REVIEW

doi:10.1038/natureXXXXX

Genomic innovation for crop improvement

Michael W. Bevan[1], Cristobal Uauy[1], Brande B. H. Wulff[1], Ji Zhou[1,2] Ksenia Krasileva[2,3] & Matthew D. Clark[2,4]

Crop production needs to increase to secure future food supplies, while reducing its impact on ecosystems. Detailed characterization of plant genome structure and genetic diversity is crucial for meeting these challenges. Advances in genome sequencing and assembly are being used to access to the large and complex genomes of crops and their wild relatives. Sequencing of wild crop relatives is identifying a wide spectrum of genetic variation, permitting the association of genetic diversity with diverse agronomic phenotypes. In combination with improved and automated phenotyping assays and functional genomic studies, genomics is providing new foundations for crop-breeding systems.

In the twentieth century, famines caused by political crises, mismanagement of food production or genocide killed an estimated 70 million people and were second only to war as the greatest man-made cause of death[1]. The father of the Green Revolution, Norman Borlaug, summarized the importance of crops: "Without food, people perish, social and political organizations disintegrate, and civilizations collapse." For thousands of years, people have dedicated considerable resources to securing food supplies; for example, grain supplies to ancient Rome were secured through an extensive network of long-distance transport, and the distribution of grain was coordinated and subsidized by the *cura annonae* ('care for the grain supply'), an important figure who contributed to the maintenance of political unity and power.

During the past 10,000 years, a period known as the Holocene, Earth's environment has been unusually stable. This probably facilitated the domestication of crops from wild species, resulting in steadily improved yields and adaptation to new agricultural areas. The production of food is now carried out on a vast scale, with 38% of Earth's surface dedicated to agriculture[2]. This increased production is having a pervasive influence on ecosystems worldwide: nitrogen production for agriculture accounts for 1.2% of global energy consumption[3]; photosynthesis can no longer maintain stable levels of atmospheric carbon dioxide; and food production consumes around 70% of freshwater supplies. The modelling of crop responses to increases in temperature predicts that there will be a considerable reduction in the yield of rice, an important crop around the world[4]. Climate change could also alter the dynamics of crop pathogenic agents by altering the range of vectors and by compromising the immune response of crops[5].

Crop production must therefore adapt to more variable environments and the substantial impact it has on the environment needs to be reduced. Productivity must also increase at a much greater rate than in the past to meet the needs of Earth's growing population[6]. Genetic improvements in crop performance continue to be crucial for increasing crop productivity, but current rates of improvement are unable to meet the demands of sustainability and food security[7]. Plant genomics has a central role in the improvement of crops, including discovery of genetic variation that underlies improved performance and increasing the efficiency of plant breeding. Both approaches are important owing to the long lead time for breeding new varieties and the need to identify new sources of genetic variation. In this Review, we describe advances in genome sequencing and assembly technologies and explain how these can be applied accurately to assemble large and complex crop genomes, as well as to access genetic diversity on an unprecedented scale. These genomic data are enabling key steps in crop improvement, such as trait identification and alteration, the breeding process and performance optimization, which can now be considered as DNA sequence analysis problems. We also show how advances in genomic and computational biology can address bottlenecks in crop improvement and make important contributions to food security.

Crop plant genomes

The extraordinary diversity of plant species is reflected in their genomes, which vary greatly in size and complexity[8]. Dramatic increases in genome size, notably in the grasses, are driven by bursts of retroelement or DNA-repeat expansion that tend to preserve an underlying conserved order and composition of genes[9]. DNA repeats have an important role in generating phenotypic diversity and plants have evolved epigenetic mechanisms to limit the parasitic expansion of repeats[10,11]. The other dominant feature of plant-genome evolution is whole-genome duplication (see Box 1 for a definition of this and other terms that are used in crop improvement), which is pervasive in most plant lineages[12]. Whole-genome duplication can lead to aneuploidy, asymmetric genome evolution[13], the rapid loss of genes, exchange between chromosomes and new gene functions, and is therefore an important driver of genetic and phenotypic diversity and adaptation. Many crops are hybrids that have been domesticated from naturally occurring polyploids or generated by breeding programmes. Figure 1 highlights the roles that genome duplication has played in the formation of the complex genomes of two major crops, wheat (*Triticum* spp.) and members of the genus *Brassica*. The large genome sizes, long tracts of related repeat sequences and the closely related homeologous genes in the large gene families of polyploid crops have presented considerable challenges for sequencing technologies. Assembling accurate and representative genomes and assessing the full range of available genetic variation are therefore central aims for crop plant genomics.

DNA sequencing and assembly technologies

Several strategies have been adopted for the sequencing and assembly of large polyploid genomes of crops. One such approach involves the reduction of genome complexity. Although the massive 22 Gb genome of the loblolly pine (*Pinus taeda* L.) is highly heterozygous, the reduced haploid genome of tissues that give rise to the germline (the gametophyte) of the tree has been successfully sequenced and assembled[14]. A

[1]John Innes Centre, Norwich Research Park, Norwich NR4 7UH, United Kingdom. [2]Earlham Institute, Norwich Research Park, Norwich NR4 7UH, United Kingdom. [3]The Sainsbury Laboratory, Norwich Research Park, Norwich NR4 7UH, United Kingdom. [4]School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, United Kingdom.

second approach involves the sequencing of diploid progenitor species. For example, strawberry (*Fragaria* × *ananassa*) is an octoploid that formed from four diploid progenitors. To access the genetic diversity of this valuable crop, a diploid variety of *F. vesca* was sequenced[15]. Oilseed rape or canola (*Brassica napus*) is an allopolyploid that formed from two diploid species of *Brassica* that are triplicated versions of an ancestral diploid (Fig. 1). Genome assemblies of *B. napus* were assigned to these two subgenomes using sequence assemblies from each diploid progenitor, but many sequence scaffolds showed ambiguous assignment to homeologous groups, owing to homeologue exchange and frequent gene loss[16]. A similar strategy was used to characterize the allotetraploid genome of the peanut (*Arachis hypogaea*), which formed from two diploid species. Essentially complete assemblies of the genomes of the probable progenitor species *A. duranensis* (1.2 Gb) and *A. ipanensis* (1.5 Gb) were generated and shown to directly align with the genetic map of a cultivated tetraploid peanut[17]. Moleculo synthetic long-read sequencing[18] of the tetraploid peanut genome showed that it was 98–99% identical to the diploid genomes, with differences due to recombination between the subgenomes. A third approach to the deconvolution of polyploid genomes involves the sequencing of DNA from purified chromosome arms. This was applied to the large allohexaploid genome of bread wheat, in which the independently maintained A, B and D genomes can tolerate the loss of entire chromosomes. Chromosomes from wheat lines that had lost an entire chromosome arm were flow-sorted to purify the small remaining chromosome arm[19]. Sequencing of the chromosome-arm DNA enabled the precise allocation of most genes to the closely related A, B or D genomes[20].

There are several examples of the successful *de novo* sequencing and assembly of large allopolyploid genomes of crops that use long-range alignments of sequence scaffolds to generate extended haplotypes that form distinctive homeologous pseudomolecules. Tobacco (*Nicotiana tabacum*; $2n = 4\times = 48$) is an allotetraploid that is derived from the diploid genomes of *N. sylvestris* and *N. tomentosiformis*. Whole-genome shotgun assemblies were aligned to physical maps to create long super scaffolds that could be assigned directly to the progenitor genomes[21]. Upland cotton (*Gossypium hirsutum*) is an allotetraploid that formed 1–2 Myr ago from two unknown diploid species. The genome complexity of upland cotton was reduced by sequencing allohaploid lines that were derived by pollen culture to a depth of coverage of 245× with Illumina short-read sequencing reads[22]. A dense genetic map was used to align and correct the scaffolds, which covered 96% of the estimated 2.5 Gb genome, and fluorescence *in situ* hybridization was used to confirm a successful allotetraploid assembly. The polyploid genome of Indian mustard (*B. juncea*) (Fig. 1) has been sequenced using a combination of Illumina short reads and PacBio single molecule, real-time long sequence reads that were aligned with optical maps from BioNano Genomics[23], which directly visualize individual molecules of tagged DNA, and dense genetic maps[24]. The genome was almost fully represented in the assembly, which was assigned to the 402 Mb A genome and the 547 Mb B genome.

BOX 1

Glossary

Terms that are used to describe genomic methods, genetic analyses and their application in crop genomic studies and improvement are briefly explained.

- **Allopolyploid** A polyploid that contains sets of chromosomes that are derived from two distinctive parents; compare with an autopolyploid, in which the chromosomes are derived from one parent.
- **Amphidiploid** A plant that has a complete set of diploid chromosomes from each parent.
- **Breeding programme** The targeted interbreeding of plant varieties and the selection of progeny with improved characteristics. Almost all crops have a long history of such genetic improvement.
- **Bread wheat** Triticum aestivum, a hexaploid that is composed of three sets of homeologous chromosomes from closely related grasses. Its grain produces a flour that when mixed with water forms a viscoelastic dough that is suitable for bread production.
- **Doubled haploid** In plant breeding, loci can be made homozygous by cultivating haploid anther cells and doubling their chromosome number by treatment with colchicine.
- **Emmer wheat** Triticum turgidum diccocum, or hulled wheat, is a tetraploid wheat that was domesticated in the Near East and is widely grown as a crop.
- **Genetic association** A method for associating genetic variation with a phenotype that uses linkage disequilibrium to identify sets of sequence polymorphisms that statistically co-occur with a given phenotype.
- **Genetic gain** The change in performance of breeding populations owing to the recurrent selection of plants with higher performance.
- **Haplotype** A co-inherited block of DNA that contains several genes with a defined order, or phasing, of sequence polymorphisms.
- **Homeologue** In polyploid cells, chromosomes that are derived from one parent are called homologues and they pair during meiosis. Homeologous chromosomes are related chromosomes from one parent that, in general, do not pair with chromosomes from the other parent.
- **Hybrid** The offspring of two distinctive parents. Hybrid offspring can have increased vigour in comparison to each parent.
- **Introgression** The introduction of a chromosome or part of a chromosome from one species into another by the production of an interspecific hybrid and repeated backcrossing to one parent to remove undesired chromosomes.
- **N50** A widely-used measure of the extent of sequence assembly. The value refers to the shortest sequence length at which 50% of the bases in the assembly are represented.
- **Near-isogenic line** Used to establish, through repeated backcrossing, a population of plants with a common genetic background that contains chromosomal regions from another plant that confer specific traits.
- **Polyploid** An organism that contains more than two pairs of homologous (pairing) chromosomes.
- **Pseudomolecule** An ordered concatenation of separate sequence scaffolds that forms a representation of the sequence of a chromosome or chromosome arm. In general, a high-density genetic map is used to achieve ordering but, more recently, chromatin proximity ligation sequence data or optical maps have been applied.
- **Quantitative trait locus** A region of a chromosome (locus) that is mapped to underlie a specific continuously varying trait. Several genes can be found at the locus, only one of which may confer the trait.
- **Scaffold** An oriented array of DNA sequence assemblies that are generated from sequence contigs that contain regions of unknown sequence but known length, represented as tracts of *n*.
- **Super scaffold** A set of scaffolds that is anchored into a longer-range order that is provided by, for example, an optical map or chromatin proximity ligation data.
- **Whole-genome duplication** The full duplication and stable inheritance of chromosomes in an organism. Leads to autopolyploidy and occurs frequently in plants.

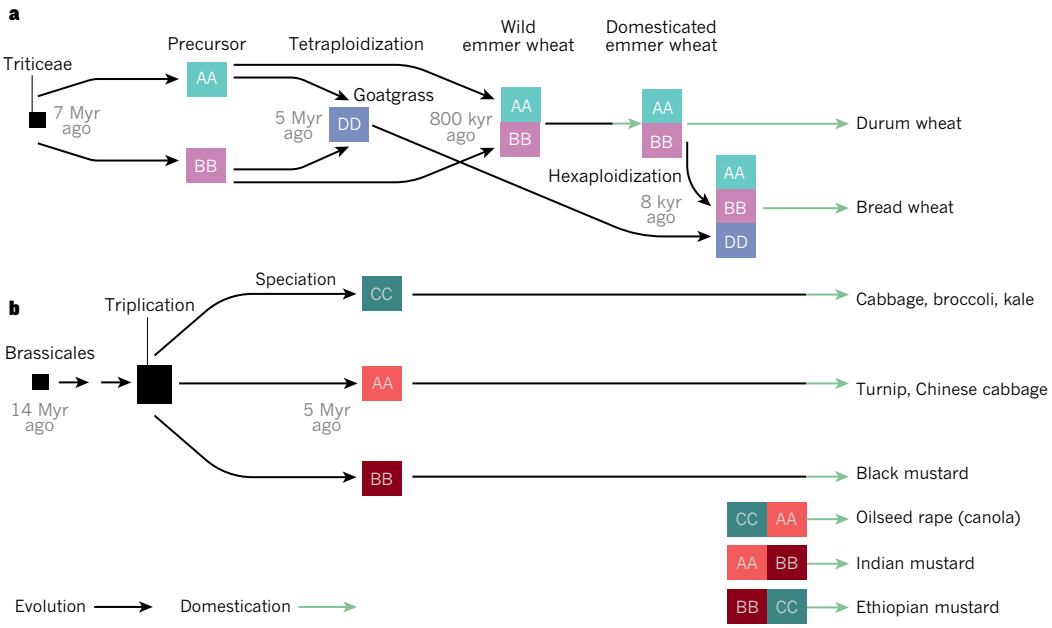

Figure 1 | Evolution and domestication of the common polyploid crops wheat and the genus Brassica. **a**, Roughly 7 Myr ago, an ancestral Triticeae gave rise to the AA and BB diploid precursor genomes of wheat. These formed the diploid precursor DD genome of goatgrass 5 Myr ago. Around 800 kyr ago, the tetraploid AABB genome of wild emmer wheat formed, which was domesticated as in the region that is now south-west Turkey. About 8 kyr ago, a random hexaploidization event occurred between domesticated emmer wheat and a wild goatgrass, which contributed its DD genome, that formed the contemporary bread wheat genome. **b**, About 14 Myr ago, an ancestral Brassicales underwent genome triplication; it then speciated to form the AA, BB and CC genomes around 5 Myr ago. The resulting species were domesticated relatively recently and also helped to form polyploid species that are important crops worldwide.

Accurate and long-read sequencing

Advances in sequencing technology are being used to resolve complex genomic regions and to determine copy number variants in the human genome (reviewed in ref. 25) that are directly relevant to resolving polyploid crop-genome assemblies and to identifying structural variants. The most important advance has been the ability to capture maximal levels of sequence variation in assemblies so that low levels of sequence polymorphism between closely related homeologues and repeats are represented in assemblies. Bias that is introduced by sequence template preparation can be reduced by making paired-end libraries without PCR amplification steps[26]. Variation can also be lost during the error-correction step of contig assembly. The DISCOVAR *de novo* genome assembler[27] captures a wider range of sequence variation during assembly by preserving sequence differences in the assembly graph. A high-quality *de novo* assembly of the mosquito genome was accomplished by the DISCOVAR assembly of a single paired-end PCR-free library[28]. Another important advance has been the increase in the length of sequence reads. Single molecule, real-time sequencing has dramatically improved the contiguity and repeat representation of mammalian[29] and large pine[30] and grass[31] genomes. This has proved useful for the production of higher-quality plant-genome assemblies and has been applied to the *de novo* assembly of smaller plant genomes such as *Arabidopsis* and the grape vine *Vitis vinifera* cv. Cabernet Sauvignon[32].

Nanopore single-molecule sequencing from Oxford Nanopore Technologies is an emerging method that could have considerable potential applications in crop genomics, including real-time selective sequencing, the identification of modified bases and in-field genomics analyses. However, this technology is at an early stage of development with respect to the sequencing of large genomes and has a fairly high rate of error. Initial nanopore-based genome assemblies used a hybrid approach with short-read Illumina sequences to correct such errors in nanopore yeast-genome assemblies[33,34]. More recent progress has been rapid: for example, the 29× coverage of the *Escherichia coli* genome that was achieved with the long-read nanopore sequencing method MinION was sufficient to generate a single 4.9 Mb contig of 99.5% accuracy[35].

Methods for linking sequences in genome assembly

A further requirement for the *de novo* assembly of polyploid plant genomes is the linking of sequence contigs into longer scaffolds and super-scaffolds while unambiguously separating very closely related chromosomes or haplotypes. Several complementary mapping approaches have been applied to the ordering of genome assemblies into longer super-scaffolds. Proximity ligation is a method that physically links segments of DNA that are in close proximity in chromatin by chemical cross-linking *in vivo*, which is followed by ligation of the adjacent DNA sequences and sequencing of the ligated regions. The proximity-based ligation of nuclear-chromatin preparations generates contact probability maps that link DNA sequence reads across Mb scales while identifying various chromosomal compartments and interactions[36,37]. The HaploSeq method used proximity ligation coupled to low-coverage (17×) Illumina sequencing to generate extended haplotypes of the human genome[38]. The bacterial-artificial-chromosome-based sequence of the 5 Gb barley genome has recently been ordered into chromosome-scale super-scaffolds using *in vivo* chromatin proximity ligation[39]. *In vitro* reconstituted chromatin provided useful links across genomic distances of up to 500 kb and enabled very large scaffolds to be generated for the human and alligator genomes[40]. The 1.4 Gb genome of quinoa (*Chenopodium quinoa*), a nutritious grain, has also been assembled[41] into very long scaffolds (90% of the genome represented by 439 scaffolds) by combining long reads, *in vitro* chromatin links and optical maps from BioNano Genomics.

Linked-read sequencing technology from 10x Genomics uses a microfluidics device to partition long DNA molecules (greater than 50 kb) into individual gel beads that contain a unique barcode consisting of a short sequence of DNA, a sequencing adaptor and semirandom priming sequences. Hundreds of thousands of individual library constructions and PCR reactions on DNA molecules can then be performed, pooled, sequenced using Illumina methods and deconvoluted into linked-read sequences from long, single molecules of DNA. In essence, this approach integrates the throughput and accuracy of Illumina sequencing with multiplexed long single molecule sequencing. When applied to the human genome, in a hybrid approach that uses linked reads to order and align a standard Illumina assembly, a 12-fold

increase in scaffold N50 was achieved[42]. Linked-read technology has been used to generate contiguous 40–200 kb blocks of genomic variants in the correct order (phased) from trio DNA (collected from a mother, father and child), a cancer cell line and a primary tumour[43]. Inheritance patterns of genomic deletions and gene rearrangements were defined, and haplotype analyses found copy number variation in a mutation in a cancer-driver gene. Linked-read sequencing has also been used to generate separate haplotypes for diploid homologous chromosomes[44]. Similarly, long reads performed using the PacBio platform coupled to the fast alignment and consensus for assembly (FALCON)-Unzip genome assembler have been used to successfully define large-scale haplotype-specific assemblies of *Arabidopsis* trios and *Vitis vinifera* cv. Cabernet Sauvignon, which is a highly heterozygous F_1 hybrid[32].

Combined approaches to genome sequencing and assembly

The genomes of several commercially important polyploid crops could now be sequenced and assembled *de novo*, and a number of ongoing sequencing projects could be supplemented, by applying the genomics technologies we have described. Linked-read and single molecule, real-time technologies can generate extended haplotypes that are specific to homologous chromosomes, which efficiently reveal large-scale genomic variation such as deletions, translocations and chromosome additions while preserving low levels of sequence variation in conserved repeats and genes.

Such approaches are important because they will establish a compendium of the multiple types of genetic variation that are required for crop improvement. Examples include the octaploid genome of strawberry. A doubled-haploid line of the highly heterozygous diploid genome of the robusta coffee plant (*Coffea canephora*; $2n = 2\times = 22$) has been sequenced[45]. It is one of the parents of the elite coffee plant *C. arabica*, the tetraploid genome of which has yet to be sequenced. The banana (*Musa* spp.) has a particularly complex polyploid breeding system. A doubled-haploid line of banana (*M. acuminata*) has been sequenced[46], assembled and integrated into 11 linkage groups. Most modern cultivars are sterile triploids that are composed of AAA genomes (Cavendish bananas, *M. acuminata*) or AAB genomes (plantains, *M. aucuminata* × *M. balbisiana*). A hexaploid wheat whole-genome shotgun assembly has been built from short-insert (250 bp) paired-end Illumina reads and long-insert mate-pair libraries using a new variation-aware contig assembler called w2rap[47]. The sequence assemblies were accurately assigned to the A, B and D genomes, and almost 14 Gb of the 17 Gb genome was represented.

Sugar cane is vegetatively propagated from hybrids of the high-sugar variety *Saccharum officinarum* ($2n = 80\times = 10$) and its vigorous wild relative *S. spontaneum* ($2n = 40–128\times = 8$): *S. spontaneum* is backcrossed to *S. officinarum* and the offspring are selected for their sugar content. Modern cultivars have $2n = 100–130$ chromosomes, which are often recombined or lost. This exceptionally complex polyploid aneuploid genome can be thought of as being composed of multiple variant copies of the closely related *Sorghum* genome[48]. To assess genetic diversity between related chromosomes and to identify lines for optimal crosses, linked-read technologies could be applied to generate a number of haplotypes that are aligned to the *Sorghum* 'baseline' genome. This method would identify copy number variation and rearrangements as well as provide a foundation for the development of DNA sequence markers for selecting desired genomic regions that would help to achieve the yield potential of sugar cane, a crop of importance worldwide.

Forage grasses (*Lolium* spp. and *Festuca* spp.) are the foundation of dairy and meat production. Perennial ryegrass (*L. perenne*) is self-incompatible and is maintained as highly heterozygous diploid or tetraploid genomes. Current *Lolium* genome resources[49] come from a homozygous inbred line. The application of new sequencing technologies to highly heterozygous commercial diploid or tetraploid lines, as well as amphidiploid introgressed lines, will enable a much broader range of genetic variation to be identified for use in improving these crucial crops. Various methods of sequencing are available for use in the analysis of crop genomes (Fig. 2). The number of analyses that each method can complete varies, ranging from multiple high-throughput analyses to a limited number of whole-genome assemblies, as does the size of the sequences that each generates.

Domestication erodes genetic diversity

Widespread archaeological evidence has identified relationships between crop domestication and the settled agriculture that formed part of the Neolithic revolution[50]. The domestication of species occurred at the geographical centres of their genetic diversity, an observation made first by Nikolai Vavilov[41,51]. Examples of this include: maize and squashes in Central America; potatoes and tomatoes in the Andean region of South America; coffee in Yemen and Ethiopia; oil palm and millets in West Africa; rice, citrus fruits and sugar cane in southeast Asia; and barley and wheat in an arc called the Fertile Crescent that extends from Israel to northern Iran and central Asia.

Crop domestication involved the identification of species of plants with desirable traits, ~~which was~~ followed by the deliberate cultivation and progressive selection, often over thousands of years, of a limited range of plant varieties for improved adaptation to settled cultivation, growing conditions and the needs of people (Fig. 3a). Recurrent selection has increased the frequency of desirable traits such as improved management (for example, harvest traits such as non-shattering grain, which keeps the seed in the flower spike so that it is not lost) and enriched characteristics such as greater nutritional value, flavour and yield; it has also reduced the frequency of undesirable traits such as toxicity, low palatability and poor adaptation to ~~the~~ agricultural environments. Genome analyses of rice, maize and wheat, which are globally important grass crops, demonstrate how rare genetic variants have been selected and established in populations by domestication.

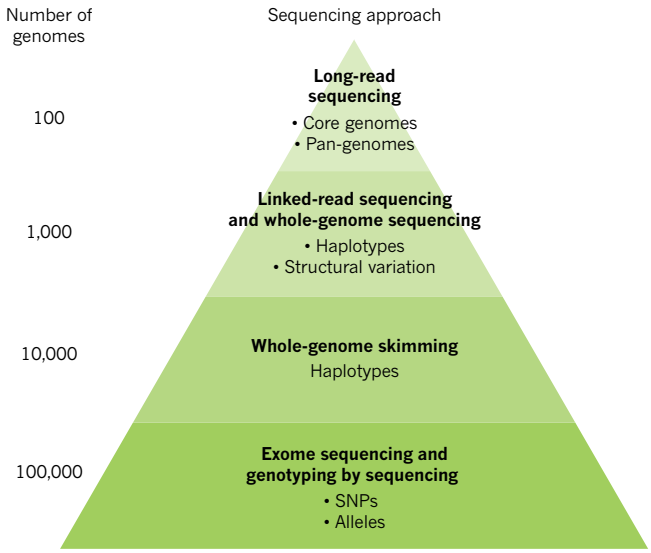

Figure 2 | Optimal sequencing systems for crop applications. A variety of sequencing methods are now available for different applications in crop impr[illegible]ent (green pyramid). The number of genomes that can be sequenced ~~cost-effectiveness~~ varies according to the method applied (left). Long-read technologies coupled with Illumina assemblies are providing accurate long-range assemblies of hundreds of complete genomes. These are used to define comprehensively the range and types of variation that are found in the genomes of a species (the pan-genome). Linked reads, either alone or coupled to Illumina sequencing, will provide cost-effective capacity for thousands of genome sequences. This can identify a wide range of useful variation, including large-scale structural variation. Skim sequencing consists of low-coverage (for example, 5–10×) Illumina reads and presents a cost-effective way of identifying genetic variation and haplotypes in populations. Exome sequencing uses sets of oligonucleotides to capture gene-coding regions, and genotyping by sequencing typically involves the sequencing of about 100–150 bases from a randomly located restriction-enzyme cleavage site in the genome.

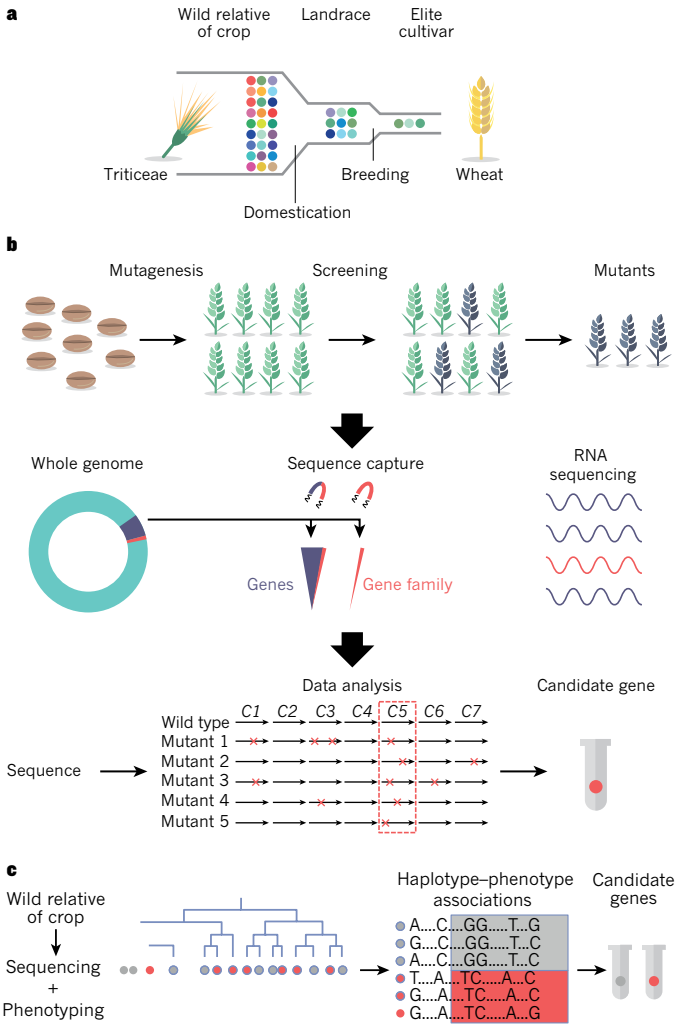

Figure 3 | Erosion of genetic diversity in cultivated crops and its re-incorporation through genomics. **a**, Genetic diversity (coloured circles) in populations of wild precursors of crops has been eroded by domestication, in which a limited range of diversity is present in landraces that were initially selected and adopted for cultivation. Subsequent breeding has drawn on only a limited range of the variation present in landraces to produce the elite cultivars that are used in modern agriculture. **b**, The identification of genes for crop improvement can use mutagenesis to introduce changes into the DNA of crops (top). Mutants with desired characteristics can be identified by screening for desired properties, known as phenotypes. In practice, this method is time-consuming and imprecise unless a specific phenotype can be measured in large populations. Mutant lines with desired phenotypes are pooled and sequenced (middle). Genomics can accelerate the process of identifying mutants by sequencing populations of mutant crops (or a range of wild relatives). Sequencing can be targeted to all genes, or specific families of genes, using sequence capture methods. RNA can also be sequenced to identify changes in gene expression that are caused by mutagenesis. Sequences of mutant lines are then compared to identify genes that are consistently mutated in the lines that exhibit the desired phenotype (bottom). **c**, Genomics can also be used to access genetic variation in populations of crop wild relatives. A population can be sequenced using a variety of approaches (described in Fig. 2). At the same time, the population is screened for a range of phenotypes of interest. Patterns of sequence variation, or haplotypes, can be associated with phenotypes to identify sequence variation that may cause the phenotype.

Archaeological evidence has shown the existence of non-shattering rice forms in the lower part of the Yangtze river basin in China between 7,000–8,000 years ago, which highlights the length of the duration of rice domestication[52]. Genome analyses of wild and domesticated populations of rice identified a dramatic loss in genetic variation that occurred on the domestication of japonica rice (*Oryza sativa* var. japonica) from wild populations[53]. About 55 domestication-related loci are known, including the genes *sh4* (seed shattering), *PROG1* (tiller angle, to maintain an upright posture during growth), *Waxy* (grain starch quality) and *qSW5* (grain width). Maize was domesticated from a wild teosinte grass that grew in Mexico roughly 9,000 years ago. Among several domestication genes, selection for the increased expression of *tb1*, which encodes a TCP-class transcription factor, reduces branching and selection for a single amino-acid change in *tga1*, which encodes another transcription factor, reduces the hard casing around kernels to improve processing[54]. The domestication of maize for growth in temperate climates has selected for variations in flowering time, enabling growth to adapt to higher latitudes. Several genetic loci have been selected, each with a relatively small effect on the flowering time, which demonstrates another important pattern of genomic change during domestication[55]. Analysis of genetic variation in modern maize, early domesticated maize[56] and wild populations of teosinte identified almost 1,200 genes that have been affected by domestication, as well as a reduction in genetic diversity such that modern maize lines have only about 57% of the diversity of progenitor populations[57]. Bread wheat arose from a chance hybridization between a cultivated tetraploid emmer wheat (*Triticum turgidum*) and a wild diploid goatgrass (*Aegilops tauschii*) (Fig. 1) in the south Caspian basin, where extant populations of progenitor species co-existed. This formed a stable hexaploid spelt (hulled) wheat, from which a free-threshing derivative was derived by selection. New characteristics such as elastic dough for bread-making were introduced from the *A. tauschii* genome, with gene dosage effects and pseudogenization contributing to polyploid-specific traits[58,59]. Genetic diversity in bread wheat was constrained during the domestication of emmer wheat and by the small number of individual plants that contributed to the formation of the new hexaploid bread-wheat species[60].

Accessing genetic diversity

The cumulative effect of thousands of years of the domestication, industrialised breeding and global trade of crops is that almost all people now depend on a few dozen species for the bulk of our nutrition and that of our domesticated animals[61]. Three approaches have emerged that address this troubling reliance on a few crops with limited genetic variation. The first is to diversify food sources through the domestication and improvement of alternative species: for example, quinoa and teff (*Eragrostis tef*), which are two grains with a high nutritional value and relatively low agronomic demands[62]. The second involves increasing the genetic diversity of elite crops: for example, by the inclusion of new traits such as disease resistance. This draws on the extensive genetic diversity of wild relatives of crops, which have evolved over long timescales and adapted to a wide range of environments. Such diversity is essential for improving many traits, including crop performance in challenging environments that are the result of water stress, high temperatures or the presence of pathogenic agents[63]. However, this irreplaceable endowment of wild species is being lost through urbanization and myriad other forms of environmental degradation. Consequently, the third solution is to halt the loss of populations of wild relatives of crops to conserve a wide range of the remaining variation.

Depending on the ease of genetic exchange, undomesticated relatives of several crops have been used to increase genetic diversity. In wheat, new polyploids made by crossing different goatgrass variants with tetraploid wheats have captured a wide range of useful genetic variation. These 'synthetic' wheats are crossed into elite wheat lines and then backcrossed to fix desired phenotypes[64]. A durum wheat (pasta wheat; *Triticum durum*) line that maintains high yields when grown in saline soil was created by introgressing a region of the einkorn wheat (*T. monococcum*) genome (a diploid progenitor of tetraploid durum wheat and hexaploid bread wheat) that contains a sodium exclusion pump[65]. In diploid species of crop, problems with compatibility and stability can limit the range of varieties that can be used, and extensive backcrossing is required to remove undesired allelic variation. Pre-breeding strategies are therefore often used to create populations, including near isogenic

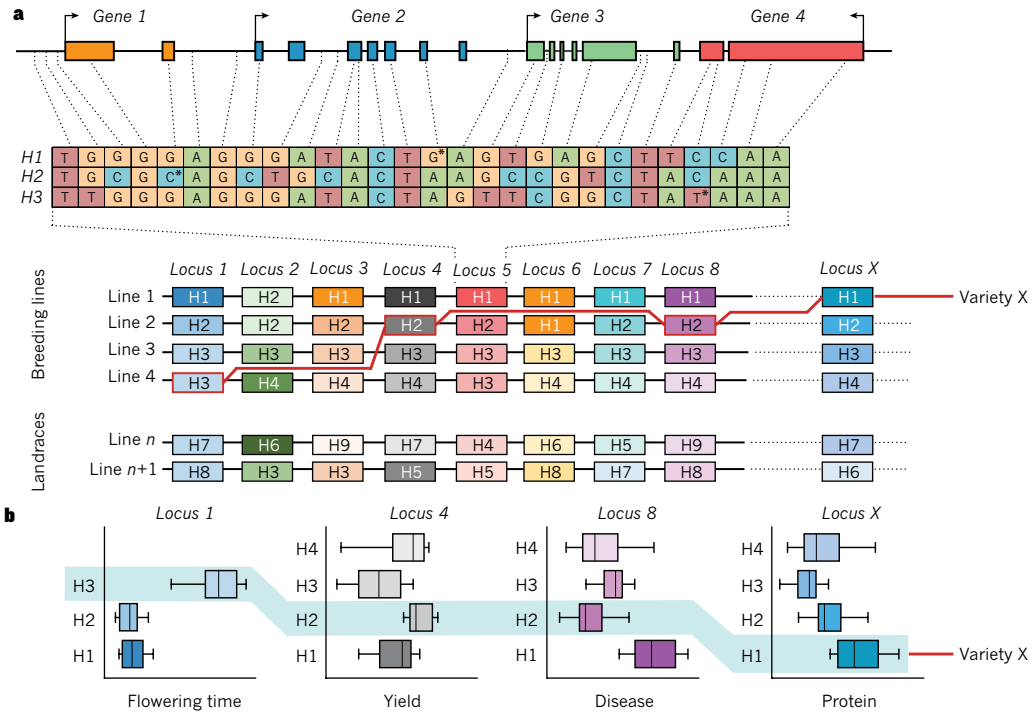

Figure 4 | The assembly of haplotypes in a crop-breeding programme. **a**, An example of a genomic region that consists of four genes and contains genetic variation that defines three haplotypes (H1, H2 and H3) at a particular locus (locus 5) on a chromosome. The position of the SNP that defines each haplotype is marked by an asterisk. An array of haplotypes (H1–H4) from the same chromosome, with the variants of four breeding lines (line 1, line 2, line 3 and line 4) aligned underneath each locus, is also shown. Line *n* and line *n*+1 are landraces (domesticated lines) that can introduce new haplotypes (H5–H9) and genetic diversity. The genomic structure, diversity and functions of haplotypes are established by the re-sequencing of lines and the analysis of quantitative trait loci. The red line traces the assembly of a new line (variety X) from component haplotypes, using markers that are specific for the haplotypes in each line, that have been chosen on the basis of desired combinations of phenotypes that are expressed by each haplotype. **b**, The performance of various haplotypes in lines 1–4 is determined in different environments, often under field conditions and over several years, using specific assays. Examples are shown for the variation in performance of four common plant traits that are influenced by genetic variation at locus 1 (time to flower), locus 4 (yield), locus 8 (resistance to disease) and locus X (protein content), with the performance of variety X highlighted in light blue.

lines, for the characterization of quantitative trait loci[66]. In rice, many useful quantitative trait loci have been identified through crosses with the wild rice *O. rufipogon* with *O. sativa*[67], after which they are introduced into elite germplasm. Genetic variation in wild teosinte maize lines[68] is being characterized by crossing several wild relatives with an elite inbred line and repeated backcrossing to form near isogenic lines. Quantitative trait loci are mapped in the near isogenic lines and DNA sequence markers are used to track the incorporation of DNA segments that underlie these loci into maize breeding lines.

Affordable, high-throughput sequencing technologies are facilitating innovative approaches to identifying and using genetic variation from the wild relatives of crops. For genes that are annotated with well-defined functions and that have highly distinctive sequences, such as those that encode plant immune receptors, genomic DNA can be captured using gene-family-specific oligonucleotide probes, then sequenced and assembled to provide a near-complete representation of a gene family in an individual line or population (Fig. 3b). This approach was used to identify a genetic locus that underlies resistance to late blight in *Solanum americanum*, a wild relative of the potato, using a combination of genetic mapping and resistance-gene enrichment sequencing (RenSeq)[69]. RenSeq is has now been used to isolate the resistance (*R*) gene repertoire from populations of wild relatives as well as sequence variation that is associated with phenotypes of disease resistance. In cases in which there was no prior knowledge of the *R* gene specificity, a modified version of RenSeq called MutRenSeq was used to isolate *R* genes from wheat that is resistant to the fungus that causes stem rust and from a number of independently-derived mutant plants that had lost resistance to stem rust and showed symptoms of infection[70]. Comparison of the mutant and wild-type assemblies of the captured DNA identified mutations in *R*-gene-containing contigs from all of the mutant symptomatic lines (Fig. 3b). In a genome-wide approach, tetraploid and hexaploid populations of wheat underwent exome sequencing to identify variation in more than 80% of protein-coding genes, which generated a unique and powerful resource for forward and reverse genetics[71]. In commercial breeding programmes, such mutant populations are coupled with genomics to identify genetic variation that is associated with important traits. Clustered regularly interspaced short palindromic repeat (CRISPR)–Cas9 technologies are being applied to crops for genome editing. The delivery of Cas9 and DNA templates to crop genomes can be carried out transiently[72] or using a viral vector[73], without an intermediate transgenic step. This challenges the present definitions of genetic engineering according to European Union regulations.

Sequencing methods — in particular, long-read and linked-read sequencing, whole genome skimming and exome sequencing — promise to increase the throughput and reduce the cost of ~~de~~ *de novo* genome assembly of hundreds of crop relatives, as well as mutant populations of crops and their relatives. This will enable the identification of a broad spectrum of genetic diversity that can be used to test for the association of variation with important traits. One such example is shown in Fig. 3c, in which a panel of plants that contain extensive genetic diversity (for example, the wild relatives of a crop) is sequenced using methods shown in Fig. 2 and also phenotyped. Genetic analyses can therefore be used to identify genetic variation that is associated with phenotypic variation of functional relevance, as well as markers that can be used in crop breeding.

Systems for crop breeding

Mendelian principles were first applied to crop improvement in the early twentieth century[74], and the framework that was established continues to be used in modern pedigree crop breeding. It involves the selection of plants with desired phenotypes that have high heritability. Selection

requires the manual inspection and specific analyses of phenotypes in many plants across a number of generations. In general, large populations are needed to bring combinations of loci that encode desired traits to homozygosity and several generations (often more than six) are required. It can therefore take more than 10 years to bring a new variety of crop to market, and there is considerable uncertainty about predicting the effects of combining phenotypes, especially those of low heritability. DNA technology such as DNA sequence markers and genome sequencing is now widely applied to improve the efficiency of breeding. Large numbers of markers for DNA sequence polymorphisms are available for use in breeding programmes; these include single nucleotide polymorphisms that can be assayed in high-throughput modes such as the fixed-content, high-density Axiom and iSelect genotyping arrays (described for wheat in ref. 75). Marker-assisted selection is used extensively in breeding programmes to monitor genomic loci that are linked to markers in breeding pedigrees and to assemble combinations of loci[76]. An approach called genomic selection is also being used to accelerate breeding programmes[77,78]. In this method, a test population that represents the genetic diversity of a larger breeding population is thoroughly genotyped and phenotyped and a breeding value that predicts phenotypic performance on the basis of marker frequencies is assigned. The larger breeding population is then genotyped and the breeding value is used to predict the phenotypes of lines in the population. By reducing the cost of and time spent on phenotyping, and by incorporating rapid generation times (known as 'speedy' breeding[79]), the amount of time that it takes to increase genetic gain is being reduced[80].

Crop breeding as a DNA-assembly problem

Genomic technologies now facilitate the rapid and cost-effective assembly of very large polyploid crop genomes, the analysis of large populations of crop wild relatives and the collation of excellent functional genomics resources for some crops, all of which enable efficient gene characterization. These capabilities can identify the genetic variation and genes that have been used by breeders for crop improvement and help to understand how genetic variation influences phenotypes, as well as accessing a wider range of genetic variation in wild relatives of crops.

Breeding programmes use recombination to integrate desired combinations of traits to form improved varieties. Conventionally, the heritability patterns of phenotypes are the main method by which genetic combinations are assessed. The identification of quantitative trait loci in both elite lines and their wild relatives helps to define genomic loci that usually contain several candidate genes and many sources of genetic variation that either do not contribute to the desired phenotype or have deleterious effects. These 'compound-effect' loci can then be incorporated into breeding programmes through marker-assisted selection. Combinations of genomic regions that confer desired traits can be considered as sets of haplotypes that are defined by underlying genetic variation. Therefore, genomic information aids breeding by defining desired combinations of haplotypes for selection in breeding programmes.

Two approaches, one retrospective and the other prospective, hold promise for the integration of genomics and breeding using haplotype selection. The retrospective approach aims to identify the haplotypes that have been used by plant breeders. It involves the sequencing and assembly of key breeding lines that have been used widely over an extended period and the re-sequencing of lines in pedigrees that were selected. As these pedigrees have been phenotyped extensively in multiple environments, the genomic regions (and the genetic variation and haplotypes at these regions) that are associated with the decisions of breeders will be revealed. These haplotypes, including the genetic variation that they carry and their phased markers, can then be used in three ways. First, the biological functions of the region of DNA that contains the haplotype (haplotig[32]) can be established from systematic genomic analyses, including studies of gene function and gene networks, expression patterns, chromatin structure[81], epigenetic modifications and the influence of genetic variation on gene function. Genetic variation that causes desired phenotypes, as well as any deleterious variation, can be identified. Second, the markers that define the haplotig can be combined with those for other functional haplotigs to form a genome-wide set of markers that can be used a priori in breeding programmes to select new combinations of haplotigs, each with well-defined phenotypic effects. Last, the haplotig-defining markers can also be used to identify lines in which linkage in the haplotig has been broken to separate desirable and undesirable genetic variation (Fig. 4). This 'haplo–genomic' approach uses specific functional and genetic relationships that are defined by genome-wide haplotypes. It also differs from genomic selection, which uses genome-wide markers as anonymous features that are statistically related to phenotypes in only a subset of the breeding population[82]. Spindel *et al.*[83] describe how genome-selection approaches can be integrated successfully with association studies to accelerate genetic gain in rice-breeding programmes, which indicates that the integration of genome-wide haplotype information may provide breeding programmes with an even greater specificity and predictability.

A prospective approach to haplotype-led breeding involves the sequencing of populations of the progenitors and wild relatives of crops to identify conserved ancestral haplotigs that contain a broader range of genetic variation. This approach is demonstrated by the identification of new loci that are associated with tolerance of salinity using low-coverage sequencing of 106 diverse soya bean (*Glycine max*) lines[84] and by exome sequencing from several wheat lines[85]. Genetically structured populations, genetic-association analyses and targeted sequencing can be used to define new haplotypes with greater variation, fewer deleterious alleles and improved phenotypes for incorporation into breeding programmes[68]. Markers that define the haplotype can then be used to track its incorporation into breeding pedigrees.

The process of haplotype assembly (Fig. 4) during breeding can be followed using sets of genetic markers that define the haplotypes under selection. To assemble multiple haplotypes, large numbers of F_2 progeny (more than 10,000) will need to be screened[77], with several markers that span each haplotype for phasing and for selecting desired variants of each haplotype. Because current marker technologies are too expensive for use in this approach, innovative methods will need to be developed. Multiple genome assemblies of the parental lines and other genomic resources will reveal the underlying order of markers and sequence variation in each haplotype. The throughput of technologies such as the Illumina HiSeq 4000 and HiSeq X Ten platforms is such that a single lane can generate 110 Gb of 150-bp sequence reads in 3 days. This equates to the sequencing of 10,000 loci in 10,000 individuals to a depth of about 1000× at a cost of less than US$1 per sample. Multiplexed primer-based assay technologies have the potential to deconvolve the sequencing read data to count and map haplotype assemblies in several lines[86]. Complementary approaches such as linked-read[43] sequencing and long-read sequencing[35] also show promise for the large-scale genetic analyses of populations in genomics-led breeding strategies.

Progress towards the application of genomics to breeding is also revealing the potential limitations of current genomic technology. In the large genomes of maize, wheat and many other important crops, recombination is highly suppressed across extensive pericentromeric regions owing to the methylation of repeat regions[87]. These regions contain a large number of functional genes, which are therefore not accessible to breeding programmes. It may be feasible to target double-strand breaks to specific regions of chromosomes and then use CRISPR–Cas9 technology to stimulate recombination. Although mitotic recombination can be targeted precisely in yeast cells, and the fine-mapping of loci is also possible[88], it is difficult to imagine using this approach in multicellular organisms such as plants. It may be possible to use a promoter that is specific to prophase 1 of meiosis or an inducible promoter to express the gene that encodes Cas9 to guide RNAs to create new, targeted patterns of double-strand breaks in the germ line and to promote recombination in new regions of interest during normal meiosis.

Automated phenotyping

Innovative specialized technologies for phenotyping are improving

the resolution, precision and scale of measurements of crop growth and development[89,90]. These approaches can also measure important physiological characteristics, including water use and photosynthetic efficiency, in the field[91]. It will be necessary to complement genomics-led breeding with an increase the scale and precision of phenotyping using drones and networks of sensors, as this will facilitate the genetic analysis of much larger populations in breeding programmes and generate multidimensional data sets of genotype–environment interactions. To meet this goal, networks of mobile and static sensors for measuring crop performance and the environment need to be established at the farm scale. Once developed, this facility can be used for standardizing, capturing and maintaining crop phenotype data[92] and for integrating the data with environmental and genetic variation data. Such an approach will help to meet the important international goal of creating more resilient crops that maintain high yields with lower agronomic inputs, and that are better able to adapt to pests, diseases and fluctuations in climate such as variations in temperature or the water availability of the soil.

Outlook

Progress in genomics technologies is now enabling the rapid and cost-effective sequencing and assembly of the largest and most complex plant genomes. Researchers can access and characterize a vast reservoir of natural genetic variation from wild or undomesticated relatives of crops. The application of improved short-read sequencing and genome assembly will continue to provide the most accurate and cost-effective solutions for the *de novo* assembly of larger plant genomes and accessing genetic diversity. However, it is clear that sequencing technologies that involve longer reads, including single molecule, real-time sequencing, and linked-read sequencing on long molecules will have a major impact by improving sequence assembly and perhaps by even supplanting the short-read sequencing of crop genomes if increases in accuracy can be maintained. Population genomics is now feasible for large populations of plants with complex genomes through exome sequencing. It is also plausible that real-time selective sequencing using the long-read nanopore device MinION[93] will be applied to the identification of both genetic and epigenetic variation[94,95] in any gene family in populations of plants. Linked-read technologies are ready to be applied to large-scale haplotype analyses such as the identification of precise regions of introgressed chromosomes and assemblies of desired haplotypes in breeding pedigrees.

Despite these important advances in genomics, the continued development of sequencing technology and computational methods is needed to improve the cost-effectiveness, quality and coverage of genome assemblies of multiple plant genomes. Studies that include the re-sequencing of populations of wild and domesticated lines will also be needed to better characterize population-level genetic variation and its association with relevant phenotypic traits.

To meet goals for food security, a future challenge will be to develop innovative genomic and bioinformatic applications and apply these to the high-throughput identification of genes for traits and for breeding systems. Interactions between the genomes of hybrids and polyploids that give rise to new and improved traits can now be understood at the genome level, and this information has been used to select parent plants with improved potential for yield increases in hybrid breeding systems. The breeding process, which has already been accelerated substantially by the application of genomics, is now poised to undergo a step change in the specificity and precision of selecting parents, identifying haplotypes, screening progeny for desired combinations of haplotypes and predicting performance on the basis of genetic information. This progress is timely because it has a central role in securing food supplies for the future. ■

Received 2 December 2016; accepted 1 February 2017.

1. Drèze, J. & Sen, A. K. *Hunger and Public Action* (Clarendon, 1989).
2. Foley, J. A. *et al.* Solutions for a cultivated planet. *Nature* **478,** 337–342 (2011).
3. Kitano, M. *et al.* Ammonia synthesis using a stable electride as an electron donor and reversible hydrogen store. *Nature Chem.* **4,** 934–940 (2012).
4. Zhao, C. *et al.* Plausible rice yield losses under future climate warming. *Nature Plants* **3,** 16202 (2016).
5. Garrett, K. A., Dendy, S. P., Frank, E. E., Rouse, M. N. & Travers, S. E. Climate change effects on plant disease: genomes to ecosystems. *Annu. Rev. Phytopathol.* **44,** 489–509 (2006).
6. Godfray, H. C. J. *et al.* Food security: the challenge of feeding 9 billion people. *Science* **327,** 812–818 (2010).
7. Tilman, D., Balzer, C., Hill, J. & Befort, B. L. Global food demand and the sustainable intensification of agriculture. *Proc. Natl Acad. Sci. USA* **108,** 20260–20264 (2011).
8. Bennett, M. D. & Leitch, I. J. Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. *Ann. Bot.* **107,** 467–590 (2011).
9. Bennetzen, J. L., Ma, J. & Devos, K. M. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* **95,** 127–132 (2005).
10. Lisch, D. How important are transposons for plant evolution? *Nature Rev. Genet.* **14,** 49–61 (2012).
11. Kim, M. Y. & Zilberman, D. DNA methylation as a system of plant genomic immunity. *Trends Plant Sci.* **19,** 320–326 (2014).
12. Jiao, Y. *et al.* Ancestral polyploidy in seed plants and angiosperms. *Nature* **473,** 97–100 (2012).
13. Woodhouse, M. R. *et al.* Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biol.* **8,** e1000409 (2010).
14. Neale, D. B. *et al.* Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* **15,** R59 (2014).
15. Shulaev, V. *et al.* The genome of woodland strawberry (*Fragaria vesca*). *Nature Genet.* **43,** 109–116 (2010).
16. Chalhoub, B. *et al.* Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* **345,** 950–953 (2014).
17. Bertioli, D. J. *et al.* The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nature Genet.* **48,** 438–446 (2016).
18. Voskoboynik, A. *et al.* The genome sequence of the colonial chordate, *Botryllus schlosseri*. *eLife* **2,** e00569 (2013).
19. Safar, J. *et al.* Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J.* **39,** 960–968 (2004).
20. International Wheat Genome Sequencing Consortium (IWGSC). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345,** 1251788 (2014).
21. Sierro, N. *et al.* The tobacco genome sequence and its comparison with those of tomato and potato. *Nature Commun.* **5,** 3833 (2014).
22. Zhang, T. *et al.* Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nature Biotechnol.* **33,** 531–537 (2015).
23. Staňková, H. *et al.* BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. *Plant Biotechnol. J.* **14,** 1523–1531 (2016).
24. Yang, J. *et al.* The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. *Nature Genet.* **48,** 1225–1232 (2016).
25. Chaisson, M. J. P., Wilson, R. K. & Eichler, E. E. Genetic variation and the *de novo* assembly of human genomes. *Nature Rev. Genet.* **16,** 627–640 (2015).
26. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods* **6,** 291–295 (2009).
27. Weisenfeld, N. I. *et al.* Comprehensive variation discovery in single human genomes. *Nature Genet.* **46,** 1350–1355 (2014).
This paper highlights the development and application of the DISCOVAR assembler, which has been of crucial importance for the creation of assemblies with improved representation of sequence variants.
28. Love, R. R., Weisenfeld, N. I., Jaffe, D. B., Besansky, N. J. & Neafsey, D. E. Evaluation of DISCOVAR *de novo* using a mosquito sample for cost-effective short-read genome assembly. *BMC Genomics* **17,** 187 (2016).
29. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods* **12,** 780–786 (2015).
This article shows how the application of hybrid assembly methods has set new standards for sequence contiguity and the representation of diversity.
30. Zimin, A. *et al.* Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics* **196,** 875–890 (2014).
31. Zimin, A. V. *et al.* Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the mega-reads algorithm. *Genome Res.* http://dx.doi.org/10.1101/gr.213405.116 (2017).
This paper shows how long-read sequencing technology coupled with the mega-reads algorithm can be used successfully to tackle a large and complex grass genome, which paves the way for the sequencing of multiple variants.
32. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* **13,** 1050–1054 (2016).
This study applies the PacBio long-read sequencing technology to resolving the highly heterozygous *Vitis vinifera* cv. Cabernet Sauvignon genome and demonstrates the importance of this technology for the assembly of complex plant genomes.
33. Koren, S. *et al.* Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nature Biotechnol.* **30,** 693–700 (2012).
34. Goodwin, S. *et al.* Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. *Genome Res.* **25,** 1750–1756 (2015).
35. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nature Methods* **12,** 733–735 (2015).

Refs 34 and 35 highlight the potential of nanopore sequencing technology, using yeast and bacterial genomes.
36. Burton, J. N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnol.* **31,** 1119–1125 (2013).
37. Session, A. M. *et al.* Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* **538,** 336–343 (2016).
38. Selvaraj, S., Dixon, J. R., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotechnol.* **31,** 1111–1118 (2013).
39. *Nature* (in the press).
40. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res.* **26,** 342–350 (2016).
41. Jarvis, D. E. *et al.* The genome of *Chenopodium quinoa*. *Nature* **542,** 307–312 (2017).
42. Mostovoy, Y. *et al.* A hybrid approach for *de novo* human genome sequence assembly and phasing. *Nature Methods* **13,** 587–590 (2016).
43. Zheng, G. X. Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnol.* **34,** 303–311 (2016).
As well as refs 42 and 44, this paper shows the considerable potential of linked-read sequencing technology for resolving the phasing of complete chromosomes.
44. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. & Jaffe, D. B. Direct determination of diploid genome sequences. Preprint at http://biorxiv.org/content/early/2016/08/19/070425 (2016).
45. Denoeud, F. *et al.* The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345,** 1181–1184 (2014).
46. D'Hont, A. *et al.* The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488,** 213–217 (2012).
47. Clavijo, B. J. *et al.* An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Preprint at http://biorxiv.org/content/early/2016/11/04/080796 (2016).
This preprint presents open-source assembly methods that preserve genetic variation and have enabled the fast and low-cost assembly of the large and complex wheat genome.
48. Grivet, L. & Arruda, P. Sugarcane genomics: depicting the complex genome of an important tropical crop. *Curr. Opin. Plant Biol.* **5,** 122–127 (2001).
49. Byrne, S. L. *et al.* A synteny-based draft genome sequence of the forage grass *Lolium perenne*. *Plant J.* **84,** 816–826 (2015).
50. Bilgic, H., Hakki, E. E., Pandey, A., Khan, M. K. & Akkaya, M. S. Ancient DNA from 8400 year-old Çatalhöyük wheat: implications for the origin of Neolithic agriculture. *PLoS ONE* **11,** e0151974 (2016).
51. Dvorak, J., Luo, M.-C. & Akhunov, E. D. N. I. Vavilov's theory of centres of diversity in the light of current understanding of wheat diversity, domestication and evolution. *Czech J. Genet. Plant Breed.* **47,** S20–S27 (2011).
52. Zheng, Y., Crawford, G. W., Jiang, L. & Chen, X. Rice domestication revealed by reduced shattering of archaeological rice from the lower Yangtze valley. *Sci. Rep.* **6,** 28136 (2016).
53. Huang, X. *et al.* A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
54. Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127,** 1309–1321 (2006).
55. Buckler, E. S. *et al.* The genetic architecture of maize flowering time. *Science* **325,** 714–718 (2009).
56. Ramos-Madrigal, J. *et al.* Genome sequence of a 5,310-year-old maize cob provides insights into the early stages of maize domestication. *Curr. Biol.* **26,** 3195–3201 (2016).
57. Wright, S. I. *et al.* The effects of artificial selection on the maize genome. *Science* **308,** 1310–1314 (2005).
58. Galili, G., Levy, A. A. & Feldman, M. Gene-dosage compensation of endosperm proteins in hexaploid wheat *Triticum aestivum*. *Proc. Natl Acad. Sci. USA* **83,** 6524–6528 (1986).
59. Zhang, Z. *et al.* Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. *Proc. Natl Acad. Sci. USA* **108,** 18737–18742 (2011).
60. Dubcovsky, J. & Dvorak, J. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* **316,** 1862–1866 (2007).
61. Khoury, C. K. *et al.* Increasing homogeneity in global food supplies and the implications for food security. *Proc. Natl Acad. Sci. USA* **111,** 4001–4006 (2014).
62. Massawe, F., Mayes, S. & Cheng, A. Crop diversity: an unexploited treasure trove for food security. *Trends Plant Sci.* **21,** 365–368 (2016).
63. Tanksley, S. D. Seed banks and molecular maps: unlocking genetic potential from the wild. Science 277, 1063–1066 (1997).
64. Jafarzadeh, J. *et al.* Breeding value of primary synthetic wheat genotypes for grain yield. *PLoS ONE* **11,** e0162860 (2016).
65. Munns, R. *et al.* Wheat grain yield on saline soils is improved by an ancestral Na^+ transporter gene. *Nature Biotechnol.* **30,** 360–364 (2012).
66. Borrill, P., Adamski, N. & Uauy, C. Genomics as the key to unlocking the polyploid potential of wheat. *New Phytol.* **208,** 1008–1022 (2015).
67. McCouch, S. R. *et al.* Through the genetic bottleneck: *O. rufipogon* as a source of trait-enhancing alleles for *O. sativa*. *Euphytica* **154,** 317–339 (2006).
68. Liu, Z. *et al.* Expanding maize genetic resources with predomestication alleles: maize–teosinte introgression populations. *Plant Genome* http://dx.doi.org/10.3835/plantgenome2015.07.0053 (2016).
69. Witek, K. *et al.* Accelerated cloning of a potato late blight-resistance gene using RenSeq and SMRT sequencing. *Nature Biotechnol.* **34,** 656–660 (2016).
70. Steuernagel, B. *et al.* Rapid cloning of disease-resistance genes in plants using mutagenesis and sequence capture. *Nature Biotechnol.* **34,** 652–655 (2016).
This paper highlights a method with great promise for capturing the diversity of large genes families in populations of crops and their wild relatives.
71. Krasileva, K. V. *et al.* Uncovering hidden variation in polyploid wheat genomes. *Proc. Natl Acad. Sci. USA* **114,** E913–E921 (2017).
This paper describes functional genome resources that have been developed for tetraploid and hexaploid wheat lines — resources that will expedite many new areas of research.
72. Zhang, Y. *et al.* Efficient and transgene-free genome editing in wheat through transient expression of CRISPR/Cas9 DNA or RNA. *Nature Commun.* **7,** 12617 (2016).
73. Gil-Humanes, J. *et al.* High efficiency gene targeting in hexaploid wheat using DNA replicons and CRISPR/Cas9. *Plant J.* http://dx.doi.org/10.1111/tpj.13446 (2016).
74. Biffen, R. H. & Engledow, F. L. *Wheat-Breeding Investigations at the Plant Breeding Institute, Cambridge* (His Majesty's Stationery Office, 1926).
75. Allen, A. M. *et al.* Characterization of a Wheat Breeders' Array suitable for high throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). *Plant Biotechnol. J.* http://dx.doi.org/10.1111/pbi.12635 (2016).
76. Barabaschi, D. *et al.* Next generation breeding. *Plant Sci.* **242,** 3–13 (2015).
77. Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R. & Crossa, J. Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). *Plant Sci.* **242,** 23–36 (2016).
78. Vivek, B. S. *et al.* Use of genomic estimated breeding values results in rapid genetic gains for drought tolerance in maize. *Plant Genome* http://dx.doi.org/10.3835/plantgenome2016.07.0070 (2017).
79. Riaz, A., Periyannan, S., Aitken, E. & Hickey, L. A rapid phenotyping method for adult plant resistance to leaf rust in wheat. *Plant Methods* **12,** 17 (2016).
80. Varshney, R. K., Terauchi, R. & McCouch, S. R. Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. *PLoS Biol.* **12,** e1001883 (2014).
81. Rodgers-Melnick, E., Vera, D. L., Bass, H. W. & Buckler, E. S. Open chromatin reveals the functional maize genome. *Proc. Natl Acad. Sci. USA* **113,** E3177–E3184 (2016).
82. Marulanda, J. J. *et al.* Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale. *Theor. Appl. Genet.* **129,** 1901–1913 (2016).
83. Spindel, J. E. *et al.* Genome-wide prediction models that incorporate *de novo* GWAS are a powerful new tool for tropical rice improvement. *Heredity* **116,** 395–408 (2016).
84. Patil, G. *et al.* Genomic-assisted haplotype analysis and the development of high-throughput SNP markers for salinity tolerance in soybean. *Sci. Rep.* **6,** 19199 (2016).
85. Jordan, K. W., Wang, S., Lun, Y. & Gardiner, L. J. A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol.* **16,** 48 (2015).
86. Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nature Methods* **4,** 931–936 (2007).
87. Saintenac, C. *et al.* Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (*Triticum aestivum* L.). *Genetics* **181,** 393–403 (2008).
88. Sadhu, M. J., Bloom, J. S., Day, L. & Kruglyak, L. CRISPR-directed mitotic recombination enables genetic mapping without crosses. *Science* **352,** 1113–1116 (2016).
89. Furbank, R. T. & Tester, M. Technologies to relieve the phenotyping bottleneck. *Trends Plant Sci.* **16,** 635–644 (2011).
90. Fiorani, F. & Schurr, U. Future scenarios for plant phenotyping. *Annu. Rev. Plant Biol.* **64,** 267–291 (2013).
91. Araus, J. L. & Cairns, J. E. Field high-throughput phenotyping: the new crop breeding frontier. *Trends Plant Sci.* **19,** 52–61 (2013).
92. Zamir, D. Where have all the crop phenotypes gone? *PLoS Biol.* **11,** e1001595 (2013).
93. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nature Methods* **13,** 751–754 (2016).
94. Rand, A. C. *et al.* Cytosine variant calling with high-throughput nanopore sequencing. Preprint at http://biorxiv.org/content/early/2016/04/04/047134 (2016).
95. Simpson, J. T. *et al.* Detecting DNA methylation using the Oxford Nanopore Technologies MinION sequencer. Preprint at http://biorxiv.org/content/early/2016/04/04/047142 (2016).

Acknowledgements This work was supported by strategic programme funding from the UK Biotechnology and Biological Sciences Research Council (BBSRC) (GRO BB/J004588/1) to M.B., a BBSRC strategic LoLa award (BB/J003913/1) to M.B. and M.C. and funding from the Gatsby Charitable Foundation to K.K. and the 2Blades Foundation to B.W.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at go.nature.com/xxxxxxx. Correspondence should be addressed to M.B. (michael.bevan@jic.ac.uk).

Reviewer Information *Nature* thanks V. Albert, J. Schmutz, T. Mitchell-Olds and the other anonymous reviewer(s) for their contribution to the peer review of this work.