Leafing through the genomes of our major crop plants: strategies for capturing unique information

Andrew H. Paterson

Abstract | Crop plants not only have economic significance, but also comprise important botanical models for evolution and development. This is reflected by the recent increase in the percentage of publicly available sequence data that are derived from angiosperms. Further genome sequencing of the major crop plants will offer new learning opportunities, but their large, repetitive, and often polyploid genomes present challenges. Reduced-representation approaches — such as EST sequencing, methyl filtration and Cot-based cloning and sequencing — provide increased efficiency in extracting key information from crop genomes without full-genome sequencing. Combining these methods with phylogenetically stratified sampling to allow comparative genomic approaches has the potential to further accelerate progress in angiosperm genomics.

Crop domestication ranks among the greatest of human achievements, and is closely correlated with human population growth and social evolution[1]. Scientific plant breeding, building on the accomplishments of domestication, predates the discovery of the Mendelian principles and is now complemented by genetic engineering. The independent but sometimes convergent[2–4] domestications of about 200 species[5] has provided a broad sampling of angiosperm diversity among our major crops. In addition, the intensive directional selection that is applied during crop breeding has produced many extraordinary models for aspects of plant morphology and development, such as the single-celled seed epidermal trichomes ('lint fibres') of *Gossypium* spp. (cotton)[6], the enlarged vegetative and floral meristems of the *Brassica* genus, and the enlarged carbohydrate-rich seeds of cereals. Crop genomes, therefore, not only hold information that can be used for further improvements, but also offer insights into angiosperm biology in general.

The recent sequencing of the first angiosperm genomes — *Arabidopsis thaliana* and two *Oryza* subspecies — foreshadowed new opportunities that could result from having the complete genetic blueprints for our major crops. Here I explore the opportunities and obstacles that are associated with obtaining and using these genomic sequences. I discuss the challenges that arise from the large and variable sizes of plant genomes, the abundance and distribution of repetitive DNA within them, and their frequent polyploidy. I also highlight new technologies and approaches that might accelerate progress in our understanding of crop genomes.

Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30602, USA.
e-mail: paterson@dogwood.botany.uga.edu
doi:10.1038/nrg1806

Applications of crop genome information

Combining comprehensive sequence information with knowledge of the morphological and physiological diversity of angiosperms and their well-understood phylogeny promises to answer many questions about crop genome evolution and function. Genetic improvement of crops is an essential means for meeting many basic needs of the developing world, and global genetic resource collections for these plants and their relatives (BOX 1) form the basis of future progress in this area. A genome sequence for one representative of these species provides a platform for the formulation of hypotheses — for example about possible relationships between specific DNA elements and phenotypes. Genetic resource collections can then be used to test such hypotheses, translating hard-won functional data that are gained from botanical (or other) models into increased crop productivity and/or quality, and perhaps also elucidating adaptations that contributed to the evolutionary success of the angiosperms.

Applications to crop improvement. Studying the relationship between molecular and morphological diversity is a key step in probing the consequences of prehistoric domestications, and is also important for further

Box 1 | Global germplasm resources

The fact that most angiosperm seeds can be preserved in a viable condition for many years by storage at low temperature and humidity has led to the establishment of genebanks for most of the world's leading crops. These provide insurance against habitat loss or crop disease epidemics, and are also useful as an archive from which scientists can obtain well-defined materials for research. Particularly noteworthy are the 11 genebanks that are currently maintained by the Consultative Group on International Agricultural Research (CGIAR), in close association with breeding programmes that study and use the germplasm. As of 2001, the CGIAR genebanks held about 666,000 germplasm accessions (plant or seed samples) of crops, forages and agroforestry trees[93].

Several countries also maintain substantial germplasm collections, often focusing on crops of national priority. The US Department of Agriculture Germplasm Resources Information Network collection is particularly extensive, including 464,864 accessions of 11,329 species and 1,837 genera. Most of these taxa are represented by few accessions — only 400 genera are present in more than 20 accessions, with the depth of representation being closely related to economic importance.

improving crops. The remarkable success that has been achieved in modifying angiosperms for agricultural use has been based largely on phenotypic data, rather than on a genetic, biochemical or molecular understanding. Independent domestications that seem to be convergent at the phenotypic level show non-random correlations in the locations of QTLs for the corresponding traits[2–4]. This indicates that Vavilov's 'law of homologous series in variation'[7] (an early recognition of the fundamental similarity between different cultivated species) might hold at the molecular level. However, frequent genome duplications in plants create opportunities for subfunctionalization and neofunctionalization — mechanisms by which genes might have come to carry out different functions from their ancestors or extant relatives in other taxa[8]. Sequence information for other angiosperms will allow us to explore how such changes in gene function have affected botanical diversity.

A better understanding of the evolutionary fates of duplicated genes could also be important for dissecting epistasis, which has an important role in the genetics of crop productivity, but has been difficult to study using molecular marker approaches[9,10]. For example, the division of ancestral functions among different genes by subfunctionalization could create 'webs' of permanent interdependence, which might be a foundation for the formation of epistatic relationships among unlinked genes. The identification of duplicated genes, and computational inferences about their respective sets of functional domains and/or regulatory motifs, could provide initial clues for how to prioritize empirical studies.

General insights into angiosperm biology. Sequencing the genomes of a diverse set of crops could provide insights into the remarkable evolutionary success of the angiosperms, which have successfully colonized an extensive range of habitats: from the tropics to near the poles, and from sea level to at least 20,100 feet in altitude[11]. Furthermore, angiosperms show a remarkable level of morphological diversity: their ramets range from floating plants of 1 mm in length (*Wolffia* spp.[1]) to trees of 100 m in height and 10 m in trunk diameter (*Eucalyptus regnans*). The largest known genet of any organism is of a *Populus deltoides* genotype, consisting of 47,000 stems that occupy 43 ha (REF. 12). Several others of a similar size are known, some as much as 43,000 years old[13].

Further sequence information could allow the deduction of the gene repertoire and organization of the last universal common ancestor of the angiosperms. The *Arabidopsis thaliana* and *Oryza* spp. sequences have allowed us to begin to unravel the consequences of genome duplications that have fragmented ancestral gene orders and functions across multiple chromosomes. Many other angiosperm genomes have undergone lineage-specific duplications of sufficient size to be discernible from EST data[14]. However, smaller duplications might also have occurred that are too small to be resolved using this approach (for examples, see REF. 15) but which might often result in phenotypic consequences[16]. Detailed genetic maps that are based on sequence-tagged sites (STSs) reveal early clues about these duplications[17], but the extent of their effect on genome structure and function awaits further sequencing.

A better understanding of the nature and duration of the euchromatic state in angiosperms could also be obtained through an increased availability of sequence information. Defined originally by differential heteropycnosis (instrinsic stainability), euchromatin and heterochromatin represent evolutionarily distinct components of genomes that can be detected by several comparative or *in silico* approaches[18]. Detailed analysis shows that most heterochromatin in rice has been restructured since its divergence from *Sorghum bicolor* (sorghum) about 42 million years ago, whereas gene order in the neighbouring euchromatin has largely been preserved[18]. A combination of cytological and sequence information from other branches of the angiosperm family tree will help to explore the duration over which the euchromatic state has been preserved, and might also provide clues about the factors that define it.

Increased sequencing could also explain why heterozygosity in self-pollinating plants persists to a higher degree than would be expected on the basis of Mendelian principles of segregation[19]. The identification of relictual heterozygosity in primary sequence is now possible, to a degree, using whole-genome shotgun (WGS) sequencing approaches (detailed below)[20], or by resequencing approaches. Such studies could determine whether there are genes for which heterozygosity imparts a fundamental advantage, and if so whether they are related across divergent taxa.

Obstacles to sequencing crop genomes

Genome size. Plant genome sizes vary over a range of at least 1,000-fold, from 125 Mb for the haploid genome of *Arabidopsis thaliana* to 125 Gb for the lily *Fritillaria assyriaca*[21]. The decision to sequence a crop genome is therefore a complex equation that balances genome size with scientific, economic and social impact; the phylogenetic distance from previously sequenced plants (that is, the new information that is likely to be yielded); relevant information from previous studies (such as the availability of genetic or physical maps); and the persuasiveness of individual (or groups of) investigators.

Germplasm
The hereditary materials within a species.

Subfunctionalization
Division of ancestral functions of a gene between duplicated copies of the original gene.

Neofunctionalization
Evolution of new function(s) for a gene, which are thought to be made possible by duplication of the gene, with one copy retaining the ancestral function.

Epistasis
Nonlinear interactions between independent genes that affect their impact on a phenotype.

Ramet
An individual plant that is part of a clump of plants that are genetically identical to a single parent.

Genet
A set of individuals that are produced by asexual reproduction from a single zygote.

Sequence-tagged site
A genetic locus that is defined by unique sequence information.

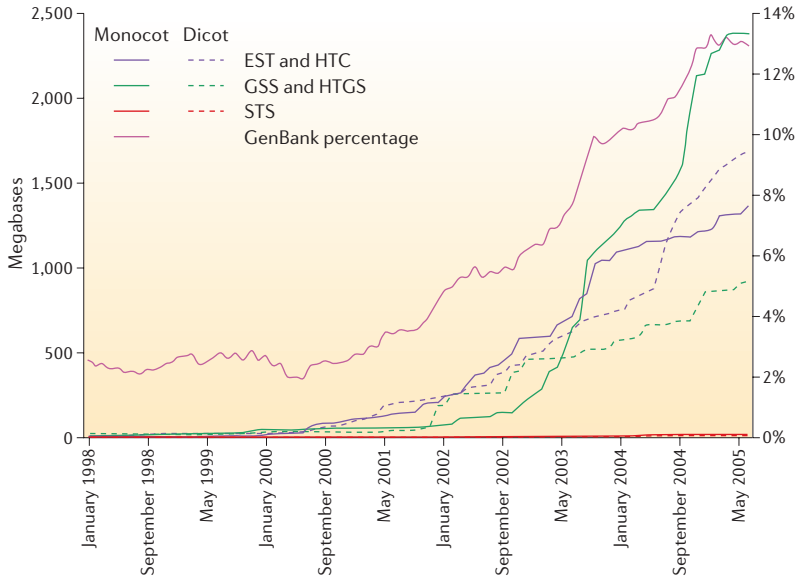


Figure 1 | **Long-term trends in angiosperm DNA-sequence generation.** The past 5 years have seen a pronounced acceleration of angiosperm sequence data accumulation, with especially large growth in genomic survey sequences (GSS) and high-throughput genomic sequences (HTGS) for monocots, reflecting the rice genome project and pilot efforts in maize. EST and high-throughput cDNA (HTC) efforts in angiosperms have also grown several-fold more rapidly than GenBank overall. STS, sequence-tagged site.

The total genome size of 70 crops for which estimates are available is about 1.48×10^{11} bp (see Supplementary information S1 (table)). Assuming that these are representative of the ~200 domesticates[5], to fully sequence just one genotype for each using current technologies (assuming 8× depth of sequence coverage) would involve 3.4×10^{12} bp of raw sequence. This is 72 times more than the 4.9×10^{10} bp of total sequence in GenBank at the time of writing. Angiosperm genomes currently comprise ~13.5% of the GenBank sequence data (FIG. 1); if this fraction is maintained — and if we sustain the 60% per year average increase in sequence information that has been realized since the 1980s — then the sequencing of 200 domesticated plants would take a relatively short 14 years. However, should rates of sequence growth drop to even 30% per year, the timetable lengthens to 24 years. Moreover, to determine the levels and patterns of polymorphism in each gene of each species, an important next step, would divert a growing share of capacity to resequencing, extending the timetable further.

Repetitive DNA. Repetitive DNA accounts for much of the remarkable variation in plant genome sizes. A comparison of DNA renaturation kinetics studies (BOX 2) for 36 plant genomes demonstrates this[22]. Kinetic complexity — an approximation of the fraction of a genome that consists of non-redundant sequence — ranges from 13% (*Allium cepa*, onion) to 77% (*Lycopersicon esculentum*, tomato) of total genome size (averaging 39%), with larger genomes having smaller fractions of non-redundant sequence (correlation coefficient $r = -0.297$). By contrast, an analysis of 24 mammals showed a narrower range of 50% (*Bos taurus*, cow) to 91% (*Cricetulus griseus*, Chinese hamster), averaging 72%, and only a tenuous relationship of kinetic complexity to genome size ($r = -0.085$).

Repetitive DNA yields little new information per additional 'family member' sequenced, and also complicates sequence assembly if family members are of such recent origin that they lack distinguishing mutations. The lower fraction of non-redundant sequence in large angiosperm genomes seems to reflect a rapid turnover of repetitive DNA families. For example, in *Zea mays* (maize) — which has an approximately average size (2.4 Gb) and percentage of kinetically unique sequence (38%) among angiosperms — most sequenced retrotransposons were inserted within the past 6 million years, leading to a doubling of genome size[23]. Therefore, not only do plants have more repetitive DNA than animals, but individual copies might have fewer distinguishing mutations and so might be more problematic for sequence assembly.

Polyploidy. In many plants, the entire genome is duplicated. Autopolyploids, such as *Saccharum* spp. (sugarcane) and *Medicago sativa* (alfalfa), contain several chromosome sets that can pair and recombine in all combinations (albeit to varying degrees[24]). These species are generally intolerant of inbreeding, as they contain heterozygosity (sequence polymorphism) within individuals, which is important to their productivity[25,26] and complicates sequence assembly. Allopolyploids such as wheat or cotton have undergone sufficient divergence that the duplicated chromosomes normally do not pair, and the sequences of gene pairs are usually distinguishable. All angiosperm genomes are palaeopolyploid[27], but the remaining 'palaeologous' gene pairs are usually well-differentiated.

Reduced-representation approaches

Although the long-term need is knowledge of the sequence, linear organization and patterns of variation in all functional elements across the major crops, the large size of crop genomes necessitates stepwise approaches: obtaining gene-related information first and progressing to genome-wide information as sequencing costs drop.

The EST approach. *En masse*, EST sequencing has been a natural first step in gene discovery that mitigates the repetitive nature of plant genomes. For many crops, collections of 10^5 or more ESTs from diverse tissues and physiological states have provided new DNA markers, revealed many candidate genes, and allowed testing of evolutionary hypotheses. Full-length cDNAs for *Arabidopsis thaliana*[28] and *Oryza* spp.[28,29] are especially useful in genome annotation, clarifying exon–intron junctions and splice variants, and identifying antisense RNA genes that might participate in gene regulation and imprinting. Their greater length than ESTs also allows more definitive comparisons of gene repertoires in diverse taxa. However, ESTs are subject to sampling bias and variation in abundance owing to different expression levels; they only include the subset of genes that are expressed in the source tissues, and exclude regulatory sequences.

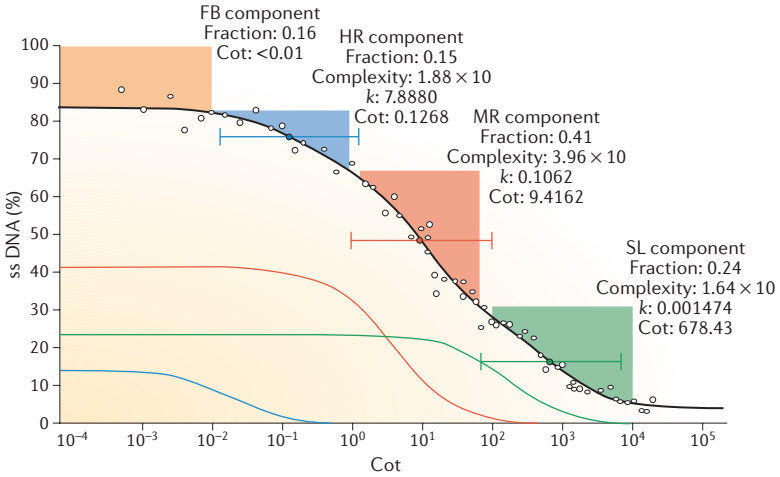
Methyl filtration and Cot-based cloning and sequencing. Transcriptome coverage in many crops is above the ~50% of genes beyond which the EST approach loses efficiency in revealing new genes[30]. Two new approaches promise to advance transcriptome coverage, and also provide information about introns and regulatory sequences from genomes for which complete sequencing is not yet justifiable.

The first approach is based on the generalization that expressed angiosperm genes tend to be hypomethylated, whereas non-expressed DNA (including some repeat families) is often methylated. This generalization has been used widely since the 1980s — for example, by using methylation-sensitive restriction enzymes to select for low-copy DNA clones that are suitable for use as locus-specific markers. Methylation filtration (MF) — the cloning of total genomic DNA into *Escherichia coli* strains that restrict methylated DNA[31,32] — similarly reduces the abundance of some repetitive DNA families. The second approach — Cot-based cloning and sequencing (CBCS)[22,33] — uses DNA renaturation kinetics to fractionate a genome into subpopulations of DNA segments that differ in their iteration frequency. These subpopulations are then cloned and sequenced to a depth that is appropriate to cover their respective sequence complexities, which are readily estimated[34,35] (BOX 2).

Although both methods enrich for genes, MF is a filter, eliminating much of a genome from consideration. It has been validated by comparing the hypomethylated DNA that is obtained to random genomic DNA, which showed a clear enrichment for known genes using MF[31,36]. However, by accessing only hypomethylated DNA, the ability of MF to extract information about genes depends on the variable relationship of methylation to expression across genes, taxa[22,33] and physiological states. For example, methylation can vary when cells are exposed to treatments such as radiation[37] or are grown in tissue culture[38,39].

Box 2 | **DNA renaturation kinetics**

DNA renaturation-kinetic analysis is an important approach for studying genome complexity. Total genomic DNA is first fragmented to a relatively small size to minimize confounding that is due to multiple classes of DNA (which might have different kinetic properties) in single molecules. This is typically 300 bp, although larger sizes can be used for specific applications. The DNA is then denatured and allowed to renature by gradual reductions in temperature. This reveals different 'Cot' values, which are the product of nucleotide concentration in moles per litre (Co), the reassociation time in seconds (t), and, if applicable, a factor that is based on the cation concentration of the buffer[94]. Careful monitoring of the percentage of total genomic DNA that has actually renatured is achieved by hydroxyapatite chromatography, which is based on the principle that hydroxyapatite (calcium phosphate) specifically binds dsDNA, whereas ssDNA will pass through a column of this material. This also provides a means to isolate the specific DNA molecules that have reannealed by a particular Cot point, for cloning and/or other studies.



Least-squares analysis of the resulting 'Cot curve'[95] allows the estimation of the overall genome size, the fraction of the genome that falls into each of 2–4 distinguishable kinetic components, and the 'kinetic complexity' of each component. This last parameter, in units of nucleotides, is an accurate approximation of the 'sequence complexity' or the quantity of non-redundant sequence in the component. On the basis of this estimate, one can determine the number of sequencing reads that are necessary to provide an essentially complete coverage of a kinetic component with a predetermined level of statistical confidence[22,33].

In the example shown in the figure, the sorghum genome was resolved into four components: a 'foldback' (FB) component that comprises 16% of the genome and is thought to consist of palindromic DNA molecules that self-reanneal at rates that are independent of DNA concentration[96,97]; a highly repetitive (HR) component that comprises 15% of the genome but only 0.0011% of the overall kinetic complexity (which is approximately equal to the sequence complexity); a middle-repetitive (MR) component that comprises 41% of the genome but only 2.4% of the overall kinetic complexity; and a single-low (SL) copy component that comprises only 24% of the genome but 97.6% of the overall kinetic complexity. Therefore, by separating this component from the remainder of genomic DNA, the majority of unique sequence can be obtained with a minimum of interference from the repetitive DNA that comprises the majority of most crop genomes. *k* represents the reassociation rate (in M^{-1} sec^{-1}) of each component. Note that *k* and Cot are mathematically 'undefined' for FB, because the components reassociate so quickly that rates are not determined by DNA concentration. Figure modified with permission from REF. 33 © (2002) Cold Spring Harbor Laboratory Press.

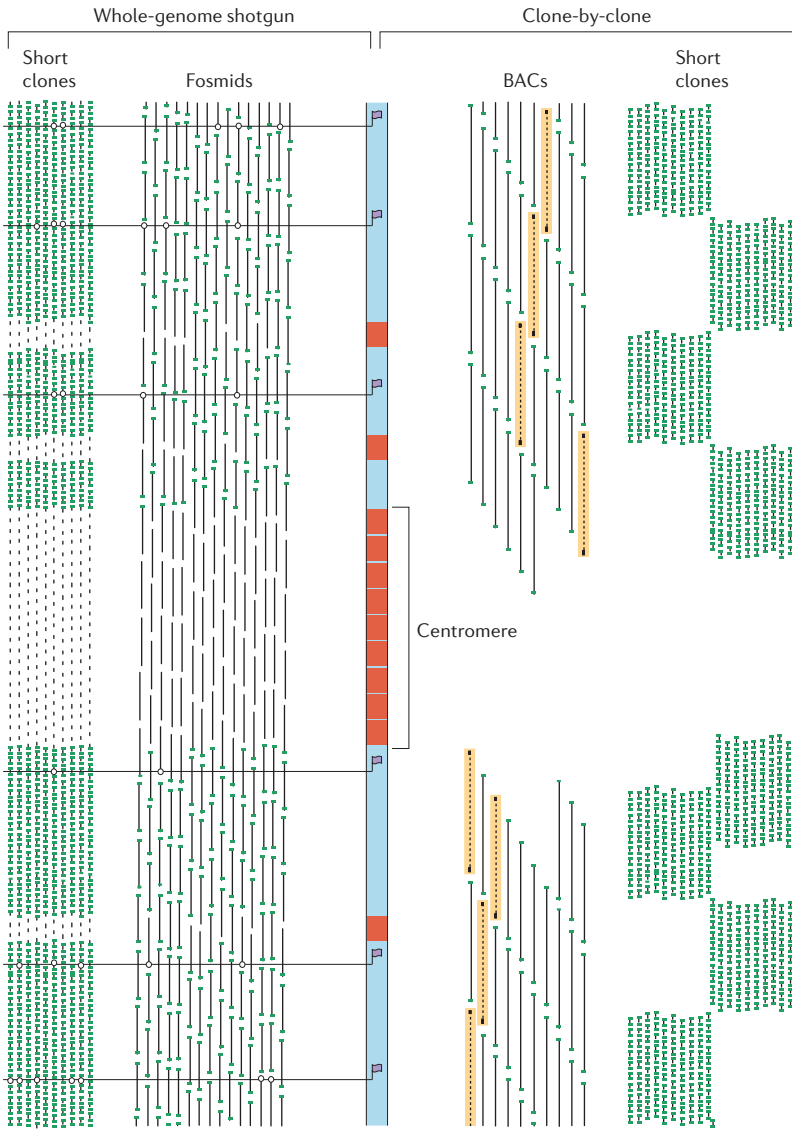


Figure 2 | **Whole-genome shotgun versus clone-by-clone sequencing strategies.** The whole-genome shotgun approach (shown on the left of the chromosome, which is shown in blue) is typically based on obtaining enough sequence so that each nucleotide in the genome is covered an average of 6 to 8 times by paired-end sequences (shown in green) for randomly chosen clones of different sizes. In the example, an approximately 7-genome equivalent coverage, consisting of short 3–4 kb clones, was obtained, together with about 0.5× coverage from paired ends of fosmids of ~36 kb. The fosmid clones from which the sequences are derived collectively provide about 15× coverage of the genome. By contrast, clone-by-clone sequencing (right) typically involves identifying a 'minimum tiling path' (hatched clones highlighted in yellow) of large-insert clones (such as BACs) of known order with respect to one another, which is achieved using a combination of genetic markers (indicated as horizontal lines extending across the figure) and physical markers (not shown). This is followed by shotgun sequencing of 6–8× coverage of shorter clones that are derived from each BAC. The minimum tiling path typically allows for about 5–10% overlap between consecutive BACs. Whole-genome shotgun samples the entire genome, but sometimes fails to provide contiguity across elements that are found at high copy number (red boxes along the chromosome; failure to provide continguity is indicated by the absence of the end sequences of clones that correspond to these regions). Clone-based strategies are at risk of failing to sample highly repetitive regions (especially regions of tandem repeats, such as those that are shown near the centre of the chromosome, and which seem to be common in large plant genomes) owing to difficulties in resolving the arrangements of BACs across these regions (if indeed they can be cloned). Both methods benefit from information such as genetically mapped sequence-tagged sites that help to reinforce physical maps and align sequence scaffolds.

By contrast, CBCS is a parser, sorting the genome into kinetically defined components, some or all of which can be selected for further study. It has been validated by comparing genomic fractions that have different renaturation rates. Both the nature and abundance of individual sequences (such as well-studied repetitive elements or genes, respectively) were consistent with the expected properties of the different fractions. There is always a possibility that repetitive CBCS clones contain parts of two or more different repeat-element families, which would prevent empirical verification of their copy number. To minimize this risk, the initial validation used DNA that was sheared to a small size (~300 nt) from which single-stranded overhangs were removed after renaturation[33]. However, there is no technical constraint on the size of clones that can be studied, and following formal verification of the method[33] subsequent studies have focused on longer low-copy clones that are more readily assembled into sequence contigs[36,40].

Empirical comparisons of MF and CBCS have been carried out in two species. In maize, CBCS was more effective than MF at filtering repetitive DNA, which comprised 68%, 33% and 14% of unfiltered, MF and low-copy Cot libraries, respectively. This is largely due to an abundance of expressed (and therefore hypomethylated) retrotransposons[36]. Clones that were obtained by MF matched known genes at rates that were similar to those obtained using CBCS[36,41]. After removing known repeats, CBCS gives a higher fraction of sequences with no significant match than MF (73.8% versus 58.9%) (REF. 36). This intriguing observation could represent either an advantage, in terms of the greater prospects for discovery of previously unknown functional sequences, or a disadvantage, owing to a lower efficiency of gene discovery. In bread wheat (*Triticum aestivum*), CBCS and MF respectively yielded 13.7× and 2.2× gene enrichment, and 3× and 1.2× reduction in repeats[42]. A report that CBCS could lead to a higher rate of occurrence of non-native sequences[43] might be due to the use of non-standard methodologies in some studies[36,40].

The use of CBCS to efficiently capture the complexity of repetitive elements[33,44] — the largest constituent in most angiosperm genomes — might circumvent some problems and create new opportunities. Approaches that are based on PCR of *Alu*-like elements[45] have been applied in human genomics to: identify DNA markers; provide 'fingerprints' of large DNA clones and individuals; and study insertion mutations, recombination, gene conversion and gene expression[46].

Whole-genome sequencing approaches

Whole-genome shotgun versus clone-by-clone sequencing. What is the best approach to sequence crop genomes? Clone-by-clone sequencing of contiguous large-insert DNA clones simplifies a large genome into small pieces and delimits uncertainties to intervals of about 100 kb. However, the costs of assembling large-insert libraries and ordering clones have motivated WGS approaches, which achieve contiguity on the basis of overlaps between paired-end sequences from random clones (FIG. 2). The respective merits of clone-by-clone and WGS-sequencing

approaches in general have previously been expertly reviewed[47]. Here we focus on their merits in the context of the distinguishing features of crop genomes.

Arguments against WGS approaches in angiosperms focus on the generally higher content and more recent origin of repetitive DNA than in animals or microbes. Consider a hypothetical element for which 1,000 identical copies are randomly dispersed throughout a genome. Using WGS, each time that one of these is sequenced, there are 999 equally likely choices (multiplied by the redundancy of coverage) for the next clone along the DNA strand. By contrast, in clone-by-clone sequencing, a given BAC clone (for example) might only contain one member of this family. This problem is especially serious in plants, in that individual members of repetitive DNA families are often recently derived and might have few distinguishing mutations. Long repetitive elements are especially problematic — for elements that are shorter than a sequencing read (often 500–1,000 nt), flanking sequence might be locus-specific.

However, several considerations mitigate these arguments. First, repetitive DNA is often defined as a collection of elements that show 70–80% matches along a predetermined length of DNA sequence. However, such inclusive definitions that are important to studying repetitive-element evolution fail to reflect the power of modern sequencing and computational methods to exclude false matches between family members. Current sequencing error rates are ~1–5% of the average polymorphism rate of about 0.1% between human alleles[48]. Sequence divergence of 0.1% reflects about 100,000 years of neutral evolution, indicating that all but the most recent sequence duplications could be resolved even in genome-wide comparisons.

A second consideration is that highly repetitive tracts of a genome might be refractory to clone-based mapping. For example, 8.3% of fingerprinted BACs (those for which alignment to one another is known on the basis of shared restriction fragments) that could not be contigged in a high-coverage sorghum library included 19.5–23.2% of all hybridizations to centromeric probes (which consist of repeated sequences), but only 3.9% of hybridizations to single-copy probes[18]. Furthermore, 81% of these BACs that could not be contigged (versus 3.5% of those that could) had ≤14 different DNA fragment sizes in their fingerprints; clones that are largely composed of tandem repeats with multiple copies of the same band size cannot be reliably assembled by fingerprinting approaches[49,50].

Although highly repetitive genomic regions tend to be gene-poor and might be physically only a small portion of a genome, their characterization could prove to be important. The ability of the WGS approach to capture these regions is therefore another of its advantages. Pericentromeric regions — which are generally rich in repetitive sequences — contain active genes in *Drosophila melanogaster*[51] and *A. thaliana*[52]. The kinetochore region of one rice (*Oryza sativa*) centromere (which was amenable to complete sequencing as it contained little highly repetitive satellite DNA) includes 14 predicted and at least 4 expressed genes[53], although another rice kinetochore seems to be devoid of non-transposon-related genes[54]. Pericentromeric regions seem to be prone to rapid restructuring[18], and might therefore be 'hotspots'

Box 3 | **Whole-genome shotgun versus clone-based sequencing of rice**

The sequencing of two *Oryza* subspecies was carried out clone-by-clone, both by an international consortium[61] and by a private company, as well as twice by whole-genome shotgun (WGS) sequencing methods[62–64]. Both of the clone-based efforts, and one of the WGS efforts, used the subspecies *japonica*, which is grown in Japan and the USA. The other WGS effort used the subspecies *indica*, which is commonly grown in China. Recent detailed analyses of both WGS-based[64] and clone-based[61] rice genome assemblies demonstrate the merits of these strategies and the controversies that occur over their use.

The clone-based effort reports an average continuous sequence length of 6.9 Mb (REF. 61). The subspecies *indica* WGS effort reports an N50 size (above which half of the total length of the sequence set can be found) of only 24.9 kb (REF. 64). However, in combination with resources from WGS sequencing of the *japonica* subspecies, and other inferences, this effort has provided 'super-scaffolds' (including gaps) of 8.3 Mb. The clone-based approach offers higher contiguity; however the WGS method can use various inferences to minimize the consequences of gaps.

Both assemblies have also yielded similar estimates of the number of genes that are not derived from transposable elements. These are estimated to be 49,088 genes for WGS of the subspecies *indica*[64] (although EST confirmation rates indicate that only about 40,216 are 'real'), 37,794 for WGS of the subspecies *japonica*[64], and 37,544 for the clone-based *japonica* assembly[61]. The *indica* WGS assembly[64] includes 97.7% of the 19,079 full-length rice cDNAs that are available[28,29]. The clone-based assembly includes 99.4% of the 8,440 physically mapped EST markers[61].

Ancient whole-genome duplication has been detected in both clone-based[98–100] and WGS[62,64] assemblies. Indeed, WGS data[101] helped to resolve an early controversy between two clone-based analyses[102], showing that this event was probably genome-wide[98] rather than representing 'ancient aneuploidy'[103].

Nonetheless, there remain significant differences in the assemblies. The WGS assembly[64] points to an overall genome size that is about 10–15% larger than the clone-based data. Moreover, the 12.3% of the WGS data that could not be aligned to the overall assembly contained only 0.7% of genes, perhaps representing gene-poor regions that are recalcitrant to clone-based mapping, and so absent from the clone-based data. The clone-based effort reports that the *indica* WGS assembly[64] covered only 69% of its assembly, that a high level of sequence mismatches and misalignments differentiated the two assemblies in a sample region that was carefully scrutinized, and that 68% of ostensibly centromere-specific CentO sequences were found outside the centromeric regions in the subspecies *indica* WGS assembly[64]. These inconsistencies await further investigation.

Gene conversion
A meiotic process of directed change in which one allele directs the conversion of a partner allele to its own form, probably by repair of heteroduplex DNA.

Box 4 | **Pending genome sequences for genera that include the major crops**

See the Further information for more information and updates on each of these projects. See also note added in proof.

Populus
About 7.5× coverage in small-insert end-sequences of the ~500-Mb genome of *Populus trichocarpa* (poplar) has been generated by the US Department of Energy Joint Genome Institute (JGI). More mapping and sequencing is ongoing, and will enhance contiguity and link the sequence to chromosomes.

Medicago
Sequencing of a minimum tiling path of about 2,100 BACs that cover the gene-rich euchromatin, estimated at 200 Mb, or 42% of the entire 470-Mb genome, is in progress for the diploid *Medicago truncatula* (a close relative of cultivated alfalfa), which is scheduled for completion in 2006. Chromosomes have been assigned to sequencing centres by an international consortium.

Sorghum
8× whole-genome shotgun coverage of the ~736-Mb genome of *Sorghum bicolour* will be generated by the JGI, combined with publicly available methylation filtration sequences[104], assembled, and integrated with sequence-tagged genetic and physical markers[105]. This will yield genetically orientated pseudomolecules that cover most chromosomes to a substantial degree. Sequencing is scheduled for completion in 2006.

Lycopersicon
Sequencing of a minimum tiling path that is estimated at 2,276 BAC clones covering the ~220 Mb of euchromatin, or 25% of the 950-Mb genome, of *Lycopersicon esculentum* (tomato) is in progress by an international consortium (including members from the United States, China, Japan, South Korea, the United Kingdom, Italy, Spain, France, Netherlands and India). This project is starting from 'seed' BAC clones, which are individually anchored to a high-density genetic map. It is anticipated that it will be completed in 2007–2008.

Zea
Substantial methylation filtration and Cot-based survey sequence, together with ESTs and BAC ends[106], are available now. A consortium of US federal agencies led by the US National Science Foundation have recently announced awards for sequencing most, if not all, of the genome.

Brassica
Sequencing a minimum tiling path of BAC clones to 'Phase 2' (sequence that is fully orientated and ordered, but contains some small sequence gaps and low-quality regions) is in progress for the ~500-Mb genome of *Brassica rapa*, subspecies *pekinensis*. The genome sequence is to be anchored to a reference genetic map by ~1,000 molecular markers.

Solanum
BAC-based sequencing of the entire potato genome, including heterochromatin, is the goal of an international consortium, with a target date for completion of 2008. A few groups have already been funded, although much of the genome remains the subject of proposals that are in review or planned.

for the evolution of new genes by transposon-mediated mechanisms[55] (although all such gene-like elements that are found in the modern rice genome seem to be non-functional)[56]. Pericentromeric regions are also recombination-poor[53], and might nurture the evolution of 'supergenes' — physically dispersed but genetically tightly linked, co-adapted gene complexes that might have been selected for during domestication[57,58].

The WGS approach should provide coverage of most genes in regions that are refractory to clone-based physical mapping, even if their assembly proves impossible. Many clone-based crop-genome sequencing efforts have targeted euchromatin, setting aside problematic heterochromatic, repetitive regions until improved technologies are available. Recently announced maize sequencing projects — one BAC-based and one WGS — intend to try to characterize these regions. Eventual sequencing of such problematic regions for these and other taxa could allow more extensive computational studies of whether such regions might represent 'evolutionary nurseries' in which coding sequences are undergoing strong positive selection[59].

The WGS strategy also has its disadvantages. Although it allows the distinction of alleles from errors by redundant sampling when high sequence coverage is obtained[20], clone-based approaches allow assembly for one allele at a time, which excludes the possibility of heterozygosity. In autopolyploids, in which each of the 4, 6 or more copies of a chromosome might contain different haplotypes of otherwise identical genes, the basic gene set might be revealed by sequencing a tiling path of clones, setting aside the task of detecting allelic variation until sequencing is cheaper or alternative approaches emerge. Such a mosaic of maternal and paternal haplotypes was the outcome of the *Ciona intestinalis* WGS sequence[20], with 1.2% rates of polymorphism resulting in some small sequence scaffolds that are in fact short divergent haplotypes. However, in autopolyploids a clone-based approach would reduce by 50% (tetraploid), 75% (hexaploid), or more, the amount of sequencing needed to reach this outcome, by only sampling one allele per locus.

As noted previously[47], WGS and clone-based strategies continue to converge. WGS-based sequence assembly clearly benefits from positional information, such as the genetic and physical maps, and associated paired BAC-end sequences, that are key to clone-based strategies. Clone-based sequencing routinely uses random shotgun approaches to sequence each clone, and the availability of WGS sequences accelerates assembly and finishing. The optimal balance between these respective strategies is perhaps as complicated as the 'equation' regarding the

Haplotype
The genetic constitution of an individual chromosome; this can refer to one locus or to an entire genome. A genome-wide haplotype would comprise half a diploid genome, including one allele from each allelic gene pair.

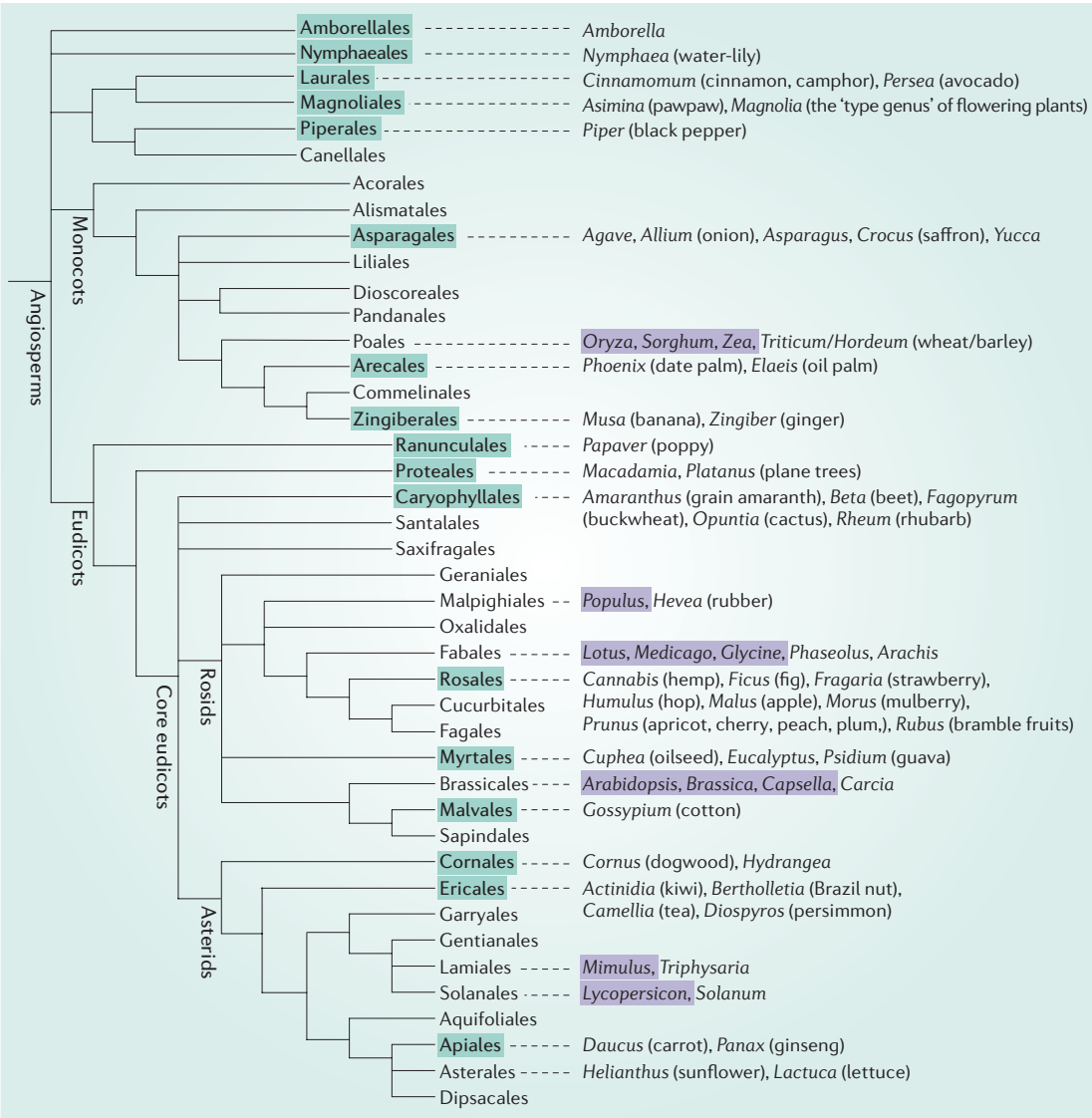


Figure 3 | **A phylogenetic view of angiosperm sequencing projects.** Phylogenetic relationships between angiosperms are shown, highlighting completed, continuing and potential sequencing projects. The positions of species that have already been sequenced or are currently being sequenced are shown (on a dark purple background). Other species that have been suggested for possible phylogenetic shadowing are also shown (taxonomic orders are on a dark green background, one or more candidate genera are listed in alphabetical order, with the common name in parentheses). Based on data from REFS 107,108.

case for when to sequence a genome: the costs of clone production, genetic and physical mapping (and relevant data that are already available), high-throughput sequencing, and directed sequence finishing all need to be considered. Features of genome organization also need to be taken into account, such as heterozygosity, polyploidy, and the abundance, distribution and homogeneity of repetitive DNA families.

Past, current and future sequencing projects. The small genome of *A. thaliana* was sequenced using a minimum tiling path of large-insert clones[60], and remains among the most completely sequenced genomes. However, it is atypical of angiosperm genomes, as it is roughly 6% of the average size and contains minimal repetitive DNA.

Minimum tiling path
A set of (usually large-insert) clones that collectively cover a genome, chromosome or target region, with a minimum of redundancy.

Sequencing of the second angiosperm, *O. sativa*, was carried out clone-by-clone both by an international consortium[61] and by a private concern, as well as twice by WGS methods[62–64]. A comparison of the WGS and clone-based rice sequencing efforts describes the trade-offs that are made between genome coverage, contiguity and accuracy of assembly (BOX 3).

Initial genome sequencing for members of 7 other genera that include the major crops are in progress or scheduled for sequencing using public funds (BOX 4). STS-based genetic maps and BAC libraries are available for most leading crops, and a growing number of genetically anchored physical maps have been constructed. In a few cases, radiation hybrid and/or chromosome-specific cell lines[65,66] allow sequences to be mapped to specific

regions of the genome. Efforts to produce high-resolution landmarks by optical mapping[67] are in progress. These collective resources enhance the ability to assemble contiguous WGS or clone-based sequences in a careful, stepwise manner to reduce experiment-wise error rates. Finally, these resources link sequence information to the results of two decades of research using STS-based genetic maps.

The potential of comparative approaches

In much the same way that clone-by-clone sequencing might start with defined seed sites and grow outwards, angiosperm sequencing has started with detailed studies of nodal model species and is spreading outwards across taxa. Many benefits of crop genome sequencing might be quickly realized by using 'phylogenetic shadowing' approaches[68] that are similar to those that have been recently proposed for mammals[69] — a taxonomic group that is a similar age to angiosperms. Only a tiny fraction of theoretically possible DNA sequences occur in nature, and natural selection quickly eliminates variants that reduce fitness by even tiny increments[70]. The demonstration that sequences have been preserved during evolution is one of the most sensitive approaches that are available to identify potential functional elements in genomes, and might prove to be the only computational means by which to identify those that are not recognizable by analogy to known elements[71].

In groups such as mammals, which have similar genome sizes and repeat abundances as one another, the number of genome comparisons that are required to reveal conserved features has been estimated. This was based on the minimum size (in nucleotides) of the feature to be identified, its relative rate of evolution, and the evolutionary distance (degree of neutral DNA substitution) of the source genome from a reference genome[72]. Sequence redundancy of about twofold (rather than the eightfold that is typical of whole-genome efforts) is suggested to offer compelling efficiency in yielding new information[69], although it also suffers various limitations as detailed below. These 2× sequences also reveal parameters such as the relatedness of repetitive element families that are important to deciding the relative merits of WGS versus clone-based approaches in eventual finished sequencing.

With modification to accommodate their more variable genome sizes, angiosperm phylogenetic shadowing might quickly provide the raw material that is needed to reveal functional elements of as small as a few nucleotides in length. In addition to *A. thaliana* and the eight crop-containing genera for which a primary sequence is done or in progress, WGS efforts are soon to begin for *Arabidopsis lyrata*, *Capsella rubella*, *Triphysaria versicolor* and *Mimulus guttatus*. Sequencing of as few as 16 other angiosperms at 2× coverage would cover most of the main branches of the family tree at least once. Conservatively anticipating that these other taxa would have an average 'branch length' of at least 0.25 neutral substitutions per site from the nearest fully sequenced reference genome[72] would provide the resolution that is required to detect conserved elements of around 8 nucleotides[69]. Sequencing a second divergent member of each branch, or a first member of some lesser branches, might approach the level of resolution that is required to identify single nucleotides that are under strong selection.

Were such 'sweet 16' angiosperms to be sequenced by a WGS approach, their average genome size of 2.1 Gb could be covered twice in 67 Gb of sequence, or about 2% of that required to fully sequence the 200 leading crops. Alternatively, CBCS could be used to capture 2× coverage of their sequence complexity (as defined in BOX 2, including low-copy, repetitive and foldback DNA), which averages about 44% of genome size based on calculations from published data[22]. This would reduce the minimal sequencing requirement to ~30 Gb, a level within the current annual capacity of some individual sequencing centres. A balance between CBCS and WGS would not only provide economy, but quantitative data about degrees of divergence among repetitive DNA families would also be obtained, which would be of value in formulating strategies for eventual completion of these genomes.

One possible set of genomes for angiosperm phylogenetic shadowing is illustrated in FIG. 3. I selected 17 taxonomic orders (highlighted in dark green), giving priority to those that contain economically important lineages, and then listed some possible genera that might be considered in choosing one within each order. Thorough sampling of the angiosperm family tree would necessarily include some ornamental (for example, *Nymphaea* spp.) and uncultivated (for example, *Amborella* spp.) representatives of ancient lineages that are important to unravelling early events in angiosperm evolution. Naturally, others would come up with different permutations regarding the exact orders and genera that are prioritized (for examples see REFS 73,74). Although identifying a consensus set would require much deliberation by the scientific community, incorporating formal quantitative measures of evolutionary divergence into prioritization of continuing sequencing efforts might accelerate progress in angiosperm comparative genomics.

Angiosperm phylogenetic shadowing would be a stepping-stone only, quickly yielding some rewards but falling short of our long-term needs in several ways. First, its value is heavily dependent on the parallel study of many taxa. For example, although many genes will have contiguity gaps in any one taxon at 2× coverage, a collective 32× average coverage of the transcriptome would provide much information about the structure and evolution of each gene across the estimated 140–180 million years[75,76] of angiosperm evolution. Second, low contiguity limits power to resolve features of genome organization, such as whole-genome duplications; however, a degree of inference about these is possible on the basis of patterns of gene sequence divergence alone[14,77]. Third, the broad phylogenetic coverage that facilitates identification of conserved features also sacrifices resolving power to associate individual features with biological or evolutionary roles. For example, sequencing *A. thaliana* and *A. lyrata* will offer greater power to attribute specific roles to DNA polymorphisms, but will reveal fewer differences than would the comparison of *Arabidopsis* spp. and *Brassica* spp. Sampling of one domestic and one wild

Radiation hybrid
A cell line that contains one or more chromosome segments from another species, which is generated by irradiation of cells from a target species, followed by fusion with normal cultured cells from a 'host' species. This allows the mapping of genes or other DNA sequences on the basis of similarities and differences in the ability of different cell lines to bind DNA probes from the target organism.

Chromosome-specific cell lines
Similar to radiation hybrids, these are generated by irradiation of cells from a target species, followed by fusion with normal cultured cells from a 'host' species. However, unlike radiation hybrids, they contain only one chromosome from the target organism. This allows mapping of genes or other DNA sequences on the basis of the binding of DNA probes from the target organism.

Optical mapping
Use of light microscopy to directly image individual DNA molecules, which are bound to specially derivatized surfaces and then cleaved by restriction enzymes.

Foldback DNA
When denatured, this DNA reassociates at a high rate that cannot be explained by bimolecular association. This is probably due to the presence on the same strand of palindromic elements that can self-anneal.

Parsimony
In systematics, parsimony refers to choosing the simplest explanation of the observed data. For example, which phylogenetic tree requires the fewest possible mutations to explain the data.

taxon within each selected order might add a new dimension to our ability to investigate genomic features that favour domestication. It would also mitigate biases, such as the possibility that domesticated taxa might be rich in loss-of-function mutations[78] at corresponding sites[2]. Using a second sample per node will also begin to address the range of complexity that must be dealt with in each taxonomic group; however, a third sample will be needed to allow parsimony-based phylogenetic inferences.

Future perspectives

Sequences for single representatives of the major crop species will only scratch the surface. Analysis of divergence among taxa might show, for example, that rapid change in particular functional groupings of genes has occurred in a particular branch of the angiosperm tree, but will not differentiate invariant attributes of a taxon from polymorphic targets for potential crop improvement. A better understanding of the relationship between divergence among taxa and diversity within the gene pool for each crop will be essential. Cataloguing the naturally occurring diversity within each gene has recently been of high priority in humans[79–83]. Revealing the suite of naturally occurring genetic polymorphisms in crop gene pools will yield many benefits — for example, suggesting possible footprints of domestication[84,85] or signals of selection that are associated with specific genes[86–88]. The predominantly self-pollinating nature of many crops has the benefit that the effective population size is relatively small, and that most allelic diversity can be captured by analysis of carefully selected 'core collections'[89].

With completed sequences in hand, diversity analysis could target genes directly, rather than through the proxy markers that have been necessary for the past 20 years. New clone-free DNA-sequencing methods that provide increased cost-efficiency and speed[90,91] present the possibility of 'transcriptome shadowing' — the rapid sequencing of large populations of cDNA to a deep level of coverage for a diverse sampling of germplasm, to shed light on levels and patterns of functional polymorphism. Once most polymorphisms in a gene pool are identified, high-throughput methods[92] promise to allow their characterization in large crop germplasm collections. Together with parallel advances in the knowledge of gene functions, this will bring crop improvement to a level of determinism that reaches well beyond the (albeit important) advances that are associated with marker-assisted selection. The freedom of crop scientists to make experimental crosses, together with the ability to know the exact genotype of each progeny individual, could foster advances that transcend those that are attainable in human genomics.

Note added in proof

It has recently been announced that the soybean genome will be fully sequenced by the US Department of Energy Joint Genome Institute.

1. Raven, P., Evert, R. & Eichhorn, S. *Biology of Plants* (Worth Publishers Inc., New York, 1992).
2. Paterson, A. *et al.* Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. *Science* **269**, 1714–1718 (1995).
3. Lin, Y., Schertz, K. & Paterson, A. Comparative analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific sorghum population. *Genetics* **141**, 391–411 (1995).
4. Hu, F. Y. *et al.* Convergent evolution of perenniality in rice and sorghum. *Proc. Natl Acad. Sci. USA* **100**, 4050–4054 (2003).
5. Clement, C. R. 1492 and the loss of Amazonian crop genetic resources. I. The relation between domestication and human population decline. *Econ. Bot.* **53**, 188–202 (1999).
6. Kim, J. K. & Triplett, B. A. Cotton fiber growth in planta and *in vitro*. Models for plant cell elongation and cell wall biogenesis. *Plant Physiol.* **127**, 1361–1366 (2001).
7. Vavilov, N. The law of homologous series in variation. *J. Genet.* **12**, 1 (1922).
8. Tocchini-Valentini, G. D., Fruscoloni, P. & Tocchini-Valentini, G. P. Structure, function, and evolution of the tRNA endonucleases of Archaea: An example of subfunctionalization. *Proc. Natl Acad. Sci. USA* **102**, 8933–8938 (2005).
9. Li, Z. K., Pinson, S. R. M., Park, W. D., Paterson, A. H. & Stansel, J. W. Epistasis for three grain yield components in rice (*Oryza sativa* L). *Genetics* **145**, 453–465 (1997).
10. Li, Z. K. *et al.* Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. I. Biomass and grain yield. *Genetics* **158**, 1737–1753 (2001).
11. Younghusband, F. *The Epic of Mount Everest* (E. P. Publishing, London, 1926).
12. Mitton, J. B. & Grant, M. C. Genetic variation and the natural history of quaking aspen. *Bioscience* **46**, 25–31 (1996).
13. Lynch, A. J. J. & Balmer, J. The Ecology, phytosociology and stand structure of an ancient endemic plant, *Lomatia tasmanica* (Proteaceae), approaching extinction. *Aust. J. Bot.* **52**, 619–627 (2004).
14. Blanc, G. & Wolfe, K. H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**, 1667–1678 (2004).
15. Lagercrantz, U. & Lydiate, D. J. Comparative genome mapping in Brassica. *Genetics* **144**, 1903–1910 (1996).
16. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
17. Rong, J. *et al.* Comparative genomics of *Gossypium* and *Arabidopsis*: unraveling the consequences of both ancient and recent polyploidy. *Genome Res.* **15**, 1198–1210 (2005).
18. Bowers, J. E. *et al.* Comparative physical mapping links retention of microsynteny to chromosome structure and recombination in grasses. *Proc. Natl Acad. Sci. USA* **102**, 13206–13211 (2005).
19. Brown, A. H. D. Enzyme polymorphism in plant-populations. *Theor. Popul. Biol.* **15**, 1–42 (1979).
20. Dehal, P. *et al.* The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**, 2157–2167 (2002).
21. Bennett, M. & Smith, J. Nuclear DNA amounts in angiosperms. *Philos. Trans. R. Soc. Lond. B* **334**, 309–345 (1991).
22. Peterson, D. G., Wessler, S. R. & Paterson, A. H. Efficient capture of unique sequences from eukaryotic genomes. *Trends Genet.* **18**, 547–550 (2002).
23. SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. & Bennetzen, J. L. The paleontology of intergene retrotransposons of maize. *Nature Genet.* **20**, 43–45 (1998).
24. Ming, R. *et al.* Detailed alignment of saccharum and sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes. *Genetics* **150**, 1663–1682 (1998).
25. Pfeiffer, T., Schrader, L. E. & Bingham, E. T. Physiological comparisons of isogenic diploid–tetraploid, tetraploid–octoploid alfalfa populations. *Crop Sci.* **20**, 299–303 (1980).
26. Ming, R., Liu, S. C., Moore, P. H., Irvine, J. E. & Paterson, A. H. QTL analysis in a complex autopolyploid: genetic control of sugar content in sugarcane. *Genome Res.* **11**, 2075–2084 (2001).
27. Bowers, J. E., Chapman, B. A., Rong, J. K. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438 (2003).
A synthesis of divergent approaches to the study of genome organization that clarified the evolutionary history of angiosperm chromosomes.
28. Haas B. J. *et al.* Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol.* **3**, research0029.1–research0029.12 (2002).
29. Kikuchi, S. *et al.* Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science* **301**, 376–379 (2003).
30. Soares, M. B. *et al.* Construction and characterization of a normalized cDNA library. *Proc. Natl Acad. Sci. USA* **91**, 9228–9232 (1994).
31. Rabinowicz, P. D. *et al.* Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nature Genet.* **23**, 305–308 (1999).
An elegant primary demonstration of the merits of the methylation filtration approach.
32. Rabinowicz, P. D., McCombie, W. R. & Martienssen, R. A. Gene enrichment in plant genomic shotgun libraries. *Curr. Opin. Plant Biol.* **6**, 150–156 (2003).
33. Peterson, D. G. *et al.* Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res.* **12**, 795–807 (2002).
An elegant primary demonstration of the merits of the Cot-based cloning and sequencing approach.
34. Britten, R. J. & Kohne, D. E. Repeated Sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* **161**, 529–540 (1968).
35. Goldberg, R. B. From Cot curves to genomics. How gene cloning established new concepts in plant biology. *Plant Physiol.* **125**, 4–8 (2001).
36. Whitelaw, C. A. *et al.* Enrichment of gene-coding sequences in maize by genome filtration. *Science* **302**, 2118–2120 (2003).
37. Kovalchuk, O. *et al.* Genome hypermethylation in *Pinus silvestris* of Chernobyl — a mechanism for radiation adaptation? *Mutat. Res.* **529**, 13–20 (2003).
38. Meng, L., Bregitzer, P., Zhang, S. B. & Lemaux, P. G. Methylation of the exon/intron region in the *Ubi1* promoter complex correlates with transgene silencing in barley. *Plant Mol. Biol.* **53**, 327–340 (2003).

39. Baurens, F. C., Nicolleau, J., Legavre, T., Verdeil, J. L. & Monteuuis, O. Genomic DNA methylation of juvenile and mature *Acacia mangium* micropropagated *in vitro* with reference to leaf morphology as a phase change marker. *Tree Physiol.* **24**, 401–407 (2004).
40. Yuan, Y. N., SanMiguel, P. J. & Bennetzen, J. L. High-Cot sequence analysis of the maize genome. *Plant J.* **34**, 249–255 (2003).
41. Springer, N. M. & Barbazuk, W. B. Utility of different gene enrichment approaches toward identifying and sequencing the maize gene space. *Plant Physiol.* **136**, 3023–3033 (2004).
A particularly balanced comparison of methylation filtration and Cot-based cloning and sequencing in maize.
42. Lamoureux, D., Peterson, D. G., Li, W., Fellers, J. P. & Gill, B. S. The efficacy of Cot-based gene enrichment in wheat (*Triticum aestivum* L.). *Genome* **48**, 1120–1126 (2005).
43. Fu, Y., Hsia, A. P., Guo, L. & Schnable, P. S. Types and frequencies of sequencing errors in methyl-filtered and high C(0)t maize genome survey sequences. *Plant Physiol.* **135**, 2040–2045 (2004).
44. Wicker, T. *et al.* The repetitive landscape of the chicken genome. *Genome Res.* **15**, 126–136 (2005).
45. Nelson, D. L. *et al. Alu* polymerase chain-reaction — a method for rapid isolation of human-specific sequences from complex DNA sources. *Proc. Natl Acad. Sci. USA* **86**, 6686–6690 (1989).
46. Batzer, M. A. & Deininger, P. L. *Alu* repeats and human genomic diversity. *Nature Rev. Genet.* **3**, 370–379 (2002).
47. Green, E. D. Strategies for the systematic sequencing of complex genomes. *Nature Rev. Genet.* **2**, 573–583 (2001).
48. Collins, F. S., Lander, E. S., Rogers, J. & Waterston, R. H. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
49. Marra, M. *et al.* High-throughput fingerprint analysis of large-insert clones. *Genome Res.* **7**, 1072–1084 (1997).
50. Soderlund, C., Humphrey, S., Dunham, A. & French, L. Contigs built with fingerprints, markers and FPC V4.7. *Genome Res.* **10**, 1772–1787 (2000).
51. Weiler, K. S. & Wakimoto, B. T. Heterochromatin and gene expression in *Drosophila*. *Annu. Rev. Genet.* **29**, 577–605 (1995).
52. Copenhaver, G. P. *et al.* Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* **286**, 2468–2474 (1999).
53. Nagaki, K. *et al.* Sequencing of a rice centromere uncovers active genes. *Nature Genetics* **36**, 138–145 (2004).
54. Zhang, Y. *et al.* Structural features of the rice chromosome 4 centromere. *Nucleic Acids Res.* **32**, 2023–2030 (2004).
55. Jiang, N., Bao, Z. R., Zhang, X. Y., Eddy, S. R. & Wessler, S. R. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**, 569–573 (2004).
56. Juretic, N., Hoen, D. R., Huynh, M. L., Harrison, P. M. & Bureau, T. E. The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Res.* **15**, 1292–1297 (2005).
57. D'Ennequin, M. L. T., Toupance, B., Robert, T., Godelle, B. & Gouyon, P. Plant domestication: a model for studying the selection of linkage. *J. Evol. Biol.* **12**, 1138–1147 (1999).
58. Paterson, A. H. What has QTL mapping taught us about plant domestication? *New Phytologist* **154**, 591–608 (2002).
59. Johnson, M. E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001).
60. Initiative, T. A. G. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
61. Matsumoto, T. *et al.* The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
62. Goff, S. A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). *Science* **296**, 92–100 (2002).
63. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). *Science* **296**, 79–92 (2002).
64. Yu, J. *et al.* The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* **3**, 266–281 (2005).
65. Gao, W. X. *et al.* Wide-cross whole-genome radiation hybrid mapping of cotton (*Gossypium hirsutum* L.). *Genetics* **167**, 1317–1329 (2004).
66. Kynast, R. G. *et al.* Dissecting the maize genome by using chromosome addition and radiation hybrid lines. *Proc. Natl Acad. Sci. USA* **101**, 9921–9926 (2004).
67. Aston, C., Mishra, B. & Schwartz, D. C. Optical mapping and its potential for large-scale sequencing projects. *Trends Biotechnol.* **17**, 297–302 (1999).
68. Boffelli, D., Nobrega, M. A. & Rubin, E. M. Comparative genomics at the vertebrate extremes. *Nature Rev. Genet.* **5**, 456–465 (2004).
70. Margulies, E. H. *et al.* An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA* **102**, 4795–4800 (2005).
A detailed consideration of the phylogenetic shadowing approach.
70. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98 (1973).
71. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
72. Eddy, S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3**, 95–102 (2005).
Describes the theoretical underpinnings of the phylogenetic shadowing approach.
73. Soltis, D. E. *et al.* Missing links: the genetic architecture of flower and floral diversification. *Trends Plant Sci.* **7**, 22–31 (2002).
74. Pryer, K. M., Schneider, H., Zimmer, E. A. & Banks, J. A. Deciding among green plants for whole genome studies. *Trends Plant Sci.* **7**, 550–554 (2002).
75. Sanderson, M. J., Thorne, J. L., Wikstrom, N. & Bremer, K. Molecular evidence on plant divergence times. *Am. J. Bot.* **91**, 1656–1665 (2004).
76. Bell, C. D., Soltis, D. E. & Soltis, P. S. The age of the angiosperms: a molecular timescale without a clock. *Evolution* **59**, 1245–1258 (2005).
77. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
78. Lin, Y. *et al.* A *Sorghum propinquum* BAC library, suitable for cloning genes associated with loss-of-function mutations during crop domestication. *Mol. Breed.* **5**, 511–520 (1999).
79. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
80. Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genet.* **32**, 135–142 (2002).
81. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
82. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
83. Ahmadi, K. R. *et al.* A single-nucleotide polymorphism tagging set for human drug metabolism and transport. *Nature Genet.* **37**, 84–89 (2005).
84. Wright, S. I. *et al.* The effects of artificial selection of the maize genome. *Science* **308**, 1310–1314 (2005).
85. Wang, R. L., Stec, A., Hey, J., Lukens, L. & Doebley, J. The limits of selection during maize domestication. *Nature* **398**, 236–239 (1999).
86. Thornsberry, J. M. *et al. Dwarf8* polymorphisms associate with variation in flowering time. *Nature Genet.* **28**, 286–289 (2001).
The seminal application of association genetics to the characterization of plant (maize) germplasm.
87. Gallavotti, A. *et al.* The role of barren stalk1 in the architecture of maize. *Nature* **432**, 630–635 (2004).
88. Wang, H. *et al.* The origin of the naked grains of maize. *Nature* **436**, 714–719 (2005).
89. Brown, A. H. D. Core collections — a practical approach to genetic-resources management. *Genome* **31**, 818–824 (1989).
90. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
91. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
References 90 and 91 describe next-generation DNA-sequencing technologies that promise further acceleration of sequence acquisition.
92. Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G. & Chee, M. S. A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genet.* **37**, 549–554 (2005).
A particularly promising genotyping assay that seems to be scalable to transcriptome- or even genome-wide applications.
93. Koo, B., Pardey, P. & Wright, B. The price of conserving agricultural biodiversity. *Nature Biotechnol.* **21**, 126–128 (2003).
94. Britten, R. J. & Davidson, E. H. Studies on nucleic-acid reassociation kinetics — empirical equations describing DNA reassociation. *Proc. Natl Acad. Sci. USA* **73**, 415–419 (1976).
95. Pearson, W. R., Davidson, E. H. & Britten, R. J. Program for least-squares analysis of reassociation and hybridization data. *Nucleic Acids Res.* **4**, 1727–1737 (1977).
96. Cech, T. R., Rosenfel. A & Hearst, J. E. Characterization of most rapidly renaturing sequences in mouse main-band DNA. *J. Mol. Biol.* **81**, 299–325 (1973).
97. Klein, H. L. & Welch, S. K. Inverted repeated sequences in yeast nuclear-DNA. *Nucleic Acids Res.* **8**, 4651–4669 (1980).
98. Paterson, A., Bowers, J., Peterson, D., Estill, J. & Chapman, B. Structure and evolution of cereal genomes. *Curr. Opin. Genet. Devel.* **13**, 644–650 (2003).
99. Paterson, A. H., Bowers, J. E. & Chapman, B. A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci. USA* **101**, 9903–9908 (2004).
100. Guyot, R. & Keller, B. Ancestral genome duplication in rice. *Genome* **47**, 610–614 (2004).
101. Wang, X., Shi, X., Hao, B. L., Ge, S. & Luo, J. Duplication and DNA segmental loss in rice genome and their implications for diploidization. *New Phytologist* **165**, 937–946 (2005).
102. Paterson, A. H., Bowers, J. E., Vandepoele, K. & Van de Peer, Y. Ancient duplication of cereal genomes. *New Phytologist* **165**, 658–661 (2005).
103. Vandepoele, K., Simillion, C. & Van de Peer, Y. Evidence that rice and other cereals are ancient aneuploids. *Plant Cell* **15**, 2192–2202 (2003).
104. Bedell, J. A. *et al.* Sorghum genome sequencing by methylation filtration. *PLoS Biol.* **3**, 103–115 (2005).
105. 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).
106. Messing, J. *et al.* Sequence composition and genome organization of maize. *Proc. Natl Acad. Sci. USA* **101**, 14349–14354 (2004).
107. Bremer, K. *et al.* An ordinal classification for the families of flowering plants. *Ann. Mo. Bot. Gard.* **85**, 531–553 (1998).
108. Soltis, D. E., Soltis, P. S., Endress, P. K. & Chase, M. W. *Phylogeny and Evolution of Angiosperms* (Sinauer Associates, Sunderland, Massachusetts, 2005).

Acknowledgements
Thanks to J. Bowers, P. Brown, C. dePamphilis, J. Estill, J. Giovannoni, S. Kresovich, R. Mauricio, J. McNeal, C. Peterson, D. Peterson, J. Shaw, P. Soltis, H. Tang, S. Tanksley, N. Young and others for helpful data and discussions, and the US National Science Foundation, US Department of Agriculture, International Consortium for Sugarcane Biotechnology and US Golf Association for financial support.

Competing interests statement
The author declares no competing financial interests.

FURTHER INFORMATION
Consultative Group on International Agricultural Research: http://www.cgiar.org
GenBank: http://www.ncbi.nlm.nih.gov/Genbank/index.html
Joint Genome Institute Community Sequencing Program — Sequencing Plans for 2006: http://www.jgi.doe.gov/sequencing/cspseqplans2006.html
Joint Genome Institute *Populus trichocarpa* sequencing project: http://genome.jgi-psf.org/Poptr1/Poptr1.home.html
***Medicago truncatula* sequencing resources:** http://www.medicago.org/genome
Multinational Brassica Genome Project: http://www.brassicagenome.org
The International Tomato Sequencing Project: http://sgn.cornell.edu/help/about/tomato_sequencing.html
The Potato Genome Sequencing Consortium: http://www.potatogenome.net
The US Department of Agriculture Germplasm Resources Information Network: http://www.ars-grin.gov

SUPPLEMENTARY INFORMATION
See online article: S1 (table)
Access to this links box is available online.