Natural Variations and Genome-Wide Association Studies in Crop Plants

Xuehui Huang and Bin Han

National Center for Gene Research, Shanghai Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China; email: xhhuang@ncgr.ac.cn, bhan@ncgr.ac.cn

Annu. Rev. Plant Biol. 2014. 65:531–51

First published online as a Review in Advance on November 20, 2013

The *Annual Review of Plant Biology* is online at plant.annualreviews.org

This article's doi: 10.1146/annurev-arplant-050213-035715

Keywords

crop diversity, genotyping, GWAS, domestication, breeding

Abstract

Natural variants of crops are generated from wild progenitor plants under both natural and human selection. Diverse crops that are able to adapt to various environmental conditions are valuable resources for crop improvements to meet the food demands of the increasing human population. With the completion of reference genome sequences, the advent of high-throughput sequencing technology now enables rapid and accurate resequencing of a large number of crop genomes to detect the genetic basis of phenotypic variations in crops. Comprehensive maps of genome variations facilitate genome-wide association studies of complex traits and functional investigations of evolutionary changes in crops. These advances will greatly accelerate studies on crop designs via genomics-assisted breeding. Here, we first discuss crop genome studies and describe the development of sequencing-based genotyping and genome-wide association studies in crops. We then review sequencing-based crop domestication studies and offer a perspective on genomics-driven crop designs.

Contents

INTRODUCTION

Natural variants in crop plants resulted mainly from spontaneous mutations in their wild progenitors. Since the beginnings of agriculture 10,000 years ago, a huge number of diverse crops adapted to various environmental conditions have been cultivated. Crop domestication and breeding have had a profound influence on the genetic diversity present in modern crops. Understanding the genetic basis of phenotypic variation and the domestication processes in crops can help us efficiently utilize these diverse genetic resources for crop improvement.

To meet the food demands of billions of people, it is critical to improve crop productivity through efficient breeding (27, 108, 133). The use of naturally occurring alleles has greatly increased grain yield. Through the use of huge germplasm resources and genetic tools such as genome sequences, genetic populations, haplotype map data sets, genome-wide association studies (GWAS), and transformation techniques, crop researchers are now able to extensively and rapidly mine natural variation and associate phenotypic variation with the underlying sequence variants. Recently, the advent of second-generation sequencing has facilitated the discovery and use of natural variation in crop design and genome-wide selection (8, 30).

ADVANCES IN CROP GENOME SEQUENCING

The reference genome sequences of crops are the basis of crop genetic studies. They are also important for rapidly investigating genetic variations in natural variants of crops. Since the rice genome was completely sequenced approximately 10 years ago (22, 28, 46, 89, 94, 130), the reference genome sequences of several other major crops—including barley, millet, maize, sorghum, potato, tomato, and *Brassica rapa*—have been reported (7, 37, 48, 70, 83, 87, 96, 97, 110, 111, 116, 132; for a review, see 8). The reference sequences can be determined through several strategies: the clone-by-clone (e.g., bacterial artificial chromosome) approach (as was done, e.g., in rice and maize), the whole-genome shotgun approach (as was done, e.g., in sorghum and foxtail millet), or a combination of the techniques (as was done, e.g., in tomato and barley). Clone-by-clone sequencing provides a way to achieve high-quality sequence assemblies for genomes of great importance. Because de novo assembly from whole-genome shotgun sequencing often results in a large number of sequence gaps, especially when using second-generation sequencing technology, it is necessary to supplement it with other information to construct long superscaffolds. These supplements may include the generation of long-insert paired-end reads, physical maps from bacterial artificial chromosome–end sequences (or fingerprinting), and high-density genetic maps. The quality of the reference sequence, which greatly affects subsequent research, depends on both

the sequencing strategy and the genome of the crop to be sequenced: Different crop species have large variations in genome size, proportion of repetitive sequences, and ploidy level (24). Generating a high-quality assembly for the large, complex crop genomes of species such as bread wheat (*Triticum aestivum*) remains a significant challenge (10, 51, 65).

Resequencing a large number of diverse varieties, facilitated by second-generation sequencing technologies and new computational biological approaches, is feasible after the release of a crop's reference genome, which is then used to characterize genome-wide variation for genetic mapping and evolutionary studies (15, 29, 41, 52, 59, 60, 127). High-throughput resequencing has rapidly expanded our knowledge of genetic variations in crops. Sequencing of six elite maize inbred lines identified more than 1.2 million single-nucleotide polymorphisms (SNPs) and more than 30,000 insertions or deletions (indels) and also uncovered hundreds of presence/absence variations of intact expressed genes (59). Among various types of sequence variants, SNPs are the most abundant (an order of magnitude greater in number than all other polymorphisms) and are also easy to identify technically (**Figure 1**), which is why only SNP markers have been widely used in most high-throughput genotyping analyses (21, 72). In second-generation sequencing, small indels (typically <6 base pairs) are usually discovered through direct alignment, and large structural variations are usually discovered by read depth, often with a relatively high level of false negatives. To capture a full catalog of sequence variants in a crop, it is best to deep sequence several representative varieties and then perform whole-genome de novo assembly and comparative genome analysis (25).

HIGH-THROUGHPUT GENOTYPING

Genome variations in crops can be defined by genotyping each individual cultivar. The genotype is an individual's full hereditary information, often defined to be the allele pattern at multiple molecular markers. An individual's phenotype is the observable characteristics that are influenced by both the genotype and the environment. Classical genetics identifies the genotype (generally adjacent to the causal variant) that is responsible for a phenotype, through linkage mapping in a recombinant population or association mapping in a natural population.

There are many types of molecular markers corresponding to different genotyping approaches. The use of polymerase chain reaction (PCR)–based markers followed by allele scoring on agarose gel laid the foundation for quantitative trait locus (QTL) mapping and gene cloning in recent years. However, the genotyping processes that are based on PCR markers are quite laborious, expensive, and time consuming when high-density genotyping is needed for a large number of individuals. The advent of high-throughput genotyping technology, coupled with the availability of reference genome sequences for multiple major crops, has greatly accelerated and facilitated genotyping procedures.

Here, we summarize five high-throughput genotyping approaches (**Table 1**). The first use of high-throughput genotyping was a microarray technology that detects SNPs by hybridizing DNA to oligonucleotides spotted on the chips. This method, known as microarray-based genotyping, enables direct scanning of allelic variation across the genome, covering hundreds to thousands of SNPs in a short time (17, 102, 121). Once a comprehensive SNP data set is available for a species, a well-designed microarray can be produced; generally, the technology is then cost efficient and the process relatively convenient. Nearly all GWAS in human genetics have used microarray technology, which can scan the human genome at 0.5–1 million SNPs. Microarray-based genotyping has been applied in crops such as rice and maize (26, 61, 72, 136). Diversity arrays technology (DArT) mapping, which is also based on microarray hybridizations, has been used in many crops, such as barley (119).

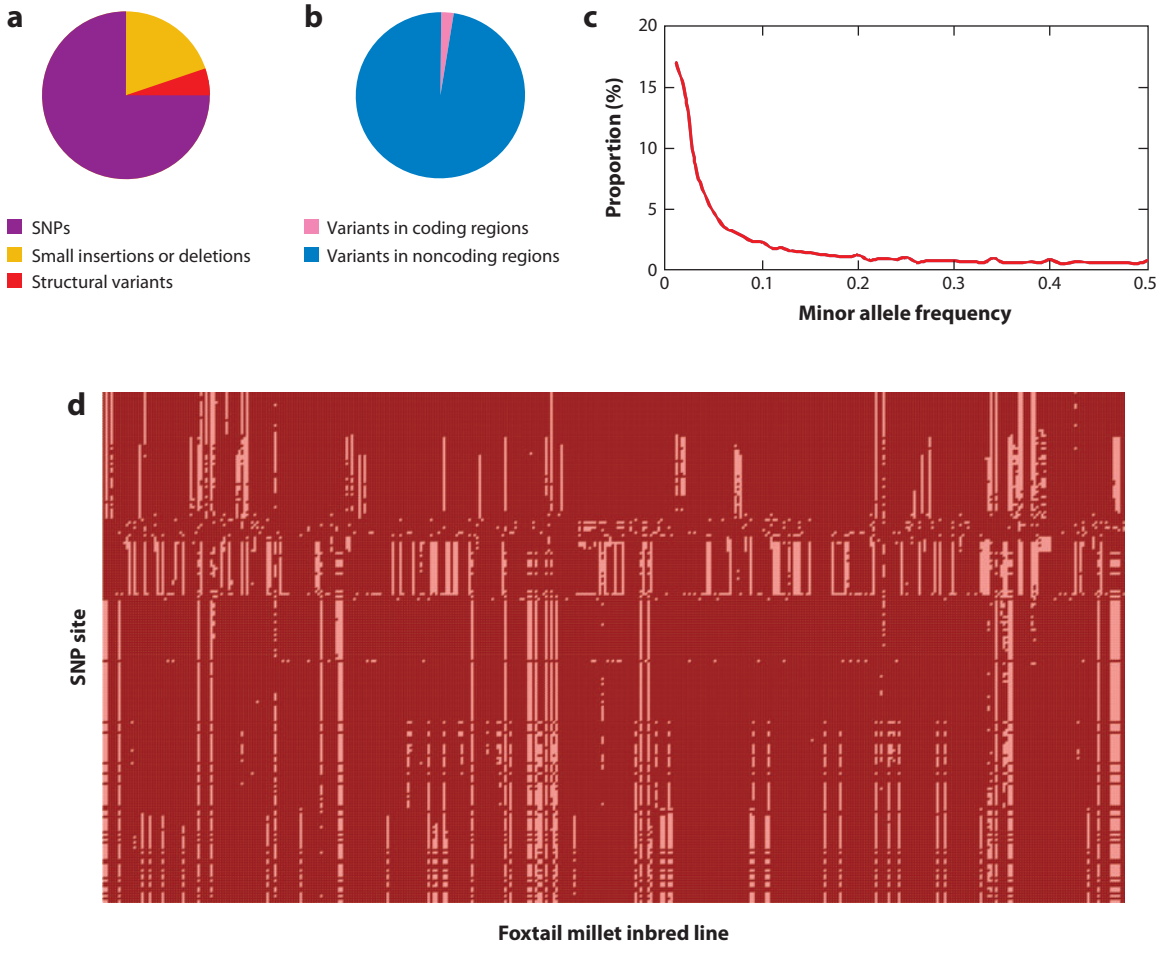

Figure 1

Genomic variation in a natural population, using real data from foxtail millet. (*a*) The proportion of three types of sequence variants in the genomes. (*b*) The proportion of sequence variants with different functional effects. (*c*) Allele frequency spectrum of all variants. (*d*) Haplotype display in a local genomic region in 400 foxtail millet varieties. Each column represents a haplotype from an inbred foxtail millet line, and each row is a single-nucleotide polymorphism (SNP) site. Dark and light red represent the major and minor alleles, respectively, at each SNP site.

The advent of second-generation sequencing technology hastened a methodological leap forward: the high-throughput sequencing-based genotyping method (99). For genotyping of recombinant populations, a method that utilizes low-coverage whole-genome resequencing data was developed (38, 137). The method was first applied in a recombinant inbred line (RIL) population that was generated from a cross between the *Oryza sativa* ssp. *japonica* Nipponbare and *Oryza sativa* ssp. *indica* 93-11 varieties (38). The genomic DNA of each line in the population was sequenced with 0.02× coverage. SNPs were identified between two parental lines, and genotype calling was based on identification of adjacent SNPs using the sliding window approach, which resulted in an ultradense genotype map. A total of 49 QTLs for 14 agronomic traits were identified at high resolution using the high-density genotype map (114). Among them, 5 QTLs of large effect were located in small genomic regions, and strong candidate genes were found in the intervals.

Table 1 Five high-throughput genotyping methods

	Microarray-based genotyping	Sequencing-based genotyping	Genotyping by sequencing	RNA-seq-based genotyping	Exon-sequencing-based genotyping
Preliminary requirement	Comprehensive SNPs available	None	A suitable restriction enzyme	None	Exon array developed
Density	Alterable	Alterable	Modest	Modest	Modest
Cost	Alterable	Alterable	Low	High	High
Experimental workload	Low	Medium	Medium	High	High
Marker distribution	Well distributed	Well distributed	Not well distributed	Not well distributed	Not well distributed
Application	Most species	Most species	Species with a large genome size	Species with a large genome size	Species with a large genome size
Additional uses	None	Identifying novel mutation variants	None	Identifying novel mutation variants and eQTL analysis	Identifying novel mutation variants

Abbreviations: eQTL, expression quantitative trait locus; SNP, single-nucleotide polymorphism.

This method now extends to different types of mapping populations in crops, including nearly isogenic lines, chromosome segment substitution lines, and F_2 populations (114, 124, 126, 129, 141). Here, we refer to this whole-genome sequencing-based genotyping (via low-pass multiplex resequencing) as sequencing-based genotyping.

The low-coverage whole-genome resequencing approach was also used in the genotyping of natural populations through completely different computation methods. In the construction of the haplotype maps in rice and foxtail millet, each variety was sequenced with low genome coverage, resulting in numerous missing genotype calls owing to low-fold sequencing. Data imputation—the process of replacing missing data with inferred values according to local linkage disequilibrium (LD)—can be used to deal with such population-scale genotype data sets. For this, a *k* nearest neighbor–based algorithm that explores local haplotype similarity to infer the missing calls was developed, and applications in both rice and foxtail millet showed high accuracy (41, 50). Many imputation pipelines are now available with different features, and researchers need to make adjustments in sequencing coverage according to the population size, diversity level, and LD decay rate of the study sample in the initial experimental designs. Some imputation methods (typically using hidden Markov model–based algorithms) are also suitable for outcrossing species whose genomes contain numerous heterozygous genotypes (11). The use and simulations of sequencing data on human populations showed that, after effective imputations, even extremely low-coverage sequencing could increase the power of GWAS when compared with the microarray method (82). Computation simulations are needed to determine the sequence coverage with different population sizes for different crop species (63).

Some crops (e.g., maize, barley, and wheat) have large genomes, and whole-genome resequencing for these crops is still expensive. There are now several alternatives to sequencing the whole genome. One method is to generate restriction-site-associated DNA tags (e.g., using *Sbf*I or *Eco*RI) and sequence them to identify polymorphic markers (6). A similar approach is to cut the genomic DNA and ligate the genomic fragments to bar-code adapters to prepare a multiplex sequencing library (19). Through the use of methylation-sensitive restriction enzymes, the regions with

transposable elements can be reduced, and the relatively low-copy regions flanking the particular restriction enzymes are enriched in the sequencing library. Genotyping by sequencing has been used in maize, sorghum, and barley, and these applications showed that this method is efficient for large-scale, low-cost genotyping despite the limitations in SNP number and distribution (19, 76, 91).

Another approach, known as RNA sequencing (RNA-seq)–based genotyping, is to retrieve genotype from RNA-seq data. Despite the great variation in genome size for different crops, there are no significant changes in number of genes or total gene sizes. Because most repetitive regions are ignored, performing transcriptome sequencing for SNP calling rather than whole-genome resequencing is quite cost efficient. For example, a recent study genotyped a large number of SNPs from 368 maize transcriptomes and then used the RNA-seq data for expression profile analysis and expression QTL (eQTL) mapping simultaneously (61). It would be much more expensive to genotype 368 maize genomes, because the repeat region occupies more than 80% of the total maize genome.

RNA-seq-based sequencing has several weak points. Because the SNP density in genic regions is much lower than that in intergenic regions, the number of SNPs called from RNA-seq data may not be large enough for GWAS, especially for high-LD crops. The existence of a strong bias in SNP distribution raises another problem: In a particular tissue at a particular time point, many genes have very low or even no expression and thus cannot be used in genotyping, but RNA preparation of multiple tissues at multiple time points for a large population would greatly increase the workload. Exon-sequencing-based genotyping, facilitated by exon capture, can also be applied to the mapping of some large, complex crop genomes (69, 78, 113) (**Table 1**).

LINKAGE MAPPING IN CROPS

To unravel the genetic basis of complex QTLs like grain yield and stress tolerance, a large sample size is key—typically thousands of individuals are needed. It is now feasible to genotype thousands of genomes using high-throughput methods. Compared with genotyping, most fieldwork is still laborious, involving the measurement of multiple traits at several time points in large-scale experiments across diverse environments. Some sensor-based platforms have been developed for measuring biomass traits, including near-infrared spectroscopy on agricultural harvesters and spectral reflectance of plant canopies (75). Future progress in phenotyping technologies will accelerate genetic mapping and gene discovery in crops.

Genetic mapping in crops is usually undertaken in segregating mapping populations, such as F_2 populations, RILs, and backcross inbred lines. Further fine mapping and gene cloning then follow, often using advanced backcross-derived populations. A mapping population derived from a cross between the *O. sativa* ssp. *indica* Kasalath and *O. sativa* ssp. *japonica* Nipponbare varieties has enabled the identification and cloning of tens of QTLs underlying a wide range of traits (32). Although this strategy has been used successfully in functional genomics studies in crops, there are two major limitations to QTL mapping in conventional recombinant populations. First, there are only a few recombination events in the mapping population; for example, typically one or two recombinations occur per chromosome in rice segregating populations, which would lead to poor mapping resolution unless very large populations are used. Second, because the sequence divergence between the selected parents in a particular segregating population represents only a small fraction of all genetic variation within a species, in a single segregating mapping population, only QTLs at which the two parents differ can be detected.

To overcome these disadvantages, some new types of populations have been constructed and used. In maize, nested association mapping (NAM) was developed to enable high power and high

resolution through joint linkage-association analysis (71). The NAM population was created by crossing 25 diverse inbred maize lines to the B73 reference line, where the NAM population in total includes ~5,000 RILs. The NAM population has been used in large-scale genetic mapping for several important traits (12, 58, 86, 109). In the model plant *Arabidopsis*, the Multiparent Advanced Generation Inter-Cross (MAGIC) population was generated, including hundreds of RILs descended from a heterogeneous stock of 19 intermated *Arabidopsis* accessions (57). Computational simulation demonstrated that the QTLs explaining 10% of the phenotypic variance can be detected with an average mapping error of approximately 300 kb when using the MAGIC population. Another team crossed eight *Arabidopsis* accessions and produced a set of six RIL populations called the *Arabidopsis* multiparent RIL (AMPRIL) population (40). QTL analysis in the AMPRIL population showed that this genetic resource was able to detect QTLs explaining 2% or more of a trait's variation.

As an alternative to conventional linkage analysis, bulk segregation analysis coupled with multiple-sample pooling sequencing can be used in the genetic mapping of simple qualitative traits or mutant mapping. Several methods have been reported for this application, including SHOREmap (98), next-generation mapping (5), MutMap (1), and MutMap-Gap (105). In *Arabidopsis*, a mutant in the Columbia (Col-0) reference background was crossed to the Landsberg *erecta* (L*er*-1) accession, and a pool of 500 mutant F_2 plants was sequenced to detect the causal mutation sites (98). To avoid potential interference from different genetic backgrounds, James et al. (49) recommended using populations derived from backcrossing the mutant line to the non-mutagenized parent for mapping by sequencing. In rice, a recessive mutant was crossed to the parental line used for the mutagenesis, and the genomes from multiple lines of mutant F_2 progeny were pooled and sequenced (1). The strategy can be further improved for application to quantitative traits in crops (e.g., grading the RILs into several sequencing pools according to a particular trait).

GENOME-WIDE ASSOCIATION STUDIES IN CROPS

GWAS take full advantage of ancient recombination events to identify the genetic loci underlying traits at a relatively high resolution. The GWAS methodology became well established in human genetics during a decade of great effort. Through global collaborations, millions of common SNPs were identified in human populations to construct a high-density haplotype map of the human genome (44, 45). Several commercial microarrays were designed for large-scale genotyping and analysis of GWAS panels, with many accompanied tool kits developed. GWAS approaches have been widely used in genetic research to identify the genes involved in human disease (2, 118). The contribution of GWAS work to understanding of the genetic basis and molecular mechanisms of common disease is evident.

With the rapid development of sequencing technologies and computational methods, GWAS are now becoming a powerful tool for detecting natural variation underlying complex traits in crops (88). Different from GWAS in humans, GWAS in crops usually use a permanent resource—a population of diverse (and preferably homozygous) varieties that can be rephenotyped for many traits and only needs to be genotyped once—and one can subsequently generate specific mapping populations for specific traits or QTLs in crops (4). Human GWAS usually adopt a case–control design: the identification of susceptibility loci for a particular disease through a population-scale genome-wide comparison between large groups of patients and healthy controls. Moreover, human GWAS have been influenced by the missing heritability problem, where most loci that they identify have a very low rate of phenotypic contribution. To detect more QTLs, the population size (up to tens of thousands of individuals) and number of markers (up to millions of SNPs or

whole genomes) must be increased continually. GWAS in crops are therefore much less costly than those in human.

GWAS have now been carried out successfully in many crops, including maize, rice, sorghum, and foxtail millet (41, 42, 50, 58, 61, 76, 109, 136). Based on the magnitude of resources already developed and published, rice and maize are the two major models for crop GWAS, and both have panels of thousands of genotyped inbreds and multiple environment trials conducted for several traits. In rice, 1,083 cultivated *O. sativa* ssp. *indica* and *O. sativa* ssp. *japonica* varieties and 446 wild rice accessions (*Oryza rufipogon*) were collected and sequenced with low genome coverage (39). A high-density haplotype map of the rice genome was constructed using data imputation, and a GWAS was then conducted to characterize the alleles associated with 10 grain-related traits and flowering time using the comprehensive data set of ~1.3 million SNPs. Several association signals were tied to previously well-characterized genes. A GWAS was also carried out in 446 *O. rufipogon* accessions for leaf sheath color and tiller angle, which would have stronger mapping power owing to higher levels of genetic diversity in the wild species. Moreover, the GWAS was performed based on the microarray-based genotyping approach. In total, 413 diverse accessions of *O. sativa* were genotyped at 44,100 SNP variants and phenotyped for 34 traits, and the result showed the complex genetic architecture of the traits in rice.

In maize, the genetic architectures of flowering time, leaf angle, leaf size, and disease resistance traits were dissected by conducting linkage mapping and GWAS jointly in the NAM panel, and multiple related candidate genes were identified (12, 58, 86, 109). The GWAS result demonstrated that the genetic architecture of these traits is dominated by many QTLs with small effects. Maize oil is an important food and energy source, and a GWAS in maize was recently performed for maize kernel oil composition (61). A total of 368 maize lines were analyzed at ~1 million SNPs genome-wide, and 74 loci were found to be associated with maize kernel oil concentration and fatty acid composition.

There are some slight differences between rice GWAS and maize GWAS. The differences are reflected in the trade-offs in power and resolution in the mapping of selfing versus outcrossing species. In the rice genome, LD generally decays at ~100 kb, which might be a result of self-fertilization and the small effective population size. Some other self-pollinated crops (e.g., foxtail millet and soybean) show similar (or slower) LD decay rates as well. Owing to extended LD, the low-coverage sequencing approach coupled with missing data imputation—which has been applied in rice GWAS—is quite efficient and powerful. The low rate of LD decay, however, also means that the resolution of GWAS in the selfing species cannot resolve a single gene. Maize, in contrast, is a standard outcrossing species because the plant has separate male and female inflorescences that differ in flowering time. Owing to its outcrossing nature, maize has rapid LD decay (within ~2 kb) and great genetic diversity, which makes it a promising model with greater power in GWAS. In most cases, the resolution of maize GWAS may reach the single-gene level. Accordingly, a typical GWAS in maize may need tens of millions of SNPs to be accurately typed genome-wide for numerous varieties, which is still challenging and costly because of the large size, abundant repeats, and paralog sequences in the maize genome.

GWAS have been successfully extended to genetic studies in other crops. A total of 916 diverse foxtail millet varieties, including both traditional landraces and modern cultivars, were genotyped through whole-genome low-coverage sequencing (50). A GWAS in foxtail millet identified multiple loci for tens of agronomic traits, which were measured in five different environments. In sorghum, 917 worldwide diverse accessions were collected, and ~0.2 million SNPs were identified through genotyping by sequencing (76). To identify the loci underlying variation in agronomic traits, a GWAS was carried out on plant height components and inflorescence architecture, and several known loci were mapped to these traits. For a barley GWAS, a team collected an

association mapping panel with 224 spring barley accessions, which were genotyped at 957 SNP sites using a genotyping microarray (81). The method identified some relevant candidate genes despite the low marker density. Another team carried out a GWAS of 32 morphologic and 10 agronomic traits in a collection of 615 barley cultivars, and the SNP density was very low as well (115). Associative transcriptomics of traits in *Brassica napus* was applied based on SNPs from mRNA data to assess the feasibility of conducting GWAS in polyploid crops (31). Genomic deletions were detected that underlie two QTLs for the glucosinolate content of seeds, which identified a strong candidate gene—the transcription factor–encoding gene *HAG1*. Bread wheat is a typical polyploid crop with a huge genome. Association analysis has been tested in *Triticum urartu*, the progenitor species of the A genome of bread wheat (65). However, implementing GWAS in wheat remains technically difficult, and great effort will be needed to overcome the challenges.

These studies show that the GWAS method in crops is a useful and robust strategy complementary to classical biparental cross mapping and has the power to genetically map multiple traits simultaneously. GWAS results are expected to be further utilized to investigate the genetic basis of plant morphology, yield, and physiology in more grasses, including close wild relatives of domesticated crops.

A matter requiring attention in crop GWAS is the need to account for the massive population structure, including weighing the trade-offs of increased false-negative and decreased false-positive rates from accounting for that structure (55, 77, 85, 112). The mixed model is the most popular method to detect genotype–phenotype associations in crop GWAS (9, 131). However, the computational burden of this model becomes impractical at large sample sizes (typically ~1,000 accessions) and numbers of markers (typically ~1 million SNPs). Some advances have substantially decreased the computation time, including the Efficient Mixed-Model Association eXpedited (EMMAX) program and the compressed mixed linear model method (53, 66, 67, 134, 135, 139). GWAPP, an interactive Web-based application for conducting GWAS in *Arabidopsis thaliana*, was developed using an efficient implementation of a linear mixed model known as an accelerated mixed model (101). These methods mainly sequentially query SNPs with the inclusion of population structure. Some additional multiple-regression approaches and nonparametric statistics have also been developed (56, 100).

It is important to note that GWAS have low power for rare alleles, which make up a substantial proportion of natural variation. In rice, ~44% of the SNPs are of low frequency (minor allele frequency < 0.05). In the case of rare alleles, either the use of a large sample size or the construction of multiple biparental cross populations (e.g., NAM or MAGIC) may be helpful.

In most cases, there are several genes in the interval of one GWAS locus (sometimes even in humans and maize), only one of which might contribute to the QTL (that is, the causal gene). Therefore, follow-up analyses of GWAS loci and additional experiments may be required to pinpoint the causal genes. Gene annotation, expression profiles, and variant catalogs can be carefully analyzed in the post-GWAS stage. For example, the nucleotide binding site–leucine-rich repeat genes are probably the causal genes in the GWAS loci for disease traits, and genes that are highly expressed in the grain-filling stage are more likely to be causal genes in GWAS of grain-related traits. The large-effect variants in the coding region of a gene can also provide important clues for detailed analysis of GWAS loci. T-DNA mutants and targeting induced local lesions in genomes (TILLING) analysis of candidate genes at the GWAS loci in crops using artificially induced mutants (e.g., ethyl methanesulfonate–induced or radiation-induced mutants) could be an effective approach for further identifying and validating gene–trait associations. More importantly and directly, additional experimental studies, including transgenic analysis, will be necessary to conclusively identify causal genes and causal variants. More information will be gained through diverse GWAS panels in crops when more traits (e.g., drought tolerance and photosynthesis efficiency)

are carefully evaluated in these crops and more functional trials are performed, which will be of great utility in addressing the important biological questions.

In addition to the phenotypic traits of ecological, agronomic, and economic interest, other molecular components reflect natural variation among individuals. The nonphenotypic variation may include diversity in gene expression level, DNA methylation, and metabolite accumulation, which may contribute to phenotypic changes through complex pathways (23, 68). Over the past decade, efforts to characterize nonphenotypic variation in natural accessions have improved our knowledge, and GWAS are the most effective way to identify eQTLs, methylation QTLs (mQTLs), and loci for various metabolism products. The rapid development of RNA-seq and MethylC-seq technologies coupled with SNP data has enabled eQTL and mQTL mapping at a previously unimaginable scale. In GWAS for eQTLs, the expression level of each gene is equivalent to the phenotype data in an ordinary GWAS. Studies in maize and *Arabidopsis* reported that a large fraction of the eQTLs were due to polymorphisms within or close to the gene (referred to as local eQTLs or *cis*-eQTLs) (54, 61, 95). In contrast, some eQTLs were mapped elsewhere in the genome (referred to as distant eQTLs or *trans*-eQTLs). In maize, the immature seeds were used for RNA extraction, and the gene expression levels were obtained by RNA-seq technology. To detect potential *cis*-eQTLs, a GWAS was performed to detect associations between the expression level of each associated gene and nearby SNPs. Among the GWAS loci, 41 exhibited a statistical correlation between DNA sequence polymorphisms and expression levels. These differences in expression levels may account for a substantial proportion of diversity in agronomic traits. Therefore, the identification of allelic expression differences and phenotypic variation will improve our understanding of how the regulatory elements affect transcript accumulation and agronomic traits afterward.

DNA methylation is a covalent base modification of nuclear genomes that can be accurately inherited during cell divisions. In plants, DNA methylation occurs on cytosines at CG, CHG, and CHH sites (where H = A, T, or C), and DNA methylation levels change in diverse natural accessions, referred to as natural epigenetic variation. Single-methylation polymorphisms are considered to be the most abundant natural epigenetic variation. A recent study carried out population-wide sequencing of the genomes and methylomes of hundreds of diverse *Arabidopsis thaliana* accessions and found that CG, CHG, and CHH single-methylation polymorphisms accounted for 23%, 13%, and 64% of all single-methylation polymorphisms, respectively (95). A GWAS was then conducted to link the genetic and methylation variants. Association analyses of the DNA methylation levels with genetic variants identified 2,739 significant mQTLs. More studies will be needed to shed light on the genetically dependent methylation variation and the underlying mechanisms.

GENOME VARIATION INSIGHTS INTO CROP DOMESTICATION

Crop domestication is the outcome of continuous selection procedures that led to increased adaptation of a plant for cultivation by humans. Human ancestors made great efforts and took a long time to improve wild plants to better suit human needs, and finally created tens of crops containing inherent differences from their progenitors (18). By a few thousand years ago, human ancestors had independently cultivated major cereal crops in various parts of the world. Based on archaeological records and genetic analysis, the geographic origins and subsequent demographic processes for most crops have been revealed (18). Bread wheat and barley were domesticated across the Fertile Crescent region in western Asia. The cultivated rice *O. sativa* was initially formed from the common wild rice *O. rufipogon* in southern China, and foxtail millet probably arose from wild green foxtail in northern China. Genetic studies have identified the wild grass teosinte in southern

Mexico as the closest relative of maize. More recently, the sweet potato (*Ipomoea batatas*) was domesticated in South America and then spread across the world, and the domestication of sorghum, now grown throughout the arid and semiarid tropics, originated in Africa. Crop domestication represents one of the most important developments in human history.

Understanding the origins and domestication processes of cultivated crops is important for modern crop breeding (for a review, see 80). The earliest agricultural practice was to grow and harvest wild plants of a favorable species, a key step from the hunter-gatherer life to agricultural civilization. Afterward, humans would select the individuals with the preferred characteristics in the wild species populations and use the favorable seeds to resow and plant the next year. During these processes, human selection and crop improvement occurred every year, and many morphological and physiological traits of the wild progenitors were reshaped. For most crops, the initial domestication may have focused primarily on converting wild types with few, smaller, and shattering seeds to those with more, larger, and nonshattering seeds, often along with changes in plant architectures. The traits under human selection in crop domestication may also include seed dormancy, flowering time, outcrossing rate (e.g., the change of mating system from outcrossing to selfing in rice and foxtail millet), and coloration. For some domestication traits (e.g., seed shattering), the crop's native characteristics are completely altered, such that the plants cannot grow or reproduce without human intervention. For most traits (e.g., seed size and flowering time), the phenotypic variation is quite broad in the wild progenitors, whereas the domesticated crops tend to have limited variation.

Detecting the phenotypic differences between a crop plant and its wild relatives is relatively easy; identifying the molecular basis for these differences is much harder. Considerable effort has been devoted to detecting QTLs and genes that are under domestication selection in crops, and tens of such genes have been identified in previous studies. The most common approach is to (*a*) construct a recombinant population from a cross between a crop plant accession and a line of its wild relatives, (*b*) perform QTL mapping to understand whether a domestication trait is controlled by a few genes of large effect or many genes of small effect, and (*c*) fine map a domestication gene with relatively large effect in a large population. It is important to note that not all QTLs for domestication traits (e.g., seed size) have been under human selection and belong to domestication loci: Some QTLs are responsible only for natural intraspecies variation, and some have alleles that exist at a very low frequency in both cultivated and wild populations (55, 107). Only genes underlying typical domestication traits whose favored alleles for this trait under cultivated conditions have been significantly different from those under natural conditions in a wild progenitor can be considered domestication-related genes.

Domestication loci can also be detected through population genetic analysis, which may rely on procedures including (*a*) use of an extensive collection of tens of diverse varieties of both a domesticated crop and its wild progenitor, (*b*) whole-genome profiling of sequence variation through direct resequencing in the populations, and (*c*) genomic screening of artificial selection signatures to detect selective sweeps (i.e., elimination of sequence variation in genomic regions linked to a recently fixed beneficial mutation). There are several computational pipelines to detect selective sweeps (e.g., simple diversity ratio, cross-population composite likelihood ratio, and cross-population extended haplotype homozygosity) (13, 79). Because it is difficult to say which method is best in all cases, knowledge of previously known loci involved in domestication of a given species is important, providing an independent clue to choose an appropriate method and parameter optimization. The threshold of sweep calling can be chosen based on a permutation test (i.e., reshuffling the species information in the combined population of the crop species and the wild species, and then performing selective sweep detection with the same parameters) and well-characterized domestication loci.

Owing to significant decreases in sequencing cost, this strategy has been successfully applied in several crops (34, 39, 43, 60, 123, 127). In rice, several reports focused on detecting selective sweeps by resequencing and analyzing diverse accessions of cultivated rice and wild rice (34, 39, 127). A systematic comparison of a large number of cultivated and wild rice accessions identified 55 domestication sweeps, including most well-characterized genes in rice (39). In maize, a comprehensive assessment of selective sweeps was implemented based on the genome-wide resequencing of 75 wild, landrace, and improved maize lines (43). Analysis of the resultant 484 domestication features revealed that maize domestication targeted hundreds of genes of diverse biological function that likely affected many unstudied aspects of phenotype.

Genetic mapping and comparative genome analysis found that the *Shattering 1* gene for seed shattering was under parallel selection during sorghum, foxtail millet, rice, and maize domestication (50, 64). Similar cases include *Waxy*, *C1*, and so on (92, 117). Because the required population-scale comparative genomics are being or will be used in more crops (such as sorghum, barley, and foxtail millet), comprehensive comparisons of the domestication sweeps in multiple crops may find more key genes under parallel selection in the grass family. However, in many cases, the same traits are controlled by different genes in the domestication of different crops (some domestication genes were identified in one crop even without orthologs in other crops).

The limitations of population genetics in analyzing domestication loci are due to the lack of biological information: We do not know why these chromosomal loci were selected during domestication. Integrating population genetic analysis and high-resolution genetic mapping may be an effective way to overcome this lack of information. A recent study constructed and genotyped a permanent mapping population from a cross between cultivated and wild rice, and identified tens of QTLs underlying a wide range of domestication traits at a relatively high resolution (39). The information on fine-mapped domestication-related QTLs enabled identification of the selective sweeps associated with the traits. Crosses between teosinte and maize have also been widely used in maize genetic studies to investigate the functions of domestication loci and identify the causal genes (18).

Wild populations of most crops have not been effectively used so far owing to difficulties in both collection and planting, but close wild relatives of domesticated crops are important systems for ecological and population genetic research. Wild relatives of crop plants, which have adapted to various climatic changes and survived until today, often contain many excellent traits of agricultural importance, including tolerance to biotic and abiotic stresses, and are thus thought to be valuable genetic resources for crop improvements. The rich gene alleles and great genetic diversity in wild progenitors may be lost because of genetic bottlenecks in crop domestication. Through comprehensive collection and assessment of wild relatives, it is possible to uncover excellent genetic resources that can then be used to discover new alleles and provide genetic donor material to breeders.

The causal variants underlying domestication traits can arise from either de novo mutation or standing variation in wild populations (84). For the alleles with standing variation, GWAS of the traits in the wild populations will be an alternative approach to identifying domestication genes. Genome assemblies of wild species are therefore quite useful in both studying domestication and determining the utility of wild resources. To unlock the genetic potential of wild rice, the International *Oryza* Map Alignment Project was initiated with the goal of completing the sequencing and assembly of all 23 species in the *Oryza* genus (47, 120). So far, the 261-Mb de novo assembly genome sequence of *Oryza brachyantha* (genome FF) and the de novo assembly of *Oryza rufipogon* (genome AA) have been reported (14, 39), and the sequence assembly and genome annotation of other wild rice species are under way. Along with the construction of advanced mapping populations

using wild species, the 23 reference genome sequences will provide an important resource for future breeding and molecular applications in the genus *Oryza*.

CROP DESIGNS BY GENOME-WIDE SELECTION

The study of natural variation and genetic mapping in crops has brought great advances in recent years, and now it is time to put more effort into applying this knowledge to crop breeding through molecular design (breeding that designs the crossing and selection procedures to generate a line with expected genotypes). Although genetic modification technology can provide an effective way to improve specific traits like insect resistance in crops, breeding programs using natural allelic variation are still the most common approach in practice because the yield traits are generally controlled by numerous QTLs. There is also public concern over the use of genetically modified crops. Therefore, the use of molecular design in breeding is clearly a feasible alternative to genetic modification technology.

Successful application of breeding through molecular design depends primarily on close and long-term interactions between breeders and geneticists. To establish a platform for such an interaction, a comprehensive digital map of the elite varieties could be constructed collectively. The digital information includes high-density genotypes and measurements of multiple agronomic traits in diverse environments for different varieties. The data set will be essential in molecular design, because the elite varieties have accumulated some favorite alleles, and most genetic improvement programs need to be based on these resources along with further supplementation of elite alleles from diverse landraces and wild relatives. For disease resistance and grain quality, the well-characterized major QTLs or genes can be screened in the elite varieties, followed by fine tuning through marker-assisted selection (3, 36, 74, 93, 106, 140). Moreover, the biparental population of a selected cross, coupled with the information on associated loci in GWAS and selective sweeps between tradition landraces and modern cultivars, can be directly applied in crop breeding through genomics-assisted selection.

The yield traits are usually highly polygenic, controlled by a large number of QTLs with small effects; QTLs with large effects would probably be under intensive selection during domestication and recent breeding, such that the deleterious alleles with large effects are rare or extinct in the gene pools of modern elite cultivars (125). Marker-assisted selection of only a few QTLs will therefore have limited effects on the grain yield. In this situation, whole-genome prediction, which is a method for predicting phenotypes using high-density genotype data, might be complementary to breeding design. This method was recently applied in an inbred line population to estimate general combining ability in maize (90). A collection of 285 diverse maize inbred lines was genotyped at 56,110 SNP sites using a microarray, and 130 metabolites were also surveyed using gas chromatography–mass spectrometry. The lines were crossed with two testers, and the F_1 individuals were phenotyped in six environments for seven traits. To predict the general combining ability of the inbred lines for seven traits of the hybrids, SNP and metabolite data were fitted into models, and both of them provided high predictive accuracy (from 0.72 to 0.81). This study demonstrates that in silico genomic design based on the combination of all superior SNPs has the potential to predict yield performance to generate new superior varieties.

Although both GWAS and whole-genome prediction rely on high-density genotypes and phenotypes, there are some differences between the two approaches (**Figure 2**). For example, the goals of the two methods are different. The GWAS method aims to identify individual loci (although there are still some problems in detecting epistatic interactions) and related genes significantly associated with a trait, and follow-up gene functional studies then analyze the molecular mechanism. Whole-genome prediction aims to provide accurate modeling from the genotype to the

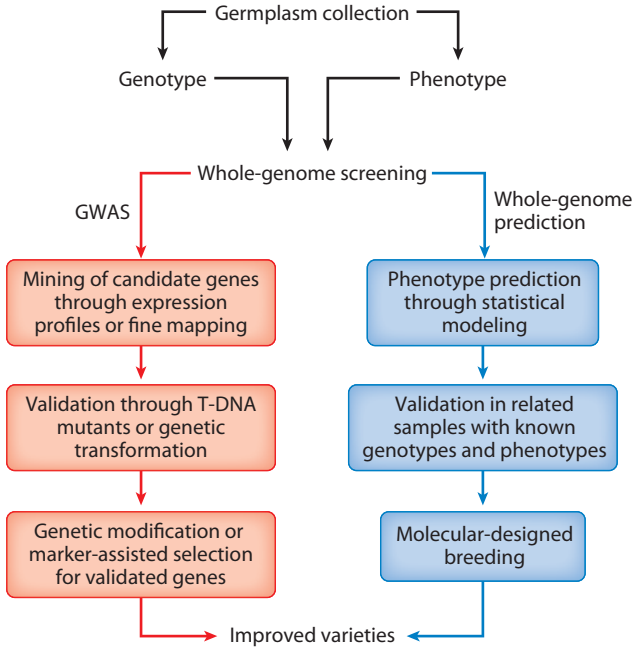

Figure 2

A schematic view of a genome-wide association study (GWAS) and whole-genome prediction using genotype and phenotype data in diverse crop varieties.

phenotype, and a breeding program with designed crosses and marker-assisted selection can then be carried out under computational guidance (73, 122). Therefore, the training population in whole-genome prediction is generally the same as the breeding population that is being used, and the traits for prediction should be closely related to crop improvement; GWAS, in contrast, generally use a population with a high level of genetic diversity, and all kinds of phenotypes can be examined. Compared with GWAS, whole-genome prediction is still in its development stage, and many questions remain, such as how to select the training population and the most appropriate statistical models (20, 30, 128).

Hybrid breeding is one of the most important aspects of crop breeding. Through crossing of different inbred lines, produced F_1 hybrids often provide higher yields than both parents, a phenomenon known as heterosis or hybrid vigor. Commercial maize and rice seeds now include a substantial proportion of such F_1 hybrids, which has contributed greatly to global food security. Crop breeders make and select many parent inbred lines every year that are then used in crosses to generate F_1 hybrids to test their yield performance. A small proportion of these hybrids have improved performance, which can then be adopted for commercialization. Although heterosis is widely used in agriculture and has been investigated for decades because of its direct and practical implications for hybrid crop breeding, its genetic basis remains poorly understood. In rice, RILs were intercrossed to create an experimental population referred to as an immortalized F_2 (35, 62, 138). The yield component traits, including tiller number, grain number, and grain weight, were then phenotyped in the immortalized F_2 population. These studies further assessed the relative contributions of dominance, overdominance, and epistasis to heterosis in hybrid rice. Whole-genome profiling of gene expression changes between F_1 hybrids and their parents was also assayed in rice and maize to detect potential heterosis-associated genes and their unique

Table 2 Levels of genetic diversity during domestication and modern breeding in three crop species

	Wild progenitors	Landraces	Modern cultivars	Reference(s)
Maize	0.0059	0.0049	0.0048	42
Rice	0.0030	0.0024	Not reported	38, 39
Foxtail millet	Not reported	0.0010	0.0009	48

expression patterns (16, 33, 103, 104). With more information on heterosis studies, integration of new genomics technologies with traditional hybrid breeding strategies will help breeders design and select the best combinations.

SYNTHESIS AND CONCLUSION

Available crop germplasm is a valuable resource for crop genetic study. Although the genetic diversity of domesticated crops has been significantly reduced compared to that of their wild progenitors, a relatively high level of genetic variation in modern crops has been maintained owing to genetic drift and introgressions between or among the domesticated crops and their close wild relatives (**Table 2**). Such diverse crop germplasm provides resources that facilitate modern breeding and crop genomic and population genetic studies. Enhancing crop improvement in the future will require continuing to collect diverse crops worldwide, constructing populations of crosses between or among wild species and cultivated crop species, and identifying the morphologies and genetic variation of natural-variant crop plants.

Exploiting the origins of cultivated crops and crop domestication histories will help us understand how breeding has influenced the genetic diversity of modern crops. The studies reviewed here demonstrate that using genome sequencing to create a comprehensive map of genome variations among cultivated species and closely related wild species will be a powerful approach for crop domestication research. This map would provide detailed information on genomic patterns to reveal the direction of introgression of the targeted genetic loci within or between cultivars and wild species.

It is now feasible to comprehensively reveal genetic variation for traits among crop species through enhanced genomic approaches. Genotyping via low-pass multiplex sequencing is a powerful way to profile genomic patterns among diverse crops. Constructing haplotype map data sets for all major crops as completely as possible is a key to genomics-assisted crop breeding. We can predict that more and more such data sets will be available to facilitate genome-wide association mapping of important agronomic traits in crops. We also believe that genetic mapping of complex traits will be improved by developing new statistical models and integrating more data sets from diverse crops. The enhanced genomic approaches will continue to reveal variability in genetic architectures among traits and among crop species. Moreover, revealing the genetic basis of heterosis in crops through an approach that integrates genomics and population genetics will be crucial for crop improvement and accelerate genomics-assisted crop breeding.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We acknowledge the assistance of Yan Zhao in preparing the figures and an anonymous reviewer for helpful suggestions. We apologize to any authors whose work may not have been cited owing to length restrictions. Research in the authors' laboratories is supported by grants from the Ministry of Science and Technology of China (2012AA10A302, 2012AA10A304, and 2013CBA01404) to B.H. and X.H., the Ministry of Agriculture of China (2013ZX08009-002 and 2013ZX08001-004) to B.H., the National Natural Science Foundation of China (31322028) to X.H., and the Science and Technology Commission of Shanghai Municipality (13QA1403900) to X.H.

LITERATURE CITED

1. Abe A, Kosugi S, Yoshida K, Natsume S, Takagi H, et al. 2012. Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* 30:174–78
2. Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. *Science* 322:881–88
3. Ashikari M, Matsuoka M. 2006. Identification, isolation and pyramiding of quantitative trait loci for rice breeding. *Trends Plant Sci.* 11:344–50
4. Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, et al. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465:627–31
5. Austin RS, Vidaurre D, Stamatiou G, Breit R, Provart NJ, et al. 2011. Next-generation mapping of *Arabidopsis* genes. *Plant J.* 67:715–25
6. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3:e3376
7. Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, et al. 2012. Reference genome sequence of the model plant *Setaria*. *Nat. Biotechnol.* 30:555–61
8. Bevan MW, Uauy C. 2013. Genomics reveals new landscapes for crop improvement. *Genome Biol.* 14:206
9. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633–35
10. Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, et al. 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 491:705–10
11. Browning BL, Browning SR. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84:210–23
12. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, et al. 2009. The genetic architecture of maize flowering time. *Science* 325:714–18
13. Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. *Genome Res.* 20:393–402
14. Chen J, Huang Q, Gao D, Wang J, Lang Y, et al. 2013. Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. *Nat. Commun.* 4:1595
15. Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, et al. 2012. Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* 44:803–7
16. Chodavarapu RK, Feng S, Ding B, Simon SA, Lopez D, et al. 2012. Transcriptome and methylome interactions in rice hybrids. *Proc. Natl. Acad. Sci. USA* 109:12040–45
17. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, et al. 2007. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* 317:338–42
18. Doebley JF, Gaut BS, Smith BD. 2006. The molecular genetics of crop domestication. *Cell* 127:1309–21
19. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379
20. Endelman JB, Jannink JL. 2012. Shrinkage estimation of the realized relationship matrix. *G3* 2:1405–13
21. Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH. 2004. An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. *Genome Res.* 14:1812–19
22. Feng Q, Zhang Y, Hao P, Wang S, Fu G, et al. 2002. Sequence and analysis of rice chromosome 4. *Nature* 420:316–20

23. Fernie AR, Schauer N. 2009. Metabolomics-assisted breeding: a viable option for crop improvement? *Trends Genet.* 25:39–48
24. Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K. 2011. Crop genome sequencing: lessons and rationales. *Trends Plant Sci.* 16:77–88
25. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, et al. 2012. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 477:419–23
26. Ganal MW, Durstewitz G, Polley A, Berard A, Buckler ES, et al. 2011. A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS ONE* 6:e28334
27. Godfray HC, Beddington JR, Crute IR, Haddad L, Lawrence D, et al. 2010. Food security: the challenge of feeding 9 billion people. *Science* 327:812–18
28. Goff SA, Ricke D, Lan TH, Presting G, Wang R, et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296:92–100
29. Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, et al. 2009. A first-generation haplotype map of maize. *Science* 326:1115–17
30. Hamblin MT, Buckler ES, Jannink JL. 2011. Population genetics of genomics-based crop improvement methods. *Trends Genet.* 27:98–106
31. Harper AL, Trick M, Higgins J, Fraser F, Clissold L, et al. 2012. Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. *Nat. Biotechnol.* 30:798–802
32. Harushima Y, Yano M, Shomura A, Sato M, Shimano T, et al. 1998. A high-density rice genetic linkage map with 2275 markers using a single F_2 population. *Genetics* 148:479–94
33. He G, Zhu X, Elling AA, Chen L, Wang X, et al. 2010. Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. *Plant Cell* 22:17–33
34. He Z, Zhai W, Wen H, Tang T, Wang Y, et al. 2011. Two evolutionary histories in the genome of rice: the roles of domestication genes. *PLoS Genet.* 7:9
35. Hua J, Xing Y, Wu W, Xu C, Sun X, 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. USA* 100:2574–79
36. Huang N, Angeles ER, Domingo J, Magpantay G, Singh S, et al. 1997. Pyramiding of bacterial blight resistance genes in rice: marker-assisted selection using RFLP and PCR. *Theor. Appl. Genet.* 95:313–20
37. Huang S, Li R, Zhang Z, Li L, Gu X, et al. 2009. The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* 41:1275–81
38. Huang X, Feng Q, Qian Q, Zhao Q, Wang L, et al. 2009. High-throughput genotyping by whole-genome resequencing. *Genome Res.* 19:1068–76
39. Huang X, Kurata N, Wei X, Wang ZX, Wang A, et al. 2012. A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490:497–501
40. Huang X, Paulo MJ, Boer M, Effgen S, Keizer P, et al. 2011. Analysis of natural allelic variation in *Arabidopsis* using a multiparent recombinant inbred line population. *Proc. Natl. Acad. Sci. USA* 108:4488–93
41. Huang X, Wei X, Sang T, Zhao Q, Feng Q, et al. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42:961–67
42. Huang X, Zhao Y, Wei X, Li C, Wang A, et al. 2012. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* 44:32–39
43. Hufford MB, Xu X, van Heerwaarden J, Pyhajarvi T, Chia JM, et al. 2012. Comparative population genomics of maize domestication and improvement. *Nat. Genet.* 44:808–11
44. Int. HapMap Consort. 2005. A haplotype map of the human genome. *Nature* 437:1299–320
45. Int. HapMap Consort. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–61
46. Int. Rice Genome Seq. Proj. 2005. The map-based sequence of the rice genome. *Nature* 436:793–800
47. Jacquemin J, Bhatia D, Singh K, Wing RA. 2013. The International *Oryza* Map Alignment Project: development of a genus-wide comparative genomics platform to help solve the 9 billion-people question. *Curr. Opin. Plant Biol.* 16:147–56

48. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–67
49. James G-V, Patel V, Nordström K-JV, Klasen J-R, Salomé P-A, et al. 2013. User guide for mapping-by-sequencing in *Arabidopsis*. *Genome Biol.* 14:R61
50. Jia G, Huang X, Zhi H, Zhao Y, Zhao Q, et al. 2013. A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). *Nat. Genet.* 45:957–61
51. Jia J, Zhao S, Kong X, Li Y, Zhao G, et al. 2013. *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* 496:91–95
52. Jiao Y, Zhao H, Ren L, Song W, Zeng B, et al. 2012. Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* 44:812–15
53. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, et al. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42:348–54
54. Keurentjes J, Fu J, Terpstra IR, Garcia JM, Ackerveken G, et al. 2007. Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proc. Natl. Acad. Sci. USA* 104:1708–13
55. Korte A, Farlow A. 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29
56. Korte A, Vilhjalmsson BJ, Segura V, Platt A, Long Q, Nordborg M. 2012. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* 44:1066–71
57. Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, et al. 2009. A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet.* 5:e1000551
58. Kump KL, Bradbury PJ, Wisser RJ, Buckler ES, Belcher AR, et al. 2011. Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* 43:163–68
59. Lai J, Li R, Xu X, Jin W, Xu M, et al. 2010. Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.* 42:1027–30
60. Lam HM, Xu X, Liu X, Chen W, Yang G, et al. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* 42:1053–59
61. Li H, Peng Z, Yang X, Wang W, Fu J, et al. 2013. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.* 45:43–50
62. Li L, Lu K, Chen Z, Mu T, Hu Z, Li X. 2008. Dominance, overdominance and epistasis condition the heterosis in two heterotic rice hybrids. *Genetics* 180:1725–42
63. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. 2011. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* 21:940–51
64. Lin Z, Li X, Shannon LM, Yeh CT, Wang ML, et al. 2012. Parallel domestication of the *Shattering1* genes in cereals. *Nat. Genet.* 44:720–24
65. Ling HQ, Zhao S, Liu D, Wang J, Sun H, et al. 2013. Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* 496:87–90
66. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, et al. 2012. GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28:2397–99
67. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. FaST linear mixed models for genome-wide association studies. *Nat. Methods* 8:833–35
68. Majewski J, Pastinen T. 2011. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.* 27:72–79
69. Mascher M, Richmond TA, Gerhardt DJ, Himmelbach A, Clissold L, et al. 2013. Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. *Plant J.* 76:494–505
70. Mayer KF, Waugh R, Brown JW, Schulman A, Langridge P, et al. 2012. A physical, genetic and functional sequence assembly of the barley genome. *Nature* 491:711–16
71. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, et al. 2009. Genetic properties of the maize nested association mapping population. *Science* 325:737–40
72. McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, et al. 2009. Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc. Natl. Acad. Sci. USA* 106:12273–78

73. Meuwissen TH, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–29
74. Miura K, Ashikari M, Matsuoka M. 2011. The role of QTLs in the breeding of high-yielding rice. *Trends Plant Sci.* 16:319–26
75. Montes JM, Melchinger AE, Reif JC. 2007. Novel throughput phenotyping platforms in plant genetic studies. *Trends Plant Sci.* 12:433–36
76. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, et al. 2012. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. USA* 110:453–58
77. Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, et al. 2009. Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21:2194–202
78. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461:272–76
79. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res.* 15:1566–75
80. Olsen KM, Wendel JF. 2013. A bountiful harvest: genomic insights into crop domestication phenotypes. *Annu. Rev. Plant Biol.* 64:47–70
81. Pasam RK, Sharma R, Malosetti M, van Eeuwijk FA, Haseneyer G, et al. 2012. Genome-wide association studies for agronomical traits in a world wide spring barley collection. *BMC Plant Biol.* 12:16
82. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, et al. 2012. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* 44:631–35
83. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457:551–56
84. Peter BM, Huerta-Sanchez E, Nielsen R. 2012. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS Genet.* 8:e1003011
85. Platt A, Vilhjalmsson BJ, Nordborg M. 2010. Conditions under which genome-wide association studies will be positively misleading. *Genetics* 186:1045–52
86. Poland JA, Bradbury PJ, Buckler ES, Nelson RJ. 2011. Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proc. Natl. Acad. Sci. USA* 108:6893–98
87. Potato Genome Seq. Consort. 2011. Genome sequence and analysis of the tuber crop potato. *Nature* 475:189–95
88. Rafalski JA. 2010. Association genetics in crop improvement. *Curr. Opin. Plant Biol.* 13:174–80
89. Rice Chromosome 10 Seq. Consort. 2003. In-depth view of structure, activity, and evolution of rice chromosome 10. *Science* 300:1566–69
90. Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, et al. 2012. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat. Genet.* 44:217–20
91. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, et al. 2013. Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.* 14:R55
92. Saitoh K, Onishi K, Mikami I, Thidar K, Sano Y. 2004. Allelic diversification at the *C* (*OsC1*) locus of wild and cultivated rice: nucleotide changes associated with phenotypes. *Genetics* 168:997–1007
93. Sasaki A, Ashikari M, Ueguchi-Tanaka M, Itoh H, Nishimura A, et al. 2002. Green revolution: a mutant gibberellin-synthesis gene in rice. *Nature* 416:701–2
94. Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, et al. 2002. The genome sequence and structure of rice chromosome 1. *Nature* 420:312–16
95. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, et al. 2013. Patterns of population epigenomic diversity. *Nature* 495:193–98
96. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, et al. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463:178–83
97. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112–15
98. Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, et al. 2009. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods* 6:550–51
99. Schneeberger K, Weigel D. 2011. Fast-forward genetics enabled by new sequencing technologies. *Trends Plant Sci.* 16:282–88

100. Segura V, Vilhjalmsson BJ, Platt A, Korte A, Seren U, et al. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44:825–30
101. Seren U, Vilhjalmsson BJ, Horton MW, Meng D, Forai P, et al. 2012. GWAPP: a web application for genome-wide association mapping in *Arabidopsis*. *Plant Cell* 24:4793–805
102. Singer T, Fan Y, Chang HS, Zhu T, Hazen SP, Briggs SP. 2006. A high-resolution map of *Arabidopsis* recombinant inbred lines by whole-genome exon array hybridization. *PLoS Genet.* 2:e144
103. Song GS, Zhai HL, Peng YG, Zhang L, Wei G, et al. 2010. Comparative transcriptional profiling and preliminary study on heterosis mechanism of super-hybrid rice. *Mol. Plant* 3:1012–25
104. Swanson-Wagner RA, Jia Y, DeCook R, Borsuk LA, Nettleton D, Schnable PS. 2006. All possible modes of gene action are observed in a global comparison of gene expression in a maize F_1 hybrid and its inbred parents. *Proc. Natl. Acad. Sci. USA* 103:6805–10
105. Takagi H, Uemura A, Yaegashi H, Tamiru M, Abe A, et al. 2013. MutMap-Gap: whole-genome resequencing of mutant F2 progeny bulk combined with *de novo* assembly of gap regions identifies the rice blast resistance gene *Pii*. *New Phytol.* 200:276–83
106. Takeda S, Matsuoka M. 2008. Genetic approaches to crop improvement: responding to environmental and population changes. *Nat. Rev. Genet.* 9:444–57
107. Tang H, Cuevas HE, Das S, Sezen UU, Zhou C, et al. 2013. Seed shattering in a wild sorghum is conferred by a locus unrelated to domestication. *Proc. Natl. Acad. Sci. USA* 110:15824–29
108. Tester M, Langridge P. 2010. Breeding technologies to increase crop production in a changing world. *Science* 327:818–22
109. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, et al. 2011. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* 43:159–62
110. Tomato Genome Consor. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485:635–41
111. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, et al. 2010. The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nat. Genet.* 42:833–39
112. Vilhjalmsson BJ, Nordborg M. 2012. The nature of confounding in genome-wide association studies. *Nat. Rev. Genet.* 14:1–2
113. Wang H, Chattopadhyay A, Li Z, Daines B, Li Y, et al. 2010. Rapid identification of heterozygous mutations in *Drosophila melanogaster* using genomic capture sequencing. *Genome Res.* 20:981–88
114. Wang L, Wang A, Huang X, Zhao Q, Dong G, et al. 2011. Mapping 49 quantitative trait loci at high resolution through sequencing-based genotyping of rice recombinant inbred lines. *Theor. Appl. Genet.* 122:327–40
115. Wang M, Jiang N, Jia T, Leach L, Cockram J, et al. 2011. Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. *Theor. Appl. Genet.* 124:233–46
116. Wang X, Wang H, Wang J, Sun R, Wu J, et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43:1035–39
117. Wang ZY, Zheng FQ, Shen GZ, Gao JP, Snustad DP, et al. 1995. The amylose content in rice endosperm is related to the post-transcriptional regulation of the waxy gene. *Plant J.* 7:613–22
118. Wellcome Trust Case Control Consort. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–78
119. Wenzl P, Carling J, Kudrna K, Jaccoud D, Huttner E, et al. 2004. Diversity Arrays Technology (DArT) for whole-genome profiling of barley. *Proc. Natl. Acad. Sci. USA* 101:9915–20
120. Wing RA, Ammiraju JS, Luo M, Kim H, Yu Y, et al. 2005. The *Oryza* Map Alignment Project: the golden path to unlocking the genetic potential of wild rice species. *Plant Mol. Biol.* 59:53–62
121. Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, et al. 1998. Direct allelic variation scanning of the yeast genome. *Science* 281:1194–97
122. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. 2013. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* 14:507–15
123. Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, et al. 2005. The effects of artificial selection on the maize genome. *Science* 308:1310–14

124. Xie W, Feng Q, Yu H, Huang X, Zhao Q, et al. 2010. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc. Natl. Acad. Sci. USA* 107:10578–83
125. Xing Y, Zhang Q. 2010. Genetic and molecular bases of rice yield. *Annu. Rev. Plant Biol.* 61:421–42
126. Xu J, Zhao Q, Du P, Xu C, Wang B, et al. 2010. Developing high throughput genotyped chromosome segment substitution lines based on population whole-genome re-sequencing in rice (*Oryza sativa* L.). *BMC Genomics* 11:656
127. Xu X, Liu X, Ge S, Jensen JD, Hu F, et al. 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* 30:105–11
128. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42:565–69
129. Yu H, Xie W, Wang J, Xing Y, Xu C, et al. 2011. Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. *PLoS ONE* 6:e17595
130. Yu J, Hu S, Wang J, Wong GK, Li S, et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. *ssp. indica*). *Science* 296:79–92
131. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38:203–8
132. Zhang G, Liu X, Quan Z, Cheng S, Xu X, et al. 2012. Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nat. Biotechnol.* 30:549–54
133. Zhang Q. 2007. Strategies for developing Green Super Rice. *Proc. Natl. Acad. Sci. USA* 104:16402–9
134. Zhang Z, Buckler ES, Casstevens TM, Bradbury PJ. 2009. Software engineering the mixed model for genome-wide association studies on large samples. *Brief. Bioinforma.* 10:664–75
135. Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, et al. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42:355–60
136. Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, et al. 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* 2:467
137. Zhao Q, Huang X, Lin Z, Han B. 2010. SEG-Map: a novel software for genotype calling and genetic map construction from next-generation sequencing. *Rice* 3:98–102
138. Zhou G, Chen Y, Yao W, Zhang C, Xie W, et al. 2012. Genetic composition of yield heterosis in an elite rice hybrid. *Proc. Natl. Acad. Sci. USA* 109:15847–52
139. Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44:821–24
140. Zong G, Wang A, Wang L, Liang G, Gu M, et al. 2012. A pyramid breeding of eight grain-yield related quantitative trait loci based on marker-assistant and phenotype selection in rice (*Oryza sativa* L.). *J. Genet. Genomics* 39:335–50
141. Zou G, Zhai G, Feng Q, Yan S, Wang A, et al. 2012. Identification of QTLs for eight agronomically important traits using an ultra-high-density map based on SNPs generated from high-throughput sequencing in sorghum under contrasting photoperiods. *J. Exp. Bot.* 63:5451–62