Sequence and analysis of the *Arabidopsis* genome

Michael Bevan*, Klaus Mayer†, Owen White‡, Jonathan A Eisen‡, Daphne Preuss§, Thomas Bureau#, Steven L Salzberg‡ and Hans-Werner Mewes†

The comprehensive analysis of the genome sequence of the plant *Arabidopsis thaliana* has been completed recently. The genome sequence and associated analyses provide the foundations for rapid progress in many fields of plant research, such as the exploitation of genetic variation in *Arabidopsis* ecotypes, the assessment of the transcriptome and proteome, and the association of genome changes at the sequence level with evolutionary processes. Nevertheless, genome sequencing and analysis are only the first steps towards a new plant biology. Much remains to be done to refine the analysis of encoded genes, to define the functions of encoded proteins systematically, and to establish new generations of databases to capture and relate diverse data sets generated in widely distributed laboratories.

Addresses
*Molecular Genetics Department, John Innes Centre, Colney Lane, Norwich NR4 7UH, UK; e-mail: michael.bevan@bbsrc.ac.uk
†GSF-Forschungszentrum für Umwelt und Gesundheit, Munich Information Center for Protein Sequences am Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, D-82152, Germany
‡The Institute for Genomic Research, 9712 Medical Centre Drive, Rockville, Maryland 20850, USA
§Howard Hughes Medical Institute, The University of Chicago, 1103 East 57th Street, Chicago, Illinois 60637, USA
#McGill University, Department of Biology, 1205 rue Dr Penfield, Montreal, Quebec, H3A 1B1, Canada
Correspondence: Michael Bevan

Current Opinion in Plant Biology 2001, 4:105–110

1369-5266/01/$ – see front matter

Abbreviations

AGI	*Arabidopsis* Genome Initiative
BAC	bacterial artificial chromosome
EST	expressed sequence tag
YAC	yeast artificial chromosome

Introduction

Arabidopsis had a small but important role as an experimental organism during the time when nearly all plant biology was conducted in the major crop species. The advent of the molecular biology era lead to the more widespread use of easily transformable species such as tobacco and petunia. However, map-based cloning strategies, which are used for the comprehensive genetic analyses of developmental and other processes, required a species with a relatively small genome, that was easily transformable, for which detailed genetic maps were available, and that was studied by a dynamic community of biologists. Two landmark studies on *Arabidopsis* set the scene for an explosive growth in our knowledge of plant gene function [1,2], and the realisation of what could be achieved by concerted focus on a single species lead to the formation of the *Arabidopsis* Multinational Steering Committee in 1990. This committee established a general framework of goals and aspirations for the *Arabidopsis* community, which included obtaining a complete gene sequence. Important precursors to the sustained international sequencing effort were established after 1991. The effort began with systematic physical mapping with cosmids and the creation of 'An *Arabidopsis thaliana* Database' (AatDB) in the Goodman Laboratory [3], the landmark commitment to physical mapping. Later, the widespread adoption of yeast artificial chromosomes (YACs) as substrates for mapping provided essentially complete coverage of most *Arabidopsis* chromosomes by the end of 1995.

Several factors converged to catalyse the establishment of a large-scale *Arabidopsis* sequencing program. These included the availability of a physical map of YACs for *Arabidopsis*, constant improvements in the efficiency of sequencing, the manifest value to the scientific community of the yeast and *Caenorhabditis elegans* genome sequences, and the emergence of initial data from pilot-scale sequencing revealing that *Arabidopsis* has a gene-rich genome [4]. Sequencing groups in the US, Japan and Europe were formed in August 1996, and a memorandum of understanding established common techniques and resources, accuracy standards, levels of analysis, and a common public release policy for sequence information. The initial objective was to complete the *Arabidopsis* genome sequence by 2004. At the end of 2000 this ambitious and difficult goal was achieved.

Technology

The substrates for sequencing were large-insert bacterial artificial chromosome (BAC) and P1 artificial chromosome (PAC) libraries that collectively covered the genome more than 15 times. Given these excellent starting materials, three components had to be developed and optimised to sequence efficiently the (then) large genome: the discovery and ordering of clones, techniques for sequencing, and techniques for the analysis of sequence data.

Whole-genome shotgun sequencing was not imaginable at the start of the project because of the relatively low efficiency of sequencing and inadequate computing resources. Since its introduction, especially for sequencing the *Drosophila* genome, the cost-effectiveness and rapid accumulation of useful data made possible by the shotgun strategy make it the method of choice, especially when coupled to targeted closure strategies. Earlier in the *Arabidopsis* project, the most

Table 1

***Arabidopsis* Genome Initiative groups, regions and sequence totals.**

Group	Chromosome	Total	Contact URLs
TIGR	2,3,1	35 Mb	http://www.tigr.org/tdb/at/atgenome/atgenome.html
KAZUSA	5,3	30 Mb	http://www.kazusa.or.jp/kaos/
SPP	1	20 Mb	http://sequence-www.stanford.edu/ara/SPP.html
EU ESSA	4,5	19 Mb	http://websvr.mips.biochem.mpg.de/proj/thal/
EU GENOSCOPE	3	9 Mb	http://www.genoscope.cns.fr/externe/English/Projets/Projet_A/organisme_A.html
CWU	4,5	13 Mb	http://nucleus.cshl.org/protarab/

important component that facilitated the sequencing of large contiguous chromosomal segments was the development of rapid and effective ways to identify BAC clones in efficient tiling paths (i.e. ordered arrays of clones with minimal overlaps representing a region of a chromosome) for sequencing using fingerprint contigs ([5•]; URL http://genome.wustl.edu/gsc/arab/arabidopsis.html), coupled to BAC-end sequencing (URL http://www.tigr.org/tdb/at/abe/bac_end_search.html). These techniques were pioneered in *Arabidopsis* and, when used effectively, generated large contigs of BAC clones with minimal overlap. Together with the improvements in sequencing technologies, such as dye-terminator chemistry and capillary electrophoresis, these strategies steadily increased the efficiency of *Arabidopsis* sequencing.

More than 98% of identified and anchored BACs, comprising 115.4 Mb net sequence, had been sequenced by October 2000. More than 100,000 expressed sequence tags (ESTs) had also been produced by the *Arabidopsis* Genome Initiative (AGI) and other groups by December 2000. By the end of 2000, most regions of the genome, including the highly repetitive centromeric regions had been accessed by BAC-clone sequencing. Subtelomeric regions of approximately 5–10 kb were sequenced using smaller clone inserts in cosmids and phage λ coupled to a genomic inverse polymerase chain reaction (IPCR). In two cases, telomeric regions that had previously been cloned in YACs were sequenced and joined to sequenced chromosome arms. Highly accurate contiguous sequence extended from telomeres deep into centromeric heterochromatin, an achievement beyond the scope of that originally considered possible.

In general, the sequencing of heterochromatic regions had been complicated by complications in identifying unambiguous clone overlaps in repetitive regions and by difficulties in the assembly of shotgun sequence of repeats. The relative difficulty of sequencing the heterochromatic regions, together with the vastly decreased rate of gene discovery in these regions relative to that in other regions, meant that they had generally become a low priority. Consequently, at least 15 heterochromatic BAC clones are still being sequenced. It will take a concerted effort using more directed methods to obtain contiguous sequences, but this should be possible. For example, a cluster of 70 kb containing 180 basepair repeats in centromere 4 has been sequenced and assembled using minor polymorphisms and retroelement sequences [6•]. Interestingly, some pericentromeric regions of reduced gene density and increased retroelement density in *Arabidopsis* resemble known regions of the maize genome [6•,7•]. Thus, the essential completion of these regions in *Arabidopsis* suggests that clone-by-clone strategies could be employed to sequence the major cereal genomes.

In most cases, once the groups comprising the AGI had established which chromosome regions each would sequence, the sequencing was completed and the sequences linked to those discovered by other groups with minimal overlap. Table 1 shows the AGI groups, the chromosome regions they worked on and their sequence totals.

Sequence assembly and analysis

The sequences of individual BACs were analysed in specialist informatics centres within each network. With minor variations, sequences were directly available to users before and after assembly and, after accuracy checks, analysis and annotation, were subsequently released via Genbank, the European Bioinformatics Institute (EBI) or the DNA Databank of Japan (DDBJ). Maps of sequenced regions that integrated genetic markers and other features with BAC identification and underlying sequence are one of the most useful forms of sequence information released, which greatly assists map-based gene isolation. The Munich Information Center for Protein Sequences (MIPS), The *Arabidopsis* Information Resource (TAIR), the Database of *Arabidopsis thaliana* Annotation (D*at*A) and the Kazusa *Arabidopsis* data Opening Site (KAOS) provide examples of integrated maps. MIPS and The Institute for Genomic Research (TIGR) have assembled pseudo-molecules representing chromosome arms by carefully assessing overlaps between adjacent BACs and ordering BACs with markers. These molecules form the foundation for future sequence analysis work by these groups. The resulting minimal gene set (i.e. that which results from removing multiple representations of genes caused by sequencing overlapping regions of clones) is renamed using chromosome codes to guide users to the chromosomal location of genes, and is used as a substrate for further analyses of genome structure and protein function. In this strategy, the original BAC-based annotation is relegated to use for the identification of substrates for gene

cloning. The total length of the ten pseudo molecules is 115 Mb, extending from either the telomeres or rDNA repeats to the 180 basepair centromeric repeats [8••]. The unsequenced centromeric and nucleolar organiser (NOR) regions are estimated to measure approximately 10 Mb (see below), yielding a total genome size of approximately 125 Mb, which is in the range of the 50–150 Mb haploid content estimated by different methods.

Throughout the project, annotation was conducted in different centres using different combinations of analysis software. These involved *in silico* gene-finding methods, comparison to EST and protein databases, and manual reconciliation of those data. Gene finding involved several steps: first, analysis of BAC sequences using a computational gene finder; second, alignment of the sequence to the protein and EST databases; and third, assignment of functions to each of the genes. Genscan [9], GeneMark.HMM [10], Xgrail [11], Genefinder (P Green, unpublished software) and GlimmerA [12] were used to analyse BAC sequences. All of these systems were specially trained for *Arabidopsis* genes. Splice sites were predicted using NetGene2 [13], Splice Predictor [14] and GeneSplicer (M Pertea, S Salzberg, unpublished software). For the second step, BACs were aligned to ESTs and to the *Arabidopsis* gene index [15] using programs such as DDS/GAP2 [16] or BLASTN [17].

Sequence was annotated and further analysed in the specialist centres using a combination of software tools, integration with EST sequence and manual annotation. The overall standard achieved was highly consistent among analysis groups, with 80% of the gene models completely consistent. Finally, the predicted functions, structure and other features of encoded proteins were compiled to facilitate experimental analysis. This analysis is being reiterated on the complete gene set using comparisons between gene family members and other approaches to refine gene predictions.

The software described above has been independently scrutinised for selectivity and sensitivity in gene-finding in *Arabidopsis* [18••]. This thorough and thoughtful analysis assessed a variety of gene-finding programs using AraSet, a group of 168 *Arabidopsis* genes (composed of 1028 exons, 860 introns and 94 entire intergenic sequences) that have experimentally validated structures. The performance of the software was assessed using specificity criteria that measured sensitivity (i.e. true positives/[true positives + false negatives]) and specificity (i.e. true negatives/[true negatives + false positives]). The identification of exons was very effective; GeneMark.hmm was the most reliable tool for predicting exons and gene structures, identifying 67 of the 128 AraSet genes completely and correctly. Common problems included the prediction of initiation sites (only 76% of which were identified correctly) and of intergenic regions (9% of which were predicted as introns). Perhaps the most important standard demanded by biologists is the prediction of the correct protein sequence, and GeneMark.hmm identified 89 of the AraSet protein sequences correctly. A combination of GeneMark and Genscan was eventually recommended as the latter program is better at finding some specific genes. It is important for users of the sequence to understand that gene predictions for all but experimentally defined genes (~10% of the total) are probabilities, not certainties. It is particularly important to define initiating and terminal exons before subsequent investigations and to attempt to verify predicted protein sequences using similarity with known protein sequences. Full-length cDNA sequences should allow the systematic experimental definition of the gene structures of up to 60% of all *Arabidopsis* genes (i.e. the proportion of genes for which ESTs have been matched to the predicted genes). An efficient way of gathering experimental data from laboratories using genome sequence data is also required to refine gene models.

The 25,498 genes predicted in *Arabidopsis* (Table 2) make up the largest gene set established to date [8••,19,20••]: 19,099 genes are predicted in *C. elegans* [19] and 13,601 in *Drosophila* [20••]. *Arabidopsis* and *C. elegans* have a similar gene density that is greater than that of *Drosophila*. *Arabidopsis* has significantly greater numbers of tandem gene duplications and segmental duplications than either *C. elegans* or *Drosophila*.

Genome structure

Each of the five *Arabidopsis* chromosomes has an essentially similar structure comprising either acro- or metacentric centromeres terminated by similar short subtelomeric repeats and telomeres. Over 5 Mb of centromeric DNA, more than for any other species, has been sequenced and assembled into large contigs. Unsequenced regions that may total approximately 3 Mb comprise large tracts of monotonous repeats as described above. Transposons account for at least 10% of the genome. Class I elements are much less abundant in *Arabidopsis* than in other plants, such as maize, and primarily occupy the centromere. In contrast, *Basho* elements and Class II transposons, such as miniature inverted-repeat transposable elements (MITEs) and Mutator-like elements (MULEs), predominate in the pericentromeric domains [21•]. Transposon-rich regions are relatively gene-poor, and have low rates of recombination and cognate ESTs, indicating correlation among low gene expression, high transposon density and low recombination.

The centromeres are responsible for spindle attachment and the proper segregation of sister chromatids during meiosis and mitosis. The centromeres of *Arabidopsis*, like those of other higher eukaryotes, contain retroelements, transposons, microsatellites and middle repetitive DNA [22••]. Nevertheless, at least 47 expressed genes encoding a diverse array of proteins are found within the genetically defined centromeres of *Arabidopsis* [8••]. Several repeats are conserved among the centromeres of the five

Table 2

Summary of the general features of the *Arabidopsis* genome.

Feature	Chromosome 1	Chromosome 2	Chromosome 3	Chromosome 4	Chromosome 5
DNA molecules					
Length (bp)*	29,105,111	19,646,945	23,172,617	17,549,867	25,953,409
Number of genes†	6543	4036	5220	3825	5874
Gene density (kb per gene)	4.0	4.9	4.5	4.6	4.4
Exons					
Number	35,482	19,631	26,570	20,073	31,226
Total length (bp)	8,772,559	5,100,288	6,654,507	5,150,883	7,571,013
Average number per gene	5.4	4.9	5.1	5.2	5.3
Average size (bp)	247	259	250	256	242
Introns					
Number	28,939	15,595	21,350	16,248	25,352
Total length (bp)	4,828,766	2,768,430	3,397,531	3,030,649	4,030,045
Average size (bp)	168	177	159	186	159
Proportion of genes with ESTs	60.8%	56.9%	59.8%	61.4%	61.4%
Number of ESTs‡	30,522	14,989	20,732	16,605	22,885

In total, the five chromosomes of *A. thaliana* contain *DNA molecules of 115,409,949 basepairs (bp) in length, †which encode 25,498 genes. ‡A total of 105,733 *A. thaliana* EST sequences can be found in EST databases.

Arabidopsis chromosomes (see URL http://preuss.bsd.uchicago.edu/arabidopsis.genome.html) that differ from conserved sequences found in the centromeres of other eukaryotes. Much work remains to be done to define the roles of different sequences in the varied functions of the centomere, and the sequence reported provides a foundation for these studies.

Gene families that are organised in tandem arrays of two or more units are relatively common. Analysis of the *Arabidopsis* genome revealed 1528 tandem arrays, containing up to 23 adjacent members and 4140 individual genes. These tandem arrays may have a functional significance by providing templates for sequence exchange that generates allelic diversity and permits new expression patterns to evolve from duplicated promoter regions. Larger segmental duplications were identified using methods that either aligned chromosomal sequences [23] or aligned proteins and searched for tracts of conserved gene order. Both methods demonstrated that about 58–60% of the *Arabidopsis* genome is present in 24 duplicated segments, each of more than 100 kb in size. The only duplicated segment found within the centromeric regions is a 375 kb segment on chromosome 4. Many duplications appear to have undergone further rearrangements, such as local inversions, after the duplication event. The extent of sequence conservation among the duplicated genes varies greatly, with 37% of the 17,193 genes found within the segmental duplication blocks being highly conserved (BLASTP score of $E<10^{-30}$) and a further 10% showing less significant ($E<10^{-5}$) similarity. The proportion of homologous genes in each duplicated segment also varies widely, and is between 20% and 47% for the highly conserved ($E<10^{-30}$) class of genes. In many cases, the number of copies of a gene and its counterpart (i.e. the matching partner of a gene or several genes in a duplicated region of the genome) differ (e.g. one copy on one chromosome and multiple copies on another); this could be caused by either tandem duplication or gene loss following the segmental duplication.

Polyploidy occurs widely in plants and is thought to be a key factor in plant evolution [24]. As the majority of the *Arabidopsis* genome is represented in duplicated (but not triplicated) segments, it appears most likely that *Arabidopsis* had a tetraploid ancestor. This has also been proposed on the basis of comparative sequence analysis conducted on the *Arabidopsis* genome and a segment of the tomato genome [25•]. The more recent duplication of the *Arabidopsis* genome postulated in that study was estimated to have occurred 112 million years ago, providing an estimate for the date of the possible tetraploidy event indicated by our data.

The extensive molecular divergence within the duplicated segments of the *Arabidopsis* genome have masked their proposed ancestry and have created a functional diploid. Nevertheless, these ancestral relationships imply that sets of genes found in duplicated segments could have redundant/duplicated gene function. For example, for the mutant phenotypes of genes encoding the SHATTERPROOF or SEPALLATA families of MADS-box transcription factors to become visible, all of the genes within the relevant family must be defective [26,27]. These functionally redundant genes are found in segmental duplications.

Protein function

The annotation of the sequenced *Arabidopsis* genes includes the definition of a functional category, which describes the cellular role of the gene product. Predicted

proteins were automatically assigned to these functional categories. These predictions were based on the yeast proteome and the assumption that sequences conserved between the yeast and *Arabidopsis* genomes reflect common functional relationships. A total of 25,498 predicted gene functions were analysed [8••].

Sixty-nine percent of *Arabidopsis* proteins had significant sequence similarity to proteins of known function in all organisms. In contrast, only 9% of the genes have been characterised experimentally. Approximately 30% of the predicted proteins, comprising both plant-specific proteins and proteins with similarity to genes of unknown function from other organisms, could not be assigned to functional categories. The substantial proportion of genes with predicted functions in metabolism, gene regulation and defence is consistent with previous analyses. Differing degrees of sequence conservation among proteins of related function in different organisms revealed both deeply conserved gene functions, such as those of genes involved in protein synthesis and chromatin maintenance, and the evolution of plant-specific proteins, which are involved in transcriptional control, signalling, and metabolism and energy transfer. Extensive lateral gene transfer from the ancestor of *Synechocystis* sp. [28] appears to have enriched the plant lineage in a variety of functions, which are predominantly carried out by proteins involved in metabolism and energy transfer but also by other proteins including those involved in signalling.

Pronounced redundancy was detected within segmental duplications and tandem arrays, and among genes scattered throughout the *Arabidopsis* genome. Using similarity exceeding a BLASTP value of $E<10^{-20}$ and extending over at least 80% of the protein length to identify protein families, a total of 11,601 protein types have been identified. Thirty-five percent of the predicted proteins are encoded only once in the genome, but the proportion of proteins present in families of more than five members is substantially higher in *Arabidopsis* (37.4%) than in *Drosophila* (24.0%) and *C. elegans* (12.1%). The absolute number of *Arabidopsis* gene families and singletons (i.e. the total number protein types) is in the same range as that of multicellular eukaryotes, indicating that a proteome of 11,000–15,000 types is sufficient for a wide diversity of multicellular life. The proportion of gene families with more than two members is considerably more pronounced in *Arabidopsis* than in the other eukaryotes whose gene sequence is known. Segmental duplication is responsible for 6303 gene duplications in *Arabidopsis* and accounts for a significant proportion of the increased family size. These features of the *Arabidopsis* genome, and presumably of other plant genomes, may indicate that the constraints on genome size are more relaxed or that the role of unequal crossing-over to generate new gene copies is more prominent in plants than in animals.

Conclusions

The sequence described in this review has the potential to change plant genetic analysis profoundly. Both forward and reverse genetics approaches are greatly simplified by the sequence and its analysis as mutations are more readily isolated and the full range of gene products is available for screening. The value of a sequenced genome for gene isolation is reviewed elsewhere [29•]. The extensive gene duplications identified in the *Arabidopsis* genome allow assessment of the likelihood of a particular gene being functionally redundant in advance of its functional analysis, thereby permitting such analyses to be designed appropriately. The high specificity conferred by nucleotide sequence, and the completeness of the survey, allows complex mixtures of RNA and protein to be resolved into their individual components using microarrays and mass spectrometry. This specificity can also be employed in the parallel analysis of genome-wide polymorphisms and quantitative traits in natural populations [30•]. Looking ahead, the challenge of determining the function of the large set of predicted genes, many of which are plant-specific, is now a clear priority [31], and multinational programs have been initiated to define gene function systematically on a large scale.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Pruitt RE, Meyerowitz EM: Characterization of the genome of ***Arabidopsis thaliana***. *J Mol Biol* 1986, 187:169-183.

2. Somerville CR, Ogren WL: Photorespiration deficient mutants of ***Arabidopsis thaliana*** lacking mitochondrial serine transhydroxymethylase activity. *Plant Physiol* 1981, 67:666-671.

3. Hwang I, Kohchi T, Hauge BM, Goodman HM, Schmidt R, Cnops G, Dean C, Gibson S, Iba K, Lemieux B *et al.*: Identification and map position of YAC clones comprising one-third of the ***Arabidopsis*** genome. *Plant J* 1991, 1:367-374.

4. Bevan M, Bancroft I, Bent E, Love K, Goodman H, Dean C, Bergkamp R, Dirkse W, Van Staveren M, Stiekema W *et al.*: Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of ***Arabidopsis thaliana***. *Nature* 1998, 391:485-488.

5. • Marra M, Kucaba T, Sekhon M, Hillier L, Martienssen R, Chinwalla A, Crockett J, Fedele J, Grover H, Gund C *et al.*: A map for sequence analysis of the ***Arabidopsis thaliana*** genome. *Nat Genet* 1999, 22:265-270.

The authors report the development of a key technology, high-throughput mapping of BAC clones, that was required for rapid and efficient sequencing of the *Arabidopsis* genome.

6. • Mayer K, Schuller C, Wambutt R, Murphy G, Volckaert G, Pohl T, Dusterhoft A, Stiekema W, Entian KD, Terryn N *et al.*: Sequence and analysis of chromosome 4 of the plant ***Arabidopsis thaliana***. *Nature* 1999, 402:769-777.

The authors announce the completion of the sequencing of chromosome 4 from *Arabidopsis*. This chromosome and chromosome 2 [7•] from *Arabidopsis* were the first plant chromosomes to be sequenced completely.

7. • Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M *et al.*: Sequence and analysis of chromosome 2 of the plant ***Arabidopsis thaliana***. *Nature* 1999, 402:761-768.

The authors announce the completion of the sequencing of chromosome 2 from *Arabidopsis*. This chromosome and chromosome 4 [6•] from *Arabidopsis* were the first plant chromosomes to be sequenced completely.

8. •• The *Arabidopsis* Genome Initiative: Sequence and analysis of the flowering plant ***Arabidopsis thaliana***. *Nature* 2000, 408:796-815.
The authors announce the completion of the sequencing of the *Arabidopsis thaliana* genome. This is the first plant genome to be completely sequenced.

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

10. Lukashin AV, Borodovsky M: GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* 1998, 26:1107-1115.

11. Uberbacher EC, Mural RJ: Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc Natl Acad Sci USA* 1991, 88:11261-11265.

12. Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H: Interpolated Markov models for eukaryotic gene finding. *Genomics* 1999, 59:24-31.

13. Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S: Splice site prediction in ***Arabidopsis thaliana*** DNA by combining local and global sequence information. *Nucleic Acids Res* 1996, 24:3439-3452.

14. Brendel V, Kleffe J: Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in ***Arabidopsis thaliana*** genomic DNA. *Nucleic Acids Res* 1998, 26:4748-4757.

15. Quackenbush J, Liang F, Holt I, Pertea G, Upton J: The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res* 2000, 28:141-145.

16. Huang X, Adams MD, Zhou H, Kerlavage AR: A tool for analyzing and annotating genomic sequences. *Genomics* 1997, 46:37-45.

17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215:403-410.

18. •• Pavy N, Rombauts S, Dehais P, Mathe C, Ramana DVV, Leroy P, Rouze PR: Evaluation of gene prediction software using a genomic data set: application to ***Arabidopsis thaliana*** sequences. *Bioinformatics* 1999, 15:887-900.
A critically important assessment of the performance of gene analysis tools used in the *Arabidopsis* genome project. This is required reading for all users of the genome sequence as it clearly explains the criteria used to define gene structures.

19. The *C. elegans* Sequencing Consortium: Sequence and analysis of the genome of ***C. elegans***. *Science* 1998, 282:2012.

20. •• Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al.*: The genome sequence of ***Drosophila melanogaster***. *Science* 2000, 287:2185-2195.
A major milestone in biology and a technical *tour de force*. The major value of sequence is only realised upon comparison with sequence from another organism, therefore this work adds greatly to the initial interpretations of the *Arabidopsis* sequence.

21. • Le QH, Wright S, Yu Z, Bureau T: Transposon diversity in ***Arabidopsis thaliana***. *Proc Natl Acad Sci USA* 2000, 97:7376-7381.
The first comprehensive analysis of transposons in a sequenced genome. The relationships between elements and the activities of mobile genetic elements in shaping the *Arabidopsis* genome are described.

22. •• Copenhaver GP, Nickel K, Kuromori T, Benito M-I, Kaul S, Lin XY, Bevan M, Murphy G, Harris B, Parnell LD *et al.*: Genetic definition and sequence analysis of ***Arabidopsis*** centromeres. *Science* 1999, 286:2468-2474.
The authors describe the first detailed analysis of higher eukaryotic centromeres.

23. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL: Alignment of whole genomes. *Nucleic Acids Res* 1999, 27:2369-2376.

24. Wendel JF: Genome evolution in polyploids. *Plant Mol Biol* 2000, 42:225-249.

25. • Ku H-M, Vision T, Liu J, Tanksley SD: Comparing sequenced segments of the tomato and ***Arabidopsis*** genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc Natl Acad Sci USA* 2000, 97:9121-9126.
A publication describing a path-finding study that used inferred gene order to establish relationships among the restless genomes of crop plants.

26. Liljegren SJ, Ditta GS, Eshed Y, Savidge B, Bowman JL, Yanofsky MF: ***SHATTERPROOF*** MADS-box genes control seed dispersal in ***Arabidopsis***. *Nature* 2000, 404:766-770.

27. Pelaz S, Ditta GS, Baumann E, Wisman E, Yanofsky MF: B and C floral organ identity functions require SEPALLATA MADS-box genes. *Nature* 2000, 405:200-203.

28. Kotani H, Tabata S: Lessons from the sequencing of the genome of a unicellular cyanobacterium, ***Synechocystis*** SP. PCC6803. *Annu Rev Plant Physiol Plant Mol Biol* 1998, 49:151-171.

29. • Lukowitz W, Gillmor CS, Scheible W-R: Positional cloning in ***Arabidopsis***. Why it feels good to have a genome initiative working for you. *Plant Physiol* 2000, 123:795-805.
A thorough primer describing how genome sequence can be used for gene isolation.

30. • Alonso-Blanco C, Koornneef M: Naturally occurring variation in ***Arabidopsis***: an underexploited resource for plant genetics. *Trends Plant Sci* 1999, 5:1360-1385.
A thoughtful review of the extent of natural variation in *Arabidopsis* and how to exploit it using genome sequence.

31. Chory J, Ecker JR, Briggs SR, Caboche M, Coruzzi G, Cook D, Dangl J, Grant S, Guerinot ML, Henikoff S *et al.*: Functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. *Plant Physiol* 2000, 123:423-425.