195–202. (1998).

35. 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).

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

37. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59,** 24–31 (1999).

38. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26,** 1107–1115 (1998).

39. Hebsgaard, S. M. *et al.* Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.* **24,** 3430–3452 (1996).

40. Huang, X., Adams, M. D., Zhou, H. & Kerlavage, A. R. A tool for analyzing and annotating genomic sequences. *Genomics* **46,** 37–45 (1997).

41. Frishman, D. & Mewes, H.-W. PEDANTic genome analysis. *Trends Genet.* **13,** 415–416 (1997).

42. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).

43. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300,** 1005–1016 (2000).

Supplementary Information is available at *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank K. Mayer and H. Schoof of MIPS for discussions; S. Rhee and E. Huala of TAIR for sequences for the RI markers; and R. Wells for editing the manuscript. This work was funded by National Science Foundation/US Department of Energy/US Department of Agriculture (NSF/DOE/USDA) grants to the SPP Consortium and TIGR.

Correspondence and requests for materials should be addressed to A.T. (e-mail: theo@nature.berkeley.edu) or J.R.E. (e-mail: ecker@salk.edu). The accession members for chromosome 1 are: northern arm, AE005172; southern arm AE005173.

Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*

European Union Chromosome 3 Arabidopsis Sequencing Consortium, The Institute for Genomic Research & Kazusa DNA Research Institute*

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

***Arabidopsis thaliana* is an important model system for plant biologists[1]. In 1996 an international collaboration (the Arabidopsis Genome Initiative) was formed to sequence the whole genome of *Arabidopsis*[2] and in 1999 the sequence of the first two chromosomes was reported[3,4]. The sequence of the last three chromosomes and an analysis of the whole genome are reported in this issue[5–7]. Here we present the sequence of chromosome 3, organized into four sequence segments (contigs). The two largest (13.5 and 9.2 Mb) correspond to the top (long) and the bottom (short) arms of chromosome 3, and the two small contigs are located in the genetically defined centromere[8]. This chromosome encodes 5,220 of the roughly 25,500 predicted protein-coding genes in the genome. About 20% of the predicted proteins have significant homology to proteins in eukaryotic genomes for which the complete sequence is available, pointing to important conserved cellular functions among eukaryotes.**

Chromosome 3 is submetacentric and represents about 20% of the *Arabidopsis* genome. It has been estimated, using yeast artificial chromosome (YAC)-based physical maps[9,10], to be 21–23 Mb long (excluding the centromeric and telomeric regions). We sequenced 330 clones (bacterial artificial chromosomes (BACs)[11,12], P1 clones[13] and transformation-competent artificial chromosomes (TACs)[14]) and eight polymerase chain reaction (PCR) products and assembled them into four contigs representing 23,172,617 base pairs (bp) of non-redundant sequence. The bottom arm contains a residual sequencing gap of around 5 kilobases (kb). Of the approximately 150 and 450 kb of sequence in the two small centromeric contigs, 340 kb (90 and 250 kb, respectively) correspond to high accuracy DNA sequence; the rest consists of unfinished highly repetitive BAC sequences.

For each chromosome arm, the canonical telomeric repeats[15] specific for chromosome ends border a long euchromatic region (~11 and ~7 megabases (Mb) for the top and bottom arms, respectively) characterized by a high and roughly uniform gene density. The gene density then gradually decreases as the retrotransposon density increases towards the peri-centromeric/centromeric heterochromatic region. The top arm contig terminates at the F15D2 BAC clone, which contains at its end a 180-bp tandem repeat characteristic of the *Arabidopsis* centromere[16,17]. The bottom arm contig begins at the F4M19 BAC clone with a 5S ribosomal DNA (rDNA) repeat cluster[18]. The two small centromeric contigs were mapped in between the two arm contigs by tetrad analysis[8]. The relative positions and orientations of these small contigs have not yet been confirmed experimentally. The probable structure of the chromosome 3 centromere is shown in ref. 7 and in Supplementary Information. The size of the genetically defined centromeric region is estimated to be around 1.7 Mb; added to the size of the chromosome arms, this indicates a size of 24 Mb for the whole chromosome. Unexpectedly, the centromeric region contains, in addition to known repetitive elements[7], a block of 40 nearly perfect telomeric repeat units and a single complete rDNA unit (over 99% identical in the 25S, 18S and 5.8S regions). A general description of the characteristics of

Table 1 Features of chromosome 3

(a) The DNA molecule	
Length	23,172,617 bp
Top arm	13,590,268 bp
Bottom arm*	9,582,349 bp
Base composition (%GC)	
Overall	35.4
Coding	44.3
Non-coding	33.0
Number of genes	5,220
Gene density	4.5 kb per gene
Average gene length	1,925 bp
Average peptide length	424 amino acids
Exons	
Number	26,570
Total length	6,654,507 bp
Average per gene	5.1
Average size	250 bp
Introns	
Number	21,350
Total length	3,397,531 bp
Average size	159 bp
Percentage of genes with ESTs†	59.8%
Number of ESTs†	20,732
(b) The proteome	
Total proteins	5,220
Proteins with INTERPRO domains	2,989 (57.8%)
Genes which contain at least one transmembrane domain	1,615 (30.9%)
Genes which contain at least one SCOP domain	1,664 (31.9)%
Secretory pathway default value‡	877
Secretory pathway >0.95 specificity	813
Chloroplast default value	754
Chloroplast >0.95 specificity	420
Mitochondria default value	554
Mitochondria >0.95 specificity	63
Functional classification	
Cellular metabolism	745
Transcription	566
Plant defence	354
Signalling	356
Growth	357
Protein fate	314
Intracellular transport	269
Transport	155
Protein synthesis	148
Total	3,264

* The size of the bottom arm included the two small centromeric contigs (~340 kb).
† EST matches were calculated using a similarity threshold of 90%.
‡ The assignation to secretory pathway, chloroplast and mitochrondria result from a TargetP analysis.

chromosome 3 is available as Supplementary Information.

To annotate chromosome 3 we combined *in silico* gene-finding methods and sequence comparisons with external databases, followed by human inspection. Chromosome 3 contains 5,220 putative genes. A 1.9-kb-long gene (from start to stop codon) is predicted on average every 4.5 kb and contains from 1 (21%) to 78 exons; therefore, 43% of the DNA potentially encodes proteins. The characteristics of the genes predicted on this chromosome are essentially similar to those described for chromosomes 2 (ref. 4) and 4 (ref. 3), and are shown in Table 1(a). To assign unambiguously *Arabidopsis* expressed sequence tags (ESTs) to their cognate genes, a whole-genome comparison was performed and for each EST the best match with the predicted gene was retained. We know that at least 2,714 (53%) of the predicted genes are expressed, because there are corresponding complementary DNAs or ESTs. Of the 39 genes represented by more than 50 ESTs, 18% are located in a 165-kb interval, mainly owing to a cluster of putative lectin genes that has nine members and accounts for 349 ESTs. Alternative splicing has been reported in *Arabidopsis* genes[19–21]. We detected potential alternative splicing for 1–2% of around 1,000 predicted genes with at least three EST and/or messenger RNA matches, which confirms that alternative splicing is rare in chromosome 3 compared with its frequency in mammalian genes[22], at least in the set of ESTs available.

We found two types of duplication, represented either by clustered gene families or by segmental duplications. Eight hundred and thirty seven predicted genes and 47 predicted pseudogenes are members of 306 clustered gene families with between 2 and 23 members each. The percentage of the predicted genes showing a match with an EST is 38%, which is significantly ($P < 10^{-3}$) lower than the percentage of matches between ESTs and the overall predicted genes on chromosome 3 (52%). The low expression of these genes suggests that some of them may be vestigial. Other possibilities such as gene silencing or the acquisition of specific expression patterns could also explain this feature. In most cases it is likely that clustering is due to simple amplification by duplication. When considering only clustered gene families with two members, 86% of the gene pairs are on the same strand. Remnants of very recent duplications (for example, AT3g24870 and AT3g24880) which show over 98% identity (including introns) have been identified. In some cases, the organization of the cluster is much more complex and its evolution is more difficult to trace (for example, AT3g59150 to AT3g59270). In addition to the clustered gene families, there are large segmental duplications between the chromosome 3 sequence and sequences on other chromosomes, such as a 4-Mb duplication between chromosomes 3 and 2 (ref. 7). The analysis of the 45 chromosome 3 clustered gene families in this duplication shows that, in about 80% of cases, the best match is found in the same cluster on chromosome 3, indicating that most of these clustered gene families were probably created after the segmental duplication.

Table 2 Homologues of human genes with best matches on chromosome 3

Human gene	Identity*	Possible function	*Arabidopsis* gene
Acute myeloid leukaemia (DEK)	36% (138 aa)	Chromatin structure	*AT3g48710*
Ataxia telangiectasia (ATM)	38% (1,042 aa)	DNA repair	*AT3g48190*
Retinoblastoma (RB1)	23% (850 aa)	Regulation of apoptotic function	*AT3g12280*
Machado-Joseph (MJD1)	36%(244 aa)	Protein folding	*AT3g54130*
Miller-Dieker lissen. (PAF)	32% (312 aa)	Phospholipase	*AT3g49660*
Myotubular myopathy 1 (MTM1)	35% (454 aa)	Signalling pathway	*AT3g10550*
Coffin-Lowry (RPS6KA3)	47% (323 aa)	Ribosomal protein S6 kinase	*AT3g08720*
Williams-Beuren (ELN)	31% (762 aa)	Composition of elastic fibres	*AT3g22140*
Carnitine deficiency (SLC22A5)	31% (453 aa)	Sodium carnitine cotransporter	*AT3g20660*
Downregulated in adenoma (DRA)	27% (506 aa)	Anion transport	*AT3g12520*

*Amino-acid identity over the homologous region of the gene; the size of the region is given in parentheses.

The characteristics of the 5,220 putative proteins encoded by this chromosome are shown in Table 1(b). Four hundred and twenty seven (8.3%) were already known at least by the presence of an mRNA in the sequence databases. For about 60% of the annotated proteins, we can predict a putative function using sequence similarity criteria and INTERPRO[23] analysis. Results from these two programs do not completely overlap, which further increases the percentage for which a putative function can be assigned. The remaining group consists of proteins predicted by gene-finding programs only and of proteins with similarity to other proteins of unknown function. The distribution of the functional categories is in good agreement with those reported for the *Arabidopsis* genome as a whole[7].

Some gene functions are over-represented in the 5,220 genes predicted on chromosome 3, in part owing to the presence of clustered gene families. Of the 20 *Skp1* homologues found in the *Arabidopsis* genome, eight are on chromosome 3, and six of these are found in two clustered gene families. In yeast *Skp1* is involved in specific protein degradation via the Skp1-Cdc53-F-box pathway, and a similar role has been proposed in *Arabidopsis*[24]. Three of the four nitrilase genes in the *Arabidopsis* genome were found on chromosome 3 as a clustered gene family.

Chromosome 3 harbours some unexpected features, such as a roughly 5-kb chloroplast DNA insertion, the complete rDNA repeat unit and the telomeric repeat, all in the genetically defined centromere. Although the biological analysis of the chromosome 3 genes has to be placed in the context of the whole genome[7], chromosome 3 contains a number of homologues to human disease genes (Table 2) which point to important conserved cellular functions in eukaryotes. Most if not all are involved in basic cellular mechanisms. The analysis of homologues of human genes can provide some clues about the function of these genes in a plant. Chromosome 3 contains a putative orthologue of the human *DEK* gene, which is involved in acute myelogenous leukaemia. This gene induces alterations of the superhelical density of DNA in chromatin[25]. We also detected homologous genes to both human *SKI2W* and yeast *Ski2* (ref. 26). The yeast gene regulates expression of non-poly-A mRNA and has antiviral activities, and it is tempting to attribute an antiviral defence role to this gene in *Arabidopsis*. Inter-genome comparisons will be of great interest for accelerating the understanding of gene function. In this respect *Arabidopsis* will be extremely useful in providing information on genes common to different organisms, owing to the relative ease of screening for mutants in a selected gene. Most of the general features are consistent with those reported for the other chromosomes. For example, the gene density is similar for all the *Arabidopsis* chromosomes. This highly conserved gene density can be explained by both the very high number of segmental duplications[7] and a continuously very high rate of reshuffling during the evolution of the *Arabidopsis* genome. □

Methods

We used two strategies to establish the sequence of chromosome 3. The first was based on the construction of a fine physical map by ordering P1, TAC and BAC clones using DNA markers and clone end sequences[27]. Under the second strategy we isolated seeded clones using STS-selected markers and sequenced them by shotgun sequencing. Seeded clones were then extended in both directions by searching for sequence identities in the BAC end sequence database, which was then cross-examined with BAC fingerprint data. When necessary PCR products were sequenced to fill gaps between two contigs. Individual BAC clones were assembled from the shotgun sequences. Gaps between contigs were closed by primer walking and low-quality regions were finished by resequencing. Individual BAC clone assemblies were checked by restriction digests and the error rate of the final sequence was estimated to be lower than 1 in 3×10^5 by comparing independent sequence overlaps representing ~950 kb. Of the 27 discrepancies found, 10 were sequencing errors and the remainder are probably due to BAC mutations.

The annotation of the sequences was performed as described[3,4,28]. Alignments were computed using a SMITH-WATERMAN[29] algorithm implemented in LASSAP[30] (large scale sequence comparison package) version 1.2.0a.

Received 20 October; accepted 15 November 2000.

1. Meinke, D. W., Cherry, J. M., Dean, C., Rounsley, S. D. & Koornneef, M. *Arabidopsis thaliana*: a model plant for genome analysis. *Science* **282,** 662, 679–682 (1998).
2. Bevan, M. *et al.* Objective: the complete sequence of a plant genome. *Plant Cell* **9,** 476–478 (1997).
3. Mayer, K. *et al.* Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana. Nature* **402,** 769–777 (1999).
4. Lin, X. *et al.* Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana. Nature* **402,** 761–768 (1999).
5. Theologis, A. *et al.* Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana. Nature* **408,** 816–820 (2000).
6. Tabata, S. *et al.* Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana. Nature* **408,** 823–826 (2000).
7. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature* **408,** 796–815 (2000).
8. Copenhaver, G. P. *et al.* Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* **286,** 2468–2474 (1999).
9. Camilleri, C. *et al.* A YAC contig map of *Arabidopsis thaliana* chromosome 3. *Plant J.* **14,** 633–642 (1998).
10. Sato, S. *et al.* A physical map of *Arabidopsis thaliana* chromosome 3 represented by two contigs of CIC YAC, P1, TAC and BAC clones. *DNA Res.* **5,** 163–168 (1998).
11. 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).
12. Mozo, T., Fischer, S., Shizuya, H. & Altmann, T. Construction and characterization of the IGF *Arabidopsis* BAC library. *Mol. Gen. Genet.* **258,** 562–570 (1998).
13. Liu, Y. G., Mitsukawa, N., Vasquez-Tello, A. & Whittier, R. F. Generation of high-quality P1 library of *Arabidopsis* suitable for chromosome walking. *Plant J.* **7,** 351–358 (1995).
14. Liu, Y. G. *et al.* Complementation of plant mutants with large genomic DNA fragments by a transformation-competent artificial chromosome vector accelerates positional cloning. *Proc. Natl Acad. Sci. USA* **96,** 6535–6540 (1999).
15. Richards, E. J. & Ausubel, F. M. Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana. Cell* **53,** 127–136 (1988).
16. Martinez, M., Estelles, A. & Somerville, C. A highly repeated DNA sequence in *Arabidopsis thaliana. Mol. Gen. Genet.* **204,** 417–423 (1986).
17. 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).
18. Fransz, P. *et al.* Cytogenetics for the model system *Arabidopsis thaliana. Plant J.* **13,** 867–876 (1998).
19. Lopato, S. *et al.* atSRp30, one of two SF2/ASF-like proteins from *Arabidopsis thaliana*, regulates splicing of specific plant genes. *Genes Dev.* **13,** 987–1001 (1999).
20. Lazar, G. & Goodman, H. M. The *Arabidopsis* splicing factor SR1 is regulated by alternative splicing. *Plant Mol. Biol.* **42,** 571–581 (2000).
21. Sablowski, R. W. & Meyerowitz, E. M. Temperature-sensitive splicing in the floral homeotic mutant apetala3–1. *Plant Cell* **10,** 1453–1463 (1998).
22. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. Frequent alternative splicing of human genes. *Genome Res.* **9,** 1288–1293 (1999).
23. Apweiler, M. *et al.* InterPro. CCP11 Newsletter 10(2000) 〈http://www.hgmp.mrc.ac.uk/CCP11/newsletter/vol3_4〉.
24. Schouten, J., de Kam, R. J., Fetter, K. & Hoge, J. H. Overexpression of *Arabidopsis thaliana* SKP1 homologues in yeast inactivates the Mig1 repressor by destabilising the F-box protein Grr1. *Mol. Gen. Genet.* **263,** 309–319 (2000).
25. Alexiadis, V. *et al.* The protein encoded by the proto-oncogene DEK changes the topology of chromatin and reduces the efficiency of DNA replication in a chromatin-specific manner. *Genes Dev.* **14,** 1308–1312 (2000).
26. Dangel, A. W., Shen, L., Mendoza, A. R., Wu, L. C. & Yu, C. Y. Human helicase gene SKI2W in the HLA class III region exhibits striking structural similarities to the yeast antiviral gene SKI2 and to the human gene KIAA0052: emergence of a new gene family. *Nucleic Acids Res.* **23,** 2120–2126 (1995).
27. Sato, S. *et al.* A sequence-ready contig map of the top arm of *Arabidopsis thaliana* chromosome 3. *DNA Res.* **6,** 117–121 (1999).
28. Sato, S. *et al.* Structural analysis of *Arabidopsis thaliana* chromosome 3. I. Sequence features of the regions of 4,504,864 bp covered by sixty P1 and TAC clones. *DNA Res.* **7,** 131–135 (2000).
29. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).
30. Glemet, E. & Codani, J. J. LASSAP, a LArge Scale Sequence compArison Package. *Comput. Appl. Biosci.* **13,** 137–143 (1997).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

This work was supported in part by a contract from the European Commission, by The Kazusa DNA Research Institute Foundation, and by the National Science Foundation, the US Department of Energy and the US Department of Agriculture. We thank G. Copenhaver for his contribution to completing the chromosome; M. Marra and M. Sekhon for mapping data; J. Hazan, H. Roest Crollius and M. Katinka for discussions; and system administrators, sequencing facility, bioinformatics department and administrative staff from TIGR, Genoscope and Kazusa Institute.

Correspondence and requests for materials should be addressed to M. Salanoubat (e-mail: salanou@genoscope.cns.fr). The annotated set of chromosome 3 genes is available at http://www.kazusa.or.jp/kaos/, http://www.mips.biochem.mpg.de/ and at http://www.tigr.org. All individual BACs have been submitted to EMBL, GB and DDBJ.

* **European Union Chromosome 3 Arabidopsis Genome Sequencing Consortium**

M. Salanoubat[1], K. Lemcke[2], M. Rieger[3], W. Ansorge[4], M. Unseld[5], B. Fartmann[6], G. Valle[7], H. Blöcker[8], M. Perez-Alonso[9], B. Obermaier[11], M. Delseny[12], M. Boutry[13], L. A. Grivell[14], R. Mache[15], P. Puigdomènech[16], V. De Simone[17], N. Choisne[1], F. Artiguenave[1], C. Robert[1], P. Brottier[1], P. Wincker[1], L. Cattolico[1], J. Weissenbach[1], W. Saurin[1], F. Quétier[1], M. Schäfer[3], S. Müller-Auer[3], C. Gabel[3], M. Fuchs[3], V. Benes[4], E. Wurmbach[4], H. Drzonek[4], H. Erfle[4], N. Jordan[5], S. Bangert[5], R. Wiedelmann[5], H. Kranz[5], H. Voss[5], R. Holland[6], P. Brandt[6,8], G. Nyakatura[6], A. Vezzi[7], M. D'Angelo[7], A. Pallavicini[7], S. Toppo[7], B. Simionati[7], A. Conrad[8], K. Hornischer[8], G. Kauer[8], T.-H. Löhnert[8], G. Nordsiek[8], J. Reichelt[8], M. Scharfe[8], O. Schön[8], M. Bargues[9], J. Terol[9], J. Climent[9], P. Navarro[10], C. Collado[10], A. Perez-Perez[10], B. Ottenwälder[11], D. Duchemin[11], R. Cooke[12], M. Laudie[12], C. Berger-Llauro[12], B. Purnelle[13], D. Masuy[13], M. de Haan[14], A. C. Maarse[14], J.-P. Alcaraz[15], A. Cottet[15], E. Casacuberta[16], A. Monfort[16], A. Argiriou[17], M. Flores[17], R. Liguori[17], D. Vitale[17], G. Mannhaupt[2], D. Haase[2], H. Schoof[2], S. Rudd[2], P. Zaccaria[2], H. W. Mewes[2] & K. F. X. Mayer[2]

The Institute for Genomic Research

Samir Kaul[18], Christopher D. Town[18], Hean L. Koo[18], Luke J. Tallon[18], Jennifer Jenkins[18], Timothy Rooney[18], Michael Rizzo[18], Avram Walts[18], Teresa Utterback[18], Claire Y. Fujii[18], Terrance P. Shea[18], Todd H. Creasy[18], Brian Haas[18], Rama Maiti[18], Dongying Wu[18], Jeremy Peterson[18], Susan Van Aken[18], Grace Pai[18], Jennifer Militscher[18], Patrick Sellers[18], John E. Gill[18], Tamara V. Feldblyum[18], Daphne Preuss[19], Xiaoying Lin[18], William C. Nierman[18], Steven L. Salzberg[18], Owen White[18], J. Craig Venter[20] & Claire M. Fraser[18]

Kazusa DNA Research Institute

T. Kaneko[21], Y. Nakamura[21], S. Sato[21], T. Kato[21], E. Asamizu[21], S. Sasamoto[21], T. Kimura[21], K. Idesawa[21], K. Kawashima[21], Y. Kishida[21], C. Kiyokawa[21], M. Kohara[21], M. Matsumoto[21], A. Matsuno[21], A. Muraki[21], S. Nakayama[21], N. Nakazaki[21], S. Shinpo[21], C. Takeuchi[21], T. Wada[21], A. Watanabe[21], M. Yamada[21], M. Yasuda[21] & S. Tabata[21]

1, Genoscope and CNRS FRE2231, 2 rue G. Crémieux, 91057 Evry Cedex, France; 2, GSF - National Research Center for Environment and Health, Munich Information Center for Protein Sequences, at Max-Planck-Institute for Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany; 3, Genotype GmbH Angelhofweg 39, D-69259 Wilhelmsfeld, Germany; 4, European Molecular Biology Laboratory, Biochemical Instrumentation Program, Meyerhofstrasse 1, D-69117 Heidelberg, Germany; 5, LION Bioscience AG, Im Neuenheimer Feld 515-517, 69120 Heidelberg, Germany; 6, MWG-Biotech AG, Anzinger Strasse 7a, 85560 Ebersberg, Germany; 7, CRIBI, Università di Padova, via G. Colombo 3, Padova 35131, Italy; 8, GBF - German Research Centre for Biotechnology, Dept. of Genome Analysis, Mascheroder Weg 1, D-38124 Braunschweig, Germany; 9, Department of Genetics, University of Valencia, 46100 Burjasot, Spain; 10, Sistemas Genomicos SL, Valencia Technology Park, Benjamin Franklin Ave 12, 46980 Paterna, Spain; 11, MediGenomix GmbH, Lochhamer Strasse 29, 82152 Planegg-Martinsried, Germany; 12, Laboratoire Génome et Développement des Plantes, UMR 5096 CNRS/Université de Perpignan, 52, Avenue de Villeneuve, 66860 Perpignan Cédex; 13, Unité de Biochimie physiologique, University of Louvain, Croix du Sud, 2/20, B-1348 Louvain-la-Neuve, Belgium; 14, Section for Molecular Biology, Swammerdam Institute for Life Sciences, University of Amsterdam, Kruislaan 318, 1098 SM Amsterdam, The Netherlands; 15, Laboratoire Plastes et Différenciation Cellulaire, Université J. Fourier et CNRS, BP 53, 38041, Grenoble Cedex 9, France; 16, Institut de Biologia Molecular de Barcelona, CID-CSIC, Jordi Girona, 18, 08034 Barcelona, Spain; 17, CEINGE and Department of Biochemistry and Medical Biotechnology, University "Federico II" of Napoli, Via Pansini 5, 80131 Napoli, Italy; 18, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA; 19, Department of Molecular Genetics and Cell Biology, University of Chicago, Chicago, Illinois 60637, USA; 20, Celera Genomics Corporation, 45 West Gude Drive, Rockville, Maryland 20850, USA; 21, Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan