Whole Genome Sequence Comparisons and "Full-Length" cDNA Sequences: A Combined Approach to Evaluate and Improve Arabidopsis Genome Annotation

Vanina Castelli, Jean-Marc Aury, Olivier Jaillon, Patrick Wincker, Christian Clepet, Manuella Menard, Corinne Cruaud, Francis Quétier, Claude Scarpelli, Vincent Schächter, Gary Temple, Michel Caboche, Jean Weissenbach and Marcel Salanoubat

Genome Res. 2004 14: 406-413
Access the most recent version at doi:10.1101/gr.1515604

Whole Genome Sequence Comparisons and "Full-Length" cDNA Sequences: A Combined Approach to Evaluate and Improve *Arabidopsis* Genome Annotation

Vanina Castelli,[1] Jean-Marc Aury,[1] Olivier Jaillon,[1] Patrick Wincker,[1] Christian Clepet,[2] Manuella Menard,[1] Corinne Cruaud,[1] Francis Quétier,[1] Claude Scarpelli,[1] Vincent Schächter,[1] Gary Temple,[3,4] Michel Caboche,[2] Jean Weissenbach,[1] and Marcel Salanoubat[1,5]

[1]*Genoscope-Centre National de Séquençage and Centre National de la Recherche Scientifique Unité Mixte de Recherche-3080, 91000 Evry, France;* [2]*Institut National de la Recherche Agronomique, Unité de Recherche en Génomique Végétale, 91000 Evry, France;* [3]*Life Technologies, a Division of Invitrogen, Carlsbad, California 92008 USA*

To evaluate the existing annotation of the *Arabidopsis* genome further, we generated a collection of evolutionary conserved regions (ecores) between *Arabidopsis* and rice. The ecore analysis provides evidence that the gene catalog of *Arabidopsis* is not yet complete, and that a number of these annotations require re-examination. To improve the *Arabidopsis* genome annotation further, we used a novel "full-length" enriched cDNA collection prepared from several tissues. An additional 1931 genes were covered by new "full-length" cDNA sequences, raising the number of annotated genes with a corresponding "full-length" cDNA sequence to about 14,000. Detailed comparisons between these "full-length" cDNA sequences and annotated genes show that this resource is very helpful in determining the correct structure of genes, in particular, those not yet supported by "full-length" cDNAs. In addition, a total of 326 genomic regions not included previously in the *Arabidopsis* genome annotation were detected by this cDNA resource, providing clues for new gene discovery. Because, as expected, the two data sets only partially overlap, their combination produces very useful information for improving the *Arabidopsis* genome annotation.

[Supplemental material is available online at www.genome.org. The cDNA sequences have been released to the EMBL. The data produced during this analysis and accession nos. are available at http://www.genoscope.cns.fr/Arabidopsis/. The GSLT cDNA clones are available at Genoscope. The results can be visualized at http://www.genoscope.cns.fr/cgi-bin/ggb/ggb?source=Arabidopsis/.]

The sequence of the *Arabidopsis thaliana* genome was completed in 2000 by the *Arabidopsis* Genome Initiative (AGI; Lin et al. 1999; Mayer et al. 1999; AGI 2000; Salanoubat et al. 2000; Tabata et al. 2000; Theologis et al. 2000). The first annotation of this sequence relied on ab initio gene prediction combined with database searches, mainly by using ESTs, mRNA sequences, and protein alignments from *Arabidopsis* and other plant species (AGI 2000; Schoof and Karlowski 2003). During the past two years, the annotation of the *Arabidopsis* genome has been updated regularly (The Institute for Genomic Research [Haas et al. 2003] and Munich Information Center for Protein Sequences [MIPS; Schoof et al. 2002]). In particular, the annotation has been greatly improved by the integration of "full-length" cDNAs produced by the community (Haas et al. 2002; Seki et al. 2002a,b). For instance, in the last version of the MIPS annotation (June 2003) about 12,000 annotated gene models were supported by "full-length" cDNA sequences. These cDNA resources have been extremely useful, but are still insufficient, as ~14,000 of the annotated *Arabidopsis* genes are supported only by EST or protein resources, and/or predicted ab initio data.

[4]Present address: Intronn, Inc. Gaithersburg, MD 20878, USA.
[5]Corresponding author.
E-MAIL salanou@genoscope.cns.fr; FAX 33-01-60-87-25-14.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1515604.

To evaluate and further improve these annotations, we used two different types of data as follows. (1) Whole genome sequence comparisons between *Arabidopsis* and rice. In this strategy, we detected evolutionarily conserved regions (ecores) between *Arabidopsis* and the available rice sequence draft, as was done between the human and pufferfish genome using Exofish (Roest Crollius et al. 2000). It has been shown previously that whole genome comparisons, on the basis of a tool like Exofish, can be used as an efficient method to evaluate quality and to improve existing annotations of insect genomes (Jaillon et al. 2003). (2) Production of new cDNA sequences from enriched full-length normalized cDNA libraries constructed using mRNAs from *Arabidopsis* tissues, some of which had not been used previously for the construction of "full-length" cDNA libraries. In this study, we show that both sets of data can be combined to provide new and reliable information that substantially improves the existing annotation of the *Arabidopsis* genome.

RESULTS

In the most recent version of the *Arabidopsis* genome annotation (MIPS, June 2003), 26,446 annotated genes were identified. Because for many annotated genes the UTR regions are not available, the CDS only will be used and referred to hereafter as annotations or annotated features. Of these, 12,165 annotated gene models were supported by "full-length" cDNAs, including most of the "full-length" cDNA analyzed in Yamada et al. (2003). An

14:406–413 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

Table 1. Distribution of Ecores in the *Arabidopsis* Sequence

	Ecores	Genes	Genes detected	Ecores within genes	Exons	Exons detected	Ecores which overlap exons	Ecores within genes which do not overlap exons	Ecores/gene
numbers	80,010	26,446	19,445	73,374	137,242	64,432	72,814	562	3.8
%	NA	NA	74	92	NA	47	91	0.7	NA

additional 5321 annotated genes were supported by ESTs only, raising to 17,486 the number of additional annotated gene models supported by biological evidence. Although ESTs provide an experimental support, they are of limited interest for constructing accurate gene models. We therefore grouped annotated genes in two categories, (1) those supported by "full-length" cDNAs (Group A), and (2) those supported by ESTs only, or with no experimental support at all (Group B).

To estimate the level of completion of the annotation of the *Arabidopsis* genome, independent of existing annotation resources, we performed genome-wide sequence comparisons between *Arabidopsis* and rice genomic sequences.

Whole Genome Sequence Comparisons

The recent availability of rice genome sequence drafts offers a unique possibility to compare *Arabidopsis* and rice sequences and to search for regions conserved during evolution. We have used the "Exofish" comparative genomics method (Roest Crollius et al. 2000) to detect, with a very low background, conserved regions between *Arabidopsis* and rice, which are separated by an evolutionary distance of 135–235 million years (Sun et al. 1998; Yang et al. 1999). Exofish is a three-step process that includes alignment computing, filtering, and design of evolutionarily conserved regions (ecores, see Methods). To minimize background, we calibrated both alignment computation and filtering conditions using the Syngenta rice draft sequence (Goff et al. 2002) on a set of 1589 manually annotated genomic regions of *Arabidopsis* (P. Rouzé and S. Aubourg, pers. commun.). The optimal conditions we defined produce a specificity close to 100% on this set and a sensitivity at the gene and exon level of 93% and 64%, respectively. These settings were applied to the whole *Arabidopsis* genome compared with the set of 3424 BACs released by the IRGSP (International Rice Genome Sequencing Program; http://rgp.dna.affrc.go.jp/IRGSP/). The comparison between the *Arabidopsis* annotation and the ecores is summarized in Table 1. A total of 74% (19,445) of the annotated genes included at least one ecore and 47% of the annotated exons are matched by at least one ecore. A total of 92% of the ecores are localized within the boundaries of an annotated gene, and ~0.7% of these ecores do not match with an annotated exon (a representation of the observed cases is shown in Figure 1, Case 1). Of the 19,445 annotated genes detected by ecores, 4808 from group B do not have experimental support. Interestingly, we find 10 times more ecores outside annotated exons in group B than in group A (Table 2). After removal of background ecores (transposons, tRNA, or pseudogenes that escape masking), a total of 3285 ecores were found in intergenic regions, as defined by the current annotation. We expect that a substantial portion of these 3285 remaining ecores, which are not found in annotated regions, correspond to gene extensions or to yet undetected genes.

To further analyze the ecores lying in the intergenic regions, we have constructed models based on ecotigs (ECOre conTIG). Such models are constructed by linking in the same model two or more ecores that are located in the same relative position on both genomes (see Methods). A fraction of the ecotigs are composed of more than one gene, reflecting the microsynteny existing between *Arabidopsis* and rice (Salse et al. 2002; Vandepoele et al. 2002). The details of the ecotig analysis are provided as Supplemental information. By selecting ecotigs composed exclusively of ecores in which at least one overlaps an annotated gene and at least one is located outside of this gene, potential gene extensions were detected for 424 annotated genes (Fig. 1, Case 2). In addition, 403 ecotigs composed exclusively of ecores that do not overlap an annotation were found (Fig. 1, Case 3).

This analysis strongly suggests that the gene catalog of *Arabidopsis* is not yet complete, and that a number of existing annotations require re-examination. To address these issues, we made use of a novel collection of full-length cDNAs.

Analysis of the cDNA Collection

We have sequenced 31,558 cDNA clones (GenoScope/LifeTechnologies [GSLT]) from four normalized cDNA libraries (~9500 clones each) that originated from (1) hormone-treated callus, (2) flower buds and flowers at various developmental stages, (3) forming siliques to the developing embryo stage, and (4) leaves and stems. After gap closure, we obtained the full-insert sequence for 21,572 GSLT clones; the remaining clones corresponded to either 5′ and 3′ unassembled sequences, or 5′ and 3′ singletons (Table 3). These full-insert sequences were used to construct gene models in a two-step process. (1) We identified the location of the cDNA sequence on the *Arabidopsis* genome using BLAST alignments. This step assigned almost all GSLT cDNA sequences (>99.9%) to a unique genomic location. (2) We built gene models using two different programs to align GSLT cDNA sequences to genomic sequences (SIM4 [Florea et al. 1998], EST_GENOME [Mott 1997]), and NETGENE, a program specialized in splice-site recognition in *Arabidopsis* (Hebsgaard et al. 1996). Such gene models (called GSLT models hereafter) were

Table 2. Distribution of Ecores According to the Type of Annotated Gene Support

Annotated genes	Genes	Ecores within genes	Exons	Genes/exons detected	Ecores which overlap exons	Ecores within genes which do not overlap exons
With FL support (Groups A)	12,165	38,959	65,658	10,335/35,485	38,919	40 (0.1%)
With EST or without support (Group B)	14,281	34,428	71,584	9119/28,957	33,906	522 (1.5%)

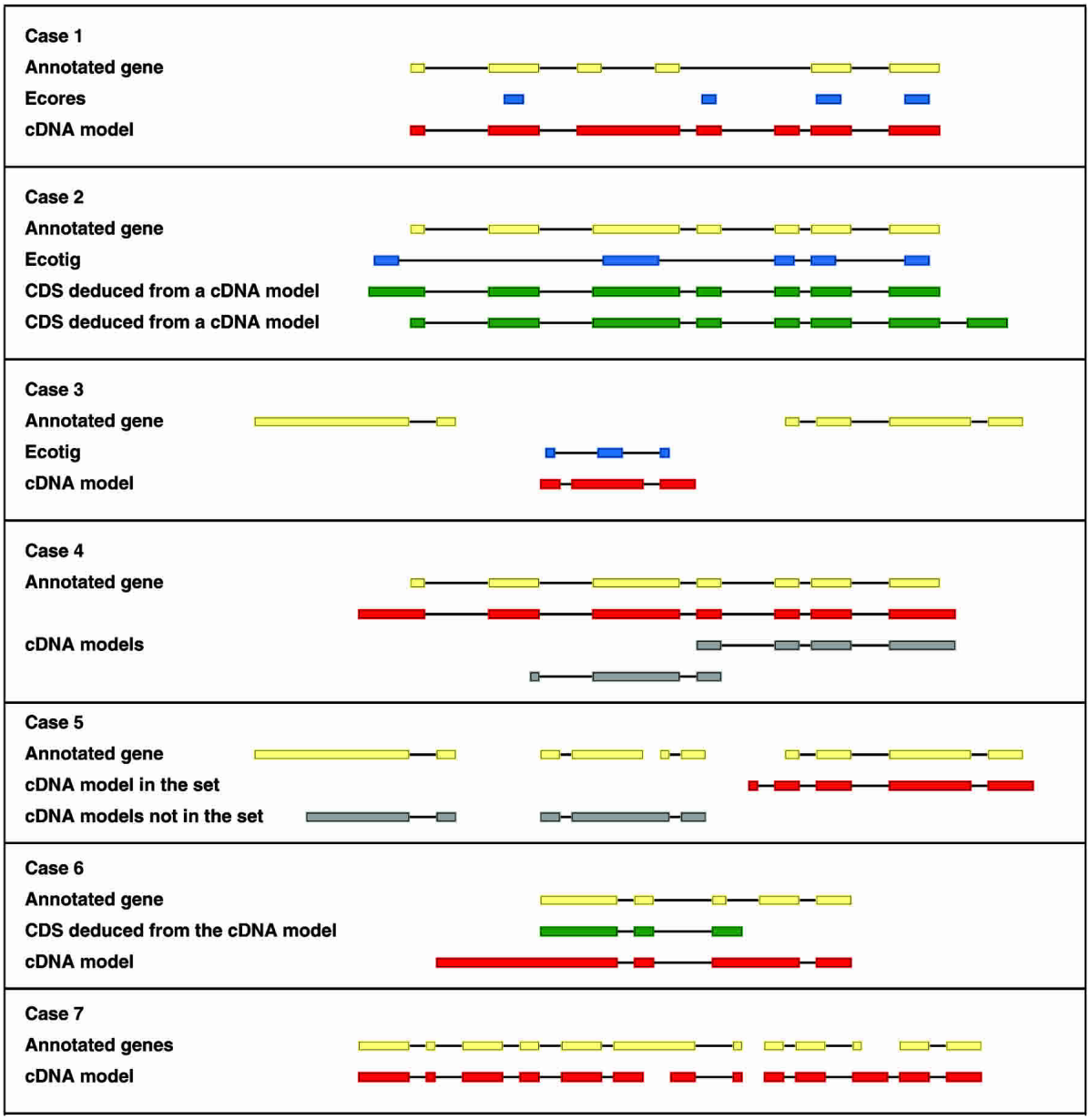

Figure 1 Graphical description of the different situations observed when comparing annotated genes with ecores/ecotigs/cDNAs. (Case 1) Missing internal exon detected by ecores or cDNA sequences. In some cases, the internal exon is only partly missing. (Case 2) Extension of an annotated gene. (Case 3) Novel gene. (Case 4) A cDNA overlapping (red) or partially overlapping (gray) an annotated gene. (Case 5) cDNAs included (red) or not (gray) in the evaluation set. (Case 6) CDS annotation extending beyond GSLT CDS. (Case 7) On the *right*, cDNA bridging two annotated genes, on the *left*, cDNAs splitting an annotated gene.

considered as validated when all of their splice sites were confirmed by at least two of the three programs (see Methods). This process yielded 18,025 validated GSLT models and 3547 non-validated GSLT models.

Because the GSLT sequences are based on the assembly of single pass reads and may contain sequencing errors, for CDS determination, we generated a cDNA sequence (virtual cDNA) using the matching *Arabidopsis* genomic sequence. CDSs for 21,572 cDNA clones for which a full-insert sequence was available were determined (see Methods). Additional information on the GSLT resource is available as Supplemental information.

We compared the length of the sequences from the GSLT resource with the 21,797 publicly available mRNA sequences with complete CDSs (GenBank, PLN section release 133) referred to hereafter as E-A-mRNA (Existing *Arabidopsis* mRNAs). A small subset of these E-A-mRNAs were not in the MIPS June 2003 annotation. The most 5′ and 3′ sequences from both data sets were selected, and their size differences calculated for 4841 and 4836 pairs, respectively. The results are shown in Figure 2. In 26% of the cases (1244), sequences from the GSLT resource extend the 5′ end sequence of the E-A-mRNA resource, and for 61 of these, at least one novel 5′ exon was detected. A list of 5′ and 3′ extensions with novel exon(s) is available at http://www.genoscope.cns.fr/Arabidopsis/file1. In some cases, the coding region was also extended. An example is shown in Figure 3. When this analysis was not restricted to the most 5′ sequence from the GSLT

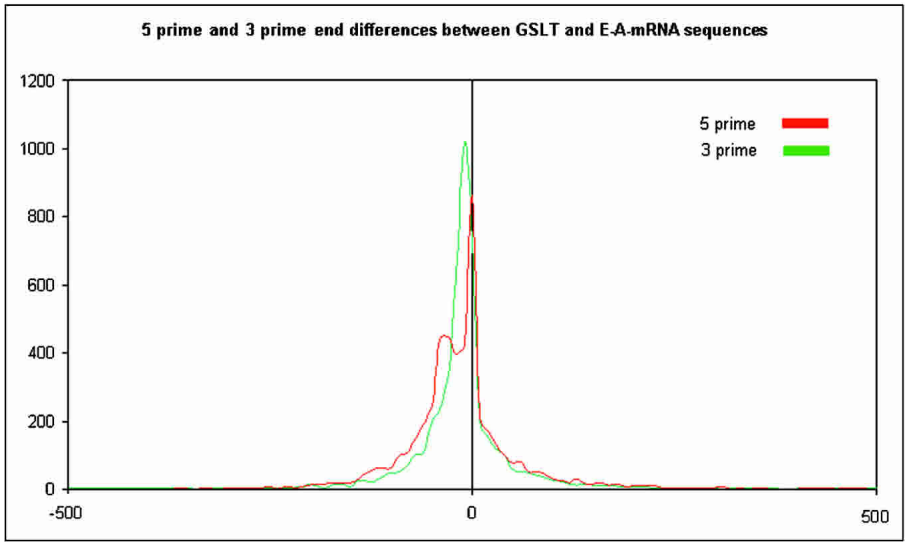

Figure 2 Comparison of the length of GenBank and GSLT cDNA. 5′ (green) 3′ (red). Positive abscissa values correspond to cases in which the GSLT cDNA extends the E-A-mRNA resource, whereas negative values correspond to longer E-A-mRNA cDNAs. The Y axis corresponds to the number of cases found at a given X value.

resource, the GSLT sequences extended the 5′-end sequence from the E-A-mRNA resource in 22% of the cases. Furthermore, 77% of the GSLT CDS sequences started either at the same ATG or at an upstream ATG, compared with the E-A-mRNA CDS sequences.

GSLT Models and Annotated Gene Structure

Of the 18,025 GSLT clones for which a validated gene model was available, 17,159 overlapped 9297 annotated genes, at least partially (see Fig. 1, Case 4), 326 overlapped 251 annotated genes, but on the opposite strand, and 540 are located in regions with no gene annotation. Additional information on nonvalidated gene models and unassembled sequences is available as Supplemental information. Of the 9297 annotated genes overlapped by the GSLT validated gene models, 6429 were already supported by a "full-length" cDNA, 1967 by ESTs, and 901 were not supported by expression data.

To evaluate the impact of the GSLT resource on the genome annotation, we used a suitable subset of these 18,025 GSLT clones. This subset (13,031 clones) is restricted to cDNA sequences covering the totality of an annotated gene, and matching this gene solely (Fig. 1, Case 5; Table 4). We then compared the CDS deduced from the GSLT models with the annotated CDSs from groups A and B defined above (results in Table 5).

As expected, the vast majority (95%) of the annotated gene models for Group A were confirmed by the GSLT clone analysis, validating these annotations and our analysis simultaneously. Conversely, ~45% of annotated gene models not supported by "full-length" cDNAs needed to be inspected for extensions, missing exons, and incorrect splice sites. Lists corresponding to dubious annotations can be found at http://www.genoscope.cns.fr/Arabidopsis/file 2 to 5 and used to explore the supporting evidence on a browser (http://www.genoscope.cns.fr/cgi-bin/ggb/ggb?source=Arabidopsis).

Existing CDS annotation goes beyond the GSLT CDS (Fig. 1, Case 6) for 388 (7%) genes from group A and 306 (15.8%) from group B, although the sequence of the clone covers the totality of the annotated gene. Manual inspection of a 10-Mb region shows that these mostly correspond to GSLT cDNAs that are probably derived from immature mRNAs. This was further confirmed by the sequence of a publicly available mRNA (E-A-mRNA) in 70% of the 388 annotated genes from group A, suggesting that these gene models are accurately annotated. However, in 36 and 43 annotated genes from the A and B groups, respectively, GSLT models from two independent clones at least disagree with the proposed annotated gene (http://www.genoscope.cns.fr/Arabidopsis/file 6 and 7). In most of the cases, (21/43) the difference between the annotated gene and the GSLT model is due to an unspliced intron located in the same position in at least two GSLT sequences. Alternative splicing is the second most frequent explanation (6/43); one example is shown in Figure 4.

Novel Genes

The GSLT cDNA resource was used to detect new *Arabidopsis* genes that were overlooked during previous annotation processes. Using an automated analysis, we detected 326 genomic regions not overlapped by an annotated gene, but covered by at least a GSLT cDNA sequence. For each region, the cDNA clone with the longest CDS was selected. These unannotated regions were classified according to the relative size of the CDS and exon number (Table 6). Of the 326 classified regions, 96 show evolutionarily conserved regions (see below) (http://www.genoscope.cns.fr/Arabidopsis/file 8)

Additional Features

One of the difficulties encountered during the annotation process is to define the correct beginning and end of a gene (Fig. 1,

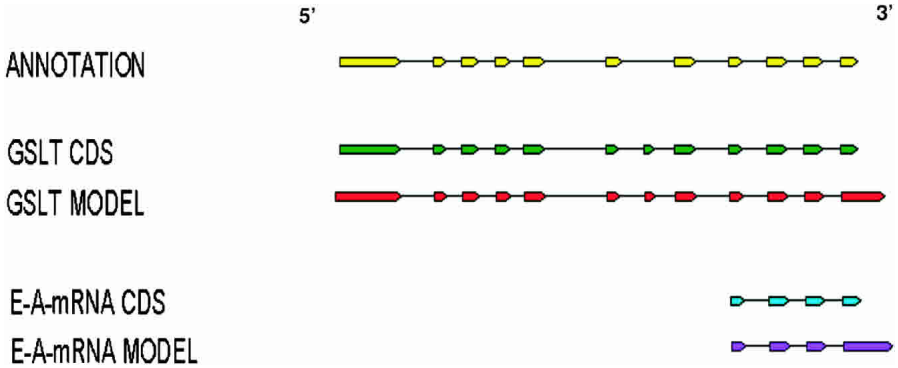

Figure 3 An example of 5′ extension detected by the GSTL resource. In this example, the gene structure of At3g58760 can also be corrected for a missing exon located between exons 6 and 7 of the annotated gene, due to longer cDNAs present in the GSLT resource.

Table 3. Sequencing Statistics

Number of cDNA clones	Number of sequences (including gap closures)	Number of clones with their full insert sequences	Number of unassembled or singleton clones
31,558	64,779	21,572	9986

Case 7). In some cases, erroneous predictions lead to a gene model that merges or splits real genes. We searched for GSLT cDNA sequences bridging two or more consecutive annotated genes, and found 93 regions (186 annotated genes) in which two genes could potentially be merged. Conversely, we found 35 cases in which two nonoverlapping GSLT sequences were included in the same gene annotation, raising the possibility that the annotation had merged two real genes (http://www.genoscope.cns.fr/Arabidopsis/file 9 and 10)

Alternative splicing is thought to be rare in plants as compared with mammals, although the number of known cases is increasing (Jordan et al. 2002; Kong et al. 2003) and may only represent the tip of the iceberg (Kazan 2003). For 702 annotated genes, the GSLT and E-A-mRNA cDNA sequences allowed the construction of more than one gene model. Of these, 276 show in-frame internal modifications of the CDS due to a different acceptor and/or donor splice-site usage, exon skipping, or an unspliced intron. For another 206 annotated genes, the alternative gene model extends the annotated CDS. In a large number of cases, the CDS extension is caused by alternative splicing as shown in Figure 5. Frequently, also the cDNA supporting the annotated gene is not "full-length" and not even "full-coding". Other types of annotated gene modifications can be found at http://www.genoscope.cns.fr/Arabidopsis/file 11 to 13.

Table 4. Data Set Used for Evaluation of the Current Annotation

Number GSLT clones corresponding to annotated genes from group A	Number of annotated genes from group A with GSLT clones	Number GSLT clones corresponding to annotated genes from group B	Number of annotated genes from group B with GSLT clones
10,196	5558	2835	1931

In addition, we found 326 GSLT sequences that overlapped 251 annotated genes, but on the opposite strand, and could correspond to antisense RNAs. In most of the cases, visual inspection did not reveal cloning artifacts. Of the 251 annotated genes, 166 were covered by both a cDNA supporting the annotated gene and an antisense cDNA. In 141 cases, the antisense cDNA was unspliced and did not permit the exclusion of possible genomic DNA contamination. Interestingly, in 12 cases, more than one unspliced antisense cDNA (GSLT and E-A-mRNA) was found for a given gene, increasing the possibility that they may correspond to antisense RNAs. In 25 cases, the antisense cDNA contained splicing events, and in all of these cases, the GT-AG splice site consensus sequence was found.

Combining Comparative Genomics and cDNA Data

Using the GSTL cDNA resource and systematic intergenome comparisons, we were able to provide biological support for an additional 3756 annotated genes. A total of 1209 of these are supported by GSLT cDNAs, and the remaining are supported by an ecotig only (http://www.genoscope.cns.fr/Arabidopsis/file 14).

Additional internal exons have been suggested by the observation of a total of 562 ecores mapping within annotated genes, but which do not match annotated exons (Fig. 1, Case 1). Most of these ecores (522) map in 380 group B gene annotations (not supported by a full-length cDNA). Fifty of these genes, as well as all of the 66 ecores that reside in these models but outside annotated exons, are matched by GSLT cDNAs. This suggests also that a vast majority of the remaining 456 ecores of group B annotations represent true exons. By combining the two data sets, potential missing exons were detected for 456 annotated genes.

Among the 424 annotated genes potentially extended by ecotigs, 192 could be matched by GSLT or E-A-mRNA. In 153 cases (80%), one ecore localized in the extension is matched by an exon of a cDNA, and in 118 cases, this extending exon is part of the CDS deduced from the cDNA (Fig. 6). A total of 947 potential gene extensions were found when systematic intergenome comparisons and full-length GSLT cDNA were used.

In addition, 403 ecotigs were composed exclusively of ecores mapping outside an annotation. A total of 87 of these ecotigs (22%) were colocalized with at least one cDNA (GSLT and/or E-A-mRNA). An example is presented in Figure 7. Some of these could correspond to novel genes (75), whereas others (12) appear to be extensions of annotated models (cDNAs that also match an annotated gene). The rest of these ecotigs (316) represent additional potential novel genes or gene extensions, and could be targeted for biological validation using reverse transcribed PCR with primers designed from the corresponding ecore sequences (http://www.genoscope.cns.fr/Arabidopsis/file 15 and 16).

Table 5. Comparison Between CDS From GSLT-Deduced and Annotated Gene Models

	Annotated gene models with a CDS reduction	Annotated genes with no CDS reduction	Annotated genes Identical to GSLT models	Annotated genes with internal modifications proposed by GSLT models	Annotated genes with a proposed extension	Annotated genes with extension including additional exons
Group A 5558	388 (7%)	5170	4904 (94.9%)	142 (2.7%)	178 (3.4%)	107 (2.1%)
Group B 1931	306 (15.8%)	1625	890 (54.8%)	298 (18.3%)	425 (26.1%)	272 (16.7%)

The sum is over 100%, as an annotated gene can belong to more than one category.

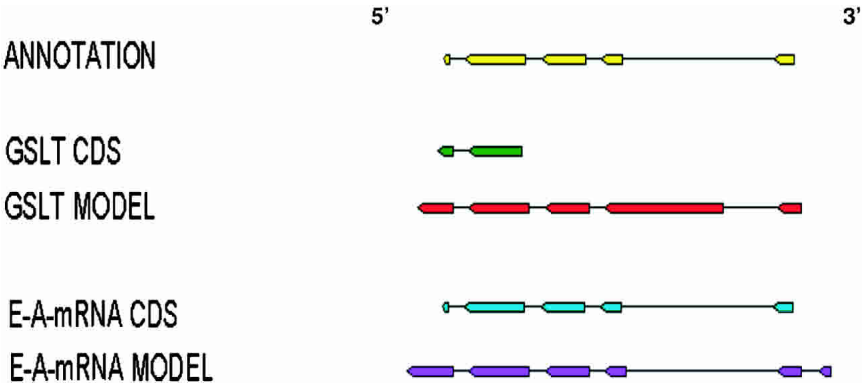

Figure 4 The gene structure of At3g61860 was confirmed by three E-A-mRNA cDNAs. In the GSLT resource, we found three other cDNAs corresponding to a different gene structure (only one of each is represented). The difference between gene models is caused by the usage of an alternative 3′ acceptor site for intron 1.

DISCUSSION

In this study, we attempted to estimate the degree of completion and accuracy of the existing annotation of the *Arabidopsis* genome and to improve it by using novel data sets on the basis of systematic intergenome comparisons and full-length enriched cDNA libraries.

Table 6. Features of the Possible Novel Genes

Number of exons	>1		=1		
CDS size/cDNA size (%)	>60	<60	>60	<60	Total
Number of corresponding genomic regions	49	98	25	154	326
Number of genomic regions with ecore(s)	29	44	3	20	96

A systematic intergenome comparison was performed between rice and *Arabidopsis* genomes. The ecore analysis provides (1) a way to monitor the degree of completion of genome annotation, (2) a method to refine the proposed gene models, and (3) a resource for novel candidate gene models. About half of the 8% of ecores that fell outside gene annotations could be ascribed to background, suggesting that the fraction of coding features that remains unannotated is very low (4%–5%). This fraction corresponds either to parts of existing genes or to novel genes.

As an example, we estimate, on the basis of the analysis of a subset of 66 nonexon-matching ecores, that most of the 562 nonexon-matching ecores, detected within the boundary of an annotated gene, correspond to missing or alternative internal exons. Of these, 40 are found in the group of annotated genes supported by "full-length" cDNAs and 522 are in the group not supported by a "full-length" cDNA (~14,300). A total of 424 potential gene extensions were detected by the ecotig construction. When these genes are overlapped by a cDNA sequence, 80% of these extensions are confirmed. In 19% of cases, ecore(s) located in the predicted extension are not part of the CDS. This probably results from the absence of a real "full-ORF" cDNA for the corresponding gene. In addition, we found 403 ecotigs that represent clues for novel gene discovery or additional gene extensions, as ecotigs could have been artificially split due to the provisional state of the rice genome sequence we used.

Of the 21,572 complete GSTL cDNA sequences that we produced, 20,407 correspond to 10,512 annotated genes. A detailed comparison between the annotated genes and the GSLT-based models shows a very high contrast between the annotated genes supported by "full-length" cDNAs and those that are not supported. Identical gene models are found in 95% and 55% of the cases, respectively. The discrepant 5% and 45% displayed either splice-site differences, exon skipping, or 5′ and/or 3′ extensions. Given the limitations in the cDNA approach, such discrepancies do not necessarily invalidate the annotated gene models. In some instances, alternative gene models were found to be due to alternatively spliced isoforms, showing the usefulness of the GSLT resource to further document genes already supported by cDNA sequences. In addition, a small number of discrepancies between annotated gene models and GSLT models could be explained by errors either in the genomic sequence or in splice-site determination in the GSLT models.

The GSLT resource also permitted discovery of yet undetected genes. We found 326 genomic regions covered by a cDNA sequence with no corresponding annotated genes, pseudogenes, or transposons. Of these regions, 147 correspond to a spliced gene model, excluding the possibility that the cDNA sequence results from the cloning of genomic DNA. Furthermore, 73 of these regions have corresponding ecores and represent good candidates for novel genes (see Table 6). The situation of the remaining regions requires further investigation, as the size of the CDS of true genes can be very short, 80 amino acids in plants (Cock and McCormick 2001) and 50 amino acids in yeast (Kessler et al. 2003), and such small ORF genes may have been systematically overlooked during the annotation process.

The GSLT cDNA sequences are being incorporated into the next release of the *Arabidopsis* genome annotation, which is in progress at TIGR and MIPS. As shown here, it will allow the validation, updating, or discovery of thousands of gene models, as well as the recognition of alternate splice sites.

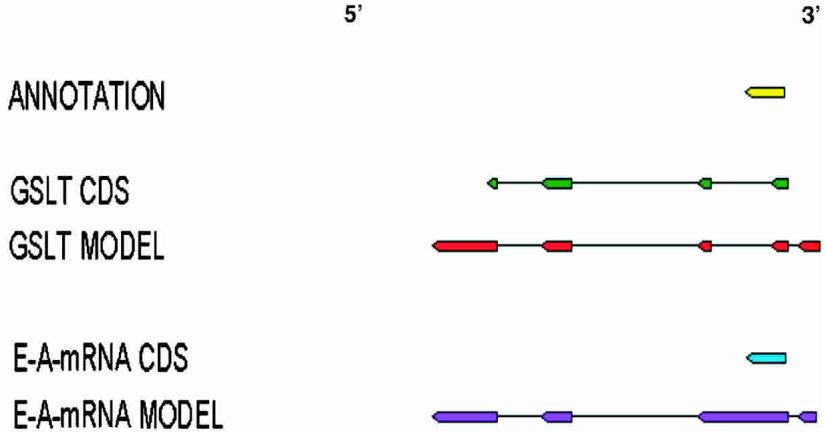

Figure 5 The At4g21215 gene structure is confirmed by an E-A-mRNA cDNA. A cDNA from the GSLT resource leads to another gene structure, due to the presence of a supplementary intron.

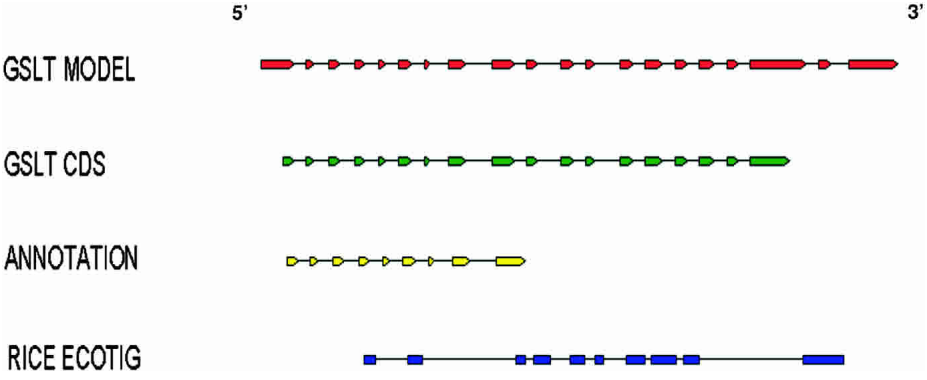

Figure 6 A 3′ extension of an annotated gene model, detected by both an ecotig and a GSLT cDNA sequence (At1g20100).

The results obtained from comparative genomics are also valuable in improving the *Arabidopsis* genome annotation. Unfortunately, the incorporation of the ecores or ecotigs in an annotation process is not as straightforward as for the cDNA sequences. Although the specificity is high, the sensitivity is estimated to be close to 50%, so Exofish is not sufficient per se to refine the gene models, and gene modeling has to incorporate additional data. For instance, ab initio gene prediction programs could be improved significantly if they could take into account the presence of ecores when constructing new gene models. Nevertheless, with the increasing number of genomic sequences generated, it is obvious that comparative genomics will be an important tool for genome annotation in the coming years.

METHODS

Library Construction

cDNA libraries were constructed with mRNA extracted from four different tissues (accession Col 0): leaf and stem, hormone-treated callus, flower buds, and flowers at various developmental stages and forming siliques in the developing embryo.

Four normalized libraries were prepared at Invitrogen Corp. as follows: First and second-strand cDNA were synthesized from poly(A)$^+$ mRNA, using Superscript II RT (Invitrogen) and an oligo-dT primer containing a NotI site, following the protocols described in the Invitrogen Manual: SuperScript Plasmid System with Gateway Technology for cDNA Synthesis & Cloning (Cat 18248013) http://www.invitrogen.com/content/sfs/manuals/18248.pdf. The cDNA was polished with T4 polymerase, digested with NotI to create 5′-blunt/3′-Not I cDNA, then size-fractionated on a gel, purified, and ligated into the pCMV-Sport6.1 vector. The libraries were normalized to Cot-10, essentially following method 2-1 of Bonaldo et al. (1996), to reduce abundant sequences and to increase the frequency of rare or novel transcripts. Quality control of the normalized library was performed by comparing hybridization between standard and normalized library to confirm an average reduction of at least 10-fold for the abundant sequences.

Sequencing Procedure

All cloned inserts were sequenced at both ends using a primer complementary to the vector sequence (for the 5′ end read) and a primer anchored to the poly(A) tail (for the 3′ end read). When the sequences obtained from end sequencing were not sufficient to cover the complete insert sequence, one or two primer walking sequences were generated (only one cDNA clone per gene).

Alignment of cDNA Sequences on the *Arabidopsis* Genome and CDS Construction

We first use BLAST to generate the alignments between the microsatellite repeat-masked cDNA sequences and the genomic sequence using the following settings: W = 20, X = 8, match score = 5, mismatch score = −4. The sum of scores of the HSPs (High-Scoring Pairs) is then calculated for each possible location, then the location with the higher score is retained if the sum of scores is more than 1000. Once the location of a cDNA sequence is determined, the corresponding genomic region is enlarged by 5 kb on each side and is used to align the cDNA sequence with Sim4 and EST_GENOME (using the following settings Sim4 : W = 15 K = 30 C = 14 R = 2 A = 4; EST_GENOME : mismatch 2, Gap penalty 3). Splice sites are also determined in this region using NETGENE. The resulting splice site positions are automatically defined and compared. If, for a defined cDNA sequence, all of the splice sites are identical for at least two of the three programs used, the model is considered as a validated gene model. When the reconciliation is not possible, the gene model proposed by EST_GENOME is used. For all of the gene models, the longest ORF was determined and the CDS constructed by using the first ATG found.

Exofish Procedure

To determine the conditions that would generate alignment in coding regions, we first tested a large range of TBLASTX conditions (W,X, scoring matrix) between a well-annotated set of 1589 genes including introns, exons, and 100 bp of intergenic region at both ends of each gene (P. Rouzé and S. Aubourg, pers. commun.) and the Syngenta Rice draft sequence. All sequences were masked against known repeats from rice and *Arabidopsis*. For each condition, a filter was applied on the basis of the length and percent identity of alignments, in order to exclusively retain those that are located in a coding region. For a given set of BLAST conditions and filter, we selected the conditions that provided the highest sensitivity (match score = 15, mismatch score = −3, W = 4, X = 13). Finally, we joined overlapping alignments to form ecores. Hence, Exofish is a three-step process—compute alignments/filter/create ecores.

We also assembled *Arabidopsis* ecores to create ecotigs. These ecotigs group the ecores together as long as they are colinear on the two genomes. Two consecutive ecores on the *Arabidopsis* genome are in the same model if these two ecores are composed of at least two consecutive HSPs, or if they are separated at most by one HSP on the rice genome (Jaillon et al. 2004). If there

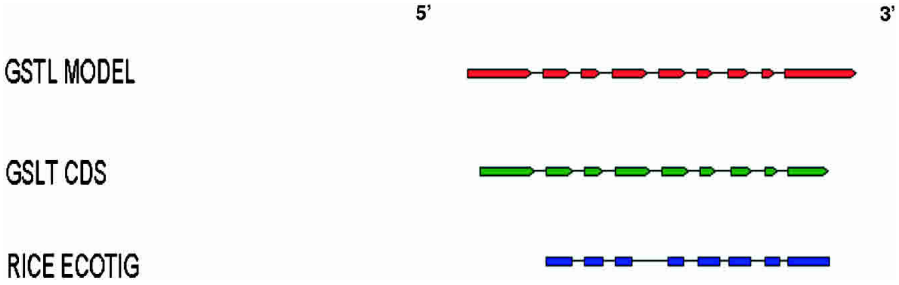

Figure 7 Novel gene detected by an ecotig and a GSLT cDNA sequence (GSLTF85ZE11, accession no. BX819512).

is little synteny between the two analyzed genomes, the ecotig corresponds mainly to real gene models, whereas in other cases, these models correspond to syntenic regions.

ACKNOWLEDGMENTS

We thank Chris Gruber and Mark Smith for library construction, Sébastien Aubourg and Pierre Rouzé for providing the set of *Arabidopsis* manually annotated genes prior to publication. We thank Nathalie Choisne, Sylvie Samain, Nadia Demange, Agnes Violet, and Susan Cure for help during the course of the project. We also thank Franck Aniere and the entire system network team.

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

Arabidopsis Genome Initiative (AGI) 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Bonaldo, M.F., Lennon, G., and Soares, M.B. 1996. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res.* **6:** 791–806.

Cock, J.M. and McCormick, S. 2001. A large family of genes that share homology with CLAVATA3. *Plant Physiol.* **126:** 939–942.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967–974.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., et al. 2002. A draft sequence of the rice genome (*Oryza sativa L. ssp. japonica*). *Science* **296:** 92–100.

Haas, B.J., Volfovsky, N., Town, C.D., Troukhan, M., Alexandrov, N., Feldmann, K.A., Flavell, R.B., White, O., and Salzberg, S.L. 2002. Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol.* **3:** RESEARCH0029.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith Jr., R.K., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31:** 5654–5666.

Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouze, P., and Brunak, S. 1996. Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.* **24:** 3439–3452.

Jaillon, O., Dossat, C., Eckenberg, R., Eiglmeier, K., Segurens, B., Aury, J.M., Roth, C.W., Scarpelli, C., Brey, P.T., Weissenbach, J., et al. 2003. Assessing the *Drosophila melanogaster* and *Anopheles gambiae* genome annotations using genome-wide sequence comparisons. *Genome Res.* **13:** 1595–1599.

Jaillon, O., Aury, J.-M., Roest Crollius, H., Salanoubat, M., Wincker, P., Dossat, C., Castelli, V., Boudet, N., Samair, S., Eckenberg, R., et al. 2004. Genome-wide analyses based on comparative genomics. In *Cold Spring Harbor Symposia on Quantitative Biology, Vol. LXVIII.* Cold Spring Harbor Laboratory Press, New York, (in press).

Jordan, T., Schornack, S., and Lahaye, T. 2002. Alternative splicing of transcripts encoding Toll-like plant resistance proteins—What's the functional relevance to innate immunity? *Trends Plant Sci.* **7:** 392–398.

Kazan, K. 2003. Alternative splicing and proteome diversity in plants: The tip of the iceberg has just emerged. *Trends Plant Sci.* **8:** 468–471.

Kessler, M.M., Zeng, Q., Hogan, S., Cook, R., Morales, A.J., and Cottarel, G. 2003. Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome. *Genome Res.* **13:** 264–271.

Kong, J., Gong, J.M., Zhang, Z.G., Zhang, J.S., and Chen, S.Y. 2003. A new AOX homologous gene OsIM1 from rice (*Oryza sativa L.*) with an alternative splicing mechanism under salt stress. *Theor. Appl. Genet.*

Lin, X., Kaul, S., Rounsley, S., Shea, T.P., Benito, M.I., Town, C.D., Fujii, C.Y., Mason, T., Bowman, C.L., Barnstead, M., et al. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402:** 761–768.

Mayer, K., Schuller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Dusterhoft, A., Stiekema, W., Entian, K.D., Terryn, N., et al. 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* **402:** 769–777.

Mott, R. 1997. EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13:** 477–478.

Roest Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., et al. 2000. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat. Genet.* **25:** 235–238.

Salanoubat, M., Lemcke, K., Rieger, M., Ansorge, W., Unseld, M., Fartmann, B., Valle, G., Blocker, H., Perez-Alonso, M., Obermaier, B., et. al. 2000. Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature* **408:** 820–822.

Salse, J., Piegu, B., Cooke, R., and Delseny, M. 2002. Synteny between *Arabidopsis thaliana* and rice at the genome level: A tool to identify conservation in the ongoing rice genome sequencing project. *Nucleic Acids Res.* **30:** 2316–2328.

Schoof, H. and Karlowski, W.M. 2003. Comparison of rice and *Arabidopsis* annotation. *Curr. Opin. Plant Biol.* **6:** 106–112.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., et al. 2002a. Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* **296:** 141–145.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., et al. 2002b. RIKEN *Arabidopsis* full-length cDNA database. *Trends Plant Sci.* **7:** 562–563.

Sun, G., Dilcher, D.L., Zheng, S., and Zhou, Z. 1998. In search of the first flower: A jurassic angiosperm, archaefructus, from Northeast China. *Science* **282:** 1692–1695.

Tabata, S., Kaneko, T., Nakamura, Y., Kotani, H., Kato, T., Asamizu, E., Miyajima, N., Sasamoto, S., Kimura, T., Hosouchi, T., et al. 2000. Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature* **408:** 823–826.

Theologis, A., Ecker, J.R., Curtis, J.P., Federspeil, N.A., Kaul, S., and Venter, C. 2000. Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature* **408:** 816–820.

Vandepoele, K., Simillion, C., and Van de Peer, Y. 2002. Detecting the undetectable: Uncovering duplicated segments in *Arabidopsis* by comparison with rice. *Trends Genet.* **18:** 606–608.

Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C., Nguyen, M., et al. 2003. Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302:** 842–846.

Yang, Y.W., Lai, K.N., Tai, P.Y., and Li, W.H. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. *J. Mol. Evol.* **48:** 597–604.

WEB SITE REFERENCES

http://www.genoscope.cns.fr/; gives direct access to the browser.

http://www.invitrogen.com/content/sfs/manuals/18248.pdf; contains protocol used for libraries construction.

http://www.genoscope.cns.fr/Arabidopsis; permits access to files listed in the text, with links to the browser.

http://rgp.dna.affrc.go.jp/IRGSP/; The International Rice Genome Sequencing Project home page.

Received May 7, 2003; accepted in revised form December 27, 2003.