ELSEVIER

Comparative genome analysis of monocots and dicots, toward characterization of angiosperm diversity

Andrew H Paterson[1],[*], John E Bowers[1], Brad A Chapman[1],
Daniel G Peterson[2], Junkang Rong[1] and Thomas M Wicker[1]

The importance of angiosperms to sustaining humanity by providing a wide range of 'ecosystem services' warrants increased exploration of their genomic diversity. The nearly completed sequences for two species representing the major angiosperm subclasses, specifically the dicot *Arabidopsis thaliana* and the monocot *Oryza sativa*, provide a foundation for comparative analysis across the angiosperms. The angiosperms also exemplify some challenges to be faced as genomics makes new inroads into describing biotic diversity, in particular polyploidy (genome-wide chromatin duplication), and much larger genome sizes than have been studied to date.

Addresses
[1]Plant Genome Mapping Laboratory, University of Georgia, Athens GA 30602, USA
[2]Mississippi Genome Exploration Laboratory, Mississippi State University, Mississippi State MS 39762, USA
[*]e-mail: paterson@plantbio.uga.edu

Abbreviations
CBCS Cot-based cloning and sequencing
EST expressed sequence tag
SNP single-nucleotide polymorphism
STS sequence-tagged site

Introduction

The angiosperms, or flowering plants, provide ecosystem services including oxygen, fuel, medicines, erosion and flood control, soil regeneration, and other benefits [1] that are absolutely essential to humanity and indeed are a cornerstone of the global ecosystem. The 'domestication' of about 200 angiosperms to provide most of the world's supply of food, feed and fibre has largely determined our ability to sustain modern human populations and has also empowered human social development [2]. A small subset of domesticates, plus a few botanical models such as *Arabidopsis thaliana*, account for most of our present knowledge of the repertoire, organization and function of plant genes.

The past two decades of plant molecular genetics research, and in particular the past few years of high-throughput genomics, have set the stage for new advances in comparative biology. For the first time, we have access to large numbers (and in some cases all) of the genes in a genome, albeit for a small subset of angiosperms. Now we can begin the long process of sifting through the many molecular-level differences that have accumulated during the approximately 170–235 million years [3] since the angiosperms diverged from a common ancestor, to seek specific changes that contribute to variation in life history traits, biochemistry, morphology and development, and adaptation to the biotic and abiotic environment.

While comparative biology offers valuable insight into divergence at many taxonomic levels, of particular interest is comparison of members of the two major angiosperm subclasses, monocots and dicots. The largely finished sequence of the dicot *Arabidopsis* [4], together with the rapidly progressing sequence of the monocot *Oryza* (rice) [5••–9••] provide a natural framework for this work. Genetic maps, physical maps and expressed sequence tag (EST) resources for a host of additional taxa permit early assessments of diversity within each of the angiosperm subclasses, and provide important contextual information by which to better relate major events in the *Arabidopsis* and *Oryza* lineages to the plant family tree. In this review, we explore early messages arising from comparison of the content and organization of monocot and dicot genomes, address key consequences of polyploidy for angiosperm comparative genomics, and compare and contrast methods that are likely to be important to further description and study of angiosperm genomic diversity.

Gene repertoire

Many functions in diverse eukaryotes are directed by genes that exhibit much similarity at the amino acid and even nucleotide level [10], including the angiosperms. The *Arabidopsis* transcriptome is currently estimated to include 30 078 genes (http://www.ncbi.nlm.nih.gov). The rice transcriptome appears to be more complex, with estimates based on genomic shotgun sequencing of 46 022–55 615 genes [9••] and 32 277–61 668 genes [5••]. Higher estimates based on finished sequencing (62 500 genes [6••]) may reflect more effective gene prediction. For example, the finished rice chromosome 10 sequence contains 3471 predicted genes, but the corresponding shotgun reads contain only 1724 [8••], although this

difference partly results from the inclusion of transposable element-related genes in the chromosome 10 count.

About 80.6% [9••] to 85% [5••] of *Arabidopsis* predicted proteins are homologous to rice predicted proteins, with average identity of 49.5% and modal identity of 33% [5••]. Curiously, these findings are not transitive — much lower frequences of rice genes show matches to *Arabidopsis* (47% [6••]; 49.4% [9••]; 43.8% [7••]; 67% [8••]). Factors that influence this may include differences in GC content between rice and *Arabidopsis* genes [9••] or a greater abundance of retroelement-like genes in rice. Most *Arabidopsis* predicted proteins that lack homology to rice are classified as 'hypothetical' (or other similar terms), suggesting the possibility of mis-annotation [5••,6••].

Although many genes and functions are widely distributed across the tree of life, there may exist substantial populations of angiosperm-specific genes, as well as monocot- and dicot-specific genes. About 8000 (30%) *Arabidopsis* genes are found in the rice ssp. japonica shotgun sequence, but not in *Drosophila, Caenorhabditis elegans, Saccharomyces* or sequenced bacterial genomes [5••]. Analysis of 33 620 unigenes for the monocot sugarcane [11•] showed that 82% had matches to the rice genome, versus 71% with matches to the *Arabidopsis* genome, perhaps suggesting that 11% (roughly 3600) may be monocot-specific.

Chromosome and genome organization
Given that the vast majority of angiosperms lack complete sequences, genetic maps continue to be a central tool for studying their chromosome organization. Most major crops, and many botanical models, enjoy detailed sequence-tagged site (STS)-based genetic recombination maps that are suitable not only for comparative biology, but also for crop improvement. While these maps have been successfully applied to many needs using traditional restriction-fragment length polymorphism or simple sequence repeat based methods, genetically mapped STSs can readily be used to discover single-nucleotide polymorphism (SNPs) or small insertion/deletion polymorphisms [12•] that can then be genotyped by a wide range of more economical SNP-based technologies. The ability to acquire such polymorphism information for corresponding loci in many genotypes increases the value of STS maps and reduces the costs associated with their wider utilization.

Limitations to the centiMorgan-scale resolution of genetic recombination maps might be improved to kilobase level by their integration with physical maps based upon large-insert clones such as bacterial artificial chromosomes (BACs) [13,14]. 'Gene mapping' by hybridization of cloned or synthetic DNA probes [15] to large-insert libraries offers many of the advantages of somatic cell genetics, in particular obviating the need for genetic

polymorphism which may impose a bias on the subsets of DNA probes that can be 'mapped.' Similarly, mapping based either upon radiation hybrids [16] or on genetic stocks containing partial deletions of individual chromosomes [17•] have accelerated progress in genomics research for taxa with few DNA polymorphisms.

Ancient polyploidy and its consequences
Comparative studies of plant chromosome evolution show important differences from early results in animals. Gene order conservation along the chromosomes of vertebrates is evident after hundreds of millions of years of divergence [18,19], but comparisons of the *Arabidopsis* sequence to partial gene orders of other angiosperms (flowering plants) sharing common ancestry ∼170–235 million years ago [3] have yielded conflicting results. Although gene order conservation is considerable in confamilial taxa such as *Arabidopsis* and *Brassica* ([20,21,22•]), and even in diverse dicots [23•], comparison of the *Arabidopsis* sequence to selected fully sequenced rice BACs or contigs have led to disparate conclusions ranging from 'scant collinearity' [24,25] to 'frameworks of conserved genes' [26].

The recurring observation of 'networks of synteny' [27], with target regions of rice [26], tomato [21,27], soybean [28,29,30•], and *Medicago truncatula* [31•] showing non-random relationships with multiple unlinked regions of *Arabidopsis* was an important clue to resolving the seeming difference in rates of genome structural evolution between plants and animals. Many angiosperm genomes have been through one or more genome-wide duplication or 'polyploidization' events [32,33]. Early hints at the possibility of duplication even in the small genome of *Arabidopsis* [34,35] were borne out by detailed analysis of the nearly finished sequence, revealing widespread duplication accompanied by loss of many duplicated gene copies [4,36–38].

Recent progress in revealing the history of ancient duplication events clarifies our understanding of plant chromosome evolution. Ancient duplication has two major consequences for comparative genomics. First, it appears to be followed by 'diploidization', or loss of many single members of homologous pairs, obscuring and complicating analysis of colinearity. This process is initially rapid [39–42], but continues for a long time [43••]. Second, knowledge of the timing of duplication events relative to divergence of taxa from a common ancestor is essential [44,45••]. Only if taxon divergence postdates duplication are traditional 'one-to-one' genomic comparisons sufficient. If duplication in one or both lineages postdates taxon divergence, more complex approaches are needed. By using a phylogenomic approach to relate specific duplication events to the plant family tree, together with finished sequence information to infer the likely gene order in hypothetical ancestors of modern duplicated

chromosomal segments, the level of gene order conservation discerned in diverse angiosperm lineages is improved [43••]. Early evidence in *Oryza* [46–48] also reflects widespread [49•], perhaps genome-wide [50•], duplication.

Analysis of large-scale duplications has necessitated the development of new bioinformatics tools, going beyond whole-genome alignments that rely on the presence of unique sequence matches [51,52]. Especially promising alternatives are FISH (Fast Identification of Segmental Homology) [53•] and ADHoRe (Automatic Detection of Homologous Regions) [54•]. Programs for homolog identification and phylogenomic analysis of specific duplication events are also available [55•].

Further insights into angiosperm genomic diversity

While botanical models provide seminal information that can be extrapolated to a degree by comparative approaches, comprehensive information about angiosperm diversity will require detailed exploration of many additional genomes. The greatest challenge to their widespread genomic analysis, and a practical motivation for many comparative genomics efforts, is that angiosperms exhibit about 1000-fold variation in genome size due mostly to repetitive DNA. EST sequencing is a first step toward further characterization of angiosperm genomic diversity. More than 20 angiosperm species, representing many diverse branches of the plant family tree, each enjoy more than 10 000 ESTs in GenBank at time of writing, and the number of species and ESTs is growing rapidly.

As EST sequencing reaches diminishing returns (typically at ~50% of the genes in a genome), two new approaches show promise toward completing the sets of gene sequences from large-genome taxa. 'Methyl filtration' based upon degrees of differential methylation of expressed versus non-expressed sequences [56,57•] reduces the abundance of repetitive DNA in plant (but not animal [58•]) genomic DNA libraries. Cot-based cloning and sequencing (CBCS) [59••–61••] involves the fractionation of a genome into 'components' based on the degree of sequence repetition (Cot analysis) [62,63•], followed by cloning and sequencing of corresponding clone libraries to a depth appropriate to represent the 'sequence complexity' of the respective component(s).

Perhaps the most important difference between these methods lies in the implicit assumptions made about the nature of the DNA that is 'filtered'. By accessing only hypomethylated DNA, methyl filtration is subject to the variable relationship between methylation and gene expression across genes and taxa, which has been reviewed in detail [59••,60••]. Differences in methylation associated with abiotic stresses (radiation [64], tissue culture [65,66]) raise new questions about the stability of this relationship. Although its validation by comparison of hypomethylated DNA to random genomic DNA shows enrichment for known genes [56], the higher 'genome reduction factor' afforded by methyl filtration [67] might reflect loss of many genes that are methylated to some degree. By contrast, CBCS provides access to the entire genome. For this reason, its validation was necessarily different from that for methyl filtration, comparing specific quantifiable properties of different genomic fractions to one another. To minimize the risk that repetitive CBCS clones contained parts of two or more different element families (thus obscuring empirical verification of their copy number), it was essential that validation be performed on DNA sheared to ~300 nucleotides, together with removal (after Cot hybridization) of single-stranded overhangs by mung bean nuclease [59••].

CBCS is flexible to a wide range of permutations [59••] based on the biology of the system and the goals of the investigator. For example, capturing the sequence complexity of the low-copy DNA in a genome would be made more efficient by using longer clones than were appropriate for validation studies [59••,60••]. In the well-studied cereals, for example, the distance between genes appears to be correlated with differences in genome size in different taxa, but with noteworthy exceptions in the form of 'gene-rich' regions that largely lack repetitive DNA [68]. Such gene-rich regions should be well-covered by Cot clones that are long enough (1–2 kb) to offer sequencing economies and sufficient information for assembly [59••,60••,69]. Depending on the (genome-specific) number and dispersion patterns of repetitive DNA families, at some point increased DNA fragment length may tend to cause under-representation of terminal (5′ and 3′) regions of many genes, plus entire short genes in close proximity to repetitive DNA. Sequencing clones from multiple Cot libraries sheared to different average fragment sizes may provide the best balance between gene discovery and sequencing economics.

In principle, one could envision superimposing CBCS on methyl filtration, or vice versa; however, more information is needed to determine the cost-benefit balance of this approach. Some have argued [57•] that the use of host cells with intermediate tolerance of methylated DNA might have affected CBCS validation studies; however, these arguments are invalid, failing to note that the efficacy of genomic fractionation by CBCS was demonstrated by empirical determination of copy number by hybridization studies, by empirical comparison of the highly repetitive (HRCot) fraction to a sampling of random genomic DNA to demonstrate that it did indeed represent the percentage of the genomic DNA estimated (from the Cot curve) to be highly repetitive, and through detailed annotation and characterization of the Cot sequences showing that they were comprised of DNA element types appropriate to the respective fractions [59••,61••].

Finally, knowledge of the sequences and distribution of the repetitive DNA that accounts for most of angiosperm genomic diversity is of much value. Knowledge of repetitive DNA improves EST and genome annotation [70••], and better understanding of its physical distribution could help to identify 'gene-rich' genomic domains that are priorities for early sequencing [68]. Use of the 'Alu' element family empowered many advances in human genome research far before the sequence was available [71]. CBCS is ideally suited to direct analysis of repetitive DNA. Only 253 sequences from the sorghum HRCot fraction were sufficient to account for 15% of its genomic DNA, exactly the fraction predicted by Cot analysis [59••]. A recent study (TM Wicker *et al.*, unpublished) describes characterization of the majority of repetitive DNA in an entire genome by CBCS. Methyl filtration eliminates some repetitive DNA from sequencing libraries, but because some plant transposable elements are hypomethylated it is less effective than CBCS at separating the two fractions, at least in maize [72•].

Conclusions

The identification of multiple polyploidization events in the *Arabidopsis* lineage, together with methods to mitigate the effects of these events on comparative genomics, sets the stage for a re-evalation of gene order conservation across diverse angiosperms. The *Oryza* sequence will provide the information needed to study the course of monocot genome evolution, and then to perform truly orthologous comparisons within and among monocots and dicots. Detailed study of these two lineages will provide a framework of gene orders and sequences valuable to future analyses of other angiosperm genomes (whether completely sequenced, or represented as STS-based genetic maps). Synteny information about monocots and dicots will permit new inferences about the probable genomic organization of common ancestors of the angiosperms, and foster exploration of possible parallels with more distantly related taxa.

Much additional information from many more taxa will be needed to elucidate the specific events responsible for the morphological and physiological diversity that adapts different angiosperms to different ecological niches, crop production systems and human needs. Selected angiosperms have been domesticated because they exhibit one or more extraordinary features, such as the large carbohydrate-rich seeds of the cultivated cereals, the remarkably long and strong single-celled fibres of cotton, the curd-like semi-sterile inflorescence of cauliflower, and the bulbous berry of tomato. Each crop is an elegant 'model' that offers unique opportunities to make new advances in (comparative) plant biology, but will ultimately require detailed genomic exploration. Efficient new methods promise that such information will grow at an accelerating rate.

References and recommended reading
Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Tilman D, Cassman KG, Matson PA, Naylor R, Polasky S: **Agricultural sustainability and intensive production practices**. *Nature* 2002, **418**:671-677.

2. Raven P, Evert R, Eichhorn S: *Biology of Plants*. New York: Worth Publishers, Inc.; 1992.

3. Yang YW, Lai KN, Tai PY, Li WH: **Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages**. *J Mol Evol* 1999, **48**:597-604.

4. Initiative TAG: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana***. *Nature* 2000, **408**:796-815.

5. Goff SA, Ricke D, Lan TH, Presting G, Wang RL, Dunn M,
•• Glazebrook J, Sessions A, Oeller P, Varma H *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp japonica)**. *Science* 2002, **296**:92-100.
See annotation for [9••].

6. Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T,
•• Katayose Y, Wu JZ, Niimura Y, Cheng ZK, Nagamura Y *et al.*: **The genome sequence and structure of rice chromosome 1**. *Nature* 2002, **420**:312-316.
See annotation for [7••].

7. Feng Q, Zhang YJ, Hao P, Wang SY, Fu G, Huang YC, Li Y,
•• Zhu JJ, Liu YL, Hu X *et al.*: **Sequence and analysis of rice chromosome 4**. *Nature* 2002, **420**:316-320.
Published simultaneously, these papers represent the first essentially completely sequenced rice chromosomes. The article by Sasaki *et al.* [6••] is distinguished in also reporting the largest rice chromosome.

8. Consortium TRCS: **In-depth view of structure, activity, and
•• evolution of rice chromosome 10**. *Science* 2003, **300**:1566-1569.
The most recent of the rice chromosome sequences, for the smallest and most heterochromatic of the chromosomes, provides an especially integrative picture of the properties of a chromosome.

9. Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, Deng YJ, Dai L,
•• Zhou Y, Zhang XQ *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp indica)**. *Science* 2002, **296**:79-92.
Published simultaneously, these articles [5••,9••] comprise the first genomic shotgun sequence of a cereal, and shed light on the genomic diversity of cereals from dicot plants such as *Arabidopsis* as well as non-plant taxa.

10. Rubin GM, Yandell MD, Wortman JR, Miklos GLG, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W *et al.*: **Comparative genomics of the eukaryotes**. *Science* 2000, **287**:2204-2215.

11. Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM,
• Ferro MIT, Henrique-Silva F, Giglioti EA, Lemos MVF, Coutinho LL *et al.*: **Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane**. *Genome Res* 2003, **13**:2725-2735.
One of the most detailed of plant EST databases.

12. Nasu S, Suzuki J, Ohta R, Hasegawa K, Yui R, Kitazawa N,
• Monna L, Minobe Y: **Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (*Oryza sativa, Oryza rufipogon*) and establishment of SNP markers**. *DNA Res* 2002, **9**:163-171.
Among the first illustrations of a genome-wide SNP map in plants, including early quantification of rates and types of SNPs within Oryza, and demonstration of an uneven genomic landscape of SNP distribution.

13. Draye X, Lin Y-R QX-y, Bowers JE, Burow GB, Morrell PL, Peterson DG, Presting GG, Ren SX, Wing RA, Paterson AH: **Toward integration of comparative genetic, physical, diversity,**

and cytomolecular maps for grasses and grains, using the *Sorghum* genome as a foundation. *Plant Physiol* 2001, **125**:1325-1341.

14. Coe E, Cone K, McMullen M, Chen SS, Davis G, Gardiner J, Liscum E, Polacco M, Paterson A, Sanchez-Villeda H *et al.*: **Access to the maize genome: an integrated physical and genetic map**. *Plant Physiol* 2002, **128**:9-12.

15. Cai WW, Reneker J, Chow CW, Vaishnav M, Bradley A: **An anchored framework BAC map of mouse chromosome 11 assembled using multiplex oligonucleotide hybridization**. *Genomics* 1998, **54**:387-397.

16. Kynast RG, Riera-Lizarazu O, Vales MI, Okagaki RJ, Maquieira SB, Chen G, Ananiev EV, Odland WE, Russell CD, Stec AO *et al.*: **A complete set of maize individual chromosome additions to the oat genome**. *Plant Physiol* 2001, **125**:1216-1227.

17. Sorrells ME, La Rota M, Bermudez-Kandianis CE,
•   Greene RA, Kantety R, Munkvold JD, Miftahudin, MA, Ma XF, Gustafson PJ *et al.*: **Comparative DNA sequence analysis of wheat and rice genomes**. *Genome Research* 2003, **13**:1818-1827
The most detailed picture to date of the alignment of the large and complex polyploid genome of wheat to rice.

18. Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GLG, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J *et al.*: **A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome**. *Science* 2002, **296**:1661-1671.

19. Smith SF, Snell P, Gruetzner F, Bench AJ, Haaf T, Metcalfe JA, Green AR, Elgar G: **Analyses of the extent of shared synteny and conserved gene orders between the genome of *Fugu rubripes* and human 20q**. *Genome Res* 2002, **12**:776-784.

20. Lan TH, DelMonte TA, Reischmann KP, Hyman J, Kowalski SP, McFerson J, Kresovich S, Paterson AH: **An EST-enriched comparative map of *Brassica oleracea* and *Arabidopsis thaliana***. *Genome Res* 2000, **10**:776-788.

21. Rossberg M, Theres K, Acarkan A, Herrero R, Schmitt T, Schumacher K, Schmitz G, Schmidt R: **Comparative sequence analysis reveals extensive microcolinearity in the Lateral suppressor regions of the tomato, *Arabidopsis*, and Capsella genomes**. *Plant Cell* 2001, **13**:979-988.

22. Lukens L, Zou F, Lydiate D, Parkin I, Osborn T: **Comparison of a
•   *Brassica oleracea* genetic map with the genome of *Arabidopsis thaliana***. *Genetics* 2003, **164**:359-372.
A more rigorous approach to *Brassica-Arabidopsis* synteny than prior approaches.

23. Stirling B, Yang ZK, Gunter LE, Tuskan GA, Bradshaw HD:
•   **Comparative sequence analysis between orthologous regions of the *Arabidopsis* and Populus genomes reveals substantial synteny and microcollinearity**. *Canadian Journal of Forest Research* 2003, **33**:2245-2251.
An albeit early glimpse at the comparative messages that are likely to derive from the second genomic sequence for a dicot plant.

24. Liu H, Sachidanandam R, Stein L: **Comparative genomics between rice and *Arabidopsis* shows scant collinearity in gene order**. *Genome Res* 2001, **11**:2020-2026.

25. Devos KM, Beales J, Nagamura Y, Sasaki T: ***Arabidopsis* — rice: will colinearity allow gene prediction across the eudicot-monocot divide?** *Genome Res* 1999, **9**:825-829.

26. Mayer K, Murphy G, Tarchini R, Wanbutt R, Volckaert G, Pohl T, Dusterhoft, Andreas, Stiekema W, Entian K-D, Terryn N, Lemcke K *et al.*: **Conservation of microstruction between a sequenced region of the genome of rice and multiple segments of the genome of *Arabidopsis thaliana***. *Genome Res* 2001:1-8.

27. Ku HM, Vision T, Liu JP, Tanksley SD: **Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny**. *Proc Natl Acad Sci USA* 2000, **97**:9121-9126.

28. Grant D, Cregan P, Shoemaker RC: **Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis***. *Proc Natl Acad Sci USA* 2000, **97**:4168-4173.

29. Lee M, Sharopova N, Beavis WD, Grant D, Katt M, Blair D, Hallauer A: **Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population**. *Plant Mol Biol* 2002, **48**:453-461.

30. Yan HH, Mudge J, Kim DJ, Larsen D, Shoemaker RC, Cook DR,
•   Young ND: **Estimates of conserved microsynteny among the genomes of Glycine max**. *Theor Appl Genet* 2003, **106**:1256-1265.
See annotation for [31•].

31. Zhu HY, Kim DJ, Baek JM, Choi HK, Ellis LC, Kuester H,
•   McCombie WR, Peng HM, Cook DR: **Syntenic relationships between *Medicago truncatula* and *Arabidopsis* reveal extensive divergence of genome organization**. *Plant Physiol* 2003, **131**:1018-1026.
These papers provide early glimpses into what is likely to be the third dicot genome to be sequenced.

32. Stebbins G: **Chromosomal variation and evolution; polyploidy and chromosome size and number shed light on evolutionary processes in higher plants**. *Science* 1966, **152**:1463-1469.

33. Masterson J: **Stomatal size in fossil plants — evidence for polyploidy in majority of angiosperms**. *Science* 1994, **264**:421-424.

34. McGrath JM, Jancso MM, Pichersky E: **Duplicate sequences with a similarity to expressed genes in the genome of *Arabidopsis thaliana***. *Theor Appl Genet* 1993, **86**:880-888.

35. Kowalski SP, Lan TH, Feldmann KA, Paterson AH: **Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveals islands of conserved organization**. *Genetics* 1994, **138**:499-510.

36. Paterson AH, Bowers JE, Burow MD, Draye X, Elsik CG, Jiang CX, Katsar CS, Lan TH, Lin YR, Ming RG *et al.*: **Comparative genomics of plant chromosomes**. *Plant Cell* 2000, **12**:1523-1539.

37. Blanc G, Barakat A, Guyot R, Cooke R, Delseny I: **Extensive duplication and reshuffling in the *Arabidopsis* genome**. *Plant Cell* 2000, **12**:1093-1101.

38. Vision T, Brown D, Tanksley S: **The origins of genomic duplications in *Arabidopsis***. *Science* 2000, **290**:2114-2117.

39. Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA: **Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat**. *Plant Cell* 2001, **13**:1749-1759.

40. Ozkan H, Levy AA, Feldman M: **Allopolyploidy-induced rapid genome evolution in the wheat (Aegilops-Triticum) group**. *Plant Cell* 2001, **13**:1735-1747.

41. Kashkush K, Feldman M, Levy AA: **Gene loss, silencing and activation in a newly synthesized wheat allotetraploid**. *Genetics* 2002, **160**:1651-1659.

42. Eckhardt N: **A sense of self: the role of DNA sequence elimination in allopolyploidization**. *Plant Cell* 2001, **13**:1699-1704.

43. Bowers JE, Chapman BA, Rong JK, Paterson AH: **Unravelling
•• angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events**. *Nature* 2003, **422**:433-438.
An early application of 'phylogenomics', a merger of phylogenetic inference with structural genomic information that is expected to be of growing importance in the analysis and interpretation of genomic data.

44. Bowers JE, Abbey C, Anderson S, Chang C, Draye X, Hoppe AH, Jessup R, Lemke C, Lennington J, Li ZK *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* 2003, **165**:367-386.

45. Kellogg EA: **It's all relative**. *Nature* 2003,
•• **422**:383-384.
A succinct and illustrative critique of new methods described by Bowers *et al.* [43••] and elsewhere for application of 'phylogenomics' to the analysis of plant chromosome evolution.

46. Kishimoto N, Higo H, Abe K, Arai S, Saito A, Higo K: **Identification of the duplicated segments in rice chromosomes 1 and 5 by**

linkage analysis of cDNA markers of known functions.
Theor Appl Genet 1994, **88**:722-726.

47. Nagamura Y, Inoue T, Antonio B, Shimano T, Kajiya H, Shomura A, Lin S, Kuboki Y, Harushima Y, Kurata N *et al.*: **Conservation of duplicated segments between rice chromosomes 11 and 12**. *Breed Sci* 1995, **45**:373-376.

48. Wang SP, Liu KD, Zhang QF: **Segmental duplications are common in rice genome**. *Acta Bot Sin* 2000, **42**:1150-1155.

49. Vandepoele K, Simillion C, Van de Peer Y: **Evidence that rice and
• other cereals are ancient aneuploids**. *Plant Cell* 2003, **15**:2192-2202.
See annotation for [50•].

50. Paterson A, Bowers J, Peterson D, Estill J, Chapman B: **Structure
• and evolution of cereal genomes**. *Curr Opin Genet Dev* 2003, **13**:644-650.
These papers provide early glimpses into the duplication history of the rice genome, albeit by different methods and with somewhat different conclusions.

51. Delcher AL, Phillippy A, Carlton J, Salzberg SL: **Fast algorithms for large-scale genome alignment and comparison**. *Nucleic Acids Res* 2002, **30**:2478-2483.

52. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL: **Alignment of whole genomes**. *Nucleic Acids Res* 1999, **27**:2369-2376.

53. Calabrese PP, Chakravarty S, Vision TJ: **Fast identification and
• statistical evaluation of segmental homologies in comparative maps**. *Bioinformatics* 2003, **19(Suppl 1)**:i74-i80.
See annotation for [55•].

54. Vandepoele K, Saeys Y, Simillion C, Raes J, Van de Peer Y:
• **The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice**. *Genome Res* 2002, **12**:1792-1801.
See annotation for [55•].

55. Chapman BA, Bowers JE, Schulze SR, Paterson AH:
• **A comparative phylogenetic approach for dating whole-genome duplications**. *Bioinformatics* 2004, **20**: in press.
Describe new computational methods for identifying genomic duplication and the application of these methods to selected test cases.

56. Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, Stein L, McCombie WR, Martienssen RA: **Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome**. *Nat Genet* 1999, **23**:305-308.

57. Rabinowicz PD, McCombie WR, Martienssen RA: **Gene
• enrichment in plant genomic shotgun libraries**. *Curr Opin Plant Biol* 2003, **6**:150-156.
A valuable recent update on the status of methyl-filtration based methods [56], although including errors and incomplete information about CBCS.

58. Rabinowicz PD, Palmer LE, May BP, Hemann MT, Lowe SW, McCombie WR, Martienssen RA: **Genes and transposons are
• differentially methylated in plants, but not in mammals**. *Genome Res* 2003, **13**:2658-2664.
Demonstrates that the methyl filtration approach is limited to plants, and is not applicable to animals.

59. Peterson DG, Schulze SR, Sciara EB, Lee SA, Bowers JE, Nagel A,
•• Jiang N, Tibbitts DC, Wessler SR, Paterson AH: **Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery**. *Genome Res* 2002, **12**:795-807.
See annotation for [61••].

60. Peterson DG, Wessler SR, Paterson AH: **Efficient capture of
•• unique sequences from eukaryotic genomes**. *Trends Genet* 2002, **18**:547-550.
See annotation for [61••].

61. Peterson DG, Schulze SR, Sciara EB, Lee SA, Bowers JE, Nagel A,
•• Ning J, Tibbitts DC, Wessler SR, Paterson AH: **Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery**. URL: http://www.ncbi.nlm.nih.gov/entrez. 2001.
Merger of a classical approach with modern methods to 'fractionate' genomes based on the degree of repetition of DNA element families. An alternative to 'genomic shotgun sequencing' that may permit us to economically access the sequence complexity (set of unique sequences) in large, highly repetitive genomes.

62. Britten RJ, Kohne DE: **Repeated sequences in DNA**. *Science* 1968, **161**:529.

63. Goldberg RB: **From cot curves to genomics. How gene cloning
• established new concepts in plant biology**. *Plant Physiol* 2001, **125**:4-8.
Much of science advances by successive generations 'standing on the shoulders of giants'. This review by one such giant, illustrates the many connections of contemporary genomics to its 30 year old roots in methods that are no longer known to many readers (see [62]).

64. Kovalchuk O, Burke P, Arkhipov A, Kuchma N, James SJ, Kovalchuk I, Pogribny I: **Genome hypermethylation in *Pinus silvestris* of Chernobyl — a mechanism for radiation adaptation?** *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis* 2003, **529**:13-20.

65. Meng L, Bregitzer P, Zhang SB, Lemaux PG: **Methylation of the exon/intron region in the Ubi1 promoter complex correlates with transgene silencing in barley**. *Plant Mol Biol* 2003, **53**:327-340.

66. Baurens FC, Nicolleau J, Legavre T, Verdeil JL, 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* 2004, **24**:401-407.

67. Barbazuk B, Whitelaw CA, Quackenbush J, Schubert K, Beachy R, Lakey N, Bennetzen JL: **Consortium for Maize Genomics — an examination of maize gene coverage obtained from shotgun sequences derived from methyl-filtered and high Cot selection libraries**. In *Proceedings of the International Plant and Animal Genomes* XII *Conference*. San Diego, CA, 10–14 Jan 2004: http://www.intl-pag.org/12/abstracts/W50_PAG12_233.html.

68. Feuillet C, Keller B: **High gene density is conserved at syntenic loci of small and large grass genomes**. *Proc Natl Acad Sci USA* 1999, **96**:8265-8270.

69. Yuan YN, SanMiguel PJ, Bennetzen JL: **High-Cot sequence analysis of the maize genome**. *Plant J* 2003, **34**:249-255.

70. Wicker T, Guyot R, Yahiaoui N, Keller B: **CACTA transposons in
•• Triticeae. A diverse family of high-copy repetitive elements**. *Plant Physiol* 2003, **132**:52-63.
An elegant example of how better understanding of repetitive DNA may improve genome annotation.

71. Batzer MA, Deininger PL: **Alu repeats and human genomic diversity**. *Nat Rev Genet* 2002, **3**:370-379.

72. Whitelaw CA, Barbazuk WB, Pertea G, Chan AP, Cheung F, Lee Y,
• Zheng L, van Heeringen S, Karamycheva S, Bennetzen JL *et al.*: **Enrichment of gene-coding sequences in maize by genome filtration**. *Science* 2003, **302**:2118-2120.
The first demonstration of genome-wide application of both methyl filtration [56,57•] and CBCS [59••–61••] in plants.