Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*

The Kazusa DNA Research Institute, The Cold Spring Harbor and Washington University in St Louis Sequencing Consortium & The European Union Arabidopsis Genome Sequencing Consortium*

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

The genome of the model plant *Arabidopsis thaliana* has been sequenced by an international collaboration, The Arabidopsis Genome Initiative. Here we report the complete sequence of chromosome 5. This chromosome is 26 megabases long; it is the second largest *Arabidopsis* chromosome and represents 21% of the sequenced regions of the genome. The sequence of chromosomes 2 and 4 have been reported previously[1,2] and that of chromosomes 1 and 3, together with an analysis of the complete genome sequence, are reported in this issue[3–5]. Analysis of the sequence of chromosome 5 yields further insights into centromere structure and the sequence determinants of heterochromatin condensation. The 5,874 genes encoded on chromosome 5 reveal several new functions in plants, and the patterns of gene organization provide insights into the mechanisms and extent of genome evolution in plants.

We determined the sequence of chromosome 5 from 403 overlapping bacterial artificial chromosome (BAC), phage (P1) and transformation-competent artificial chromosome (TAC) clones[6–8], comprising two contigs representing the chromosome arms. The order of clones was selected using fingerprint contigs, BAC end sequencing and Southern blotting as described previously[2,9], starting from a set of clones anchored to the genetic map by marker content. Regions of sequence have been published previously as part of the sequence release policy of The Arabidopsis Genome Initiative[10,11]. At present, three centromeric BACs and roughly 7 kilobases (kb) in two regions of significant sequence complexity are still being sequenced.

We assigned 30 sequenced markers unambiguously to chromosome 5 and their order was confirmed on a recombinant inbred map[12]. We sequenced telomeric regions and integrated them with genome sequence using genomic DNA in polymerase chain reactions (gPCR). The sequence of heterchromatic regions flanking the centromere was determined up to and including regions of 180-base-pair (bp) centromeric repeats. The features of chromosome 5 are summarized in Table 1; for a graphical image of chromosome 5 displaying genes, repeats and other features, see Supplementary Information Fig. 1. The total sequenced regions comprise 25,953,409 bp; together with roughly 250 kb of 5S repeat sequence and around 1,000 kb of unsequenced 180-bp repeats[9,13], this yields an estimated chromosome size of 27.2 megabases (Mb). This is less than the 28.4 Mb predicted by physical mapping[9]; the difference is probably due to overlaps of unknown size between clones. The average gene density on chromosome 5 is close to that described for the entire genome[5], with about 1 gene every 4.4 kb (Table 1(a)). Gene density varies along the chromosome, with a lower average gene density close to the centromeric region, similar to that reported for chromosomes 2 and 4 (ref. 14).

The sequence extends from centromeric repeats into telomeric regions, resembling that achieved for the arms of other chromosomes[1–4]. Subtelomeric repeats, characterized by short direct repeats and degenerate telomeric sequences, separated regions of normal gene density from the telomeric repeats (T_3AG_3) proper. Yeast artificial chromosome (YAC) and BAC contigs covering the centromeric region[9,13] have been constructed, and the repeat content and distribution have been analysed in detail[14]. The interspersion pattern of 180-bp repeat clusters was similar to that on the centromere of chromosome 4 (CEN4; ref. 2). We found two clusters of 5S ribosomal DNA flanking the central 180-bp repeat domain in the sequence, confirming the two 5S loci near CEN5 revealed by fluorescence *in situ* hybridization (FISH, Fig. 1) in several ecotypes[15]. The short 5S cluster on the long arm is entirely contained

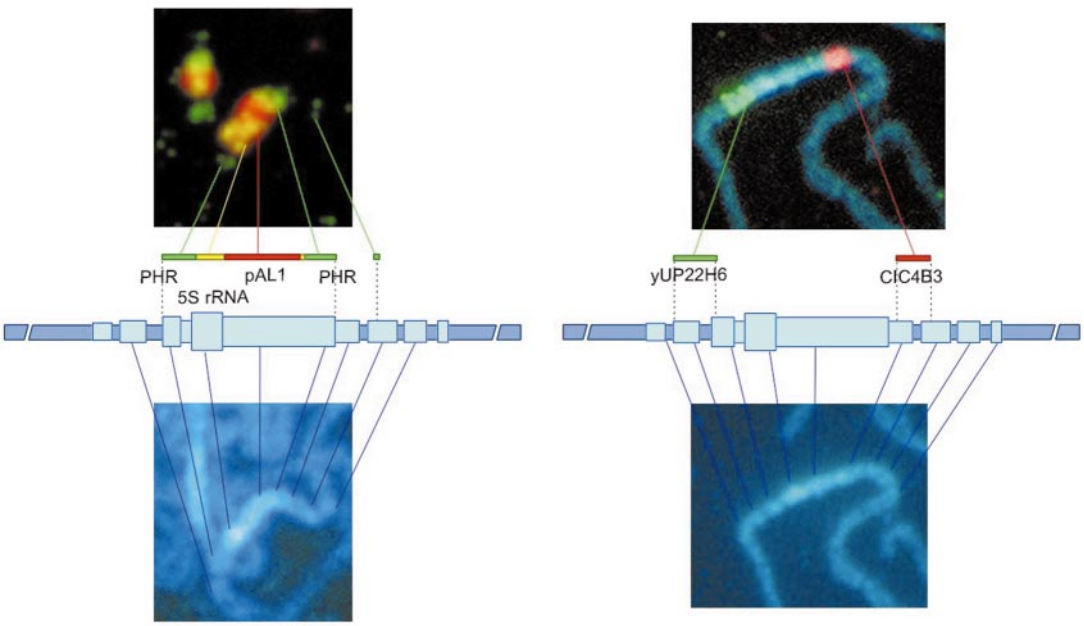

Figure 1 FISH analysis of heterochromatin showing CEN5 features. The positions of 5S rDNA (yellow), pAL1 180-bp repeats (red) and pericentromeric Athila retrolements (green) are shown in the left panel. These features are linked to the physical map using YACs yUP22H6 (green) and CIC4B3 (red) to define the short and long arm, respectively (right panel). DAPI-stained pachytene cells reveal several knobs with variable intensity near CEN5. The heterochromatic knob hk5L is visible as a small island of green hybridization signals (left panel).

within BAC T25B21 and the long cluster is flanked by BACs T26N4 and T3P1. Contiguous sequence extends from both 5S clusters into the flanking regions of the central 180-bp domain bounded by BACs T21M13 and F13C19.

Sequence from the central heterochromatic domain is characterized by a relatively low gene density, increased repeat density (including the centromere-associated *Athila* retrolement) and increased pseudogene density (see Supplementary Information Fig. 1). Expressed genes in this central region of CEN5 include those encoding galactinol synthase and a phosphate/phosphoenolpyruvate translocator precursor. More extensive heterochromatin flanks the 5S clusters, as shown in Fig. 1. Within these heterochromatic tracts, gene density varies substantially, unlike the gradual reduction in gene density observed in centromere-proximal regions of chromosomes 2 and 4. This patchy distribution of heterochromatin can be seen in images stained with 4,6-diamidino-2-phenylindole (DAPI; Fig. 1), consistent with a proposed chromomere (differentially stained chromosomal region) model[29].

On the long arm, retroelements are found in patches that correlate with heterochromatin features. FISH analysis with a BAC probe containing Athila retroelements revealed a heterochromatic knob, named hk5L, in the region of YAC CIC5B3, which underlies the sequence containing the repeats (Fig. 2). There were eight DNA transposons and eighteen retrotransposons in the region corresponding to hk5L. We found 30 tandem repeats of a 2,200-bp sequence, with an insertion of *En-Spm*-like transposon. This 2,200-bp element had no significant similarity to the 1,950-bp tandem repeat element in the heterochromatic knob hk4S on chromosome 4 (refs 16, 17) and did not include the terminal direct repeat structure that was found in the 1,950-bp element. Tandem repeats are generally associated with heterochromatic domains, such as the nucleolar organizers, heterochromatic knobs and pericentromeric heterochromatin[16–18]. This structural conservation indicates the importance of tandem repeat clusters in heterochromatin formation, and suggests that this sequence organization, rather than sequence content, is important in chromosome condensation.

Chromosome 5 contains 4,110 genes encoding proteins of predicted function. Proteins involved in metabolism (21.1%), transcription (18.6%) and defence (11.9%) comprise the major classes (Table 1(b)), and these proportions are consistent with similar analysis of the whole genome[5]. Chromosome 5 contains 37 families of genes encoding proteins of predicted function, present as tandem arrays of more than three members, consistent with the high proportion of gene families observed in the other chromosomes. The largest cluster contains an array consisting of five germin (oxalate oxidase) genes, four receptor-like kinase genes, a single transporter gene and three acyl transferase genes (one of which is interrupted by a retroelement), followed by another nine germin genes.

A number of genes on chromosome 5 exhibit a high degree of overall similarity to genes of known function in other organisms that have not been previously identified in *Arabidopsis*. These include the gene for the minD septum-site determining protein, which is highly conserved in *Chlorella*, *Synechocystis* and *Escherichia coli*. It is required, together with the minC and MinE proteins, for proper placement of the septum in cell division[19]. It may be involved in organelle replication, although it is not reliably predicted to possess a transit peptide. A protein very similar to the product of the *Drosophila* separation anxiety gene encodes a possible *N*-acetyl transferase. Mutation of the *Drosophila* and human genes impairs chromosome spindle assembly and formation, leading to non-disjunction at first meiosis, which causes behavioural differences in flies and humans. Chromosome 5 encodes a homologue of Notchless[20], which modifies Notch activity in cell signalling that determines cell fate in *Drosophila*. This suggests related functions in signal transduction in plants, although proteins related to Notch are not found in *Arabidopsis*[5].

Eighty-eight genes on chromosome 5 have high overall similarities ($E < 10^{-15}$) to the 289 genes involved in human disease syndromes established for comparison with the *Drosophila* genome[21]. Most of these are also highly conserved between *Drosophila* and *Caenorhabditis elegans*, revealing a significant potential for *Arabidopsis* biology to contribute to our knowledge of human disease conditions. Several of these genes belong to classes

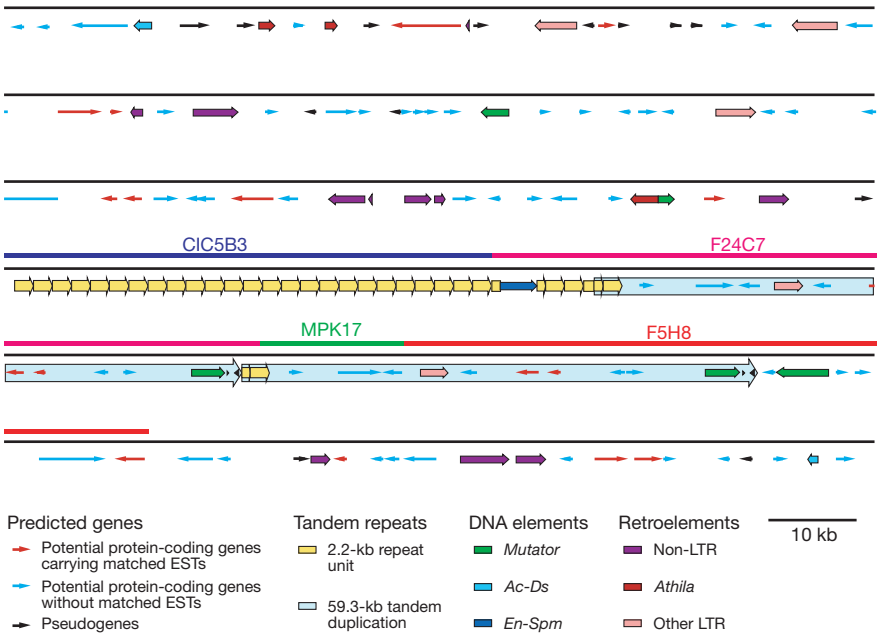

Figure 2 Sequence features of the heterochromatic knob region. The diagram shows a 0.6-Mb region surrounding the heterochromatic knob. Predicted protein coding genes are shown as thin arrows. Tandem repeats and transposons are indicated by colour-coded block arrows.

Table 1 Features of chromosome 5

(a) The DNA molecule	
Length	25,953,409 bp
Top arm	11,132,192 bp
Bottom arm	14,803,217 bp
Base composition (%GC)	
Overall	34.5
Coding	44.1
Non-coding	32.5
Number of genes	5,874
Gene density	4.4 kb per gene
Average gene length	1,974 bp
Average peptide length	429 amino acids
Exons	
Number	31,226
Total length	7,571,013 bp
Average per gene	5.3
Average size	242 bp
Introns	
Number	25,352
Total length	4,030,045 bp
Average size	159 bp
Percentage of genes with ESTs	61.4%
Number of ESTs	22,885
(b) The proteome	
Total proteins	5,874
With INTERPRO domains	3,136 (53.4%)
Proteins containing ≥1 transmembrane domain	1,940 (33.0%)
Proteins containing ≥1 SCOP domain	2,121 (36.1%)
Secretory pathway (default value)	1,014 (17.3%)
Secretory pathway >0.95 specificity	964 (16.4%)
Chloroplast (default value)	887 (15.1%)
Chloroplast >0.95 specificity	475 (8.1%)
Mitochondria (default value)	627 (10.7%)
Mitochondria >0.95 specificity	65 (1.1%)
Functional classification	
Cellular metabolism	868 (21.1%)
Transcription	763 (18.6%)
Plant defence	490 (11.9%)
Signalling	420 (10.2%)
Growth	469 (11.4%)
Protein fate	395 (9.6%)
Intracellular transport	334 (8.1%)
Transport	206 (5.0%)
Protein synthesis	165 (4.0%)
Total	4,110

encoding proteins of deeply conserved function, such as DNA excision repair genes (implicated in xeroderma pigmentosum), a retinis pigmentosa RPGR gene homologue encoding a characterized *Arabidopsis* UVB resistance gene, and ATP-dependent copper transporters (implicated in Wilson's and Menke's diseases). In the latter case the *Arabidopsis* homologues are more similar to the human counterpart than to the *Drosophila, C. elegans* or yeast homologues.

The 5,874 predicted genes on chromosome 5 include many genes with significant similarity to genes from other organisms, including human. This indicates the potential utility of *Arabidopsis* sequence in applications beyond crop plant improvement. The large-scale application of functional genomics tools is required for the systematic identification of the cellular roles of 1,764 predicted genes encoding proteins of no known function. The community of *Arabidopsis* researchers possesses the organization and techniques[22] to accomplish this task efficiently. □

Methods

We used a clone-based strategy to assemble sequence from the Columbia ecotype as described[2,9]. We assessed sequence accuracy by comparison of overlapping sequence of adjacent BACs, and 15 mismatches were found and corrected. A 233-kb region was sequenced by two laboratories independently to assess accuracy and consistency of assembly, and we found no differences. We analysed sequence assemblies for frame shifts and corrected them if necessary. The DNA sequence was analysed as described[2] except that the gene-prediction tools Genemark.hmm[23] and GENSCAN-2 (ref. 24) were used. We integrated the output of a combination of gene-prediction programs with predicted exon–intron structures[25] to derive polypeptide sequences. Similarities between the encoded polypeptides and other proteins were established and classified using BLAST[26] and INTERPRO[27] and structural motifs classified using PEDANT[28]. We determined the predicted functions of genes by significant homology with genes of known function from other organisms.

Received 20 October; accepted 15 November 2000.

1. Lin, X. *et al.* Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402,** 761–768 (1999).
2. Mayer, K. *et al.* Sequence and analysis of chromsome 4 of the plant *Arabidopsis thaliana*. *Nature* **402,** 769–777 (1999).
3. Theologis, A. *et al.* Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature* **408,** 816–820 (2000).
4. Salanoubat, M. *et al.* Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature* **408,** 820–822 (2000).
5. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
6. 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).
7. Lui, Y.-G., Mitsukawa, N., Vazquez-Tello, A. & Whittier, R. F. Generation of a high-quality P1 library of *Arabidopsis* suitable for chromosome walking. *Plant J.* **7,** 351–358 (1995).
8. Lui, 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).
9. Kotani, H., Hosouchi, T. & Tsuruoka, H. Structural analysis and complete physical map of *Arabidopsis thaliana* chromosome 5 including centromeric and telomeric regions. *DNA Res.* **6,** 381–386 (1999).
10. Sato, S. *et al.* Structural analysis of *Arabidopsis thaliana* chromosome 5. I. Sequence features of the 1.6 Mb regions covered by twenty physically assigned P1 clones. *DNA Res.* **4,** 215–230 (1997).
11. Sato, S. *et al.* Structural analysis of *Arabidopsis thaliana* chromosome 5. X. Sequence features of the regions of 3,076,755 bp covered by sixty P1 and TAC clones. *DNA Res.* **7,** 31–63 (2000).
12. Lister, C. & Dean, C. Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. *Plant J.* **4,** 745–750 (1993).
13. Tutois, S. *et al.* Structural analysis and physical mapping of a pericentromeric region of chromosome 5 of *Arabidopsis thaliana*. *Chrom. Res.* **7,** 143–156 (1999).
14. Copenhaver, G. P. *et al.* Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* **286,** 2468–2474 (1999).
15. Fransz, P. F. *et al.* Cytogenetics for the model system *Arabidopsis thaliana*. *Plant J.* **13,** 867–876 (1998).
16. Fransz, P. F. *et al.* Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: structural organisation of heterochromatic knob and centromere region. *Cell* **100,** 367–376 (2000).
17. The Cold Spring Harbor, Washington University in St Louis Genome Sequencing Centre and PE Biosystems Arabidopsis Genome Sequencing Consortium. The complete sequence of a heterochromatic island from a higher eukaryote. *Cell* **100,** 377–386 (2000).
18. Ananiev, E. V., Philips, R. L. & Rines, H. W. Complex structures of knob DNA on maize chromosome 9: retroelement invasion into heterochromatin. *Genetics* **149,** 2025–2037 (1998).
19. Wakasugi, T. *et al.* Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulgaris*: the existence of genes possibly involved in chloroplast division. *Proc. Natl Acad. Sci. USA* **94,** 5967–5972 (1997).
20. Royet, J., Bouwmeester, T. & Cohen, S. M. Notchless encodes a novel WD40 repeat containing protein that modulates notch signalling activity. *EMBO J.* **17,** 7351–7360 (1998).
21. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).
22. Meinke, D. W., Cherry, J. M., Dean, C., Rounsley, S. D. & Koornneef, M. *Arabidopsis thaliana*: A model plant for genome analysis. *Science* **282,** 662–681 (1998).
23. 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).
24. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
25. 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).
26. 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).
27. Apweiler, R. *et al.* INTERPRO. *CCP 11 Newsletter* **10** (cited March 2000) ⟨http://www.ebi.ac.uk/interpro/⟩ (2000).
28. Frishman, D. & Mewes, H.-W. PEDANTic genome analysis. *Trends Genet.* **13,** 415–416 (1997).
29. Lima-de-Farier, A. *Molecular Evolution and Organisation of the Chromosome* (Elsevier, Amsterdam, 1983).

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 by the Kazusa DNA Research Institute Foundation, the National Science Foundation (NSF), the US Department of Agriculture (USDA) and the US Department of Energy (DOE), the USDA NRI Plant Genome Program, and the European Commission. Additional support came from the BBSRC (Biotechnology and Biological Sciences Research Council), GSF-Forschungszentrum f. Umwelt u. Gesundheit, BMBF (Bundesministerium f. Bildung, Forschung und Technologie), Plant Research International, Wageningen, Westvaco Corporation and David L. Luke III.

Correspondence and requests for materials should be addressed to M.B. (e-mail: michael.bevan@bbsrc.ac.uk). The annotated set of chromosome 5 genes is available at

http://www.kazusa.or.jp/kaos/, http://www.mips.biochem.mpg.de/proj/thal/ and http://www.tigr.org/tdb/ath1/htmls/ath1.html.

Kazusa DNA Research Institute
S. Tabata[1], T. Kaneko[1], Y. Nakamura[1], H. Kotani[1], T. Kato[1], E. Asamizu[1], N. Miyajima[1], S. Sasamoto[1], T. Kimura[1], T. Hosouchi[1], K. Kawashima[1], M. Kohara[1], M. Matsumoto[1], A. Matsuno[1], A. Muraki[1], S. Nakayama[1], N. Nakazaki[1], K. Naruo[1], S. Okumura[1], S. Shinpo[1], C. Takeuchi[1], T. Wada[1], A. Watanabe[1], M. Yamada[1], M. Yasuda[1] & S. Sato[1]

The Cold Spring Harbor and Washington University Sequencing Consortium
M. de la Bastide[2], E. Huang[2], L. Spiegel[2], L. Gnoj[2], A. O'Shaughnessy[2], R. Preston[2], K. Habermann[2], J. Murray[3], D. Johnson[3], T. Rohlfing[3], J. Nelson[3], T. Stoneking[3], K. Pepin[3], J. Spieth[3], M. Sekhon[3], J. Armstrong[2], M. Becker[3], E. Belter[3], H. Cordum[3], M. Cordes[3], L. Courtney[3], W. Courtney[3], M. Dante[3], H. Du[3], J. Edwards[3], J. Fryman[3], B. Haakensen[3], E. Lamar[3], P. Latreille[3], S. Leonard[3], R. Meyer[3], E. Mulvaney[3], P. Ozersky[3], A. Riley[3], C. Strowmatt[3], C. Wagner-McPherson[3], A. Wollam[3], M. Yoakum[3], M. Bell[2], N. Dedhia[2], L. Parnell[2], R. Shah[2], M. Rodriguez[2], L. Hoon See[2], D. Vil[2], J. Baker[2], K. Kirchoff [2], K. Toth[2], L. King[2], A. Bahret[2], B. Miller[2], M. Marra[3], R. Martienssen[4], W. R. McCombie[2] & R. K. Wilson[3]

The European Union Arabidopsis Genome Sequencing Consortium
G. Murphy[5], I. Bancroft[5], G. Volckaert[6], R. Wambutt[7], A. Düsterhöft[8], W. Stiekema[9], T. Pohl[10], K.-D. Entian[11], N. Terryn[12], N. Hartley[5], E. Bent[5], S. Johnson[5], S.-A. Langham[5], B. McCullagh[5], J. Robben[6], B. Grymonprez[6], W. Zimmermann[7], U. Ramsperger[8], H. Wedler[8], K. Balke[8], E. Wedler[8], S. Peters[9], M. van Staveren[9], W. Dirkse[9], P. Mooijman[9], R. Klein Lankhorst[9], T. Weitzenegger[10], G. Bothe[10], M. Rose[11], J. Hauf[11], S. Berneiser[11], S. Hempel[11], M. Feldpausch[11], S. Lamberth[11], R. Villarroel[12], J. Gielen[12], W. Ardiles[12], O. Bents[13], K. Lemcke[13], G. Kolesov[13], K. Mayer[13], S. Rudd[13], H. Schoof[13], C. Schueller[13], P. Zaccaria[13], H. W. Mewes[13] & M. Bevan[5]

Institute of Plant Genetics and Crop Plant Research (IPK)
P. Fransz[14]

1, Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292, Japan; 2, Lita Annenberg Hazen Genome Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; 3, Genome Sequencing Center, Washington University, School of Medicine, 4444 Forest Park Boulevard, St. Louis, Missouri 63108, USA, 4, Plant Biology Group, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; 5, John Innes Centre, Colney Lane, Norwich NR4 7UH, UK; 6, Katholieke Universiteit Leuven, Laboratory of Gene Technology, Kardinaal Mercierlaan 92, B-3001 Leuven, Belgium; 7, AGOWA GmbH, Glienicker Weg 185, D-12489 Berlin, Germany; 8, QIAGEN GmbH, Max-Volmer-Strasse 4, D-40724 Hilden, Germany; 9, Greenomics, Plant Research International, Droevendaalsesleeg 1, NL 6700 AA Wageningen, The Netherlands; 10, GATC GmbH, Fritz-Arnold Strasse 23, D-78467 Konstanz, Germany; 11, SRD GmbH, Oberurseler Strasse 43, Oberursel 61440, Germany; 12, Department for Plant Genetics,(VIB), University of Gent, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium; 13, GSF-Forschungszentrum f. Umwelt u. Gesundheit, Munich Information Center for Protein Sequences am Max-Planck-Institut f. Biochemie, Am Klopferspitz 18a, D-82152, Germany; 14, Institute of Plant Genetics and Crop Plant Research (IPK), Correnstrasse 3, D-06466 Gatersleben, Germany