PublisherInfo		
PublisherName	:	BioMed Central
PublisherLocation	:	London
PublisherImprintName	:	BioMed Central

Arabidopsis chromosome 4 sequence

ArticleInfo		
ArticleID	:	3564
ArticleDOI	:	10.1186/gb-2000-1-1-reports030
ArticleCitationID	:	reports030
ArticleSequenceNumber	:	55
ArticleCategory	:	Paper report
ArticleFirstPage	:	1
ArticleLastPage	:	5
ArticleHistory	:	RegistrationDate : 2000–2–8 Received : 2000–2–8 OnlineDate : 2000–4–27
ArticleCopyright	:	BioMed Central Ltd2000
ArticleGrants	:	
ArticleContext	:	130591111

Todd Richmond

Abstract

Members of the *Arabidopsis* Genome Initiative, including the European Union *Arabidopsis* Genome Sequencing Consortium and The Cold Spring Harbor, Washington University in St Louis and PE Biosystems *Arabidopsis* Sequencing Consortium, have completed sequencing one of the first two plant chromosomes.

Significance and context

Arabidopsis thaliana is the model organism of choice for modern plant biologists. Its small genome, the excellent genetic and physical maps of the genome and the lack of large amounts of repetitive DNA made it the first choice for plant genome sequencing. In 1996, a multinational organization, the *Arabidopsis* Genome Initiative (AGI), was formed to coordinate the worldwide effort to sequence the first higher plant. Made up of labs from the United States, Europe and Japan, AGI set a goal for the completion of the *Arabidopsis* genome by 2004. The members divided up the five chromosomes between them and began sequencing. Advances in sequencing and computing technology have pushed forward their initial timetable, however, and two papers report the completion of the first two plant chromosomes. Mayer *et al*. report the sequencing of chromosome 4, which represents about 17% of the genome. In a companion paper in the same issue of *Nature*, Lin *et al*. report the sequencing of chromosome 2, which represents approximately 15% of the *Arabidopsis* genome. These sequences are two of the largest pieces of DNA sequence ever assembled and together represent almost a third of the *Arabidopsis* genome. The complete sequences of chromosomes 2 and 4 offer unique insights into large-scale genomic organization, plant heterochromatic DNA, non-coding regions, gene duplication events, and gene family organization.

Key results

The paper summarizes years of work by hundreds (if not thousands) of people in dozens of labs spread over three continents. The key features of chromosome 4 are as follows. The long arm of chromosome 4 is 14.5 Mb, the short arm is 3.0 Mb (plus nearly 3.5 Mb of ribosomal DNA repeats). Nearly 50% of the sequence encodes for protein, for a total of 3,744 predicted proteins. Each gene is about 4.6 kb in length, containing an average of 5.2 exons. The actual or potential cellular function for approximately 60% of the genes can be predicted on the basis of similarity to other characterized

proteins. Only 33% of the predicted genes are represented among the available 45,000 *Arabidopsis* expressed sequence tags (ESTs). Of these, 6% of the genes match 75% of the ESTs. Note that it is not clear if the authors are referring at this point only to the chromosome 4 sequence or to all *Arabidopsis*sequence available; it is clearly important to sequence normalized EST libraries in order to maximize the amount of non-redundant sequence gathered. Almost 8% of the predicted genes have no ESTs and no similarity to other proteins; these may represent spurious gene predictions or plant-specific genes expressed at low levels.

The authors give some statistics on various motifs and structural topologies found in the predicted proteins. They also attempt to classify the proteins into major functional categories (such as metabolism and transcription). The only major surprise is the large number of genes involved in disease and defense responses. This is largely due to several large clusters of leucine-rich repeat genes, including one family of 15 contiguous genes. A surprisingly large number of genes are arranged in tandem copies. Of genes with products that have significant similarity to other proteins in *Arabidopsis*, 12% are arrayed in tandem clusters, ranging from pairs of genes to the 15 leucine-rich repeat genes. This hints at the underlying mechanism of how plants generate sequence diversity and evolve new metabolic and regulatory functions. The final section covers the centromeric heterochromatin region of chromosome 4, a roughly 4 Mb region. The majority of repetitive elements in the *Arabidopsis* genome are found in the centromeric regions of the chromosomes, and include dispersed repeats, representatives of most of the transposons found in other plant species, long terminal repeat (LTR) and non-LTR retroelements, and *Athila* retroelements. Although every description of *Arabidopsis thaliana* as a model organism cites the fact that it has very low amounts of repetitive DNA, this paper and the accompanying description of chromosome 2 offer a highly detailed characterization of its repetitive elements. These results are, for the most part, consistent with previous findings and estimates.

Links

Information on *Arabidopsis* and its genome sequence is available from The *Arabidopsis* Information Resource (TAIR), MIPS *Arabidopsis thaliana* database (MATDB), Kazusa *Arabidopsis* data opening site (KAOS), The Institute for Genomic Research's *Arabidopsis thaliana* annotation database. Sequence-based, genetic and physical maps of the *Arabidopsis* genome are available from the Cold Spring Harbor Laboratory.

Reporter's comments

This paper, and its companion, just begin to scratch the surface of the overwhelming amount of information contained in the sequence of two plant chromosomes. Comparing this paper with the chromosome 2 paper, the interest of the primary authors is clear. While Lin *et al*. emphasize chromosome structure and organization, Mayer *et al*. appear more interested in protein-coding sequences and the functional classification of the predicted proteins. Once the entire *Arabidopsis* genome (ecotype Columbia) is completed and the complete sequence of the Landsberg ecotype is

released to researchers, a flood of papers is likely to overwhelm the plant community. In a sense, this is somewhat frustrating, as we will be forced to rely on others to analyze and summarize the important information, even though not all researchers agree on what data is important, how to present it or how to interpret it. As a case in point, both the chromosome 2 and the chromosome 4 papers make reference to the large duplication event shared by the two chromosomes. Lin *et al*. report that this duplication is 4.6 Mb long, with several translocations or inversions, encompassing a total of 1,100 genes. Mayer *et al*. report that the two chromosomes share four blocks of conserved sequence, two of which are inverted, totaling 2.5 Mb in length. Which interpretation is correct? Another source of frustration is the lack of consistency in reporting results. While Lin *et al*., reporting on chromosome 2, place more emphasis on the overall chromosome structure and the organization and distribution of various elements, Mayer *et al*. place more emphasis on the genes, predicted functions and structural components. It makes this reporter wish that the two teams had coordinated with one another, divided up the various areas of interest and done a complete report on those areas for both chromosomes. For now, we must be satisfied with a partial analysis. The upcoming completion of the *Arabidopsis* genome will truly be a landmark. For the first time, we will have the complete genetic blueprint for a flowering plant. The initial data, especially from Lin *et al*., suggest that all plants have a common set of genes for many functions. It is already clear that many of the genes found in other plants are present in *Arabidopsis*, even when comparing across the monocot/dicot and angiosperm/gymnosperm divisions. But until another plant, such as rice, is completely sequenced, it will be difficult to evaluate the size of that set of common genes. With the *Arabidopsis* genome expected to be finished by the end of the year, we can then move on to the more complex area of functional genomics and begin to elucidate the function of the estimated 25,000 proteins that make up a flowering plant.

Table of links

References

1. Mayer K, Schuller C, Wambutt R, Murphy G, Volckaert G, Pohl T, Düsterhoft A, Stiekema W, Entian K-D, Terryn N, et al: Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature. 1999, 402: 769-777. 0028-0836