GENOMIC ANALYSIS OF ARABIDOPSIS THALIANA

Takayuki Kohchi
Nara Institute of Science and Technology, Japan

Keywords: *Arabidopsis thaliana*, chromosome structure, genome project, molecular genetics, physical map, nuclear genes, chloroplast DNA, mitochondrial DNA

Contents

Summary

Genetics began with the rediscovery of Mendel's rules of inheritance at the beginning of the twentieth century, for which the garden pea was used as the model organism. Another fundamental finding in genetics was the discovery of transposons, movable elements explaining epigenetics that were first studied in maize by Barbara McClintock in the mid-twentieth century. Finally, the twentieth century ended with the determination of the complete nucleotide sequences of the nuclear, plastid, and mitochondrial genomes for the small crucifer weed, *Arabidopsis thaliana*. *Arabidopsis* has a nuclear genome 125 Mb in length, encoding 25 500 genes on five chromosomes. The completeness of the genome information of *Arabidopsis* represents the first such accomplishment in plants, and provides an excellent resource for understanding the genetic basis of plant behavior and for the molecular breeding of crops. This achievement makes *Arabidopsis* a very attractive model for plant molecular genetics, development, metabolism, cellular signaling, and physiology. In this review, the history, current status, and perspectives of *Arabidopsis* genomics are comprehensively examined.

1. Introduction: Why is Arabidopsis a Model Plant Species?

The small crucifer *Arabidopsis thaliana* is among the most popular and useful model plant species in modern biology due to its remarkable genetic tractability (Figure 1). Intuitively, the direct study of human crops may seem preferable to the analysis of weeds such as *Arabidopsis*, since such work would be immediately beneficial to agriculture and horticulture. However, crops have been modified and domesticated for

many generations to select for increased utility as food and agricultural resources compared to wild-type progenitors. The selection of traits was intended for practical use, not for plant scientists. As a result, crops have a genetic bias for agricultural purposes, such as increased yield, growth, and development. Typical crop plants have one generation a year and require large fields to grow in numbers sufficient for genetic analysis. Polyploidy and allopolyploidy confuses the genetics of crops. In contrast, *Arabidopsis* has the opposite characteristics, such as a short generation time and small size, making it a good choice for laboratory study. The first seed sets are obtained six weeks after germination. For some purposes, such as M2 seed harvesting, hundreds of plants can be grown in a growth chamber, facilitating genetic screening. Mutant screening can be done even in Petri dishes if the mutant phenotype can be scored at germination or in seedlings. It is easy to maintain mutant lines, since *Arabidopsis* is self-fertile and has more than 10 000 seed sets in a single plant. These features promote efficient genetic analysis of various aspects of plant growth, development, and proliferation.

Figure 1. *Arabidopsis thaliana* (L.) Heynh. *Arabidopsis* is an annual long-day plant belonging to a family of the crucifers. In the vegetative stage, its leaves form rosette structures. During the transition to the reproductive phase, bolting of inflorescence stems can be observed from the shoot apex and from leaf axils. Flowers are formed as lateral organs in inflorescence stems. In each flower, there are typically four sepals, four white petals, six stamens, and two fused carpels. This picture was kindly provided by Ken Nakahara (Nara Institute of Science and Technology).

Transgenic experiments are important for analyzing gene function *in vivo*. *Arabidopsis* is susceptible to infection with *Agrobacterium tumefaciens*, an important soil bacterium harboring a Ti-plasmid utilized for genetic manipulation of plants. Previously, the transformation of *Arabidopsis* was not easy because its regeneration efficiency in tissue culture was not as high as that of other dicots such as tobacco and carrot. Recently, *in planta* transformation, performed simply by dipping inflorescence into *Agrobacteruium* culture, has become routine in *Arabidopsis*, making it one of the host plants most amenable to introduction of foreign genes for genetic manipulation.

Arabidopsis has many advantages, not only in basic biology but also in applied studies in agriculture. For agricultural applications, *Arabidopsis* is a weed, not a crop. However, most of the metabolic, physiological, and developmental processes as well as their underlying genetic mechanisms in *Arabidopsis* are also prevalent in crop plants. (One should note, however, that *Arabidopsis* does not adequately model certain unique properties, such as nitrogen-utilization in legumes and C4 photosynthesis in maize.) Although there are reasons to study individual species, many genetic traits of crop domestication and further breeding potential can be studied using *Arabidopsis* as a model.

2. Small Size and Simple Organization of the *Arabidopsis* Genome

In addition to the many advantages described above, the rapid rise of *Arabidopsis* as a preferred plant model is based upon its small genome size and few repetitive sequences.

The framework for *Arabadopsis* genomics was set by early kinetic analyses of total *Arabidopsis* DNA in 1984. Most of the genome was found to be composed of single-copy sequences (50–57%), with repetitive sequences (23–27%) derived primarily from chloroplast DNA. The haploid *Arabidopsis* nuclear genome was calculated to be 70 Mb. Compared to more recent estimates of 115 Mb for single-copy DNA, and 125 Mb for the entire genome, this value calculated in the 1980s was underestimated, but it attracted the attention of plant molecular biologists. The content of the *Arabidopsis* genome was also analyzed at the molecular level by quantitative gel blotting using randomly selected lambda clones as probes. DNA sequences in the *Arabidopsis* genome were composed predominantly of single-copy sequences with small repetitive sequences (0.8%) and rDNA repeats. The deduced genome size, 70 Mb, correlated roughly to the genome size calculated from nuclear volume, and represented the smallest published size in flowering plants. The contrast between its size and that of familiar plants was striking (see Table 1).

	Gene	Gene product	Function
Protein-coding genes	*cob*	Apocytochrome *b*	Respiratory electron transport
	coxI	Cytochrome oxidase subunit I	Respiratory electron transport
	nad1	NADH dehydrogenase subunit	Respiratory electron transport
	nad2	NADH dehydrogenase subunit	Respiratory electron transport
	nad4	NADH dehydrogenase subunit	Respiratory electron transport
	nad5	NADH dehydrogenase subunit	Respiratory electron transport
	nad6	NADH dehydrogenase subunit	Respiratory electron transport
	rtl	Reverse transcriptase-like protein	
rRNA genes	L1-L8	Large rRNA gene	Protein synthesis
	S1-S4	Small rRNA gene	Protein synthesis
tRNA genes	*trnW*	Tryptophan tRNA	Protein synthesis
	trnQ	Glutamine tRNA	Protein synthesis
	trnM	Methionine tRNA	Protein synthesis

* Respiratory electron transport

Table 1. Haploid genome size in various plants. The genome sizes were calculated from kinetic complexity measurements when molecular information was not available. The value for human genome size is also included for comparison.

The benefit of a small genome is clear. For example, 20 000 lambda clones in a genomic library is statistically and practically sufficient to isolate genomic clones in *Arabidopsis*, while millions of clones must be screened for plant species with a huge genome such as wheat and pea.

For Southern blotting, less than 1 microgram of DNA is sufficient to detect a single-copy sequence in *Arabidopsis*, making it easy to perform RFLP (restriction fragment length polymorphism) mapping, an important method used to link recombinant DNA techniques to genetics. Although researchers have preferred recently to use polymerase chain reaction (PCR)-based molecular markers for mapping, PCR amplification of DNA from small amounts of tissue, such as a single leaf, is also easier due to the small genome. This characteristic makes *Arabidopsis* one of the best model systems for plant molecular study.

3. Chromosomes and Maps

The initial work in *Arabidopsis* genetics was performed in 1907 in Germany by Leibach, who established that *Arabidopsis* is a diploid plant with two sets of five different chromosomes (2n = 10). The first boom in the isolation of visible mutations occurred in the 1960s and 1970s, when classical genetic mapping with mutational markers was carried out in *Arabidopsis*. These loci were assigned to five linkage groups in 1983 by Koornneef et al. in the Netherlands. A genetic map for the five chromosomes was constructed based on recombination frequencies between markers, resulting in genetic lengths of chromosomes from 51 cM (chromosome 2) to 123 cM (chromosome 1).

Although mutation markers are simple and require no molecular analysis, few markers can be used in a single cross. In contrast, molecular markers, such as RFLPs, random amplified polymorphic DNAs (RAPDs), simple sequence length polymorphisms (SSLPs), and cleaved amplified polymorphic sequences (CAPSs) are codominant and easy to use in combination. Initial genetic maps with RFLP markers were generated in the late 1980s by two laboratories in the United States. After the development of more markers, a single genetic map was obtained by the use of “recombinant inbred” lines, populations descended from a single seed. Molecular markers can be integrated into a physical map. These mapping data as well as other useful information for *Arabidopsis* studies can be accessed through the *Arabidopsis* Information Resource (TAIR; <http://www.arabidopsis.org/>).

The effort to construct a physical map started in the mid-1980s. Initially, a cosmid-fingerprinting strategy was employed. It succeeded partially, generating around 700 cosmid contigs covering 91 Mb after collection of fingerprints from approximately 20 000 cosmid clones by Goodman’s group in the United States. Since the practical limit of a cosmid-based strategy was reached at that point, yeast artificial chromosomes (YACs) and bacterial artificial chromosomes (BACs) were incorporated into the mapping effort. In the 1990s, using molecular markers as anchoring probes, a physical map using YACs and BACs was also constructed as part of the Multinational Coordinated *Arabidopsis* Genome Research Project established in 1990. Due to sufficient accumulation of anchor sequences, YAC/BAC contigs were assembled into a

useful physical map with substantial continuity. Construction of a physical map accelerated the launch of the whole genome sequence project.

-
-
-

Bibliography

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature* **408**, 796–815. [This paper is a milestone in which completion of *Arabidopsis* genome sequencing is reported. Features of the whole genome are well documented.]

Bowman J. (1993). *Arabidopsis, An Atlas of Morphology and Development*. New York, NY: Springer-Verlag. 450 pp. [The growth and development of *Arabidopsis* is introduced with numerous pictures of the wild type and many mutants. The arrangement of various aspects of continuous morphogenesis in plants make this one of the best reference books even for biologists already familiar with plant anatomy.]

Chang C., Bowman J.L., DeJohn A.W., Lander E.S., and Meyerowitz E.M. (1988). Restriction fragment length polymorphism linkage map *for Arabidopsis thaliana. Proeedings of National Academy of Sciences* USA **85**, 6856–6860. [The first RFLP map of Arabidopsis is reported using mainly randomly selected lambda clones. This was the first important contribution to map-based cloning of mutant genes.]

Hwang I., Kohchi T., Hauge B.M., Goodman H.M., Schmidt R., Cnops G., Dean C., Gibson S., Iba K., Lemieux B., Arondel V., Danhoff L., and Somerville C. (1991). Identification and Map Position of YAC Clones Comprising One-Third of the Arabidopsis Genome. *Plant Journal* **1**, 367–374. [This report describes the first step toward the assembly of an *Arabidopsis* physical map using YAC clones corresponding to RFLP markers.]

Koornneef M., van Eden J., Manhart C.J., Stam P., Braaksma F. J., and Feenstra W. J. (1983). Linkage map of *Arabidopsis thaliana. Journal of Heredity* **74**, 265–272. [76 loci were integrated to five linkage groups corresponding to five chromosomes. Map lengths and centromere positions were determined genetically.]

Ku H.-M., Vision T., Liu J., and Tanksley S.D. (2000). Comparing sequenced segments of the tomato and Arabidopsis genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. *Proceedings of National Academy of Sciences*, USA **97**, 9121–9126. [Gene arrangements on chromosomes were compared between *Arabidopsis* and tomato. A conserved structural relationship known as synteny was seen often among species, making *Arabidopsis* research more applicable to crops.]

Leutwiler L.S., Hough-Evans B.R., and Meyerowitz E.M. (1984). The DNA of *Arabidopsis thaliana. Molecular and General Genetics* **194**, 15–23. [The *Arabidopsis* genome is characterized by using DNA reassociation kinetics. The simple structure of the genome was first recognized by molecular biologists in this report. The blossoming of *Arabidopsis* as a model system began with this discovery.]

Meyerowitz E.M. and Somerville, C.R. (1994.) *Arabidopsis.* Cold Spring Harbor Laboratory Press: Plainview, NY. 1300 pp. [*Arabidopsis* molecular genetics and methodology are well-summarized in this book. Various topics in plant biology are discussed, and major contributions of molecular genetics in *Arabidopsis* studies in the preceding decade are documented.]

Pruitt, R.E. and Meyerowitz, E.M. (1986). *Characterization of the Genome of Arabidopsis thaliana. Journal of Molecular Biology* **187**, 169–183. [The *Arabidopsis* genome is characterized by using quantitative gel blot analysis of DNA. The simple structure of the genome was confirmed by this first molecular study.]

Sato S., Nakamura Y., Kaneko T., Asamizu E., and Tabata S. (1999). Complete structure of the chloroplast genome of *Arabidopsis thaliana. DNA Research* **6**, 283–90. [The chloroplast genome of *Arabidopsis* was completely sequenced. The deduced gene organization was extremely well conserved as reported in various land plants from liverwort to higher plants.]

Unseld M., Marienfeld J.R., Brandt P., and Brennicke A. (1997). The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. *Nature Genetics* **15**, 57–61. [The mitochondrial genome of *Arabidopsis* was completely sequenced, the first of the three genomes to be so. The deduced gene content was similar to that of liverwort mitochondria, in which the whole mitochondrial genome structure was reported although extensive gene rearrangements were found.]

Biographical Sketch

Takayuki Kohchi graduated as D.Agric. from the School of Agriculture, Kyoto University in 1989. He served as research fellow at the Department of Genetics, Harvard Medical School, and the Department of Molecular Biology, Massachusetts General Hospital, till 1992. He was Research Assistant at the Faculty of Agriculture, Kyoto University, till 1994, and at the Nara Institute of Science and Technology till 1995. Since then, he has been Associate Professor at the Nara Institute of Science and Technology.