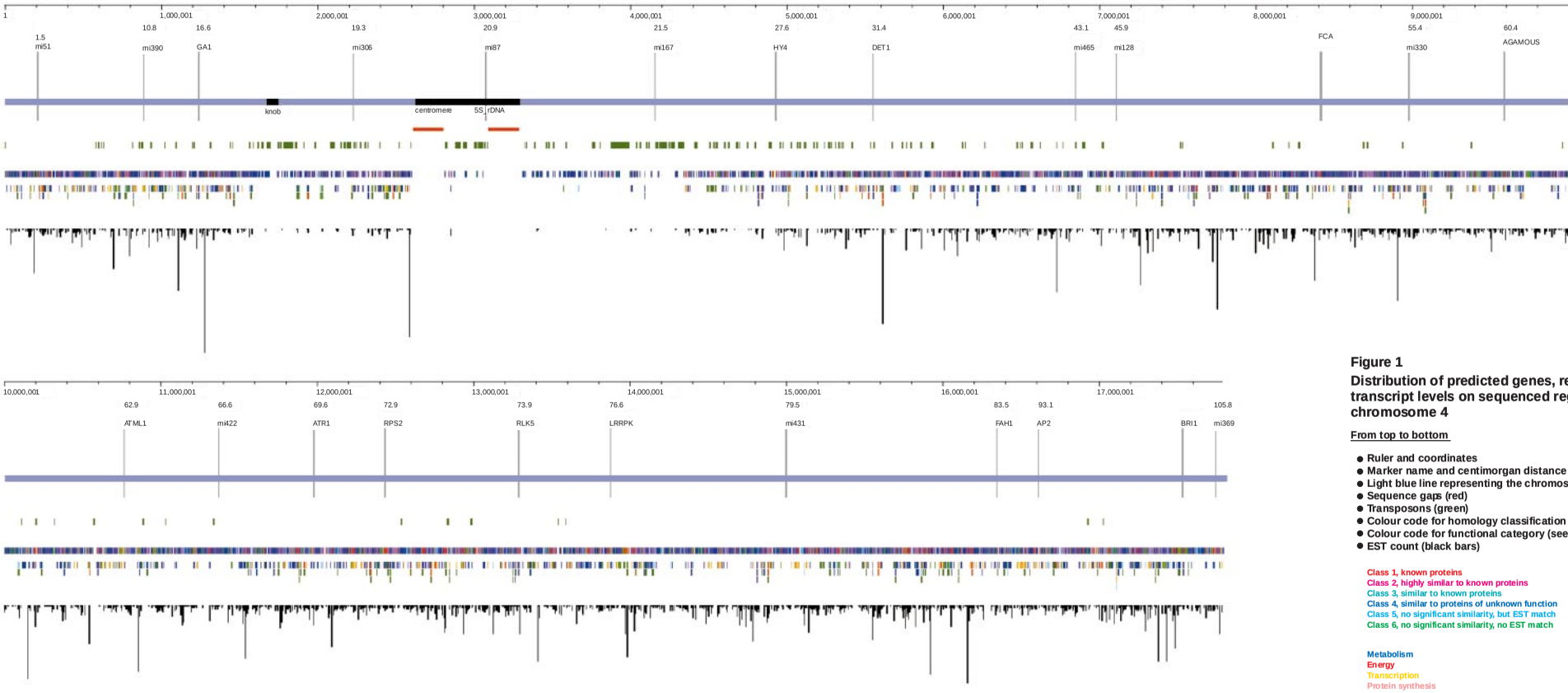

Figure 1

Distribution of predicted genes, repeats and transcript levels on sequenced regions of chromosome 4

From top to bottom

- Ruler and coordinates
- Marker name and centimorgan distance
- Light blue line representing the chromosome
- Sequence gaps (red)
- Transposons (green)
- Colour code for homology classification (see below)
- Colour code for functional category (see below)
- EST count (black bars)

Class 1, known proteins
Class 2, highly similar to known proteins
Class 3, similar to known proteins
Class 4, similar to proteins of unknown function
Class 5, no significant similarity, but EST match
Class 6, no significant similarity, no EST match

Metabolism
Energy
Transcription
Protein synthesis
Protein destination
Transport facilitation
Intracellular transport
Cellular biogenesis
Signal transduction
Disease/defence
Cellular organization

Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*

The European Union Arabidopsis Genome Sequencing Consortium & The Cold Spring Harbor, Washington University in St Louis and PE Biosystems Arabidopsis Sequencing Consortium*

* *A full list of authors appears at the end of this paper*

The higher plant *Arabidopsis thaliana* (Arabidopsis) is an important model for identifying plant genes and determining their function. To assist biological investigations and to define chromosome structure, a coordinated effort to sequence the *Arabidopsis* genome was initiated in late 1996. Here we report one of the first milestones of this project, the sequence of chromosome 4. Analysis of 17.38 megabases of unique sequence, representing about 17% of the genome, reveals 3,744 protein coding genes, 81 transfer RNAs and numerous repeat elements. Heterochromatic regions surrounding the putative centromere, which has not yet been completely sequenced, are characterized by an increased frequency of a variety of repeats, new repeats, reduced recombination, lowered gene density and lowered gene expression. Roughly 60% of the predicted protein-coding genes have been functionally characterized on the basis of their homology to known genes. Many genes encode predicted proteins that are homologous to human and *Caenorhabditis elegans* proteins.

Green plants capture solar energy through photosynthesis, which supports most other forms of life on Earth, as well as providing oxygen for aerobic organisms. Plants also synthesize a wide variety of fibres, chemicals, pharmaceuticals and raw products which provide the basis of many industrial processes. Several features of plant development and interactions with their environment are different from those of other multicellular organisms, and the molecular mechanisms of these processes are not known in detail. The elucidation of these activities at a molecular level will provide genes and knowledge for both improving the efficiencies of food and raw material production and for providing deeper insights into new developmental processes and environmental responses. *Arabidopsis thaliana* (Arabidopsis), a member of the Brassica family of dicotyledonous plants, has become an important model species for the study of many aspects of plant biology[1]. It has a relatively small nuclear genome (for plants) of roughly 130 megabases (Mb) in five chromosomes, and low repetitive DNA content. These features have lead to the development of a wide range of resources for map-based gene isolation in *Arabidopsis*. Here we describe the sequence of chromosome 4 of *Arabidopsis*.

Sequence acquisition and verification

Chromosome 4 is acrocentric, with the shorter arm tipped by near-homogeneous ribosomal DNA repeats extending for 3.5 Mb to the telomere[2]. Tiling paths of predominantly bacterial artificial chromosomes (BACs) covering chromosome 4 of the Columbia ecotype were assembled and optimized for minimal clone overlap by introducing gaps that were then spanned by polymerase chain reaction (PCR) products amplified from genomic DNA or an underlying BAC clone if available. The total sequence of 19,360,101 base pairs (bp) was derived from a composite of 131 BACs, 4 P1s, 56 cosmids and 10 PCR products. Sequence assemblies were verified by comparing the restriction profile predicted from the sequence with experimentally determined restriction profiles of each clone. Overlaps of sequence of adjacent clones were used to assess potential sequence differences between clones. *Escherichia coli* insertion elements were edited from the final sequence contig using sequence derived from PCR-amplified genomic DNA. Consistent sequence differences were resolved by sequencing PCR products amplified from genomic DNA. The collinearity of assembled sequences with the chromosome was confirmed by integrating 70 sequenced and 16 non-sequenced genetic markers into the sequence, and by comparison of this marker order with that derived from the Columbia × Landsberg *erecta* recombinant inbred (RI) linkage map[3]. All sequenced genetic markers were identified unambiguously on the chromosome; there was a good correlation between the sequence-based maps and the RI genetic map, with only one genetic marker changed in its relative position.

The predicted overall accuracy standard was less than 1 error in 10,000 bases, and the experimentally determined accuracy derived from independently sequenced overlap regions was 1 mismatch every 12,546 bp. These mismatches have been corrected. The main sequence differences encountered between clones and genomic DNA were in the number of bases, especially A and T, in simple sequence repeats. The total sequence of 19,360,101 bp occurs in three contigs: two representing the short and long arms; and a third, smaller one in the centromeric heterochromatin. The short-arm sequence of 2,608,702 bp starts in the last copy of the rDNA repeat cluster in clone T15P10 and terminates in clone T32N4. The long-arm sequence of 14,498,507 bp extends from the clone F14G16 and terminates with a telomere-proximal cosmid clone LC47K1.

The 17,385,623-bp non-redundant sequence encodes 3,744 genes, 4 small nucleolar RNAs (SnRNAs), and 81 transfer RNAs (Table 1). Nearly 50% of the sequence is protein coding, and only 19% of the genes do not have introns. The average gene density is 4,643 bp per gene, with a gene density as low as 150 kb per gene in pericentromeric heterochromatin. This average gene density is similar to that observed previously in regions of chromosome 4 (ref. 4) and 5 (ref. 5), and in the genomes of other small multicellular eukaryotic genomes, such as *C. elegans*[6] and the protozoan *Plasmodium falciparum*[7,8]. Table 1 shows other general features of the chromosome.

The 600–750 rDNA genes are arranged into two near-homo-

Figure 1 Distribution of predicted genes, repeats and transcript levels on sequenced regions of chromosome 4. Gene models and repeats are identified as single lines coloured to show the degree of homology (see Table 1 for definition) and assigned functional category. Matches of ESTs to the chromosome were performed using BLAST[44]. Searches were restricted to previously assigned coding sequences plus an additional 200 bp at the flanking 5′ and 3′ regions to detect ESTs matching to non-translated DNA. ESTs had to exceed 90% identity the genomic sequence to be considered as transcripts of the respective loci. The number of independently sequenced ESTs matching a locus is shown by the height of the black line.

geneous megabase-sized clusters at the tip of the short arms of chromosomes 2 and 4 which comprise the nucleolar organizers (NORs)[2]. NOR4 is about 3.6–4.0 Mb, and the repeat units are essentially identical except for three spacer length variants. The distal rDNA repeats are separated from the characteristic T3AG4 heptad telomeric repeat[9] by a 13-bp duplicated sequence, with no intervening subtelomeric repeats[10]. At the proximal end of the rDNA cluster, BAC T15P10 contains a single copy of the rDNA gene and is contiguous with the short-arm sequence. Cosmid LC47K1 terminates the clone tiling path on the long arm. No further clones from BAC or cosmid libraries have been found to extend the tiling path, possibly because of the proximity of this cosmid to telomeric sequences. Sequences in LC47K1 have significant similarity ($P = 9.6 \times 10^{-6}$) with the pAtT27 repeat that is associated with telomeres and centromeres[11]. Dispersed repeats consist predominantly of long terminal repeats (LTR), non-LTR retroelements and simple sequence repeats that are also found much more frequently in heterochromatin. These features are annotated in the MATDB database[12] (http://websvr.mips.biochem.mpg.de/proj/thal/).

Protein-coding sequences

The classification of predicted genes on chromosome 4 according to homology criteria is shown in Table 1. Only 8% have been characterized experimentally to date, 23% have highly significant homology to known genes including those from other organisms, and 32% potentially have significant similarity to known genes. Therefore, the actual or potential cellular role of about 60% of the genes on chromosome 4 can be predicted using sequence similarity criteria. In the remaining set, 26% share homology to predicted proteins of unknown function in other organisms, 3% have no significant homology but have an expressed sequence tag (EST) match, and 8% have no significant homology with other proteins and no EST matches. A number of these may be spurious gene predictions, but some may represent plant-specific genes. The distribution of genes along the chromosome according to homology with known genes and functional categories is shown in Fig. 1. There is no readily discernable pattern of gene distribution along the chromosome according these criteria except that heterochromatin contains a higher proportion of genes with no significant homologies to genes from other organisms and several probable pseudogenes. A searchable list of predicted genes, significant homologies and functional categories can be found in the MATDB database[12].

The number of ESTs matching predicted genes at ⩾90% sequence similarity was used to assess the transcription levels of predicted genes on chromosome 4, and the results of this analysis are shown in Fig. 1. Only 34% of the predicted genes have an EST match, with 6% of the predicted genes matching 75% of the ESTs, revealing the proportion of genes that are highly expressed in the tissues sampled. The highly expressed genes carry out a wide variety of cellular functions; those that encode components of the two photosystems and ancillary proteins are especially prominent. The assessment of expression levels using the EST set derived from messenger RNA made from a mixture of tissues and treatments showed a wide range of expression levels along gene-rich regions of the chromosome arms, with a significant reduction in the proportion of highly expressed genes in the heterochromatic region.

Table 1 Summary of predicted features of chromosome 4

Total non-overlapping sequence	17,385,623 bp
Long arm	14,498,507 bp
Short arm	2,608,702 bp
Centromeric contig	278,414 bp
G+C content	
Overall	36.02%
Exons	44.08%
Introns	33.08%
Intergenic regions	32.23%
Protein-coding DNA	46.07%
Number of encoded proteins	3,744
Integrated sequenced markers	70
Base pairs per gene	4,643
Average exons per gene	5.24 (1–41)
Average exon length	256 bp
Average intron length	188 bp
Known genes (class 1)	8%
Strong similarity to known genes (class 2, >1/3 FASTA self-score)	23%
Similarity to known genes (class 3, FASTA score >150)	32%
Similar to predicted genes (class 4, FASTA score >150)	26%
No similarities, but EST match (class 5)	3%
No similarities, no EST match (class 6, hypothetical proteins)	8%
Predicted chloroplast/mitochondrial targeted	18%
tRNA	81
snRNA	4
LTR retroelements	87
Non-LTR retroelements	103
Repeat unit (for example, inverted, flanking)	816
Repeat region (for example, satellites, tandem repeats)	563

Structural analysis of proteins

Nearly 40% of the proteins studied have at least partially known 3D structures, 35% have significant similarity to at least one SCOP[13] domain, and 11% are completely covered by SCOP domains. Analysis of general protein-folding class occurrence showed that multicellular organisms tend to avoid proteins with mixed α and β structure in favour of all-α proteins. Analysis of the *Arabidopsis* data set indicates that the percentage of proteins possessing multidomain arrangements is gradually increasing with the complexity of multicellular life. A more detailed SCOP analysis revealed a number of folding topologies shared between *Arabidopsis* and other organisms (Table 2). Phosphate loop NTPases, TIM barrel, and Rossmann fold have been previously shown to be among the top five most common folds in all three kingdoms of life[14]. The high occurrence of protein kinases and α/α superhelices is apparently limited to eukaryotic organisms. Conversely, some of the protein architecture frequently found in *C. elegans*, yeast and *E. coli* is rarely found in the *Arabidopsis* dataset. For example, one of the most abundant *C. elegans* folds, the immunoglobulin-like β-sandwich, ranks only 67 in *Arabidopsis*. The C3HC4 RING finger domain and the cytochrome P450 domain rank 5 and 6 on chromosome 4, but rank 16 and 112 in yeast, respectively, and are completely missing in *E. coli*. Although it is possible that other chromosomes code for a disproportionate number of proteins possessing these folds, the established role of P450-containing proteins in a wide range of metabolic pathways and detoxification systems[15] probably accounts for a large proportion of proteins containing this domain.

The RING finger motif is found in functionally diverse proteins, in which facilitating protein–protein interactions, for example in protein degradation[16], is thought to be a common role for the domain[17]. The highly represented EF-hand-like domain is a conserved Ca^{2+}-binding loop found in Ca^{2+}-binding proteins with diverse functions[18]. This domain was encoded by 33 chromosome-4 genes encoding 4 calmodulins, 11 Ca^{2+}-dependent protein kinases, 2 putative Ca^{2+}-binding proteins, 1 caltractin protein, and several other diverse proteins. Calcium-dependent protein kinases containing four EF-hand domains are uniquely found in plants, and the Ca^{2+} cation is a second messenger involved in transducing environmental responses[19], explaining the relative abundance of the domain. Several protein folds unique to *Arabidopsis* have been found, such as in the C-terminal region of glutathione *S*-transferases and in phospholipase C/P1. The functions of these new domains need to be assessed experimentally.

Functional characterization and intergenome comparisons

Genes were assigned to functional categories on the basis of their homology to experimentally characterized genes and detectable domain signatures. The number of genes in each functional cate-

Table 2 The top ten folds in *Arabidopsis* chromosome 4 proteins and their rank in other model organisms

Fold	Percentage of proteins with at least one domain of a given type (rank)			
	Arabidopsis	*C. elegans*	*Yeast*	*E. coli*
Protein kinases, catalytic core	14.6 (1)	9.25 (1)	7.3 (2)	0.08 (180)
α/α superhelix	6.4 (2)	9.25 (1)	3.8 (6)	0.9 (30)
Phosphate-loop NTPases	5.8 (3)	6.5 (2)	10.5 (1)	3.9 (5)
TIM barrel	4.5 (4)	2.6 (13)	5.1 (4)	8.3 (1)
RING finger domain, C3HC4	3.7 (5)	1.9 (16)	1.5 (16)	0 (–)
Cytochrome P450	3.7 (5)	1.8 (17)	0.2 (109)	0 (–)
DNA/RNA-binding 3-helical bundle	3.3 (7)	4 (7)	1.9 (15)	1.8 (14)
7-bladed β-propeller	3.1 (8)	2.9 (11)	6.8 (3)	0.4 (87)
NAD(P)-binding Rossmann fold	2.9 (9)	2.7 (12)	3.6 (7)	6.2 (3)
α/β-Hydrolases	2.6 (10)	3.5 (9)	2.5 (9)	2.5 (9)

gory is shown in Fig. 2a. The large number of genes encoding the cytochrome P450 motif were assigned to the metabolism category because of their well-characterized functions, although their substrates are not yet known. The large proportion of proteins involved in cellular communication, signalling and transcription are typical of higher multicellular eukaryotes. The high proportion of genes involved in defence and disease is due in part to several clusters of leucine-rich repeat (LRR)-resistance specificity genes, and reflects the multitude of proteins for specifically detecting pathogens and the complexity of the systems used by plants to combat abiotic and biotic stresses.

Chromosome-4 genes with their associated functional category were compared with the complete set of translations from *E. coli*, *Saccharomyces cerevisae*, *C. elegans*, *Synechocystis sp.PCC6803* and the *Homo sapiens* non-redundant protein database. The proportion of *Arabidopsis* genes in each functional category with a match in each genome is shown in Fig. 2b. The functional categories of metabolism, energy, transport facilitation and cellular biogenesis contained roughly equal proportions of genes with significant homology (cutoff $\leqslant 10^{-30}$) to genes from all five classes of organisms. A clearly defined eukaryotic set of genes involved in transcription, cell division and DNA synthesis, signal transduction and other categories can be seen. Generally, the highest proportion of matches were with human genes in most functional categories. But the

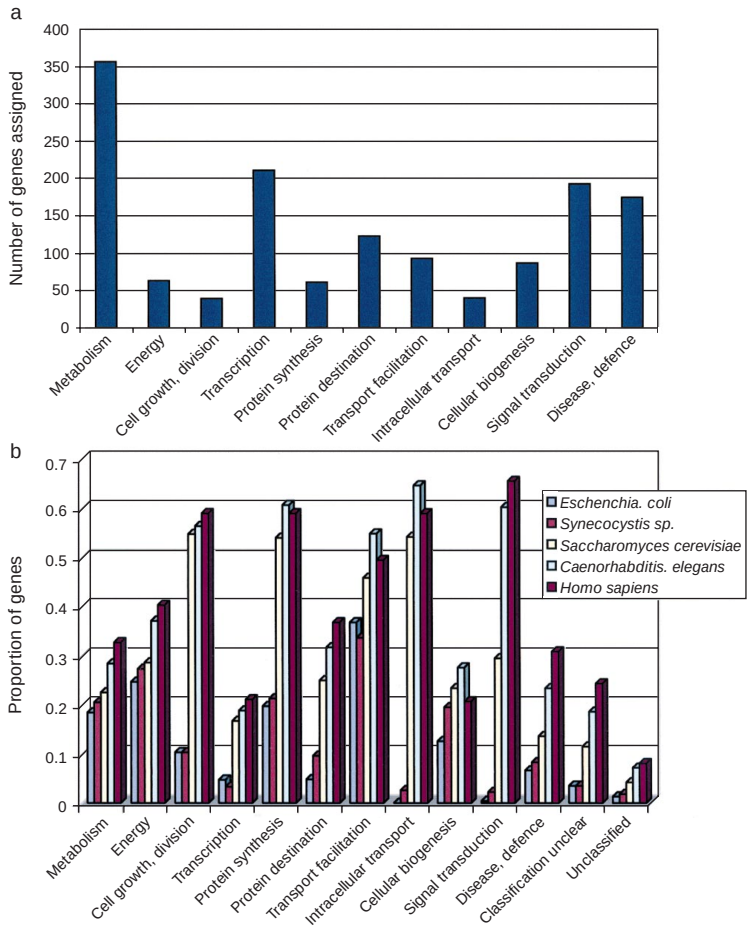

Figure 2 Functional analysis of genes. **a**, The extracted protein sequence was compared against the PIR-International protein database using the FASTA algorithm[50]. On the basis of the significant homologies detected, genes were classified and assigned to functional categories within the MIPS functional catalogue[12]. The number of genes in each category is shown on the *y* axis. **b**, All chromosome-4 genes with associated functional categories were searched against the complete set of translations from *E. coli*, *S. cerevisae*, *C. elegans*, *Synechocystis sp.PCC6803* and the *H. sapiens* non-redundant protein database. The proportion of genes *Arabidopsis* genes with BLASTX matches of $\leqslant 10^{-30}$ in each genome is shown (0.1 = 10%) on the *y* axis.

Table 3 List of selected homologues of vertebrate genes

Arabidopsis gene	Homologue	Identity (%)*	Possible function
AT4g00020	BRCA2 breast cancer	38.9% (126)	Cellular responses to DNA damage
AT4g13870	Werner syndrome	37.4% (190)	Helicase involved in replication fork formation
AT4g38350	Nieman–Pick C disease	42.7% (330)	Intracellular cholesterol transport
AT4g21150	Human ribophorin II	29.7% (405)	Ribosome binding to ER
AT4g20410	Bovine SNAP	24.2% (252)	Intra-Golgi transport
AT4g24880	Human snurportin-1	30.5% (233)	Cap-specific nuclear import receptor
AT4g04210	Human p47 cofactor	30.7% (322)	Adaptor for Golgi-vacuole transport facilitation
AT4g34430	Human 170K subunit of hSWI/SNF complex	43.7% (87)	Chromatin remodelling transcription complex
AT4g32790	Human multiple exotosis 2	32.2% (87)	Skeletal disorder
AT4g27580	Mouse stomatin	31.9% (204)	Membrane-associated ion channel regulator
AT4g08180	Human oxysterol binding protein	28.2% (593)	Feedback regulation of sterol synthesis?
AT4g31480	Human β-COP	49.0% (953)	Non-clathrin membrane trafficking ER-Golgi
AT4g34450	Human γ-COP	49.0% (866)	Non-clathrin membrane trafficking ER-Golgi
AT4g08520	Human ζ-COP	46.3% (121)	Non-clathrin membrane trafficking ER-Golgi
AT4g27890	Human NUDC	43.9% (255)	Nuclear distribution protein
AT4g26600	Human proliferating cell nucleolar antigen p120	48.5% (583)	Cell division
AT4g25840	Human GS1 protein	48.8% (217)	Protein can escape X-chromosome inactivation
AT4g08960	Human PTPA	44.3% (309)	Phosphotyrosyl phosphatase activator
AT4g04930	Human membrane fatty acid desaturase	52.0% (306)	Overexpression of human gene inhibits EGF receptor Biosynthesis, *Drosophila* homologue is involved in spermatogenesis
AT4g19210	Human RNase L inhibitor	74.5% (601)	Antiviral 2-5A RNase L system
AT4g12620	Human replication control protein 1	36.8% (481)	DNA replication

Selected *Arabiderpsis* homologues of vertebrate genes are shown. Similarities were determined using the FASTA algorithm[50]. ER, endoplasmic reticlum.
*Number of amino acids is given in parenthesis.

significant proportion of matches to *C. elegans* and yeast indicates the great antiquity of the last common ancestor of plants and metazoans. At a reduced stringency of $\leqslant 10^{-10}$ (data not shown), a larger proportion of *Arabidopsis* proteins matched those of *Synechocystis* rather than those of *E. coli*, especially in the categories of signal transduction and cellular biogenesis.

Chloroplasts and mitochondria are thought to have originated from ancient symbiotic relationships with photosynthetic eubacteria[20]. The original genomes of these putative endosymbionts are thought to have been greatly reduced by the loss of genes, many of which are now in the nuclear genome, and the encoded proteins are transported into plastids and mitochondria by characteristic amino-terminal signal peptides. To extent our knowledge of organelle function and show the extent of gene transfer from organelle to nucleus, we determined the number and relationships of proteins with the potential for organelle targeting[21]. Approximately 18% of predicted chromosome-4 proteins have a potential N-terminal chloroplast and mitochondrial transit peptides, which are indistinguishable using the available bioinformatics tools. Twenty-six per cent of the proteins with predicted transit peptides are significantly similar to proteins from the cyanobacterium *Synechocystis*, compared with 13% of all predicted proteins, which is consistent with the bias towards *Synechocystis* homologies described above and supports the derivation of chloroplast-targeted nuclear genes from a putative endosymbiont related to present-day *Synechocystis*. The frequent insertions of chloroplast and mitochondrial DNA in the nuclear genome shows the continuation of this process. A higher than average representation of hypothetical proteins (those with no significant similarities to any other protein, and with no EST match) were found in this class, providing some evidence of functionality to the hypothetical peptides, and perhaps indicating that a significant proportion of potential plant-specific genes are targeted to organelles.

Several homologues of human proteins implicated in disease have been found in *Arabidopsis*, including BRCA2, which is linked to breast cancer and thought to be involved in cellular responses to DNA damage (Table 3). The highly conserved homologue of human phosphotyrosyl phosphatase activator (PTPA) suggests that tyrosine phosphorylation may have a role in plants, in which serine/threonine and histidine phosphorylation are the known modes of phospho-transfer in signalling pathways. The homologue of an RNaseL inhibitor gene, which is involved in interferon-mediated virus resistance in humans, indicates a potential role for RNaseL-mediated resistance to plant viruses. These examples illustrate a remarkable sequence conservation of proteins involved in many aspects of cellular function, and permit the transfer of biological knowledge between diverse organisms.

There are significant similarities with genes from different plant groups, notably with the nodulin class of genes[22], which are defined by their expression during development of nodules that establish symbiotic relationships with N-fixing bacteria such as *Rhizobium*. The limitation of symbiotic N-fixing relationships to the Leguminosae may be specified by legume-specific genes, or by the nodule-specific expression of genes common to many plants. The functional analysis of nodulin-like genes in *Arabidopsis* could therefore contribute important information to the study of symbiotic N-fixation.

Multigene families

A significant proportion (12%) of the genes with significant similarities to known proteins are in clusters on genes on the same DNA strand, ranging from the frequent occurrence of pairs of related genes, to a family of 15 contiguous receptor kinase-like proteins (AT4g 18680–18800) which are over 95% identical at the sequence level. Figure 3a shows the distribution of coding-region repeats containing three or more family members on chromosome 4. Most genic repeats are in adjacent regions of the chromosome (shown by orthogonal connecting lines). Many of the clustered genes in the pericentromeric heterochromatin are retroelement polyprotein genes ('T'). Duplications of regions containing a variety of genes were also detected (diagonal lines). Although a wide range of functions are encoded by genes in these clusters, several classes appear to predominate, such as LRR-disease-resistance proteins, putative receptor kinases, Ser/Thr protein kinases, and cytochrome-P450-like proteins. One consequence of this arrangement of genes may be the potential to co-regulate gene expression; for example, the cluster of five small-auxin-upregulated (SAUR) proteins encoded by F11I11 may be co-regulated by auxin. Another consequence is the potential for generating sequence diversity, as shown in the *Cf9* family of LLR-disease-resistance genes in tomato[23], in which multiple-resistance specificities are generated by sequence exchange between gene family members. This could occur in other gene families, generating new cytochrome P450 substrate specificities, or new ligand binding and signalling specificities in families of putative receptor kinase genes.

Four blocks of genes (labelled 1–4 in Fig. 3b) totalling 2.5 Mb are duplicated with significant conservation of sequence between chromosomes 2 and 4, of which two are in an inverted form (Fig. 3b). These data show that collinear duplicated clusters occur

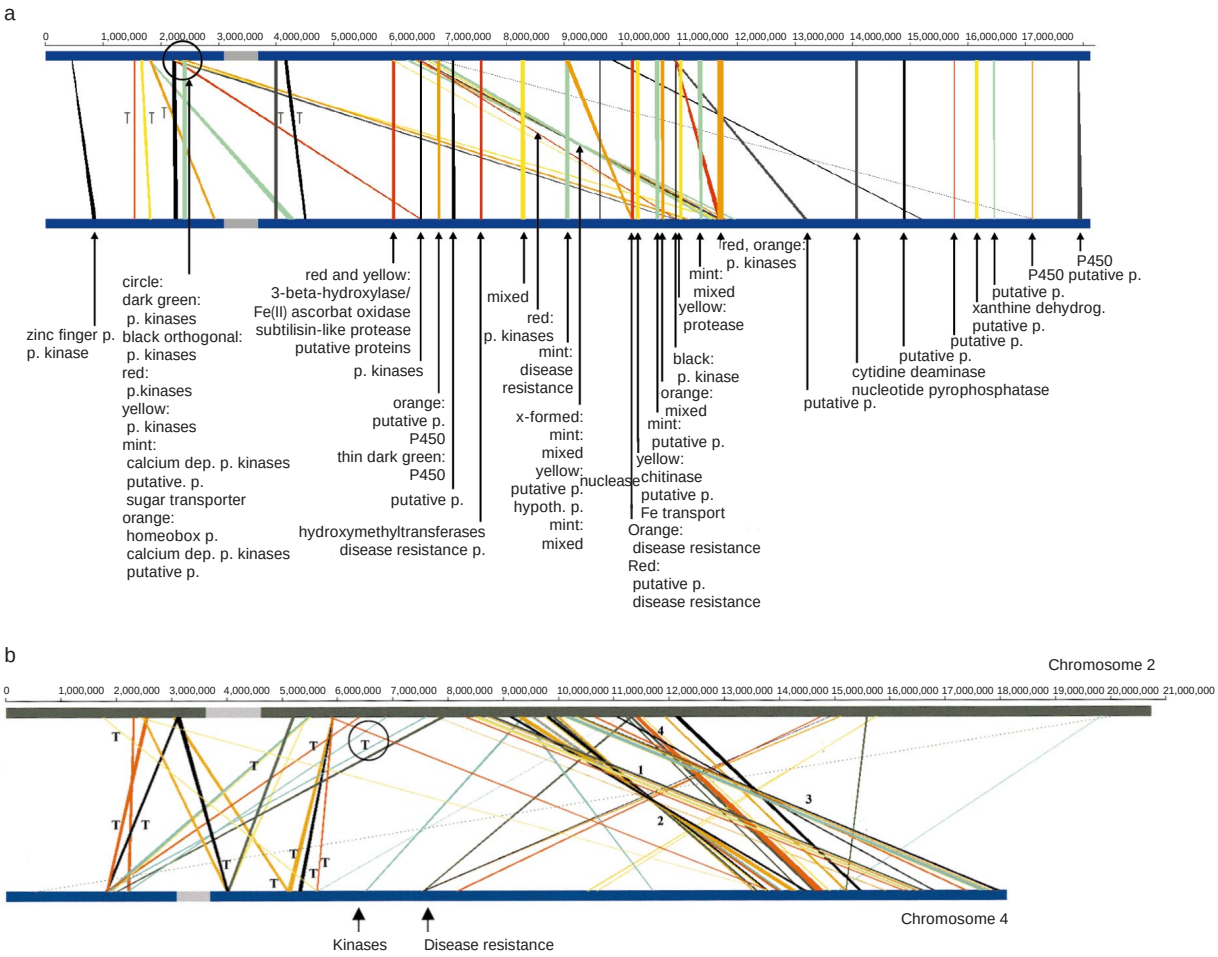

Figure 3 Distribution of multigene families and gene clusters on *Arabidopsis* chromosomes. **a**, Distribution of multigene families and gene clusters on chromosome 4. The output of a BLASTX self-comparison of chromosome 4 sequences (cutoff 10^{-10}) was filtered for coding regions that contained at least three genes in a ⩽30-kb interval that matched a different cluster within a ⩽60-kb interval. Lines are drawn between matching sites. The colours illustrate different matches, and the identity of genes in several of the clusters is shown below the diagram. **b**, Distribution of orthologous gene clusters between chromosome 2 and chromosome 4. The output of a BLASTX comparison between chromosomes 2 and 4 (cutoff 10^{-10}) was filtered for coding regions that contained at least three genes in a ⩽30-kb interval that matched a different cluster within a ⩽60 kb interval on the other chromosome. Lines are drawn between matching sites. The colours illustrate different matches. The identity of genes in two of the clusters are shown. T, transposon.

in plant genomes as well as yeast[24]. There are also large replications of transposon genes within the pericentromeric heterochromatin of chromosomes 2 and 4. Another segmental duplication of 37 contiguous genes between the centre of the long arm of chromosome 4 and the top arm of chromosome 5 has also been identified (data not shown). If extrapolated, about 10–20% of the low-copy regions of the *Arabidopsis* genome could be assigned to duplicated structures, supporting the hypothesis that intragenome duplication is an important evolutionary process. It will be interesting to assess both the full extent of segmental duplication within the *Arabidopsis* genome and its evolutionary and biological significance.

The heterochromatic region

The centromeric heterochromatin region of chromosome 4 has been mapped cytogenetically to a roughly 4 Mb interval, extending into the short arm to yeast artificial chromosome (YAC) CIC7C3 and to YAC CIC11H10 in the long arm (P. Fransz *et al.*, manuscript in preparation). It consists of a central zone of 200 kb of 5S rDNA and 1 Mb of pAL1-rich repeat sequence[25,26] flanked by dispersed retroelements and other repeats. To integrate cytogenetic and sequence analyses for future analysis of centromere function, and to delineate the boundaries of sequence in the centromeric region, low-copy regions from BACs adjacent to the boundaries of the pericentromeric heterochromatin were used, together with the pAL1 repeats, as fluorescent *in situ* hybridization (FISH) probes on pachytene spreads of *Arabidopsis* chromosomes. Figure 4 shows the results of a FISH analysis using sequences from several sequenced BACs. Selected low-copy sequences from F28D6 on the long arm gave FISH signals dispersed in the inner centromeric domain of all chromosomes. T1J1 from the short arm hybridizes close to the heterochromatin boundary. This shows the end of short arm sequence in T1J1 ends at the boundary of the pericentromeric heterochromatin, and F28D6 contains multicopy repeats specific for the inner heterochromatin of all chromosomes. Bacterial artificial chromosome F21I2 low-copy sequences are found immediately adjacent to a cluster of pAL1 repeats, confirming the boundary of this contig at the end of the central zone[26]. Bacterial artificial chromosome T4B21 and pAL1 repeat hybridization reveals the short arm heterochromatin is substantially smaller than the long arm heterochromatin, and provides independent evidence for the position of the central sequence contig adjacent to the central domain defined by pAL1 hybridization. Further FISH analyses using 5S rDNA sequences as a probe confirm the position of sequenced contigs in the centromeric region (data not shown). This analysis confirms the integration of YAC and cytogenetic mapping, reveals the extent of heterochromatin sequenced, and defines the end points of sequenced contigs adjacent to the inner centromeric domain. The sequence gap remaining between F21I2 and F14G16 probably contains the 5S rDNA tract and approximately 1 Mb of pAL1-rich repeat sequence[25].

A repeat motif comprising 22 tandemly arranged copies of a 1950-bp element flanked by two 31-bp direct repeats spans 43,472 bp of BAC T5H22. This is bordered by roughly 80 kb of gene-poor, retroelement (including *Athila* retroelements) enriched sequences. This region probably represents the core of the knob or chromomere, a brightly staining cytogenetic feature that maps to this region of chromosome 4 and that is a prerequisite for chromosome condensation in this region. Heterochromatic knobs composed of tandem arrays of 180-bp repeats can form neocentromeres in maize[27], indicating that homogeneous arrays of repeats can support some of the functions of a centromere, although this has not been observed in *Arabidopsis* to our knowledge.

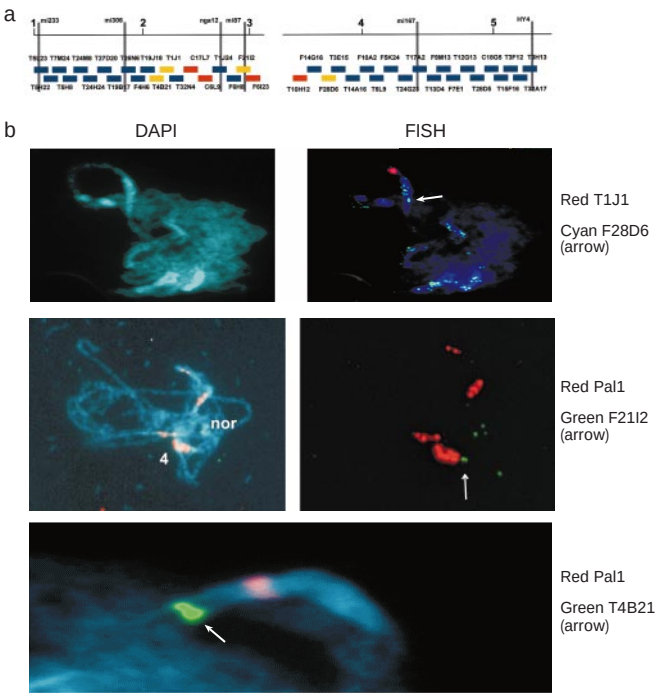

Figure 4 Fluorescent *in situ* hybridization analysis of sequences in the heterochromatic region. **a**, Tiling path of sequenced BACs and integrated genetic markers in the heterochromatic region of chromosome 4. Blue BACs are sequenced and annotated, red BACs are being sequenced, and yellow BACs are sequenced clones used as templates for preparation of FISH probes. **b**, DAPI stained and FISHs of the same pachytene spread are shown in adjacent panels. Arrows point to the position of hybridization signals from DNA derived from yellow-coloured BACs shown in Fig. 4a, and other sequences.

The distribution of repeat classes along chromosome 4, and those shared between chromosomes 2 and 4, is shown in Fig. 5. At the stringency used (BLASTN cutoff $<10^{-40}$, $\geq$150-bp with $\geq$85% identity), two clear patterns of intrachromosomal repeat distribution are apparent. Orthogonal connecting lines reveal frequent replication of repeats in adjacent regions in the heterochromatic region of chromosome 4, notably at the chromomere and at a cluster of sequenced pAL1 repeats, and much less frequently along the length of the chromosome arms (Fig. 5a). Related repeat units, characterized by dispersed repeats, LTR and non-LTR retroelements, MuDR-like sequences, En-like and TNP2-like retroelements are found at higher density towards the central region of heterochromatin, and at lower density along the chromosome arms. *Athila* retroelements are found in increasingly large clusters of up to 40 kb towards the centromeric gap, and are not found outside the heterochromatin zone, consistent with cytogenetic analyses[26]. Chromosomes 2 and 4 share both similar types and distributions of repeats in the heterochromatic region, with the exception of chromomere-associated repeats (Fig. 5b). The short-arm heterochromatin of both chromosomes has a lower repeat density than long-arm heterochromatin.

The central contig of BACs in chromosome 4 is particularly repeat rich, notably a 60 kb array of 177 bp pAL1-like repeats, interrupted by an LTR retroelement on F6H8. This repeat is characteristic of the core region of *Arabidopsis* centromeres defined by FISH[26], and is consistent with the cytogenetic analysis shown in Fig. 4. The extent and distribution of islands of complex sequence in a sea of simple sequence repeats in the sequenced regions in or near the putative centromere of chromosome 4 is reminiscent of the centromere structures defined in *Drosophila*[28] and human[29], although there is no apparent sequence conservation of the known elements.

The analysis of gene density, expression levels and recombination

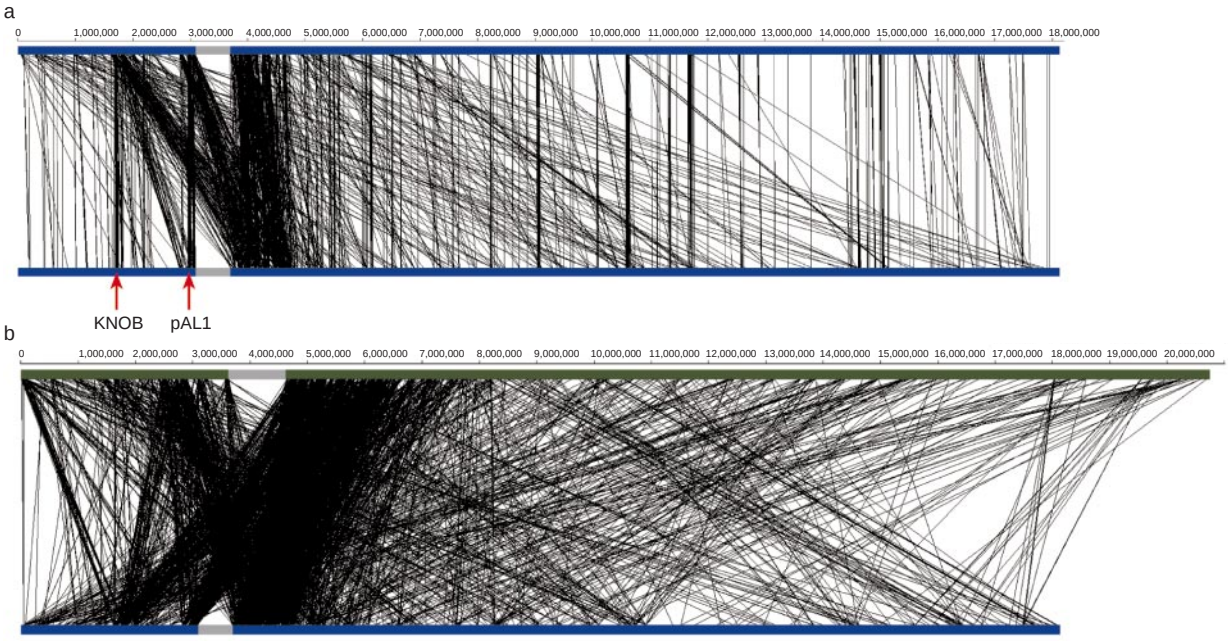

Figure 5 Distribution of sequence repeats. **a**, Sequence comparisons within chromosome 4 were conducted using an all-against-all BLASTN comparison using a cutoff of 10^{-40}. These results were filtered for sequences of $\geq$150 bp and $\geq$85% identity. The black lines depict direct repeats, and the red lines depict inverted repeats. Arrows point to the knob (chromomere) and pAL1 repeat clusters. **b**, Sequence comparisons between chromosomes 2 and 4 were conducted using an all-against-all BLASTN comparison using a cutoff of 10^{-40}. These results were filtered for sequences of $\geq$150bp and $\geq$85% identity.

frequency shows significant differences in these parameters between euchromatin and heterochromatin on chromosome 4 (Fig. 1). The heterochromatin has a roughly tenfold reduction in gene density compared with distal regions of both arms (Fig. 1). A higher proportion of genes with no significant similarity to other genes or probable pseudogenes are found in the heterochromatin as compared with other sequenced regions. Gene expression levels are significantly reduced in this region, as determined by the number of ESTs matching gene models (Fig. 1). Nevertheless, expressed known genes, such as the fatty-acid desaturase encoded by *AT4g04930*, are close to in the central region, far from other highly expressed genes. Four out of the ten genes close to the central region with significant similarity to known genes encode proteins that are involved in DNA replication (*Xenopus* replication protein A1, *AT4g07340*, and rice replication protein A1, *AT4g07450*); and a pair of putative centromeric proteins encoded by *AT4g07890* and *AT4g07900*. *Arabidopsis* genes involved in DNA replication, such those encoding MCM-like replication licence factors, are specifically expressed late in replication; thus, it is interesting to speculate that the chromatin domain surrounding these genes, which are replicated late in the cell cycle, may influence gene expression levels specifically during replication.

Recombination frequencies, deduced from the Col/Ler recombinant inbred lines, vary between 50–200 kb cM^{-1} on the chromosome arms, and decrease to 1,000 kb cM^{-1} in a region from mi233 in the centre of the short arm through the centromeric region to marker HY4 on the long arm[30]. These markers precisely encompass the heterochromatin of the long arm, but extend distally to the chromosomere through a region of lowered gene density and gene expression (Fig. 1). In *Schizosaccharomyces pombe*, recombination suppression and transcriptional silencing are conferred by centromeric-like repeats in the *mat2–mat3* mating-type interval[31], and recombination and gene expression are repressed in the centromere of *S. pombe*[32]. A positive correlation between recombination frequency and gene density in plants can be proposed from the mapping of recombination hot-spots at the 5′ end of transcribed genes in maize[33]. Until now, the limited nucleotide sequencing of DNA repeats in *Arabidopsis* and other plants has revealed little of their higher order organization and overall composition. These results will permit the contribution of specific DNA elements to chromatin condensation, recombination, DNA replication, chromosome pairing, gene expression and transgene stability to be studied with respect to a specific sequence context.

Conclusions

The sequence strategies adopted have resulted in single sequence contigs representing each arm of chromosome 4. Although the long-arm telomere is not represented in the clone contigs analysed, direct approaches such as functional isolation using YACs can be used to obtain telomeric sequences[9]. The unsequenced central heterochromatic region probably contains the 5S rDNA gene cluster and pAL1 repeat clusters and components of the centromere, and obtaining accurate contiguous sequence in this region is a high priority. The sequence of this chromosome, which represents about 17% of the Arabidopsis genome, will be valuable for defining the mechanisms of chromosome maintenance and change. Together with the sequence of chromosome 2 (ref. 34), this work is the first major milestone of the Arabidopsis Genome Initiative, an international collaboration to complete the 130-Mb genome sequence before the end of the year 2000. The functions of a large number of new genes can now be defined systematically, generating the potential for a deeper understanding of plant development and environmental interactions, and providing new knowledge for crop plant improvement.

Note added in proof: The short arm sequence, now extended to join the centromeric contig, is 3,052,402 bp, and the total chromosome sequence is now 17,545,799 bp. □

Methods

Sequence strategy

A clone-based strategy was used to assemble sequence from the Columbia ecotype. Clones representing most of the short arm were identified by the assembly of contigs of BACs from the TAMU[35] and IGF BAC[36] libraries using fingerprint-based approaches[37]. Contigs were identified using RFLP markers mapping to the short arm of the chromosome as hybridization probes, followed by experimental confirmation of predicted overlaps within the contigs. Gaps between contigs were covered by PCR products amplified from genomic DNA and using BAC end sequence data (http://www.tigr.org/tdb/at/atgenome/bac_end_search/bac_end_search.html). Clones representing most of the long are of the chromosome were identified as described[38]. Variations on this approach included iterative hybridization of selected clones to colony filters, and the analysis of AFLP[39] content to establish clone overlaps. Additional clones were identified using the end-sequence database. Clones were also identified by hybridization from the Choi BAC (S. Choi, unpublished data) and LC cosmid libraries (I. Bancroft, unpublished data). To minimize sequence overlaps, tiling-path gaps up to 15–20 kb were introduced into tiling paths, and PCR products spanning the gap were sequenced by primer walking or shotgun sequencing. RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html) was used to define low-copy sequences.

Sequence analysis

Contigs of top and bottom arms were assembled from BAC-sequences submitted by sequencing laboratories. Overlaps, as well as the consistency of the assembly with respect to frameshifts, were inspected, and inconsistencies corrected. The DNA sequence was initially analysed with the gene prediction tools Genemark[40], XGrail[41], Genefinder (P. Green, unpublished data) and GENSCAN[42]. Netplantgene[43] was applied for the prediction of potential splice sites. In all cases, parameters were adjusted to *Arabidopsis* sequence. Similarities to the EMBL, PIR-International databases, *Arabidopsis* and plant ESTs were calculated using the BLAST algorithm[44]. Results from the gene and splice-site prediction tools were combined, and gene models were built incorporating the result of the homology searches using the SPLICE programme (Zaccharia and Mewes, manuscript in preparation). The software accurately predicted 70–80% of internal exons and 50–60% of terminal exons, on the basis of trials with 100 genes with experimentally determined structures (L. Parnell *et al.*, unpublished data). About 95% of potential genes were identified in the genome sequence. TRNAs were identified using tRNA Scan[45]. Analysis for potential chloroplast targeting signal peptides was performed using ChloroP[21] (cut-off 0.50). The gene naming convention is AT (Arabidopsis), chromosome 1–5, g = gene, and a five-digit numbering system from the top of the top arm, in increments of ten. Similarity-based structure prediction and fold assignments were conducted by iterative similarity searches with PSI-BLAST[44] using 5,345 non-redundant structural domain sequences from the SCOP database[13] and 5,464 non-redundant PDB sequences as query. FISH analysis was done as described[26], with minor modifications.

To determine the ranking of protein folds encoded by chromosome 4 genes, searches were conducted against a non-redundant protein sequence database from which sequences enriched in low-complexity, coiled-coil, and transmembrane regions were excluded through application of the SEG[46], COILS[47] and ALOM[48] (version 2 by K. Nakai) programs, respectively. The resulting position-specific score matrices (PSSM) were saved. Chromosome-4 sequences were compared with the PDB and SCOP PSSM libraries using the IMPALA[49] suite of programs. SCOP provides semi-manual classification of individual protein domains from PDB into a four-level hierarchy, starting from the most general folding classes. A more detailed classification was used to group domains with the same arrangement and topological connection of major secondary-structure elements, such as TIM barrels. A protein was considered to possess a known 3D structure if a significant IMPALA hit (E-value < 0.01) with at least one PDB sequence was found. SCOP structural domains were mapped on query sequences through IMPALA hits. If there were several hits, the highest scoring non-overlapping alignments were retained. A sequence was considered to be covered by SCOP alignments completely if the maximal distance between any two adjacent SCOP domains in it was less than 70 amino acids.

Received 5 October; accepted 27 October 1999.

1. Meinke, D. W., Cherry, J. M., Dean, C. D., Rounsley, S. & Koornneef, M. *Arabidopsis thaliana*: a model plant for genome analysis. *Science* **282,** 662–682 (1998).
2. Copenhaver, G. C. & 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).
3. Lister, C. & Dean, C. Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. *Plant J.* **4,** 745–750 (1993).
4. Bevan, M. *et al.* Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. *Nature* **391,** 485–488 (1998).
5. Kotani, H., Nakamura, Y., Sato, S., Kaneko, T., Asamizu, E. *et al.* Structural analysis of *Arabidopsis thaliana* chromosome 5. II. Sequence features of 1,044,062 bp covered by thirteen physically-assigned P1 clones. *DNA Res.* **4,** 291–300 (1997).
6. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018 (1999).
7. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malarial parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
8. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400,** 532–538 (1999).
9. Richards, E. J. & Ausubel, F. M. Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. *Cell* **53,** 127–136 (1988).
10. Copenhaver, G. C. & Pikaard, C. S. RFLP and physical mapping with an rDNA-specific endonuclease reveals that nucleolus organiser regions of *Arabidopsis thaliana* adjoin the telomeres on chromosomes

2 and 4. *Plant J.* **9,** 259–272 (1996).
11. Richards, E. J., Goodman, H. M. & Ausubel, F. M. The centromeric region of *Arabidopsis thaliana* chromosome 1 contains telomere-similar sequences. *Nucleic Acids Res.* **19,** 3351–3357 (1991).
12. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **27,** 44–48 (1999).
13. Hubbard, T. J. P., Ailey, B., Brenner, S. E., Murzin, A. G. & Chothia, C. SCOP: a structural classification database of proteins. *Nucleic Acids Res.* **27,** 254–256 (1999).
14. Gerstein, M. A structural census of genomes: comparing bacterial, eukaryotic and archeal genomes in terms of protein structure. *J. Mol. Biol.* **274,** 562–576 (1997).
15. Mizutani, M., Ward, E. & Ohta, D. Cytochrome P-450 superfamily in *Arabidopsis thaliana*: isolation of cDNAS, differential expression, and RFLP mapping of multiple cytochromes P-450. *Plant Mol. Biology* **37,** 39–52 (1998).
16. Joazeiro, C. A. P. *et al.* The tyrosine kinase negative regulator c-Cbl as a RING-type, E2-dependent ubiquitin-protein ligase. *Science* **286,** 309–312 (1999).
17. Jensen, R. B., Jensen, K. L., Jespersen, H. M. & Skriver, K. Widespread occurrence of a highly conserved RING-H2 zinc finger motif in the model plant *Arabidopsis thaliana*. *FEBS Lett.* **436,** 283–287 (1998).
18. Moncrief, N. D., Kretsinger, R. H. & Goodman, M. Evolution of EF-hand calcium-modulated proteins 2. Domains of several sub-families have diverse evolutionary history. *J. Mol. Evol.* **30,** 522–562 (1991).
19. McAinsh, M. R. & Hetherington, A. M. Encoding specificity in Ca^{2+} signaling systems. *Trends Plant Sci.* **3,** 32–36 (1998).
20. Douglas, S. E. Plastid evolution: origins, diversity, trends. *Curr. Opin. Genet. Dev.* **8,** 655–661 (1998).
21. 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).
22. Gamas, P., de Carvalho Niebel, F., Lescure, N. & Cullimore, J. V. Use of a subtractive hybridisation approach to identify new *Medicago truncatula* genes induced during nodule development. *Mol. Plant Microbe Interaction* **9,** 233–242 (1996).
23. Parniske, M. *et al.* Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the *Cf-4/9* locus in tomato. *Cell* **9,** 821–832 (1997).
24. Mewes, H.-W. *et al. Nature* **387** (Suppl.) 7–8 (1997).
25. Round, E., Flowers, S. K. & Richards, E. J. *Arabidopsis thaliana* centromere regions: genetic map positions and repetitive DNA structure. *Genome Res.* **7,** 1045–1054 (1997).
26. Fransz, P. F. *et al.* Cytogenetics for the model system *Arabidopsis thaliana*. *Plant J.* **13,** 867–876 (1998).
27. Richards, E. J. & Dawe, R. K. Plant centromeres: structure and control. *Curr. Opin. Plant Biol.* **1,** 130–135 (1998).
28. Murphy, T. D. & Karpen, G. H. Localization of centromere function in a *Drosophila* minichromosome. *Cell* **82,** 599–609 (1995).
29. Henning, K. A. *et al.* Human artificial chromosomes generated by modification of a yeast artificial chromosome containing both human alpha satellite and single-copy sequences. *Proc. Natl Acad. Sci. USA* **96,** 592–597 (1999).
30. Copenhaver, G. P., Browne, W. E. & Preuss, D. Assaying genome-wide recombination and centromere functions with *Arabidopsis* tetrads. *Proc. Natl Acad. Sci. USA* **95,** 247–252 (1998).
31. Grewal, S. I. S. & Klar, A. J. S. A recombinationally repressed region between *mat2* and *mat3* loci shares homology to centromeric repeats and regulates directionality of mating type switching in fission yeast. *Genetics* **146,** 1221–1238 (1997).
32. Allshire, R. C., Nimmo, E. R., Ekwall, K., Javerzat, J.-P. & Cranston, G. Mutations derepressing silent centromeric domains in fission yeast disrupt chromosome segregation. *Genes Dev.* **9,** 218–233 (1995).
33. Xu, X. J., Hsai, A.-P., Zhang, L., Nikolau, B. J. & Schnable, P. S. Meiotic recombination breakpoints resolve at high rates at the 5′ end of a maize coding sequence. *Plant Cell* **7,** 2151–2161 (1995).
34. Lin, X. *et al.* Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402,** 761–768 (1999).
35. Choi, S. D., Creelman, R., Mullet, J. & Wing, R. A. Construction and characterisation of a bacterial artificial chromosome library from *Arabidopsis thaliana*. *Weeds World* **2,** 17–20 (1995).
36. Mozo, T. *et al.* A complete BAC-based physical map of the *Arabidopsis thaliana* genome. *Nature Genet.* **22,** 271–275 (1999).
37. Marra, M. *et al.* A map or sequence analysis of the *Arabidopsis thaliana* genome. *Nature Genet.* **22,** 265–270 (1999).
38. Bent, E., Johnson, S. & Bancroft, I. BAC representation of two low-copy regions of the genome of *Arabidopsis thaliana*. *Plant J.* **13,** 849–855 (1998).
39. Vos, P. *et al.* AFLP, a new technique for DNA fingerprinting. *Nucleic Acids Res.* **23,** 4407–4414 (1995).
40. Borodovsky, M. & Peresetsky, A. Deriving non-homogeneous DNA Markov chain models by cluster analysis algorithm minimizing multiple alignment entropy. *Comput. Chem.* **18,** 259–267 (1994).
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.* **88,** 1261–1265 (1991).
42. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
43. Hebsgaard, S. M. *et al.* Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.* **24,** 3439–3452 (1996).
44. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
45. Fichant, G. A. & Burks, C. Identifying potential tRNA genes in genomic DNA sequences. *J. Mol. Biol.* **220,** 659–671 (1991).
46. Wootton, J. C. & Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17,** 149–163 (1993).
47. Lupas, A. N., van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252,** 1162–1164 (1991).
48. Klein, P., Kanehisa, M. & DeLisi, C. The detection and classification of membrane-spanning proteins. *Biochim. Biophys. Acta* **815,** 468–476 (1985).
49. Schäffer, A. A. *et al.* IMPALA: Software to match a protein sequence against a collection of PSI-BLAST-constructed position–specific score matrices. *Bioinformatics*, in the press.
50. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci.* **85,** 2444–2448 (1988).

Acknowledgements

We wish to thank A. Schäffer for invaluable assistance with the IMPALA software and S. Brenner for providing an up-to-date version of the SCOP database. We are grateful to S. Choi for a copy of his large-insert BAC library. Scientists at the John Innes Centre are acknowledged for their help in interpreting gene function. This work was funded in part by Contracts from the European Commission, by the National Science Foundation (NSF) Cooperative Agreement (funded by the NSF, US Department of Agriculture and the US Department of Energy), and by a grant from the USDA NRI Plant Genome Program. Additional support from the Biotechnology and Biological Sciences Research Council, Bundesministerium f. Bildung, Forschung und Technologie, Groupe de Recherche et d'etude des Genomes, Plan Nacional de Investigacion Cientifica y Technica, Westvaco Corporation and D. L. Luke III is gratefully acknowledged.

Correspondence and requests for materials should be addressed to M. Bevan (e-mail: bevan@bbsrc.ac.uk). The sequence and preliminary analysis of clones and PCR products were made available immediately after completion through the MATDB database[12]. The results of computational analyses, including the functional and structural characterization of the protein sequences involved, are available at the PEDANT-pro genome analysis server (http://pedant.mips.biochem.mpg.de). Underlying recombinant clones can be obtained from the NASC (http://www.nasc.ac.uk/). The accession numbers for chromosome 4 are: short arm, AJ270058; long arm, AJ270060.

*** The European Union Arabidopsis Genome Sequencing Consortium:**
K. Mayer[1], C. Schüller[1]†, R. Wambutt[2], G. Murphy[3], G. Volckaert[4], T. Pohl[5], A. Düsterhöft[6], W. Stiekema[7], K.-D. Entian[8], N. Terryn[9], B. Harris[10], W. Ansorge[11], P. Brandt[12]†, L. Grivell[13], M. Rieger[14], M. Weichselgartner[15], V. de Simone[16], B. Obermaier[17], R. Mache[18], M. Müller[19], M. Kreis[20], M. Delseny[21], P. Puigdomenech[22], M. Watson[23], T. Schmidtheini[24], B. Reichert[25], D. Portatelle[26], M. Perez-Alonso[27], M. Boutry[28], I. Bancroft[3], P. Vos[29], J. Hoheisel[30], W. Zimmermann[2], H. Wedler[2], P. Ridley[3], S.-A. Langham[3], B. McCullagh[3], L. Bilham[3], J. Robben[4], J. Van der Schueren[4], B. Grymonprez[4], Y.-J. Chuang[4], F. Vandenbussche[4], M. Braeken[4], I. Weltjens[4], M. Voet[4], I. Bastiaens[4], R. Aert[4], E. Defoor[4], T. Weitzenegger[5], G. Bothe[5], U. Ramsperger[6], H. Hilbert[6], M. Braun[6], E. Holzer[6], A. Brandt[6], S. Peters[7], M. van Staveren[7], W. Dirkse[7], P. Mooijman[7], R. Klein Lankhorst[7], M. Rose[8], J. Hauf[8], P. Kötter[8], S. Berneiser[8], S. Hempel[8], M. Feldpausch[8], S. Lamberth[8], H. Van den Daele[9], A. De Keyser[9], C. Buysshaert[9], J. Gielen[9], R. Villarroel[9], R. De Clercq[9], M. Van Montagu[9], J. Rogers[10], A. Cronin[10], M. Quail[10], S. Bray-Allen[10], L. Clark[10], J. Doggett[10], S. Hall[10], M. Kay[10], N. Lennard[10], K. McLay[10], R. Mayes[10], A. Pettett[10], M.-A. Rajandream[10], M. Lyne[10], V. Benes[11], S. Rechmann[11], D. Borkova[11], H. Blöcker[12], M. Scharfe[12], M. Grimm[12], T.-H. Löhnert[12], S. Dose[12], M. de Haan[13], A. Maarse[13], M. Schäfer[14], S. Müller-Auer[14], C. Gabel[14], M. Fuchs[14], B. Fartmann[15], K. Granderath[15], D. Dauner[15], A. Herzl[15], S. Neumann[15], A. Argiriou[16], D. Vitale[16], R. Liguori[16], E. Piravandi[17], O. Massenet[18], F. Quigley[18], G. Clabauld[18], A. Mündlein[19], R. Felber[19], S. Schnabl[19], R. Hiller[19], W. Schmidt[19], A. Lecharny[20], S. Aubourg[20], F. Chefdor[20], R. Cooke[21], C. Berger[21], A. Montfort[21], E. Casacuberta[21], T. Gibbons[23], N. Weber[24], M. Vandenbol[26], M. Bargues[27], J. Terol[27], A. Torres[27], A. Perez-Perez[27]†, B. Purnelle[28], E. Bent[3], S. Johnson[3], D. Tacon[3], T. Jesse[29], L. Heijnen[29], S. Schwarz[30], P. Scholler[30], S. Heber[30], P. Francs[31], C. Bielke[1], D. Frishman[1], D. Haase[1], K. Lemcke[1], H. W. Mewes[1], S. Stocker[1], P. Zaccaria[1] & M. Bevan[3]

The Cold Spring Harbor, Washington University in St Louis and PE Biosystems Arabidopsis Sequencing Consortium:
R. K. Wilson[32], M. de la Bastide[33], K. Habermann[33], L. Parnell[33], N. Dedhia[33], L. Gnoj[33], K. Schutz[33], E. Huang[33], L. Spiegel[33], M. Sehkon[32], J. Murray[32], P. Sheet[32], M. Cordes[32], J. Abu-Threideh[32], T. Stoneking[32], J. Kalicki[32], T. Graves[32], G. Harmon[32], J. Edwards[32], P. Latreille[32], L. Courtney[32], J. Cloud[32], A. Abbott[32], K. Scott[32], D. Johnson[32], P. Minx[32], D. Bentley[32], B. Fulton[32], N. Miller[32], T. Greco[32], K. Kemp[32], J. Kramer[32], L. Fulton[32], E. Mardis[32], M. Dante[32], K. Pepin[32], L. Hillier[32], J. Nelson[32], J. Spieth[32], E. Ryan[33], S. Andrews[32], C. Geisel[32], D. Layman[32], H. Du[32], J. Ali[32], A. Berghoff[34], K. Jones[34], K. Drone[34], M. Cotton[34], C. Joshu[34], B. Antonoiu[34], M. Zidanic[34], C. Strong[34], H. Sun[34], B. Lamar[34], C. Yordan[34], P. Ma[35]†, J. Zhong[35]†, R. Preston[33], D. Vil[33], M. Shekher[33], A. Matero[33], R. Shah[33], I'K. Swaby[33],

A. O'Shaughnessy[33], M. Rodriguez[33], J. Hoffman[33], S. Till[33], S. Granat[33], N. Shohdy[33], A. Hasegawa[33], A. Hameed[33], M. Lodhi[33]†, A. Johnson[33]†, E. Chen[35]†, M. Marra[33], R. Martienssen[34] & W. R. McCombie[33]

1, GSF-Forschungszentrum f. Umwelt u. Gesundheit, Munich Information Center for Protein Sequences am Max-Planck-Institut f. Biochemie, Am Klopferspitz 18a, D-82152, Germany; 2, AGOWA GmbH, Glienicker Weg 185, D-12489 Berlin, Germany; 3, John Innes Centre, Colney Lane, Norwich NR4 7UH, UK; 4, Katholieke Universiteit Leuven, Laboratory of Gene Technology, Kardinaal Mercierlaan 92, B-3001 Leuven, Belgium; 5, GATC GmbH, Fritz-Arnold Strasse 23, D-78467 Konstanz, Germany; 6, QIAGEN GmbH, Max-Volmer-Str.4, D-40724 Hilden, Germany; 7, CPRO-DLO, Droevendaalsesleeg 1, NL 6700 AA Wageningen, The Netherlands; 8, SRD GmbH, Oberurseler, Str. 43, Oberursel 61440, Germany; 9, Department of Genetics, University of Gent, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium; 10, Wellcome Trust Genome Campus, Hinxton, Cambs. CB10 1SA, UK; 11, EMBL Biochemical Instrumentation Programme, Meyerhofstr. 1, D-69117 Heidelberg, Germany; 12, GBF, Mascheroder Weg 1, D-38124 Braunschweig, Germany; 13, Section for Molecular Biology, Swammerdam Institute of Life Sciences, University of Amsterdam, Kruislaan 318, 1098 SM Amsterdam, The Netherlands; 14, Genotype GmbH, Angelhofweg 39, D-69259 Wilhelmsfeld, Germany; 15, MWG AG Biotech, Anzinger Str. 7, 85554 Ebersberg, Germany; 16, CEINGE and Dipartimento di Biochimica e Biotecnologie Mediche, Università "Frederico II" di Napoli, Via Pansini 5, 80131 Napoli, Italy; 17, MediGenomix GmbH, DNA-Analytics and Genomics, Locharmer Str. 29, D-82152 Planegg/ Martinsried, Germany; 18, Laboratoire Plastes et Différenciation cellulaire, UMR5575, Université Joseph Fourier et CNRS BP53 F-38041 Grenoble, France; 19, Vienna Biocenter, Institute of Microbiology & Genetics, Dr. Bohr-Gasse 9, A-1030 Vienna, Austria; 20, Institute de Biotechnologie des Plantes (IBP), UMR/CNRS 8618, University de Paris-Sud, F-91405 Orsay, France; 21, Lab. Physiologie et Biologie Moleculaire des Plantes, UMR CNRS 5545, Universite de Perpignan, 52 Avenue de Villeneuve, 66860 Perpignan Cedex, France; 22, Department de Genètica Molecular, Institut de Biologia Molecular de Barcelona, CSIC, Barcelona, Spain; 23, Department of Biological Sciences, University of Durham, Durham, DH1 3LE, UK; 24, Microsynth GmbH, Schutzenstr. 15, CH-9436 Balgach, Switzerland; 25, Baseclear, PO Box 1336 Leiden, The Netherlands; 26, Faculté Universitaire des Sciences Agronomiques, Unité de Microbiologie, 6, avenue Maréchal Juin, B-5030 Gembloux, Belgium; 27, Departament de Genètica, University of Valencia, 46100 Burjasot, Valencia, Spain; 28, UCL-FYSA, Croix du Sud, 2-20, B-1348 Louvain-la-Neuve, Belgium; 29, Keygene NV, PO Box 216, 6700 AE Wageningen, The Netherlands; 30, Functional Genome Analysis, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 506, D-69120 Heidelberg, Germany; 31, Institute of Plant Genetics and Crop Plant research (IPK), Correnstr. 3, D-06466 Gatersleben, Germany; 32, Genome Sequencing Center, Washington University, School of Medicine, 4444 Forest Park Blvd., St. Louis, MO 63108, USA; 33, Lita Annenberg Hazen Genome Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; 34, Cold Spring Harbor Plant Biology Group, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; 35, Applied Biosystems, 850 Lincoln Centre Drive, Foster City, California 94494, USA

† *Present addresses: MWG AG Biotech, Anzinger Str. 7, 85554 Ebersberg, Germany (P. Brandt); Sistemas Genomicos SL, Valencia Technology Park, Benjamin Franklin Ave 12, 46980 Parterna, Spain (A. Perez-Perez); BIOMAX Informatics GmbH, Locharmer Str.11, D 82152 Martinsried, Germany (C. Schüller); Axys, 11099 N. Torrey Pines Rd., Suite 160, La Jolla, California 92037, USA (M. Lodhi); Department of Forestry, NC State University, Box 8008, Raleigh, North Carolina 27695, USA (A. Johnson); Celera Genomics, 850 Lincoln Center Drive, Foster City, California 94494, USA (P. Ma, J. Zong, E. Chen).*