Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*

Xiaoying Lin*, Samir Kaul*, Steve Rounsley†, Terrance P. Shea†, Maria-Ines Benito, Christopher D. Town, Claire Y. Fujii, Tanya Mason, Cheryl L. Bowman, Mary Barnstead, Tamara V. Feldblyum, C. Robin Buell, Karen A. Ketchum, John Lee, Catherine M. Ronning, Hean L. Koo, Kelly S. Moffat, Lisa A. Cronin, Mian Shen, Grace Pai, Susan Van Aken, Lowell Umayam, Luke J. Tallon, John E. Gill, Mark D. Adams†, Ana J. Carrera, Todd H. Creasy, Howard M. Goodman†, Chris R. Somerville†, Greg P. Copenhaver†, Daphne Preuss†, William C. Nierman, Owen White, Jonathan A. Eisen, Steven L. Salzberg, Claire M. Fraser & J. Craig Venter†

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA
** These authors contributed equally to this work*

***Arabidopsis thaliana* (Arabidopsis) is unique among plant model organisms in having a small genome (130–140 Mb), excellent physical and genetic maps, and little repetitive DNA. Here we report the sequence of chromosome 2 from the Columbia ecotype in two gap-free assemblies (contigs) of 3.6 and 16 megabases (Mb). The latter represents the longest published stretch of uninterrupted DNA sequence assembled from any organism to date. Chromosome 2 represents 15% of the genome and encodes 4,037 genes, 49% of which have no predicted function. Roughly 250 tandem gene duplications were found in addition to large-scale duplications of about 0.5 and 4.5 Mb between chromosomes 2 and 1 and between chromosomes 2 and 4, respectively. Sequencing of nearly 2 Mb within the genetically defined centromere revealed a low density of recognizable genes, and a high density and diverse range of vestigial and presumably inactive mobile elements. More unexpected is what appears to be a recent insertion of a continuous stretch of 75% of the mitochondrial genome into chromosome 2.**

Arabidopsis thaliana (thale cress), a member of the mustard family, is a dicotyledenous flowering plant with a diploid number of 10 that has become a widely used model for the study of plant biology because of its small size, short generation time, facile genetics and ease of transformation. Whole-genome analysis began ten years ago with the production of cosmid and yeast artificial chromosome (YAC)-based physical maps[1], while early large-scale sequencing efforts focused on ordered cosmids[2]. On August 20–21 1996, representatives of six research groups from Japan, Europe and the USA met in Washington DC to discuss strategies for facilitating international cooperation in completing the sequencing of the *Arabidopsis* genome (the *Arabidopsis* Genome Initiative (AGI)) by the year 2004. The group agreed on a workplan to ensure that the sequencing of all five chromosomes was completed except for the difficult to sequence repetitive regions: the nucleolar organizer regions (NORs) and centromeres (http://genome-www.stanford.edu/Arabidopsis/AGI/AGI_memo.html)[3].

The main molecular and cytological features of the *A. thaliana* chromosomes had been defined by the time high-throughput sequencing began (Fig. 1). The NORs in *A. thaliana* are found on the upper ends of chromosomes 2 and 4. Each region, 3.5–4 Mb in length, consists of an array of tandemly repeated ribosomal RNA genes[4]. Using a combination of restriction analysis, polymerase chain reaction (PCR) and sequencing, it has been shown that the rDNA of the NOR on chromosome 4 immediately abuts the telomeric repeats with less than 500 base pairs (bp) of intervening sequence[5]. Restriction analysis is consistent with a similar arrangement on chromosome 2, with the rDNA terminating at a different point within the repeat unit. By contrast, the genomic organization of *A. thaliana* centromeres is not well characterized. These regions are heterochromatic[6] and contain large tandem arrays of a family of 180-bp elements[7–9] that are reminiscent of the 170-bp alphoid repeats found in primate centromeres[10]. Large blocks of DNA consisting primarily of 180-bp repeats have been characterized by 2D gel electrophoresis and assigned to specific centromeres on the basis of restriction-fragment-length polymorphisms[11]. The 180-bp repeat block on chromosome 2 is 820–830 kb in size[11], which should be considered as a minimum estimate for the size of the unsequenced centromeric region on this chromosome.

Features of chromosome 2

Chromosome 2 is acrocentric and was originally estimated to be 13.5 Mb in length from the YAC-based physical map[12]. Sequencing was initiated using bacterial artificial chromosomes (BACs) that had been placed onto the physical map of chromosome 2 by hybridization to YACs and genetic markers. Subsequently, BAC end sequences and BAC fingerprint data[13] allowed extension from these initial seed points and completion of the entire chromosome. A total of 257 BAC and P1 clones (including 5 BACs completed by other groups) were sequenced to produce over 24 Mb of finished sequence, which has been assembled as two contigs terminating in blocks of 180-bp repeats. These repeats represent the inner boundaries of our finished sequence.

The upper (short) arm, measured from the lower end of the NOR to the centromeric 180 bp repeats, is 3.6 Mb. The northern-most BAC (F23H14) in this contig contains a single rDNA repeat unit that is oriented in the same direction as the telomere-proximal repeat, suggesting that all the rDNA units are arranged in a head-to-tail fashion running 5′ to 3′ from the telomere towards the centromere[14]. Immediately adjoining the rDNA is about 60 kb of highly repetitive sequence containing many transposons. The lower (long) arm from centromere to telomere is 16 Mb. The sequence of the southern-most BAC in this contig (F11L15) is joined by a small PCR fragment to the *A. thaliana* sequence in pAtT51, a telomere-containing clone previously mapped to TEL2S[15] (http://genome-www.stanford.edu/Arabidopsis/ww/Nov98RImaps/index.html). Within this region, a putative transcriptional co-activator gene was identified. Because pAtT51 was derived from the Landsberg ecotype, the last 2.5 kb of this contig may not be identical to the sequence in the Columbia ecotype.

† Present addresses: Cereon Genomics, 270 Albany Street, Cambridge, Massachusetts 02139, USA (S.R., T.P.S.); Celera Genomics Corporation, 45 West Gude Drive, Rockville, Maryland 20850, USA (M.D.A., J.C.V.); Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachussets 02114, USA (H.M.G.); Department of Plant Biology, Carnegie Institution of Washington, 260 Panama St., Stanford, California 94305, USA (C.R.S.); Department of Molecular Genetics and Cell Biology, University of Chicago, Chicago, Illinois 60637, USA (G.P.C., D.P.).

At 19.6 Mb, the overall length of chromosome 2 (excluding the NOR and centromere) is 45% larger than the original estimate. This difference is due in part to gaps in the YAC-based physical map and to a deletion in one of the YACs (see Fig. 1). Similar increases in the actual physical lengths of the other *A. thaliana* chromosomes indicate that the genome size is actually 130–140 Mb, rather than the earlier estimates of 70–100 Mb.

Using sequence information for 37 markers from the *Arabidopsis* Recombinant Inbred genetic map (http://genome-www.stanford.edu/Arabidopsis/ww/Nov98RImaps/index.html), the relationship between physical and genetic distances along the chromosome was examined. As the centromere represents a discontinuity, regression analysis for the two arms was performed separately. The ratios of physical to genetic distance on the short arm and long arm were not significantly different, with values of 244 kb cM^{-1} and 223 kb cM^{-1}, respectively (Fig. 2a). Although these two values indicate a similar overall rate of recombination along the two arms of the chromosome, local distortions are evident which may reflect chromosomal rearrangements between the two ecotypes used to construct the mapping population (Fig. 2b).

Gene content

Using a combination of gene prediction programs and database searches, the chromosome was annotated (Fig. 3 (facing page 768); and Table 1). Statistics of the chromosomal composition are listed in Table 2, along with corresponding data from other eukaryotic genome projects. A graphical distribution of various features is shown in Fig. 4. Of the 4,037 genes identified, 2,078 (51.5%) were assigned either a definite or putative function, 865 (21.4%) were labelled as unknown genes and 1,094 genes (27.1%) were designated as hypothetical. In addition, 400 pseudogenes, most of which are related to proteins found in retrotransposons and are located near the centromere, and 79 genes encoding structural RNAs (73 tRNA, 4 tRNA and 2 small nucleolar (snRNA)) were found.

On average, a gene occurs every 4.4 kb, is about 2 kb in length (from start to stop codon) and contains 4.6 exons. More than 50% of the chromosome is therefore intergenic sequence. Twenty-three per cent of the genes consist of a single exon, whereas the largest gene (*T20F21.17*) contains 52 predicted exons and encodes a protein that is 50% identical to the human ch-TOG protein[16]. At least 1,352 (33.5%) genes are expressed, as reflected by the presence of a matching complementary DNA or expressed sequence tag (EST) entry in GenBank. Genes that are most highly represented by ESTs include a glycine-rich RNA-binding protein, aquaporins, ribulose bisphosphate carboxylase activase, a chlorophyll *a*/*b* binding protein and lipid transfer proteins. The distribution of genes and ESTs along the chromosome is shown in Fig. 4b, c.

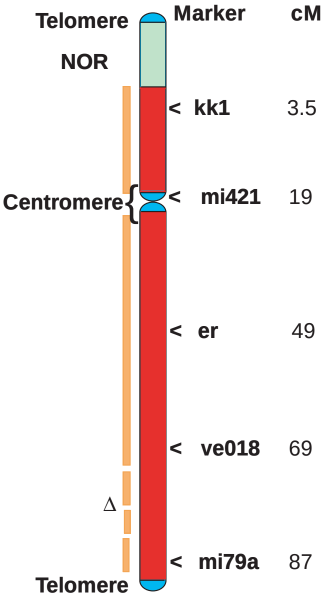


Figure 1 Features of *Arabidopsis thaliana* chromosome 2. The chromosome is roughly 23 Mb long (including the NOR), spans a genetic map distance of 90 cM and contains 15% of the unique sequence of the *A. thaliana* genome. The locations of a number of genetic markers are shown. The original YAC contigs are represented by the orange line, and the location of the deletion noted in the text is shown (Δ).

Classification by biological role or biochemical function of the known and putative genes shows that most major cellular processes appear to be represented, with genes involved in regulatory functions and signal transduction (DNA-binding proteins/transcription factors and protein kinases) comprising the largest functional groups (Table 1). These homology-based assignments are consistent with results from searching the hidden Markov models (HMMs) of protein domains in the pFam database[17]. Of the 4,037 genes predicted to encode a protein, 1,436 match at least one of the 434 distinct pFam HMMs (22% of the total HMMs). The most frequently seen matches are to the LRR (leucine-rich repeat, 239 matches), the pkinase (eukaryotic protein kinase, 161 matches) and the zf-C3HC4 (zinc finger, C3HC4 type, 46 matches) HMMs. The cellular destinations of the proteins were also predicted. Over 17% (698) contained signal peptides, whereas over 10% (416) were predicted to be organelle-targeted (135 for the chloroplast and 281 for the mitochondrion). In addition, more than 16% of the proteins are probably located in cellular membranes, with 665 sequences predicted to have 4 or more transmembrane segments (TMSs). Furthermore, about 25% of the proteins with 1, 2 or 3 predicted TMSs (1,438) are predicted to be localized to the membrane.

Gene and chromosomal duplications

Chromosome 2 contains many duplicated genes. More than 60% of

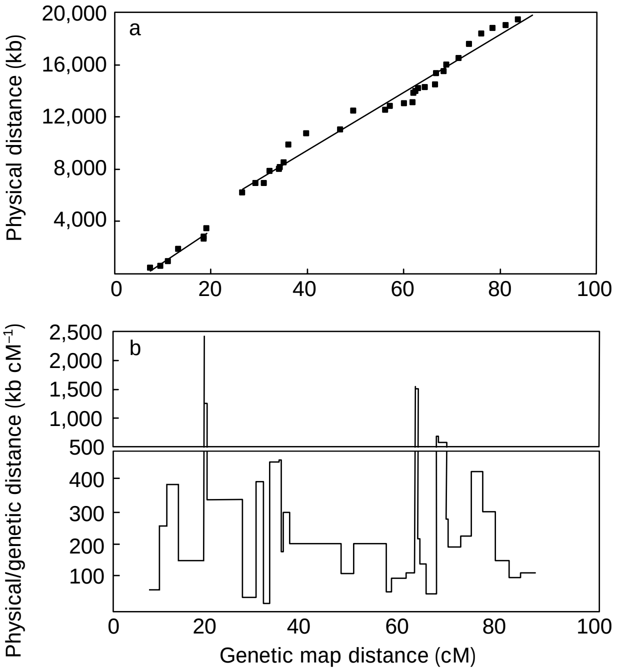


Figure 2 Representation of physical/genetic distance along chromosome 2. **a**, Physical distance (in kb) is plotted against position of each marker on the Recombinant Inbred (RI) map. Data above and below the centromere are treated as separate sets for regression analysis. Note that genetic distance is measured from the telomere (TEL2N), whereas physical distance is measured from the bottom of the NOR. **b**, The ratio of physical to genetic distance between successive pairs of markers is plotted against map position for each marker pair on the RI map.

the predicted proteins (2,542 out of 4,037) have a significant match (FASTA, $P < 10^{-10}$) to another *A. thaliana* protein, with 2,138 matching at least one other protein encoded on chromosome 2. Of these, 593 are found in 239 tandem duplications that range in size from 2 to 9 genes (average 2.5). A particularly striking example is found in BAC F16P2, which contains tandem repeats of three gene families (Fig. 5). Of the 12 intact copies of a putative short-chain dehydrogenase/reductase gene in this region, 9 are found in tandem. This gene family is most similar to tropinone reductase, a branch point in the biosynthesis of tropane alkaloids including such medicinally important compounds as atropine, scopolamine and cocaine[18,19]. Although these alkaloids are not known to occur in the Brassicaceae, at least one of these genes is expressed, suggesting a role in secondary metabolism which has yet to be elucidated. This same BAC also contains a tandem array of seven genes encoding glutathione *S*-transferase (GST), which has been implicated in numerous plant processes including stress and pathogen responses, xenobiotic tolerance and vacuolar sequestration[20]. In total, 13 GST genes have been found on chromosome 2, with 4 of the other genes occurring as tandem pairs. The GSTs within the long tandem repeat are more closely related to one another than to any other of the 25 GSTs known in *Arabidopsis*, indicating that this expansion is of relatively recent origin. There is also a smaller duplication consisting of three pumilio-like genes, two of which are in tandem. The proximity of these gene duplications suggests either that this is a repeat-prone region of the chromosome or that they arose as part of the same expansion event.

Table 1 Classification of chromosome-2 proteins according to role or function

	Percentage of chromosome-2 proteins
Regulatory functions	7.8
Signal transduction	4.0
Cellular structure, organization and biogenesis	3.6
Transport and binding proteins	3.6
Protein fate	3.0
Energy metabolism	2.7
Growth and development	2.6
Secondary metabolism	2.5
Cellular processes	2.4
Pathogen responses	2.0
Protein synthesis	1.9
Fatty acid and phospholipid metabolism	1.5
General transcription	1.4
Environmental response	1.0
DNA metabolism	0.7
Central intermediary metabolism	0.7
Amino-acid biosynthesis	0.6
Purine and pyrimidine base, nucleoside and nucleotide metabolism	0.6
Biosynthesis of cofactors, prosthetic groups and carriers	0.4
Other categories	6.9
Unclassified	1.4
Unknown protein	21.4
Hypothetical protein	27.1

4,037 genes are included in this analysis.

Some of the duplicated genes are found within large chromosomal duplications (Fig. 6). A segment of the bottom arm of chromosome 2 (13.4–14.1 Mb, containing 170 genes) from clone F20M17 to F4P9 was found to be duplicated on chromosome 1 (from clone F19P19 to F3F20) with an inversion in the middle (13.8–13.9 Mb on chromosome 2). Within this region, 57 gene pairs (33%) are duplicated between the two chromosomes. Several instances of tandem duplications on just one of the two chromosomes were also observed. For example, F4P9.8 encodes an auxin-inducible protein that matched two adjacent copies of a similar protein-coding sequence (F19P19.31 and F19P19.32) on chromosome 1; and a single copy of a gene encoding an ion-channel protein on chromosome 1 (YUP8H12.9) matched two adjacent copies on chromosome 2 (T32F6.9 and T32F6.8). Sequence comparison of the proteins in these asymmetric duplications suggests that the tandem duplication events occurred after the large-scale duplication. An additional smaller section near the top of chromosome 2 (200–500 kb) shares 16 genes with a region on chromosome 1 (T10B6). An even larger duplication (4.6 Mb) between chromosome 2 (6.7–11.3 Mb, from clone F9O13 to F18A8) and chromosome 4 (from clone F20O9 to the bottom end of the chromosome) was found. In this region, 39% of the genes (430 out of 1,100) are duplicated between the two chromosomes, including some singleton-tandem duplication pairs, as described above. In addition, although this region contains megabase-scale rearrangements (translocations or inversions), within each rearranged segment gene order is still preserved between the two chromosomes. A small part of this duplication (45 kb) has been described recently[21].

Comparative and evolutionary genomics

Comparison of the predicted proteins on chromosome 2 to other *A. thaliana* proteins, to proteins from other plants, and to all available complete genome sequences provides insight into the evolution of plant genomes. More specifically, the timing of gene duplications can be better understood. For example, as most of the 2,542 chromosome-2 proteins (83%) that have paralogues within the *A. thaliana* genome are more similar to their paralogue than to any protein in the available completed genomes, it is likely that these gene duplications occurred after the separation of plants from the animal and fungal lineages. This analysis also suggests that all plants have a common set of genes for many functions. Even though sequencing in other plants has focused primarily on particular types of genes, more than 40% (1,656) of the proteins encoded on chromosome 2 have a significant match to a protein previously

Table 2 Compositional analysis of *Arabidopsis* chromosome 2 and comparison with other eukaryotic chromosomes and genomes

Feature	*A. thaliana**	*P. falciparum**	*S. cerevisiae*†	*C. elegans*†
Overall length	Total: 19.6 Mb Short arm: 3.6 Mb Long arm: 16.0 Mb	0.95 Mb	13.4 Mb	97 Mb
Base composition (%GC)	Overall: 35.8% Coding: 43.6% Non-coding: 32.1%	19.7% 24.3% 13.3%	38% 40% 35%	36% § §
Number of genes	4,037	209	6,339	19,099
Gene density	4.4 kb per gene	4.5 kb per gene	2.1 kb per gene	5.0 kb per gene
Exon statistics‡	20,352 exons 4.6 exons per gene 295 bp per exon	355 exons 1.7 exons per gene 728 bp per exon	6,561 exons 1.04 exons per gene §	§ 5 exons per gene §
Intron statistics	15,926 introns 3.6 introns per gene 178 bp per intron	144 introns 1.6 introns per gene 207 bp per intron	§ § §	§ § §

* Chromosome 2 only.
† Complete genome.
‡ Includes intronless genes.
§ Not published.

identified in other plant species. Of these, 48% (803) do not have a match in any complete genome sequence. In addition, comparison with other species indicates that many of the genes on chromosome 2 probably originated within the plant lineage. More than 65% (2,745) of the predicted proteins do not show significant similarity to any other completed genome, suggesting that they are unique to plants. Of the remaining proteins (1,293), the majority are most similar to proteins from *Saccharomyces cerevisiae* (368 or 28%) or *Caenorhabditis elegans* (633 or 49%). The number of proteins whose best match is to one of these two proteomes is proportional to the total number of proteins encoded in each genome, suggesting that *A. thaliana* is roughly equally evolutionarily distant from animals and fungi. The remaining 292 proteins (23%) are most similar to bacterial and archaeal proteins, which may be due in part to transfers of genes from organellar genomes to the nucleus (see below).

Lateral transfers of genes from organelles to the nucleus have been well established[22]. In most cases, such genes were identified because their products were known to function in the organelle. The sequence of chromosome 2 allows a reverse type of analysis, identification of nuclear genes that originated in the chloroplast by their evolutionary ancestry. Genes most similar to cyanobacterial genes relative to genes from other complete genomes were considered to have been acquired from the chloroplast genome. Using this methodology, 135 putative chloroplast-derived genes were identified. Of these, 71 (52.5%) are predicted to have chloroplast-targeting peptides. The putative functions of these genes include DNA repair, protein synthesis, regulatory functions, cell division, photosynthesis, transport and oxidation–reduction. As these genes are not found in any of the completed plant chloroplast genomes, most of these gene transfers probably occurred soon after the symbiosis with the chloroplast evolved. Many other genes are of unknown function and will be worth pursuing as potential novel chloroplast-targeted functions.

A mitochondrial genome in the nucleus

Completion of the chromosome has revealed a large and

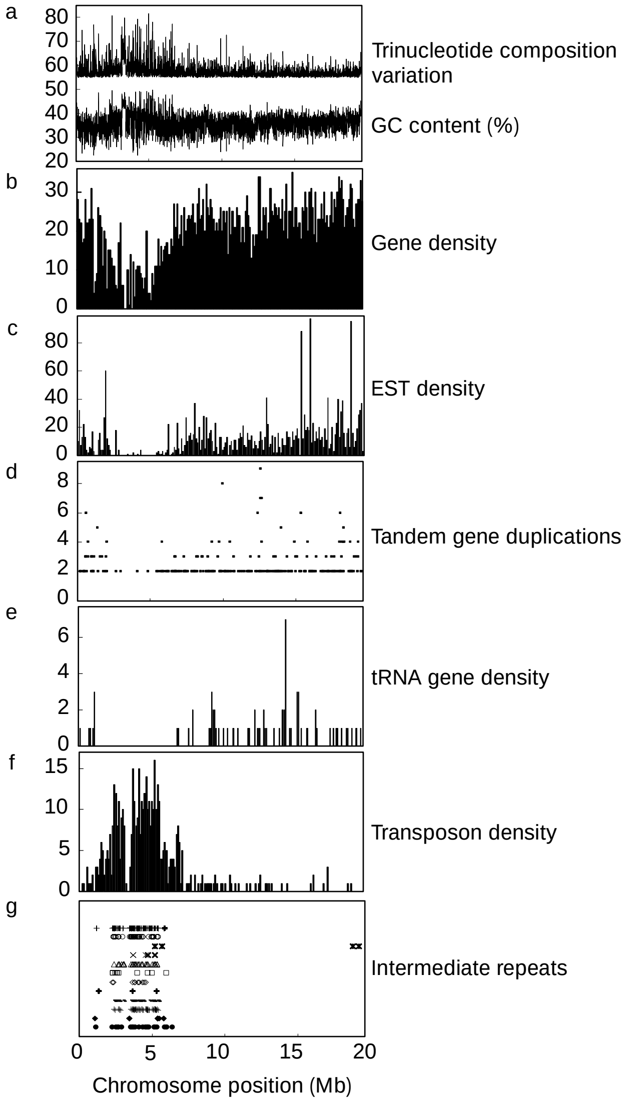


Figure 4 Distribution of various features along chromosome 2. The two arms of the chromosome are represented as a continuous sequence of DNA, beginning immediately after the NOR (position 1) and continuing to the centromere (position 3,606,929), where it is joined to the long arm (which terminates with the most telomeric base (position 19,647,005)) by a 120-bp spacer. The various features are plotted relative to the nucleotide position on the chromosome. **a**, Chi-squared statistics of trinucleotide frequency[50] and GC composition in a 10-kb sliding window. **b**, Gene density per 100 kb; **c**, EST density per 100 kb. **d**, Tandem repeats, showing repeat number as well as location. **e**, tRNAs per 100 kb. **f**, Density of transposon-related sequences per 100 kb. **g**, Position of pericentromeric repeat classes along chromosome 2. From top to bottom they are 180 bp, Athila, 106B, 11B7RE, 163A, 164A, 278A, mi167, pAtT27, new repeats 1–4. The GC peak (45%) and the gaps in gene and transposon density from 3.2 to 3.5 Mb reveal the site of the mitochondrial DNA insertion.

unexpected organellar-to-nuclear gene-transfer event. Within the genetically defined centromere, there is a stretch of 270 kb of sequence that is nearly identical to that of the *Arabidopsis* mitochondrial genome. The authenticity of this insertion in the Columbia ecotype was confirmed by PCR amplification across the junctions of mitochondrial and unique nuclear DNA, followed by the sequencing of the corresponding fragments. This insertion is much larger than any of the previously reported organellar-nuclear transfers[23], and is 99% identical to the mitochondrial genome, suggesting that the transfer event was very recent. As the published sequence is derived from the C24 ecotype[24], it is not possible determine whether the observed sequence polymorphisms are due to divergence of the insertion from the mitochondrial genome or to differences among ecotypes. The organization of the mitochondrial

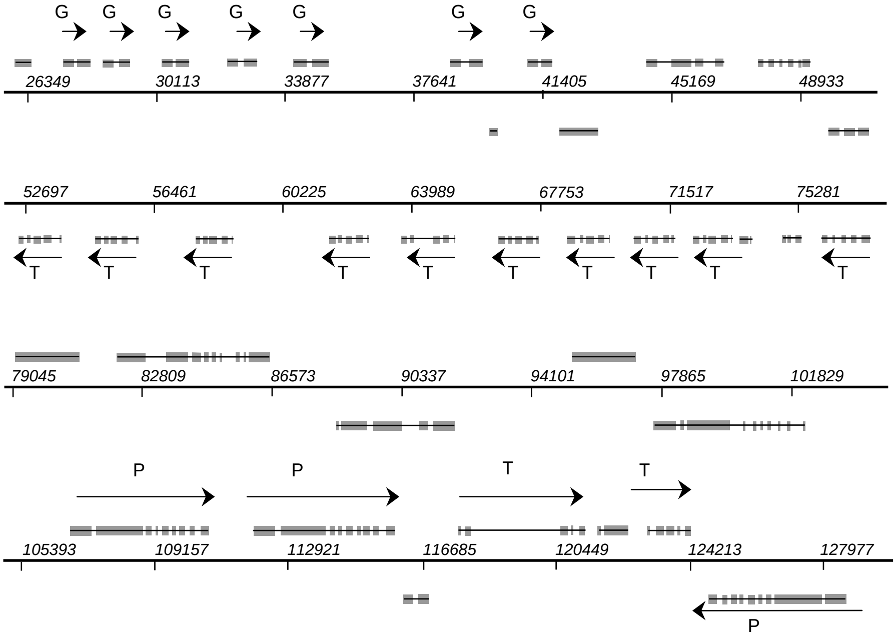


Figure 5 Organization of genes on BAC F16P2 showing the three tandem gene duplications. The display from TIGR Annotator shows the exon–intron structure of the annotated genes. The glutathione *S*-transferase and tropinone reductase genes are labelled G and T, respectively. A smaller duplication of pumilio-like protein (P) is also present.

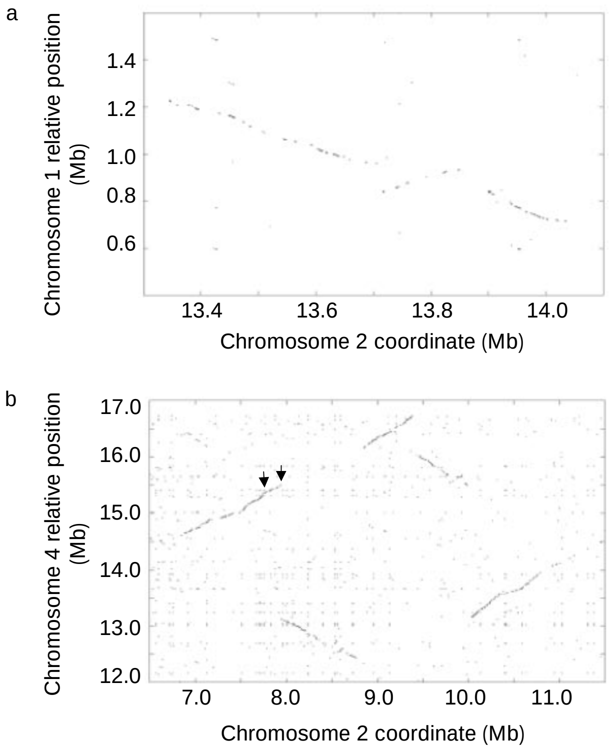


Figure 6 Large chromosomal duplications. Regions of similarity were initially detected using the MUMmer program after which sequences were aligned using the dds program with criteria of at least 75% identity over 100 bp. Points in the dot plot represent the coordinates of each such match. Values show positions (in bp) within each contig assembled from all available sequence. **a**, Duplication between chromosomes 1 and 2. **b**, Duplication between chromosomes 4 and 2. Arrows indicate the small, previously described duplication[21].

DNA in the nucleus differs from that of the published mitochondrial genome and its predicted alternate forms[25]; it corresponds to another possible isoform with an internal deletion (Fig. 7). This deletion may have occurred during or after transfer or may represent an alternate form of the Columbia mitochondrial genome.

The pericentromeric region

Until the completion of chromosome 2, it was unclear what features would be discovered in the regions around the centromere (the pericentromeric region). In addition to the detection of the mitochondrial genome insertion, an analysis of this region has revealed the presence, distribution and diversity of specific repetitive elements on a single chromosome. Transposons were identified by searching for predicted amino-acid sequences that are similar to characterized transposases, reverse-transcriptases, polyproteins or other transposon-related coding sequences. Of the 563 transposon proteins annotated, 50% are pseudogenes found either as fragments of previously intact elements or as genes interrupted by stop codons, frameshifts or other mutations. A breakdown of the classes of transposons identified is shown in Table 3. Representative members of most of the large transposon families seen in other plants (*Mutator*, *En-Spm* and *Ac-Ds* (*haT/mariner*)) are present. Retroelements including members of the *LINE*-like Ta11 elements, long terminal repeat (LTR) elements in both the *Ty3-gypsy* and the *Ty1-copia* family, and the *Arabidopsis Athila* family represent the largest class of transposons. This accumulation of retroelements and other transposons may be due to either preferential insertion into the regions flanking the centromere or their elimination from the rest of the genome at a higher rate[26]. In addition, representatives of previously described intermediate repeats as well as some new repeats were identified (Fig. 4f), a distribution consistent with previous hybridization data[27–30].

As shown in Fig. 4, the number of recognizable genes decreases close to the centromere, whereas the number of transposable elements and various classes of intermediate repeats increases. Nevertheless, within this region there are some identifiable genes: a protein kinase (T25N22.14); a plasma membrane proton ATPase (F9A16.7); an ABC transporter (T5E7.1); a DNA-replication licensing factor (T12J2.1), an RNA helicase (T12J2.7) and a 40S ribosomal protein S16 (F7B19.13). The RNA helicase, protein kinase and 40S ribosomal protein appear to be actively expressed genes, as each is represented in the EST collection.

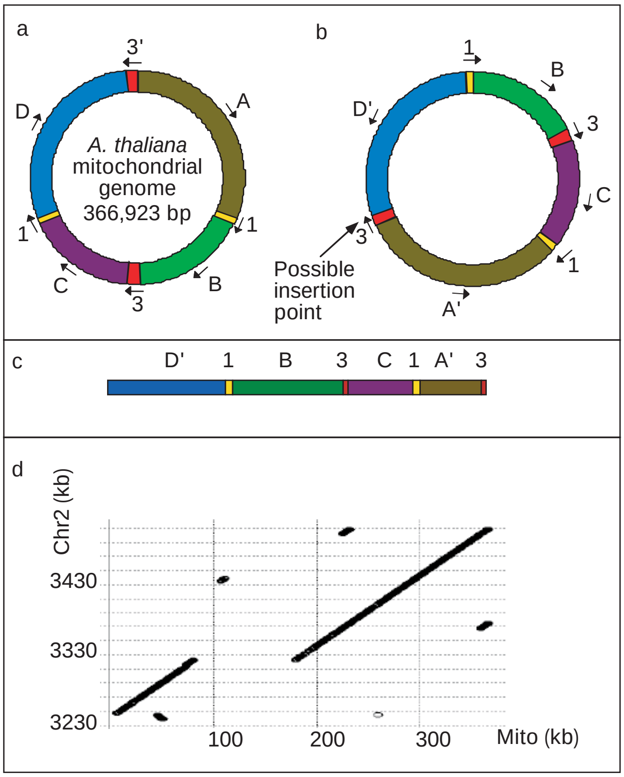


Figure 7 Proposed model for insertion of mitochondrial DNA into chromosome 2. **a**, Diagram of the major form of the published *A. thaliana* C24 mitochondrial genome[25]. The different colours correspond to the major segments of the genome: A (nucleotide 297,579–44,698); B (48,895–112,146); C (118,737–178,862); D (183,060–290,989); and two major repeats, 1 (44,698–48,894, 178,863–183,059) and 3 (112,147–118,736, 297,579–290,990). A small repeat (repeat number 2) is not shown. The orientation of one of the three repeats is inverted relative to the other. The arrows are used to indicate orientation for comparison with the alternative form. **b**, Hypothetical alternative form, generated by *in silico* rearrangements around some of the repeats. Arrows and primes indicate inversions. **c**, Linearized form of genome shown in **b** if inserted at the point between repeat 3 and segment D′. **d**, Dot plot of the linearized alternate form against the region of chromosome 2 containing the mitochondrial DNA. The gap in the diagonal indicates that there was a deletion of a large segment (see text).

Conclusions

Arabidopsis has been considered to be not only a tractable species for plant biology studies, but also a suitable model from which a fundamental understanding of basic plant processes could be applied to agriculturally relevant species. The analysis of the complete sequence of chromosome 2 validates this concept because many genes that have previously been characterized in other plants have been found. As observed with other eukaryotic genome projects, however, roughly half of the predicted genes have no known function. This represents a large, untapped resource of information that will require a combination of knockout approaches, antisense technologies and global gene-expression analyses to understand their functions.

Both gene duplication and large-scale chromosomal duplications are observed, events that apparently occurred after the divergence of plants from other lineages. These duplications are larger than any reported in the *C. elegans* genome[31], but less extensive than the complete genome duplications proposed for maize[32] and *S. cerevisiae*[33]. As gene duplication is often accompanied by functional divergence[34], the identification of such duplications in the plant lineage may help identify new plant functions. Many of the duplications occurred in tandem, generating closely linked gene families which may provide some of the allelic diversity that breeders and agricultural researchers are seeking. Some new families, such as the family of genes that may encode one step in alkaloid biosynthesis, could provide a reservoir of molecular tools to modify plant metabolism and productivity in a beneficial manner.

Table 3 Representaton of transposable elements on chromosome 2

Transposon class	Subclass	Family	Number of open reading frames
DNA elements			
		Mutator	73
		En-Spm/Tam1/Psl	71
		Ac-Ds	12
		mariner	2
		hAT	1
Retroelements			
	Non-LTR *LINE*-like retrotransposons		
		Ta11-like	152
		TSCL	1
	LTR retrotransposons		
		Athila	75
		Other *Ty1-copia*-like (*Ta1*) *Ty3-gypsy*-like (*Tat1*)	141
	Retroviral-like	Replication protein	28
		Helicase	7
		Total	563

The completion of chromosomes 2 and 4 (ref. 35), which accounts for 30% of the entire *Arabidopsis* genome, represents a landmark achievement in eukaryotic genomics. The 16-Mb contig on the lower arm of chromosome 2 is the largest sequence assembly published to date. Sequencing within the genetically defined centromere has shown not only an abundance of degenerate transposable elements, but also a number of recognizable intact genes, some of which appear to be expressed. As researchers move on to sequence larger and more complex genomes, it remains to be seen whether the level of genome closure achieved in this model plant can be duplicated.

Note added in proof: Since this paper was accepted for publication, a 23Mb contig from human chromosome 22 has been published (See *Nature* **402,** 489–495 (1999). □

Methods

Sequencing

Three libraries made from the Columbia ecotype of *A. thaliana* were used in sequencing chromosome 2: the TAMU BAC library (clones prefixed with T[36]); the IGF BAC library (clones prefixed with F[37]); and the Mitsui P1 library (clones prefixed with M[38]). Sheared BAC DNA (1.6–2 kb) was ligated to a modified pUC19 vector and transformed into *Escherichia coli*. Sequencing reactions were performed using either BigDye primers, BigDye Terminators or Dichlorohodamine Terminators (P.E. Biosystems), and were run on ABI 377 and ABI 3700 sequencers (P.E. Biosystems). Shotgun clones were sequenced to generate 7–8-fold coverage of each BAC. In total, 590,165 reactions were carried out to generate the sequence of the chromosome. In collaboration with Celera Genomics, 99,072 reactions were run on ABI 3700 sequencers, which accounted for nearly 20% of the overall sequence for the chromosome. The discrepancy rate on these machines, as determined by comparison of overlapping BACs, was comparable to that of ABI 377 machines (1 unresolved discrepancy per 32,000 bases). A base difference is possibly due to residual heterozygosity or to different DNA sources used to make the various libraries. Individual BACs were assembled from the shotgun sequences using the TIGR Assembler program[39]. Following assembly into contigs, the Grouper program (TIGR, unpublished software) was used to link individual assemblies within the BACs. The BACs were then closed using a combination of BAC walking, directed PCR and resequencing of individual shotgun clones. To confirm the assembly of the entire BAC sequence, predicted and actual restriction digest patterns were compared. The collinearity of the BACs with the *Arabidopsis* chromosome 2 sequence was confirmed by alignment of the sequence of chromosome 2 with markers obtained from the *Arabidopsis* Recombinant Inbred map (http://genome-www.stanford.edu/Arabidopsis/ ww/Nov98RImaps/index.html) and with YAC end sequences. Several markers that clearly mapped to single genetic loci were excluded from this analysis because they contained repetitive sequence and could not be assigned a unique location on the sequence map.

Two BACs (T5M2 and T5E7) within the genetically defined centromere[26] were found to contain long stretches of sequence with very high similarity to the *Arabidopsis* C24 mitochondrial genome (69 kb on T5M2, and 74 kb on T5E7) joined directly to nuclear sequence. Two overlapping BACs provided an additional 122 kb of purely mitochondrial sequence to close the gap, suggesting that the total size of the organellar DNA insertion is about 270 kb. Verification of the mitochondrial insertion involved PCR amplification of the two flanking junction sequences from *Arabidopsis* Columbia genomic DNA. PCR products with the expected size were amplified that had sequence identical to the junction BACs. Similar sized junction PCR products were obtained using additional BACs from both the TAMU and IGF libraries. Determination of the exact size and sequence of the insertion will be facilitated once the sequence of the Columbia mitochondrial genome is known, and it becomes possible to distinguish between nuclear- and organellar-derived mitochondrial sequences.

The sequence of chromosome 2 identified by the BAC clone F11L15 did not contain the telomeric repeat consensus sequence 5′-TTTAGGG-3′ (ref. 15). However, the sequence of the lower telomere of chromosome 2 had been previously identified in the clone pAtT51 (ref. 15) (http://genome-www.stanford.edu/Arabidopsis/ww/Nov98RImaps/index.html). pAtT51 was sequenced to identify primer sites that could be used to walk inwards from the telomere. Southern hybridization using probes derived from F11L15 and pAtT51 suggested that the gap between the BAC contig and pAtT51 was less than 1 kb. A 702-bp PCR product that spanned this physical gap was amplified from *Arabidopsis* Columbia genomic DNA and sequenced.

Analysis

Annotation of chromosome 2 involved both DNA and protein database searches and gene prediction programs. Gene predictions were made by Genscan[40], Genefinder (P. Green and L. Hillier, unpublished data) and GRAIL[41]. Splice sites were identified by these programs, by NetPlantGene[42] and by alignment with EST and protein sequences using the dds and dps programs[43]. Predicted protein sequences were searched against a non-redundant amino-acid database and against HMMs of the protein domains from pFAM3 (1,407) and TIGR built HMMs (502) with the HMMer2 program[17]. tRNAs were identified using tRNAscan-SE[44]. Repeats were identified using RepeatMasker (A. F. A. Smit and P. Green, http://ftp.genome.washington.edu/RM/RepeatMasker.html) and a modified version of the suffix tree algorithm used in the MUMmer system[45]. Output from the gene finding and signal detection programs was displayed using the Annotator genome viewer (L. Zhou, unpublished data) and manually edited. The accuracy of computational gene finders for *Arabidopsis* is quite variable and all gene predictions that are not supported by independent evidence, such as protein or EST homology, should be regarded as tentative, pending further research.

Putative membrane-spanning proteins were identified using TopPred[46]. SignalP[47] was used to detect signal peptide sequences, while Predotar (N. Peeters *et al.*, unpublished) searched for organellar targets. In addition, ChloroP[48] and Mitoprot[49] were used to find proteins targeted for the chloroplast and mitochondria, respectively. The results presented are the concordance of the predictions.

We analysed the proteome of chromosome 2 by comparing each predicted protein with a database of all proteins from chromosome 2, all available completed genomes and all proteins from *A. thaliana* and *C. elegans* using described methods[50]. To identify genes on other chromosomes that were related to chromosome-2 genes, nucleotide searches were also used as much of the nucleotide sequence in GenBank is unannotated. Each predicted transcript was searched against all available genomic sequence using a cut-off of 75% identity over 50% of the transcript length in order to take intron sequences into account.

To search for large-scale duplications within the genome, we assembled pseudomolecules representing all available sequence of chromosomes 1 and 4. The sequence of chromosome 2 was then compared with chromosomes 1 and 4 using the MUMmer program[45] with a word size of 20. Long-range duplications were found by plotting the results, and a local alignment program dds[43] was used to identify matches between the exact repeats within the duplicated regions (stringency >75%). Because of insufficient sequence or overlap information, it is difficult at the present time to construct chromosome-size pseudomolecules of chromosomes 3 and 5 and assess the possibility of other large duplications.

Received 5 October; accepted 29 October 1999.

1. Schmidt, R. & Dean, C. Towards construction of an overlapping YAC library of the *Arabidopsis thaliana* genome. *Bioessays* **15,** 63–69 (1993).
2. Bevan, M. *et al.* Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. *Nature* **391,** 485–488 (1998).
3. Bevan, M. *et al.* Objective: the complete sequence of a plant genome. *Plant Cell* **9**, 476–478 (1997).
4. Copenhaver, G. P. & Pikaard, C. S. Two-dimensional RFLP analyses reveal megabase-sized clusters of rRNA gene variants in *Arabidopsis thaliana*, suggesting local spreading of variants as the mode for gene homogenization during concerted evolution. *Plant J.* **9,** 273–282 (1996).
5. Copenhaver, G. P. & Picaard, C. S. RFLP and physical mapping with an rDNA-specific endonuclease reveals that nucleolus organizer regions of *Arabidopsis thaliana* adjoin the telomeres on chromosomes 2 and 4. *Plant J.* **9,** 259–272 (1996).
6. Schweizer, D., Loidl, J. & Hamilton, B. Heterochromatin and the phenomenon of chromosome banding. *Results Probl. Cell Differ.* **14,** 235–254 (1987).
7. Martinez-Zapater, J. M., Estelle, M. A. & Somerville, C. R. A highly repeated DNA sequence in *Arabidopsis thaliana*. *Mol. Gen. Genet.* **204,** 417–423 (1986).
8. Maluszynska, J. & Heslop-Harrison, J. S. Localization of tandemly repeated DNA sequences in *Arabidopsis thaliana*. *Plant J.* **1,** 159–166 (1991).
9. Simoens, C. R., Gielen, J., Van Montagu, M. & Inze, D. Characterization of highly repetitive sequences of *Arabidopsis thaliana*. *Nucleic Acids Res.* **16,** 6753–6766 (1988).
10. Pluta, A. F., Mackay, A. M., Ainsztein, A. M., Goldberg, I. G. & Earnshaw, W. C. The centromere: hub of chromosomal activities. *Science* **270,** 1591–1594 (1995).
11. Round, E. K., Flowers, S. K. & Richards, E. J. *Arabidopsis thaliana* centromere regions: genetic map positions and repetitive DNA structure. *Genome Res.* **7,** 1045–1053 (1997).
12. Zachgo, E. A. *et al.* A physical map of chromosome 2 of *Arabidopsis thaliana*. *Genome Res.* **6,** 19–25 (1996).
13. Marra, M. A. *et al.* High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7,** 1072–1084 (1997).
14. Copenhaver, G. P., Doelling, J. H., Gens, J. S. & Pikaard, C. S. Use of RFLPs larger than 100 kbp to map the position and internal organization of the nucleolus organizer region on chromosome 2 in *Arabidopsis thaliana*. *Plant J.* **7,** 273–286 (1995).
15. Richards, E. J., Chao, S., Vongs, A. & Yang, J. Characterization of *Arabidopsis thaliana* telomeres isolated in yeast. *Nucleic Acids Res.* **20,** 4039–4046 (1992).
16. Charrasse, S. *et al.* Characterization of the cDNA and pattern of expression of a new gene over-expressed in human hepatomas and colonic tumors. *Eur. J. Biochem.* **234,** 406–413 (1995).
17. Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* **26,** 320–322 (1998).
18. Leete, E. Recent developments in the biosynthesis of the tropane alkaloids. *Planta Med.* **56,** 339–352 (1990).
19. Yamad, Y. *et al.* in *Secondary Products from Plant Tissue Culture* (eds Charlwood, B. V. & Rhodes, M. J. C.) 227–242 (Clarendon, Oxford, 1990).
20. Marrs, K. A. The functions and regulation of glutathione *S*-transferases in plants. *Annu. Rev. Plant Physiol.* **47,** 127–158 (1996).
21. Terryn, N. *et al.* Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. *FEBS Lett.* **445,** 237–245 (1999).
22. Martin, W. & Herrmann, R. G. Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiol.* **118,** 9–17 (1998).
23. Blanchard, J. L. & Schmidt, G. W. Pervasive migration of organellar DNA to the nucleus in plants. *J. Mol. Evol.* **41,** 397–406 (1995).
24. Unseld, M., Marienfeld, J. R., Brandt, P. & Brennicke, A. The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. *Nature Genet.* **15,** 57–61 (1997).
25. Klein, M. *et al.* Physical mapping of the mitochondrial genome of *Arabidopsis thaliana* by cosmid and YAC clones. *Plant J.* **6,** 447–455 (1994).
26. Copenhaver, G. P. *et al.* Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* (in the press).
27. Thompson, H. L., Schmidt, R. & Dean, C. Identification and distribution of seven classes of middle-repetitive DNA in the *Arabidopsis thaliana* genome. *Nucleic Acids Res.* **24,** 3017–3022 (1996).

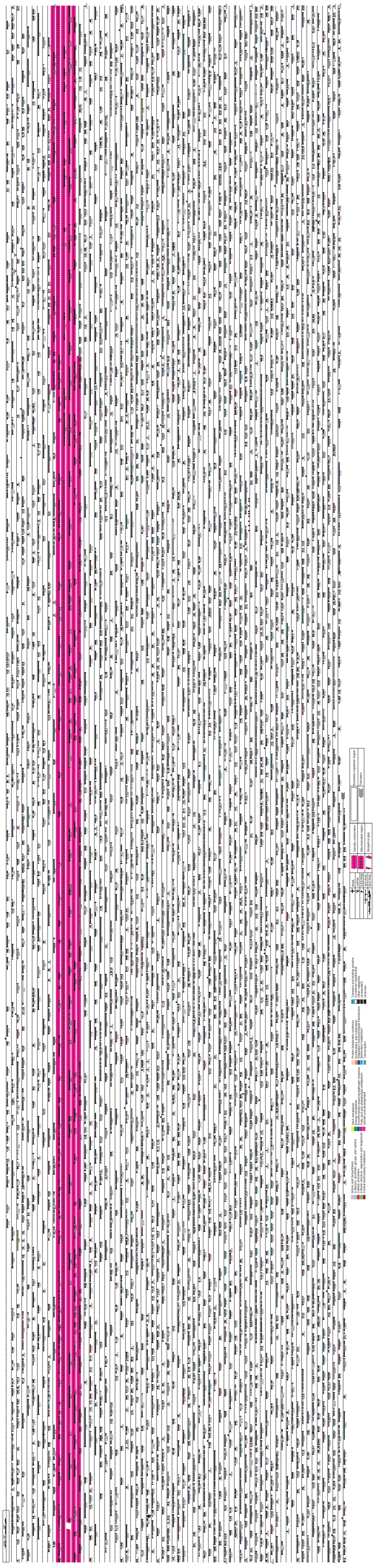
28. Thompson, H., Schmidt, R., Brandes, A., Heslop-Harrison, J. S. & Dean, C. A novel repetitive sequence associated with the centromeric regions of *Arabidopsis thaliana* chromosomes. *Mol. Gen. Genet.* **253,** 247–252 (1996).
29. Thompson, H. L., Schmidt, R. & Dean, C. Analysis of the occurrence and nature of repeated DNA in an 850 kb region of *Arabidopsis thaliana* chromosome 4. *Plant Mol. Biol.* **32,** 553–557 (1996).
30. Brandes, A., Thompson, H., Dean, C. & Heslop-Harrison, J. S. Multiple repetitive DNA sequences in the paracentromeric regions of *Arabidopsis thaliana* L. *Chromosome Res.* **5,** 238–246 (1997).
31. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).
32. Gale, M. D. & Devos, K. M. Comparative genetics in the grasses. *Proc. Natl Acad. Sci. USA* **95,** 1971–1974 (1997).
33. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387,** 708–713 (1997).
34. Hughes, A. L. The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. Lond. B Biol. Sci.* **256,** 119–124 (1994).
35. The European Union Arabidopsis Genome Sequencing Consortium & The Cold Spring Harbor, Washington University in St Louis and PE Biosystems Arabidopsis Sequencing Consortium. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* **402,** 769–777 (1999).
36. Choi, S., Creelman, R. A. Mullet, J. E. & Wing, R. A. Construction and characterization of bacterial artificial chromosome library of *Arabidopsis thaliana*. *Plant Mol. Biol. Rep.* **13,** 124–128 (1995).
37. Mozo, T., Fischer, S., Meier-Ewert, S., Lehrach, H. & Altmann, T. Use of the IGF BAC library for physical mapping of the *Arabidopsis thaliana* genome. *Plant J.* **16,** 377–384 (1998).
38. Liu, Y.-G., Mitsukawa, N., Vasquez-Tell, A. & Whittier, R. F. Generation of a high-quality P1 library of *Arabidopsis* suitable for chromosome walking. *Plant J.* **7,** 351–358 (1995).
39. Sutton, G. G., White, O., Adams, M. D. & Kerlavage, A. R. TIGR Assembler: a new tool for assembling large shotgun sequencing projects. *Genome* **1,** 9–19 (1995).
40. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
41. Uberbacher, E. C. & Mural, R. J. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl Acad. Sci. USA* **88,** 11261–11265 (1991).
42. Hebsgaard, S. M. *et al.* Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. *Nucleic Acids Res.* **24,** 3439–3452 (1996).
43. Huang, X., Adams, M. D., Zhou, H. & Kerlavage, A. R. A tool for analyzing and annotating genomic sequences. *Genomics* **46,** 37–45 (1997).
44. Lowe, T. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
45. Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res.* **27**, 2369–2376 (1999).
46. Claros, M. G. & von Heijne, G. TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci.* **10,** 685–686 (1994).
47. Nielsen, H., Brunak, S. & von Heijne, G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* **12,** 3–9 (1999).
48. Emanuelsson, O., Nielsen, H. & von Heijne, G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein. Sci.* **8,** 978–984 (1999).
49. Claros, M. G. & Vincens, P. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241,** 779–786 (1996).
50. Nelson, K. E. *et al.* Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399,** 323–329 (1999).

Acknowledgements

We thank S. Jackson, J. Jiang, K. Meyer, E. Richards and S. Tabata for helpful contributions in completing the chromosome. In addition, we would like to thank the TIGR faculty, system administrators, sequencing facility, bioinformatics department and administrative staff. This work was supported by the National Science Foundation, the Department of Energy and the US Department of Agriculture.

Correspondence and requests for materials should be addressed to J. C. Venter (e-mail: jcventer@celera.com).

Figure 3 Gene map of *Arabidopsis thaliana* chromosome 2. Predicted gene models are shown with arrowheads indicating the direction of transcription. Genes are colour coded according to broad role categories as shown in the key. The region highlighted in pink indicates the genetically defined centromere within which the cross-hatched gene-free region represents the mitochondrial insertion. The gap indicates the highly repetitive centromeric region. Note that the NOR is not drawn to scale. Gene identification numbers follow the recently adopted nomenclature (At2gNNNNN), where At is *Arabidopsis thaliana*, 2 is chromosome number, g indicates that the feature is a gene and numbering is from top to bottom of each chromosome in increments of 10. ►



Amino acid biosynthesis
Cofactors, prosthetic groups, and carriers
Cellular processes
Cellular structure, organization and biogenesis
Central intermediary metabolism
DNA metabolism
Energy metabolism
Environmental response/pathogen response
Fatty acid and phospholipid metabolism
Growth and development
Protein fate/protein synthesis
Nucleosides, and nucleotides
Regulatory functions/signal transduction
Secondary metabolism
Transcription
Transport and binding proteins
Conserved hypothetical
Other categories
Unclassified
Unknown
tRNA
snRNA
5.8S RNA
18S RNA
26S RNA
Genetic centromere
Mitochondrial region
Sequence gap
Nucleolar organizer region
Telomere