ELSEVIER

Sequencing the maize genome

Robert A Martienssen[1], Pablo D Rabinowicz, Andrew O'Shaughnessy and W Richard McCombie

Sequencing of complex genomes can be accomplished by enriching shotgun libraries for genes. In maize, gene-enrichment by copy-number normalization (high C_0t) and methylation filtration (MF) have been used to generate up to two-fold coverage of the gene-space with less than 1 million sequencing reads. Simulations using sequenced bacterial artificial chromosome (BAC) clones predict that 5× coverage of gene-rich regions, accompanied by less than 1× coverage of subclones from BAC contigs, will generate high-quality mapped sequence that meets the needs of geneticists while accommodating unusually high levels of structural polymorphism. By sequencing several inbred strains, we propose a strategy for capturing this polymorphism to investigate hybrid vigor or heterosis.

Addresses
Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA
[1]e-mail: martiens@cshl.org

Current Opinion in Plant Biology 2004, **7**:102–107

This review comes from a themed issue on
Genome studies and molecular genetics
Edited by Joseph R Ecker and Doug Cook

1369-5266/$ – see front matter

DOI 10.1016/j.pbi.2004.01.010

Abbreviations

BAC	bacterial artificial chromosome
C_0	DNA concentration at time zero
EST	expressed sequence tag
MF	methylation filtration
t	re-association time

Introduction: the genome of maize

The maize genome has been characterized at cytogenetic, genetic and molecular levels, and the emerging picture is one of highly dynamic chromosomes. Interstitial heterochromatin in maize is highly polymorphic, in both size and location in the chromosome complement [1]. These chromosome 'knobs' resemble pericentromeric heterochromatin, consisting of tens of Mbp of satellite repeat sequence interspersed with retrotransposons [2]. Euchromatic regions are more complex, but actually have a higher density of transposon insertions. Most retrotransposons prefer to insert within each other, resulting in nested groups of transposons in between genes [3]. By contrast, recently active DNA transposons, such as *Mutator* and miniature inverted-repeat transposable elements (MITEs), insert preferentially into genes, and can be used in gene-enrichment sequencing strategies [4–6].

1–2% of random shotgun reads from maize match annotated exons in GenBank [7,8] and, with an average coding region of 1 kbp [9–11] and a genome size of 2500 Mbp [12], this indicates that there is a total of between 25 000 and 50 000 maize genes. Sequencing of approximately 5 Mbp of maize DNA cloned into bacterial artificial chromosomes (BACs) has revealed that maize genes are organized into islands of 10–20 kbp, each containing 3–4 genes on average [3,13••,14]. Some BACs have no genes, indicating that the transposon 'ocean' that separates gene-islands can have wide uninterrupted straits.

Maize is a segmental allotetraploid, generated approximately 11 million years ago (Mya) by the combination of two progenitor genomes that had diverged 10 million years before that [15]. More recently, a major retrotransposon expansion occurred, doubling genome size [16], and the homoeologous chromosome pairs were rearranged into segmental duplications that range from an entire chromosome arm to just a few centiMorgans (cM). The resulting pattern of duplicated linkage groups has been statistically verified [17,18], but sequence analysis of homoeologous regions corresponding to the *ALCOHOL DEHYDROGENASE1* (*ADH1*) gene showed that more than 40% of the surrounding genes have been deleted from one or other duplicate region, consistent with progress towards functional diploidy [19•]. A similar situation has been found in the homoeologous regions corresponding to the *LIGULELESS2* gene, in which 13 predicted homoeologous gene pairs have been reduced to 12 unique genes [20•].

The current picture has been further complicated by the analysis of inbred strains. In a gene-rich genomic region sequenced from two different inbred lines, there were substantial differences in transposon identity and copy number. Surprisingly, 4 out of 10 genes were also missing in one of the inbred lines. Southern hybridization allowed the classification of several inbred lines into groups depending on the gene content in this region [13••].

Mapping and sequencing strategies

The physical and genetic maps

Many thousands of simple sequence repeat (SSR), restriction fragment length polymorphism (RFLP) and single nucleotide polymorphism (SNP) markers have been mapped onto the maize genome (http://www.maizegdb.org), far exceeding the resolution of most genetic maps. For

example, assuming one recombination breakpoint per chromosome arm, a population of 100 F_2 mapping individuals would require only 2000 markers to resolve them. Several physical mapping strategies have been applied to take advantage of this high marker density. The progeny of wide crosses with oat retain maize chromosome fragments, allowing a mapping strategy similar to those involving radiation hybrids [21]. In addition, high-throughput fluorescent *in situ* hybridization (FISH) strategies are being used for physical mapping of the maize genome (http://www.fastlane.nsf.gov/servlet/showaward?award=0321639). Most important though, has been the construction of a framework physical map of fingerprinted BAC clones, which were anchored genetically using SSR, RFLP and other markers [22]. This map has the density and resolution to provide a substrate for sequencing: more than 900 contigs are longer than 1 Mb, and most have been placed on the genetic map. Each contig spans 1–2 cM and contains 20–40 genes. 3600 smaller contigs contain the remainder of the genome and these are being joined and edited.

Drafting the sequence

In many respects, the human draft genome sequence provides a useful model for the maize genome. The sizes of the human (2900 Mbp) and maize (2500 Mbp) genomes are similar, so efforts to produce draft sequences are on a similar scale. The minimal tile of BAC clones has yet to be constructed from the fingerprint contigs, but we can estimate that approximately 25 000 BACs, of average size 130 kbp, would be required. This is 15 times the number employed in sequencing the *Arabidopsis* genome [23].7 Each BAC would require a shotgun library of approximately 1000 clones for 5× coverage (assuming 750 bp reads and an 85% success rate), giving a total of 18 million sequencing reads. Assuming that a single sequencing machine can perform 500 000 reads per year, 36 machine-years would be required, at a cost of approximately US$36 million.

5× coverage only results in short contigs of a few kbp [24], so each BAC would require extensive finishing before accurate annotation was possible. This is important, as accurate annotation of unfinished cereal genome sequence is difficult [11,25]. As for the human genome map, filling gaps in the maize map would occupy a significant portion of the finishing phase. Assuming that the resources and expertise devoted to the human draft project could be committed to maize, the finishing phase could be completed within 2–3 years of the end of the draft phase.

There are some differences between the human and maize genomes that are relevant to finishing. On the one hand, maize genes are much smaller than human genes [26], making the annotation and finishing of maize genes much easier. On the other hand repeats are more prevalent and more similar to each other in maize, often confusing the finishing of entire BACs. Finally, the high GC content of the maize sequence often results in sequencing failure because of strong stops. Thus, the finishing costs for the maize genome would be comparable to those in the human genome project, amounting to at least another US$50 million and the full-time commitment of one or more genome centers over the next few years.

Dodging the draft: whole-genome shotgun sequencing

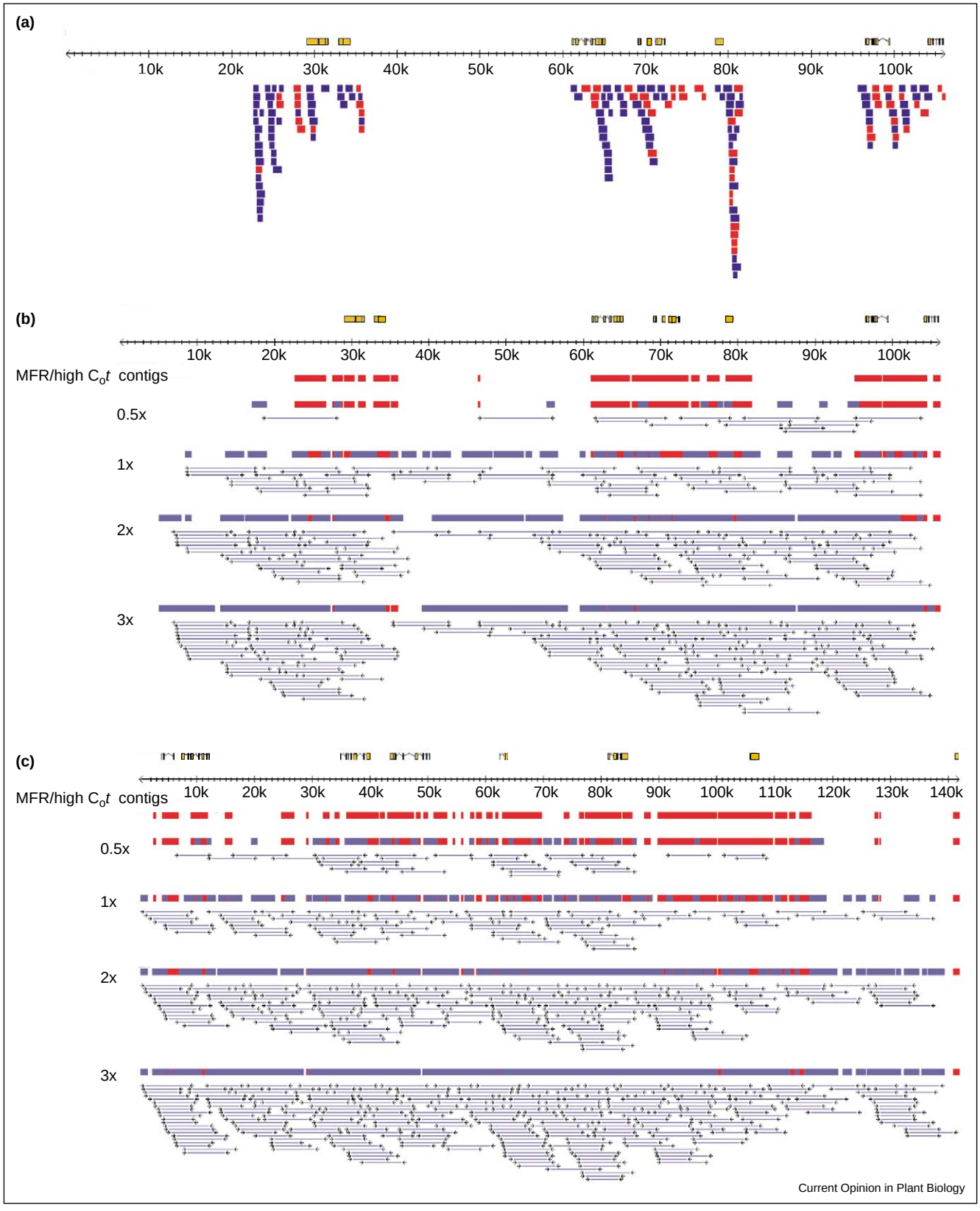
5× coverage of the entire maize genome could be achieved through a whole-genome shotgun sequencing strategy [27]. 2500 Mbp would require in the order of 25 million reads. To map the reads, two strategies have been suggested. Either the scaffold sequencing of end-reads from large clones could be used to link shotgun contigs or some form of draft sequence of a fingerprint map could be used. The relative merits of whole-genome shotgun sequencing and draft BAC-by-BAC sequencing were hotly debated in both the human and rice genome projects. Commercial interests preferred the rapid gene discovery provided by whole-genome shotgun sequencing, whereas geneticists preferred the map-based approach. More recently, private–public partnerships have employed both strategies in hybrid approaches to sequencing the rat and mouse genomes [28]. In the rat project, which is arguably the most efficient mammalian genome project to date, a relatively high level of whole-genome coverage was merged with very low coverage sequencing of mapped BACs (http://www.hgsc.bcm.tmc.edu/projects/rat).

Whole-genome shotgun sequencing would be preferable to BAC-by-BAC sequencing if it were possible to sequence only the portion of the genome that is of most interest, namely the genes and flanking regions, minimizing (although perhaps not completely excluding) heterochromatic repeats and transposons. Fortunately, the structure of the maize genome permits this type of strategy to be employed.

Gene-enrichment strategies

High C_0t sequencing [29] uses normalization to remove high-copy sequences, and enriches for genes in much the same way as cDNA normalization [30]. It also suffers from the same problems, namely the reduced representation of multigene families. Methylation filtration (MF) is an alternative method, in which undermethylated sequences are selectively cloned as 1–2 kbp fragments and sequenced from both ends [8,31]. As 95% of all maize exons, and 100% of maize genes, are at least partly undermethylated [32], MF greatly enriches for genes. However, transposons and repeats that have lost methylation, either through CpG suppression or through transcriptional activation, are also included in these libraries. Palmer *et al.* [33••] have compared gene-enriched shotgun

Figure 1


(a)
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
(b)
MFR/high C_0t contigs
0.5x
1x
2x
3x
(c)
110k
120k
130k
140k
Current Opinion in Plant Biology

sequencing with cDNA or expressed sequence tag (EST) sequencing using the rice genome as a guide. They demonstrated that MF achieves a higher level of gene coverage than that provided by ESTs, even discounting the extensive promoter and intron coverage in MF sequences that is not found in ESTs. This is because ESTs over-sample abundant transcripts and fail to sample rare or conditional transcripts.

A careful analysis of the composition of MF sequences generated by Cold Spring Harbor Laboratory and by the Maize Genomics Consortium (http://maize.danforthcenter.org) together with a set of 'unfiltered' (shotgun) libraries reveals that the unmethylated portion of the genome comprises only 16–17% [33••]. This portion requires only one sixth as many reads as the genome as a whole, or just 3 million reads for 5× coverage. More than half a million MF reads, and 400 000 high C_0t sequences have already been deposited in GenBank, and so this level of coverage could be readily achieved in 1 year with just four sequencing machines. With enough sequencing power, two or more different inbred strains could be sequenced to this level of coverage within this timeframe.

A recent publication by the Maize Genomics Consortium compares MF and high C_0t technologies [34••]. Using a mathematical model, the authors estimate that gene-enrichment strategies achieve a six-fold reduction in the effective genome size of maize.

The high degree of coverage that is achievable using gene-enrichment strategies can be readily demonstrated by aligning MF and high C_0t sequence reads to a finished maize BAC clone, followed by assembly into contigs (Figure 1a). In this scenario, each of the gene-islands is covered by MF sequencing clones, with only three small gaps in genes. These results can be extrapolated to any randomly selected region of the maize genome, and anecdotal reports indicate extensive coverage of known maize genes [32]. We estimate from this analysis that the 1 million gene-enriched reads currently in GenBank represent approximately two-fold coverage of the gene-space. Given an average read length of 750 bp, this means that the maize gene space is 350–400 Mbp.

Ordering and orienting gene islands with the genetic map

With any shotgun-sequencing strategy, contigs need to be anchored to the physical map. This can be achieved with the use of a low-pass draft sequence that is based on fingerprinted BAC clones. We simulated this procedure computationally by randomly generating 'shotgun' read-pairs from 5–10 kbp 'clones' derived from each sequenced BAC (Figure 1b,c). With as little as 1× coverage (200 reads per BAC) most of the MF-read and high C_0t contigs matched multiple end-reads, anchoring that contig unequivocally to this BAC. As gene-enriched shotgun contig coverage is increased, an even lower density of 'low-pass' reads will be required for anchoring. By using relatively large subclones from each BAC (5–10 kbp), paired end-reads will provide scaffolds to enhance the assembly. BAC ends would also provide additional anchor points and additional sequence coverage, whereas end sequences from BACs constructed with methylation-sensitive enzymes or methylation spanning linker libraries (MSLL) will help to link gene islands [35].

Low-pass sequencing can be streamlined by reducing the number of subclone libraries that are required. This could be achieved by pooling BACs that correspond to each contig and by preparing DNA for subclone libraries. This strategy works because heterozygosity in inbreds is low, so that BAC overlaps have the same sequence. 1000 paired subclone reads, as well as 100 BAC ends from each 1 Mbp contig, would provide a 1× framework for the anchoring of gene islands. Scaffolds constructed from the subclones (Figure 1b,c), BAC-ends and mapped genetic markers would provide a rich structure for each BAC fingerprint contig. This low-pass BAC sequencing would also aid contig editing and tiling-path selection, the most time-consuming steps in fingerprint mapping.

1× coverage of each BAC contig (including BAC ends) would consume 2 million reads, making a total of 5 million reads necessary to provide a high-quality sequence of every gene island mapped to a genetically defined physical contig. Such an unfinished sequence would provide sufficient resolution for positional cloning and quantitative trait locus (QTL) mapping because the number of contigs would be comparable to the number of recombination events in mapping populations (see above). Finishing would then proceed by amplifying gaps from BAC clones using rice gene models as a guide.

Conclusions and strategic issues: which maize genome and when?

Maize has unique biological properties that need to be considered in formulating the optimal genome sequencing strategy. Maize was domesticated recently, and the haplotype pool includes a substantial portion of the

(Figure 1 Legend) **(a)** MF (blue) and high C_0t (red) sequences mapped onto the BAC that contains the *bronze1* (*bz1*) gene. With an average read length of 750 bp, these reads achieve 2× coverage of gene islands. Gene-enriched reads matching the BACs that contain **(b)** *bz1* and **(c)** the *teosinte branched2* gene (*tb2*) were assembled using PHRAP (red boxes). Simulated read-pairs (arrows) from 5–10 kbp subclones (thin gray lines) were then generated and assembled (gray boxes). At 0.5× subclone read-pair coverage, most gene-enriched shotgun contigs are tagged by multiple subclone read-pairs, which map them to the BAC and provide scaffolds for further assembly as well as templates for finishing. Only subclone read pairs that are anchored to a contig on both ends are shown. Annotated genes are shown in yellow. MFR, MF read.

diversity represented in wild ancestors [36]. As a result, different inbred strains have very different genome structures. Intergenic regions in maize are hypervariable and are essentially unrelated between inbred lines, but even genic regions vary substantially [13••]. The optimal sequencing strategy should therefore take into account this variability, which is in marked contrast to the colinearity and high level of conservation revealed by human/primate comparisons [37].

Gene-enriched shotgun reads, assembled and then linked by low-pass coverage of a framework physical map, will meet the requirements of maize geneticists with respect to trait and gene discovery. Results from the first year indicate that the maize genome project could be completed in 1–2 years if moderate resources were devoted to it. Further, it could be scaled up to include half-a-dozen inbred strains for the same cost as a draft sequence of just one strain. The inclusion of several inbred strains will be important for positional cloning, as mutants have been isolated in several different backgrounds that differ in gene order and content. The variation uncovered in the sequences of inbred strains may underlie heterosis [13••], an enduring mystery of plant breeding [38].

As sequencing costs fall, draft sequencing of inbred strains will become feasible. At that time, gene-enriched shotgun sequences could be incorporated into the draft, reducing the number of reads required (Figure 1). Thus, gene-enrichment offers a powerful shortcut to gene discovery in a genetic context, without distracting from a whole-genome project. Even after a comprehensive genome project is initiated, gene-enrichment will continue to provide important information from diverse strains at a fraction of the cost of conventional sequencing strategies.

Acknowledgements

Work in the authors' laboratories was supported by grant number DBI-0110143 from the National Science Foundation Plant Genome Research Program to WRM and RAM. We apologize to those whose work was not cited for lack of space.

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. McClintock B: **Chromosome constitutions of Mexican and Guatemalan races of maize**. In *Genes Cells and Organisms*. Edited by Moore JA. New York: Garland Publishing, Inc.; 1987:395-406.

2. Ananiev EV, Phillips RL, Rines HW: **A knob-associated tandem repeat in maize capable of forming fold-back DNA segments: are chromosome knobs megatransposons?** *Proc Natl Acad Sci USA* 1998, **95**:10785-10790.

3. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z *et al.*: **Nested retrotransposons in the intergenic regions of the maize genome**. *Science* 1996, **274**:765-768.

4. Mao L, Wood TC, Yu Y, Budiman MA, Tomkins J, Woo S, Sasinowski M, Presting G, Frisch D, Goff S *et al.*: **Rice transposable elements: a survey of 73,000 sequence-tagged-connectors**. *Genome Res* 2000, **10**:982-990.

5. Raizada MN, Nan GL, Walbot V: **Somatic and germinal mobility of the *RescueMu* transposon in transgenic maize**. *Plant Cell* 2001, **13**:1587-1608.

6. May BP, Liu H, Vollbrecht E, Senior L, Rabinowicz PD, Roh D, Pan X, Stein L, Freeling M, Alexander D *et al.*: **Maize targeted mutagenesis: a knockout resource in maize**. *Proc Natl Acad Sci USA* 2003, **100**:11541-11546.

7. Meyers BC, Tingey SV, Morgante M: **Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome**. *Genome Res* 2001, **11**:1660-1676.

8. 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.

9. Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z, Nagamura Y *et al.*: **The genome sequence and structure of rice chromosome 1**. *Nature* 2002, **420**:312-316.

10. Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y, Hu X *et al.*: **Sequence and analysis of rice chromosome 4**. *Nature* 2002, **420**:316-320.

11. The Rice Chromosome 10 Sequencing Consortium: **In-depth view of structure, activity, and evolution of rice chromosome 10**. *Science* 2003, **300**:1566-1569. [Published erratum appears in *Science* 2003, **301**:1327.]

12. Arumuganathan K, Earle ED: **Nuclear DNA content of some important plant species**. *Plant Mol Biol Rep* 1991, **9**:208-218.

13. •• Fu H, Dooner HK: **Intraspecific violation of genetic colinearity and its implications in maize**. *Proc Natl Acad Sci USA* 2002, **99**:9573-9578.
Substantial differences in gene and transposon content can be found between different varieties of maize, and may underlie hybrid vigor.

14. Song R, Llaca V, Linton E, Messing J: **Sequence, regulation, and evolution of the maize 22-kD alpha zein gene family**. *Genome Res* 2001, **11**:1817-1825.

15. Gaut BS, Doebley JF: **DNA sequence evidence for the segmental allotetraploid origin of maize**. *Proc Natl Acad Sci USA* 1997, **94**:6809-6814.

16. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL: **The paleontology of intergene retrotransposons of maize**. *Nat Genet* 1998, **20**:43-45.

17. Gaut BS, Le Thierry d'Ennequin M, Peek AS, Sawkins MC: **Maize as a model for the evolution of plant nuclear genomes**. *Proc Natl Acad Sci USA* 2000, **97**:7008-7015.

18. Gaut BS: **Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses**. *Genome Res* 2001, **11**:55-66.

19. • Ilic K, SanMiguel PJ, Bennetzen JL: **A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes**. *Proc Natl Acad Sci USA* 2003, **100**:12265-12270.
This work shows that the two maize subgenomes have different gene contents and are unstable in comparison to the sorghum and rice genomes.

20. • Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M: **Genomic duplication, fractionation and the origin of regulatory novelty**. *Genetics* 2004, in press.
The differences between the homoeologous regions that correspond to the *LIGULELESS2* gene in maize provide further evidence of the instability of the maize genome.

21. Kynast RG, Okagaki RJ, Rines HW, Phillips RL: **Maize individualized chromosome and derived radiation hybrid lines and their use in functional genomics**. *Funct Integr Genomics* 2002, **2**:60-69.

22. Cone KC, McMullen MD, Bi IV, Davis GL, Yim YS, Gardiner JM, Polacco ML, Sanchez-Villeda H, Fang Z, Schroeder SG *et al.*: **Genetic, physical, and informatics resources for maize. On the road to an integrated map**. *Plant Physiol* 2002, **130**:1598-1605.

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

24. Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*)**. *Science* 2002, **296**:79-92.

25. Palmer LE, McCombie WR: **On the importance of being finished**. *Genome Biol* 2002, **3**:COMMENT2010.

26. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al.*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.

27. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA *et al.*: **A whole-genome assembly of *Drosophila***. *Science* 2000, **287**:2196-2204.

28. Mouse Genome Sequencing Consortium: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**:520-562.

29. Yuan Y, SanMiguel PJ, Bennetzen JL: **High-C_0t sequence analysis of the maize genome**. *Plant J* 2003, **34**:249-255.

30. Bonaldo MF, Lennon G, Soares MB: **Normalization and subtraction: two approaches to facilitate gene discovery**. *Genome Res* 1996, **6**:791-806.

31. Rabinowicz PD: **Constructing gene-enriched plant genomic libraries using methylation filtration technology**. *Methods Mol Biol* 2003, **236**:21-36.

32. 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.

33. Palmer LE, Rabinowicz PD, O'Shaughnessy A, Balija V,
•• Nascimento L, Dike S, de la Bastide M, Martienssen RA, McCombie WR *et al.*: **Maize genome sequencing by methylation filtration**. *Science* 2003, **302**:2115-2117.
An extensive analysis of the largest maize genomic-sequencing project to date shows that MF achieves comprehensive coverage of the maize gene space.

34. 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 authors of this paper report the sequence and analysis of 95 000 methylation filtration clones and 100 000 high C_0t clones. Assembly and comparative analysis suggest reciprocal coverage with the two methods. The authors recommend a combined approach for sequencing complex plant genomes.

35. Yuan Y, SanMiguel PJ, Bennetzen JL: **Methylation-spanning linker libraries link gene-rich regions and identify epigenetic boundaries in *Zea mays***. *Genome Res* 2002, **12**:1345-1349.

36. Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E, Doebley J: **A single domestication for maize shown by multilocus microsatellite genotyping**. *Proc Natl Acad Sci USA* 2002, **99**:6080-6084.

37. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC *et al.*: **Comparative analyses of multi-species sequences from targeted genomic regions**. *Nature* 2003, **424**:788-793.

38. Martienssen RA, Colot V: **DNA methylation and epigenetic inheritance in plants and filamentous fungi**. *Science* 2001, **293**:1070-1074.