Comparing low coverage random shotgun sequence data from *Brassica oleracea* and *Oryza sativa* genome sequence for their ability to add to the annotation of *Arabidopsis thaliana*

Manpreet S. Katari, Vivekanand Balija, Richard K. Wilson, et al.

Genome Res. 2005 15: 496-504
Access the most recent version at doi:10.1101/gr.3239105

Letter

Comparing low coverage random shotgun sequence data from *Brassica oleracea* and *Oryza sativa* genome sequence for their ability to add to the annotation of *Arabidopsis thaliana*

Manpreet S. Katari,[1,2] Vivekanand Balija,[1] Richard K. Wilson,[3] Robert A. Martienssen,[1] and W. Richard McCombie[1,4]

[1]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA;* [2]*Graduate Program in Genetics, State University of New York at Stony Brook, Stony Brook, New York 11794, USA;* [3]*The Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri 63108, USA*

Since the completion of the *Arabidopsis thaliana* genome sequence, there is an ongoing effort to annotate the genome as accurately as possible. Comparing genome sequences of related species complements the current annotation strategies by identifying genes and improving gene structure. A total of 595,321 *Brassica oleracea* shotgun reads were sequenced by TIGR (The Institute for Genome Research) and the collaboration of Washington University and Cold Spring Harbor. Vicogenta (a genome viewer based on GMOD and GBrowse) was created to view the current annotation and sequence alignments for *Arabidopsis*. *Brassica* reads were compared with the *Arabidopsis* genome and proteome databases using BLAST. Hypothetical genes and conserved unannotated regions on the short arm of chromosome 4 from *Arabidopsis* were experimentally verified using RT–PCR. We were able to improve the *Arabidopsis* annotation by identifying 25 genes that were missed, and confirming expression of 43 hypothetical genes in *Arabidopsis*. We were also able to detect conservation in genes whose transcription is normally suppressed due to methylation. We also examined how useful the *O. sativa* genome and ESTs from other species are, compared with *Brassica*, in improving the *Arabidopsis* annotation.

[Supplemental material is available online at www.genome.org. Vicogenta is available at http://mccombielab.cshl.org/katari/vicogenta.]

Arabidopsis thaliana is one of the most widely used model organisms for plant molecular biology. Reasons for its popularity include its short life cycle, small size, small genome size—125 Mb (Meyerowitz and Somerville 1994) and its low repetitive content (10%). These make *Arabidopsis* an attractive organism for genome analysis and comparative plant genomics (Meinke et al. 1998; Martienssen and McCombie 2001). Shortly after the completion of the *Arabidopsis* genome sequence (*Arabidopsis* Genome Initiative 2000), it was proposed that the function of all genes in *Arabidopsis* be determined by 2010 (Chory et al. 2000; Somerville and Dangl 2000). To accomplish this task, we need a complete set of genes and their correct gene structure.

The current annotation of the *Arabidopsis thaliana* genome is composed of predictions from gene-finding programs, alignment of expressed sequences; ESTs (Expressed Sequenced Tags), and full-length cDNA clones (Haas et al. 2002; Seki et al. 2002). Recently, high-density genomic tiling arrays were also used to improve the *Arabidopsis* annotation (Yamada et al. 2003). All of these methods have been very useful, but each has their limitations. We believe that an appropriate application of comparative genomics would complement them and help to provide a more accurate annotation.

Several studies have tested the accuracy of gene-prediction programs in *Arabidopsis* (Pavy et al. 1999) and plants in general (Pertea and Salzberg 2002), and found that one of five exons are predicted incorrectly and <50% of the gene models are completely correct. In addition, Macintosh et al. (2001) demonstrated the limitation of the current annotation methods for *Arabidopsis* by identifying putative-coding and noncoding genes that were not previously annotated (MacIntosh et al. 2001). Many gene-predicting algorithms miss most noncoding genes (ncRNA), because their training set lacks ncRNA, and many of the features that pertain to protein-coding genes are not useful for predicting noncoding genes. For example, ncRNA are not translated and, thus, codon usage and related information would not be helpful in identifying such genes.

New gene-prediction algorithms, such as Genomescan (Yeh et al. 2001) and Twinscan (Korf et al. 2001), incorporate homology information to improve gene prediction. Both algorithms use Genscan (Burge and Karlin 1997) to identify protein-coding regions in the DNA sequence, but apply homology information differently. In our study, we analyzed results from Twinscan and show Twinscan can be used to improve the *Arabidopsis* genome annotation if appropriate comparative data are available.

ESTs and full-length cDNA clones are very useful in identifying genes and accurately annotating their structure (Haas et al. 2002; Seki et al. 2002; Castelli et al. 2004). In fact, full-length cDNAs are the "gold standard" of experimental data to support gene models in annotation. However, not all genes are represented in the cDNA libraries.

Recently, a high-density oligonucleotide array has been used to make improvements to the *Arabidopsis* annotation (Ya-

[4]Corresponding author.
E-mail mccombie@cshl.edu; fax (516) 422-4109.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3239105.

15:496–504 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

mada et al. 2003). One limitation of this method is that RNA samples from every possible condition would be required to obtain a complete set of potential genes. Another disadvantage is that the gene structure is difficult to resolve using expression data. For example, if transcription is detected from two neighboring regions on a chromosome, it is difficult to determine whether the regions are from the same gene or two different genes. The same limitation applies to sequence comparisons, because two neighboring conserved sequences may not be from the same gene. However, sequence comparison does not rely on expression data and, thus, can provide candidate regions, which can be verified using sensitive techniques such as RT–PCR. Another limitation of sequence comparison is that some genes may diverge very fast, making them difficult to detect using current comparative algorithms. Combining results from all methods discussed above will result in a more complete set of genes, which can be targeted for full-length sequencing.

Complete genome sequences are available for several multicellular model organisms and their close relatives as follows: *Caenorhabditis elegans* (*C. elegans* Sequencing Consortium 1998) and *C. briggseae* (Stein et al. 2003), *Drosophila melanogaster* (Adams et al. 2000) and *Anopheles gambiae* (Holt et al. 2002), and *Homo sapiens* (Lander et al. 2001; Venter et al. 2001) and *Mus musculus* (Waterston et al. 2002). The rationale is that conserved regions have been maintained by purifying selection (Kimura 1983), so that sequences from a close relative can be used to identify biologically significant regions in the model organism. *Brassica oleracea* and *Arabidopsis thaliana* belong to the same family, *Brassicaceae*, also known as *Cruciferea*, or the mustard family. *Brassica*'s putative haploid genome size is ~600–660 Mb (O'Neill and Bancroft 2000; Paterson et al. 2001) Previous studies of proteins and mitochondrial DNA have predicted that *Arabidopsis* and *Brassica* diverged 16 to 19 Mya (million years ago) (Yang et al. 1999; Koch et al. 2000). While a "perfect" organism for comparative genomics may not exist, *Brassica* seems well placed due to its relative closeness to *Arabidopsis* (Quiros et al. 2001). Also important, *Brassica oleracea* is agronomically valuable. Some common crops that belong to the species *Brassica oleracea* are cauliflower, broccoli, kale, turnip, and cabbage (Paterson et al. 2001). The previous examples also illustrate the diversity in phenotype of *Brassica oleracea*. We also carried out a comparative analysis of *Arabidopsis* using the more distantly, but more completely sequenced genome of the monocot *Oryza sativa* (Goff et al. 2002; Yu et al. 2002).

In our analysis, we used the March, 2003 and February, 2004 version of the *Arabidopsis* annotation by MIPS (Munich Information Center for Protein Sequence) (Schoof et al. 2004). Of the 26,639 predicted genes in the February, 2004 version, 3898 (15%) are annotated as hypothetical. For the purposes of this analysis, they are predicted by computational methods rather than homology to known proteins and have, at most, one EST match. One of the questions we ask is whether hypothetical genes that are conserved in *Brassica* are more likely to be expressed compared with hypothetical genes that are not conserved.

Our computational study was performed on the entire genome; however, we only experimentally verified genes from the short arm of chromosome 4 of *Arabidopsis thaliana* using the March, 2003 version of the annotation. We chose this region of chromosome 4 for several reasons. We had sequenced it (Mayer et al. 1999; CSHL/WashU/PEB 2000) and were familiar with it and had clone resources available. In addition, the presence of the knob region provided a sample of both euchromatin and heterochromatin. Lastly, these analyses were synergistic with our work with genomic tiling microarrays in this region (Lippman et al. 2004). This region is 3 Mb long, roughly 1/40th of the entire genome, and contains 599 predicted genes. The heterochromatic knob contains many repeats and transposons, which are heavily methylated. Compared with the rest of the genome, this region contains many fewer genes that are expressed. We examined the utility of *Brassica* sequences in identifying functional elements in the knob region.

As a result two of the groups, TIGR (The Institute for Genome Research) and the CSHL/WU (Cold Spring Harbor Laboratory/Washington University) consortium sequenced 595,321 random *Brassica oleracea* shotgun reads. The sequences were aligned against the *Arabidopsis* genome sequence using BLAST, and the results were compared with the annotation. In our comparison, we had three main goals as follows: (1) identify missed genes, (2) identify incorrect gene structure, and (3) determine whether conserved hypothetical genes are more likely to be expressed than other hypothetical genes.

Results

A total of 595,321 *Brassica oleracea* shotgun reads, 415,093 sequenced by TIGR and 180,228 by CSHL and Washington U., were downloaded from GenBank and analyzed. The *Arabidopsis* genome sequence, genome annotation, and protein sequences were downloaded from MIPS (Munich Information Center for Protein Sequence) (ftp://ftpmips.gsf.de/cress) versions March, 2003 and February, 2004. The *Brassica* reads were subjected to an initial screening by first comparing them to known *Arabidopsis* repeats, transposable elements, and organelle DNA sequences. BLASTN (Altschul et al. 1990, 1997) and TBLASTX were used to align the *Brassica* reads against the *Arabidopsis* nucleotide sequences, and BLASTX was used to align the *Brassica* reads against the *Arabidopsis* protein sequences. Alignments with an E-value <1e-10 were considered as significant matches.

Brassica reads were categorized according to their top BLAST hit (see Fig. 1). Nearly 30% of the *Brassica* reads contain repeat elements or organelle DNA, and 45% of the reads do not have significant matches to the *Arabidopsis* sequences. The reads are most likely from intergenic or intronic regions, where the level of sequence conservation is much less. The *Arabidopsis* protein database is the translation of all protein-coding genes in the annotation; thus, the reads with only Nucleotide matches (7% of the *Brassica* reads) are aligning to unannotated regions. These may represent undetected genes or exons, nonprotein-coding transcriptional units, regulatory regions, or regions conserved for unknown reasons. The level of conservation (E-value < 1e-10) suggests that these unannotated regions are biologically significant.

The average length of the sequencing reads is 677 bp. If we do not consider the reads that match to the organelle DNA, and assuming the size of the *Brassica* genome is 600 Mb, we estimate coverage of 0.60× of the *Brassica oleracea* genome. According to the Lander-Waterman model (Lander and Waterman 1988) the reads cover 45% of the genome. Despite the low coverage, the *Brassica* reads match nearly 74% of the predicted *Arabidopsis* proteome; top BLASTX hits of the *Brassica* reads against the *Arabidopsis* proteome were considered as a match. This is likely due to the large percentage of *Arabidopsis* genes that are duplicated and the possible triplication of the *Brassica* diploid genome (Lan et al. 2000; O'Neill and Bancroft 2000).

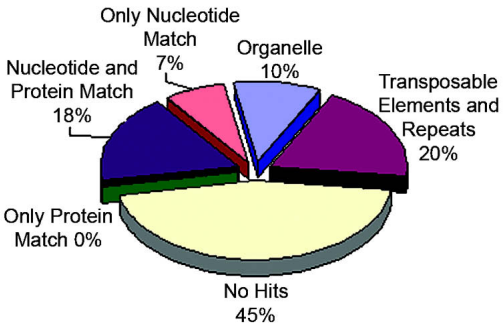

Figure 1. BLASTN, TBLASTX, and BLASTX were used to align 595,321 *Brassica oleracea* reads against databases of *Arabidopsis thaliana* chloroplast and mitochondrial DNA, known repeats and transposable elements, and finally, the *Arabidopsis thaliana* genome and protein sequences. Top BLAST hit was used to classify the *Brassica oleracea* reads. The reads had to match with an e-value < 1e-10 to be considered as a significant match. The protein database consists of translations of all predicted protein-coding genes in the genome annotation. Therefore, the 7% of the *Brassica* reads that match only the genome, represent either noncoding RNA or genes that have not yet been annotated.

Comparatively, we used more *O. sativa* sequence in our analysis. A total of 3657 BACs were downloaded, of which 1312 are finished (high-quality) sequences. The total number of bases in these overlapping BACs is 510,899,921, and the *O. sativa* genome is estimated to be 450 Mb.

We focused our experiments on the short arm of chromosome 4 of *Arabidopsis thaliana*. A total of 116 genes of 599 (19%) are predicted as hypothetical genes, which is slightly higher compared with the entire proteome, where 17% are annotated as hypothetical. In the February, 2004 version of the annotation, there are 600 predicted genes in this region, of which 108 are annotated as hypothetical. Several of the genes that are no longer annotated as hypothetical in the February, 2004 version were also confirmed by our experimental results.

One of the limitations of our analysis is that we have only used one source of RNA, i.e., whole-plant, above-ground tissue of wild-type *Arabidopsis thaliana*. Many genes for which we do not detect transcription are most likely expressed in different developmental stages or environmental conditions. Therefore, our results provide a minimum number of genes that are expressed. Another limitation is that we do not have the complete genome sequence of *Brassica*. Therefore, the number of genes and regions that are conserved between *Brassica* and *Arabidopsis* is higher than we observe in our analysis.

Conserved vs. nonconserved hypothetical genes

Hypothetical genes are predicted by at least two gene-prediction algorithms and match, at most, one EST. Primer3 was used to design Primers for hypothetical genes and, if possible, in areas where *Brassica* conservation flanked a splice site. One caveat is that primers are being designed using annotation that may not be correct. Thus, a result that does not contain a transcript may not necessarily mean the gene is not expressed; one of the primers could be in an intron, in which case, transcription would not be detected. Another possibility is that the gene is expressed in a different condition or developmental stage than the one tested.

We wanted to check whether hypothetical genes conserved between *Brassica* and *Arabidopsis* were more likely to be expressed compared with nonconserved hypothetical genes. Of the 93 conserved hypothetical genes that we tested, from the March, 2003 version of the annotation, we detected expression for 42 (46%) hypothetical genes, and of the 17 nonconserved hypothetical genes we tested, we only detected expression from 1 (6%) gene (see Table 1; gel pictures are provided in the Supplemental data). We calculated the Fisher's exact test to determine the statistical significance of our results using a DOS executable program Fisher.exe (Zhang et al. 1998). We obtained a *P*-value of 0.00077 for our data. This shows that there is significant correlation between hypothetical genes in *Arabidopsis* that are conserved in *Brassica*, and their likelihood of expression.

Yamada et al. (2003) provided a list of *Arabidopsis* genes that were detected by their microarray experiments. They detected 44 of the 110 hypothetical genes that we tested, 26 of which were also detected by our RT–PCR experiments. A total of 41 of the 44 are conserved in *Brassica*, and 27 of the 44 are conserved in *O. sativa* (see Table 1). These results are very similar to our RT–PCR experiments, and provide additional support to our observation that hypothetical genes conserved in *Brassica* are more likely to be expressed than hypothetical genes that are not conserved.

In the February, 2004 version of the *Arabidopsis* annotation, several hypothetical genes were no longer annotated as hypothetical. A total of 16 of the 110 hypothetical genes we tested were no longer annotated as hypothetical. All 16 are conserved in *Brassica*, of which 13 were detected by our analysis. This suggests that despite the considerable improvements of the *Arabidopsis* annotation in the past year, comparing *Brassica* sequences can improve the annotation even more.

We also wanted to know how useful *O. sativa* is in identifying genes in *Arabidopsis*. Of the 63 hypothetical genes that are conserved in *O. sativa*, we were able to detect expression for 34 (54%), and of the 47 that are not conserved in *O. sativa*, we were able to detect expression from nine (19%) (Table 1), indicating enrichment ($P = 0.00017$). Hypothetical genes that are conserved in *O. sativa* are more likely to be expressed than those conserved in *Brassica*, however, despite comparing to an almost complete *O. sativa* sequence, there are several expressed genes that show conservation with *Brassica*, but not *O. sativa*. *O. sativa* is useful in improving the *Arabidopsis* annotation; however, its conservation does not cover all biologically significant regions in *Arabidopsis*. Thus, *Brassica* is more useful than *O. sativa* in determining the complete set of *Arabidopsis* genes.

Table 1. Analysis of transcripts from hypothetical genes

	Total hypothetical genes	*Brassica* conservation		*O. Sativa* conservation		Comparative ESTs (1e-10)	
		Conserved genes	Nonconserved	Conserved genes	Nonconserved	Match	Do not match
Tested	110	92 (30)	18 (0)	63 (24)	47 (6)	79 (28)	31 (2)
Yield PCR Product	43	42 (22)	1 (0)	34 (17)	9 (5)	40 (20)	3 (2)
Detected by Yamada et al. (2003)	44	41 (18)	3 (0)	27 (13)	17 (5)	35 (18)	9 (0)

Finally, we also looked at how useful ESTs from species, other than *Arabidopsis*, are in determining whether a hypothetical gene is likely to be expressed and compare that with what we have learned from hypothetical genes conserved in *Brassica*. We queried the CDS sequence from the 110 hypothetical genes using TBLASTX against the est_others database, which, in August, 2004, contained over 13 million sequences. We filtered out all hits to *Arabidopsis* ESTs, and found 79 of the 110 hypothetical genes (74%) have a match to an EST from a species other than *Arabidopsis*, with an e-value of 1e-10 or better. There are 18 hypothetical genes that are conserved in *Brassica*, but do not match any ESTs from different species, whereas only five hypothetical genes have matches to ESTs from other species and are not conserved in *Brassica*. The last five may be conserved in *Brassica*, but due to the low sequence coverage, the region may be missed.

ESTs from other species are useful in identifying hypothetical genes that are likely to be expressed, because 40/43 hypothetical genes that we detected have a match to other species. However, the ESTs did not provide any more information about the hypothetical genes than what we already knew from the sequence conservation in *Brassica*. In fact, there are 13 more hypothetical genes that are conserved in *Brassica*, which are likely to be expressed in *Arabidopsis*.

Correcting gene structure of hypothetical genes

A total of 30 of the 43 PCR products, discussed above, resulted in a spliced transcript; 22 of the spliced products match the annotation correctly, and eight show a different gene model. All eight of these genes are conserved in *Brassica*, and in most cases, the BLASTN alignment of the read against the *Arabidopsis* genome does not contradict with our experimental results. The discrepancies include different exon borders, presence of an additional exon, and deletion of an exon. Three of the eight genes were correctly annotated in the new version of the TIGR annotation v4.0. The remaining five were incorrect, possibly due to the lack of EST matches to the hypothetical gene, for example, gene At4g02030 (see Fig. 2). Twinscan correctly predicted five of the eight genes. The remaining three genes that contain errors lacked matches to *Brassica* reads in the regions containing the error. This suggests that if we obtain enough *Brassica* reads to cover the entire length of the gene, Twinscan predictions would be able to predict genes much more accurately. In the February, 2004 version of the MIPS annotation, only one of the eight genes (At4g00420) has been corrected. This suggests that despite the improvements in genome annotation in the past year, there are still many genes that are incorrectly annotated.

A total of 13 PCR products did not result in spliced transcripts, due to the fact that 10 are predicted to be one-exon genes. There is only one case where the unspliced transcript is different from the annotation, i.e., At4g00640. There are no ESTs in the region, and so it is difficult to conclude whether this is an alternative splicing event, or an incorrect gene structure.

Looking at conserved unannotated regions

There are regions of conservation where there is no annotated gene. In this study, we are not concerned about conservation of gene order, so we don't have to limit ourselves with only reads that appear to be orthologs (top match) to the *Arabidopsis* genome. Therefore, we looked at all significant BLASTN and TBLASTX matches (E-value < 1e-10) from all reads, and compared it with the annotations provided with the genomic sequence. Regions in the *Arabidopsis* sequence that are conserved in *Brassica*, but do not contain any annotation, will be referred to as CURs (Conserved Unannotated Regions). To screen out regulatory regions or alternative exons, all CURs within 500 bp of a predicted gene were removed. Remaining CURs that were within 2 kb from each other were grouped as putative gene models and will be referred to as Cluster of Conserved, Unannotated Regions (CCURs). Primer3 was used to design primers in a similar fashion as described for hypothetical genes with the individual CURs serving as putative exons.

A total of 9040 CCURs were found throughout the *Arabidopsis* genome, with an average size of 717 bp; 266 of them reside in the short arm of chromosome 4. A total of 106 of these CCURs were not considered, because they matched known repeats and 48 were too small (average size 168 bp) to design primers. The small CCURs could be due to lack of *Brassica* reads to extend the conservation, or they may be small genes that require other methods of verification.

A total of 112 CCURs from the short arm of chromosome 4 were tested, and we detected 25 transcripts using RT–PCR (seven spliced and 18 nonspliced) (see Table 2 for details; gel pictures are

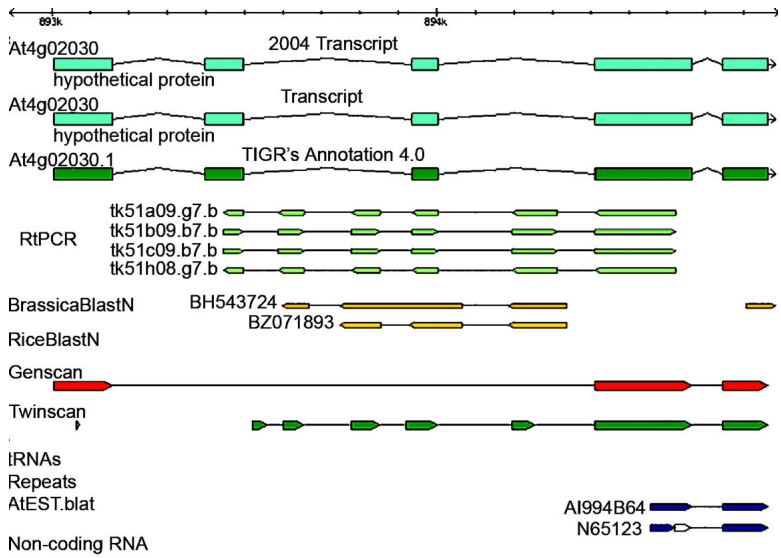

Figure 2. This figure shows a mistake in the *Arabidopsis* annotation. The image was created by Gbrowse (Stein et al. 2002). The *x*-axis on the *top* is the *Arabidopsis* genome coordinates. The label of each track is listed on the *left* side, in bold, over the figure. The track "Transcript" corresponds to the March, 2003 annotation. The At4.0 annotation does not contain the correct annotation, possibly because there are no ESTs (track AtEST.blat) in the region where the discrepancy occurs. The figure also demonstrated that the BLASTN alignment of *Brassica* sequences is often enough to tell whether an exon is present or missing. Twinscan was able to predict some of the missing exons; however, it failed to annotate them all correctly.

Table 2. Analysis of transcripts from CCURs

	Total PCR products	Match A.T. 4.0	Match EST	Match *O. sativa* sequences	Average CCUR size
Spliced Transcripts	7	6 (86%)	6 (86%)	5 (71%)	1003 bp
Nonspliced Transcripts	18	5 (28%)	5 (28%)	11 (61%)	518 bp
Total	25	11 (44%)	11 (44%)	16 (64%)	654 bp

provided in Supplemental data). Six of the seven spliced transcripts are annotated by TIGR's new annotation and have matches to ESTs and *O. sativa* (see Table 2). In one case, cluster5663, the PCR product, and the new annotation have different gene models (data not shown). The AT 4.0 annotation correlates well with the BLASTN alignments, but there is no EST evidence to support this model.

The 18 nonspliced transcripts are from CCURs that are much smaller (518 bp) compared with the CCURs that resulted in spliced transcripts (1003 bp). The small CCURs could be from genes containing one exon, or from one exon of a multiexon gene. Only five of these CCURs have matches to a new At 4.0 annotation, and only five have matches to ESTs. However, 11/18 (61%) have matches to *O. sativa* sequences, suggesting that the CCURs are biologically significant. For the past few years, many small ncRNA (noncoding RNAs) are being discovered in both animals and plants, which play a very important role in development (Bernstein et al. 2003; Carrington and Ambros 2003; Hunter and Poethig 2003). Considering the level of conservation of the short regions between both *Brassica* and *O. sativa*, it is quite possible that these CCURs represent ncRNAs or small protein-coding genes (<100 amino acids). Further experiments are required to determine whether this is true.

In the February, 2004 MIPS annotation of *Arabidopsis*, there is one gene annotation that overlaps a CCUR. The gene structure of At4g03260 is extended in the 5′ end to include a portion of Cluster5715. One primer lies in the gene structure, and the other is upstream of the annotation, which is one possible reason why the CCUR was not experimentally detected by our analysis.

Heterochromatic knob

The heterochromatic knob on chromosome 4 of *Arabidopsis* is located in the YAC CIC8B1 (CSHL/WashU/PEB 2000), which is roughly from coordinate 1,600,000 bp to 2,330,000 bp on chromosome 4, and contains genes At4g03590–At4g04620 (Fransz et al. 2000). The complete sequence of this region represents one of the first complete sequences of a heterochromatic region. The knob region is heavily methylated, so most genes in the region are not expressed (CSHL/WashU/PEB 2000). In the MIPS 2003 annotation, there are 34 hypothetical genes in the knob region, and we tested 31. A recent study carefully reannotated the knob region and identified 19 of these 31 hypothetical genes to be either DNA transposons or retrotransposons (Lippman et al. 2004). The study identified 15 total hypothetical genes in the knob region; however, two are annotated as putative in the MIPS 2003 annotation and another is not even annotated. Thus, for the analysis of hypothetical genes in the knob region, we will only discuss the 12 found in MIPS (see Supplemental data Table 3). Only one of the hypothetical genes in this region has a match to ESTs, At4g04330. However, using sequence-specific primers, we were able to detect expression in six CCURs and six hypothetical genes in this region.

Of the 12 hypothetical genes in the knob region, 11 are conserved in *Brassica* and six are conserved in *O. sativa*. We detected expression in six (50%) genes, five of which are conserved in both *Brassica* and *O. sativa* (see Table 3). Yamada et al. (2003) detected five hypothetical genes, of which all are conserved in *Brassica* and three are conserved in *O. sativa* (see Table 3).

In a previous study of the knob region (Gendrel et al. 2002), 6/12 hypothetical genes were tested for expression, all of which are conserved in *Brassica* and two of which are conserved in *O. sativa* (see Table 3). The same two genes are also expressed in wild-type *Arabidopsis* seedlings (Gendrel et al. 2002). However, expression of two more hypothetical genes was detected in *ddm12* mutant seedlings. *DDM1* is required for methylation in *Arabidopsis* and is responsible for gene silencing in methylated regions. Thus, two of the hypothetical genes are under epigenetic control, both of which are conserved in *Brassica*, but neither of which are conserved in *O. sativa*.

There are a total of 24 CCURs predicted in this region. Five match a gene annotated in TIGR's annotation version 4.0 and two match ESTs. All six of the CCURs that we detected using our sequence-specific primers were not spliced. A total of 12 of the 24 CCURs are conserved in *O. sativa*, which provides further evidence that many of these regions are biologically significant (see Supplemental data).

Discussion

Arabidopsis thaliana and *Brassica oleracea* are in the same family, *Cruciferae*, and previous studies have showed an average of 87% conservation in the coding region. Our results show that genes without experimental evidence, hypothetical genes, are more likely to be expressed if they are conserved in *Brassica* sequences compared with hypothetical genes that are not conserved. We also observe a correlation between hypothetical genes that are conserved with *Brassica* and hypothetical genes that are expressed by analyzing data from Yamada et al (2003). A total of 26 of the 44 hypothetical genes detected by microarray are shared among the 43 hypothetical genes verified using RT–PCR, suggesting that the two methods are complementary and equally useful for gene discovery. If we generalize our results to include the entire *Arabidopsis* genome, we can expect to detect expression from ~2517 hypothetical genes. This is a minimal number, because we are only considering experiments from limited sources of RNA for transcriptional verification.

In addition to identifying hypothetical genes, we were also able to identify incorrect gene structure for 21% (9/43) hypothetical genes, suggesting that there are still many hypothetical

Table 3. Conservation of hypothetical genes in the heterochromatic knob

	Wild-type whole plant	Wild-type seedlings	*ddm1* mutant seedlings	Yamada et al. (2003)
Total Genes	12	6	6	12
Detected Expression	6	2	4	5
Conserved with *Brassica*	11 (6)	6 (2)	6 (4)	11 (5)
Conserved with *O. sativa*	6 (5)	2 (2)	2 (2)	3 (3)

genes that are incorrectly annotated. We also found that gene structures of many hypothetical genes can be improved using Twinscan (see Supplemental data). However, in areas with no conservation, the algorithm makes the same mistakes as Genscan. A better coverage of the *Brassica* genome can make Twinscan a very powerful tool in annotating the *Arabidopsis* genome.

Arabidopsis thaliana and *Oryza sativa* are the first completely sequenced plant genomes. There have been several efforts to improve the *Arabidopsis* annotation using the *O. sativa* sequence most recently by Castelli et al. (2004). Our studies show that the *O. sativa* genome is useful in identifying genes in *Arabidopsis*, however it often misses some areas that code for genes, mainly because the sequences have diverged considerably. However, most genes that are conserved in *O. sativa* are conserved in *Brassica* 60/63 (95%). The remaining 5% could be due to *Brassica*'s low coverage.

Similarly, ESTs from other species are also useful in identifying hypothetical genes that are likely to be expressed; however, the ESTs did not provide any more information about the hypothetical genes than what we already knew from the sequence conservation in *Brassica*, which has been sequenced at a fairly low level. We expect that a deeper level of sequencing of the *Brassica* genome will be more informative, with respect to the *Arabidopsis* annotation, than sequencing ESTs from other plant species.

The heterochromatic knob region in the short arm of chromosome 4 is mostly transcriptionally silent (CSHL/WashU/PEB 2000). However, all of the hypothetical genes from this region that are expressed are conserved in *Brassica* and *O. sativa*, comparable to the rest of the short arm of chromosome 4. In addition, two hypothetical genes were only expressed in *ddm1* mutants, both of which are conserved in *Brassica*, but not *O. sativa*. This suggests that sequence conservation in *Brassica* can help identify genes under epigenetic control. Only one of the hypothetical genes in the heterochromatic region had matches to ESTs, indicating that comparative genomics is a more powerful way to identify epigenetically regulated genes. Considering their relatively low level of conservation, we cannot exclude the possibility that these are novel transposons, as almost all are the known targets of DDM1.

We were also able to detect transcripts in conserved regions that do not contain annotation (CCURs). Most CCURs with spliced transcripts also match ESTs and are present in version 4.0 of the *Arabidopsis* annotation released by TIGR. However, nonspliced transcripts from CCURs are much smaller, and the majority don't match ESTs, and subsequently, are not present in the TIGR's 4.0 annotation. The small size of the CCURs may be due to the lack of ample *Brassica* reads needed to extend the CCUR, or simply because they are smaller genes. cDNA libraries are often size-selected before EST sequencing, thus making it difficult to find corresponding ESTs for smaller genes. In addition, as previously shown by MacIntosh et al. (2001), gene-prediction algorithms tend to miss genes that code for <100 amino acids. The level of sequence conservation of the smaller CCURs in *Brassica* and the fact that the majority of the smaller CCURs are also conserved with *O. sativa*, suggests that these CCURs are biologically significant regions in *Arabidopsis*.

Our results suggest an increase of 850 genes in the *Arabidopsis* transcriptome. This also is a minimal number, because presumably more *Brassica* sequences would create more CCURs. Yamada et al. (2003) reported expression in 2000 (23%) of their 9043 IGR (Inter Genic Regions) and 519 (54%) of their 953 NAE (Not Annotated but Expressed) genes. Their total number of unannotated regions is also nearly 1000 more than the number of CCURS, which may be due to the low coverage of the *Brassica* genome. One possible reason why they detected many more transcripts in the unannotated genes is that they used multiple tissue samples.

In the past year, several improvements have been made in *Arabidopsis* annotation by sequencing more ESTs, full-length cDNA clones, and using genomic microarrays. However, the majority of hypothetical genes from which we detected transcripts are still annotated as hypothetical. In addition, nearly all of the gene structures that we have found to be incorrect have not changed. None of the detected CCUR transcripts are annotated as genes in the February, 2004 version. Our study shows that comparative information from *Brassica oleracea* sequences can help fill the gap and improve the current annotation considerably.

Comparative genomics, EST sequencing and analysis, full-length cDNA sequencing, gene-prediction algorithms, and genome tiling arrays are all useful for improving the *Arabidopsis* annotation, and they complement each other very well. The complementarity is demonstrated by our analysis. Comparative genomics complements gene-prediction algorithms, such as Twinscan, by providing sequence information that enables the algorithms to perform more accurately. Comparative genomics complements expression data, such as ESTs and genome tiling arrays, by providing targets missed by expression analysis. These targets can be detected using more sensitive methods, such as RT–PCR.

The ideal pipeline for genome annotation includes all methods. First, use gene-prediction algorithms with sequence-conservation information to find the majority of the genes. Second, use expression data from genome tiling microarrays and EST sequencing and sequence alignments from comparative studies to identify genes that were missed by gene prediction. EST sequencing alone will not detect transcription of many genes, they require directed methods such as RT–PCR. RT–PCR can use available information, such as sequence conservation, to design the proper primers. This is followed by full-length cDNA sequencing of all predicted genes using RNA from many different sources. Finally, use tools such as PASA to identify different splice forms (Haas et al. 2003).

The mission of the 2010 project is to determine the function of all plant genes in the genome. One of the plans to achieve this goal is "Survey genomic sequencing, and deep EST sampling from phylogenetic node species." (Somerville and Dangl 2000) We have shown that *Brassica oleracea* is a crucial species to compare for improving *Arabidopsis* annotation. These sequences have not only helped us identify and correct gene structures in *Arabidopsis*, but others have also used the sequences to identify regulatory elements (Colinas et al. 2002). A current limitation is the amount of *Brassica* sequence available. The Lander-Waterman model suggests that a 3× coverage of shotgun reads will cover 95% of the genome. This would be helpful in identifying a more complete set of genes that can be used to achieve the goals of the 2010 project.

Methods

Sequencing of the *Brassica* reads

Brassica oleracea genomic DNA from doubled haploid strains (T. Osborn, University of Wisconsin) was nebulized and the 3–5-Kb

fractions were isolated from the sheared DNA. The 3–5-Kb fragments were cloned into pBluescript or pUC19 and plated. These double-stranded subclones were then used to initiate overnight cultures using the QPix automated colony picker, which can inoculate 96-well growth plates for overnight growth. Cultures grown in the 96 growth boxes were archived using the Biomek FX automated platform (Beckman Coulter). Plasmid DNA was then isolated from these cultures using a modified SPRI protocol (Hawkins et al. 1994). Random DNA samples from each plate were picked by the Span-8 pod on the Biomek FX for DNA quantiation and quality control.

The Biomek FX and the TomTec Quadra 354 were used to set-up 7 µL of 1/16 Big Dye Terminator (v. 3) sequencing reactions in a 384-well format. Reaction plates were cycled using the MJ Research Thermal cyclers fitted with 384-well α units. We also performed reaction clean up by ethanol precipitation in a 384-well format with the TomTec. Sequencing products were stored dry in the precipitated state at −20°C until they are required for loading. Resuspension of the sequencing products in water was performed using the TomTec as well as the Biomek FX. Samples were then loaded on the ABI 3700.

Alignment of *Brassica* reads against *Arabidopsis thaliana*

A total of 595,321 *Brassica oleracea* shotgun reads were downloaded from GenBank and were used in the analysis (Entrez query: "*Brassica oleracea* [organism] AND GSS"). The *Arabidopsis thaliana* genome sequence, genome annotation, organelle sequences, and protein sequences were downloaded from MIPS (Munich Information Center for Protein Sequence) (ftp://ftpmips.gsf.de/cress) version v110303 (March 11, 2003) and v110204 (February, 2004) (Schoof et al. 2004). Sequences for transposable elements, known repeats, and *Arabidopsis* ESTs were downloaded from TAIR (Rhee et al. 2003) (http://www.Arabidopsis.org). A total of 174,275 ESTs were used in the analysis.

Only the top BLAST matches were considered when categorizing the *Brassica* shotgun reads (Fig. 1). The reads were also screened against mitochondria, chloroplast, and known repeats nucleotide database, and a transposable element amino acid database. The reads were aligned against the *Arabidopsis thaliana* nucleotide and protein databases from MIPS using all of the three programs, BLASTN, BLASTX, and TBLASTX. Default parameters and an e-value cutoff of 1e-10 were used for all BLAST programs when aligning *Brassica* sequences.

Each individual HSP (High Scoring Pair) from BLAST alignment that were 500 bp away from an annotation is called a CUR (Conserved Unannotated Region). CURs within 2 kb of each other were grouped to form CCURS (Cluster of Conserved Unannotated Regions). BLAST was performed using Amdec facility (http://amdec-bioinfo.cu-genome.org/html/index.html).

RT–PCR and sequencing hypothetical genes and CCURs

Where possible, primers for hypothetical genes were designed in adjacent exons so that the PCR product shows evidence of spliced product. In conserved hypothetical genes, primers were picked from the conserved areas. Similarly, primers for CCURs were designed on two different CURs in the hope to get a spliced transcript. Primer3 (Rozen and Skaletsky 2000) was used to choose the primers, and they were tested using e-PCR (Schuler 1997) against the *Arabidopsis* genome sequence. Default parameters were used for Primer3 (Tm = 60°C and GC = 50%). The settings for e-PCR were M = 1000, N = 2, and W = 7.

RNA was extracted from wild-type, whole-plant, above-ground tissue, above ground, using Trizol. Before use, the RNA was treated with DNAse. This was followed by using Reverse Transcriptase for first-strand cDNA synthesis with the Reverse Primer for hypothetical genes, and a mixture of both primers for the CCURs. The RT step was performed at 44 and 47°C. For the PCR amplification step, the negative control for each primer pair was RNA instead of the RT product as template. Other negative controls used per 96-well plate were as follows: no primers and no Taq DNA polymerase. Reagents from the Qiagen Hot Start TAQ Kit were used for the PCR reactions. Positive controls included Actin (At5g59370), GCR1 (At1g48270), and R18. See Supplemental data for gel images.

Two different methods were used for sequencing. The first was to treat amplified fragments with Exonuclease 1 and Shrimp Alkaline phosphatase, followed by sequencing using Big Dye Terminator chemistry with gene/CCUR-specific primers. Fragments were separated and detected on an ABI 3700. The second strategy was to clone and then sequence the PCR products. The PCR products were cloned into pCR TOPO 2.1 vector (Invitrogen) and transfected into DH10 B cells by electroporation or heat shock. They were plated on LB/AMP/IPTG/X-Gal and then picked and grown in LB medium; −21 M13 Forward and Reverse Universal primers were used to sequence the clones using Big Dye Terminator chemistry.

Querying rice sequences

A total of 3657 *O. sativa* BACs were downloaded from GenBank, of which 1312 are finished. Each BAC was cut into 5-kb segments with 500 bp overlapping. Each segment was aligned to the *Arabidopsis* genome using BLASTN and TBLASTX with default parameters and an e-value cutoff of 1e-5.

Alignment of RT–PCR products, *Arabidopsis* ESTs, and AT 4.0 using BLAT

A total of 174,275 *Arabidopsis* ESTs were downloaded from TAIR's ftp site ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/. The CDS sequences of TIGR's latest version of the *Arabidopsis* annotation were downloaded from TIGR's ftp site (Haas et al. 2003). The sequences of the PCR products were base-called using Phred, and then trimmed for vector sequences. All were aligned to the *Arabidopsis thaliana* genome v110204 from MIPS using BLAT (Kent 2002). Alignments that are >95% identical and are the top match of the Sequence, EST, or CDS were loaded into the GMOD database (Stein et al. 2002).

Acknowledgments

We thank Bruce May, Juana Arroyo, and Zach Lippman for providing *Arabidopsis* tissue and a protocol for RNA extraction, Tom Osborn for providing us with the *Brassica oleracea* BAC and with genomic DNA from doubled haploid strains, and our colleagues at Washington University and TIGR for additional *Brassica* reads. We also thank The AMDeC Bioinformatics Core Facility at the Columbia Genome Center, Columbia University for their use of the server to do our BLAST searches. This work was supported by the National Science Foundation (DBI9813578). Accession numbers for sequences from RT–PCR experiments are provided in online Supplemental data. Our software and data is available upon request.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

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

Bernstein, E., Kim, S.Y., Carmell, M.A., Murchison, E.P., Alcorn, H., Li, M.Z., Mills, A.A., Elledge, S.J., Anderson, K.V., and Hannon, G.J. 2003. Dicer is essential for mouse development. *Nat. Genet.* **35:** 215–217.

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

Carrington, J.C. and Ambros, V. 2003. Role of microRNAs in plant and animal development. *Science* **301:** 336–338.

Castelli, V., Aury, J.M., Jaillon, O., Wincker, P., Clepet, C., Menard, M., Cruaud, C., Quetier, F., Scarpelli, C., Schachter, V., et al. 2004. Whole genome sequence comparisons and "Full-Length" cDNA sequences: A combined approach to evaluate and improve *Arabidopsis* genome annotation. *Genome Res.* **14:** 406–413.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Chory, J., Ecker, J.R., Briggs, S., Caboche, M., Coruzzi, G.M., Cook, D., Dangl, J., Grant, S., Guerinot, M.L., Henikoff, S., et al. 2000. National Science Foundation-Sponsored Workshop Report: "The 2010 Project" functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. *Plant Physiol.* **123:** 423–426.

Colinas, J., Birnbaum, K., and Benfey, P.N. 2002. Using cauliflower to find conserved non-coding regions in *Arabidopsis*. *Plant Physiol.* **129:** 451–454.

CSHL/WashU/PEB. 2000. The complete sequence of a heterochromatic island from a higher eukaryote. *Cell* **100:** 377–386.

Fransz, P.F., Armstrong, S., de Jong, J.H., Parnell, L.D., van Drunen, C., Dean, C., Zabel, P., Bisseling, T., and Jones, G.H. 2000. Integrated cytogenetic map of chromosome arm 4S of *A. thaliana:* Structural organization of heterochromatic knob and centromere region. **100:** 367–376.

Gendrel, A.V., Lippman, Z., Yordan, C., Colot, V., and Martienssen, R.A. 2002. Dependence of heterochromatic histone H3 methylation patterns on the *Arabidopsis* gene DDM1. *Science* **297:** 1871–1873.

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:** research0029.1–research0029.12

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.

Hawkins, T.L., O'Connor-Morin, T., Roy, A., and Santillan, C. 1994. DNA purification and isolation using a solid-phase. *Nucleic Acids Res.* **22:** 4543–4544.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298:** 129–149.

Hunter, C. and Poethig, R.S. 2003. miSSING LINKS: miRNAs and plant development. *Curr. Opin. Genet. Dev.* **13:** 372–378.

Kent, W.J. 2002. BLAT—the BLAST-like alignment tool. *Genome Res.* **12:** 656–664.

Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, New York.

Koch, M.A., Haubold, B., and Mitchell-Olds, T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (*Brassicaceae*). *Mol. Biol. Evol.* **17:** 1483–1498.

Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17:** S140–S148.

Lan, T.H., DelMonte, T.A., Reischmann, K.P., Hyman, J., Kowalski, S.P., McFerson, J., Kresovich, S., and Paterson, A.H. 2000. An EST-enriched comparative map of *Brassica oleracea* and *Arabidopsis thaliana*. *Genome Res.* **10:** 776–788.

Lander, E.S. and Waterman, M.S. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2:** 231–239.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lippman, Z., Gendrel, A.V., Black, M., Vaughn, M.W., Dedhia, N., McCombie, W.R., Lavine, K., Mittal, V., May, B., Kasschau, K.D., et al. 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430:** 471–476.

MacIntosh, G.C., Wilkerson, C., and Green, P.J. 2001. Identification and analysis of *Arabidopsis* expressed sequence tags characteristic of non-coding RNAs. *Plant Physiol.* **127:** 765–776.

Martienssen, R. and McCombie, W.R. 2001. The first plant genome. *Cell* **105:** 571–574.

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.

Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D., and Koornneef, M. 1998. *Arabidopsis thaliana*: A model plant for genome analysis. *Science* **282:** 662, 679–682.

Meyerowitz, E.M. and Somerville, C.R. 1994. *Arabidopsis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

O'Neill, C.M. and Bancroft, I. 2000. Comparative physical mapping of segments of the genome of *Brassica oleracea* var. alboglabra that are homoeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. *Plant J.* **23:** 233–243.

Paterson, A.H., Lan, T., Amasino, R., Osborn, T.C., and Quiros, C. 2001. *Brassica* genomics: A complement to, and early beneficiary of, the *Arabidopsis* sequence. *Genome Biol.* **2:** reviews1011.

Pavy, N., Rombauts, S., Dehais, P., Mathe, C., Ramana, D.V., Leroy, P., and Rouze, P. 1999. Evaluation of gene prediction software using a genomic data set: Application to *Arabidopsis thaliana* sequences. *Bioinformatics* **15:** 887–899.

Pertea, M. and Salzberg, S.L. 2002. Computational gene finding in plants. *Plant. Mol. Biol.* **48:** 39–48.

Quiros, C.F., Grellet, F., Sadowski, J., Suzuki, T., Li, G., and Wroblewski, T. 2001. *Arabidopsis* and *Brassica* comparative genomics: Sequence, structure and gene content in the ABI-Rps2-Ck1 chromosomal segment and related regions. *Genetics* **157:** 1321–1330.

Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., et al. 2003. The *Arabidopsis* Information Resource (TAIR): A model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.* **31:** 224–228.

Rozen, S. and Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132:** 365–386.

Schoof, H., Ernst, R., Nazarov, V., Pfeifer, L., Mewes, H.W., and Mayer, K.F. 2004. MIPS *Arabidopsis thaliana* Database (MAtDB): An integrated biological knowledge resource for plant genomics. *Nucleic Acids Res.* **32:** D373–D376.

Schuler, G.D. 1997. Sequence mapping by electronic PCR. *Genome Res.* **7:** 541–550.

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

Somerville, C. and Dangl, J. 2000. Genomics. Plant biology in 2010. *Science* **290:** 2077–2078.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., et al. 2002. The generic genome browser: A building block for a model organism system database. *Genome Res.* **12:** 1599–1610.

Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. 2003. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol.* **1:** E45.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

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.
Yeh, R.F., Lim, L.P., and Burge, C.B. 2001. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11:** 803–816.
Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. 2002. A draft sequence of the rice genome (*Oryza sativa L. ssp. indica*). **296:** 79–92.
Zhang, J., Rosenberg, H.F., and Nei, M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci.* **95:** 3708–3713.

Web site references

http://mccombielab.cshl.org/katari/vicogenta; Viewer for comparing genomes to *Arabidopsis*.
http://www.Arabidopsis.org; TAIR.
http://amdec-bioinfo.cu-genome.org/html/index.html; AMDeC Bioinformatics Core Facility at the Columbia Genome Center.
ftp://ftpmips.gsf.de/cress/; MIPS FTP site.
ftp://ftp.tigr.org/pub/data/a_*thaliana*/ath1/; TIGR FTP site.

Received September 8, 2005; accepted in revised form February 3, 2005.