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

Arabidopsis chromosome 2 sequence

ArticleInfo		
ArticleID	:	3563
ArticleDOI	:	10.1186/gb-2000-1-1-reports029
ArticleCitationID	:	reports029
ArticleSequenceNumber	:	54
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, primarily from The Institute for Genomic Research, 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. Lin *et al*. report the sequencing of chromosome 2, which represents approximately 15% of the *Arabidopsis* genome. In a companion paper in the same issue of *Nature*, Mayer *et al*. report the sequencing of chromosome 4, which represents about 17% of the 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 2 are as follows. The long arm of chromosome 2 is 16.0 Mb, the short arm 3.5 Mb. Nearly 50% of the sequence codes for protein, with a total of 4,037 predicted proteins. Each gene is about 4.4 kb in length, containing an average of 4.6 exons. The largest gene contains 52 predicted exons and is 50% identical to a human protein. The actual or potential cellular function for approximately 52% 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 45,000 available *Arabidopsis* expressed sequence tags (ESTs). After classifying the predicted proteins

into functional classes, the largest functional groups were genes involved in regulatory function and signal transduction (including DNA-binding proteins, transcription factors and protein kinases). The most frequent protein domains were leucine-rich repeats, protein kinases and zinc-finger domains. More than 60% of the predicted gene products (2,542) on chromosome 2 have significant similarity to another *Arabidopsis* protein. The products of most of the genes that have paralogs (83%) within the *Arabidopsis* genome are more similar to their paralog than to proteins from other completed sequences. Of these, 593 are found in tandem duplications that range in size from two to nine genes. The same phenomenon was seen in the analysis of chromosome 4. Lin *et al.* present a number of graphs that show the distribution of features along chromosome 2, summarizing predicted gene density, EST density, tandem duplications and repetitive elements. As expected, gene density decreases as you approach the centromere and the amount of repetitive DNA increases. There is a table of the transposable elements found on chromosome 2, broken down into class, subclass and family. Unfortunately, there is no equivalent table for other types of repetitive DNA element.

Most interesting, however, are the discoveries that can only come from the analysis of the genome as a whole. Of these, the large duplication events between chromosomes are the most unexpected. The largest is a 4.6 Mb region in chromosomes 2 and 4, in which 39% of the genes (430 out of 1,100) are duplicated between the two chromosomes. Another duplication, 0.7 Mb long, occurs on chromosomes 1 and 2, in which 33% of the genes (57 out of 170) are duplicated. These duplications account for part, but not all, of the high percentage of genes with paralogs. Chromosome 2 has another unusual feature. It is well established that individual genes can be transferred from organelles to the nucleus. It is nonetheless surprising to find a stretch of 270 kb of mitochondrial sequence in the genetically defined centromere region of chromosome 2. This inserted sequence is larger than any sequence previously reported to have undergone organelle-to-nuclear transfer (almost 75% of the mitochondrial genome). The sequence identity (99%) suggests that this transfer event was very recent. When the sequence of the Landsberg *Arabidopsis* ecotype is made available later this year, it will be interesting to see if this same insertion event is present.

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) and TIGR'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 4 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. 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

Nature

The *Arabidopsis* Information Resource

MIPS *Arabidopsis thaliana* database

Kazusa *Arabidopsis* data opening site

Arabidopsis thaliana annotation database

Sequence-based, genetic and physical maps of the *Arabidopsis* genome

References

1. 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. 0028-0836