Unit 5

Genomes: Organization and Comparisons
Genome

• the sum of all genes and intergenic DNA on all the chromosomes of a cell represents the cellular genome.

• Plants possess a plastid, a mitochondrial, and a nuclear genome while animals have only the latter two.

• Like other eukaryotes, plants have linear chromosomes, each containing hundreds or thousands of genes.

• Nuclear chromosomes consist primarily of non-coding, repetitive DNA. Genes coding for proteins make up only a small percentage of genomes of most higher plant species.
<table>
<thead>
<tr>
<th>Organism</th>
<th>Chromosome number (n)</th>
<th>Genome size (bp)</th>
<th>Gene No.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>5</td>
<td>$1.19 \times 10^8$</td>
<td>25,400</td>
</tr>
<tr>
<td><em>Frageria vesca</em></td>
<td>7</td>
<td>$2.80 \times 10^8$</td>
<td>25,050</td>
</tr>
<tr>
<td><em>Brassica rapa</em></td>
<td>10</td>
<td>$2.84 \times 10^8$</td>
<td>41,174</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>12</td>
<td>$4.66 \times 10^8$</td>
<td>58,000</td>
</tr>
<tr>
<td><em>Glycine max</em></td>
<td>20</td>
<td>$1.10 \times 10^9$</td>
<td>46,430</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>10</td>
<td>$2.80 \times 10^9$</td>
<td>63,000</td>
</tr>
<tr>
<td><em>Triticum aestivum</em></td>
<td>21</td>
<td>$1.70 \times 10^{10}$</td>
<td>41,910</td>
</tr>
</tbody>
</table>
Dependence and independence of cellular genome components

- separate transcription and translation machinery for each organelle

- interdependent in activity
  - Anterograde (Nucleus to Chloroplast/ Mitochondria)
  - Retrograde signalling (Chloroplast/Mitochondria to Nucleus)

- nuclear genome possess overall control

Relative size (average)

<table>
<thead>
<tr>
<th>Genome Type</th>
<th>Relative Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>plastid genome</td>
<td>1</td>
</tr>
<tr>
<td>mitochondrial genome</td>
<td>3</td>
</tr>
<tr>
<td>nuclear genome</td>
<td>30,000</td>
</tr>
</tbody>
</table>

*bp/haploid genome
C-value paradox

1C value refers to the haploid genome size

bacteria < fungi < plants/animals

The paradox: Organism’s complexity does not follow genome size completely.

But Arabidopsis is not 100 times less complex than wheat and 1000 times less than lily

<table>
<thead>
<tr>
<th>Genome size</th>
<th>Gene#</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>~4 x 10^6 bp</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>1.20 x 10^8 bp</td>
</tr>
</tbody>
</table>

Gene # estimates for higher plants 24,000 - 60,000 genes
Genome size comparison among organisms

- Bacteria
  - E. coli

- Fungi
  - Yeast

- Plants
  - Arabidopsis
  - Rice
  - Maize
  - Pea
  - Barley
  - Wheat
  - Lily

- Animals
  - C. elegans
  - Drosophila
  - Human
  - Amphibian

Nucleotide pairs per haploid genome

- $10^5$
- $10^6$
- $10^7$
- $10^8$
- $10^9$
- $10^{10}$
- $10^{11}$

Species mentioned:
- Arabidopsis
- Rice
- Maize
- Pea
- Barley
- Wheat
- Lily
- E. coli
- Yeast
- C. elegans
- Drosophila
- Human
- Amphibian
What portion of the genome is made up by genes?

Assume the average gene size = 3000 bp (promoter, coding region but not introns)

\[
\frac{3000 \text{ bp} \times 25,400 \text{ genes}}{1.20 \times 10^8 \text{ bp}} \times 100 = 61\% \quad \text{Arabidopsis}
\]

\[
\frac{4000 \text{ bp} \times 35,000 \text{ genes} \times 100}{4.5 \times 10^9 \text{ bp}} = 3.1\% \quad \text{Pea}
\]

pea genome size
What accounts for the extra DNA among different species?

1. Repetitive sequences (coding and non-coding)

2. Polyploidy
   - Autopolyploid (all chromosome sets are homologous)
     Eg. Sugar beet autotriploid (AAA)
   - Allopolyploid (subgenome sets of chromosomes arising from 2 or more species or genera)
     Eg. canola  *Brassica napus* (n=19) allotetraploid (AACC)
     *B. oleracea* (n=9) CC
     *B. rapa* (n=10) AA
     (wheat, oat, tobacco)

3. Different size and frequency of introns in genes
Overview of species relationships in the genus 'Brassica'.
Original work by Mike Jones, for Wikipedia.
Major Classes of Genomic DNA

Genic and non-genic DNA

A. Genic DNA (<5% of total genome in large genomes)

1. “Single” copy genes
   - between 1 - 10 copies per haploid genome
   - represents the majority of genes

2. Duplicated and divergent genes
   - functional multi-gene families
   - may occur at a single or several loci on different chromosomes
   - in Arabidopsis (fully sequenced) [996 gene families] and [8,331 genes]

   Specific examples: Calcium dependent protein kinase (1 family - 34 genes)
   Cytochrome P450 (69 families with 256 genes)
   wheat seed storage proteins (gliadens) (40-60)
Major Classes of Genomic DNA

A. Genic DNA (cont’d)

3. **Tandemly repeated genes** (non-divergent)
   a) Histones
      5 histone genes are clustered in a 5-6 kb unit

      Unit is tandemly repeated

      Most multicellular organisms (50-500 tandem repeats/copies)
b) Pre-rRNA genes

Repeat unit includes the 18S, 5.8S and 28S RNA coding regions plus a non-coding spacer region

Repeat unit 8-11 kb depending on species

135 bp minor repeats (10-15 in wheat)
b) Pre-rRNA genes (cont’d)

• Copies / haploid genome
  – Arabidopsis        570 copies
  – Most animals       100-200 copies
  – Hexaploid wheat    1A  1000 copies
                           1B  2600
                           6B  6250

c) 5S-rRNA genes

  100-2000 copies/haploid genome

  2 size variants in wheat  410 bp and 500 bp
                           in flax  340 & 360 bp
Impact of repetitive genic elements

‘Single’ copy genes

\[
\text{30,000 genes} \times 3000 \text{ bp (avg)/gene} = 9 \times 10^7
\]

Repetitive genes

<table>
<thead>
<tr>
<th>Gene Type</th>
<th>Copies</th>
<th>Repeat Size (bp)</th>
<th>Total ((10^6))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histones</td>
<td>500</td>
<td>6000</td>
<td>3 \times 10^6</td>
</tr>
<tr>
<td>Wheat pre rRNA</td>
<td>9850</td>
<td>9000</td>
<td>8.8 \times 10^7</td>
</tr>
<tr>
<td>5S RNA</td>
<td>1000</td>
<td>500</td>
<td>0.5 \times 10^6</td>
</tr>
</tbody>
</table>
Major Classes of Genomic DNA

A. Non-genic sequence (>95% of genome)

1. Tandem repeats
   simple sequence repeats ("satellite" sequences)
   3-10 bp repeats repeated $10^4 - 10^6$ times [TTTAGGG]_n
   (Arabidopsis) GAA (wheat) common at telomeres
   centromere repeats are often multiple elements of short repeats

2. Dispersed repeats
   Many families of repeats spread throughout the chromosomes
   eg 150-300+ bases repeated $10^3 - 10^4$ times throughout genome
   eg Alu element (humans) (~300b)
   Afa family of elements (wheat, barley, oat + related species
   (340 b)
Major Classes of Genomic DNA

Major source of dispersed repeats are transposable elements:

Transposable elements - sequences of DNA that can move around to different positions within the genome of a single cell (mobile genetic elements) (- Nobel Prize 1983  Barbara McClintock 1940s)

2 major classes

- **transposons** - replicate and move as DNA elements
- **retrotransposons** - (originally viral sequences) that replicate via an RNA intermediate method and are copied back into DNA by a reverse transcriptase
Two classes of mobile elements

Retro-transposons  Transposons
Retrotransposons

• Particularly abundant in plants
• Maize  49-78% of genome composed of retrotransposons
• Wheat  (90% repetitive DNA,  68% transposable elements

Common elements (many are also found in plants, fungi and protists)
• LTR retrotransposons (with long terminal repeats 100bp - 5 kbp)
  – Ty1-copia retrotransposons
  – Ty3-gypsy retrotransposons
• Non-LTR retrotransposons  (up to 250,000 copies in plant species)
  – LINEs  long interspersed nuclear elements
  – SINEs  short interspersed nuclear elements

• Human  42% of genome made up of retrotransposons
  – while DNA transposons account for about 2-3%
Maize genome
$2 \times 10^9$ bp

from Schnable PS et al. 2009
Science 326:1112-1115.

Main take home lesson from C,E and F:

Genes are islands in a sea of repetitive sequences.

E- repeats from previously characterized repeats, found by RepeatMasker or other programs.
Non-genic DNA is not necessarily non-functional

Functions of some repeat regions:

Scaffolding (Matrix) attachment regions (~$10^4$ in large genome)

Spindle fiber attachment sites (centromere)

Protection of chromosome termini (telomeres)

Origins of replication (in yeast ~250 bp long)

Estimate of the # of origins in the human genome:

~50,000 (Genome Res. 2013. 23: 1-11)
Plastid Genome

All plant cells contain plastids of some type - same genome (chloroplasts, amyloplast, chromoplast, elaioplast) all from etioplast

Replicate by division. When a cell divides plastid copies are shared

In angiosperms (all common crop plants) during fertilization both plastid and mitochondrial genomes come from the ovary never from pollen. Basis of maternal inheritance of certain traits.

Significant consideration in some genetic engineering as it is a way of preventing transgene gene transfer by pollen flow.
Plastid Genome

Similar in most plant species

- 120-270 kb per plastid genome
- Circular dsDNA

- Encodes 120-140 genes
  - mostly involved in protein synthesis or photosynthesis

- Majority of proteins in the chloroplast are nuclear-encoded

- Typical leaf 10-20 chloroplasts
- 5-10 copies of the plastid genome/chloroplast

Thus 50-200 copies of the genome/leaf cell
Plastid Genome

Genes transcribed on the forward strand are on the outer circle. Those transcribed on the reverse strand are on the inner circle.

Regions found in plastid genomes of higher plants:

LSC - large single-copy region

SSC - small single-copy region

IRa, IRb - inverted repeats; duplicate copies of genes are seen in inverted orientation
Mitochondrial genome

Plant mitochondrial genome range 200-2500 kbp

Fewer genes than the chloroplast genome <100

Many mitochondria/per cell

Individual mitochondria divide and multiple organelles can refuse

Multiple copies of the genome/per mitochondrion

Genome exists as a series of subgenomic* circles of dsDNA molecules which can interchange parts of the circles by recombination (unlike mammalian and fungal mitochondria)

The majority of proteins in the mitochondria are nuclear-encoded

subgenomic* - portions of the full mitochondrial genome sometimes exist as separate circular molecules within the mitochondrion.
Mitochondrial genome

circle 1: Large repeats containing blocks of genes with related functions, R1 - R4.

circles 2 and 3: Transcribed regions on forward and reverse strands.

circle 4: GC content

from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-163