Comparison of genes among cereals

Doreen Ware* and Lincoln Stein

Comparison of partially sequenced cereal genomes suggests a mosaic structure consisting of recombinationally active gene-rich islands that are separated by blocks of high-copy DNA. Annotation of the whole rice genome suggests that most, but not all, cereal genes are present within the rice genome and that the high number of reported genes in this genome is probably due to duplications. Within the cereals, macrocolinearity is conserved but, at the level of individual genes, microcolinearity is frequently disrupted. Preliminary evidence from limited comparative analysis of sequenced orthologous genomic segments suggests that local gene amplification and translocation within a plant genome may be linked in some cases.

Addresses
Cold Spring Harbor Laboratory, Bungtown Road, Cold Spring Harbor, New York 11724, USA
*e-mail: ware@cshl.org

Current Opinion in Plant Biology 2003, **6**:121–127

This review comes from a themed issue on
Genome studies and molecular genetics
Edited by Takuji Sasaki and Ronald R Sederoff

1369-5266/03/$ – see front matter

DOI 10.1016/S1369-5266(03)00012-8

Abbreviations
EST expressed sequence tag
IRGSP International Rice Genome Sequencing Project

Introduction

Cereals are a large and diverse set of agronomically important crops from the family *Gramineae*. They are the main source of dietary calories for most human populations whether they are consumed directly, as is rice, or indirectly by way of animal feed. The structures of the cereal genomes and the genes contained within them are key to understanding the evolutionary relationships within this important family. This in turn will aid geneticists and molecular biologists in their quest to understand cereal biology and help plant breeders in their goal of developing better and hardier strains.

This review describes the current state of knowledge regarding cereal genome structure. This knowledge is based largely on the results of the recent sequencing of long genomic segments from cereals [1–4,5••–7••,8•,9••,10••,11•] and the whole-genome analyses of *Arabidopsis* [12] and rice [13••–16••]. These projects have provided insight into the high number of genes in cereal genomes relative to those of other eukaryotes, and into the conserved macro- and microcolinearity at the sequence level in grass genomes.

Cereal genome colinearity

During the past 10 years, comparative mapping of cereal genomes using cross-hybridizing genetic markers has provided compelling evidence for a high level of conservation of gene order across regions spanning many megabases (i.e. macro-colinearity). This phenomenon was first summarized for rice, oats, maize, sorghum, sugar cane, foxtail millet, wheat, and finger millet using what is now known as the 'Circle Diagram' [17], and was later extended to include ten grass species using fewer than 30 rice linkage groups [18]. These findings were noteworthy in light of known differences in the size of the grass genomes, rice (430 Mb), sorghum (770 Mb), maize (2700 Mb), and wheat (16 000 Mb) [19] and the evolutionary divergence time of 60 million years for these species [17,20].

The initial work on the colinearity of genetic markers was reinforced when it was discovered that quantitative loci for agronomic traits such as dwarfing were also colinear between grass species [21]. More recently, the sequencing of long regions of the cereal genomes has allowed microcolinearity, or colinearity across gene clusters, to be investigated. Several recent studies in the cereals have demonstrated incomplete microcolinearity at the sequence level [4,6••,7••,10••,22]. The largest and most complete study of microcolinearity in the cereals to date is that by Song *et al.* [7••]. This study identified orthologous regions from maize, sorghum, and two subspecies of rice. Song *et al.* found that gross macrocolinearity is maintained but that microcolinearity is incomplete among these cereals. Deviations from gene colinearity are attributable to micro-rearrangement or small-scale genomic changes, such as gene insertions, deletions, duplications, or inversions [23]. In the region under study, the orthologous region was found to contain six genes in rice, 15 genes in sorghum (of which three have been amplified, producing a total of 29 genes), and 13 genes in maize (of which one has been amplified, resulting in a total of 34 genes) [7••]. In maize and sorghum, gene amplification caused a local expansion of conserved genes but did not disrupt their order or orientation.

As expected, there is a high degree of gene conservation between the two shotgun-sequenced subspecies of rice, *japonica* and *indica*, which diverged more than 1 million

years ago [3]. On careful inspection, however, narrow regions of divergence can be found in these genomes [7••]. These regions correspond to areas of increased divergence among rice, sorghum and maize, suggesting that the alignment of the two rice subspecies might be useful for identifying regions of cereal genomes that are prone to rapid evolution. The differences between sorghum and maize genes originated after the ancestral genomes of the two species diverged from each other 16.5 million years ago [24].

Where microcolinearity is broken and a gene that is present in one cereal is 'missing' from its orthologous position in another, it is often possible to find a matching gene homologue in a non-orthologous location [7••,25••]. The putative mechanism for this phenomenon is an ancient gene duplication in the common ancestor followed by the loss of one gene copy in the first modern species and the loss of the other copy in the second species. A second example of gene duplication and movement is provided by a study of tandem repeats of glutathione *S*-transferase genes in wheat. The best-conserved copy of the wheat gene was not found in the previously identified orthologous position in rice. Instead, it was found in a non-orthologous position on rice chromosome 10 where the gene was also amplified [25••]. This genomic segment on rice chromosome 10 also contained a large insertion of the chloroplast genome [26•], suggesting that this portion of the rice genome is fluid. From the study of orthologous regions in cereals [7••], Song *et al.* concluded that gene amplification and gene translocation are functionally associated with each other, and that the differential divergence of non-conserved 'hot spots' along chromosomes is manifested during speciation. A mosaic of conserved segments that are interspersed with non-conserved segments becomes apparent when orthologous regions of different species are compared [7••].

Work on gene conservation is of practical importance. The identification of orthologous regions (i.e. regions believed to be derived from the same segment of an ancestral genome) has been used to select candidate genes that are associated with certain functions or to select molecular markers to increase the map density at specific genetic loci, thereby facilitating map-based cloning.

Transcriptome estimates in cereals

Despite the exceptions described above, overall macrocolinearity is well established. Until recently, it was not known whether this phenomenon also implies a high degree of microcolinearity (i.e. colinearity at the level of genes and gene clusters). To answer this question, it is necessary to extend the analysis to the nucleotide sequence level, and for this we turn to analysis of the gene content (or transcriptome) of cereals.

A central aim of genome analysis is to identify and classify all of the genes of a particular species and to understand the redundant and unique functions of each gene. In the absence of a whole-genome sequence, gene discovery must rely on other tools, the most important of which are expressed sequence tag (EST) libraries. ESTs derived from a diverse set of cDNA libraries can provide information on the transcript abundance, tissue location, and developmental expression of genes. They are limited, however, by the initial biological sample (tissue type, developmental stage and environmental conditions of growth) and the sampling of the cDNA library. In addition to expression information, cDNA sequences provide a much-needed resource for annotating the *in silico* gene predictions generated from whole-genome analysis. As of October 2002, the total number of cereal ESTs available from the National Center for Biotechnology Information (NCBI) was more than 1 million: 111 315 for rice, 338 004 for barley, 266 203 for wheat, 190 301 for maize and 107 609 for sorghum [27].

ESTs are random sequences. A typical EST contains only a portion of the coding region of the original gene transcript and typically overlaps with other ESTs derived from the same gene. It is common practice to cluster ESTs to provide the longest contiguous sequence for each transcript and to provide a core set of unique genes for each species. Examples of these clusters are The Institute for Genomic Research (TIGR) gene indexes [28] and the NCBI UniGene sets. The utility of these clustered EST sequences has been further extended to include tentative groups of orthologues from different species and to establish putative orthologues and paralogues for the partial gene sequences [29]. In a whole-genome analysis of rice *japonica*, cereal clusters were generated and compared to the rice gene predictions. Nearly all of the cereal clusters were found to have a significant similarity to one or more rice gene predictions [13••]. Goff *et al.* [13••] suggest that most cereal genes are conserved across species and that phenotypic variation is due to a small number of genes or functional differences within similar genes.

Estimates of the number of cereal genes that are based on partial genomic sequencing and EST clusters range from 44 700 for rice [30] to 55 000 for maize [31]. Two independent analyses that were based on the shotgun genomic sequencing of the rice genome estimate the number of predicted rice genes to be 32 000–50 000 [13••] and 46 022–55 615 [14••]. More recent estimates, obtained by extrapolating the annotation of the finished rice chromosomes 1 and 4, predict a slightly higher gene number of 57 000–62 500 [15••,16••]. Together, these estimates suggest that rice has a transcriptome that contains nearly twice the number of genes in *Arabidopsis* [12] and humans [32], and more than three times the number in *Caenorhabditis elegans* [33] and *Drosophila* [34].

Gene amplification in cereals

Why might the number of genes be higher in cereals than other eukaryotes? Whole-genome analyses from *Arabidopsis* and rice suggest that cereals have a high number of genes because of a combination of factors that includes gene amplification and inaccurate gene predictions.

Analysis of the *Arabidopsis* genome revealed that a large portion of the genome was subjected to a polyploidization 50–200 million years ago [12,35–37]. In *Arabidopsis*, 65% of genes belong to a gene family and 35% belong to gene families that have more than five members [12]. Local duplication is also common in *Arabidopsis*, involving 17% of genes. The analysis of the rice *japonica* shotgun sequence reveals only slightly higher rates of duplication than those found in *Arabidopsis*, with globally duplicated genes estimated at 77% [13••] and locally duplicated genes at 15–30% depending on the chromosome [13••]. These findings are supported by reports from the International Rice Genome Sequencing Project (IRGSP) on rice chromosomes 1 and 4 [15••,16••].

Other research supports a model in which genome duplications beget new genes to create an extensive gene pool. In some cases, the duplicates will degenerate into pseudogenes; in others, the parent and daughter copies evolve distinct functions [35]. Polyploidization and segmental chromosomal duplication are believed to be common events in higher plant evolution [37]. The maize genome is believed to have undergone the most recent duplication 11.4 million years ago [24]. Local clusters of genes have been identified in the cereals and most likely reflect local gene duplication. Clusters of disease resistance genes are among the most common types of gene cluster reported in the literature [8•,38–40,41•,42–44]. The most highly represented gene family on rice chromosome 4 is a receptor serine/threonine kinase with 132 members. These genes are found in clusters and also as individuals, and represent nearly 2% of the 6756 genes found on chromosome 4 [15••]. The expansion of this gene family demonstrates how local gene expansion can increase the number of genes within a species.

Whole-genome duplication, through polyploidization, segmental duplication, and local gene amplification, increases the number of paralogous gene sequences found in plants. Together, these forms of duplication explain why the numbers of genes in grass genomes exceed those in the other eukaryotic genomes sequenced so far, including the human genome [45]. The use of all three duplication mechanisms in the evolution of grass genomes suggests that grasses, like other plants, may have evolved more rapidly than could be predicted by comparing only coding-sequence substitution rates [46]. This rapid evolution may be one of the reasons for the degree of speciation within certain large families of *Gramineae* [45]. It will be interesting to determine in evolutionary terms how the cereal genomes have maintained a level of macro-colinearity in light of the dynamic plant genome.

Gene-rich islands in cereal genomes

Gene distribution has been studied in animal systems, and it appears that all animal genomes and chromosomes are, to some degree, divided into gene-rich and gene-poor compartments [47,48]. In plants, this pattern is even more marked. Cytological and genetic studies have demonstrated that wheat genes are clustered into islands rather than being distributed evenly through the genome [49–52]. Cereal genomic sequences have confirmed these findings and support the existence of a mosaic structure in which low-copy, gene-rich regions, known as 'gene islands', are interspersed among high-copy retrotransposon-rich sequences [10••,53]. Gene islands have been found in barley, sorghum, maize, rice, and wheat. Within an island, genes occur with a density of 24–217 genes per Mbp [4,5••,9••,10••,15••,16••,25••,39,40,41•,42,54,55], which is similar to the density of genes in the relatively small *Arabidopsis* genome [12]. By contrast, genes are scant in retrotransposon-rich regions (10 genes per Mbp [56]). In fact, retrotransposon-rich regions frequently undergo heavy DNA and histone methylation to form cytologically distinct heterochromatin [57,58].

In marked contrast to the great disparity in the sizes of cereal and *Arabidopsis* genomes, the size, number, and distribution of introns and exons are similar within the sequenced genes of different cereals and of *Arabidopsis* [12,13••,15••,16••,30]. This finding supports the notion that the genome expansion of cereals is based upon the recent invasion and expansion of retrotransposon elements [53,59].

Cereal genes and 'hot spots' for recombination

It has been suggested that genic regions in the cereals are associated with hot spots of recombination. Supporting evidence for this hypothesis can be found in barley, wheat, rice, and maize. In wheat and barley, comparison of the physical maps with the genetic-linkage maps has shown that recombination is confined to gene-rich regions [8•]. Analysis of rice chromosome 4 showed that repeats were most often located in the heterochromatic regions, and that these regions have a low recombination frequency with a ratio of physical distance to gene distance of 1.5 cM per Mb. In contrast, the euchromatic (i.e. gene-rich) region has a physical distance to gene distance of 4.79 cM per Mb [16••]. A recent study of the maize a1–sh2 interval, physically positioned 101 meiotic breakpoints to three locations within the maize genome. Two of these locations were found to be in gene-rich regions and one was found in a region that contained no gene content. Interestingly, the non-genic region of the maize genome is a low-copy sequence [9••].

Together, these studies support the existence of a mosaic structure within cereal genomes, which contain low- and high-copy sequences. The low-copy sequences tend to be gene rich and are recombinationally active, whereas the high-copy DNA is gene-poor and recombinationally inactive.

Classification of cereal gene predictions

Whole-genome analysis of rice [13••,14••] has provided the most robust source of cereal gene predictions to date. These gene predictions are based on prediction algorithms. Predicting genes is still an imperfect science [60], and no single gene-prediction algorithm is completely accurate. The annotation of the shotgun sequence of rice *japonica* that is discussed in this section used several gene-prediction algorithms [13••], and scored the different gene models on their homology and length of match. The resulting predictions were classified as high (H), medium (M) and low (L) confidence on the basis of their homology with known genes from rice or other species, and their homology with Prosite motifs [61] and Pfam domains [61,62]. Many of the gene models are incomplete because of the nature of shotgun sequence. The predictions were also classified on the basis of minimum lengths of 300 or 500 bases. The largest set of predictions is 61 668 and includes all HML genes longer than 300 bp. This gene set is referred to as HML_{300} [13••]. Goff *et al.* [13••] compared their gene predictions to the 25 554 predicted genes from *Arabidopsis* [12], and found that 85% of the *Arabidopsis* genes are significantly homologous to the HML_{300} predicted genes. Four thousand *Arabidopsis* genes did not have any homology to the HML_{300} genes, however, suggesting that these genes could be dicot specific or simply inaccurate predictions.

One-third of the *Arabidopsis* genes that are also found in rice have no detectable homologues in *Drosophila*, *C. elegans*, *Saccharomyces*, or the sequenced bacterial genomes. Homologues of more than 13 000 HML_{300} genes are not found in other non-plants but are found in *Arabidopsis*. Many of these genes are likely to encode plant-specific proteins [13••]. No significant homology for 3886 H_{300} and 31 387 HML_{300} genes has been found in the *Arabidopsis* genome. However, most of these were low-evidence predictions (ML_{300}). The H_{300} genes represent 6% of the total predicted genes from rice and are likely to represent cereal- or rice-specific genes.

In the same study [13••], Goff *et al.* assigned functional classification to the HML_{300} gene predictions using Interpro [63,64] and GeneOntology (GO) [65] terms. The largest functional category was metabolism, which involved 25% of the predicted genes. Goff *et al.* [13••] also looked at differences in the type and number of gene families in the predicted sequences of rice and other eukaryotes by examining homology between predicted rice genes and known genes from other species. In *Arabidopsis*, many gene families are either absent or, conversely, over-represented in comparison to those families in the yeast, *C. elegans* or *Drosophila* genomes [12], the same is true for the rice HML_{300} genes. Two classes of genes that fall into the 'over-represented in plants' class include the genes that encode the RING zinc-finger proteins (840 HML_{300} predicted genes) and the F-box domain proteins (150 HML_{300} predicted genes). These proteins function in intracellular protein-degradation pathways and their regulation. The over-representation of these gene products is consistent with speculation that protein turnover and the regulation of protein degradation play an important role in the maintenance of homeostasis in plants.

Whole-genome analysis of rice suggests that the *Arabidopsis* genes appear to be a subset of the genes found in rice. *Arabidopsis* genes, along with genes from other eukaryotes, are useful in assigning functions to cereal genes. The analysis of available coding sequences from other cereals suggests that most cereal genes are present within the rice genome. The catalogue of predicted genes and their position within the rice genome will greatly enhance our ability to identify genes that are involved in agronomic traits because of macro-colinearity between rice and other cereals.

Conclusions

A finished rice genome will provide a complete index of potential rice genes but will not tell us which of these genes are important in providing desired traits in cereal crops. Genome sequences are just a beginning: they provide a necessary resource for powerful methods for proteome analysis that require sequence knowledge. In the future, techniques such as gene, protein and metabolic profiling will provide insights into the function and expression patterns of genes and into how these genes ultimately contribute to a crop's ability to react to an environment and reproduce.

Acknowledgements

We acknowledge Susan McCouch, Leonid Teytelman and Bonnie Harvey for critical reading of the manuscript, and the US Department of Agriculture (USDA) Agricultural Research Service (ARS) and USDA Cooperative State Research Education and Extension Service (CSREES) for financial support.

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. Chen M, SanMiguel P, de Oliveira AC, Woo SS, Zhang H, Wing RA, Bennetzen JL: **Microcolinearity in *sh2*-homologous regions of the maize, rice, and sorghum genomes**. *Proc Natl Acad Sci USA* 1997, **94**:3431-3435.

2. Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z: **Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum**. *Proc Natl Acad Sci USA* 1999, **96**:7409-7414.

3. Bennetzen JL: **Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions**. *Plant Cell* 2000, **12**:1021-1029.

4. Tarchini R, Biddle P, Wineland R, Tingey S, Rafalski A: **The complete sequence of 340 kb of DNA around the rice *Adh1–adh2* region reveals interrupted colinearity with maize chromosome 4**. *Plant Cell* 2000, **12**:381-391.

5. •• Fu H, Park W, Yan X, Zheng Z, Shen B, Dooner HK: **The highly recombinogenic *bz* locus lies in an unusually gene-rich region of the maize genome**. *Proc Natl Acad Sci USA* 2001, **98**:8903-8908.
The authors investigated the high rate of recombination that had been identified previously at the *bz* locus of maize. This is one of the first gene-dense regions to be identified in maize. It contains a 32-kb region of the genome that includes ten genes, eight of which are transcribed, with an average distance of 1 kb of intergenic space between the genes. The gene-rich island is bordered on both ends by retrotransposons blocks.

6. •• Dubcovsky J, Ramakrishna W, SanMiguel PJ, Busso CS, Yan L, Shiloff BA, Bennetzen JL: **Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes**. *Plant Physiol* 2001, **125**:1342-1353.
This is one of the first detailed examinations of the level of microcollinearity between barley and rice. The authors selected bacterial artificial chromosomes for genomic sequencing from a previously identified collinear region from barley chromosome 5 and rice chromosome 3. They found four conserved regions containing four predicted genes with similar exon number, size, and location. However, the orientation and number of the genes differed between rice and barley in some cases. The authors did not observe extensive similarity beyond the exon structures, untranslated regions, and promoter sequences. The differences in distances between genes in barley and rice were explained by the insertion of different transposable retroelements into each species.

7. •• Song R, Llaca V, Messing J: **Mosaic organization of orthologous sequences in grass genomes**. *Genome Res* 2002, **12**:1549-1555.
This is the most comprehensive comparison of sequence among the cereals to date. The authors expand on their previous work in maize [11•] to look at collinear regions in sorghum and two cultivars of rice. Their results support a mosaic organization of the orthologous regions in which conserved sequences are interspersed with non-conserved sequences. Gene amplification, gene movement, and retrotransposition account for the majority of the non-conserved sequences. The analysis also suggests that gene amplification is frequently linked with gene movement.

8. • Sandhu D, Gill KS: **Gene-containing regions of wheat and the other grass genomes**. *Plant Physiol* 2002, **128**:803-811.
Deletion lines were used to localize gene-containing regions of the wheat genome. The work suggests that gene-containing regions are found in about 10% of the wheat genome. The locations of the gene-containing regions on genetic and physical maps show that recombination is confined to these regions. The density and number of genes in each region vary along with the recombination frequency. Comparison with orthologous regions in rice suggests that the gene density in wheat is about half that in rice, and that the difference between the rice and wheat genomes is due to the amplification of gene-poor regions in wheat.

9. •• Yao H, Zhou Q, Li J, Smith H, Yandeau M, Nikolau BJ, Schnable PS: **Molecular characterization of meiotic recombination across the 140-kb multigenic *a1–sh2* interval of maize**. *Proc Natl Acad Sci USA* 2002, **99**:6157-6162.
The authors looked at recombinant-rich regions of the maize genome and mapped the recombination hot spots. Three of the hotspots are genic, and one is not. The non-genic spot was present at low copy number in the maize genome. This finding indicates that recombination-active portions of the genome are low copy, and that the retrotransposon portion of the maize genome is relatively inactive.

10. •• SanMiguel PJ, Ramakrishna W, Bennetzen JL, Busso CS, Dubcovsky J: **Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m)**. *Funct Integr Genomics* 2002, **2**:70-80.
The work described in this paper follows up on work done in rice and barley [1] to include a sequenced bacterial artificial chromosome from wheat. In a region that is conserved between rice and barley [11•], one gene was found to be present in the same orientation in wheat and rice but inverted in barley. The results showed that the wheat genome is a complex mixture of different sequence elements but has a general pattern of content that is similar to those of rice and barley.

11. • 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.
The authors sequenced all 23 members of the 22-kD α-zein gene family of maize, forming the largest contiguous maize region sequenced to date. On the basis of cDNA database analysis, seven of the α-zein genes were expressed, and these expressed genes were interspersed with non-expressed genes. The authors suggest that the amplification of blocks of genes explains the rapid and compact expansion of the α-zein cluster during the evolution of maize.

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

13. •• Goff SA, Ricke D, Lan TH, Presting G, Wang R, 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.
The shotgun sequences of two rice cultivars were published concurrently. This paper reports on Syngenta's analysis of the Nipponbare sequence, which covers 93% of the genome. Gene predictions from the assembly are compared to cereal ESTs and the *Arabidopsis* predictions. The paper reports on synteny with other cereals and with *Arabidopsis*, and on the genes classes found in rice. The authors find that 98% of known maize, wheat, and barley proteins are also found in rice.

14. •• 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.
The shotgun sequences of two rice cultivars were published concurrently. The authors of this paper report on the shotgun sequence of the cultivar *Indica* by the Beijing Genomics Institute. The authors compare the gene predictions for rice with those for *Arabidopsis*, and summarize information on the structure of the genes and the functions assigned to them by homology.

15. •• 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.
The completed sequencing of chromosomes 1 and 4 by the IRGSP was reported concurrently. In this paper, the authors describe the structure of longest rice chromosome, chromosome 1. They describe biological insights provided by the completely sequenced chromosome, including the presence of gene families and the location of the gene families on the whole chromosome. The authors also discuss the significance of finishing the sequence.

16. •• 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.
The completed sequencing of chromosomes 1 and 4 by the IRGSP was reported concurrently. In this paper, the authors describe the finished sequence of chromosome 4. They compare this sequence to that of chromosome 4 of the rice subspecies *Indica*, and find that there is an overall conserved synteny that diverges with single nucleotide polymorphisms, insertions, and deletions.

17. Gale MD, Devos KM: **Comparative genetics in the grasses**. *Proc Natl Acad Sci USA* 1998, **95**:1971-1974.

18. Devos KM, Gale MD: **Genome relationships: the grass model in current research**. *Plant Cell* 2000, **12**:637-646.

19. Arumuganathan K, Earle ED: **Nuclear DNA content of some important plant species**. *Plant Mol Biol Reporter* 1991, **25**:208-218.

20. Keller B, Feuillet C: **Colinearity and gene density in grass genomes**. *Trends Plant Sci* 2000, **5**:246-251.

21. Peng J, Richards DE, Hartley NM, Murphy GP, Devos KM, Flintham JE, Beales J, Fish LJ, Worland AJ, Pelica F *et al.*: **'Green revolution' genes encode mutant gibberellin response modulators**. *Nature* 1999, **400**:256-261.

22. Tikhonov AP, Bennetzen JL, Avramova ZV: **Structural domains and matrix attachment regions along colinear chromosomal segments of maize and sorghum**. *Plant Cell* 2000, **12**:249-264.

23. Bancroft I: **Insights into the structural and functional evolution of plant genomes afforded by the nucleotide sequences of chromosomes 2 and 4 of *Arabidopsis thaliana***. *Yeast* 2000, **17**:1-5.

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

25. Xu F, Lagudah ES, Moose SP, Riechers DE: **Tandemly duplicated safener-induced glutathione *S*-transferase genes from *Triticum tauschii* contribute to genome- and organ-specific expression in hexaploid wheat**. *Plant Physiol* 2002, **130**:362-373.

•• The authors describe the isolation of two glutathione *S*-transferase genes from wheat chromosome 6. The two genes have different expression levels and patterns in the different genomes of hexaploid wheat. An analysis of the available rice genome found the regions that were most homologous to the wheat glutathione *S*-transferase genes were found on rice chromosome 10 rather than chromosome 2. Rice chromosome 2 had previously been identified as the syntenic region for wheat chromosome 6.

26. Yuan Q, Hill J, Hsiao J, Moffat K, Ouyang S, Cheng Z, Jiang J, Buell CR: **Genome sequencing of a 239-kb region of rice chromosome 10L reveals a high frequency of gene duplication and a large chloroplast DNA insertion**. *Mol Genet Genomics* 2002, **267**:713-720.

• The authors discuss the high density of duplicated genes found in a 239-kb region of rice chromosome 10L, and the insertion of a large piece of the chloroplast genome into this region. This segment of the rice genome is the same as that described by Xu *et al.* [25••] who reported that it contains the closest homologues of the glutathione *S*-transferase genes found on wheat chromosome 6.

27. **dbEST: database of expressed sequence tags**. http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

28. Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J: **The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species**. *Nucleic Acids Res* 2001, **29**:159-164.

29. Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Antonescu V, White J *et al.*: **Cross-referencing eukaryotic genomes: TIGR orthologous gene alignments (TOGA)**. *Genome Res* 2002, **12**:493-502.

30. Wu J, Maehara T, Shimokawa T, Yamamoto S, Harada C, Takazaki Y, Ono N, Mukai Y, Koike K, Yazaki J *et al.*: **A comprehensive rice transcript map containing 6591 expressed sequence tag sites**. *Plant Cell* 2002, **14**:525-535.

31. Brendel V, Kurtz S, Walbot V: **Comparative genomics of *Arabidopsis* and maize: prospects and limitations**. *Genome Biol* 2002, **3**:REVIEWS1005.

32. 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.

33. The *C. elegans* Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology**. *Science* 1998, **282**:2012-2018.

34. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al.*: **The genome sequence of *Drosophila melanogaster***. *Science* 2000, **287**:2185-2195.

35. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes**. *Science* 2000, **290**:1151-1155.

36. Ku HM, Vision T, Liu J, 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.

37. Vision TJ, Brown DG, Tanksley SD: **The origins of genomic duplications in *Arabidopsis***. *Science* 2000, **290**:2114-2117.

38. Hulbert SH, Webb CA, Smith SM, Sun Q: **Resistance gene complexes: evolution and utilization**. *Annu Rev Phytopathol* 2001, **39**:285-312.

39. Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P: **A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion**. *Genome Res* 2000, **10**:908-915.

40. Panstruga R, Buschges R, Piffanelli P, Schulze-Lefert P: **A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome**. *Nucleic Acids Res* 1998, **26**:1056-1062.

41. Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B: **Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution**. *Plant J* 2001, **26**:307-316.

• The authors sequenced the largest contiguous block of the wheat genome to date. They found gene clusters, isolated genes and several new retrotransposon elements within this genomic sequence. One of the new retrotransposon elements was found to be closely associated with gene sequences.

42. 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.

43. Wei F, Wing RA, Wise RP: **Genome dynamics and evolution of the Mla (powdery mildew) resistance locus in barley**. *Plant Cell* 2002, **14**:1903-1917.

44. Meyers BC, Chin DB, Shen KA, Sivaramakrishnan S, Lavelle DO, Zhang Z, Michelmore RW: **The major resistance gene cluster in lettuce is highly duplicated and spans several megabases**. *Plant Cell* 1998, **10**:1817-1832.

45. Messing J: **Do plants have more genes than humans?** *Trends Plant Sci* 2001, **6**:195-196.

46. Gaut BS, Morton BR, McCaig BC, Clegg MT: **Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL***. *Proc Natl Acad Sci USA* 1996, **93**:10274-10279.

47. Clay O, Bernardi G: **The isochores in human chromosomes 21 and 22**. *Biochem Biophys Res Commun* 2001, **285**:855-856.

48. Sumner AT, de la Torre J, Stuppia L: **The distribution of genes on chromosomes: a cytological approach**. *J Mol Evol* 1993, **37**:117-122.

49. Werner JE, Endo TR, Gill BS: **Toward a cytogenetically based physical map of the wheat genome**. *Proc Natl Acad Sci USA* 1992, **89**:11307-11311.

50. Gill KS, Gill BS, Endo TR, Taylor T: **Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat**. *Genetics* 1996, **144**:1883-1891.

51. Faris JD, Haen KM, Gill BS: **Saturation mapping of a gene-rich recombination hot spot region in wheat**. *Genetics* 2000, **154**:823-835.

52. Sandhu D, Champoux JA, Bondareva SN, Gill KS: **Identification and physical localization of useful genes and markers to a major gene-rich region on wheat group 1S chromosomes**. *Genetics* 2001, **157**:1735-1747.

53. 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.

54. Brooks SA, Huang L, Gill BS, Fellers JP: **Analysis of 106 kb of contiguous DNA sequence from the D genome of wheat reveals high gene density and a complex arrangement of genes related to disease resistance**. *Genome* 2002, **45**:963-972.

55. Rahman S, Abrahams S, Abbott D, Mukai Y, Samuel M, Morell M, Appels R: **A complex arrangement of genes at a starch branching enzyme I locus in the D-genome donor of wheat**. *Genome* 1997, **40**:465-474.

56. Rostoks N, Park YJ, Ramakrishna W, Ma J, Druka A, Shiloff BA, SanMiguel PJ, Jiang Z, Brueggeman R, Sandhu D *et al.*: **Genomic sequencing reveals gene content, genomic organization, and recombination relationships in barley**. *Funct Integr Genomics* 2002, **2**:51-59.

57. Bennetzen JL, Schrick K, Springer PS, Brown WE, SanMiguel P: **Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA**. *Genome* 1994, **37**:565-576.

58. Gruenbaum Y, Naveh-Many T, Cedar H, Razin A: **Sequence specificity of methylation in higher plant DNA**. *Nature* 1981, **292**:860-862.

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

60. Hogenesch JB, Ching KA, Batalov S, Su AI, Walker JR, Zhou Y, Kay SA, Schultz PG, Cooke MP: **A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes**. *Cell* 2001, **106**:413-415.

61. Hofmann K, Bucher P, Falquet L, Bairoch A: **The PROSITE database, its status in 1999**. *Nucleic Acids Res* 1999, **27**:215-219.

62. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL: **The Pfam protein families database**. *Nucleic Acids Res* 2000, **28**:263-266.

63. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD *et al.*: **InterPro — an integrated documentation resource for protein families, domains and functional sites**. *Bioinformatics* 2000, **16**:1145-1150.

64. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD *et al.*: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites**. *Nucleic Acids Res* 2001, **29**:37-40.

65. Lewis S, Ashburner M, Reese MG: **Annotating eukaryote genomes**. *Curr Opin Struct Biol* 2000, **10**:349-354.