Minireview

Insights into cereal genomes from two draft genome sequences of rice

Ian Bancroft

Address: John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK. E-mail: ian.bancroft@bbsrc.ac.uk

Published: 28 May 2002

Genome ***Biology*** 2002, **3(6)**:reviews1015.1–1015.3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/6/reviews/1015

Abstract

Draft genome sequences have been reported for two subspecies of rice. The drafts include the sequences of an estimated 99% of all rice genes and provide major advances in our understanding of the content and complexity of cereal genomes in general and the rice genome in particular.

A third of the human population depends on rice as a staple food [1]. Rice is also an important model for other cereal crops that, along with rice, account for more than 60% of worldwide agricultural production [2]. Its small genome and long history of genetic studies led to rice being an early target for complete genome sequencing. Draft sequence has recently been reported for two subspecies of rice: *japonica* [3], which is widely grown in Japan, and *indica* [4], which is widely grown in China and elsewhere. These sequence data can now be analyzed and compared with the published genome sequence of the widely adopted model species for dicotyledonous flowering plants, *Arabidopsis thaliana* [5].

Gene content of rice

The rice draft genomic sequences were generated by whole-genome shotgun (WGS) sequencing, largely starting from plasmid clones of nuclear DNA. This strategy contrasts with that employed by the International Rice Gene Sequencing Project (IRGSP) [6], which produces high-quality sequence from large-insert clones that are individually selected from a complete physical map.WGS sequencing rapidly and cost-effectively generates sequence from nearly all of the genes in the genome, but the lengths of contiguous sequence are shorter, errors are more numerous and integration with genetic maps is poorer. The *japonica* WGS sequencing [3] was conducted to 6X redundancy - that is, each base was sequenced an average of six times. Repeated sequences (which are numerous in the rice genome) were removed, permitting the assembly of 42,109 contiguous sequences, or contigs, representing 390 megabases of total sequence (thus, the mean length of contigs was 9.2 kb). The *indica* WGS sequencing [4] was conducted to 4X redundancy. In this case, repeated sequences were masked, permitting the assembly of 127,550 sequence contigs, representing 361 Mb total length (the mean length of contigs was thus 3.5 kb). The total size of the *indica* genome was estimated as 466 Mb [4] and that of *japonica* should be very similar. On the basis of comparisons with individually sequenced rice genes, finished sequence from the IRGSP, and other rice sequences, the *japonica* and *indica* sequences provide 99% and 92% coverage of genes, respectively [3,4]. The sequence probably contains relatively few mis-assemblies; this was tested for the *indica* data by using cDNA sequences to screen for artificial exon-sized rearrangements, and putative mis-assemblies were identified in only 1.1% of the genes tested [4]. Thus, although highly fragmented, the draft sequence generated by WGS allows us to gain early insights into many characteristics of the rice genome.

Automated gene prediction within the rice genome sequence is complicated by gradients in patterns of the usage of codons and amino acids. This phenomenon was analyzed in the *indica* sequence [4], where the 5′ ends of genes were found to have a higher GC content than the 3′ ends by up to 25%. This difference extended approximately a kilobase from the 5′ ends of genes. These gradients are not observed in *A. thaliana*, so assessing total gene numbers in rice is more difficult than in *A. thaliana,* and the numbers reported should be considered very much as approximations. Using

various measures of estimation of true gene numbers, the range for *japonica* was reported as being 32,000-50,000 [3] and for *indica* as 46,000-56,000 [4]. Both estimates are considerably more than the 25,500 or so genes predicted in the *A. thaliana* genome. They are also greater than for *Caenorhabditis elegans* and *Drosophila melanogaster* (which have around 19,100 and 13,500 genes, respectively) [7,8] and, probably, than for human (which has perhaps 38,000) [9]. The relatively larger number of genes encoded by plant genomes is likely to be a consequence of the frequency with which polyploidy occurs and the apparent selective advantage of polyploidy.

The functions of rice genes

The proteins encoded by the predicted genes in both the *indica* and *japonica* sequences were analyzed and classified using the InterPro [10,11] and Gene Ontology (GO) [12,13] databases. The results are not directly comparable with the previously published classification of *A. thaliana* proteins [5], which used InterPro and PENDANT [14,15]. But when 25,426 *A. thaliana* genes were classified using GO, along with 53,398 predicted complete *indica* genes, allowing direct comparison of their functional classifications, the proportions of genes in each functional category were found to be almost identical between the two plants [4]. The major difference, as shown in Figure 1, is that a larger proportion of rice genes remain unclassified. The classified predicted proteins (10,893 from *indica* and 9,230 for *A. thaliana*) are very likely to represent the products of genuine genes, but what about the genes predicted to encode unclassifiable proteins? For the predicted *A. thaliana* proteins, 80-85% (around 21,000 genes) have homologous predicted genes in the rice *indica* and *japonica* sequences [3,4]. This includes about 8,000 proteins predicted to be in rice but not in *D. melanogaster*, *C. elegans*, *Saccharomyces cerevisiae* or sequenced bacterial genomes, so these probably represent the plant-specific set of genes. The approximately 4,000 predicted proteins of *A. thaliana* that did not have homologies in rice are either artefacts or are unique to dicotyledonous plants. When the predicted rice gene sequences are compared with the *A. thaliana* genome, less than half (about 30,000) have significant homologs. This represents the minimum number of genes in rice, therefore, although it should be recognized that they are not necessarily all functional. The predicted rice proteins with no homologs in *A. thaliana* are either artefacts of the automated annotation (probably most of them) or are unique to monocotyledonous plants. Although the number of genes in rice is relatively large, the number of distinct gene families that appear to be present (15,000 reported for *japonica* [3]) is similar to the number for *A. thaliana, D. melanogaster* and *C. elegans* (13,382, 10,736 and 14,177, respectively [5]). The increased number of genes per family in plants is clearly a consequence of gene duplication, but it is not necessarily the case that the functions of members of a family are redundant.

Comparative analyses

Analysis of the *japonica* sequence has revealed extensive gene duplication in rice [3]. The sequences of more than 2,000 mapped rice cDNA markers were used to identify homologs in the genome sequence. Using a threshold value of 80% identity over 100 base-pairs, the mean number of homologous loci per cDNA was over 1.94. Evidence of extensive duplication of genomic segments was also detected. Most of the segments identified were very small (four markers or fewer), suggesting that extensive recombination and/or rearrangement has occurred since the duplication event(s). Dating of the duplications suggested that whole-genome duplication occurred 40 to 50 million years ago. The genome of *A. thaliana* also appears to be the result of whole

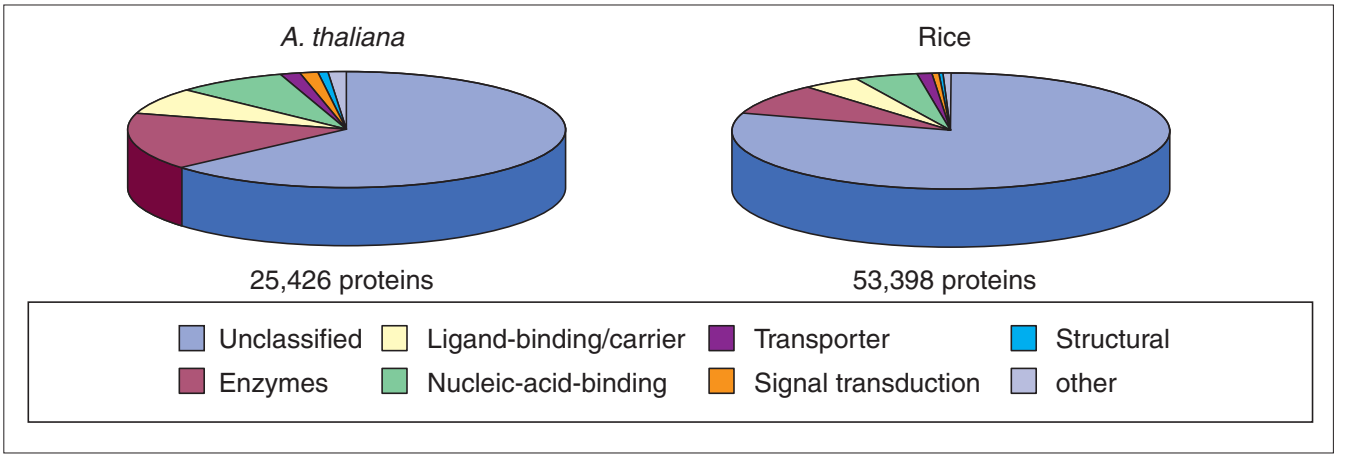

Figure 1
Comparison of the functional classifications of predicted rice and *A. thaliana* proteins (see text for further details; the figure is redrawn from [3,4]).

genome duplication - a tetraploidy event - in its ancestry, followed by extensive rearrangement [16]. This underlines the importance of polyploidy in plant evolution and as a key feature of plant genomes to be recognized when planning comparative analyses.

There is extensive conservation in the *japonica* genome sequence [3] of genes previously identified in other cereals. The alignment of cereal genomes using sequence-based markers from cereals other than rice showed extensive conservation of gene order (conserved synteny), confirming previous results [17]. But can the draft rice sequence be used to assess the extent of conservation of synteny between the rice and *A. thaliana* genomes? Although the fragmented nature of WGS-derived sequence data makes such data poorly suited to the comparative analysis of distantly related genomes, conservation of genome structure was identified between *japonica* and *A. thaliana* [3]. At least 137 blocks of conserved synteny (defined as three or more *A. thaliana* genes from the same chromosome mapping to one rice bacterial artificial chromosome, or BAC, contig) were identified. This finding supports the general applicability of previous data from a specific region of the rice genome for which conserved genome microstructure could be identified in *A. thaliana* [18]. Several of the syntenic blocks in *japonica* aligned to multiple regions of the *A. thaliana* genome, providing further confirmation for observations that the *A. thaliana* genome is the result of multiple rounds of duplication [18,19]. More extensive analyses were reported [3] for adjacent pairs of genes present in *A. thaliana,* for which the positions in the *japonica* sequence were assessed. This analysis showed that many pairs of genes show synteny, but most pairs had intervening genes in rice. This pattern of interspersion of conserved and non-conserved genes confirms previous similar findings in comparisons of the microstructure of the genomes of rice and *A. thaliana* [18,20] and provides support for the generality of this feature of plant genome microstructure [21].

Towards a high-quality rice genome sequence

The published draft rice genome sequences contain at least parts of almost all rice genes. The annotation of those genes is presently poor, and their number could be anywhere in the range 30,000 to 60,000. Future rice genome sequencing plans include the integration of the draft sequence data recently reported [3,4] with data being generated by the IRGSP [6], which already incorporates sequence data generated by Monsanto [22]. This will allow the most rapid progress towards the completion of fully contiguous, high quality genome sequence, which will be needed to underpin comparative genomics and gene discovery in cereals. Verification and functional analysis of the predicted genes will be important if they are to fulfil their potential for aiding the development of our understanding of plant science and the advancement of agricultural production worldwide.

References

1. Khush GS: **Origin, dispersal, cultivation and variation of rice.** *Plant Mol Biol* 1997, **35:**25-34.
2. Harlan JR: *The Living Fields: Our Agricultural Heritage*. New York: Cambridge University Press; 1995: 30-31.
3. 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.
4. 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.
5. The Arabidopsis Genome Initiative**: Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2002, **408:**796-815.
6. Sasaki T, Burr B: **International Rice Genome Sequencing Project: the effort to completely sequence the rice genome.** *Curr Opin Plant Biol* 2000, **3:**138-41.
7. The *C. elegans* Sequencing Consortium: **Sequence and analysis of the genome of *C. elegans:* a platform for investigating biology.** *Science* 1998, **282:**2012-2018.
8. 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.
9. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
10. 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.
11. **InterPro** [http://www.ebi.ac.uk/interpro/]
12. Lewis S, Ashburner M, Reese, MG: **Annotating eukaryote genomes.** *Curr Opin Struct Biol* 2000, **10:**349-354.
13. **Gene Ontology Consortium** [http://www.geneontology.org/]
14. Frishman D, Albermann K, Hani J, Heumann K, Metanomski A, Zollner A, Mewes HW: **Functional and structural genomics using PEDANT.** *BioInformatics* 2001, **17:**44-57.
15. **PEDANT** [http://pedant.gsf.de/]
16. Blanc G, Barakat A, Guyot R, Cooke R, Delseny M: **Extensive duplication and reshuffling in the *Arabidopsis* genome.** *Plant Cell* 2000, **12:**1093-1101.
17. Gale MD, Devos KM: **Comparative genetics in the grasses.** *Proc Natl Acad Sci USA* 1998, **95:**1971-1974.
18. Mayer K, Murphy G, Tarchini R, Wambutt R, Volckaert G, Pohl T, Dusterhoft A, Stiekema W, Entian KD, Terryn N, *et al.*: **Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of *Arabidopsis thaliana*.** *Genome Res* 2001, **11:**1167-1174.
19. Ku H-M, 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.
20. van Dodeweerd A-M, Hall CR, Bent EG, Johnson SJ, Bevan MW, Bancroft I: **Identification and analysis of homoeologous segments of the genomes of rice and *Arabidopsis thaliana*.** *Genome* 1999, **42:**887-892.
21. Bancroft I: **Duplicate and diverge: the evolution of plant genome microstructure.** *Trends Genet* 2001, **17:**89-93.
22. Barry GF: **The use of the Monsanto draft rice genome sequence in research.** *Plant Physiol* 2001, **125:**1164-1165.