# Crop genomics: advances and applications

*Peter L. Morrell[1], Edward S. Buckler[2] and Jeffrey Ross-Ibarra[3]*

Abstract | The completion of reference genome sequences for many important crops and the ability to perform high-throughput resequencing are providing opportunities for improving our understanding of the history of plant domestication and to accelerate crop improvement. Crop plant comparative genomics is being transformed by these data and a new generation of experimental and computational approaches. The future of crop improvement will be centred on comparisons of individual plant genomes, and some of the best opportunities may lie in using combinations of new genetic mapping strategies and evolutionary analyses to direct and optimize the discovery and use of genetic variation. Here we review such strategies and insights that are emerging.

Genome-wide association studies
(GWASs). Studies that search for a statistical association between a phenotype and a particular allele by screening loci (most commonly by genotyping SNPs) across the entire genome.

[1]*Department of Agronomy and Plant Genetics, University of Minnesota, St Paul, Minnesota, 55108.*
[2]*US Department of Agriculture–Agriculture Research Service (USDA–ARS) and Institute for Genomic Diversity and the Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York 14853, USA.*
[3]*Department of Plant Sciences and the Genome Center, University of California Davis, California, 95616, USA.*
*Correspondence to P.L.M. and J.R.-I.*
*e-mail: pmorrell@umn.edu; rossibarra@ucdavis.edu*
doi:10.1038/nrg3097
Published online 29 December 2011

The completion of reference genome sequences for many important crops and model plants has the potential to aid in the realization of the long-standing promise of plant genomics to dramatically accelerate crop improvement[1]. Since the late 1960s, it has been possible to survey molecular markers across a plant genome[2], but for decades the number of markers that could be readily assayed placed limits on the genetic resolution that could be achieved using either experimental or comparative genetic approaches. Only a few years ago, the highest-density genetic maps required the laborious assay of several thousand markers (for example, REF. 3). Experimental populations were generally limited to simple crosses between two parents; more elaborate study designs that might provide an assessment of the genomic distribution of agronomically important mutations and their frequency in the relevant germplasm were proscribed by limits on marker technologies and the analytical approaches that could be used to distinguish the contribution of multiple parents. Comparative approaches for the identification of functionally important mutations based on analysis of marker frequency among populations had also been proposed[4], but the high variance in expected allele frequency between populations[5] made the discovery of functionally important variants among the high number of loci surveyed highly improbable.

A reference genome is now available for a number of crops (FIG. 1), and progress is being made towards references for crops with large genomes[6] (for example, see links in Further information). In addition, reference genomes have been published for a number of other model plant systems, including *Arabidopsis thaliana* and *Brachypodium distachyon*[7,8]. Comparative genomics — which is traditionally thought of as the analysis of synteny (gene order) and sequence comparisons among related species — is now being redefined by the rapid publication of increasing numbers of reference genomes, by estimation of sequence diversity from high-throughput resequencing, by the examination of the genomic distribution of large insertions and deletions (indels) and copy number variants (CNVs) and by the emergence of a new generation of experimental and computational approaches. From genetic mapping to evolutionary analysis, the future of crop improvement will revolve around the comparisons of individual plant genomes. Maximizing the use of this genomic data for crop improvement is of fundamental importance if we are to continue increasing crop production in the face of growing human populations and changing climates while minimizing the environmental impact of agricultural activity.

In this Review, we begin by addressing the challenges for comparative crop genomics that are posed by the complex organization of plant genomes and the high levels of nucleotide and structural diversity that are found in many crop species. We then discuss the importance of understanding domestication, as the origin and demography of a crop affect the genetic basis of agronomic traits and influence patterns of nucleotide diversity genome-wide. We examine the ways in which our understanding of the genetics of agronomic traits is being fundamentally reshaped by genomic data. High-density genetic markers are being used in genome-wide association studies (GWASs) and can also be exploited for genomic selection.

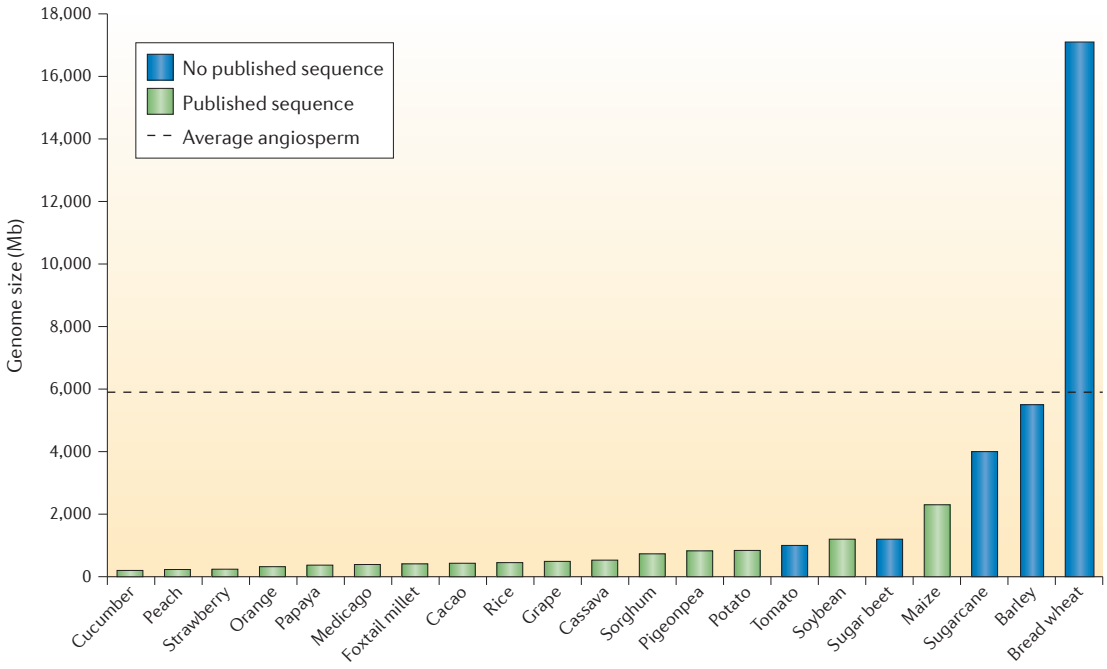



Figure 1 | **Crop genome size.** Genome size of all published crop genomes (shown in green) and the five most important production crops with unpublished genome sequences (shown in blue). The average angiosperm genome size of ~6 Gb is shown by the dotted line for comparison.

Understanding of agronomic traits is also being improved by a new generation of multiparent genetic mapping populations (or next-generation populations). As we discuss, higher-throughput resequencing and marker genotyping will also enable new approaches towards crop improvement, such as the identification and selective elimination of deleterious mutations.

## Challenges of plant genomes

The genomic tools that are applied to plants are often developed for and tested against data from humans or other model systems, such as fruitflies or mice[9,10], but the size and dynamic nature of plant genomes adds to or exacerbates challenges that are faced in other systems (FIG. 1). Plants tend to have a larger number of multigene families[11] and a higher frequency of polyploidy than occurs in mammals. This makes paralogy a more substantive issue because the short sequence reads that are typical of high-throughput sequencing may not map uniquely to a reference genome, and allelic variation cannot then be distinguished from differences among closely related gene family members (FIG. 2). Paralogy remains a problem even in plant species that have a high-quality reference genome owing to the prevalence of extensive copy number variation[12,13]. For instance, estimates suggest that the maize reference genome accounts for only ~70% of the low-copy-number sequences that are present in the parents of a diverse set of maize inbreds and that this copy number variation leads to a high percentage of false-positive variants[14]. It seems likely that continued improvement in sequence read length, along with methodological approaches that assess allelic segregation among lines[14] and that make use of local patterns of linkage disequilibrium (LD), will be useful for identifying paralogous reads in complex crop genomes. Although there may be no simple solution to the complexity of polyploid genomes, sequencing diploid relatives[15,16] or double haploid lines[17] can provide a baseline for future genome-level research in polyploid crops.

The high levels of nucleotide diversity in some crop genomes pose a challenge for comparative analyses, as higher numbers of mismatches between a sample and a reference will result in reduced sequence read mapping (FIG. 2) or reduced hybridization to oligonucleotide arrays. For example, the maize and human genomes are similar in size, but an average pair of maize individuals differs at tenfold more sites than any two humans do[18]. Although many crops do not have high levels of diversity, the difficulties of a diverse genome are not unique to maize as an outcrossing species: diversity is also high in the clonally propagated grape[19] and even in self-fertilizing ('selfing') species, such as barley[20].

Another challenge in plant comparative genomics is genome size (FIG. 1). Plant genome size varies by more than three orders of magnitude in currently characterized species[21], largely owing to the prevalence of transposable elements[22]. Size alone makes genomic analysis more difficult: shotgun sequencing reads that are sufficient to provide deep (25×) coverage of four *Drosophila melanogaster* genomes — enabling the identification of heterozygous sites and structural variation — would provide a meagre ~1× coverage of the wheat genome. The density of transposable elements in plant genomes also means that a large fraction of shotgun sequencing data is of limited use for reference-based genomic analysis, as

Paralogy
Unlike orthologous genes, which trace their common origin to a locus in an ancestral species, paralogous loci consist of gene copies that trace their common origin to a duplication event within a genome.

Linkage disequilibrium
(LD). Nonrandom association of alleles at two or more loci. The pattern and extent of LD in a genomic region is affected by mutation, recombination, genetic drift, natural selection and demographic history.

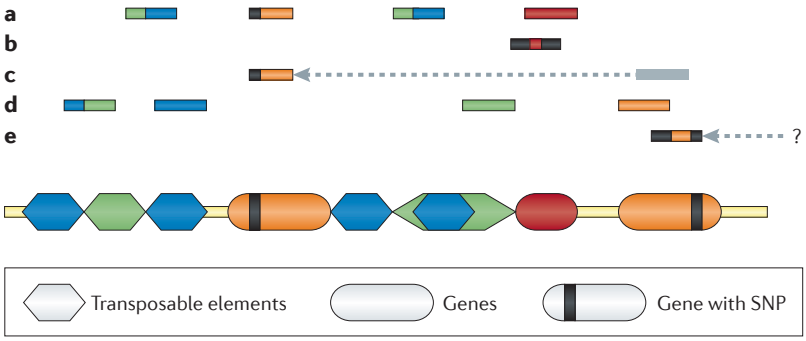



Figure 2 | **Challenges of read mapping in plant genomes.** The mapping of short sequence reads to a reference plant genome is shown with the genome at the bottom and with sequencing reads above. Coloured shapes represent transposable elements or genes; the two orange ovals represent a pair of paralogous genes. Short sequence reads are shown directly above where they would map to the reference. Different scenarios are shown in lines **a**–**e**. **a** | Uniquely mapping reads, including junctions between sequence repeats. **b** | A sequence from a diverse genome that would fail to map to the reference owing to an excess of SNP differences. **c** | A read from one paralogue that maps incorrectly owing to a sequence error or a SNP. The correct mapping is shown with a grey read. **d** | Reads that would map multiply and are usually filtered from further analysis. **e** | A read from a third copy of the orange gene that is incorrectly mapped to one of the reference copies, leading to a false SNP. This is likely to be the result of a copy number variant that was not included in the reference genome (as indicated by the question mark).

reads map with equal probability to multiple positions in the reference (FIG. 2). It is not surprising that the crop genomes that have been sequenced to date have all been relatively small — the largest crop genome sequenced, maize, is less than half the size of the average angiosperm genome (FIG. 1; TABLE 1).

Although plant genomes pose a number of challenges for genomic analysis, they do offer some advantages. Unlike most animals, crops can be propagated clonally or maintained as inbred lines, and the seeds of many species can be stored indefinitely, which effectively immortalizes genotypes of interest. This makes it possible to sequence a line once but to phenotype the line many times, and it allows replication across environments[23]. Inbred lines or specially created double haploids also avoid the difficulties of sequencing highly heterozygous genomes. Sequencing of the grape genome has provided a useful comparison of the advantages and difficulties of sequencing a diploid outcrossing accession[13] or an inbred line[24].

## Origin and evolution of crops

Understanding the origins and domestication of crop plants is of substantial evolutionary interest, as domesticated plants provide a model system for studying adaptation[25,26]. An understanding of crop origins has long been held as central to the identification of useful genetic resources for crop improvement[27]. Domestication shapes the genetic variation that is available to modern breeders as it influences levels of nucleotide diversity and patterns of LD genome-wide. The demographic history of domestication also informs our expectations of the genetic architecture of traits and thus our ability to identify causal genetic variants for crop improvement.

***Demographic history and geographic origins.*** Genome-wide polymorphisms make it possible to examine the demographic history and geographic origins of crops. Domestication is an evolutionarily recent phenomenon, and most of the genealogical history at any locus will be shared between a domesticate and its wild progenitor[28]. Comparisons of alleles within and between domesticated and wild taxa will reveal divergence times that greatly predate the origin of the cultivated form[29,30], reflecting the time to most recent common ancestor of the species rather than the time of divergence of the domesticate. A detailed understanding of domestication history requires a large number of loci in conjunction with modelling of population demography. Some of the earliest work on demographic modelling in plants used mean patterns of genetic diversity to fit a bottleneck model of domestication[31], an approach that was later extended to include an explicit likelihood framework[32,33]. More recently, investigators have used methods that incorporate more detailed information, such as the site frequency spectrum[34,35], to distinguish among different evolutionary models.

One of the most fundamental issues that influences the genetic architecture of agronomic traits and the levels of genetic diversity in crop genomes is the number of times that a species has been domesticated. There are compelling examples for both single domestications (such as maize and soybeans)[36,37] and multiple domestications (such as avocados, common beans and barley)[38–40], but the number and location of domestication events for most crops remain unresolved. Simple statistical methods that cluster individuals or populations based on genetic diversity within the domesticate can be misleading, as the number of genetic groupings is not necessarily reflective of domestication history[41,42]. For example, although genetic evidence suggests two domestications of the common bean[39], genetic drift in cultivated populations leads to the identification of multiple genetic groups[43].

The details of even the simplest of domestication scenarios are likely to be complex. For example, geographical spread of the domesticate followed by admixture with wild relatives can obscure geographic origins[44,45]. Extensive admixture may be one explanation for the continued controversy regarding the origins of the domesticated *indica* and *japonica* subspecies of rice. Analyses from recent genome-wide resequencing have failed to reach a consensus on the number of domestications of rice: modelling of genetic differentiation supports separate domestications followed by introgression at agronomically important loci[46], whereas the site frequency spectrum and phylogenetic analysis of multiple data sets argue for a single origin[35]. As whole-genome data become available for more crops and their wild relatives, application of methods that make better use of additional information from detailed haplotype structure and patterns of admixture across the genome (for example, REFS 47,48) will improve insight into the complex demographic histories of many crops.

**Bottleneck**
A temporary marked reduction in population size.

**Site frequency spectrum**
The distribution of allele frequencies in a population: essentially a count of the number of alleles in a population at a given frequency.

**Genetic drift**
Fluctuations in allele frequencies that are due to the effects of random sampling.

**Admixture**
The mixing of two or more genetically differentiated populations.

**Introgression**
The incorporation of genetic material from one population or species into another by hybridization and backcrossing.

**Haplotype**
The combination of alleles or genetic markers found on a single chromosome of an individual.

Table 1 | **Crop genome characteristics**

| Crop | Genome size (Mb) | Gene number | Transposable element content (%) | Refs |
|---|---|---|---|---|
| Cucumber | 200 | 21,000 | | Phytozome |
| Peach | 230 | 28,000 | | Phytozome |
| Strawberry* | 240 | 35,000 | 22 | 16 |
| Orange | 320 | 25,000 | | Phytozome |
| Papaya | 370 | 29,000 | 52 | 149 |
| Medicago* | 375 | 48,000 | 30 | 15 |
| Foxtail millet | 410 | 35,000 | | Phytozome |
| Cacao | 430 | 29,000 | 24 | 150 |
| Rice | 450 | 41,000 | 25 | 151 |
| Grape | 490 | 30,000 | 41 | 24 |
| Cassava | 530 | 31,000 | | Phytozome |
| Sorghum | 730 | 28,000 | 63 | 109 |
| Pigeonpea | 833 | 49,000 | 52 | 152 |
| Potato | 840 | 39,000 | 62 | 17 |
| Tomato | 1,000 | | | Kew |
| Soybean | 1,200 | 46,000 | 59 | 153 |
| Sugar beet | 1,200 | | | Kew |
| Maize | 2,300 | 33,000 | 85 | 154 |
| Sugarcane | 4,000 | | | Kew |
| Barley | 5,500 | | | Kew |
| Bread wheat | 17,100 | | | Kew |
| Average angiosperm | 5,900 | | | Kew |

The table shows the genome size, gene and transposable element content for the world's ten top production crops and all other crops with sequenced genomes. In the 'Refs' column, 'Phytozome' refers to http://www.phytozome.net and 'Kew' refers to http://data.kew.org/cvalues. *Sequence of a species related to the main crop.

## The genomic basis of domestication

Plant and animal domestication motivated some of the earliest thought and research on evolution and natural selection[49], and the past several decades have seen widespread application of molecular markers to the study of plant domestication and improvement. Until recently, however, our understanding of the genetics of crop evolution has been limited by a reliance on genetic mapping methods that require *a priori* identification of a phenotype of interest. Studies have primarily focused on a suite of traits that make up the 'domestication syndrome'[50], including decreased dispersal, reduced branching, loss of seed dormancy, reduced natural defences and increased size of certain morphological features. Mapping strategies cannot uncover loci that are responsible for phenotypes that have not been measured, regardless of the evolutionary importance of the phenotypes, and the view that domestication involves few loci[51] is probably a result of this limitation. Population genetic approaches, however, have the potential to identify loci that are subject to selection even without a known phenotype[25]. These methods also permit the detection of alleles at extreme frequencies that are difficult to detect by genetic mapping.

One emerging result from population genetic investigation is that crop evolution has probably involved many loci. Some of the first population-genomics analyses of domestication[33,52,53] identified dozens of loci involved in a variety of functions, presumably representing numerous phenotypes that were not amenable to study using mapping approaches. Wright *et al.*[33] used demographic modelling to identify loci that had lost more diversity than would be expected owing to population bottlenecks alone. They estimated that 2–4% of loci in the maize genome — as many as 1,300 genes — were involved in either domestication or subsequent improvement. Whole-genome resequencing studies in a number of crops continue to add to the list of putatively selected loci[14,46,54]. By contrast, in spite of the prevalence of CNVs in cultivated and wild genomes (for example, REFS 30,55), there is little evidence to date to support a prominent role for CNVs in domestication[56].

The involvement of a large number of loci in domestication is predicted both by the complex nature of many domestication traits and by models that incorporate the effects of unconscious selection[49]. Flowering time adaptation, for example, may be controlled by many genes[57], and adaptation to changes in soil nutrients[58] or loss of seed dormancy[59] are examples of traits that are unlikely to have been consciously selected by early farmers.

Although molecular evidence of selection at a locus is important for confirmation that the locus was involved in adaptation, it is important to consider that many agronomic traits reflect a complex evolutionary history. Introgression, for example, can lead to unusual histories for individual loci. In barley, in which many traits show evidence of multiple origins — a finding that is consistent with the domestication history of this species[60,61] — allelic variation at a flowering time locus in European cultivars appears to have arisen by introgression from barley that was independently domesticated in Central Asia[62]. Whereas early population genetic inference focused on identification of selective sweeps, there is now considerable evidence that selection does not always follow this simple model. Selection histories can include incomplete sweeps, local adaptation, multiple mutational origins and adaptation from standing variation (for example, REFS 63–65). The interplay of these factors is complex, and identifying loci of adaptive interest[66] or distinguishing evolutionarily important phenomena, such as local adaptation from multiple mutational origins[67], will require careful analysis of genome-wide data from a geographically diverse set of samples.

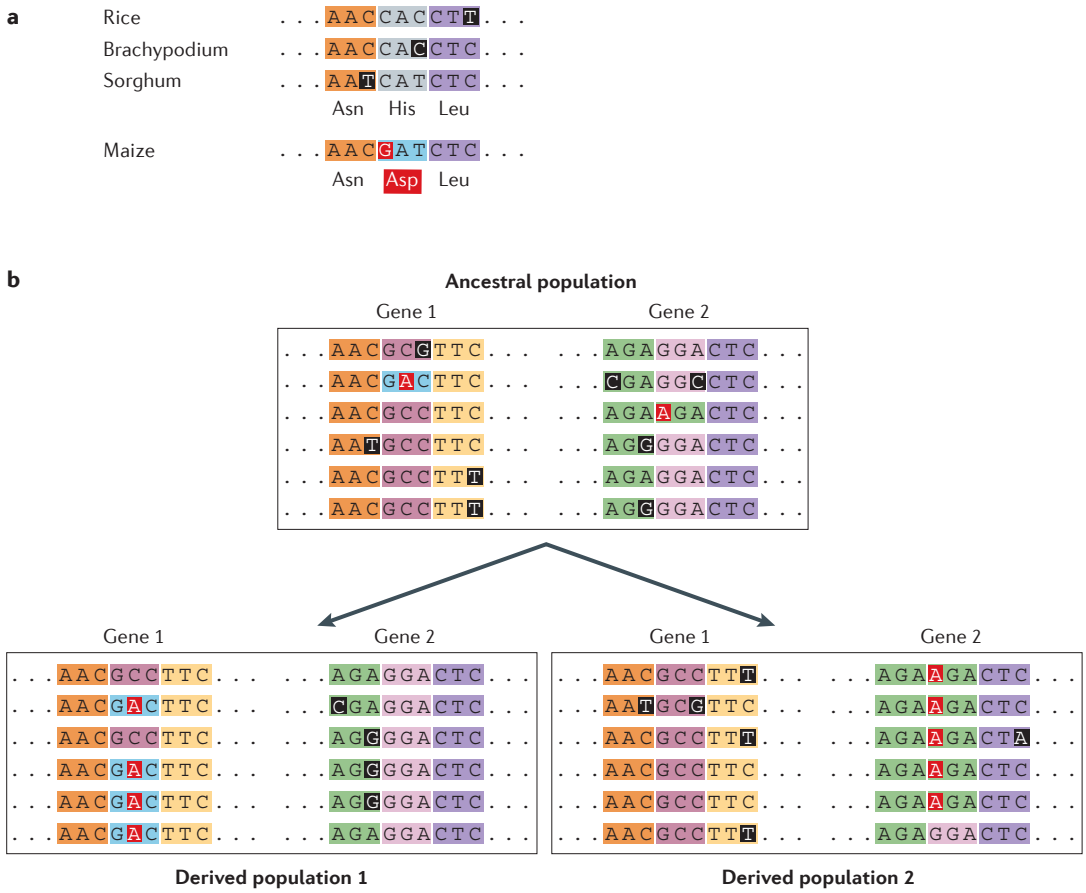
Finally, although comparative genetic mapping studies between species suggested some similarity in the genetic basis of domestication syndrome traits[68], it is perhaps too early to determine whether there are strong commonalities among genes that have been selected during crop evolution. Even so, selected loci in both maize and sunflowers appear to be enriched for functions related to amino acid biosynthesis[33,53]. Given that strong selection may increase the frequency of linked deleterious mutations (BOX 1), it is possible that selection on amino acid biosynthesis during domestication contributes to observations that genes involved in protein metabolism have a role in heterosis[69].

Selective sweeps
Increases in frequency of an allele and closely linked chromosomal segments that are due to positive selection. Sweeps initially reduce variation and subsequently lead to a local excess of rare alleles as new unique mutations accumulate.

Standing variation
Variation for a locus or trait that is polymorphic in a population.

Heterosis
Otherwise known as 'hybrid vigour', heterosis is the phenomenon whereby progeny of a cross between genetically distinct parents have greater fitness than either of the parental types.

## Box 1 | **Genetic load**




Genetic load refers to the reduction in fitness caused by suboptimal genotypes in a population[121]. Genetic load can arise in a number of ways, including directional selection, recombination or mutation. Mutational load — the presence of deleterious mutations segregating in a population — is of particular interest for crop genomics. Deleterious mutations are most readily detected in protein-coding genes and can take several forms, including premature stop codons, splice site variants or insertions and deletions (indels) that result in the loss or impairment of protein function. These types of mutations are frequently associated with Mendelian disorders in humans, providing direct evidence that loss-of-function changes tend to be deleterious, particularly when homozygous[122]. Although most nonsynonymous mutations in plants are strongly deleterious, a sizable proportion are only slightly so, and these mutations may segregate at appreciable frequencies[123].

Unambiguously deleterious mutations are fairly common in crop genomes[17,54,124]. Statistical analysis of homologous sequence from multiple genomes can identify amino acid changes that are likely to be disadvantageous (for example, REF. 125), but these comparative analyses benefit from transcriptomic data, as transcript variation among individuals may render some putatively deleterious mutations inconsequential[120]. Part **a** of the figure shows a hypothetical alignment of coding sequence from multiple grass species. The conserved nature of the histidine amino acid across species suggests that the nonsynonymous change (indicated by the red 'G') observed in maize is likely to be deleterious. Synonymous changes are shown in black.

Selection against deleterious mutations is hindered by Hill–Robertson effects — because of linkage, selection can only act on the net effect of both beneficial and deleterious mutations. Deleterious mutations should thus be enriched in regions of the genome in which recombination is suppressed and around the targets of strong positive selection[126,127]. Although neither prediction has yet been explicitly demonstrated in crops, patterns of residual heterozygosity in the maize genome support the first prediction[14], and evidence from humans[128] bears out the second. Whereas inbreeding can act to purge deleterious mutations[129,130], drift can increase the frequency of deleterious mutations in small populations[131,132]. Drift is a stochastic process, and unique sets of deleterious alleles would be expected to increase in frequency in different breeding populations (for example, REF. 124). This is illustrated in part **b** of the figure, in which two nonsynonymous mutations (indicated by the red 'A's) in the ancestral population increase in frequency in two derived populations. Because drift operates independently in isolated populations, different breeding programs are likely to have a number of distinct, high-frequency deleterious mutations. Given that most deleterious mutations are at least partially recessive, crosses between lines from different breeding populations should exhibit complementation at these loci, explaining, at least in part, the widespread observation of heterosis.

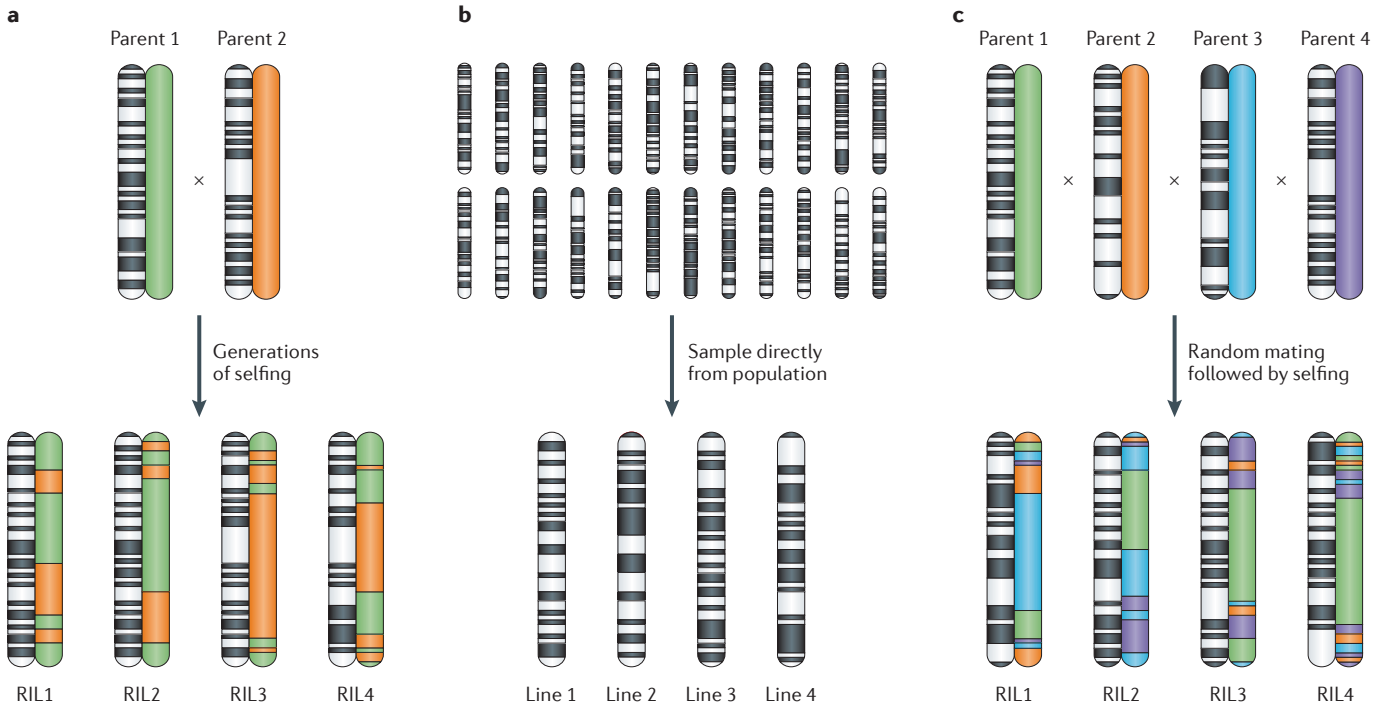



Figure 3 | **Mapping populations.** Comparison of three mapping strategies. A single chromosome is shown per individual. On the left, black and white bars indicate the genotype at each of 50 biallelic SNPs. The parental origin of each chromosomal segment is shown on the right in colour. **a** | Quantitative trait locus (QTL) mapping, in which two lines are crossed and the resulting $F_1$ generations are self-fertilized for several generations, resulting in homozygous recombinant inbred lines (RILs). **b** | Four lines from an association mapping panel. Lines are directly sampled from a population, such that no crosses are required but the parent-of-origin is unknown. **c** | A next-generation mapping population with four parents, which are randomly mated and then self-fertilized for several generations.

## Genetic architecture and plant genomes

Although an understanding of domestication history provides useful insight into species history and may help to identify loci of agronomic interest, plant-breeding efforts tend to be much more focused on the immediate needs of farmers and end users of crops. Yield, disease resistance, agronomic performance and product quality (for example, fruit or grain quality) are the typical areas of focus. Effective use of genetic variation for plant breeding requires an understanding of the genetic architecture of traits that have immediate applications to plant breeding.

***Quantitative trait locus mapping.*** Current understanding of genetic architecture is largely derived from quantitative trait locus (QTL) mapping[70]. QTL-mapping approaches generally begin with two parental inbred lines that are crossed for a number of generations to form a population of recombinant homozygous lines (FIG. 3a); typically, the $F_1$ generation is self-fertilized, but backcrossing and other strategies are also used. QTL approaches have proved to be enormously useful for plant breeding and have been successful in identifying loci of large effect and dissecting the genetic basis of fairly simple traits. The primary disadvantages of QTL mapping are the time involved in creating populations, the limited inference that can be made from alleles in two parental lines, the small number of recombination events captured in most mapping populations and a necessary focus on traits that can be readily and accurately phenotyped.

***Association mapping.*** The development of high-throughput, dense genotyping has led to a shift from traditional QTL mapping to association or LD mapping. Rather than focusing on two parental lines that differ strongly in phenotype, LD-mapping approaches assess the correlation between phenotype and genotype in populations of unrelated individuals (FIG. 3b). Association-mapping[71] panels sample more genetic diversity, can take advantage of many more generations of recombination and avoid the generations of time-consuming crosses that are necessary for QTL mapping[72]. When combined with dense, genome-wide marker coverage, association mapping can considerably improve the genetic resolution at which causative variants can be identified. Whereas QTLs identified by biparental mapping strategies can span tens of megabases, the long history of recombination events that is captured in most association panels enables a much greater genetic resolution[73]. With a large panel and with sufficiently dense genome-wide marker coverage, association mapping can potentially map causative loci to individual nucleotide changes.

Association mapping is widely used in a variety of crops, including those without extensive genomic resources, such as sugar beets[74] or pearl millet[75]. Only recently have plant studies incorporated the extremely high marker density needed for GWASs. Perhaps the best example to date is that of Huang *et al.*[76]. The authors used low-coverage resequencing of the genomes of a panel of more than 500 rice landraces and found 80 loci associated with 14 agronomic traits. Several of these associations were to previously characterized loci, which lent credence to the results. They identified a mix of genetic architectures among the traits studied: variation in colour and grain traits showed associations to few loci of large effect, but drought response, flowering and other morphological traits were explained by many loci of small effect. Although overall genetic resolution was quite high (~25 kb or 1–3 genes), the authors found several occasions in which known causative loci showed weaker signals than nearby markers. Similar results have been observed in GWASs conducted in *A. thaliana*[77] and suggest limits to the precision available in association-mapping studies, particularly in inbreeding organisms[77,78].

***Next-generation populations.*** A new generation of genetic-mapping populations has been designed with the goal of overcoming many of the limitations of biparental QTL mapping and association mapping. These populations combine the controlled crosses of QTL mapping with multiple parents and multiple generations of intermating. Next-generation populations are often larger than traditional QTL populations, and many lines are crossed in parallel; this increases the rate of effective recombination per generation and maximizes 'genetic map expansion', thereby improving genetic resolution compared to traditional biparental mapping[79]. As with association panels, next-generation populations will more effectively sample rare alleles than typical biparental populations. Because of the controlled nature of the crosses involved, next-generation populations can also overcome some of the difficulties of association mapping, including population structure and the unknown frequency of causative mutations. Owing to the approximately even contribution of all parents, next-generation designs also allow for better estimation of allelic effects than is possible under standard association-mapping approaches[80].

There are many potential designs for next-generation mapping populations[79], but all of them involve the crossing of multiple parents and advancement of populations through several generations to improve resolution in genetic mapping (FIG. 3c). The nested association mapping (NAM) population in maize is based on crossing diverse strains to a reference parent[81] and has already been successfully applied to the study of numerous traits[57,82–84]. Other designs have involved intercrossing multiple parents, forming a single large population[85]. These populations have been referred to as multiparent advanced generation intercross (MAGIC) populations[85,86] or recombinant inbred advanced intercross line (RIAIL) populations[79]. Individual crops may require different strategies: for example, designs that involve repeated outcrossing are difficult to implement for selfing species[85]. Multiparent crosses have a long history in plant breeding[87,88] and have been a source of considerable insight into evolutionary processes in crops[89,90].

***The problem of rare variants.*** Several challenges of association mapping in plants have been discussed previously (for example, REF. 25), including marker density and population structure, but other limitations are less well-understood. One important limitation concerns the frequency of causative mutations. Because individual SNPs that are at a very low frequency (approaching $1/n$, where $n$ is the sample size) probably explain a small fraction of the total trait variation within a population, most association-mapping studies have relied on the assumption that causative variants are fairly common. This idea, when it is applied to studies of the genetic basis of human disease, is known as the 'common disease–common variant' hypothesis[91]. Many studies in humans have been unable to explain much of the heritable genetic variation for traits, a result that may be due to polymorphisms that contribute to trait variation but that are kept at a low frequency in a population owing to the action of purifying selection[92]. Models for mutation–selection balance postulate that many of the mutations that contribute to quantitative genetic variation are unconditionally deleterious and would thus be found at low frequencies (reviewed in REF. 93). Identifying low-frequency causative polymorphisms is difficult, requiring much larger sample sizes (and thus more phenotyping), and the trait variance explained by rare SNPs is poorly estimated by the small number of phenotypes associated with rare genotypes. The ascertainment bias that is frequently present in association mapping SNP data also results in fewer low-frequency polymorphisms being genotyped[94]. Determination of the extent to which rare variants contribute to trait variation is a major goal of genome-resequencing projects, including the 1000 Genomes Project[95] in humans, the 1001 Genomes Project in *A. thaliana*[96], the *Drosophila* Genetic Reference Panel (which is resequencing 192 genomes of *D. melanogaster*) and the Maize Hapmap Phase II project (which is resequencing 103 maize genomes).

One recently reported consideration for association mapping is the potential for synthetic associations. Synthetic associations occur when a low-frequency causative mutation of large effect is not genotyped, leading more common, linked markers to appear to be associated with the trait[97]. Ascertainment bias in genotyped markers can lead to synthetic associations, but factors such as the Hill–Robertson effect[98] may also have a role. Evidence for synthetic associations has been recently documented in humans[99], but observations regarding genetic load (BOX 1) in crops suggest that it may be a widespread phenomenon.

## Applying evolution

Plant breeding is a form of applied evolution, and evolutionary analyses can provide hypotheses and models

Purifying selection
Selection against a deleterious allele.

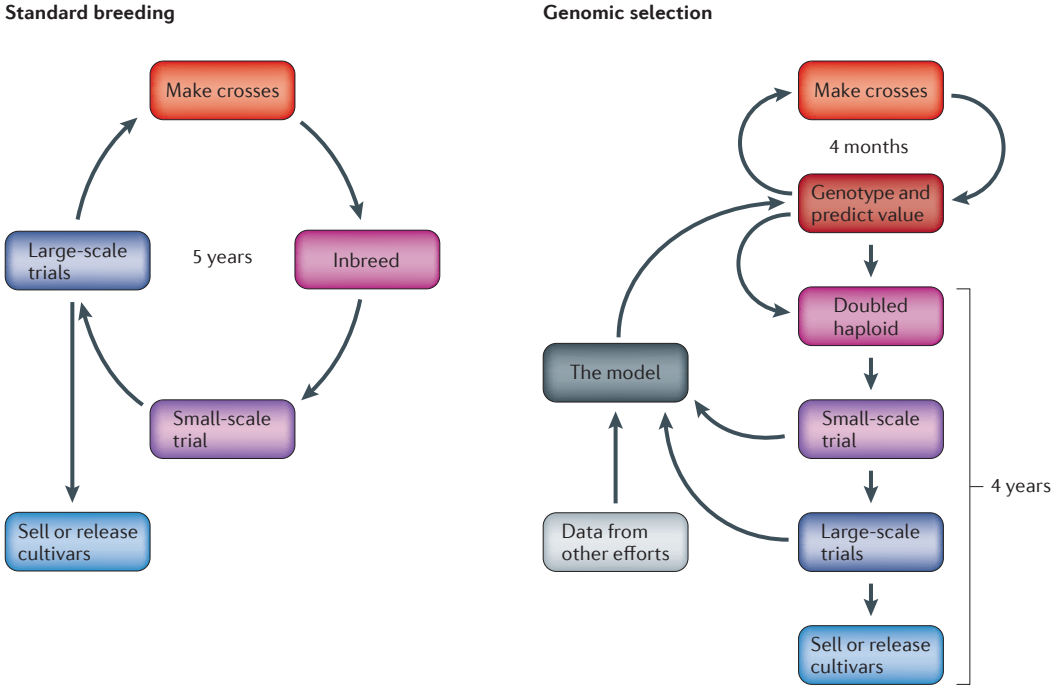
Ascertainment bias
Sampling bias that arises from how SNPs are chosen for inclusion on SNP arrays; SNPs that are known to be polymorphic in a particular population will have frequencies that are higher than would be expected by random sampling alone.

Hill–Robertson effect
The reduction in efficacy of selection at a locus owing to selection at linked loci.

that may substantially enrich and accelerate the search for useful variation. Clegg[100] pointed to three major areas of focus in evolutionary genetics that each have important implications for plant breeding — the genetic basis of adaptation, the quantification of variation and the processes of genetic transmission.

***Adaptation.*** The identification of loci that are responsible for adaptive evolution has long been a goal in evolutionary genetics (for example, REF. 4), and many of the approaches developed in the field (reviewed in REFS 101,102) are directly transferable to the identification of loci of agronomic interest[103].

Box 2 | **Genomic selection**




Genomic selection is a form of indexed, marker-assisted selection in which a marker data set is used to make phenotypic predictions[133,134]. Genomic selection and genome-wide association studies (GWASs) can use the same genotypic and phenotypic data, but genomic selection models de-emphasize the identification of individual polymorphisms that control complex traits in favour of weighted prediction of phenotypic values based on a training data set. Like GWASs, genomic selection has traditionally been limited by the cost and availability of dense genome-wide marker data, but recent developments in high-throughput genotyping allow for inexpensive genome-wide marker data to be rapidly collected in large numbers for even non-model taxa[135,136].

The key to genomic selection is the creation of training sets that have sufficient genetic and phenotypic diversity to permit selection to be applied in a meaningful way. Although most genomic selection has focused on rather narrow breeding efforts within one breeding programme[137], the long-term goal should be to produce models that encompass the worldwide diversity of a species, incorporating information on phenotype and performance in numerous environments. The details of genomic selection models are likely to vary by species and by breeding programme, and factors such as the genetic architecture of a trait will also be important in structuring the equations and priors of genomic selection models. Although trait architecture will undoubtedly differ among species, there also appear to be generalities that are worth predicting: for example, comparisons of flowering time between *Arabidopsis thaliana* and maize[138].

It is expected that genomic selection will revolutionize breeding in the next decade. The figure describes the breeding cycle that is common to maize in the twentieth century compared to a hypothetical breeding cycle that implements genomic selection. Whereas cultivar and hybrid trials and release in such a scenario would still take considerable time (4 years), the time between cycles of crossing could be up to 15 times faster (4 months versus 5 years) using genomic selection to choose lines for continued breeding. For example, although maize breeding and agronomy in the last century were tremendously successful and increased yield nearly eightfold in 70 years (US Department of Agriculture (USDA) National Agricultural Statistics Service)[139], adaptive evolution still occurred fairly slowly. Crosses, recombination and opportunities for allele frequency change only occurred every 5 or more years, and as much as half of the yield gain came from improved management practices[140].

Today, genomic selection efficiency falls far short of the goal suggested in the figure — its accuracy is limited by inefficiencies in the prediction of phenotype from genotype. In spite of these issues, current genomic selection methods are likely to be 2–3 times faster than the traditional breeding cycle. A continuing goal of crop genetics and breeding should be to improve methods of connecting phenotype to genotype — ideally, genomic selection will become indistinguishable from GWASs — until the pace of improvement is only limited by the biology of the species.

Box 3 | **Targeted genome editing**

A number of approaches from evolutionary and quantitative genetics can be used to identify the genomic location and genetic effect of loci of agronomic importance. However, validation of the genetic effects and use of individual alleles in plant-breeding programmes is expensive and time consuming[141]. It is hoped that genomic selection (BOX 2) will accelerate the introgression of multiple favourable alleles into breeding populations. Even the most precise marker-assisted introgression programme will introgress large chromosomal segments and will require multiple generations of backcrossing. Introgression of large regions limits the use of backcrosses for testing the genetic effect of individual alleles and increases the risk of introgression of unwanted linked variation.

The recent development of targeted genome-editing technologies, such as zinc finger nucleases[142] and transcription activator-like effector (TALE) nucleases[143] offers exciting potential to resolve these issues. These technologies make use of sequence-specific designer nucleases that cleave targeted loci, enabling creation of small insertions and deletions (indels), insertion of novel DNA or even replacement of individual alleles. Both methods have primarily been developed in non-plant systems, and facile and inexpensive application of these methods is not yet a reality in crop plants. Nonetheless, both zinc finger nucleases[144] and TALE nucleases[145] have been successfully applied to crops, and their potential impacts for plant breeding are enormous. Testing candidate loci may become a straightforward task, as locus-specific knockouts or allelic replacement allow both functional validation and a direct means of estimating effect sizes of individual alleles. Alleles at loci of known agronomic interest could be directly edited into individual lines, entirely bypassing the process of backcrossing. It is even possible to imagine targeted replacement of deleterious mutations (BOX 1) in elite breeding lines.

Although analysis of transgenic lines created using loci that have been functionally characterized in other systems provides a high-throughput means of testing genes that may contribute to agronomic phenotypes[146], exogenous genes are more likely to have undesired interactions with the native genomic background. Similar to the situation for Dobzhansky–Muller effects that are associated with incipient speciation[147], such genic interactions will probably differ among individuals or breeding populations. By contrast, improvement efforts based primarily on extant diversity may be less likely to encounter such negative interactions[148]. Because targeted transgenesis and the ability to use extant variation are likely to prove to be more effective than random incorporation of transgenic events, genomic editing is likely to offer an attractive alternative to current transgenic technologies.

Loci that have been identified as the targets of selection for adaption to the agronomic environment can be incorporated into breeding in two ways. First, identification of selected genes may provide direct targets for future improvement. The rice 'green revolution' gene semidwarf1 provides a particularly compelling example: although it was selected during domestication[104], variation at this locus has played an important part in modern rice breeding[105]. In the case of waxy maize, improvement occurred through selection on a locus in the same pathway as a gene targeted by selection during domestication[106].

Second, selection during historical adaptation is often accompanied by a considerable loss of diversity at the target locus and at linked genes and non-coding sequence[14,46,54]; this limits the variability that breeders have to work with in modern breeding populations. Modern breeding could thus benefit from the careful reintroduction of diversity in these regions. Genomic selection (BOX 2) could be used to introgress variation in these regions from traditional cultivars or wild taxa while selecting for acceptable agronomic phenotypes. Alternatively, targeted genome-editing technologies (BOX 3) provide exciting opportunities for changing individual nucleotides and small regions of native genes.

Dobzhansky–Muller effects
Intrinsic reductions in viability or fertility resulting from epistatic interactions between multiple substitutions, typically observed in the offspring of a cross between individuals from genetically distinct populations.

***Recombination and variation.*** There are two fundamental sources of genetic diversity: mutation and recombination. Mutations, which are broadly defined as novel heritable variants, are reassorted or rearranged along a chromosome into new combinations by recombination. Next-generation sequencing has made it possible to readily identify all common mutations in populations, particularly in single-copy genes and in well-annotated portions of genomes. The remaining challenge in understanding diversity then revolves around the haplotypes that are formed by combinations of mutations, and many questions become contingent on rates and patterns of historical recombination[107].

Recombination rates can vary dramatically across regions of a chromosome; in many crop genomes, extended pericentromeric and centromeric regions show substantially reduced levels of recombination[6,14,108]. In maize, >20% of low-copy-number genes are in these low-recombination regions[14], and Hill–Robertson effects are likely to have prevented breeding efforts from effectively exploiting variation at these loci. There is substantial evidence that much of maize heterosis results from persistence of deleterious mutations that are difficult to eliminate owing to a lack of recombination[14,109,110]. Genomic selection (BOX 2) models could be modified to select for crossovers in low-recombination regions of the genome, exposing new haplotypes and allelic combinations and allowing for more efficient purging of genetic load (BOX 1).

***Mating system.*** Finally, because genetic transmission is mediated by the pattern of mating among individuals, almost all forms of genetic change are affected by the mating system[100]. Mating system is likely to have played a fundamental part in crop evolution owing to the advantage of lines that breed true for favourable traits[111]. The most readily evident consequence of mating-system variation is its effect on levels of heterozygosity and effective recombination, which affect patterns of LD genome-wide[112]. High levels of inbreeding do not always translate to high levels of LD[113,114], and the impact of a selfing mating system depends to a large degree on how recently selfing arose in a given lineage[115].

There is emerging empirical evidence for the impact of mating system on trait architecture. For example, genetic mapping in the NAM population has found that flowering time in outcrossing maize is controlled by a large number of loci of small effect, and there is little evidence of genetic interactions among loci or between loci and the environment[57]. This differs dramatically from GWASs in the selfing species *A. thaliana*, in which a number of well-characterized loci explain a large portion of the variance for the trait[77,116]. Selfing crops, including barley, rice and sorghum, also have individual loci with large additive effects on flowering time[117–119]. Mating-system-derived differences in the fundamental architecture of traits thus have important implications for breeding strategies that are required to select on these traits effectively.

## Conclusions

As it continues to be redefined by applications of genome resequencing, high-density genetic markers and a new generation of experimental designs that more readily relate mutational diversity to agronomic phenotypes, comparative genomics will become increasingly relevant to crop improvement. Recent studies of *A. thaliana* suggest that sequencing and annotation of additional genomes to reference quality[120] will provide a much better assessment of the functional content of genomes. Some of the best opportunities may lie in using evolutionary analyses to direct and optimize the discovery of variation and to provide enhanced recombination. Limited effective recombination may have constrained selection efficiency; efforts to eliminate putatively deleterious mutations and directed efforts at restoration of diversity around loci that are subject to strong selection or historical trait introgression are potentially novel applications of genome-wide diversity data.

A combination of GWASs and next-generation-mapping populations will improve our ability to connect phenotypes and genotypes, and genomic selection can take advantage of this data for rapid selection and breeding. Evolutionary analyses can identify signals of historical selection at loci with unknown phenotypic effects or with a lack of diversity and may direct breeding towards particular loci or genomic regions that could most benefit from improvement. The combination of these approaches with the promise of improved genomic modification technologies provides an opportunity for comparative genomics to apply our understanding of the past to the future of crop improvement.


1. Paterson, A. H., Freeling, M. & Sasaki, T. Grains of knowledge: genomics of model cereals. *Genome Res.* **15**, 1643–1650 (2005).
2. Brown, A. H. D. Enzyme polymorphism in plant populations. *Theor. Popul. Biol.* **15**, 1–42 (1979).
3. Bowers, J. E. *et al.* A high-density genetic recombination map of sequence-tagged sites for *Sorghum*, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. *Genetics* **165**, 367–386 (2003).
4. Lewontin, R. C. & Krakauer, J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175–195 (1973).
5. Nei, M. & Maruyama, T. Lewontin–Krakauer test for neutral genes — comment. *Genetics* **80**, 395–395 (1975).
6. Schulte, D. *et al.* The International Barley Sequencing Consortium—at the threshold of efficient access to the barley genome. *Plant Physiol.* **149**, 142–147 (2009).
7. Kaul, S. *et al.* Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
8. Initiative, T. I. B. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010).
9. Pool, J. E., Hellmann, I., Jensen, J. D. & Nielsen, R. Population genetic inference from genomic sequence variation. *Genome Res.* **20**, 291–300 (2010).
10. Metzker, M. L. Sequencing technologies — the next generation. *Nature Rev. Genet.* **11**, 31–46 (2010).
11. Lockton, S. & Gaut, B. S. Plant conserved non-coding sequences and paralogue evolution. *Trends Genet.* **21**, 60–65 (2005).
12. Haun, W. J. *et al.* The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. *Plant Physiol.* **155**, 645–655 (2011).
13. Velasco, R. *et al.* A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**, e1326 (2007).
14. Gore, M. A. *et al.* A first-generation haplotype map of maize. *Science* **326**, 1115–1117 (2009).
15. Young, N. D. *et al.* The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* 16 Nov 2011(doi:1001038/nature10625).
16. Shulaev, V. *et al.* The genome of woodland strawberry (*Fragaria vesca*). *Nature Genet.* **43**, 109–116 (2011).
17. Xu, X. *et al.* Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189–195 (2011).
**This is an excellent example of both the challenges and promise of comparative genomics in crop plant genomes. To overcome polyploidy and high levels of heterozygosity, the authors use a combination of traditional Sanger and next-generation methods to sequence and annotate the genome of a doubled-monoploid potato line.**
18. Rafalski, A. & Morgante, M. Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. *Trends Genet.* **20**, 103–111 (2004).
19. Lijavetzky, D., Cabezas, J. A., Ibanez, A., Rodriguez, V. & Martinez-Zapater, J. M. High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics* **8**, 424 (2007).
20. Caldwell, K. S., Russell, J., Langridge, P. & Powell, W. Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. *Genetics* **172**, 557–567 (2006).
21. Gaut, B. S. & Ross-Ibarra, J. Selection on major components of angiosperm genomes. *Science* **320**, 484–486 (2008).
22. Tenaillon, M. I., Hollister, J. D. & Gaut, B. S. A triptych of the evolution of plant transposable elements. *Trends Plant Sci.* **15**, 471–478 (2010).
23. Nordborg, M. & Weigel, D. Next-generation genetics in plants. *Nature* **456**, 720–723 (2008).
24. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
25. Ross-Ibarra, J., Morrell, P. L. & Gaut, B. S. Colloquium papers: plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc. Natl Acad. Sci. USA* **104** (Suppl. 1), 8641–8648 (2007).
26. Brown, A. H. D. Variation under domestication in plants: 1859 and today. *Phil. Trans. R. Soc. B* **365**, 2523–2530 (2010).
27. Harris, D. R. Vavilov's concept of centres of origin of cultivated plants: its genesis and its influence on the study of agricultural origins. *Biol. J. Linn. Soc.* **39**, 7–16 (1990).
28. Rosenberg, N. A. & Nordborg, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Rev. Genet.* **3**, 380–390 (2002).
**This is a very accessible introduction to genealogical histories and coalescent theory that are pertinent to interpretation of sequence polymorphism data.**
29. Gaut, B. S. & Clegg, M. T. Molecular evolution of the *Adh1* locus in the genus *Zea*. *Proc. Natl Acad. Sci. USA* **90**, 5095–5099 (1993).
30. Kim, M. Y. *et al.* Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. *Proc. Natl Acad. Sci. USA* **107**, 22032–22037 (2010).
31. Eyre-Walker, A., Gaut, R. L., Hilton, H., Feldman, D. L. & Gaut, B. S. Investigation of the bottleneck leading to the domestication of maize. *Proc. Natl Acad. Sci. USA* **95**, 4441–4446 (1998).
32. Haudry, A. *et al.* Grinding up wheat: a massive loss of nucleotide diversity since domestication. *Mol. Biol. Evol.* **24**, 1506–1517 (2007).
33. Wright, S. I. *et al.* The effects of artificial selection on the maize genome. *Science* **308**, 1310–1314 (2005).
**This paper discusses a comparative resequencing study that used an original likelihood ratio test to model demography and selection. The paper was unique in providing an estimate of the proportion of loci in the genome involved in domestication and/or improvement.**
34. Caicedo, A. L. *et al.* Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* **3**, 1745–1756 (2007).
35. Molina, J. *et al.* Molecular evidence for a single evolutionary origin of domesticated rice. *Proc. Natl Acad. Sci. USA* **108**, 8351–8356 (2011).
36. Matsuoka, Y. *et al.* A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl Acad. Sci. USA* **99**, 6080–6084 (2002).
37. Li, Y. H. *et al.* Genetic diversity in domesticated soybean (*Glycine max*) and its wild progenitor (*Glycine soja*) for simple sequence repeat and single-nucleotide polymorphism loci. *New Phytol.* **188**, 242–253 (2010).
38. Chen, H., Morrell, P. L., Ashworth, V. E., de la Cruz, M. & Clegg, M. T. Tracing the geographic origins of major avocado cultivars. *J. Hered.* **100**, 56–65 (2009).
39. Gepts, P. & Bliss, F. A. Phaseolin variability among wild and cultivated common beans (*Phaseolus vulgaris*) from Colombia. *Econ. Bot.* **40**, 469–478 (1986).
40. Morrell, P. L. & Clegg, M. T. Genetic evidence for a second domestication of barley (*Hordeum vulgare*) east of the Fertile Crescent. *Proc. Natl Acad. Sci. USA* **104**, 3289–3294 (2007).
41. Allaby, R. G., Fuller, D. Q. & Brown, T. A. The genetic expectations of a protracted model for the origins of domesticated crops. *Proc. Natl Acad. Sci. USA* **105**, 13982–13986 (2008).
42. Ross-Ibarra, J. & Gaut, B. S. Multiple domestications do not appear monophyletic. *Proc. Natl Acad. Sci. USA* **105**, E105; author reply E106 (2008).
43. Kwak, M. & Gepts, P. Structure of genetic diversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., Fabaceae). *Theor. Appl. Genet.* **118**, 979–992 (2009).
44. Myles, S. *et al.* Genetic structure and domestication history of the grape. *Proc. Natl Acad. Sci. USA* **108**, 3530–3535 (2011).
45. van Heerwaarden, J. *et al.* Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proc. Natl Acad. Sci. USA* **108**, 1088–1092 (2011).
46. He, Z. *et al.* Two evolutionary histories in the genome of rice: the roles of domestication genes. *PLoS Genet.* **7**, e1002100 (2011).
47. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* **5**, e1000519 (2009).
48. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
49. Darwin, C. *The Variation of Animals and Plants under Domestication* (Appleton, New York, 1876).
50. Hammer, K. Das Domestikationssyndrom. *Kulturpflanze* **32**, 11–34 (1984).
51. Burger, J. C., Chapman, M. A. & Burke, J. M. Molecular insights into the evolution of crop plants. *Am. J. Bot.* **95**, 113–122 (2008).
52. Vigouroux, Y. *et al.* Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proc. Natl Acad. Sci. USA* **99**, 9650–9655 (2002).
53. Chapman, M. A. *et al.* A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (*Helianthus annuus*). *Plant Cell* **20**, 2931–2945 (2008).

54. Lam, H. M. *et al.* Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genet.* **42**, 1053–1059 (2010).
55. Hurwitz, B. L. *et al.* Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus *Oryza*. *Plant J.* **63**, 990–1003 (2010).
56. Swanson-Wagner, R. A. *et al.* Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.* **20**, 1689–1699 (2010).
57. Buckler, E. S. *et al.* The genetic architecture of maize flowering time. *Science* **325**, 714–718 (2009). **This study, using the NAM population, found that flowering time in maize provides a good fit to classic models of a quantitative trait and that a large number of loci contribute additively to the phenotype.**
58. Vielle-Calzada, J. P. *et al.* The Palomero genome suggests metal effects on domestication. *Science* **326**, 1078–1078 (2009).
59. Sugimoto, K. *et al.* Molecular cloning of Sdr4, a regulator involved in seed dormancy and domestication of rice. *Proc. Natl Acad. Sci.* **107**, 5792–5797 (2010).
60. Takahashi, R. The origin and evolution of cultivated barley. *Adv. Genet.* **7**, 227–266 (1955).
61. Morrell, P. L. & Clegg, M. T. in *Wild Crop Relatives: Genomic and Breeding Resources: Cereals* (ed. Kole, C.) 309–320 (Springer, Berlin, 2011).
62. Jones, H. *et al.* Population-based resequencing reveals that the flowering time adaptation of cultivated barley originated east of the Fertile Crescent. *Mol. Biol. Evol.* **25**, 2211–2219 (2008).
63. Bellon, M. R., Hodson, D. & Hellin, J. Assessing the vulnerability of traditional maize seed systems in Mexico to climate change. *Proc. Natl Acad. Sci. USA* **108**, 13432–13437 (2011).
64. Komatsuda, T. *et al.* Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proc. Natl Acad. Sci. USA* **104**, 1424–1429 (2007).
65. Purugganan, M. D., Boyles, A. L. & Suddith, J. I. Variation and selection at the *CAULIFLOWER* floral homeotic gene accompanying the evolution of domesticated *Brassica oleracea*. *Genetics* **155**, 855–862 (2000).
66. Teshima, K. M., Coop, G. & Przeworski, M. How reliable are empirical genomic scans for selective sweeps? *Genome Res.* **16**, 702–712 (2006).
67. Ralph, P. & Coop, G. Parallel adaptation: one or many waves of advance of an advantageous allele? *Genetics* **186**, 647–668 (2010).
68. Paterson, A. H. *et al.* Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. *Science* **269**, 1714–1718 (1995).
69. Goff, S. A. A unifying theory for general multigenic heterosis: energy efficiency, protein metabolism, and implications for molecular breeding. *New Phytol.* **189**, 923–937 (2011).
70. Mauricio, R. Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. *Nature Rev. Genet.* **2**, 370–381 (2001).
71. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
72. Myles, S. *et al.* Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* **21**, 2194–2202 (2009).
73. Mackay, T. F., Stone, E. A. & Ayroles, J. F. The genetics of quantitative traits: challenges and prospects. *Nature Rev. Genet.* **10**, 565–577 (2009).
74. Wurschum, T. *et al.* Genome-wide association mapping of agronomic traits in sugar beet. *Theor. Appl. Genet.* **123**, 1121–1131 (2011).
75. Saidou, A. A. *et al.* Association studies identify natural variation at PHYC linked to flowering time and morphological variation in pearl millet. *Genetics* **182**, 899–910 (2009).
76. Huang, X. *et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genet.* **42**, 961–967 (2010).
77. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
78. Hamblin, M. T., Buckler, E. S. & Jannink, J. L. Population genetics of genomics-based crop improvement methods. *Trends Genet.* **27**, 98–106 (2011).
79. Rockman, M. V. & Kruglyak, L. Breeding designs for recombinant inbred advanced intercross lines. *Genetics* **179**, 1069–1078 (2008). **This study provides an examination of breeding designs that maximize genetic resolution in intercross populations.**
80. Macdonald, S. J. & Long, A. D. Joint estimates of quantitative trait locus effect and frequency using synthetic recombinant populations of *Drosophila melanogaster*. *Genetics* **176**, 1261–1281 (2007). **The authors of this paper provide a strong rationale for the development of next-generation populations. The study design permits estimation of QTL location, effect and frequency. Comparison of effect size of alleles contributed by founders of the population is particularly compelling.**
81. Yu, J. M., Holland, J. B., McMullen, M. D. & Buckler, E. S. Genetic design and statistical power of nested association mapping in maize. *Genetics* **178**, 539–551 (2008).
82. Brown, P. J. *et al.* Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS Genet.* **7**, e1002383 (2011).
83. Kump, K. L. *et al.* Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nature Genet.* **43**, 163–168 (2011).
84. Tian, F. *et al.* Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature Genet.* **43**, 159–162 (2011).
85. Cavanagh, C., Morell, M., Mackay, I. & Powell, W. From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Curr. Opin. Plant Biol.* **11**, 215–221 (2008).
86. Kover, P. X. *et al.* A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet.* **5**, e1000551 (2009).
87. Harlan, H. V. & Martini, M. L. A composite hybrid mixture. *J. Am. Soc. Agron.* 487–490 (1929).
88. Suneson, C. A. An evolutionary plant breeding method. *Agron. J.* **48**, 188–191 (1956).
89. Allard, R. W., Kahler, A. L. & Weir, B. S. The effect of selection on esterase allozymes in a barley population. *Genetics* **72**, 489–503 (1972).
90. Clegg, M. T., Allard, R. W. & Kahler, A. L. Is the gene the unit of selection? Evidence from two experimental plant populations. *Proc. Natl Acad. Sci. USA* **69**, 2474–2478 (1972).
91. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
92. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
93. Johnson, T. & Barton, N. Theoretical models of selection and mutation on quantitative traits. *Phil. Trans. R. Soc. Lond. B* **360**, 1411–1425 (2005).
94. Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H. & Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15**, 1496–1502 (2005).
95. Altshuler, D. L. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
96. Weigel, D. & Mott, R. The 1001 Genomes Project for *Arabidopsis thaliana*. *Genome Biol.* **10**, 107 (2009).
97. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
98. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294 (1966).
99. Huff, C. D. *et al.* Crohn's disease and genetic hitchhiking at IBD5. *Mol. Biol. Evol.* 4 Aug 2011 (doi:10.1093/molbev/msr151).
100. Clegg, M. T. Measuring plant mating systems. *Bioscience* **30**, 814–818 (1980).
101. Nielsen, R. Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197–218 (2005).
102. Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. & Clark, A. G. Recent and ongoing selection in the human genome. *Nature Rev. Genet.* **8**, 857–868 (2007).
103. Walsh, B. Using molecular markers for detecting domestication, improvement, and adaptation genes. *Euphytica* **161**, 1–17 (2008).
104. Asano, K. *et al.* Artificial selection for a green revolution gene during *japonica* rice domestication. *Proc. Natl Acad. Sci. USA* **108**, 11034–11039 (2011).
105. Spielmeyer, W., Ellis, M. H. & Chandler, P. M. Semidwarf (sd-1), "green revolution" rice, contains a defective gibberellin 20-oxidase gene. *Proc. Natl Acad. Sci. USA* **99**, 9043–9048 (2002).
106. Fan, L. *et al.* Post-domestication selection in the maize starch pathway. *PLoS ONE* **4**, e7612 (2009).
107. Stumpf, M. P. & McVean, G. A. Estimating recombination rates from population-genetic data. *Nature Rev. Genet.* **4**, 959–968 (2003).
108. Paterson, A. H. *et al.* The *Sorghum* bicolour genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
109. McMullen, M. D. *et al.* Genetic properties of the maize nested association mapping population. *Science* **325**, 737–740 (2009).
110. Schon, C. C., Dhillon, B. S., Utz, H. F. & Melchinger, A. E. High congruency of QTL positions for heterosis of grain yield in three crosses of maize. *Theor. Appl. Genet.* **120**, 321–332 (2010).
111. Allard, R. W. History of plant population genetics. *Annu. Rev. Genet.* **33**, 1–27 (1999).
112. Nordborg, M. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**, 923–929 (2000).
113. Morrell, P. L., Toleno, D. M., Lundy, K. E. & Clegg, M. T. Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization. *Proc. Natl Acad. Sci. USA* **102**, 2442–2447 (2005).
114. Morrell, P. L., Toleno, D. M., Lundy, K. E. & Clegg, M. T. Estimating the contribution of mutation, recombination and gene conversion in the generation of haplotypic diversity. *Genetics* **173**, 1705–1723 (2006).
115. Charlesworth, D. Effects of inbreeding on the genetic diversity of populations. *Phil. Trans. R. Soc. Lond. B* **358**, 1051–1070 (2003).
116. Zhao, K. Y. *et al.* An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.* **3**, e4 (2007).
117. Turner, A., Beales, J., Faure, S., Dunford, R. P. & Laurie, D. A. The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. *Science* **310**, 1031–1034 (2005).
118. Yano, M. *et al.* Identification of quantitative trait loci controlling heading date in rice using a high-density linkage map. *Theor. Appl. Genet.* **95**, 1025–1032 (1997).
119. Lin, Y. R., Schertz, K. F. & Paterson, A. H. Comparative analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific *Sorghum* population. *Genetics* **141**, 391–411 (1995).
120. Gan, X. *et al.* Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**, 419–423 (2011). ***De novo* sequencing and annotation along with trascriptome sequencing of 18 reference genomes from the founders of a next-generation population are discussed in this paper. Re-annotation of individual genes suggests that many genes that appear to have lost function in simple comparisons with the original *A. thaliana* reference genome contain compensatory mutations that restore function at the locus.**
121. Muller, H. J. Our load of mutations. *Am. J. Hum. Genet.* **2**, 111–176 (1950).
122. Ng, S. B. *et al.* Exome sequencing identifies the cause of a Mendelian disorder. *Nature Genet.* **42**, 30–35 (2010).
123. Gossmann, T. I. *et al.* Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol. Biol. Evol.* **27**, 1822–1832 (2010).
124. Lai, J. S. *et al.* Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature Genet.* **42**, 1027–1030 (2010).
125. Gunther, T. & Schmid, K. J. Deleterious amino acid polymorphisms in *Arabidopsis thaliana* and rice. *Theor. Appl. Genet.* **121**, 157–168 (2010).
126. Lu, J. *et al.* The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends Genet.* **22**, 126–131 (2006).
127. Tang, H. B., Sezen, U. & Paterson, A. H. Domestication and plant genomes. *Curr. Opin. Plant Biol.* **13**, 160–166 (2010).
128. Chun, S. & Fay, J. C. Evidence for hitchhiking of deleterious mutations within the human genome. *PLoS Genet.* **7**, e1002240 (2011).
129. Lande, R. & Schemske, D. W. The evolution of self-fertilization and inbreeding depression in plants. I. Genetic models. *Evolution* **39**, 24–40 (1985).

130. Charlesworth, D. & Willis, J. H. The genetics of inbreeding depression. *Nature Rev. Genet.* **10**, 783–796 (2009).
131. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78**, 737–756 (1974).
132. Lynch, M., Conery, J. & Burger, R. Mutational meltdowns in sexual populations. *Evolution* **49**, 1067–1080 (1995).
133. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).
134. Heffner, E. L., Sorrells, M. E. & Jannink, J. L. Genomic selection for crop improvement. *Crop Sci.* **49**, 1–12 (2009).
135. Andolfatto, P. *et al.* Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.* **21**, 610–617 (2011).
136. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**, e19379 (2011).
137. Heffner, E. L., Jannink, J.-L. & Sorrells, M. E. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Gen.* **4**, 65–75 (2011).
138. Salome, P. A. *et al.* Genetic architecture of flowering-time variation in *Arabidopsis thaliana*. *Genetics* **188**, 421–433 (2011).
139. Troyer, A. F. Adaptedness and heterosis in corn and mule hybrids. *Crop Sci.* **46**, 528–543 (2006).
140. Duvick, D. N. The contribution of breeding to yield advances in maize (*Zea mays* L.). *Adv. Agron.* **86**, 83–145 (2005).
141. Bernardo, R. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Sci.* **48**, 1649–1664 (2008).
142. Weinthal, D., Tovkach, A., Zeevi, V. & Tzfira, T. Genome editing in plant cells by zinc finger nucleases. *Trends Plant Sci.* **15**, 308–321 (2010).
143. Bogdanove, A. J. & Voytas, D. F. TAL effectors: customizable proteins for DNA targeting. *Science* **333**, 1843–1846 (2011).
144. Shukla, V. K. *et al.* Precise genome modification in the crop species *Zea mays* using zinc-finger nucleases. *Nature* **459**, 437–441 (2009).
**This paper is an impressive demonstration of the potential power of targeted genomic editing. The authors use a custom zinc finger nuclease to modify two traits in maize and show that the method is sufficiently precise to target only one of the two paralogues of the enzyme of interest.**
145. Morbitzer, R., Romer, P., Boch, J. & Lahaye, T. Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors. *Proc. Natl Acad. Sci. USA* **107**, 21617–21622 (2010).
146. Century, K., Reuber, T. L. & Ratcliffe, O. J. Regulating the regulators: the future prospects for transcription-factor-based agricultural biotechnology products. *Plant Physiol.* **147**, 20–29 (2008).
147. Presgraves, D. C. The molecular evolutionary basis of species formation. *Nature Rev. Genet.* **11**, 175–180 (2010).
148. Gepts, P. A comparison between crop domestication, classical plant breeding, and genetic engineering. *Crop Sci.* **42**, 1780–1790 (2002).
149. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**, 991–996 (2008).
150. Argout, X. *et al.* The genome of *Theobroma cacao*. *Nature Genet.* **43**, 101–108 (2011).
151. Goff, S. A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92–100 (2002).
152. Varshney, R. K. *et al.* Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nature Biotech.* 6 Nov 2011 (doi:10.1038/nbt.2022).
153. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
154. Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).

**Acknowledgements**
We thank the US Department of Agriculture (USDA)–US National Institute for Food and Agriculture (NIFA) (2011-68002-30029) for partial support for P.L.M., USDA–NIFA (2009-01864) for J.R.-I. and the USDA Agriculture Research Service and the National Science Foundation (NSF)–Plant Genome Research (0820691) and NSF–Division of Biological Infrastructure (0965342) for support to E.S.B. The authors are grateful to J. Gerke, A. Gonzales, M. Hufford, D. Segal, R. Stupar and three anonymous referees for comments on the manuscript.

**Competing interests statement**
The authors declare no competing financial interests.

## FURTHER INFORMATION

**Peter L. Morrell's homepage:** http://faculty.agronomy.cfans.umn.edu/pmorrell
**Edward S. Buckler's homepage:** http://www.maizegenetics.net
**Jeffrey Ross-Ibarra's homepage:** http://www.rilab.org
**1000 Genomes Project (in humans):** http://www.1000genomes.org
**1001 Genomes Project (in *A. thaliana*):** http://www.1001genomes.org
***Drosophila* Genetic Reference Panel:** http://www.hgsc.bcm.tmc.edu/project-species-i-Drosophila_genRefPanel.hgsc
**Gramene:** http://www.gramene.org
**International Barley Sequencing Consortium:** http://www.public.iastate.edu/~imagefpc/IBSC%20Webpage/IBSC%20Template-home.html
**International Tomato Genome Sequencing Project:** http://solgenomics.net/organism/solanum_lycopersicum/genome
**International Wheat Genome Sequencing Consortium:** http://www.wheatgenome.org
**Phytozome:** http://www.phytozome.net
**US Department of Agriculture (USDA) National Agricultural Statistics Service:** http://www.nass.usda.gov

**ALL LINKS ARE ACTIVE IN THE ONLINE PDF**