Plant Genome Sequencing

Traditional Sanger Sequencing Genome Sequencing Approach

1. Create sequencing libraries of different insert sizes
   - 2kb
     - Bulk of sequencing is performed on these libraries
   - 10kb
     - Used for linking contigs during assembly
   - 40kb
     - Used to link larger contigs assembly
   - Bacterial artificial chromosomes
     - Used to link ever larger contigs assembly

2. Paired-end sequencing data collected for libraries

3. Contigs created by looking for overlapping reads

4. Contigs assembled based on homology to 10kb, 40kb and BAC sequence data; these large assemblies are called scaffolds

5. Pseudochromosomes assembled based on homology of scaffolds to the markers located on a high-density genetic map
Modern Long Read PacBio Sequencing Genome
Sequencing Approach

1. Create 20kb insert libraries

2. Sequence with PacBio single molecule technology
   • Reads generally 10-15 kb in length

3. Add short read (150bp) paired end data to correct for inherent PacBio errors

4. Assembly reads into contigs
   • Contigs MUCH longer than with Sanger sequencing

5. Scaffolds developed by long-range scaffolding methods
   • BioNano restriction enzyme mapping
   • Hi-C cross-linked DNA library sequencing
   • 10X linked read sequencing

6. Pseudochromosomes assembled based on homology of scaffolds to the markers located on a high-density genetic map
Scaffold Assembly
Building a Scaffold Using Paired-end Reads of Different Sized Sequences

**Step 1**: Build a contig with overlapping 2-kb paired-end reads

2-kb read

**Step 2**: Link two contigs with 10-kb paired-end reads

10-kb read

**Step 3**: Link three 10-kb contigs with 40-kb paired-end reads

40-kb read

**Step 4**: Link two 40-kb contigs with 100-kb BAC end sequences (BES)

BES read

**Step 5**: Here link two 100-kb BAC sized contigs with a 40-kb paired-end read; other sized reads can also be used for this linking

**Step 6**: Continue linking larger blocks of sequences until the block can not be linked with another block. This block is defined as a scaffold.
Genome Assembly
Linking Scaffolds to a Dense Genetic Map

Step 1: Place scaffold relative to sequence complementarity of marker

Step 2: Sequentially place other scaffolds relative to complementarity of markers

Step 3: If no scaffold is complementary to a marker, a gap is inserted relative to the sequence of genetic map. These are represented as “Ns” in the sequence.

Step 4: Repeat steps 1-3 until a chromosome length sequence is developed. The overlapping sequences of each of the linked scaffolds defines a pseudochromosome.

Sequence-based genetic linkage map of a chromosome

Pseudochromosome Sequence
Phaseolus vulgaris
Summary Genome Sequencing and Assembly

Short read production information
• Sequence technology: Sanger, Roche 454, Illumina
• Number of libraries: 21 (15 paired, 6 unpaired)
• Total Reads: 49,214,786 (10,696,722 successful paired-end reads; 2.3% failed)
• Coverage: 21.02x total (18.64X linear, 3.38X paired-end)

Long read production information
• PacBio technology
• 83.2x sequence coverage
• Illumina data from short read project added to PacBio data
Assembly Notes

- **Initial assembly**
  - Arachne
    - Short read
  - Falcon
    - Long read

- **Final assembly**
  - Breaks of initial assembly
    - 71 breaks applied to initial short read assembly
    - 43 breaks applied to initial long read assembly
  - Genetic map and soybean/common bean synteny applied to data
    - Genetic map
      - 7,015 markers used
    - Soybean/common bean synteny used to detect misjoins
      - 248 joins applied to broken assembly short read assembly
      - 577 joins applied to the long read assembly
  - 11 pseudochromosomes assembled
    - 98.8% of short read assembled sequence is found in the pseudochromosomes
    - 95.3% of long read assembled sequence is found in the pseudochromosomes
Estimated genome coverage from Kew Gardens C-value Database

- *P. vulgaris* = 0.6 picograms
  - 1 pg = 978 megabases
    - *P. vulgaris* = \(586.8\) Mb

Coverage

- Short read
  - 521.1 Mb/586.8 Mb = \(88.8\)% coverage
- Long read
  - 537.2/586.8 Mb = \(91.5\)% coverage

<table>
<thead>
<tr>
<th>Summary information</th>
<th>Short read</th>
<th>Long read</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main genome scaffold total</td>
<td>708</td>
<td>478</td>
</tr>
<tr>
<td>Main genome contig total</td>
<td>41,391</td>
<td>1,044</td>
</tr>
<tr>
<td>Main genome scaffold sequence total</td>
<td>521.1 Mb</td>
<td>537.2 Mb</td>
</tr>
<tr>
<td>Main genome contig sequence total</td>
<td>472.5 Mb</td>
<td>531.6 Mb</td>
</tr>
<tr>
<td></td>
<td>(9.3% gap)</td>
<td>(1.1% gap)</td>
</tr>
<tr>
<td>Main genome scaffold N50/L50</td>
<td>5/50.4 Mb</td>
<td>5/49.7 Mb</td>
</tr>
<tr>
<td>Main genome contig N50/L50</td>
<td>3,273/39.5 Mb</td>
<td>73/1.9 Mb</td>
</tr>
<tr>
<td>Number of scaffolds &gt; 50 Kb</td>
<td>28</td>
<td>87</td>
</tr>
<tr>
<td>% main genome in scaffolds &gt;50 Kb</td>
<td>99.3%</td>
<td>99.1%</td>
</tr>
</tbody>
</table>

Loci

27,433 total loci containing 36,995 protein-coding transcripts

Alternative Transcripts

9,562 total alternatively spliced transcripts
N50 and L50: Measures of the Quality of Genomes

Contig
- An aligned group of reads that represent one section of the genome
  - No missing sequence data

Scaffolds
- Groups of contigs that define a section of the genome
  - Larger than contigs
  - Can contain gaps (missing sequence) that are filled in with Ns
  - Number of scaffolds is always smaller than the number of contigs

Pseudochromosome
- Group of scaffolds that represent one chromosome of the species

N50
- The number of contigs (or scaffolds) whose collective distance equals 50% of the genome length
  - This is a NUMBER

L50
- The length of the smallest contig (or scaffolds), of the collection of the contigs (or scaffolds) that comprise the set of N50 contigs (or scaffolds)
  - This is a LENGTH

IMPORTANT NOTE
Today, the L50 length is almost always reported as the N50
<table>
<thead>
<tr>
<th>Species name</th>
<th>Common name</th>
<th>Genotype</th>
<th>Year</th>
<th>Publication</th>
<th>Technical method</th>
<th># Chrom</th>
<th>Est. genome size/assembled size (Mb)</th>
<th>Repeat content (%)</th>
<th>Chrom size range (Mb)</th>
<th># genes/transcripts</th>
<th>Contig N50/L50 (#/kb)</th>
<th>Scaffold N50/L50 (#/kb)</th>
<th>Genomic duplication history</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>Arabidopsis Columbia</td>
<td>2000</td>
<td>Nature 408:796</td>
<td>HSS/S</td>
<td>5</td>
<td>125/135</td>
<td>20³</td>
<td>18-29</td>
<td>27,416/35,386</td>
<td>Eudicot 3x + Brassicaceae (2x+2x)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Populus trichocarpa</td>
<td>Poplar Nisqualy 1</td>
<td>2006</td>
<td>Science 313:1596</td>
<td>WGS/S</td>
<td>19</td>
<td>485/423</td>
<td>40³</td>
<td>11-36</td>
<td>41,335/73,013</td>
<td>??/126</td>
<td>??/3,100</td>
<td>Eudicot 3x + (2x)</td>
<td></td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>Grape PN40024</td>
<td>2007</td>
<td>Nature 449:463</td>
<td>WGS/S</td>
<td>19</td>
<td>475/487</td>
<td>22³</td>
<td>10-22</td>
<td>/ 26,346</td>
<td>??/126</td>
<td>??/2,065</td>
<td>Eudicot 3x</td>
<td></td>
</tr>
<tr>
<td>Carica papaya</td>
<td>Papaya Sunup</td>
<td>2008</td>
<td>Nature 452:991</td>
<td>WGS/S</td>
<td>9</td>
<td>372/370</td>
<td>52</td>
<td></td>
<td>27,332/27,996</td>
<td>??/11</td>
<td>??/1,000</td>
<td>Eudicot 3x</td>
<td></td>
</tr>
<tr>
<td>Zea mays</td>
<td>Maize B73</td>
<td>2009</td>
<td>Science 326:1112</td>
<td>HSS/S</td>
<td>10</td>
<td>7/3,234</td>
<td>84</td>
<td>150-301</td>
<td>39,475/137,208</td>
<td>??/227</td>
<td>??/1,140</td>
<td>Eudicot 3x</td>
<td></td>
</tr>
<tr>
<td>Cucumis sativus</td>
<td>Cucumber 9930</td>
<td>2009</td>
<td>Nat Genet 41:1275</td>
<td>WGS/S,i</td>
<td>7</td>
<td>??/244</td>
<td>22³</td>
<td></td>
<td>21,491/32,528</td>
<td>??/227</td>
<td>??/1,140</td>
<td>Eudicot 3x</td>
<td></td>
</tr>
<tr>
<td>Malus x domestica</td>
<td>Apple Golden Delicious</td>
<td>2010</td>
<td>Nat Genet 42:833</td>
<td>WGS/S</td>
<td>17</td>
<td>742/604</td>
<td>36</td>
<td>21-47</td>
<td>63,538/63,541</td>
<td>16,171/13</td>
<td>102/1,542</td>
<td>Eudicot 3x + Rosaceae 2x</td>
<td></td>
</tr>
<tr>
<td>Jatropha curcas</td>
<td>Jatropha</td>
<td>2010</td>
<td>DNA Res 18:65</td>
<td>WGS/S</td>
<td></td>
<td>380/285</td>
<td>37</td>
<td></td>
<td>40,929/??</td>
<td>??/4</td>
<td>Eudicot 3x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fragaria vesca</td>
<td>Strawberry H4x4</td>
<td>2011</td>
<td>Nat Genet 43:109</td>
<td>WGS/S, 454,</td>
<td>7</td>
<td>240/220</td>
<td>23</td>
<td></td>
<td>32,831/??</td>
<td>??/1,300</td>
<td>Eudicot 3x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis lyrata</td>
<td>Lyrata MN47</td>
<td>2011</td>
<td>Nat Genet 43:476</td>
<td>WGS/S</td>
<td>8</td>
<td>??/207</td>
<td>30</td>
<td>19-33</td>
<td>32,670/??</td>
<td>1,309/5,200</td>
<td>Eudicot 3x + Brassicaceae (2x+2x)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phoenix dactylifera</td>
<td>Date palm Khalas</td>
<td>2011</td>
<td>Nat Biotech 29:521</td>
<td>WGS/I</td>
<td>18</td>
<td>658/381</td>
<td>29</td>
<td></td>
<td>28,890/??</td>
<td>??/6</td>
<td>??/30</td>
<td>Eudicot 3x</td>
<td></td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>Potato DM1-3 S16 R44</td>
<td>2011</td>
<td>Nature 475:189</td>
<td>WGS/S, 454,</td>
<td>12</td>
<td>844/727</td>
<td>62</td>
<td></td>
<td>35,119/51,472</td>
<td>6,446/31</td>
<td>121/1,782</td>
<td>Eudicot 3x + Solanaceae 3x</td>
<td></td>
</tr>
<tr>
<td>Thellungiella parvula</td>
<td>Thellungiella</td>
<td>2011</td>
<td>Nat Genet 43:913</td>
<td>WGS/S, 454,</td>
<td>7</td>
<td>160/137</td>
<td>8</td>
<td></td>
<td>30,419/??</td>
<td>8/5,290</td>
<td>Eudicot 3x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cucumis sativus</td>
<td>Cucumber B10</td>
<td>2011</td>
<td>PLoS ONE 6:e22728</td>
<td>WGS/S, 454</td>
<td>7</td>
<td>??/323</td>
<td>26,587/??</td>
<td>??/23</td>
<td>??/323</td>
<td>Eudicot 3x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brassica rapa</td>
<td>Cappage Chiifu-401-42</td>
<td>2011</td>
<td>Nat Genet 43:1035</td>
<td>WGS/I</td>
<td>10</td>
<td>??/283</td>
<td>40</td>
<td></td>
<td>41,174/??</td>
<td>2,778/27</td>
<td>39/1,971</td>
<td>Brassicaceae 2x + (2x)</td>
<td></td>
</tr>
<tr>
<td>Cajanus cajan</td>
<td>Pigeon pea ICPL 87119</td>
<td>2011</td>
<td>Nat Biotech 30:83</td>
<td>WGS/S,</td>
<td>11</td>
<td>808/606</td>
<td>52</td>
<td>10-48</td>
<td>40,071</td>
<td>7815/23</td>
<td>380/516</td>
<td>Eudicot 3x + Legume 2x</td>
<td></td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>Medicago</td>
<td>2011</td>
<td>Nature 480:520</td>
<td>WGS/S, 454,</td>
<td>8</td>
<td>454/384</td>
<td>35-57</td>
<td></td>
<td>44,135/45,888</td>
<td>53/1270</td>
<td>Eudicot 3x + Legume 2x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes:
- ³ indicates that the genome size is estimated from transcriptome data.
- ?? indicates data is not available.
- (2x) indicates a genome duplication history.
- Eudicot and Poales are plant families.
- Legume refers to the family Fabaceae.
<table>
<thead>
<tr>
<th>Species name</th>
<th>Common name</th>
<th>Genotype</th>
<th>Year</th>
<th>Publication</th>
<th>Technical method</th>
<th># Chrom</th>
<th>Est. genome size/assembled size (Mb)</th>
<th>Repeat content (%)</th>
<th>Chrom size range (Mb)</th>
<th># genes/transcripts</th>
<th>Contig N50/L50 (#/kb)</th>
<th>Scaffold N50/L50 (#/kb)</th>
<th>Genome duplication history</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solanum lycopersicon</td>
<td>Tomato</td>
<td>Heinz 1706</td>
<td>2012</td>
<td>Nature 485:635</td>
<td>WGS/S,So</td>
<td>12</td>
<td>900/760</td>
<td>63</td>
<td>45-65</td>
<td>34,727/??</td>
<td></td>
<td></td>
<td>Eudicot 3x + Solanaceae 3x</td>
</tr>
<tr>
<td>Linum usitatissimum</td>
<td>Flax</td>
<td>CDC Bethune</td>
<td>2012</td>
<td>PI Journal 72:461</td>
<td>WGS/I</td>
<td>15</td>
<td>373/318</td>
<td>24</td>
<td>43,484</td>
<td>4,427/20</td>
<td>132/693</td>
<td></td>
<td>Eudicot 3x + (2x)</td>
</tr>
<tr>
<td>Musa acuminata</td>
<td>Banana</td>
<td>DH-Pahang, ITC1511</td>
<td>2012</td>
<td>Nature 488:213</td>
<td>WGS/S, 454, I</td>
<td>11</td>
<td>??/523</td>
<td>44</td>
<td>22-35</td>
<td>36,542</td>
<td>/43</td>
<td>/1,311</td>
<td>Zingiberales 2x + (2x + 2x)</td>
</tr>
<tr>
<td>Gossypium raimondii</td>
<td>Cotton (B genome diploid)</td>
<td>2012</td>
<td>Nat Genet 44:1098</td>
<td>WGS/I</td>
<td>13</td>
<td>775/567</td>
<td>57</td>
<td>25-69</td>
<td>40,976/??</td>
<td>4,918/45</td>
<td>2,284/95</td>
<td>Eudicot 3x + Gossypium 2x</td>
<td></td>
</tr>
<tr>
<td>Azadirachta indica</td>
<td>Neem</td>
<td>Local tree</td>
<td>2012</td>
<td>BMC Genomics 13:464</td>
<td>WGS/I</td>
<td>??/364</td>
<td>20,169/??</td>
<td>??/0.7</td>
<td>??/452</td>
<td>??/452</td>
<td></td>
<td></td>
<td>Eudicot 3x + Gossypium 2x</td>
</tr>
<tr>
<td>Prunus mume</td>
<td>Chinese plum</td>
<td>2 genotypes</td>
<td>2012</td>
<td>Nature Communications 3:1318</td>
<td>WGS/I</td>
<td>8</td>
<td>??/237</td>
<td>45</td>
<td>31,390/??</td>
<td>2009/32</td>
<td>120/578</td>
<td></td>
<td>Eudicot 3x + Rosaceae 2x</td>
</tr>
<tr>
<td>Pyrus bretschneideri</td>
<td>Pear</td>
<td></td>
<td>2013</td>
<td>Genome Research</td>
<td>HSS+WGS/I</td>
<td>17</td>
<td>528/512</td>
<td>53</td>
<td>11-43</td>
<td>42,812/??</td>
<td>??/36</td>
<td>??/7698</td>
<td>Eudicot 3x + Rosaceae 2x</td>
</tr>
<tr>
<td>Citrullus lanatus</td>
<td>Watermelon</td>
<td>97103</td>
<td>2013</td>
<td>Nat Genet 45:51</td>
<td>WGS/I</td>
<td>11</td>
<td>425/354</td>
<td>45</td>
<td>24-34</td>
<td>24,828/??</td>
<td>??/26</td>
<td>??/2380</td>
<td>Eudicot 3x</td>
</tr>
<tr>
<td>Morus notabilis</td>
<td>Mulberry</td>
<td></td>
<td>2013</td>
<td>Nature Communications 4:2445</td>
<td>WGS/I</td>
<td>7</td>
<td>357/330</td>
<td>47</td>
<td>29,338/??</td>
<td>2,638/34</td>
<td>245/390</td>
<td></td>
<td>Eudicot 3x</td>
</tr>
</tbody>
</table>
Evolution of Plant Genomes

Introduction

Modern plant genomes are quite variable
- ~150 megabase (Mb) *Arabidopsis thaliana* genome.
- 18,000 Mb hexaploidy wheat genome.

Why understanding the evolutionary history of genomes?
- Applied genetics perspective
- Application of comparative genomics for gene discovery.
  - *Arabidopsis* terminal flower 1 (*tfl1*)
    - Encodes a transcription factor
    - It controls indeterminacy/determinacy phenotype
    - *Arabidopsis* *tfl1* as a reference gene
    - Homolog of this gene also controls the phenotype in other
      - Dicot species
        - Snapdragon (*Antirrhinum*)
        - Pea (*Pisum sativum*)
      - Monocot
        - Rice (*Oryza sativum*)
    - Mutations all results in a determinate phenotype
The relevant question

- To what degree are functional genes in one plant species conserved in another species?
  - Important to trace
    - Evolutionary events
    - Related to current organization of plant genomes
Polyploidy and the Construction of Plant Genomes

**Whole genome duplication (WGD)**
- Common event in the evolution of plant species
  - Entire genome doubles in size
  - Duplicates the same genome
- Two related diploid species merge
  - During mitosis
    - Chromatids migrate to separate daughter cells
  - If they move to only one cell
    - The cell will be a tetraploid
- If the 2x duplicate cell is involved in reproduction
  - Resulting gamete
    - 2x the normal number of cells
  - If 2x gamete unites
    - Offspring will be tetraploid

**Polyploidy**
- An organism that contains extra sets of chromosomes.
  - Tetraploids
    - Cultivated potato
    - Alfalfa
- For a success of any polyploidy
  - It must generate balanced gametes.
    - The same number of chromosomes as other gametes
- Embryos from gametes with the same number of gametes
  - Successfully survive
Other Polyploids

- Allopolyploids
  - Two species with very similar chromosomal structure and number intermate.
  - After chromosomal doubling organism, genome will have
    - Number of chromosomes equal to the sum of the number of chromosomes from each of the parent species.
- Examples of allopolyploid species
  - Tetraploid durum wheat (x=14)
  - Hexaploid bread wheat (x=21).
- Durum wheat arose from
  - Union of two diploid species (x=7) species
- Bread wheat arose from
  - Diploid wheat species with the tetraploid wheat species
Constructing the *A. thaliana* genome as a model for eudicot genome evolution

- With the whole genome sequence
  - Study the duplication history of the *A. thaliana* genome.
  - Ancestral duplication signatures could be inferred
    - Blastp analysis
      - Protein vs. protein comparison
      - Identifies gene pairs
        - E-value < -10 used in Fig. 1
    - *Suggests genes are ancestrally related*
  - Duplicates are mapped relative position in the genome
    - Displayed using a dot blot
    - Blocks observed
      - Linear arrayed dots
      - Form a diagonal in the dot blot,
        - Signatures of a duplication event
Figure 1A

- Early comparison of the proteins in the *A. thaliana* genome
  - Red and green diagonals in the upper right panel
    - Block $\alpha_3$
      - Chromosome 1 vs. chromosome 1 block
      - Signature of a duplicated block of genes
      - Genes that have the same conserved order
      - At two ends of the *A. thaliana* chromosome 1
    - Block $\alpha_5$
      - Another pairs of duplicated genes on chromosome 1
    - Block $\alpha_8$
      - Shared block on chromosomes 1 and 3
  - Block, $\alpha_{11}$
  - Largest block
  - Ends of chromosomes 3 and 2
- Total
  - 27 major duplicated blocks
    - Strong signals
    - *Signals of a recent duplication*

*So how does this relate to the mechanism of genome construction?*

- *A. thaliana* underwent a WGD
  - Chromosomes were broken
  - Rearranged into new chromosomes
  - New chromosomes developed
    - *Represent blocks of DNA from the progenitor species*
Figure 1. Dot blot display revealing duplication events. (from Bowers et al. 2003. Nature 422:433)

Figure 1: Arrangement of duplicated protein-encoding *Arabidopsis thaliana* in the composition of the 26 large duplications (at left and bottom). Twenty-three right duplications. Randomly chosen from 26,028 genes in their chromosome duplication (see text) are highlighted. Colours show how the duplication order. The best-matching gene pairs are plotted, colour-coded to indicate chromosome segments contributing duplications, distinguishing contributions to opposite (green) transcriptional orientations. For further analysis, 57 adjacent duplications at left and bottom respectively from the (1) lower-numbered chromosomes regions with opposite orientation and order replicable by localized inversion (2) higher- and lower-numbered chromosomes (light blue); (3) lower- and combined into 26 'large' duplications (4 260 that each included 1% (260) of the higher-numbered chromosomes (dark blue); (4) higher-numbered chromosomes (green). Eight shorter duplications were pooled lower left and duplication. Higher-resolution versions of the figure and lists of gene orders are available (see both and axes represent 21,749 genes, in an inferred ancestral order that accounts for the evolutionary Information).
Progenitor *Arabidopsis* genome

- How it was modified by the duplication event
- Compare to species that is evolutionary close.
  - *A. lyrata*
    - 8 chromosomes
  - *A. thaliana*
    - 5 chromosomes
- Genetic maps developed using shared loci were

Fig. 2

- Five *A. thaliana* chromosomes
  - Constructed from ancestral genome with eight chromosomes
- At Chr I
  - Blocks of AlyLG1 + AlyLG2
- At Chr II
  - Blocks of AlyLG3 + AlyLG4.
- Conclusion
  - *Two species with different chromosome numbers consist of the same chromosomal blocks*
Figure 2. Comparative physical map of *A. thaliana* and the genetic map of *A. lyrata*. (from: Yogeeswaran et al. Genome Research 15:505)

Figure 2. Colinearity of *A. lyrata* linkage map with the *A. thaliana* genome. *A. thaliana* chromosomes (At Chr I–V) are represented as patterned bars (drawn to scale, 1 unit = 1 Mb; gray rectangles, centromeres; gray circles, heterochromatic knobs). *A. lyrata* linkage groups (Aly LG 1–8) are shown in black (drawn to scale, 1 unit = 5cM). Sixteen colinear blocks are highlighted with the same pattern as the At chromosome to which they correspond. Markers defining the ends of each colinear block are shown on the map in black lettering. Markers mapping with LOD score less than 3.0 are featured in parentheses. Italicized markers map to translocated or nonsyntenic regions in *A. lyrata*. Translocations T1 and T2 are highlighted by arrows whose patterns correspond to the At chromosome where their colinear region lies. Major inversions I1 and I2 and minor inversion I3 are highlighted in light gray. Three chromosomal fusions are denoted as F1–F3.
Fig. 1B – Early duplication events
- Shows evidence of more ancient duplications
  - 27 α duplications reoriented
    - Notice block α5
    - Two duplicates blocks in the same order
    - Two in an opposite orientation
      - Presumed ancestral order derived from these four blocks
    - Same procedure that uncovered the α blocks.
- Two types of blocks discovered.
  - 22 β blocks
    - Another duplication event in the A. thaliana lineage

The 7 γ blocks
- Controversial
  - Hypothesis 1
    - Early duplication in the angiosperm lineage
  - Hypothesis 2
    - Duplication after the split of monocots and dicots
- Grapevine genome sequenced
  - Evidence from the genome appears to have resolved this question
    - Grape
      - Ancestor of the rosids
        - Group of species included A. thaliana.
    - Blast and dot blot analysis of grape genome
Figure 3
- Any genes shared with two other regions of the genome
  - Grape genome has a hexaploid history
- How about other species
  - Signal of hexaploidy is detected
  - **Figure 4**
    - Grape and poplar genomes were compared
    - Only triplicated regions in grape used
      - Triplicated regions
        - Two copies in poplar
      - *Hexaploid ancestry concept is supported*
      - *Poplar under went an additional WGD after its divergence from the grape lineage*

**Shared duplications in dicot and monocot analysed**
- Grape and rice orthologs analyzed
  - Hypothesis 1
    - Rice shared the hexaploid ancestry
      - 3-to-3 relationship
        - Not observed
  - Hypothesis
    - Rice does not share the same hexaploid ancestry
      - 3-to-1 relationship observed
  - Conclusion
    - *Monocots and dicots do not share the same hexaploid history.*

*(Note: See Tang et al. 2008. Genome Research18:1944 for an alternative perspective.)*
**Figure 3.** Dot blot representation of duplicate regions of the grapevine genome. (from: Jaillon et al. 2007. Nature 449:463)

**Figure S5.** The grape genome originated from a polyploidy event that joined three ancestral genomes. The nineteen chromosomes of grape are represented on both the x and y axis. Dots represent the positions of paralogous pairs of genes. For clarity, intrachromosomal paralogs are not shown. Clusters of paralogs form a succession of dots, that indicate that the gene order of the ancestral genome was locally maintained. These clusters are painted in seven colours. Each colour marks paralogous blocks, that were colinear in the ancestors of the three constituents of the grape genome. Some regions are not painted in triplicate in this grid, either because a whole region is not visible in synteny with two others in the present-day grape genome (too many rearrangements or gene loss), or because one or two syntenic regions lie in supercontigs which are still not anchored.
**Figure 4.** Comparison of the triplicated blocks and the Poplar genome. (from: Jaillon et al. 2007. Nature 449:463)

**Figure S6.** The distribution of 8,604 orthologous genes between *Vitis vinifera* (x axis) and *Populus trichocarpa* (y axis) chromosomes.
Summary of Eudicot Evolution

- Two diploid mate
  - Tetraploid species developed
- Tetraploid species mated to another diploid
  - Produce the ancestral hexaploid
    - All subsequent eudicots derived from this ancestor
  - Signatures of the same duplications
    - Should be observed in their genome history

Monocot genome evolution.

- Monocots also have a duplication history.
  - Figure 5
    - Compared rice and maize.
      - Maize chromosomes (y-axis) as the reference
        - Most rice genes found in two copies
      - Rice chromosomes (x-axis) as the reference
        - Blocks found three or four times in maize.
    - Conclusion
      - WGD event in the history of monocots
      - An additional duplication occurred in the maize lineage.
Figure 5. A comparison of maize and rice duplication events. (from: Wei et al. (2007) PLoS Genetics 3(7):e123, 1254)

Figure 1. Dotplot Analysis of the Integrated Maize Map against Rice Pseudomolecules
Synteny blocks were detected, and background noise was filtered with SyMAP [37]. The interactive dotplot can be viewed at http://www.agcol.arizona.edu/sympat. When clicking the related synteny block, the detailed window with contig number will pop up. The viewer can select the preferred area and double click the selection, and then a graphic alignment is displayed.
doi:10.1371/journal.pgen.0030123.g001
Unified model of grass evolution – developing the ancestor

- Based on sequences of genome sequences of
  - Rice
  - Sorghum
  - Brachypodium (a model grass species)
  - Maize
- 56-73 MYA
  - Ancestral grass species containing five chromosomes
    - Duplicated
    - Genome with ten chromosomes appeared
  - Then
    - A4 and A6 fractionated
      - Chromosomes A4, A6, and A2 appear
    - A7 and A10 fractionated
      - Chromosomes A7, A10, and A3 appear
  - Paleopolyploid developed
    - 12 chromosomes
      - *Progenitor of all of the modern grasses*

Unified model of grass evolution – developing the lineages

- Rice genome structure
  - Represents the ancient paleotetraploid.
    - Basic set of chromosomes
      - Building blocks for other genomes
Figure 6
- Breakage/translocation/fusion events
  - Involve chromosomal fragments from the n=12 ancestor.
    - Developed
      - Brachypodium
      - Poideae (representing the wheat lineage)
      - Panicoideae (representing the maize/sorghum lineage)
    - Panicoideae
      - Simplest history
      - Arose from only four breaks
  - Other lineages
    - More complex patterns of evolution
      - Maize genome
        - Underwent additional duplication
        - Additional breakage/translocation/fusion events
  - * Constructed the modern maize chromosomes*
Figure 6. A unified model of grass genome evolution. (from: Vogel et al. 2010. Nature 463:763.)

Supplementary Figure 18. Grass chromosome evolution model. The monocot chromosomes (r1-r12 for rice, t1-t7 for Triticeae, bd1-bd5 for Brachypodium, s1-s10 for sorghum, and m1-m10 for maize) are represented with a five colour code to illustrate the evolution of segments from a common ancestor with five proto-chromosomes and a n=12 intermediate as described in 62, and are named according to the rice nomenclature. The events that have shaped the structure of the 5 different grass genomes including the 7 Brachypodium chromosome nested insertion events during their evolution from the common ancestor are indicated as whole genome duplication, ancestral chromosome translocations and fusions, and lineage- specific nested chromosome insertions.
Summary

- Plant genomes
  - A long history of genome duplications
    - Unlike animal and fungal genoemes,
  - Figure 7
    - Illustrates the duplication history
      - (The γ event should be moved to the origin of the eudicot lineage.)
    - Significant role of WGD in development of plant species
      - Many duplications appear 55-70 MYA
        - Transition point
          - Cretaceous and Tertiary periods
            - Mass extinction of species
        - Hypothesis
          - Duplications gave plants the needed gene repertoire
            - To survive this extinction
            - Flourish on earth

(see Fawcett et al. 2009. PNAS USA 106:5737)

- Figure 8
  - Additional species were analyzed
  - Extended the analysis to deeper phylogeny
  - Additional duplication events determined
    - Ancestral seed plants
      - ζ at ~330 MYA
    - Ancestral angiosperms
      - ε at ~220 MYA
**Figure 7.** A summary of the duplication history of plants. (from Van de Peer et al. 2009. Trends in Plant Sciences 14:680)
Figure 8: Ancestral polyploidy events in seed plants and angiosperms. [Jiao et al (2011) Nature 473:97]

Original figure legend from manuscript. Two ancestral duplications identified by integration of phylogenomic evidence and molecular time clock for land plant evolution. Ovals indicate the generally accepted genome duplications identified in sequenced genomes (see text). The diamond refers to the triplication event probably shared by all core eudicots. Horizontal bars denote confidence regions for ancestral seed plant WGD and ancestral angiosperm WGD, and are drawn to reflect upper and lower bounds of mean estimates from Fig. 2 (more orthogroups) and Supplementary Fig. 5 (more taxa). The photographs provide examples of the reproductive diversity of eudicots (top row, left to right: Arabidopsis thaliana, Aquilegia chrysantha, Cirsium pumilum, Eschscholzia californica), monocots (second row, left to right: Trillium erectum, Bromus kalmii, Arisaema triphyllum, Cypripedium acaule), basal angiosperms (third row, left to right: Amborella trichopoda, Liriodendron tulipifera, Nuphar advena, Aristolochia fimbriata), gymnosperms (fourth row, first and second from left: Zamia vazquezii, Pseudotsuga menziesii) and the outgroups Selaginella moellendorfii (vegetative; fourth row, third from left) and Physcomitrella patens (fourth row, right). See Supplementary Table 4 for photo credits.
Dicot Paleohistory

- A duplication history is common in both dicots and monocots
  - Revealed by comparisons within different species

- Rice used as a reference for monocots
- Grape used as a reference for dicots
The ancestral dicot contained seven chromosomes

- The ancestor underwent a duplication to produce a paleohexaploid ancestor of modern dicots

Subsequent events

- Whole genome duplications within some lineages
- Breakage, fusion to generate new chromosomes in all lineages
a, Paralogous gene pairs in *Eucalyptus* for the identified palaeohexaploidization (bottom) and palaeotetraploidization (top) events. Each line represents a duplicated gene, and colours reflect origin from the seven ancestral chromosomes (A1, A4, A7, A10, A13, A16, A19). b, Number of synonymous substitutions per synonymous site ($K_s$) distributions of *Eucalyptus* paralogues (top) and *Eucalyptus*–*Vitis* orthologues (bottom). Blue bars (top) indicate $K_s$ values for 378 gene pairs from the palaeotetraploidization WGD event (red dot), and red bars show $K_s$ values for 274 gene pairs of the palaeohexaploidization event (red star). c, Evolutionary scenario of genome rearrangements from the Eudicot ancestor to *Eucalyptus* and other sequenced plant genomes; palaeohistory modified from ref. 49.
The Gene-based Evolution of Duplicated Genes

If duplications are a major signature of plant genomes
  • Copy number of genes should equal the number of rounds of duplication.

Table 1
  • Number of genes found within plant species
    o Complete genome sequence
      ▪ If the hexoploidy concept is true for dicots, and
      ▪ Grape only contains this hexaploid event
    • Estimate
      o Ancestral dicot contains ~10,000 genes
        (=30,000/30).
Table 1. The estimated number of genes in sequenced plant genomes.

<table>
<thead>
<tr>
<th>Species</th>
<th>Estimated # of Genes (from <a href="http://www.phytozome.net">www.phytozome.net</a>)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Eudicots</strong></td>
<td></td>
</tr>
<tr>
<td>Cucumber</td>
<td>21,491</td>
</tr>
<tr>
<td>Cassava</td>
<td>47,164</td>
</tr>
<tr>
<td>Poplar</td>
<td>41,000</td>
</tr>
<tr>
<td>Medicago</td>
<td>50,692</td>
</tr>
<tr>
<td>Soybean</td>
<td>66,153</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>27,343</td>
</tr>
<tr>
<td>Papaya</td>
<td>27,332</td>
</tr>
<tr>
<td>Grape</td>
<td>30,434</td>
</tr>
<tr>
<td>Mimulus</td>
<td>25,530</td>
</tr>
<tr>
<td><strong>Monocots</strong></td>
<td></td>
</tr>
<tr>
<td>Sorghum</td>
<td>34,496</td>
</tr>
<tr>
<td>Maize</td>
<td>32,540</td>
</tr>
<tr>
<td>Brachypodium</td>
<td>25,532</td>
</tr>
<tr>
<td>Rice</td>
<td>31,500</td>
</tr>
</tbody>
</table>
Similarly

- Poplar underwent an additional duplication,
  - Theoretically # of genes = 60,000 genes
- *A. thaliana* underwent two duplications
  - Theoretically # of genes = 120,000 genes
- *Not observed*

Monocot calculations

- Rice, Brachypodium, and sorghum only contain a duplication event
  - Number of ancestral monocot genes
    - 15,000 (=30,000/2).
  - Maize
    - Additional duplication event
      - But has undergone a reduction to ~30,000 genes
- Conclusion
  - Necessary to reduce the number of genes to ensure the success of the species.
Diploidization.

The polyploid past history of plants
- Surprising result for Arabidopsis and rice genomes
  - Why??
    - Selected for sequencing because of their small genome sizes

Consequences of polyploidy?
- Doubling or tripling of the number of chromosomes
  - Evident for monocots.

Fate of the additional gene set from the WGD
- Concept
  - Species cannot maintain the entire set of duplicate chromosomes
  - New genes a problem
    - Generate deleterious mutations
    - Compromises the fitness of a genome
  - Genome must transition back to its original state.
    - Process is called
      - Diploidization.
To revert back to the diploid state

- Many duplicate genes must be eliminated from the gene set
  - But a recently duplicated genome
    - Soybean
      - Withstands the extra copies
      - Genome about 2X the basic set of 30,000 genes of hexoploid ancestral eudicot

Events associated with diploidization

- Duplicate genome must change its chromosome pairing pattern
  - After the duplications,
    - Four chromosomes pair
    - Form quadrivalents
  - Chromosomal structure must be changed so
    - Bivalents must be formed
  - Result
    - Doubling of the chromosome number
      - Seen for the monocot lineage
  - Once bivalents are formed
    - Gene sets can evolve
      - Processes
        - Deletions and chromosomal rearrangements
Duplicate genes can undergo specific changes

- Common fate
  - Gene death of new copies
    - Loses associated with
      - Chromosomal breakage
      - Rearrangements.
  - Result
    - New basic set of chromosomes and genes will have appeared

Duplicate genes fate differs

- Some are retained as multicopy
  - Up to the ploidy level for that species
- Other reduced to only a single copy

“Deletion resistant” genes

- Not reduced to single copy
  - Dosage dependent
    - Mainly encode
      - Transcription factors
    - May lead to
      - Complex morphologies

“Duplication resistant” genes

- Must be maintained as single copy
  - Mainly encode
    - Enzymes or genes of unknown function
Table 1. The estimated number of genes in sequenced plant genomes.

<table>
<thead>
<tr>
<th>Species</th>
<th>Estimated # of Genes (from <a href="http://www.phytozome.net">www.phytozome.net</a>)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Eudicots</strong></td>
<td></td>
</tr>
<tr>
<td>Cucumber</td>
<td>21,491</td>
</tr>
<tr>
<td>Cassava</td>
<td>47,164</td>
</tr>
<tr>
<td>Poplar</td>
<td>41,000</td>
</tr>
<tr>
<td>Medicago</td>
<td>50,692</td>
</tr>
<tr>
<td>Soybean</td>
<td>66,153</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>27,343</td>
</tr>
<tr>
<td>Papaya</td>
<td>27,332</td>
</tr>
<tr>
<td>Grape</td>
<td>30,434</td>
</tr>
<tr>
<td>Mimulus</td>
<td>25,530</td>
</tr>
<tr>
<td><strong>Monocots</strong></td>
<td></td>
</tr>
<tr>
<td>Sorghum</td>
<td>34,496</td>
</tr>
<tr>
<td>Maize</td>
<td>32,540</td>
</tr>
<tr>
<td>Brachypodium</td>
<td>25,532</td>
</tr>
<tr>
<td>Rice</td>
<td>31,500</td>
</tr>
</tbody>
</table>
Developing new functions

**Duplicate set of genes cannot be maintained**

- Deleterious mutations can arise
- Duplicate genes are modified
  - Changes will provide
    - New functions
    - Altered altered functions
  - New functions may lead to the evolution of the species
    - Higher level of fitness
    - Evolutionary modifications of duplicate genes

**Neofunctionalization.**

- One duplicate gene maintains its original function
- Second gene evolves a function
  - May increase the adaptability of an individual

**Subfunctionalization**

- Modifies the duplicates
- Basic structure of both copies altered
  - Expression pattern of the gene changes
    - Results in a higher level of the protein production
- Alternately, the function of the original gene is maintained
  - Structure of both copies is significantly changed.
    - New copies retains
      - Part of the original function
    - Two genes work together
      - Function of the original gene maintained
Synteny: The Result of WGD and Reconstructing Plant Genomes

Synteny among plant species.

- Major result of the duplication history
  - Synteny
    - Maintenance of gene order between two species
  - Classic approach to synteny
    - Based on shared markers mapped onto two different species.
      - Macrosynteny is detected by
        - Large scale chromosomal blocks shared by two species.

**Fig. 9**

- Example of macrosynteny
  - Tomato and eggplant
    - Eggplant linkage group 4
      - Evolutionarily related to tomato
    - Linkage groups 10S and 4L.
      - Highly conserved marker order over many centimorgans of the two genomes
Figure 9. Macrosynteny between tomato and eggplant, including a QTL for a shared domestication trait. (from: Doganlar et al. 2002. Genetics 161:1713.)
Genetic mapping of shared genes

- First method of comparing species
- Only way to compare species that have not been sequenced
- Many examples of synteny mapping in plants.
- The power of synteny mapping
  - Discovery of shared loci from two species
    - Control the same phenotype
    - Map to the same genetic location.

Fig. 9 again

- Major QTL for fruit striping
  - Eggplant linkage 4.
  - Previous work with tomato
    - Major QTL
      - Linkage group 10 of tomato
  - Syntenic marker and QTL observed here
- Hypothesis
  - Multiple loci are shared in the same macrosyntenic order
    - Same ancestral gene is controlling this trait in these two species.
Leveraging knowledge in one species for gene discovery in a second species

- Phenotypic traits mapped extensively in one species
  - Points a researcher working on a second species
  - Likely location of a similar gene in second species.
  - Leverage is
    - Great aid for genetic discovery
    - For species in where the discovery of important genetic factors are limited by a lack of funding