

MOLECULAR ORGANIZATION OF PLANT CHROMOSOMES

References: The Arabidopsis Genome Initiative (2000)

A lesson from the Arabidopsis genome sequences:

I. Sequencing strategy:

- A. Cloned genomic fragments into various vectors—most of these were large fragments (~100 kb or greater)
- B. Assembled clones into “contigs”—a contig is a set of overlapping clones that span a portion of the genome. Methods used include:
 1. BAC fingerprinting—that is, cut BAC clones into fragments with some enzyme, run each clone on a separate lane in a gel, and then try to match fragment sizes to get them to fit together
 2. Genetic mapping in the centromeric regions to identify contig positions and orientation.
 3. Other procedures—hybridization, PCR, etc. 22 PCR products directly sequenced.
 4. YACs, phage clones, PCR for cloning telomere sequences.
- C. Sequence both strands of clones and assemble individual clone sequences to give the entire genome sequence. Accuracy of sequencing: 99.9-99.999%.

II. General statistics

- A. Genome size: 125 Mb (megabase pairs) (about 10 Mb of this are rDNA and centromeric areas that haven't been completely sequenced).
- B. Number of genes: 25,498 (some are from computer program based prediction).
 1. 11,601 different types or families of genes. (That is, many genes are duplicated, so of the total number of genes, perhaps half that number actually encode different proteins. However, the function of these proteins is almost completely unknown, so simply because two genes have high sequence similarity does not necessarily mean that they will have similar functions.)
- C. Unique genes: *Arabidopsis* has about 150 gene families not found in other eukaryotes. They are: transcription factors, enzymes and proteins- with unknown function.
- D. Organization
 1. Tandem arrays—17% of genes in *Arabidopsis* (Figure 3)—up to 23 adjacent members.
 2. Segmental duplications
 - a. Determined by using:
 - i. Conserved sequences (i.e., aligning sequence data directly) or
 - ii. Conserved gene order (i.e., aligning predicted protein sequences—this is somewhat more robust because DNA sequences can vary quite a bit but still code for essentially the same amino acids)
 - b. Very common—58% - 60% of genome duplicated in 24 segments of >100 kb (Figure 4).
 - c. Have further rearranged after duplication- local inversions.
 - d. Evidence for previous tetraploidization event—i.e., an entire genome duplication that has subsequently been rearranged, with the loss of some of the duplicated fragments (as in maize), could have a tetraploid ancestor.
 - e. The genome has stabilized, though, resulting in regular diploid segregation, divergence of the functions of duplicated genes.

3. **Telomeres**

- a. Consist of tandemly repeated DNA blocks (CCCTAAA).
- b. Are 2-3 kb in length, separated from the coding sequences by repetitive subtelomeric regions (<4 kb).
- c. Imperfect telomere-like sequences up to 24 kb found elsewhere in the genome (e.g. centromeres).

4. Nucleolar organizers (**NORs**)

- a. 3.5-4 Mb of tandem rDNA repeats (18S, 5.8S, and 25S units); ~350-400 units.
- b. Are next to telomeres on chromosomes 2 and 4.

- Note: Transcription of the rRNA genes by RNA polymerase I initiates the formation of a nucleolus, the subnuclear region where ribosomes are assembled.

5. **Centromere**

- a. Contains tandemly repeated DNA, primarily 180 bp repeats and 5S rDNA
- b. At least 47 expressed genes in the genetically defined centromeres.

5. **Gene density**

1. Tends to be generally evenly spaced along the length of the chromosome except in the heterochromatic regions near the centromeres, telomeres, or in the rDNA repeat region.
2. Some regions are relatively more gene dense.
3. 4.5 kb per protein/gene.

6. **ESTs (Expressed sequence tags)** are essentially cDNA clones that have been partially sequenced at each end. About 40% of predicted genes do not have an EST match.

1. Low transcription rates-most likely reason.
2. Mistakes in gene prediction-may be.
3. Sequencing error-unlikely.

7. **Transposable elements (TE)**

- a. Tend to be concentrated near the centromeres, relatively few in telomeres.
- b. 10% of the genome (1/5 of the intergenic DNA).
- c. Type I (replicate through RNA intermediate); retrotransposons 2,109.
- d. Type II (jump as a DNA element): several groups = 1,209.
- e. Large diversity of TE families present.
- f. TE rich regions are gene poor and have low recombination

E. Comparison (92.1 Mb) between sequence of Columbia and Landsberg *erecta* genotypes

1. Two major types of sequence differences:
 - a. Single nucleotide polymorphisms (SNPs)
 - b. Insertions-deletions (Indels)
2. Average density: 1 SNP/3.3 kb and 1 indel/6.1 kb.
3. In coding and non-coding regions
4. Genes were translocated between the two accessions—may be by TE.