospital, F. (2001) Size of donor chromosome segments around introgressed loci and reduction of linkage drag in marker-assisted backcross programs. *Genetics* 158, 1363–1379

7 Aharoni, A. and Vorst, O. (2002) DNA microarrays for functional plant genomics. *Plant Mol. Biol.* 48, 99–118

8 Potokina, E. *et al*. (2004) Functional association between malting quality trait components and cDNA array based expression patterns in barley (*Hordeum vulgare* L.). *Mol. Breed.* 14, 153–170

9 Chen, W. *et al*. (2002) Expression profile matrix of *Arabidopsis* transcription factor genes suggests their putative functions in response to environmental stresses. *Plant Cell* 14, 559–574

10 Czechowski, T. *et al*. (2004) Real-time RT-PCR profiling of over 1400 *Arabidopsis* transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J.* 38, 366–379

11 Horak, C.E. and Snyder, M. (2002) ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol.* 350, 469–483

12 Borevitz, J.O. *et al*. (2003) Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* 13, 513–523

13 Rostoks, N. *et al*. (2005) Single-feature polymorphism discovery in the barley transcriptome. *Genome Biol.* 6, R54

14 Powell, W. and Langridge, P. (2004) Unfashionable crop species flourish in the 21st century. *Genome Biol.* 5, 233

15 Tan, P.K. *et al*. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* 31, 5676–5684

16 Miklos, G.L.G. and Maleszka, R. (2004) Microarray reality checks in the context of a complex disease. *Nat. Biotechnol.* 22, 615–621

17 Larkin, J.E. *et al*. (2005) Independence and reproducibility across microarray platforms. *Nat Methods* 2, 337–343

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

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

20 Schadt, E.E. *et al*. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422, 297–301

21 Bystrykh, L. *et al*. (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using ‘genetical genomics’. *Nat. Genet.* 37, 225–232

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

23 Brem, R.B. *et al*. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* 296, 752–755

24 Kirst, M. *et al*. (2004) Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. *Plant Physiol.* 135, 2368–2378

25 Wright, S.I. *et al*. (2005) The effects of artificial selection on the maize genome. *Science* 308, 1310–1314

26 Jiang, C. and Zeng, Z-B. (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140, 1111–1127

27 Monks, S.A. *et al*. (2004) Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* 75, 1094–1105

28 Morley, M. *et al*. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* 430, 743–747

29 Khlestkina, E.K. *et al*. (2004) Genetic diversity in cultivated plants – loss or stability? *Theor. Appl. Genet.* 108, 1466–1472

30 Fu, Y.B. *et al*. (2005) Allelic reduction and genetic shift in the Canadian hard red spring wheat germplasm released from 1845 to 2004. *Theor. Appl. Genet.* 110, 1505–1516

31 Friebe, B. *et al*. (1996) Characterization of wheat – alien translocations conferring resistance to diseases and pests: current status. *Euphytica* 91, 59–87

32 Tanksley, S.D. and Nelson, J.C. (1996) Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theor. Appl. Genet.* 92, 191–203

33 Bernacchi, D. *et al*. (1998) Advanced backcross QTL analysis in tomato. I. Identification of QTLs for traits of agronomic importance from *Lycopersicon hirsutum*. *Theor. Appl. Genet.* 97, 381–397

34 Xiao, H. *et al*. (1998) Identification of trait-improving quantitative trait loci alleles from a wild rice relative, *Oryza rufipogon*. *Genetics* 150, 899–909

35 Ho, C. *et al*. (2002) Improvement of hybrid yield by advanced backcross QTL analysis in elite maize. *Theor. Appl. Genet.* 105, 440–448

36 Frary, A. *et al*. (2004) Advanced backcross QTL analysis of a *Lycopersicon esculentum*×*L. pennellii* cross and identification of possible orthologs in the *Solanaceae*. *Theor. Appl. Genet.* 108, 485–496

37 Wang, D. *et al*. (2004) Identification of putative QTL that underlie yield in interspecific soybean backcross populations. *Theor. Appl. Genet.* 108, 458–467

38 Septiningsih, E.M. *et al*. (2003) Identification of quantitative trait loci for grain quality in an advanced backcross population derived from the *Oryza sativa* variety IR64 and the wild relative *O. rufipogon*. *Theor. Appl. Genet.* 107, 1433–1441

39 Gur, A. and Zamir, D. (2004) Unused natural variation can lift yield barriers in plant breeding. *PLoS Biol.* 2, e245

40 Tenaillon, M.I. *et al*. (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp *mays* L.). *Proc. Natl. Acad. Sci. U. S. A.* 98, 9161–9166

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

42 Flint-Garcia, S.A. *et al*. (2003) Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.* 54, 357–374

43 Gupta, P.K. *et al*. (2005) Linkage disequilibrium and association studies in higher plants: Present status and future prospects. *Plant Mol. Biol.* 57, 461–485

44 Weiss, K.M. and Clark, A.G. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* 18, 19–24

45 Hinds, D.A. *et al*. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* 307, 1072–1079

46 Pritchard, J.K. *et al*. (2000) Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170–181

47 Thornsberry, J.M. *et al*. (2001) *Dwarf8* polymorphisms associate with variation in flowering time. *Nat. Genet.* 28, 286–289

48 Palaisa, K. *et al*. (2004) Long-range patterns of diversity and linkage disequilibrium surrounding the maize *Y1* gene are indicative of an asymmetric selective sweep. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9885–9890

49 Comai, L. *et al*. (2004) Efficient discovery of DNA polymorphisms in natural populations by EcoTILLING. *Plant J.* 37, 778–786

50 Morgante, M. and Salamini, F. (2003) From plant genomics to breeding practice. *Curr. Opin. Biotechnol.* 14, 214–219

51 Yedid, G. and Bell, G. (2002) Macroevolution simulated with autonomously replicating computer programs. *Nature* 420, 810–812

52 Lewis, B.P. *et al*. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 15–20

53 Kidner, C.A. and Martienssen, R.A. (2003) Macro effects of microRNAs in plants. *Trends Genet.* 19, 13–16

54 Axtell, M.J. and Bartel, D.P. (2005) Antiquity of microRNAs and their targets in land plants. *Plant Cell* 17, 1658–1673

55 Grotewold, E. (2005) Plant metabolic diversity: a regulatory perspective. *Trends Plant Sci.* 10, 57–62

56 Mochida, K. *et al*. (2003) Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags. *Mol. Genet. Genomics* 270, 371–377

57 Madlung, A. and Comai, L. (2004) The effect of stress on genome regulation and structure. *Ann. Bot.* 94, 481–495

58 Young, N.D. and Tanksley, S.D. (1989) Restriction fragment length polymorphism maps and the concept of graphical genotypes. *Theor. Appl. Genet.* 77, 95–101

59 Peleman, J.D. and van der Voort, J.R. (2003) Breeding by design. *Trends Plant Sci.* 8, 330–334

60 Gupta, P.K. and Varshney, R.K. (2004) *Cereal Genomics*, Kluwer Academic Publishers

61 Van den Bosch, K.A. and Stacey, G. (2003) Summaries of legume genomics projects from around the globe. Community resources for crops and models. *Plant Physiol.* 131, 840–865

62 Stein, N. and Graner, A. (2004) Map-based gene isolation in cereal genomes. In *Cereal Genomics* (Gupta, P.K. and Varshney, R.K., eds), pp. 331–360, Kluwer Academic Publishers

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

64 Goff, S. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296, 92–100

65 Yu, J. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296, 79–92

66 International Rice Genome Sequencing Project. (2005) The map-based sequence of the rice genome. *Nature* 436, 793–799

67 Yu, J. *et al*. (2005) The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* 3, e38

68 Feng, Q. *et al*. (2002) Sequence and analysis of rice chromosome 4. *Nature* 420, 316–320

69 Bedell, J.A. *et al*. (2005) Sorghum genome sequencing by methylation filtration. *PLoS Biol.* 3, e13

70 Ahn, S. *et al*. (1993) Homoeologous relationships of rice, wheat and maize chromosomes. *Mol. Gen. Genet.* 241, 483–490

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

72 Gaut, B.S. (2002) Evolutionary dynamics of grass genomes. *New Phytol.* 154, 15–28

73 Sorrells, M.E. *et al*. (2003) Comparative DNA sequence analysis of wheat and rice genomes. *Genome Res.* 13, 1818–1827

74 Salse, J. *et al*. (2004) New *in silico* insight into the synteny between rice (*Oryza sativa* L.) and maize (*Zea mays* L.) highlights reshuffling and identifies new duplications in the rice genome. *Plant J.* 38, 396–409

75 Klein, P.E. *et al*. (2003) Sequence-based alignment of sorghum chromosome 3 and rice chromosome 1 reveals extensive conservation of gene order and one major chromosomal rearrangement. *Plant J.* 34, 605–621

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

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

78 Brunner, S. *et al*. (2005) Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* 17, 343–360

79 Edwards, D. and Batley, J. (2004) Plant bioinformatics: from genome to phenome. *Trends Biotechnol.* 22, 232–237