ELSEVIER

Journal of Biotechnology 78 (2000) 281–292

JOURNAL OF Biotechnology

www.elsevier.com/locate/jbiotec

Progress in *Arabidopsis* genome sequencing and functional genomics

R. Wambutt [a], G. Murphy [b], G. Volckaert [c], T. Pohl [d], A. Düsterhöft [e], W. Stiekema [f], K.-D. Entian [g], N. Terryn [h], B. Harris [i], W. Ansorge [j], P. Brandt [k,1], L. Grivell [l], M. Rieger [m], M. Weichselgartner [n], V. de Simone [o], B. Obermaier [p], R. Mache [q], M. Müller [r], M. Kreis [s], M. Delseny [t], P. Puigdomenech [u], M. Watson [v], T. Schmidtheini [w], B. Reichert [x], D. Portatelle [y], M. Perez-Alonso [z], M. Boutry [aa], I. Bancroft [bb], P. Vos [cc], J. Hoheisel [dd], W. Zimmermann [a], H. Wedler [a], P. Ridley [b], S.-A. Langham [b], B. McCullagh [b], L. Bilham [b], J. Robben [c], J. Van der Schueren [c], B. Grymonprez [c], Y.-J. Chuang [c], F. Vandenbussche [c], M. Braeken [c], I. Weltjens [c], M. Voet [c], I. Bastiaens [c], R. Aert [c], E. Defoor [c], T. Weitzenegger [d], G. Bothe [d], U. Ramsperger [e], H. Hilbert [e], M. Braun [e], E. Holzer [e], A. Brandt [e], S. Peters [f], M. van Staveren [f], W. Dirkse [f], P. Mooijman [f], R. Klein Lankhorst [f], M. Rose [g], J. Hauf [g], P. Kötter [g], S. Berneiser [g], S. Hempel [g], M. Feldpausch [g], S. Lamberth [g], H. Van den Daele [h], A. De Keyser [h], C. Buysshaert [h], J. Gielen [h], R. Villarroel [h], R. De Clercq [h], M. Van Montagu [h], J. Rogers [i], A. Cronin [i], M. Quail [i], S. Bray-Allen [i], L. Clark [i], J. Doggett [i], S. Hall [i], M. Kay [i], N. Lennard [i], K. McLay [i], R. Mayes [i], A. Pettett [i], M.-A. Rajandream [i], M. Lyne [i], V. Benes [j], S. Rechmann [j], D. Borkova [j], H. Blöcker [k], M. Scharfe [k], M. Grimm [k], T.-H. Löhnert [k], S. Dose [k], M. de Haan [l], A. Maarse [l], M. Schäfer [m], S. Müller-Auer [m], C. Gabel [m], M. Fuchs [m], B. Fartmann [n], K. Granderath [n], D. Dauner [n], A. Herzl [n], S. Neumann [n], A. Argiriou [o], D. Vitale [o], R. Liguori [o], E. Piravandi [p], O. Massenet [q], F. Quigley [q], G. Clabauld [q], A. Mündlein [r], R. Felber [r], S. Schnabl [r],

* Corresponding author. Tel.: +44-1603-452835; fax: +44-1603-505725.

E-mail address: bevan@bbsrc.ac.uk (M. Bevan)

[1] Present address. MWG AG Biotech, Anzinger Str. 7, 85554 Ebersberg, Denmark.

[2] Present address. Sistemas Genomicos SL, Valencia Technology Park, Benjamin Franklin Ave 12, 46980 Parterna, ES.

[3] Present address. BIOMAX Informatics GmbH, Locharmer Str.11, D 82152 Martinsried, DE.

PII: S0168-1656(00)00195-4

R. Hiller [r], W. Schmidt [r], A. Lecharny [s], S. Aubourg [s], I. Gy [s], R. Cooke [t], C. Berger [t], A. Monfort [u], E. Casacuberta [u], T. Gibbons [v], N. Weber [w], M. Vandenbol [y], M. Bargues [z], J. Terol [z], A. Torres [z], A. Perez-Perez [z,2], B. Purnelle [aa], E. Bent [b], S. Johnson [b], D. Tacon [b], T. Jesse [bb], L. Heijnen [bb], S. Schwarz [cc], P. Scholler [cc], S. Heber [cc], C. Bielke [dd], D. Frishmann [dd], D. Haase [dd], K. Lemcke [dd], H.W. Mewes [dd], S. Stocker [dd], P. Zaccaria [dd], K. Mayer [dd], C. Schüller [dd,3], M. Bevan [b,*]

[a] *AGOWA GmbH, Glienicker Weg 185, D-12489 Berlin, Germany*
[b] *John Innes Centre, Colney Lane, Norwich NR4 7UH, UK*
[c] *Katholieke Universiteit Leuven, Laboratory of Gene Technology, Kardinaal Mercierlaan 92, B-3001 Leuven, Belgium*
[d] *GATC GmbH, Fritz-Arnold Strasse 23, D-78467 Konstanz, Germany*
[e] *QIAGEN GmbH, Max-Volmer-Str. 4, D-40724 Hilden, Germany*
[f] *CPRO-DLO, Droevendaalsesleeg 1, NL 6700 AA Wageningen, The Netherlands*
[g] *Institut für Mikrobiologie, Marie-Curie-Str. 9, D-60439 Frankfurt/M., Germany*
[h] *Department of Genetics, University of Gent, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium*
[i] *Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK*
[j] *EMBL Biochemical Instrumentation Programme, Meyerhofstr. 1, D-69117 Heidelberg, Germany*
[k] *GBF, Mascheroder Weg 1, D-38124 Braunschweig, Germany*
[l] *Section for Molecular Biology, Swammerdam Institute of Life Sciences, University of Amsterdam, Kruislaan 318, 1098 SM Amsterdam, The Netherlands*
[m] *Genotype GmbH, Angelhofweg 39, D-69259 Wilhelmsfeld, Germany*
[n] *MWG AG Biotech, Anzinger Str. 7, 85554 Ebersberg, Germany*
[o] *CEINGE and Dipartimento di Biochimica e Biotecnologie Mediche, Università 'Frederico II' di Napoli, Via Pansini 5, 80131 Napoli, Italy*
[p] *MediGenomix GmbH, DNA-Analytics and Genomics, Locharmer Str. 29, D-82152 Planegg/Martinsried, Germany*
[q] *Laboratoire Plastes et Différenciation cellulaire, UMR5575, Université Joseph Fourier et CNRS BP53 F-38041 Grenoble, France*
[r] *Vienna Biocenter, Institute of Microbiology & Genetics, Dr. Bohr-Gasse 9, A-1030 Vienna, Austria*
[s] *Institute de Biotechnologie des Plantes (IBP), UMR/CNRS 8618, University de Paris-Sud, F-91405 Orsay, France*
[t] *Lab. Physiologie et Biologie Moleculaire des Plantes, UMR CNRS 5545, Universite de Perpignan, 52 Avenue de Villeneuve, 66860 Perpignan Cedex, France*
[u] *Department de Genètica Molecular, Institut de Biologia Molecular de Barcelona, CSIC, Barcelona, Spain*
[v] *Department of Biological Sciences, University of Durham, Durham, DH1 3LE, UK*
[w] *Microsynth GmbH, Schutzenstr. 15, CH-9436 Balgach, Czech Republic*
[x] *Baseclear, PO Box 1336 Leiden, The Netherlands*
[y] *Faculté Universitaire des Sciences Agronomiques, Unité de Microbiologie, 6, Avenue Maréchal Juin, B-5030 Gembloux, Belgium*
[z] *Departament de Genètica, University of Valencia, 46100 Burjasot, Valencia, Spain*
[aa] *UCL-FYSA, Croix du Sud, 2-20, B-1348 Louvain-la-Neuve, Belgium*
[bb] *Keygene NV, PO Box 216, 6700 AE Wageningen, The Netherlands*
[cc] *Functional Genome Analysis, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 506, D-69120 Heidelberg, Germany*
[dd] *GSF-Forschungszentrum f. Umwelt u. Gesundheit, Munich Information Center for Protein Sequences am Max-Planck-Institut f. Biochemie, Am Klopferspitz 18a, D-82152, Germany*

Received 26 November 1998; received in revised form 2" December 1999; accepted 3" December 1999

Abstract

Arabidopsis thaliana has a relatively small genome of approximately 130 Mb containing about 10% repetitive DNA. Genome sequencing studies reveal a gene-rich genome, predicted to contain approximately 25 000 genes spaced on

average every 4.5 kb. Between 10 to 20% of the predicted genes occur as clusters of related genes, indicating that local sequence duplication and subsequent divergence generates a significant proportion of gene families. In addition to gene families, repetitive sequences comprise individual and small clusters of two to three retroelements and other classes of smaller repeats. The clustering of highly repetitive elements is a striking feature of the *A. thaliana* genome emerging from sequence and other analyses.

Keywords: *Arabidopsis thaliana*; Genome sequencing; Genes

1. Introduction

There are many fundamental differences in developmental strategies, metabolism and environmental interactions between the plant and animal Kingdoms, and a deeper understanding of plant-specific processes will add new facets to the greatly expanding knowledge emerging from studies of the model eukaryotic organisms such as yeast, *Drosophila*, *C. elegans* and mouse. In addition, a much deeper understanding of higher plant biology is urgently needed to address the significant challenges facing agriculture in the coming years. Nutritious food needs to be provided to the ever-increasing world population, and agricultural technology can provide technical options to place into the significant political and economic solutions that must be found. The necessary increases in crop production need to be achieved in ways that avoid environmental degradation and that preserves different habitats. Finally, significant losses in productivity from plants adapted for growth in the world's most productive regions have been predicted due to climate changes. These challenges require improved plant productivity, adaptability and utility that can be achieved through molecular genetic approaches.

Studies in maize (*Zea mays*) and *Arabidopsis thaliana* (Arabidopsis) have provided the best solutions to acquiring a large amount of information on many aspects of plant biology at the molecular level. Maize is a member of the monocot group of plants that includes the cereals, the primary source of nutrition for humans and their domesticated animals. The dicot group of plants is much more diverse that the monocots, and *A. thaliana* has emerged as a suitable model for this group. It has had a long and important history of genetic analysis, but the most significant impetus for adopting it as a model was the suitability of its relatively small genome for map-based gene isolation (Pruitt and Meyerowitz, 1986). This has led to the development of resources for gene isolation such as genetic and physical maps, contributing to the large increase in studies of a wide range of topics in *A. thaliana*, ranging from pathology to development, that are enhancing our understanding of complex phenomena in molecular detail.

Progress in sequencing genomes of organisms such as bacteria, yeast and *C. elegans* have provided the plant community with a clear example of the power of systematic approaches to defining gene function on a whole-genome scale. The relatively small 130 Mb genome of *A. thaliana* (see Fig. 1 for a comparison of plant genome sizes), together with the large cohort of scientists working on diverse areas of *A. thaliana* biology, suggested this plant as the obvious choice for determining the sequence of the entire genome. Large-scale cDNA sequencing was initiated in the US and France in 1992, and a unigene set of 11 000 sequences has been assembled from the 30 000 EST sequences. Genome sequencing began in the EC in 1994 on a pilot scale, and a plan to extend the scope of this work into an international large-scale effort was endorsed in 1994. The Arabidopsis Genome Initiative (AGI) was formed in 1996 after funding was secured in the US, Japan and the EU to scale up the rate of sequencing. The present goal of AGI is to complete the genome sequence before the end of 2000.

This review describes our present knowledge of chromosome structure, the approaches taken to genome-wide sequencing and progress made, features of genome composition revealed at the sequence level, and the approaches taken to determine the biological roles of genes.

2. The *A. thaliana* genome

The karyotype of ecotype Columbia, the strain being sequenced, is shown in Fig. 2. 5S RNA loci were identified in the pericentromeric regions of chromosomes 4 and 5. In the ecotype Landsberg erecta (Ler) a third 5S locus was found in the middle of the long arm of chromosome 3. The pericentromeric region was shown to contain a central domain flanked by interspersed variable blocks of pAL1 180 bp tandem repeats, the 106B repeat with homology to LTR sequences, and the uncharacterized 17A10 repeats (Fransz et al., 1998). The 730 ± 100 rDNA genes are arranged into two near-homogeneous megabase-sized clusters at the tip of the short arms of chromosomes 2 and 4 that comprise the nucleolar organisers (NORs). NOR2 and NOR4 are both about 3.6–4.0 Mb, and the repeat units are essentially identical in each locus except for three spacer length variants in NOR4 (Copenhaver and Pikaard, 1996a,b). The proximal end of both the NORs have been identified in sequenced BAC clones identified with chromosome-specific markers.

The telomeres of *A. thaliana* are composed of tandemly-repeated approximate 3.5 kb blocks of 5′-C(C/T)CTAAA-3′ conforming to the consensus found in eukaryotes. More extensive subtelomeric repeats may be found and these are predicted to contain interspersed low-copy sequences (Richards and Ausubel, 1988; Richards et al., 1992). However, no extensive subtelomeric repeat sequences were found at the junctions between NOR2 and NOR4 and the telomere (Copenhaver and Pikaard, 1996a).

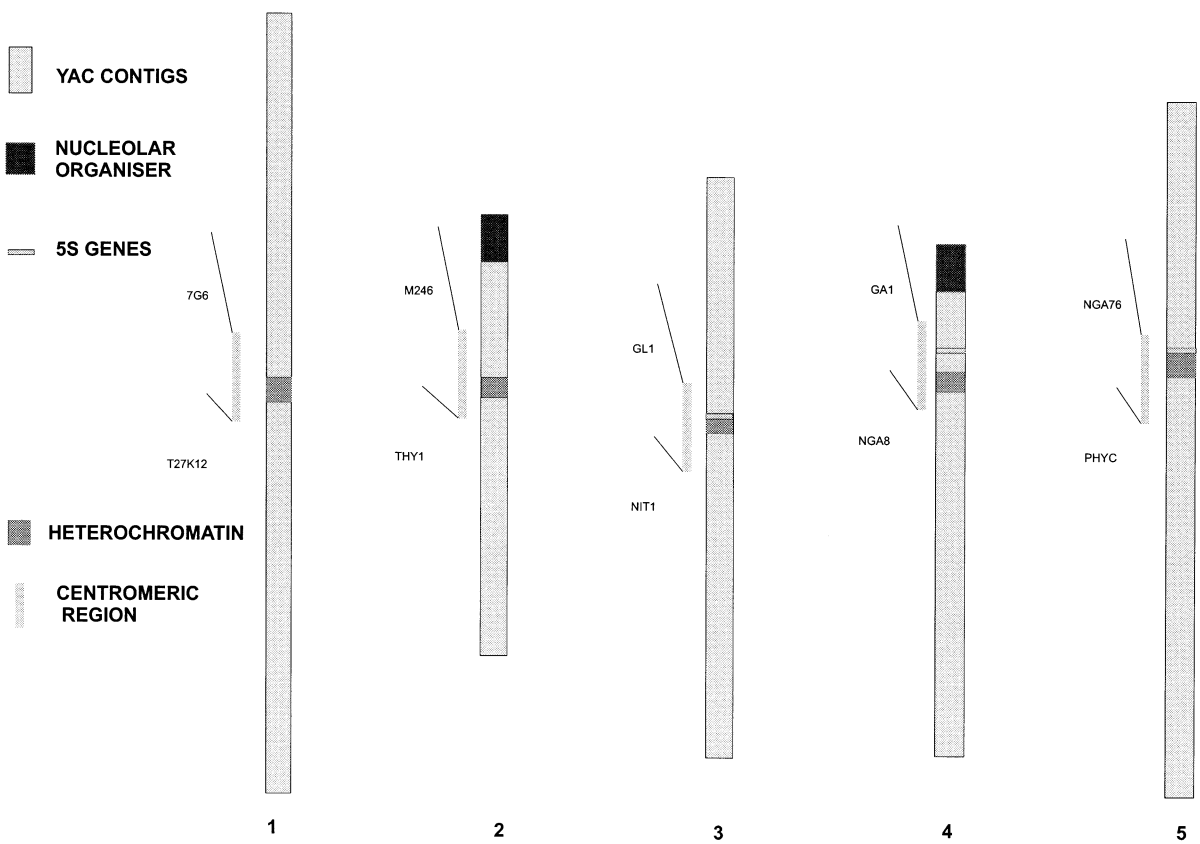

Fig. 1. The relative lengths of the five chromosomes of *A. thaliana* are represented by genetic map units and synaptonemal spread measurements. The genetic markers defining functional centromeres are shown. The extent of YAC contig cover is shown, but this is not necessarily contiguous coverage.

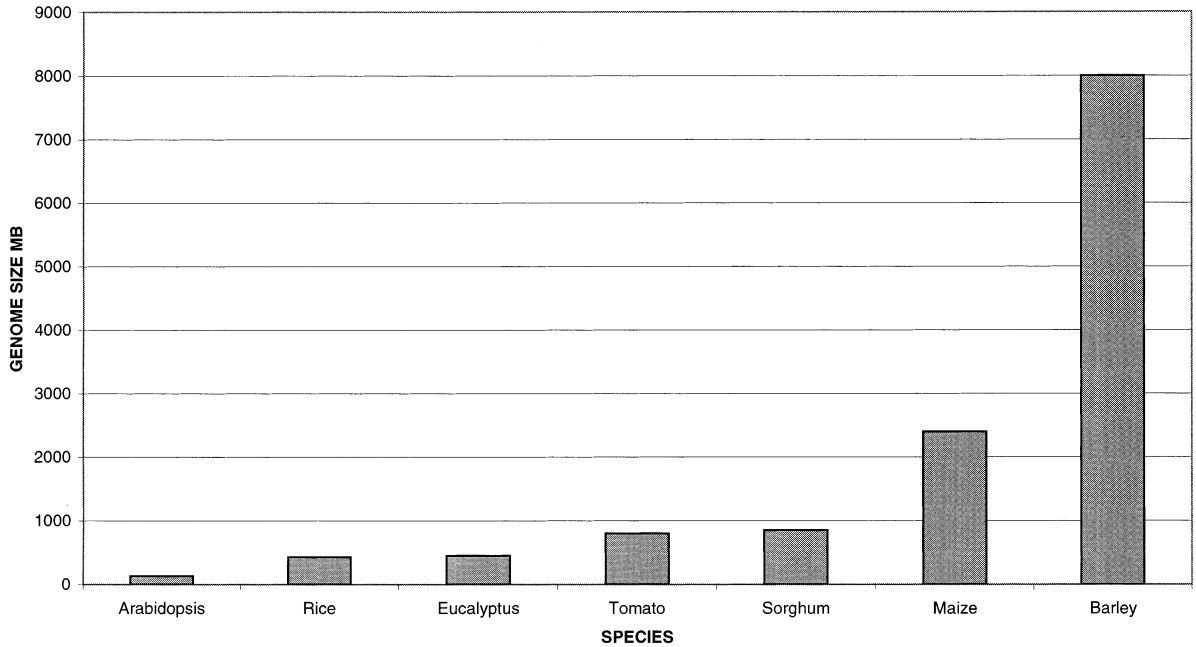

Figure 2. The relative sizes of selected diploid plant genomes is shown, based on Feulgen staining for DNA content.

Centromeres have been mapped functionally on the five chromosomes using tetratype frequencies (Copenhaver et al., 1998). Polymorphisms in the clusters of 180 bp pAL1 repeats have been used to map their position on each chromosome (Round et al., 1997). These corresponded to regions of significant complexity identified by physical mapping (see below), to regions of interspersed repeat structures observed cytogenetically, and to the regions of low cross-over frequencies described above. In situ hybridisation showed the 106B and pAL1 repeat classes were part of projections of the chromosome pulled pole-wards by the kinetochore fibres, and were thus part of a functional centromere. The location of centromeres to large regions of complex interspersed repetitive DNA is typical of centromeres found in higher eukaryotes.

The most common mapping population used is the recombinant inbred (RI) lines generated from a cross of ecotypes Landsberg erecta and Columbia and taken by single seed descent to the F_8 generation (Lister and Dean, 1993). This provided 300 lines that are essentially homozygous and can be replicated indefinitely. Currently over 500 markers of different types have been mapped on this population, and this map is maintained by the Arabidopsis Genome Resource (AGR) at http://synteny.nott.ac.uk/agr/ri.html.

The first generation of physical maps of chromosome 2 (Zachgo et al., 1996), chromosome 3 (Camilleri et al., 1998; Sato et al., 1998), chromosomes 4 (Schmidt et al., 1995) and chromosome 5 (Kotani et al., 1997a,b; Schmidt et al., 1997) represented chromosomes as contigs of YACs anchored onto the genetic map using molecular markers. This strategy yielded extensive coverage of the chromosomes, with few gaps in representation. To date about 90% of the genome is represented by a minimally redundant set of 350 YACs anchored by one or more markers to the genome.

Detailed physical maps of all five chromosomes are being constructed to establish a set of minimally-overlapping clones for sequencing. This is accomplished using a range of methods, exemplified here in chromosome 4. The TAMU and IGF BAC libraries together with the Mitsui P1 library together provide a 10–12-fold coverage of the chromosome YACs from the physical map were used as hybridisation probes to select BACs, which were then digested and analysed on South-

ern blots using iterative hybridisations with selected BACs. This approach yielded a minimally overlapping set of clones with experimentally determined overlaps together with a restriction map (Bent et al., 1998). This map has proved essential for identifying chimaeric and deleted clones and sequence assembly errors. Gaps between contigs were covered using the BAC end sequence database (see below), and direct sequencing of PCR products amplified from genomic DNA or underlying clones. More recently a semi-automated BAC fingerprinting system (fingerprint contig or fpc) that matches high-throughput sequencing capacity has been developed (Marra et al., 1997) (http://www.sanger.ac.uk/Users/cari/fpc.shtml). Genome wide contig assembly using high stringency parameters in fpc has resulted in 396 contigs of 300 kb average size containing an average of 45 clones each (http://genome.wustl.edu/gsc/arab/arabsearch.shtml). Minimal tiling paths are being selected for sequencing using these contigs by several sequencing groups (see Table 1). A BAC end-sequence database containing 36 574 end sequences from 20 000 TAMU and IGF BAC clones has been developed (http://www.tigr.org/tdb/at/atgenome/bac_end_search/bac_end–search.html). In this strategy, the complete sequence of a 'seed' BAC, anchored by marker content, is searched against a database of end-sequences from a BAC library to select the minimally overlapping clones to be sequenced in each direction.

A variety of approaches have been adopted by most groups to asemble chromosomal BAC contigs (Table 1) to utilize the strengths of different approaches based on local representation, ability to walk through repetitive regions, and the ability to generate chromosome-arm sized contigs. The identification of clones for sequencing in chromosome arms need not be a rate-limiting step for sequence acquisition, but as more complex regions are dealt with, such as centromeres, and telomeres, innovative strategies may need to be developed to achieve contiguous coverage of entire chromosomes.

3. Sequence analysis

The goal of AGI groups has been to produce highly accurate sequence representing as much of the genome as possible in chromosome arm-sized contigs. Most future uses of the sequence, such as

Table 1
The Arabidopsis Genome Initiative: regions, strategies and progress

Chromosome	Group	Link	Progress	
			Clones sequenced	Clones remaining
1	SPP Consortium TIGR	http://sequence-www.stanford.edu/ http://genome.bio.upenn.edu/ATGCUP.html http://pgec-genome.pw.usda.gov/ http://www.tigr.org/tdb/at/atgenome/	122 (12 Mb)	262
2	TIGR	http://www.tigr.org/tdb/at/atgenome/	249 (21 Mb)	Completed
3	KAZUSA GENOSCOPE TIGR	http://www.kazusa.or.jp/arabi/ http://genoscope.cns.fr/externe/arabidopsis/ http://www.tigr.org/tdb/at/atgenome	105 (10Mb)	200
4	EU Consortium CSHL/WU Consortium	http://websvr.mips.biochem.mpg.de/proj/thal/ http://nucleus.cshl.org/protarab	228 (18 Mb)	Completed
5	KAZUSA EU Consortium CSHL/WU Consortium	http://www.kazusa.or.jp/arabi/ http://websvr.mips.biochem.mpg.de/proj/thal/ http://nucleus.cshl.org/protarab	307 (20 Mb)	54

Table 2
General features of a chromosome 4

Long arm	14, 498, 507
Short arm	3 052 402 bp
Overall G+C content	36.02%
Protein-coding DNA	46.07%
Number of encoded proteins	3744
Base-pairs per gene	4643
Average exons per gene	5.24 (1-41)
Average exon length	256 bp
Average intron length	188 bp
Known genes	8%
Strong similarity to known genes	23%
Similarity to known genes	32%
Similar to predicted genes	26%
Predicted chloroplast/mitochondrial targeted	18%
tRNA	81

mass-spectrometric analysis of proteins of the sequence will require highly accurate sequence. Chip technology for expression analyses and mapping requires highly accurate sequence for designing discriminatory elements in arrays. To achieve this accuracy, standard shotgun sequencing approaches are used. The assembly of multiple sequences into a contig representing the BAC insert is generally carried out by programs such as Phrap, and approximately 10% of the clones are problematic due to repetitive DNA and require specific solutions.

Independent sequencing of a small proportion of the genome indicate that the overall accuracy levels reached are less than one error every 10 000 bases. In addition, most of the *A. thaliana* sequence data has a quantitative assessment of base calling accuracy associated with the consensus sequence obtained from the *Phred* base-calling program (Ewing et al., 1998).

Assessing the assembly of contigs to ensure they represent the underlying genome is another critical quality control measure that have revealed gross errors in the assembly of several clones.

Approximately 40–50% of the predicted genes match EST and cDNA sequence (see Table 2), and this provides the only experimental evidence for modeling genes encoding proteins having no significant similarity to proteins from other organisms.

A variety of programmes are used to analyse chromosome 4 sequence. Splice sites are identified by NETPLANTGENE and coding regions by XGRAIL. Two gene modelling programmes, GENEFINDER (P. Green, unpublished) and GENESCAN (Burge and Karlin, 1997), are used to identify potential genes. In *A. thaliana* GENESCAN is generally more reliable in predicting exon–intron boundaries, and GENEFINDER provides more reliable data on the number of exons in the gene model. Gene models are investigated for consistency with known protein families and cDNA and models adjusted accordingly. Putative errors such as frameshifts can be identified by these analyses. The position of predicted exons and introns is shown in embl entries.

Sequence similarity searches using a variety of algorithms (e.g. BLAST, FASTA) against the latest releases of the public nucleotide and protein databases, together with available EST sequence data, are carried out on the protein sequence predicted by the gene models. A FASTA score of >150 is used as an adequate similarity score across conserved motifs or the entire protein sequence to annotate a gene as 'similar to protein X from organism Y'. This information and related material for chromosome 4 sequence can be found at the *A. thaliana* database (AtDB) at http://genome-www.stanford.edu/Arabidopsis/ is a convenient link to genome databases, such as those supported by sequencing groups. These web sites are shown in Table 1.

4. Sequence features

By November 1999 approximately 104 Mb of the predicted 130 Mb genome have been sequenced (see Table 1). It is predicted that the remaining sequence will be completed before the end of 2000. Chromosomes 2 and 4 are completed except for 1 Mb of complex repeats associated with the central region of the centromeres. Some general features of chromosome 4 sequence are shown in Table 2. This analysis reveals an information-rich genome, with a gene-space of over 50%, and where about half of the genes have

sufficient similarity with genes from other organisms to infer a putative cellular role for the encoded proteins. The repeat sequences encountered were primarily of LTR-and non-LTR retroelements occurring as singletons up to large clusters in pericentromeric heterochromatin. These elements comprise less than 10% of the gene features, in contrast to the distribution of retroelements in the maize genome where they are present as complex interspersions of different classes in large clusters occupying over 50% of the genome (San Miguel et al., 1996). An interesting feature of the *A. thaliana* genome is the frequent occurrence of clusters of closely related protein-coding genes adjacent on the same strand. These comprise 12% of the chromosome 4 sequence to date.

Gene regulatory systems appear to be standard for eukaryotes, with long promoter elements containing multiple binding sites for transcription factors.

Cellular role categories of plant genes have been established based on those used for yeast (Mewes et al., 1997). Additional categories to accommodate new classes of proteins found in plants were devised, such as for secondary products to cope with the large number of genes in this category, for components of the photosynthetic apparatus not connected with electron transport, for cell wall components and polysaccharide metabolism, intracellular traffic into chloroplasts and vacuoles, chloroplast components, and a category named disease and defense, containing many diverse genes. Each protein with a significant similarity to a characterized protein is assigned a single category based on functional similarity inferred from any structural similarity using the PEDANT software system (http://pedant.mips.biochem.mpg.de/index.html).

The proportion of predicted genes encoding products in each cellular role has been determined for a small number of genes. About 1/3 of the predicted genes with significant similarities to other genes are involved in primary and secondary metabolism. This reflects a possible bias towards these classes of genes possessing conserved motifs with established functions, but it also reflects the wide range of metabolism found in higher plants. Genes in this class in general have a greater degree of similarity to genes of eubacterial origin, reflecting the possible prokaryotic origin of various biosynthetic pathways. These types of genes are now encoded and regulated by the nuclear genome of plants and have transit peptides to direct them to chloroplasts. The proportion of genes involved in cellular regulation, such as transcription and signal transduction, are typical of multicellular eukaryotes, and are generally equally remotely related to genes from human, *C. elegans* and *Drosophila*, reflecting the antiquity of the last common ancestor of plants and animals. Finally, genes in some classes, such as those involved in intracellular trafficking and some aspects of transcriptional control, show a high degree of conservation among these diverse organisms, probably reflecting conserved mechanisms.

Fifty percent of the predicted genes have no significant similarity to any other genes. This may be because similar genes in other organisms have not yet been sequenced, or this class of genes may carry out plant-specific functions with protein motifs not found in other organisms. With the yeast genome and most of the *C. elegans* genome completed, the former possibility is becoming less likely. Determining the functions of this set of genes is an important goal.

5. Gene function

There are several complexities associated with systematic function search in *A. thaliana* and other higher plants. The most significant is the complex haploid phase of the angiosperm life-cycle that leads to the loss of a significant proportion of genes required for cell viability in 'standard' loss-of-function genetic screens.

The large proportion of closely related genes in the *A. thaliana* genome indicates a significant potential for functional redundancy among related genes, such that a loss-of-function mutation may be compensated by the activity of another family member.

Furthermore, the best frequencies of gene replacement by homologous recombination re-

ported in *A. thaliana* are less that 1/3000, which is not suitable for routine use (Kempin et al., 1997).

Therefore strategies involving mis-expression of genes are being developed that may reveal the functions of both individuals and family members more efficiently, leading to phenotypes in the heterozygous state that reveal the function of genes necessary for cell viability.

Alternative strategies have been developed for *A. thaliana* that use either *Agrobacterium tumefaciens* T-DNA or heterologous transposons for insertional mutagenesis. A minimum of 100 000 independent random insertions into the *A. thaliana* genome are needed to achieve a 90% probability of disrupting a 2 kb gene. Many insertions of the large T-DNA sequence generally cause null mutations when inserted in ORFs or introns, and large populations of insertions have been developed and used with significant success (Bouchez and Hofte, 1998; Winkler et al., 1998). T-DNA insertion populations can reveal unlinked mutations, and rearrangement of insertion sites and T-DNA.

Two transposon tagging systems, the Ac–Ds and Spm–En-1 transposons from maize, have been developed, each with significant advantages in terms of the ease of generating large populations, making 'clean' insertions that can be readily sequenced, and making informative insertions. An autonomous En-1 element has been transformed into *A. thaliana* and propagated by single seed descent for five generations (Wisman et al., 1998) to yield a population of 3000 lines with a total of 15 000 insertions. Screening for insertions is carried out by PCR from sites at the ends of the transposons in pools of lines that give a unique address for each line. Speulman et al. (1999) created populations containing 100 000 elements at 30–40 copies of En-1 per genome. The SLAT collection (Tissier et al., 1999) contains 28 000 plants contains a single copy of dSpm per genome that has been segregated from the transposase source to stabilize insertions. These populations provide a 90% chance of detecting an insertion in an average-sized 5 kb genes in *A. thaliana*. The pooling strategies adopted in the En–Spm systems enable rapid and comprehensive reverse screens to detect inserts in genes by PCR or filter hybridisation, and the high copy number lines are particularly suitable for forward screens. Disadvantages of lines containing several copies of autonomous elements include somatic reversion events and the requirement for time-consuming segregation to obtain a line with a single insertion of interest. The size of these populations will increase steadily such that En–Spm systems, especially with their increased copy number per genome, provide an excellent resource for forward screens and large-scale gene knockout in *A. thaliana* for the next few years.

Ac–Ds transposons from maize have small terminal repeats that permit substantial modifications to the element without compromising transposition functions, consequently these have been modified to detect plant gene expression patterns by incorporating a splice acceptor in three reading frames with a GUS reporter gene (Sundaresan et al., 1995). Gene expression can be monitored when the Ds insertion is heterozygous, enabling detection of genes in which mutations cause embryo and gametophyte lethality.

The excision frequency and selection for re-insertion of Ds elements can be controlled to both maximize the generation of insertions and to define the regions of the genome in which insertions are preferentially recovered (Bancroft and Dean, 1993; Balcells et al., 1994). Thus launching pads evenly distributed throughout the genome provide both a high probability of generating a spectrum of insertions that may be quite different from that obtained using other strategies and a means of directing insertions at high-frequency into specific regions of the genome containing genes of interest.

Ds has also been modified with the constitutive high-level 35S promoter from cauliflower mosaic virus directed outwards from the terminus of the transposon. Insertion adjacent to a gene can drive expression that may lead to phenotypes in the heterozygous state, thus revealing functions of apparently redundant genes or genes necessary for cell viability (Schaffer et al., 1998). The flexibility of the Ac–Ds system in

terms of the types of element that can be engineered, and the genetic strategies that can be used to direct the elements to different areas of the genome makes this a powerful system for functional genomics in *A. thaliana*. Presently large populations of both gene trap and activation-tagged lines are being generated.

The full variety of insertional mutation populations are being used by plant scientists in reverse screens to determine the functions of sequenced genes, and in forward screens to isolate genes contributing to a wide variety of phenotypes. Plants have been grown in a wide variety of different and tightly controlled conditions to detect phenotypes not normally observable under 'standard' optimal growth conditions (for example Hirsch et al., 1998). These conditions provide the potential for a systematic approach to plant functional genomics.

6. Relevance to crop plant research and agricultural improvement

One of the most important future developments in plant sciences is to ensure the systematic application of the knowledge gained of biological processes in *A. thaliana* to crop plants. Sequences are substantially conserved among flowering plants because of their relatively recent divergence approximately 200 million years ago (Wolfe et al., 1989), but several considerations complicate the straight forward definition of an orthologous relationship as that between genes encoding the most similar proteins. For example, the number of members of gene families (paralogs) can be different between even closely related plants because frequent genome duplications are often found in cultivated species. Conserved gene order, together with sequence similarity, may be a valuable additional tool for defining functional relationships such as orthology among genes, because if recombination has not broken up the local order of genes between plants, ancestral homeologous relationships are revealed. In relatively closely related plants such as cereals these relationships have been established (reviewed by Gale and Devos, 1998), but to date there is no evidence of conserved gene order between cereals such as rice and distantly-related *A. thaliana*. Finally, the flowering plants are characterised by very diverse morphology, development and environmental interactions, and it is likely that members of gene families from even closely related species have evolved to perform different functions.

An effective way to harness *A. thaliana* genome sequence and functional genomics is to sequence reference crop plant genomes and make links between these genomes and *A. thaliana*. The most important group of plants to consider are the grasses, because of their supreme importance for food production, legumes, because of their symbiotic relationship with N-fixing bacteria, and trees, because of their economic and ecological importance. The relatively small and simple genome of rice compared to maize (450 Mb vs. 2400 Mb) makes it a obvious candidate for the reference grass genome. The rice genome sequence will form a bridge between the relatively well-characterised *A. thaliana* genome, where it is feasible to determine the cellular roles of large numbers of genes, and the other cereals, whose genome complexity and present relative intractability to biological investigation preclude them from detailed analysis.

7. Conclusions and prospects

Completing the sequence of *A. thaliana* in the next year will provide one of the main foundations for establishing a quantitative and predictive plant biology, by revealing all of the genes necessary for the full range of plant function. More incisive experiments can be planned based on knowledge derived from the functions of large sets of genes. Finally, comparing gene function between organisms will reduce this large body of information to sets of core cellular processes. It is likely that plants will contribute several new biological processes to this set because of the great evolutionary distance between plants and other eukaryotes.

References

Balcells, LI., Sundberg, E., Coupland, G., 1994. A heat-shock promoter fusion to the Ac transposase gene drives inducible transposition of a Ds element during Arabidopsis embryo development. Plant J. 5, 755–764.

Bancroft, I., Dean, C., 1993. Transposition pattern of the maize element Ds in *Arabidopsis thaliana*. Genetics 134, 1221–1229.

Bent, E., Johnson, S., Bancroft, I., 1998. BAC representation of two low-copy regions of the genome of *Arabidopsis thaliana*. Plant J. 13, 849–855.

Bouchez, D., Hofte, H., 1998. Functional genomics in plants. Plant Physiol. 118, 725–732.

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

Copenhaver, G.P., Pikaard, C.S., 1996a. RFLP and physical mapping with an rDNA-specific endonuclease reveals that nucleolus organizer regions of *Arabidopsis thaliana* adjoin the telomeres on chromosomes 2 and 4. Plant J. 9, 259–272.

Copenhaver, G.P., Pikaard, C.S., 1996b. 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.

Copenhaver, C.S., Browne, W.E., Preuss, D., 1998. Assaying genome-wide recombination and centromere functions with Arabidopsis tetrads. Proc. Natl. Acad. Sci. USA. 95, 247–252.

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998. Base-calling of automated sequence traces using Phred. I. Accuracy assessment. Genome Res. 8, 175–186.

Fransz, P., Armstrong, S., Alonso-Blanco, C., Fischer, T.C., Torres-Ruiz, R.A., Jones, G., 1998. Cytogenetics for the model system *Arabidopsis thaliana*. Plant J. 13, 867–876.

Gale, M., Devos, K., 1998. Plant comparative genetics after 10 years. Science 282, 656–658.

Hirsch, R.E., Lewis, B.D., Spalding, E.P., Sussman, M.R., 1998. A role for the AKT1 potassium channel in plant nutrition. Science 280, 918–921.

Kempin, S.A., Liljegren, S.J., Block, L.M., Rounsley, S.D., Yanofsky, M.F., Lam, E., 1997. Targeted disruption in Arabidopsis. Nature 389, 802–803.

Kotani, H., Sato, S., Fukuami, M., Hosouchi, T., Nakazaki, N., Okumura, S., Wada, T., Lui, Y.-G., Shibata, D., Tabata, S., 1997a. A fine physical map of *Arabidopsis thaliana* chromosome 5: Construction of a sequence-ready contig map. DNA Res. 4, 371–378.

Kotani, H., Nakamura, Y., Sato, S., Kaneko, T., Asamizu, E., Miyajima, N., Tabata, S., 1997b. 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.

Lister, C., Dean, C., 1993. Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. Plant J. 4, 745–750.

Marra, M., Kucaba, T.A., Dietrich, N.L., Green, E.D., Brownstein, B., Wilson, R.K., McDonald, K.M., Hillier, L.W., McPherson, J.D., Waterston, R.H., 1997. High-throughput fingerprinting of large-insert clones. Genome Res. 7, 1072–1084.

Mewes, H.-W., et al., 1997. Overview of the yeast genome. Nature 387, 7–84.

Pruitt, R.E., Meyerowitz, E.M., 1986. Characterization of the genome of *Arabidopsis thaliana*. J. Mol. Biol. 187, 169–183.

Richards, E.J., Chao, S., Vongs, A., Yang, J., 1992. Characterization of *Arabidopsis thaliana* telomeres isolated in yeast. Nucl. Acids Res. 20, 4039–4046.

Richards, E.J., Ausubel, F.M., 1988. Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. Cell 53, 127–136.

Round, E., Flowers, S.K., Richards, E.J., 1997. *Arabidopsis thaliana* centromere regions: genetic map positions and repetitive DNA structure. Genome Res. 7, 1045–1054.

San Miguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P., Edwards, K., Lee, M., Avramova, Z., Bennetzen, J., 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274, 765–768.

Sato, S., Kotani, H., Hayashi, R., Liu, Y.-G., Shibata, D., Tabata, S., 1998. A physical map of *Arabidopsis thaliana* chromosome 3. DNA Res. 4, 215–230.

Schaffer, R., Ramsay, N., Samach, A., Corden, S., Putterill, J., Carre, I.A., and Coupland, G. (1998) The late elongated hypocotyl mutation of Arabidopsis disrupts circadian rhythms and the photoperiodic control of flowering.

Schmidt, R., West, J., Love, K., Lenehan, Z., Lister, C., Thompson, H., Bouchez, D., Dean, C., 1995. Physical map and organization of *Arabidopsis thaliana* chromosome 4. Science 270, 480–483.

Schmidt, R., Love, K., West, J., Lenehan, Z., Dean, C., 1997. Description of 31 YAC contigs spanning the majority of *Arabidopsis thaliana* chromosome 5. Plant J. 11, 563–572.

Speulman, E., Metz, P.L.J., van Arkel, G., te Lintel Hekkert, B., Stiekema, W., Pereira, A., 1999. Plant Cell 11, 1853–1866.

Sundaresan, V., Springer, P., Volpe, T., Haward, S., Jones, J.D.G., Dean, C., Ma, H., Martienssen, R., 1995. Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. Genes Devel. 9, 1797–1810.

Tissier, A.F., Marillonnet, S., Klimyuk, V., Patel, K., Torres, M.A., Murphy, G., Jones, J.D.G., 1999. Multiple independent defective suppressor–mutator transposon insertions in Arabidopsis: a tool for functional genomics. Plant Cell 11, 1841–1852.

Winkler, R.G., Frank, M.R., Galbraith, D.W., Feyereisen, R., Feldmann, K.A., 1998. Systematic reverse genetics of transfer-DNA-tagged lines of Arabidopsis. Plant Physiol. 118, 743–750.

Wisman, E., Cardon, G.H., Fransz, P., Saedler, H., 1998. The behavior of the autonomous maize transposable element En–Spm in *Arabidopsis thaliana* allows efficient mutagenesis. Plant Mol. Biol. 37, 989–999.

Wolfe, K., Gouy, M., Yang, Y.-W., Sharp, P., Li, W.-H., 1989. Date of the monocot–dicot divergence estimated from chloroplast DNA sequence data. Proc. Natl. Acad. Sci. USA 86, 6201–6205.

Zachgo, E.A., Wang, M.L., Dewdney, J., Bouchez, D., Camilleri, C., Belmonte, S., Huang, L., Dolan, M., Goodman, H.M., 1996. A physical map of chromosome 2 of *Arabidopsis thaliana*. Genome Res. 6, 19–25.