Conservation of Microstructure between a Sequenced Region of the Genome of Rice and Multiple Segments of the Genome of Arabidopsis thaliana

Klaus Mayer,1 George Murphy,2 Renato Tarchini,3,11 Rolf Wambutt,4 Guido Volckaert,5 Thomas Pohl,6 Andreas Düsterhöft,7 Willem Stiekema,8 Karl-Dieter Entian,9 Nancy Terryn,10 Kai Lemcke,1 Dirk Haase,1 Caroline R. Hall,2 Anne-Marie van Dodeweerd,2 Scott V. Tingey,3 Hans-Werner Mewes,1 Michael W. Bevan,2 and Ian Bancroft2,12

1National Research Center for Environment and Health, Institute for Bioinformatics, Munich Information Centre for Protein Sequences, 85764 Neuherberg, Germany; 2John Innes Centre, Colney, Norwich, NR7 4UH, United Kingdom; 3DuPont Agricultural Biotechnology, Newark, Delaware 19711, USA; 4AGOWA GmbH, D-12489 Berlin, Germany; 5Katholieke Universiteit Leuven, Laboratory of Gene Technology, B-3001 Leuven, Belgium; 6GATC GmbH, D-78467 Konstanz, Germany; 7QIAGEN GmbH, Max-Volmer-Str.4, D-40724 Hilden, Germany; 8Plant Research International, NL 6708 PB, Wageningen, The Netherlands; 9Institut für Mikrobiologie, D-60439 Frankfurt/M., Germany; 10Department of Genetics, University of Ghent, B-9000 Ghent, Belgium

The nucleotide sequence was determined for a 340-kb segment of rice chromosome 2, revealing 56 putative protein-coding genes. This represents a density of one gene per 6.1 kb, which is higher than was reported for a previously sequenced segment of the rice genome. Sixteen of the putative genes were supported by matches to ESTs. The predicted products of 29 of the putative genes showed similarity to known proteins, and a further 17 genes showed similarity only to predicted or hypothetical proteins identified in genome sequence data. The region contains a few transposable elements: one retrotransposon, and one transposon. The segment of the rice genome studied had previously been identified as representing a part of rice chromosome 2 that may be homologous to a segment of Arabidopsis chromosome 4. We confirmed the conservation of gene content and order between the two genome segments. In addition, we identified a further four segments of the Arabidopsis genome that contain conserved gene content and order. In total, 22 of the 56 genes identified in the rice genome segment were represented in this set of Arabidopsis genome segments, with at least five genes present, in conserved order, in each segment. These data are consistent with the hypothesis that the Arabidopsis genome has undergone multiple duplication events. Our results demonstrate that conservation of the genome microstructure can be identified even between monocot and dicot species. However, the frequent occurrence of duplication, and subsequent microstructure divergence, within plant genomes may necessitate the integration of subsets of genes present in multiple redundant segments to deduce evolutionary relationships and identify orthologous genes.

Rice (Oryza sativa) is a widely grown crop, and is the staple food for over one-half of the world’s population. Extensive classical and molecular genetic maps have been constructed to assist biological analyses and plant breeding applications (Kinosita 1995; Kurata et al. 1994). The genome size of rice, ~440 Mb (Arumuganathan and Earle 1991), is one of the smallest of the cereals. It has been postulated that the genes within the genome of rice, as with the genes of other Gramineae, are clustered in gene-rich regions separated by gene-poor DNA (Barakat et al. 1997). A high degree of conservation of the order of gene-specific markers (conserved synteny) has been observed between the genomes of most cereals, including rice (Moore et al. 1995). The sequences of exons and exon–intron structures of orthologous genes in the sh2/a1-homologous regions of rice and sorghum have been shown to be conserved (Chen et al. 1998). However, more divergence of gene content has been found in the Adh1 regions of the genomes of maize and sorghum (Tikhonov et al. 1999). Nevertheless, rice is being developed as the key model monocot species for molecular genetic investigations, with the expectation that, by exploiting conserved synteny, the identification and functional assign-
moment of genes in rice will lead to the identification of the equivalent genes in other cereal species. These applications are being supported by the Rice Genome Project, which commenced in 1991. The main aim of this project is the determination of the complete nucleotide sequence of the rice genome.

A 340-kb region around the rice Adh1–Adh2 region has been sequenced (Tarchini et al. 2000), and is predicted to contain 33 protein-coding genes. Fourteen of these predicted genes were supported by the identification of corresponding transcripts, and 15 genes were similar in structure to genes with known functions. Nineteen of the 33 genes were members of gene families within the sequenced region, although some copies were predicted to be nonfunctional pseudogenes.

The key model dicot plant species is *Arabidopsis thaliana* (*Arabidopsis*). Extensive classical genetic, molecular genetic, and physical maps have been developed, along with numerous genome analysis and gene cloning strategies (Koornneef 1990; Lister and Dean 1993; Feldmann et al. 1989; Giraudat et al. 1992; Bancroft et al. 1993; Schmidt et al. 1995; http://nasc.nott.ac.uk/new_ri_map.html). The *Arabidopsis* genome has been completely sequenced (The *Arabidopsis* Genome Initiative 2000). It is very gene-rich, containing 25,498 genes, with an average density of one gene per 4.5 kb. Conservation of gene order has been observed between segments of the genome of *Arabidopsis* and those of its closest relatives among crops, the cultivated *Brassica* species (Kowalski et al. 1994; Cavell et al. 1998; Lagercrantz 1998).

It had been predicted that conserved synteny between the genomes of *Arabidopsis* and cereals, which diverged ca. 200 million years ago (Wolfe et al. 1989), would be detectable for segments of ∼3 cM (Paterson et al. 1996). Such conservation could lead to the use of positional approaches to integrate functional genomics information from both monocot and dicot species. The results of comparative genetic mapping efforts have provided little evidence for conserved gene organization (Gale and Devos 1998). Although some conserved synteny can be detected between the genomes of rice and *Arabidopsis* using physical mapping and sequence analysis approaches, the extent of conservation appears low (Devos et al. 1999; Han et al. 1999; van Dodeweerdt et al. 1999). In the present study we report the results of a pilot-scale rice genome sequencing project and the use of the data to further study aspects of genome organization in *Arabidopsis*.

**RESULTS**

**Gene Prediction**

Four overlapping BACs representing the 340-kb region to be sequenced had been identified previously (van Dodeweerdt et al. 1999). A shotgun sequencing strategy was used and annotation performed on a 339,972-bp contiguous assembly as submitted to EMBL (accession no. AJ307662). Four gene prediction programs were used for modeling exon structure: Genemark.hmm, FGENESH, Genscan, and GeneFinder. Comparisons of the outputs from these programs with gene structures determined using EST matches and protein homologies for three genes are shown in Figure 1. Although all programs correctly predicted the presence of a gene, none of the predictions accurately identified the exon–intron structures of the genes.

Gene prediction in rice is complicated by the fact that a rice species setting is available only for Genemark.hmm. Nevertheless, our data suggest that even Genemark.hmm output is not reliable enough to perform an in silico whole-genome analysis in rice. Further adjustment and refinement of gene prediction programs is necessary for large-scale automated genome analysis. Similarities of genomic sequences with EST sequences, matches of predicted protein products with known proteins, and matches with transposable elements were also used to derive the final gene modeling, as shown in Figure 2. In total, 56 potential protein-coding genes were identified, along with a region containing a retrotransposon and a region showing homology to transposon Tnr1. Two tRNAs were also identified. A summary of the positions of the identified genes and other features is presented in Table 1.

**Identification of Homologous ESTs and Proteins**

The rice genomic nucleotide sequence was used to query a rice EST database to identify ESTs corresponding to modeled genes. A threshold of at least 90% sequence identity over at least 150 bp was applied. Each predicted gene was used to query all available
nucleotide and protein databases for homologous identified or predicted protein sequences. The results of both analyses are summarized in Table 1. Sixteen of the 56 modeled genes match ESTs, supporting the prediction of the presence of a gene. The predicted proteins of 29 putative genes (52%) match known proteins, and the predicted proteins of 17 putative genes (30%) match proteins predicted from genome sequence data. The predicted proteins of the 10 remaining putative genes (18%) show no similarity to known proteins, so may either represent new types of proteins or be the result of false gene predictions.

The predicted proteins for all 56 putative genes were analyzed for the presence of functional domains using Interpro (The InterPro Consortium 2000). The results are shown in Table 2. Characterized functional domains of protein products were identified for 33 of the 56 putative genes (59%). These allowed us to identify a four-member gene family (C635, W700, W940, and C1190) encoding protein products with AP2 domain/ethylene responsive element binding protein functional domains, which was the largest gene family we identified in the sequenced region.

Analysis of Genome Organization in Rice and Arabidopsis

The extent of conservation of both the presence and the position of genes in the sequenced segment of the rice genome and the corresponding segments of the genome of Arabidopsis was analyzed. BLASTP analyses were performed using the extracted amino acid sequences of the annotated rice genes to query a database of all predicted Arabidopsis protein sequences, using a P-value of ≤e−5 as a cutoff. The results were then filtered to remove adjacent matches (indicative of tandem duplications) and clusters of three or more nearby matches recorded. The results for this analysis of the 340-kb region are summarized in Table 3. The relative coding strand for each gene model is denoted by W or C. Five segments of the Arabidopsis genome contained conserved subsets of the rice genes, as shown in Figure 3. These segments represented regions of the Arabidopsis genome containing approximately 22, 27, 20, 15, and 23 genes, for the chromosome 4(a), 5, 2, 4(b), and 3 segments, respectively, shown in Figure 3. Overall, 22 of the 56 rice genes are represented in the five Arabidopsis chromosome segments, counting both copies of three pairs of related genes (W495/W505, C635/W700, and W940/C1190) that show homology to common Arabidopsis genes. The most highly conserved segment, chromosome 4(a), contains eight conserved genes, with one pair (W950-AT4g17340 and W1050-AT4g17350) reversed. The relative coding strand orientation of the genes is also conserved, except for the reversal of the final pair, which is consistent with the inversion of the segment containing the genes. This region of the Arabidopsis genome had been shown previously to be related to the sequenced segment of the rice genome (van Dodewaard et al. 1999). The Arabidopsis chromosome 5 segment contains seven conserved genes. These are also in conserved order and orientation, except the same reversed pair of genes. This region of the Arabidopsis genome had been shown previously to be related to the chromosome 4(a) segment (Bancroft 2000). The remaining segments contain seven, five, and five conserved genes for the chromosome 2, 4(b), and 3 segments, respectively, all in conserved order. However, the orientation of several of the individual genes is reversed, indicating possible small-scale inversion events.
| Feature | Position | Protein similarity/EST matches |
|---------|----------|--------------------------------|
| C15     | C/2504-3864 | hypothetical protein F30H5.2, Caenorhabditis elegans, 1.6e-05 |
| W30     | W/8619-9131 | glycin-rich protein GRP22, rape, PIR:S31415, 1.6e-36 |
| W50     | W/12306-12944 | PREC like protein, Arabidopsis thaliana, EMBL:A012239, 5.2e-38 |
| C75     | C/16037-17967 | DNA-binding protein ABF1—wild oat, PIR:S61413, 5.0e-28 |
| C99     | C/20929-24200 | 27k vesicle-associated membrane protein-associated protein (VAMP), curled-leaved tobacco, EMBL:A018117, 4.8e-84 |
| C110    | C/3382-31466 | predicted protein, Arabidopsis thaliana, PIR:T49221, 8.7e-133/D25127 |
| C140    | C/47976-49075 | predicted protein, Orzya sativa, EMBL:AC026815, 1.9e-23 |
| C200    | C/30060-56958 | glycine-rich protein 2, Arabidopsis thaliana, PIR:T05494, 7.5e-09 |
| C202    | C/38901-61517 | predicted protein, Spinacia oleracea, EMBL:AF250384, 5.3e-59/C74047 |
| W230    | W/62697-64415 | extensin-like protein—maize, PIR:S49915, 1.2e-06/AU030763, AU070696, AU030762 |
| tRNA    | C/8486-65017 | tRNA predict as a tRNA-Arg : anticodon acg |
| tRNA    | C/8486-69457 | tRNA predict as a tRNA-Pro : anticodon ggg |
| W265    | W/65017-67265 | cAMP inductive 2 protein, Mus musculus, EMBL:AF121081, 1.5e-123 |
| C285    | C/70640-71035 | transposon |
| W325    | W/83990-84706 | glycine-rich protein 2, Arabidopsis thaliana, PIR:T05494, 7.5e-09 |
| C345    | C/85408-87116 | nucleolin 3, Arabidopsis thaliana, PIR:T49223, 3.3e-36/BE040770, BE040904, AU071152, AU071239, AU070738, AU071112, AU071195, AU070659, AU070774, AU071224, AU095618, AU163454, AU071095, AU070835, AU071193, AU070840, C25085 |
| W380    | W/90519-91174 | predicted protein, Arabidopsis thaliana, PIR:T48510, 3.7e-28 |
| C395    | C/91818-95108 | predicted protein, Arabidopsis thaliana, PIR:T48510, 3.7e-28 |
| W420    | W/96931-97865 | predicted protein, Arabidopsis thaliana, PIR:T48510, 3.7e-28 |
| W435    | W/100319-102879 | predicted protein, Orzya sativa, EMBL:A082695, 8.5e-12 |
| C440    | C/104459-104992 | predicted protein, Arabidopsis thaliana, PIR:T48510, 3.7e-28 |
| W455    | W/106602-110890 | glycine-rich protein 2, Arabidopsis thaliana, PIR:T48510, 3.7e-28 |
| W495    | W/122503-126090 | predicted protein, Arabidopsis thaliana, PIR:T48510, 3.7e-28 |
| W505    | W/12671-129864 | predicted protein, Arabidopsis thaliana, PIR:T48510, 3.7e-28 |
| W535    | W/13424-136831 | protein kinase (EC 2.7.1.37) S, Arabidopsis thaliana, PIR:T05050, 4.8e-138 |
| C555    | C/137352-134234 | predicted protein, Arabidopsis thaliana, EMBL:AF008267, 3.1e-81 |
| W595    | W/145530-149814 | ubiquitin thiolesterase (EC 3.1.2.15) L3, human, PIR:A40085, 3.3e-52/AU058175, C25025 |
| C600    | C/150409-152412 | predicted protein, Arabidopsis thaliana, PIR:T49227, 5.4e-73/AU163243 |
| C635    | C/159754-160665 | CD8BP Mesembrinanthemum crystallinum AP2-related transcription factor, EMBL:AF245119, 7.9e-51/AU057740, AU083516 |
| C650    | C/164858-168510 | predicted protein, Arabidopsis thaliana, PIR:T04048, 2.0e-77 |
| W700    | W/180768-181805 | predicted protein, Arabidopsis thaliana, EMBL:AF245119, 7.9e-51/AU057740, AU083516 |
| C740    | C/192640-193413 | predicted protein, Arabidopsis thaliana, PIR:T04048, 2.0e-77 |
| W760    | W/195850-197387 | predicted protein, Arabidopsis thaliana, 1.8e-10 |
| C780    | C/200773-205219 | predicted protein, Arabidopsis thaliana, EMBL:A082404 |
| C785    | C/209473-210472 | predicted protein, Arabidopsis thaliana, EMBL:A082404 |
| W815    | W/211666-217666 | predicted protein, Arabidopsis thaliana, EMBL:A082404 |
| C825    | C/219153-221701 | predicted protein, Arabidopsis thaliana, EMBL:A082404 |
| C855    | C/224272-224625 | predicted protein, Arabidopsis thaliana, EMBL:T05050, 4.8e-138 |
| C870    | C/225607-226540 | predicted protein, Arabidopsis thaliana, EMBL:A082404, 1.5e-98 |
| W920    | W/235513-239385 | predicted protein, Arabidopsis thaliana, EMBL:A082404, 1.5e-98 |
| W930    | W/244097-247111 | predicted protein, Arabidopsis thaliana, EMBL:A082404, 1.5e-98 |
| W940    | W/248749-249612 | predicted protein, Arabidopsis thaliana, PIR:T01076, 7.4e-39 |
| W950    | W/249771-250715 | predicted protein, Arabidopsis thaliana, PIR:T01076, 7.4e-39 |
| Retrotransposon | Retrotransposon | Retrotransposon |

**Table 1. Features Identified in Rice Sequence Data**
substitution. For P0436E04, 26 genes and three transposons were identified in 145 kb of sequence, compared with 24 genes and two transposons recorded with the submission. For P0406H10, 25 genes and one transposon were identified in 156 kb of sequence, compared with 26 genes and one transposon recorded with the submission. The densities of putative genes identified, one per 5.6 kb and one per 6.2 kb for P0436E04 and P0406H10, respectively, are very similar to the density found in the 340-kb region analyzed (one per 6.1 kb). Although the annotation accompanying database submissions of rice genome sequence suggested significantly different gene structures to those predicted by our protocols, the overall gene density predicted is very similar. The gene densities of these clones are typical of those accompanying the rice BAC sequences presently in the public databases. These results suggest that a typical gene density for the rice genome is around one gene per 6 kb.

Searches were conducted for segments of the *Arabidopsis* genome that contain conserved gene content and order for each of BAC clones P0436E04 and P0406H10. The same methods and recording criteria were used. The results are summarized in Table 4 and Figure 4 for P0406H10, and Table 5 and Figure 5 for P0436E04. In both cases multiple conserved segments were identified. Only three or four conserved genes were identified in each segment; there was one reversal of gene order (W1600-AT5g07380 and W3350-AT5g07690), and the strand orientation of several of the genes was not conserved (i.e., C3852-AT1g80360, C2000-AT4g32610, W399-AT5g63880, C3900-AT5g70800). However, these results indicate that it may be feasible to align much of the rice genome with duplicated segments of the genome of *Arabidopsis*.

**DISCUSSION**

Using a combination of approaches, 56 genes were predicted in the 340 kb of rice genome sequence data we generated and analyzed, indicating a density of one gene per 6.1 kb. This density is close to that found for the genome of *Arabidopsis*; that is, one gene per 4.76 kb (Bancroft 2000), but higher than that found near the ADH1 locus of rice, one gene per 10.3 kb (Tarchini et al. 2000). Extrapolation to the 440-Mb genome of rice, using the gene densities of one per 6.1 kb or 10.3 kb,

### Table 2. Functional Domains of Predicted Rice Proteins

| Predicted gene | Interpro domains |
|----------------|------------------|
| C15            | IPR000104 Type I antifreeze protein |
| W50            | IPR001687 ATP/GTP-binding site motif A (P-loop) |
| W230           | IPR002952 Eggshell protein |
| C90            | IPR001472 Bipartite nuclear localization signal |
| W265           | IPR00694 Proline-rich region |
| W325           | IPR001878 Zn-finger CCHC type |
| C345           | IPR002965 Proline rich extensin |
| W485           | IPR000734 Lipase |
| W495           | IPR001753 Enoyl-CoA hydratase/isomerase |
| W505           | IPR001753 Enoyl-CoA hydratase/isomerase |
| W535           | IPR002290 Serine/Threonine protein kinases active-site |
| C600           | IPR006694 Proline-rich region |
| C635           | IPR001471 AP2 domain/ethylene responsive element binding protein |
| W700           | IPR001471 AP2 domain/ethylene responsive element binding protein |
| W760           | IPR002293 Permease for amino acids and related compounds, family I |
| C780           | IPR003079 Esterase/lipase/thioesterase active site |
| W815           | IPR001578 Ubiquitin carboxyl-terminal hydrolases family 1 |
| C825           | IPR001245 Tyrosine kinase catalytic domain |
| W875           | IPR001477 AP2 domain/ethylene responsive element binding protein |
| C900           | IPR001471 AP2 domain/ethylene responsive element binding protein |
| W920           | IPR001440 TPR repeat |
| W930           | IPR001623 DnaJ-N-terminal domain |
| W940           | IPR001778 Pollen allergen Poa pl signature |
| W950           | IPR001472 Bipartite nuclear localization signal |
| W1050          | IPR001778 Pollen allergen Poa pl signature |
| W1110          | IPR001305 TATA binding protein |
| W1125          | IPR001472 Bipartite nuclear localization signal |
| W1145          | IPR001477 AP2 domain/ethylene responsive element binding protein |
| W1190          | IPR001471 AP2 domain/ethylene responsive element binding protein |
| C1225          | IPR002086 Aldehyde dehydrogenase family |

### Analysis of Additional Regions of the Rice Genome

To assess the generality of our findings of gene density in the rice genome and the conservation of microstructure with the genome of *Arabidopsis*, we selected for analysis two further BACs that had been sequenced and submitted to public databases. One of these, P0436E04 (accession no. ap002818), was selected as the sequenced clone nearest to a telomere (map position 0.3 cm on chromosome 1). The other, P0406H10 (accession no. ap002524), was near the middle of a chromosome arm (20.2 cm on chromosome 1). We implemented our annotation protocols using these data and compared the putative genes derived with those accompanying the database.
would predict the presence of $\sim 72,000$ or $\sim 43,000$ genes in the rice genome, respectively. However, there is evidence of gene-rich and gene-poor isochores in the rice genome based on bulk sequence composition (Barakat et al. 1997), and both regions analyzed are likely to be characteristic of the gene-rich regions. If we estimate that the rice ESTs presently in dbEST represent $\sim 10,000$ nonredundant genes, our observation that 16 of the 56 predicted genes identified (29%) have EST matches leads us to predict a total gene number for rice of $\sim 35,000$. This would be consistent with a model in which the majority of the rice genes are contained in gene-rich regions comprising 50% of the genomic DNA of rice (220 Mb), with these gene-rich regions typically containing a gene density of one per $\sim 6$ kb, as we have observed.

The composition of the region we have analyzed differs significantly from that near the ADH1 locus (Tarchini et al. 2000). In addition to more predicted genes in a region of almost identical size (56, compared with 33), we identified fewer transposons and retrotransposons (2, compared with 15). There are smaller gene/pseudogene families; for example, the largest gene family we identified contained four members, compared to 13 members. The region around the ADH1 locus contains several genes with homology to genes known to be involved in plant disease resistance. It has a complex structure, including a large family of genes, some of which do not encode a full and functional protein, and several retrotransposons. This resembles the structure of the Arabidopsis ecotype Columbia allele of the RPP5 disease-resistance locus identified on chromosome 4 (Bevan et al. 1998). However, this is an unusual genome organization, and is not representative of the genome structure as a whole (Lin et al. 1999; Mayer et al. 1999).

Sixteen of the 56 modeled genes (29%) match EST sequences, supporting the predicted gene models. Further support for the authenticity of our predicted genes come from the highly significant homology that the predicted products of many of them show to known or predicted proteins in other species. Forty-six of the 56 predicted genes (82%) show such homology. These data suggest that the majority of our gene

Table 3. Homology Scores for Conserved Arabidopsis and Rice Genes: Rice Chromosome 2 Segment

| Rice gene | Arabidopsis genes (P value) |
|-----------|-----------------------------|
| C75       | AT2g44740 (8.1e-37) C        |
| C135      | AT2g44780 (1.2e-83) C; AT3g60260 (1.5e-83) W |
| C200      | AT5g47180 (4.3e-54) C        |
| C202      | AT4g17560 (1.0e-22) W; AT5g47190 (2.3e-39) C |
| W265      | AT4g17550 (2.1e-143) C       |
| C345      | AT3g60270 (1.0e-25) C        |
| W455      | AT4g17530 (6.0e-96) C; AT5g47200 (1.1e-71) W |
| W485      | AT2g44810 (3.7e-70) W; AT4g16820 (1.3e-105) W |
| W495      | AT4g16800 (3.7e-54) C        |
| W505      | AT4g16800 (2.0e-58) C        |
| W535      | AT2g44830 (1.4e-170) W       |
| W595      | AT4g17510 (3.8e-71) C        |
| C600      | AT3g60320 (9.6e-74) C        |
| C635      | AT4g17500 (6.6e-35) W; AT5g47220 (4.7e-31) C; AT2g44840 (4.1e-30) W |
| C650      | AT4g44850 (8.9e-63) W        |
| W700      | AT4g17490 (3.1e-21) C; AT5g47230 (9.2e-27) W |
| W760      | AT4g17486 (2.7e-44) C; AT5g47310 (1.1e-50) W |
| C870      | AT4g16790 (5.5e-7) C; AT3g60380 (4.4e-9) C |
| W940      | AT3g44940 (1.7e-36) C; AT4g16750 (3.1e-31) C |
| W950      | AT4g17340 (6.7e-97) W; AT5g47450 (4.1e-97) C |
| W1050     | AT4g17350 (8.8e-46) W; AT5g47440 (1.8e-46) C; AT4g16670 (1.3e-36) W |
| C1190     | AT2g44940 (1.6e-31) C; AT3g60490 (3.7e-31) W |

Figure 3 Comparison of the organization of conserved putative genes in the 333-kb rice DNA sequence and five segments of the Arabidopsis genome sequence. The relative coding strands of the genes are indicated by up- or down-pointing polygons. Duplicated genes in the rice sequence that detected homology to common Arabidopsis genes are indicated next to each other at the position of the gene with the lower reference number.

Table 4. Homology Scores for Conserved Arabidopsis and Rice Genes: BAC P0406H10

| Rice gene | Arabidopsis genes (P value) |
|-----------|-----------------------------|
| W750      | AT3g24480 (1.5e-106) C      |
| W1200     | AT3g24460 (2.9e-32) C       |
| C1700     | AT4g32620 (8.1e-100) W      |
| C1702     | AT4g80180 (8.1e-08) W; AT4g32620 (3.7e-32) W |
| C2000     | AT4g32610 (4.1e-09) C       |
| C2002     | AT3g43450 (4.8e-94) W       |
| C2702     | AT4g80400 (3.7e-96) W; AT4g32600 (8.9e-102) W |
| W265      | AT3g80390 (7.7e-24) C       |
| C3852     | AT1g80360 (7.3e-68) C       |
contributors to “exon shuffling” in mammalian systems (Boeke 1999). This may involve mobile genetic elements, as has been found to achieve the observed divergence of genome fine structure genetic maps (Paterson et al. 1996). Mechanisms that might rearrange, permitting the construction of unified genome organization might have evolved primarily by gross evolutionary. This is inconsistent with the suggestion that plant genomes (Lin et al. 1999; Mayer et al. 1999). It is also likely that unequal crossing over results in extensive genome reshuffling during plant evolution. This is inconsistent with the suggestion that plant genome organization might have evolved primarily by gross rearrangements, permitting the construction of unified genetic maps (Paterson et al. 1996). Mechanisms that might achieve the observed divergence of genome fine structure may involve mobile genetic elements, as has been found to contribute to “exon shuffling” in mammalian systems (Boeke and Pickeral 1999). It is also likely that unequal crossing over contributes to both tandem duplications of genes, and deletion of single or small groups of genes (Bancroft 2001).

Many duplicated regions have been identified within the genome of Arabidopsis (Lin et al. 1999; Mayer et al. 1999; Bancroft 2000). It has been suggested that these may have been the result of an ancestral tetraploidy event (Blanc et al. 2000; The Arabidopsis Genome Initiative 2000), or multiple duplication events (Vision et al. 2000). Our data support the hypothesis that there have been multiple duplication events during the evolution of the genome of Arabidopsis. These duplicated segments appear to have diverged extensively by the loss of different subsets of interspersed genes. The relationships of such highly diverged duplicated segments is revealed most clearly by comparative sequence analysis with relatively distantly related species, such as tomato (Ku et al. 2000) or rice. By integrating the data from multiple duplicated segments of the Arabidopsis genome we have been able to align segments of the rice and Arabidopsis genomes and deduce the ancestral relationships of sets of genes. It is not known whether the 340-kb rice genome segment studied is also the product of genome duplication events during the ancestry of rice. When the rice genome sequence data become available, it should be possible to analyze complex relationships within the rice genome by extensive analysis using the Arabidopsis genome sequence. By taking due account of the mechanisms of the evolution of plant genome structure, it may be possible to make extensive use of comparative genome analysis to integrate structural and functional genomics of dicot and monocot species.

**METHODS**

**Sequencing of BAC Clones**

Individual BAC clones were sequenced by standard methods using a shot-gun approach (Bodenteich et al. 1993). Cesium chloride-purified BAC DNA was sheared by nebulization (Roe et al. 1996). After end-filling, DNA fragments were size fractionated and cloned into the HindIII site of pUC18 or HindIII site of pUC19 (Amersham Pharmacia Biotech). Clones were sequenced using the ABI PRISM Dye Terminator Cycle Sequencing Ready Reaction kit with FS AmpliTaq DNA Polymerase (PE Applied Biosystems) and analyzed on ABI 377 (PE Applied Biosystems) sequencing gels. The sequence data were assembled using PHRED/PHRAP software (Green 1996).

**Analysis of Sequence Data**

The sequence was subjected to a modified analysis procedure based on that established for genome analysis of Arabidopsis thaliana (Mayer et al. 1999). BLAST (Altschul et al. 1997) analysis of the sequence against the EMBL nucleotide database and MIPS in-house databases (a nonredundant protein database, a plant transposon database, a rice EST database, and an all-plant EST database) was performed. Gene predictions were performed using GeneScan (Burge and Karlin 1997), GeneFinder (P. Green and L. Hillier, unpubl. software),

---

**Table 5. Homology Scores for Conserved Arabidopsis and Rice Genes: BAC P0436E04**

| Rice gene models | Arabidopsis gene models (P value) |
|------------------|----------------------------------|
| W399             | AT5g03880 (5.5e-08) W             |
| W1600            | AT5g07380 (8.5e-08) W             |
| W2200            | AT5g03790 (5.3e-53) C             |
| W3350            | AT5g07680 (1.9e-60) W             |
| C3900            | AT5g03660 (1.6e-46) W; AT5g07080 (1.2e-30) W |
| C4600            | AT5g03660 (1.7e-08) W; AT5g07080 (2.1e-06) W |

---

**Figure 4** Comparison of the organization of conserved putative genes in rice BAC P0406H10 and three segments of the Arabidopsis genome sequence. Representation is as shown in Figure 3.

**Figure 5** Comparison of the organization of conserved putative genes in rice BAC P0436E04 and two segments of the Arabidopsis genome sequence. Representation is as shown in Figure 3.
ACKNOWLEDGMENTS

This work was funded under the BBSRC GAIT Initiative (grant 208/3/G10960) and the EU Arabidopsis Genome Sequencing Project (CT97-0274).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Arun muganathan, K. and Earle, E.D. 1991. Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. 9: 208–218.

Altschul, S.F., et al. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

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

Barakat, A., Carels, N., and Bernardi, G. 1997. The distribution of genes in the genomes of Gramineae. Proc. Natl. Acad. Sci. 94: 6857–6861.

Bevan, M., et al. 1998. Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of Arabidopsis thaliana. Nature 391: 485–488.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Deleney, M. 2000. Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell 12: 1093–1101.

Bodenteich, A., Chissoe, S., Wang, Y.F., and Roe, B.A. 1993. Shot-gun cloning as the strategy of choice to generate templates for high-throughput deoxyxynucleotide sequencing. In Automated DNA sequencing and analysis techniques (ed. J.C. Venter), pp. 42–50. Academic Press, London.

Boeke, J.D. and Pickeral, O.K. 1999. Retroshuffling the genomic deck. Nature 398: 108–111.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268: 78–94.

Cavell, A.C., Lydiate, D.J., Parkin, I.A.P., Dean, C., and Trick, M. 1998. Colinearity between a 30-centimorgan segment of Arabidopsis thaliana chromosome 4 and duplicated regions within the Brassica napus genome. Genome 41: 62–69.

Chen, M., SanMiguel, P., and Bennetzen, J.L. 1998. Sequence organization and conservation in 2H2A-homologous regions of sorghum and rice. Genetics 148: 435–443.

Devis, K.M., Beales, J., Nagamara, Y., and Sasaki, T. 1999. Arabidopsis rice: Will colinearity allow gene prediction across the eudicot-monocot divide? Genome Res. 9: 815–829.

Feldmann, K.A., Marks, M.D., Christianson, M.L., and Quatrano, R.S. 1989. A dwarf mutant of Arabidopsis generated by T-DNA insertional mutagenesis. Science 243: 1351–1354.

Gale, M.D. and Devis, K.M. 1998. Plant comparative genetics. Science 282: 656–659.