Meeting report

Follow that plant!

Pablo D Rabinowicz and Mary E Byrne

Address: Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA.

Correspondence: Pablo D Rabinowicz. E-mail: rabinowi@cshl.org

A report on the talks presented at the Cold Spring Harbor 2000 Meeting on *Arabidopsis* Genomics, New York, 7-10 December, 2000.

It may be difficult to convince a lay person that the genome sequence of a little weed called *Arabidopsis thaliana* is not only providing an invaluable resource for understanding plant biology but is also serving as a model system for improvement of economically important crop species. An even more difficult task is to persuade people that the *Arabidopsis* Genome Initiative became an example of technology development, later used for the Human Genome Project, and that the sequence of the *Arabidopsis* genome might be useful as a model to study (and eventually, even help cure) human diseases. Researchers called to the National Science Foundation headquarters to give a press conference highlighted these issues when making the public announcement of the completion of the *Arabidopsis* genome sequence. One week before this announcement, these issues were discussed at the Cold Spring Harbor 2000 Meeting on *Arabidopsis* Genomics.

Databases and centralized resources

As for other complete genomes, annotation of the *Arabidopsis* genome generated substantial discussion. The case of *Drosophila melanogaster* was described by the plenary speaker Michael Ashburner (EMBL-EBI, Cambridge, UK). Because a whole-genome shotgun-sequencing approach was used for *Drosophila*, the annotation process was different from the procedure used for *Arabidopsis*. Initially, a 3 Mbp region of the *Drosophila* genome was annotated independently by several groups over 18 months. This genome annotation assessment project (GASP) allowed the consortium to decide which tools would be most useful for annotating the whole genome when it was finished.

In the case of *Arabidopsis,* the annotation will be now centralized at The Institute for Genome Research (TIGR, Bethesda, USA), the Munich Information Center for Protein Sequences (MIPS, Martinsried, Germany) and the Kazusa DNA Research Institute (KDRI, Japan). These institutes have received accurate sequences from bacterial artificial chromosome (BAC) clones, often annotated in a detailed way after manual curation. This strategy resulted in richer and more precise information than would a fully automated approach carried out on the complete genome. The current annotation is heterogeneous, posing a problem for global electronic analysis of the annotated data. A more automated system to add uniformity to the annotation is being developed, although a complete 're-annotation' will not be possible. An alternative to the centralized annotation was proposed by Lincoln Stein (Cold Spring Harbor Laboratories (CSHL), New York, USA). The distributed annotation system (DAS) incorporates expert information on each gene or genomic feature derived from all members of the community. In this system, each investigator curates the annotation of his or her favorite gene using a unified format and all this information is collated in a reference computer server.

The *Arabidopsis* Information Resource (TAIR) [http://www.arabidopsis.org/home.html], as discussed by Margarita Garcia-Hernandez (Carnegie Institution, Stanford, USA) is an attempt to provide the *Arabidopsis* community with a comprehensive and integrated database. TAIR contains extensive genomic information including clones, genes, both older genetic and visible marker maps and more recent AGI sequence maps, as well as some plant literature. In the future, additional features to be incorporated are gene function, gene and protein expression data, and so on. The idea of collecting, in a single website, data from many different labs, institutes and resource centers, raised the issue of intellectual property. In this regard, a concern was expressed

by TIGR and MIPS, who feel their effort should receive the proper credit if used by others.

Over the past few years, a growing number of plant populations mutated with T-DNA, transposons or ethylmethane sulfonate (EMS) have been generated to facilitate functional genomics. Many of these are available through the *Arabidopsis* Biological Resource Center (ARBC, Ohio State University, Columbus, USA), as outlined by Randy Scholl (Ohio State University). Resources created more recently were also presented. Ottoline Leyser (University of York, UK), introduced GARNet [http://garnet.arabidopsis.org.uk], the UK *Arabidopsis* functional genomics network. Michel Caboche (INRA, Versailles, France) described GENOPLANTE, a comprehensive program that includes genomic sequencing as well as functional genomics involving several plant species. Rob Martienssen (CSHL) introduced a database for transposon-based enhancer- and gene-trap systems, which includes systematic sequencing of transposon insertion sites.

Tools for gene identification

One powerful tool for gene identification is transposon tagging. Using this technique, Michael Snyder (Yale University, New Haven, USA) showed that 600 previously non-annotated genes were found in the genome of *Saccharomyces cerevisiae*, many of which had less than 100 amino acids. Such small open reading frames are typically excluded from annotation routines, and this criterion was also used during *Arabidopsis* annotation. As in yeast, transposon tagging is an important tool for the identification of new genes in *Arabidopsis*, and a number of collections of transposon lines are available. Moreover, Dick McCombie (CSHL) proposes to sequence other Brassica species, which will provide a resource for gene discovery. Still to be decided is which Brassica species would be most useful to sequence. For evolutionarily distant species, however, comparative genomics may not be so promising. In comparing the partial tomato sequence with that of *Arabidopsis,* Steve Tanskley (Cornell University, Ithaca, USA) found that a high proportion of tomato genes have a significant match in the *Arabidopsis* genome. But most of the 'hits' correspond to members of gene families, which complicates the identification of orthologous genes. Orthology could be determined for only 4% of the genes analyzed, and little colinearity was observed between the two genomes with respect to these genes.

As pointed out by Hans-Werner Mewes (MIPS, Martinsried, Germany), somewhat unexpectedly, the sequence of the *Arabidopsis* genome showed a high degree of gene duplication. Some genes are found in tandem duplications or multiple copies, and large chromosomal regions are found more than once in the same or different chromosomes. This no doubt contributes a degree of genetic redundancy. Ashburner argued, however, that truly redundant genes (ones with identical function) are unlikely, because they would not be maintained by natural selection. Redundancy may be observed in the case of recent evolutionary events. Whatever the cause, Owen White (TIGR, Bethesda, USA) proposed taking advantage of gene duplication to correct gene modeling annotation errors.

Functional studies

Arabidopsis is now well into the 'post-genome' era, as shown by the substantial number of presentations describing the use of genomic tools to study a wide variety of biological processes. For example, Jeffery Dangl (University of North Carolina, Chapel Hill, USA) used *Arabidopsis* DNA microarrays to identify groups of genes coordinately regulated during the onset of systemic acquired resistance and then, taking advantage of the genome sequence, determined the regulatory sequences in the promoters responsible for each pattern of expression. Stacey Harmer (Scripps Research Institute, La Jolla, USA) applied similar resources to identify the mechanism of circadian control of genes and their regulatory regions. Phil Benfey (New York University, USA) developed an algorithm to draw transcription factor networks using microarrays and sequence data; this will be applied to different stages of root development. Pam Green and Rodrigo Gutiérrez (Michigan State University, East Lansing, USA) are using cDNA microarrays from the *Arabidopsis* Functional Genomics Consortium (AFGC [http://afgc.stanford.edu], which includes a facility providing *Arabidopsis* cDNA microarrays containing 11,000 genes) to classify genes regulated by mRNA stability. Daphne Preuss (University of Chicago, USA) could identify centromeric genes and analyze in detail their expression and methylation patterns thanks to the availability (unique to *Arabidopsis*) of deep coverage of heterochromatic sequence. David Galbraith (University of Arizona, Tucson, USA) is attempting to assign a function to each of the cytochrome P450 proteins, which are encoded by a family of nearly 300 genes in *Arabidopsis*, by reverse genetics and microarray expression profiling under different biotic and abiotic treatments.

The private sector presented a focus on the implementation of functional genomics. Ken Feldman (Ceres Inc., Malibu, USA) described progress on obtaining full-length cDNA sequences from *Arabidopsis,* which are being used for several functional genomics approaches. Information from a database of 8,000 cDNA sequences has provided knowledge of general sequence features and will be useful for modifying gene-prediction programs, as sequence from cDNAs indicates that only 60% of genes are correctly predicted. In another cDNA sequencing program, Gary Temple (Life Technologies/Invitrogen Corp., Rockville, USA) in collaboration with GENOSCOPE (Evry, France) described a versatile system to normalize a cDNA library and generate full-length cDNA sequences, which will be publicly available. In order to overcome the problem that many mutations may

only create subtle phenotypic effects, Keith Davis (Paradigm Genetics, Inc., Research Triangle Park, USA) described a high-throughput platform collating phenotypic data from 100 traits measured at predetermined stages of plant growth and development. Jack Okamuro (Ceres) presented a similarly detailed phenotypic analysis of fruit development.

Biological diversity

The complete genome sequence has made it possible to generate many more markers for mapping quantitative trait loci, and a number of groups are identifying loci that contribute to natural variation among *Arabidopsis* strains and close relatives, with the expectation that the genes identified will provide information not readily derived by mutagenesis. Insect resistance is one such variable among naturally occurring populations of *Arabidopsis*. Thomas Mitchell-Olds (Max Planck Institute, Jena, Germany) described genotypic variation in enzymes of the glycosinolate biosynthetic pathway, products of which confer insect resistance. One enzyme in this pathway, which controls glycosinolate chain length, was found to be absent in the Landsberg *erecta* strain and may have undergone gene conversion or recombination with a closely related gene. Another trait that varies significantly between different strains is hypocotyl length, as described by Detlef Weigel (Salk Institute, La Jolla, USA). Interestingly, the variation in hypocotyl length of different strains under various light conditions could be correlated with the incident sunlight where a strain typically grows. Cluster analysis also showed that several determinants of hypocotyl length map to genes known from their mutant phenotypes to affect hypocotyl length. Ben Bowen (Lynx Therapeutics Inc., Hayward, USA) described quantitative trait loci associated with nitrogen utilization. Several candidate genes were defined using massively parallel signature sequencing (MPSS) technology, in which short signature sequences are obtained from cDNAs and attached to microbeads. All these studies should contribute more to our understanding of evolutionary processes, defining whether mechanisms of adaptation principally involve changes in enhancers or protein coding regions and whether such changes predominantly occur in regulatory or basic cellular components.

New technologies

Keeping up with the leadership of *Arabidopsis* in technology development, two new technologies are being applied to it. One, presented by Michael Sussman (University of Wisconsin, Madison, USA), is a microarray technology called the maskless array synthesizer (MAS). It is an oligonucleotide microarray construction system based on a digital micromirror device designed by Texas Instruments. This device generates successive virtual masks on a slide for solid-phase oligonucleotide synthesis. These arrays are faster to build and cheaper than commercially available

ones. The second is called targeting induced local lesions in genomes (TILLING) and was presented by Steve Henikoff (Fred Hutchinson Cancer Research Center, Seattle, USA). This EMS mutagenesis system allows identification of point mutations in known genes after denaturing high-performance liquid chromatography of denatured and re-annealed PCR fragments from mutant and wild-type plants. Henikoff is carrying out high-throughput TILLING and will provide the results to the community.

Perspectives

Undoubtedly, the *Arabidopsis* genome sequence is, so far, the most comprehensive compared with other higher eukaryotes for which the genome sequence has been completed. Although its sequence shows the deepest coverage of heterochromatic regions, it was repeatedly referred to as 'almost' or 'nearly' complete during the meeting. This was not because of the few remaining sequence gaps, but because of the determination of the researchers to hold their excitement until the release of the 14 December 2000 issue of *Nature* in which the annotated sequence was to be published. Holding back the celebration of the achievement of this milestone for five days was not too difficult, considering that the job was completed five years earlier than had been initially planned.

The milestone achieved by this group of plant biologists will certainly change the way we study not only plant biology but also biology in general. The approach for sequencing the *Arabidopsis* genome resulted in the availability of the BAC-based physical map, which was an invaluable tool long before the genome sequence was finished. The approach proved so successful that it became an example followed by the Human Genome Project. The idea that a weed can improve human health is not far fetched. Plants are not only a source of food but a source of drugs for treating diseases. Moreover, the discovery of *Arabidopsis* genes homologous to mammalian cancer-related genes opens up the possibility of using a plant as a model to study the basis of human diseases as complex as cancer. The outstanding success of the *Arabidopsis* Genome Initiative is to be followed by sequencing projects aimed at increasingly complex plant genomes.