Genome Sequencing

Principals of a genome project for a species

- Identify the reference line
  - Typically the line has been used historically in genetic research
  - Or, a resource is available for the line
    - Clone library
    - Mutant lines
    - Mapping parent
- Isolate high quality DNA
- Apply some sequencing approach that fits the budget
- Collect DNA
- Assembly the reads into
  - Contigs
  - Scaffolds
- Order scaffolds using a genetic map into pseudochromosomes
- Annotate the genes
- Use the reference genome in research
  - Discover candidate gene(s) that control a phenotype
  - Develop markers to tract important genes/regions of the genome
Approach to the Actual Genome Sequencing
- Fragment the genomic DNA
- Clone those fragments into a cloning vector
- Isolate many clones
- Sequence each clone

Sequencing Techniques Were Well Established
- Sanger was used for the past twenty years
- Helped characterize many different individual genes.
- Previously, the most aggressive efforts
  - Sequenced 40,000 bases around a gene of interest

How is Genomic Sequencing Different???
- The scale of the effort
  - Example
    - Public draft of human genome
      - Hierarchical sequencing
      - Based on 23 billion bases of data
    - Private project (Celera Genomics) draft of human genome
      - Whole genome shotgun sequencing approach
      - Based on 27.2 billion clones
      - 14.8 billion bases

Result:
- Human Genome = 2.91 billion bases
Changes That Facilitated Genomic Sequencing

- Sequencing
  - Basic technique is still the same
- Major changes
  - Thermostable polymerase enzymes
    - Improves quality of sequencing products
  - Fluorescently labeled nucleotides for the reaction
    - Allows for laser detection
  - Laser-based detection systems
    - 8, 16 or 96 samples analyzed simultaneously
  - Results for a single run
    - 500-700 bases of high quality DNA sequence data
  - Human Project peak output
    - 7 million samples per month
    - 1000 bases per second
- Robotics
  - Key addition to genomic sequencing
  - Human hand rarely touches the clone that is being sequenced
  - Robots
    - Pick subclones
    - Distribute clones into reaction plates
    - Create the sequencing reaction
    - Load the plates onto the capillary detection system
  - Result
    - Increased the quality and quantity of the data
    - Decreasing the cost
      - Dropped over 100 fold since 1990
    - Improvements felt in small research lab
  - Sequence reads today
    - $2.50 vs. $15 in the early 1990s.
Hierarchical Shotgun Sequencing of Genomes

A. The Concept

Hierarchical shotgun sequencing requires that large insert libraries be constructed. A series of these clones are ordered by several techniques. Once these clones are ordered, each clone is separately fractionated into small fragments and cloned into plasmid vectors. The plasmid clones are sequenced, and the sequence is assembled. This is the procedure used to sequence the *Arabidopsis* genome, and by the public project to sequence the human genome.

*Large insert libraries* of nuclear DNA are created in BAC, PAC or YAC vectors.

Large inserts clones are placed in order using hybridization, fingerprinting, and end sequencing. This *tiling path* consists of individual large insert clones that will be sequenced.

Each ordered clone is fractionated into small fragments and cloned into a plasmid vector. This is called *shotgun cloning*.

Each clone is then *end-sequenced*, and the sequences of all the clones are aligned.

```plaintext
ATTCGTTAGCGATTA
AGCGATTATTAGATAC
```
Hierarchical Shotgun Sequencing

- Two major sequencing approaches
  - Hierarchical shotgun sequencing
  - Whole genome shotgun sequencing
- Hierarchical shotgun sequencing
  - Historically
    - First approach
  - Why???
    - Techniques for high-throughput sequencing not developed
    - Sophisticated sequence assembly software not availability
- Concept of the approach
  - Necessary to carefully develop physical map of overlapping clones
    - Clone-based contig (contiguous sequence)
  - Assembly of final genomic sequence easier
  - Contig provides fixed sequence reference point
- But
  - Advent of sophisticated software permitted
    - Assembly of a large collection of unordered small, random sequence reads might be possible
  - Lead to Whole Genome Shotgun approach
Steps Of Hierarchical Shotgun Sequencing

- Requires large insert library
  - Clone types
    - YAC (yeast artificial chromosomes)
      - Megabases of DNA
      - Few (several thousand) overlapping clones necessary for contig assembly
    - But
      - YACs are difficult to manipulate
      - Most research skilled with bacteria but not yeast culture
  - Rarely, if ever, used today
    - BAC or P1 (bacterial artificial chromosomes)
      - Primary advantages
        - Contained reasonable amounts of DNA
          - about 75-150 kb (100,000 – 200,000) bases
        - Do not undergo rearrangements (like YACs)
        - Could be handled using standard bacterial procedures
Developing The Ordered Array of Clones

- Using a *Molecular Map*
  - DNA markers
  - Aligned in the correct order along a chromosome
  - Genetic terminology
    - Each chromosome is defined as a *linkage group*
  - Map:
    - Is reference point to begin ordering the clones
    - Provides first look at sequence organization of the genome

**Tomato High Density Marker Collection**

<table>
<thead>
<tr>
<th>Chromosome</th>
<th># Markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>363</td>
</tr>
<tr>
<td>2</td>
<td>310</td>
</tr>
<tr>
<td>3</td>
<td>242</td>
</tr>
<tr>
<td>4</td>
<td>238</td>
</tr>
<tr>
<td>5</td>
<td>158</td>
</tr>
<tr>
<td>6</td>
<td>202</td>
</tr>
<tr>
<td>7</td>
<td>191</td>
</tr>
<tr>
<td>8</td>
<td>173</td>
</tr>
<tr>
<td>9</td>
<td>184</td>
</tr>
<tr>
<td>10</td>
<td>160</td>
</tr>
<tr>
<td>11</td>
<td>149</td>
</tr>
<tr>
<td>12</td>
<td>136</td>
</tr>
</tbody>
</table>

**Marker Type**

- CAPS (1088), RFLP (1342), SNP (19), SSR (155)

**Mapping population**

- 88 F2 individuals

**Today**

- Exclusively SNP markers (n>6,000 SNPs)
- Populations larger (n>200 individuals)
Tomato EXPEN 2000
Chromosome 1
Linkage Map
Developing a Minimal Tiling Path

- Definition
  - Fewest clones necessary to obtain complete sequence
- Caution is needed
  - Clones must be authentic
  - Cannot contain chimeric fragments
    - Fragments ligated together from different (non-contiguous) regions of the genome
  - How to avoid chimeras and select the minimum path
    - Careful fingerprinting

- Overlapping the clones
  - Maps not dense enough to provide overlap
  - Fingerprinting clones
    - Cut each with a restriction enzyme (*HindIII*)
    - Pattern is generally unique for each clone
    - Overlapping clones defined by
      - Partially share fingerprint fragments
  - Overlapping define the *physical map* of the genome

**BAC clone fingerprinting**

- Restriction enzyme digestion
- Digital, imaging, scoring and aligning
Gel Photograph of digested BAC clones

(https://www.researchgate.net/profile/T_Mirkov/publication/11896126/figure/fig3/AS:394567875612700@1471083720903/Fig-3-HindIII-fingerprinting-gel-of-the-major-BAC-clones-in-the-12-Mb-contig-that.ppm)
Digital Image of Clones
(https://www.researchgate.net/figure/6535067_Fig-5-Example-of-the-clone-order-fingerprints-of-a-BAC-contig-of-the-apple-physical)
Genomic Physical Maps
- Human
  - 29,298 large insert clones sequenced
    - More than necessary
    - Why???
      - Genomic sequencing began before physical map developed
      - Physical map was suboptimal
- Arabidopsis
  - 1,569 large insert clones defined ten contigs
    - Map completed before the onset of sequencing
    - Smaller genome
      - about 125 megabases
- Yeast
  - 493 cosmid (smaller insert clones) clones
    - Relatively high number of clones for genome size

Other Uses Of A Physical Maps
- Rich source of new markers
- Powerful tool to study genetic diversity among species
- Prior to whole genome sequencing
  - Markers can locate a target gene to a specific clone
  - Gene can be sequenced and studied in depth
Sequencing Clones Of The Minimal Tiling Path

- Steps
  - Physically fractionate clone in small pieces
  - Add restriction-site adaptors and clone DNA
    - Allows insertion into cloning vectors
      - Plasmids current choice
  - Sequence data can be collected from both ends of insert DNA
    - Read pairs or mate pairs
      - Sequence data from both ends of insert DNA
      - Simplifies assembly
      - Sequences are known to reside near each other
Assembly of Hierarchical Shotgun Sequence Data

- Process
  - Data collected
  - Analyzed using computer algorithms
  - Overlaps in data looked for

- Accuracy levels
  - Analyzing full shotgun sequence data for a BAC clone
    - Goal: 99.9% accuracy
    - 100 kb BAC clone
      - 2000 sequence reads
      - Equals 8-10x coverage of clone
      - Typical level of accuracy that is sought
    - Primary software used is Phrap
      - Phrap = f(ph)ragment assembly program
      - Efficient for a “small” number of clones
        - Small relative to number from a whole genome shotgun approach
      - Each sequence read is assessed for quality by the companion software Phred
      - Assembles sequence contigs only from high quality reads
  - Working draft sequence
    - 93-95% accuracy
    - 3-5 x coverage of 100 kb BAC clone
Viewing the Quality Score Data
Here the Phred scores are overlaid on the chromatogram of a Sanger sequencing output.

- This is just one format the data can be visualized.
- The visualization comes from a quality score data file generated by base-call machine.

**Finishing The Sequence**

- Gaps need to be filled
  - Study more clones
    - Additional subclones sequenced
    - Tedious and expensive
  - Directed sequencing of clones or genomic DNA used
    - Create primers near gaps
      - Amplify BAC clone DNA
        - Sequence and analyze
      - Amplify genomic DNA
        - Sequence and analyze

**Confirming the Sequence**

- Molecular map data
  - Molecular markers should be in proper location
- Fingerprint data
  - Fragment sizes should readily recognized in sequence data
Tomato Genome Project

- Distribution of BAC sequencing effort across the genome
- BAC selection for the tomato genome project

New BACs mapped on tomato chromosomes using CAPS markers

IL-bin Mapping completed for 50 BACs
Visualizing the progress in the Tomato Genome Project

![Graph showing progress in Tomato Genome Project]

Headlines of Human Genome Sequencing Project

- February 2001
  - Working draft announced
  - Major worldwide news event
- April 2003
  - Finished draft announced
  - Little fanfare
  - Data more useful
Whole Genome Shotgun Sequencing

A. The Concept

Shotgun sequencing requires that random, small insert libraries are created from the total nuclear DNA of the species of interest. A plasmid cloning vector is used for this step. These clones are then sequenced. This step is analogous to the shotgun cloning and sequencing step used for each large-insert clone used in hierarchical shotgun. The sequences of the clones are then aligned. This is the procedure used to sequence the *Drosophila* genome, and by Celera to sequence the human genome.

Nuclear DNA is fractionated into several size classes. The most abundant class used for clone creation is 2kb in size. Large inserts (10-50 kb) are also used.

The fragments are then cloned into a plasmid vector.

Each clone is then end-sequenced from both ends to develop read pairs or mate pairs. The sequences of all the clones are aligned.

ATT CGTTAGC GATTA AGCGATTATTAGATAC
**Whole Genome Shotgun Sequencing (WGS)**

- Hierarchical sequencing approach
  - Begins with the physical map
  - Overlapping clones are shotgun cloned and sequenced
- **WGS**
  - Bypasses the mapping step
- Basic approach
  - Take nuclear DNA
  - Shear the DNA
  - Modify DNA by adding restriction site adaptors
  - Clone into plasmids
    - Plasmids are then directly sequenced
    - Approach requires read-pairs
      - Especially true because of the repetitive nature of complex genomes

**WGS**

- Proven very successful for smaller genomes
  - Essentially the only approach used to sequence smaller genomes like bacteria
- Is WGS useful for large, complex genomes?
  - Initially consider a bold suggestion
  - Large public effort dedicated to hierarchical approach
  - *Drosophila*
    - Sequenced using the WGS approach
Rice
- Early publications, not definitive
- Definitive reference genome developed from hierarchialy shotgun sequencing
- Two different rice genomes sequenced using WGS approach
  - Only developed a working draft though
  - Public hierarchical sequence available; publication released in August 2005

WGS – Major Challenge 1
- Assembly of repetitive DNA is difficult
  - Retrotransposons (RNA mobile elements)
  - DNA transposons
  - Alu repeats (human)
  - Long and Short Interspersed Repeat (LINE and SINE) elements
  - Microsatellites
- Solution
  - Use sequence data from 2, 10 and 50 kb clones
    - Data from fragments containing different types of sequences can be collected
    - Paired-end reads collected
  - Assembly Process
    - Repeat sequences are initially masked
    - Overlaps of non-repeat sequences detected
    - Contigs overlapped to create supercontigs
  - Software available but is mostly useful to the developers
    - Examples: Celera Assembler, Arcane, Phusion, Atlas
WGS – Major Challenge 2
- For the two sequences approaches
- Assembly is a scale issue
  - WGS approach
    - Gigabytes of sequence data
  - Hierarchical approach
    - Magnitudes less
  - On-going research focuses on developing new algorithms to handle and assembly the huge data sets generated by WGS

Mouse WGS Data
- 29.7 million reads
- 7.4x coverage
- Newer software
- Assembled without mapping or clone data
  - Human WGS had access to this data from the public project
- 225,000 contigs
  - Mean length = 25 kilobases in length
- Super contig subset
  - Mean length =16.9 megabases
- 200 largest supercontigs
  - Anchored using mapping data
  - Represents 96% (9187 Mb) of the euchromatic region of genome
Rat Genome Project: A combined approach

- Combination of hierarchical shotgun and whole genome shotgun sequencing

- WGS sequence reads
  - 36 million quality reads (34 million used for assembly)
  - 7X coverage
  - 60%: Whole genome shotgun data
  - Insert size: <10 kb, 10 kb, 50 kb, >150 kb
  - 40% BAC data
  - Small insert clones from the BAC

- BAC Skim
  - A low density sequence analysis of a BAC
  - 21,000 clones analyzed
  - 1.6X coverage

- Enriched BACS
  - Sequences developed by combining WGS data and BAC skim data

- BAC Fingerprinting
  - 200,000 BACs fingerprinted
  - 12X coverage
  - 11,274 fingerprint contigs (FPC) developed
  - Clones selected from contigs for BAC skim

- Bactig
  - Overlapping BACs
  - 1MB in length
• Superbactigs
  o Bactigs joined by paired-end reads
  o Mean = 5MB in length
  o 783 total for the genome

• Ultrabactigs
  o Mean = 18 MB
  o 291 total for genome
  o Synteny data, marker data, and other data used to define the ultrabactig
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 N5
Short-read Sequencing Projects

The Panda Project (genome = 2.4 Gb)

- First genome assembled fully from short reads

Sequencing
- 37 paired-end libraries
  - 150 bp, 500 bp, 2 kb, 5 kb, 10 kb in size
- 176 Gb usable sequence data
  - 73x coverage

Assembly
- SOAPdenovo used for assembly
  - Part of SOAP software package
    - Short Oligonucleotide Analysis Package
      - "SOAPdenovo uses the de Bruijn graph algorithm and applies a stepwise strategy to make it feasible to assemble the panda genome using a supercomputer (32 cores and 512 Gb random access memory (RAM)."
- Poor library and low quality reads excluded
  - 134 Gb sequence data used
    - 56x coverage
Step 1. Contig building
- Data from 500 bp or smaller libraries used first
- Assembly halted when repeat region encountered
  - 39X coverage achieved
  - N50 = 1.5 kb
  - Length = 2.0 Gb

Step 2. Scaffold building
- Paired-end data from all libraries used
  - N50 = 1.3 Mb
  - Total length = 2.3 Gb

Step 3. Closing the gaps
- Local assembly (within a specific gap) using paired end read with one end in a contig and the other in a gap
  - 223.7 Mb gaps closed
  - 54.2 remained unclosed

Step 4. Compare with other carnivores
- Determined that gaps most likely repetitive elements
Genome Assembly

Goal of assembly
- Create contigs based on similarity of sequence reads

Issues that make assembly difficult
- Sequencing errors
  - Hard to ascertain, so ignored during assembly
- Repetitive sequences
  - Some found 100,000 times
  - Repeats will lead to incorrect assembly
  - Hard to know which sequences overlap
  - Brings two regions together that are not in fact together
  - Resolving some repeats
    - If repeat is shorter than read length, assembly is possible
- Unclonable sequences
  - Some sequences lethal to bacteria
    - Cannot be cloned, so sequence data is missing
    - Not an issue with massively parallel sequencing
- How to overcome these problems
  - Finishing
  - But finishing is expensive
    - Want to ensure that most of the sequence data available is used in assembly
Result of errors, suboptimum coverage, repeats

- Many more contigs than expected

Assembly problem

- Finding the shortest superstring \((T)\) from a set of strings \((s_1, s_2, \ldots, s_n)\)

Features of original assembly algorithms

- Greedy Algorithm Approach
  - Compute all possible overlaps between strings and assign a quality score
  - Merge strings with highest score
  - Continue until no other strings can be merged
  - Uses greedy algorithm
- Fastest method to a solution
- Doesn’t guarantee optimum solution
  - Approach doesn’t work for large genomes
  - Large RAM memory requirements
Next generation assembly algorithms

- Graph theory approach
  - Graph definition
    - A mathematical structure that models pairs of objects from a collection of objects
    - For sequencing the objects are sequence reads

- Overlap-layout consensus approach
  - Set a sequence as a node
  - Overlaps are edges
  - Contig is a path of nodes and edges

- Process
  - Find all possible alignments
  - Remove overlap duplications
  - Construct consensus to create contig
ARACHNE Assembly Program

Data Preparation

1. Trim low quality sequences at ends of reads
2. Drop entire reads with low overall read scores
3. Trim vector sequences and any known contaminant sequences

Alignment of reads

1. Create table of k-mer (k=24) sequences
   a. each entry associated with a read and position in read
2. High frequency k-mers dropped (repetitive sequences)
3. Read pairs sharing k-mers identified
4. Overlapping k-mers are merged
5. Shared k-mers extended
6. Alignments refined

Error Correction and Quality Scoring

1. Multiple alignments of overlapping reads created
2. Low frequency errors (20 C vs. 1 T) converted to consensus sequence
3. Insertions/deletions are corrected
4. Quality score attached to alignment
Building the Contig and Repeat Contig

1. Plasmids (of same insert size) containing paired reads from both ends are identified; these are called **paired reads**
2. Paired reads are merged into contigs that do not cross a repeat region
3. Contig built until a repeat boundary is confronted
   a. These are called **unitig** (**unique contig**)
4. Repeat contig, formed by collapsing identical sequences from unique regions, are marked
   a. Repeat contigs have high copy number
   b. Repeat contigs are difficult to assembly with other contigs

Supercontigs

1. Unitigs containing two forward and two reverse links are merged
   a. Contigs with the most links and over the shortest distance are preferred
2. Process repeated by merging previously merged contigs into **supercontigs**
3. Repeat contigs used in an attempt to fill gaps between supercontigs

Arachne2

1. Extended supercontigs
2. Tested for weak and strong supercontigs with misassembly
3. Reassembled these questionable supercontigs
1. Breaks 600 nt read into 24 nt sequences and note read origin of the sequence
2. Create database with each sequence as main entry
   • Each sequence entry contains frequency and read identifier data
4. Discard high copy reads (these are repeats)
5. Align reads from low frequency sequences
6. Discover mate pairs represented in two plasmids of same length
   • These are paired pairs
7. Find a mate pair that matches only one end of the paired pair
   • Sequences are considered to be a single large read
8. Process continues until a repeat is encountered
9. Assembly stops and a unique contig is declared
10. Overlaps of unique contigs discovered
11. Supercontigs are declared
## Common Bean: 454-based Project

### Sequencing Libraries

<table>
<thead>
<tr>
<th>Library</th>
<th>Sequencing Platform</th>
<th>Average Read/Insert Size</th>
<th>Read Number</th>
<th>Assembled Sequence Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>454 XLR &amp; FLX+</td>
<td>362</td>
<td>38,107,155</td>
<td>18.64x</td>
</tr>
<tr>
<td>GPNB</td>
<td>454 XLR paired</td>
<td>2,798 ± 1,047</td>
<td>589,346</td>
<td>0.11x</td>
</tr>
<tr>
<td>GGAS</td>
<td>454 XLR paired</td>
<td>3,922 ± 643</td>
<td>1,940,576</td>
<td>0.41x</td>
</tr>
<tr>
<td>GXSF</td>
<td>454 XLR paired</td>
<td>3,991 ± 337</td>
<td>467,414</td>
<td>0.07x</td>
</tr>
<tr>
<td>HYFA</td>
<td>454 XLR paired</td>
<td>4,729 ± 497</td>
<td>1,648,022</td>
<td>0.25x</td>
</tr>
<tr>
<td>HYFC</td>
<td>454 XLR paired</td>
<td>4,736 ± 504</td>
<td>1,491,648</td>
<td>0.24x</td>
</tr>
<tr>
<td>HYFB</td>
<td>454 XLR paired</td>
<td>4,759 ± 528</td>
<td>1,196,104</td>
<td>0.17x</td>
</tr>
<tr>
<td>HXTI</td>
<td>454 XLR paired</td>
<td>8,022 ± 1,016</td>
<td>1,364,808</td>
<td>0.22x</td>
</tr>
<tr>
<td>GXXN</td>
<td>454 XLR paired</td>
<td>9,192 ± 1,058</td>
<td>878,832</td>
<td>0.16x</td>
</tr>
<tr>
<td>HXWF</td>
<td>454 XLR paired</td>
<td>11,903 ± 1,928</td>
<td>724,196</td>
<td>0.13x</td>
</tr>
<tr>
<td>HXWH</td>
<td>454 XLR paired</td>
<td>12,231 ± 1,902</td>
<td>413,396</td>
<td>0.08x</td>
</tr>
<tr>
<td>VUK (Fosmid-end)</td>
<td>Sanger</td>
<td>34,956 ± 4,536</td>
<td>240,384</td>
<td>0.20x</td>
</tr>
<tr>
<td>VUL (Fosmid-end)</td>
<td>Sanger</td>
<td>36,001 ± 4,632</td>
<td>88,320</td>
<td>0.08x</td>
</tr>
<tr>
<td>PVC (BAC-end)</td>
<td>Sanger</td>
<td>121,960 ± 16,572</td>
<td>81,408</td>
<td>0.08x</td>
</tr>
<tr>
<td>PVA (BAC-end)</td>
<td>Sanger</td>
<td>126,959 ± 25,658</td>
<td>89,017</td>
<td>0.09x</td>
</tr>
<tr>
<td>PVB (BAC-end)</td>
<td>Sanger</td>
<td>135,292 ± 21,487</td>
<td>92,160</td>
<td>0.09x</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>N/A</td>
<td></td>
<td>49,412,786</td>
<td>21.02x</td>
</tr>
</tbody>
</table>
Genome Assembly Statistics

Comparison of Two Legume Species and Sanger and Short Read Sequence Data Collection

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Soybean</th>
<th>Common Bean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencing method</td>
<td>WGS, Sanger</td>
<td>WGS, 454 &amp; Illumina</td>
</tr>
<tr>
<td>Genome size (contig)</td>
<td>955 Mb (1.9% gap)</td>
<td>473 Mb (9.3% gap)</td>
</tr>
<tr>
<td>Genome size (scaffold)</td>
<td>973 Mb</td>
<td>521 Mb</td>
</tr>
<tr>
<td>Contig number</td>
<td>16,311</td>
<td>41,391</td>
</tr>
<tr>
<td>Contig N50</td>
<td>1,492</td>
<td>3.273</td>
</tr>
<tr>
<td>Contig L50</td>
<td>189 kb</td>
<td>39.5 kb</td>
</tr>
<tr>
<td>Scaffold number</td>
<td>1,168</td>
<td>708</td>
</tr>
<tr>
<td>Scaffold N50</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Scaffold L50</td>
<td>47.8 Mb</td>
<td>50.4 Mb</td>
</tr>
<tr>
<td>Genetic map loci</td>
<td>1,536 SNPs</td>
<td>7,015 SNPs</td>
</tr>
</tbody>
</table>
Genome Assembly Statistics

Comparison of Two Sequencing Methods for One Species

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Common Bean</th>
<th>Common Bean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencing method</td>
<td>WGS, 454 &amp; Illumina</td>
<td>WGS, PacBio</td>
</tr>
<tr>
<td>Genome size (contig)</td>
<td>473 Mb (9.3% gap)</td>
<td>532 Mb (1.1% gap)</td>
</tr>
<tr>
<td>Genome size (scaffold)</td>
<td>521 Mb</td>
<td>537 Mb</td>
</tr>
<tr>
<td>Contig number</td>
<td>41,391</td>
<td>1,044</td>
</tr>
<tr>
<td>Contig N50</td>
<td>3,273</td>
<td>73</td>
</tr>
<tr>
<td>Contig L50</td>
<td>39.5 kb</td>
<td>1.9 Mb</td>
</tr>
<tr>
<td>Scaffold number</td>
<td>708</td>
<td>478</td>
</tr>
<tr>
<td>Scaffold N50</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Scaffold L50</td>
<td>50.4 Mb</td>
<td>49.7 Mb</td>
</tr>
<tr>
<td>Genetic map loci</td>
<td>7,015 SNPs</td>
<td>7,015 SNPs</td>
</tr>
<tr>
<td>% genome in scaffolds &gt;50 kb</td>
<td></td>
<td>99.1%</td>
</tr>
</tbody>
</table>

PacBio Scaffold Sets

<table>
<thead>
<tr>
<th>Set</th>
<th>Scaffolds</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main genome</td>
<td>478</td>
<td>537 Mb</td>
</tr>
<tr>
<td>Mitochondrion</td>
<td>7</td>
<td>448 Kb</td>
</tr>
<tr>
<td>Chloroplast</td>
<td>29</td>
<td>662 Kb</td>
</tr>
<tr>
<td>Unanchored rDNA</td>
<td>7</td>
<td>296 Kb</td>
</tr>
<tr>
<td>Alternative haplotypes</td>
<td>442</td>
<td>9 Mb</td>
</tr>
<tr>
<td>Repeat scaffolds</td>
<td>275</td>
<td>10 Mb</td>
</tr>
<tr>
<td>Excluded (&lt;1kb)</td>
<td>5</td>
<td>4 Kb</td>
</tr>
</tbody>
</table>
## Comparison of Plant Genome Species

<table>
<thead>
<tr>
<th>Species (Sequencing method)</th>
<th>N50</th>
<th>L50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brachypodium (Sanger)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contigs</td>
<td>252</td>
<td>347.8 Kb</td>
</tr>
<tr>
<td>Scaffolds</td>
<td>3</td>
<td>9.3 Mb</td>
</tr>
<tr>
<td>Sorghum (Sanger)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contigs</td>
<td>958</td>
<td>195.4 Kb</td>
</tr>
<tr>
<td>Scaffolds</td>
<td>6</td>
<td>62.4 Mb</td>
</tr>
<tr>
<td>Soybean (Sanger)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contigs</td>
<td>1,492</td>
<td>189.4 Kb</td>
</tr>
<tr>
<td>Scaffolds</td>
<td>10</td>
<td>47.8 Mb</td>
</tr>
<tr>
<td>Common Bean (454)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contigs</td>
<td>3,273</td>
<td>39.5 Kb</td>
</tr>
<tr>
<td>Scaffolds</td>
<td>5</td>
<td>50.4 Mb</td>
</tr>
<tr>
<td>Canola (454/Sanger/Illumina)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contigs</td>
<td></td>
<td>38.9 kb</td>
</tr>
<tr>
<td>Scaffolds</td>
<td></td>
<td>763.7 Kb</td>
</tr>
</tbody>
</table>
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

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

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

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

**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.
<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>Arabidopsis thaliana</td>
<td>Arabidopsis</td>
<td>Columbia</td>
<td>2000</td>
<td>Nature 408:796</td>
<td>HSS/S</td>
<td>5</td>
<td>125/135</td>
<td>20^i</td>
<td>18-29</td>
<td>27,416/35,386</td>
<td></td>
<td></td>
<td>Eudicot 3x + Brassicaceae (2x+2x)</td>
</tr>
<tr>
<td>Populus trichocarpa</td>
<td>Poplar</td>
<td>Nisqualy 1</td>
<td>2006</td>
<td>Science 313:1596</td>
<td>WGS/S</td>
<td>19</td>
<td>485/423</td>
<td>40^i</td>
<td>11-36</td>
<td>41,335/73,013</td>
<td></td>
<td></td>
<td>Eudicot 3x + (2x)</td>
</tr>
<tr>
<td>Vitus vinifera</td>
<td>Grape</td>
<td>PN40024</td>
<td>2007</td>
<td>Nature 449:463</td>
<td>WGS/S</td>
<td>19</td>
<td>475/487</td>
<td>22^i</td>
<td>10-22</td>
<td>/ 26,346</td>
<td></td>
<td></td>
<td>Eudicot 3x</td>
</tr>
<tr>
<td>Carica papaya</td>
<td>Papaya</td>
<td>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>27,332/27,996</td>
<td>39/1,000</td>
<td></td>
<td></td>
<td>Eudicot 3x</td>
</tr>
<tr>
<td>Zea mays</td>
<td>Maize</td>
<td>B73</td>
<td>2009</td>
<td>Science 326:1112</td>
<td>HSS/S</td>
<td>10</td>
<td>/3,234</td>
<td>84</td>
<td>150-301</td>
<td>39,475/137,208</td>
<td></td>
<td></td>
<td>Poales (2x+2x) + (2x)</td>
</tr>
<tr>
<td>Cucumis sativus</td>
<td>Cucumber</td>
<td>9930</td>
<td>2009</td>
<td>Nat Genet 41:1275</td>
<td>WGS/S</td>
<td>7</td>
<td>??/244</td>
<td>22^i</td>
<td>21,491/32,528</td>
<td>??/227</td>
<td></td>
<td></td>
<td>Eudicot 3x</td>
</tr>
<tr>
<td>Malus x domestica</td>
<td>Apple</td>
<td>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>
</tr>
<tr>
<td>Jatropha curcas</td>
<td>Jatropha</td>
<td></td>
<td>2010</td>
<td>DNA Res 18:65</td>
<td>WGS/S</td>
<td></td>
<td>380/285</td>
<td>37</td>
<td>40,929/??</td>
<td>??/4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fragaria vesca</td>
<td>Strawberry</td>
<td>H4x4</td>
<td>2011</td>
<td>Nat Genet 43:109</td>
<td>WGS/S, 454, 1, So</td>
<td>7</td>
<td>240/220</td>
<td>23</td>
<td>32,831/??</td>
<td>??/1,300</td>
<td></td>
<td></td>
<td>Eudicot 3x</td>
</tr>
<tr>
<td>Arabidopsis lyrata</td>
<td>Lyra</td>
<td>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></td>
<td>Eudicot 3x + Brassicaceae (2x+2x)</td>
</tr>
<tr>
<td>Phoenix dactylfera</td>
<td>Date palm</td>
<td>Kharas</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>28,890/??</td>
<td>??/6</td>
<td>??/30</td>
<td></td>
<td>Eudicot 3x + Brassicaceae (2x+2x)</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>Potato</td>
<td>DM1-3 516 R44</td>
<td>2011</td>
<td>Nature 475:189</td>
<td>WGS/S, 454, 1, So</td>
<td>12</td>
<td>844/727</td>
<td>62</td>
<td>35,119/51,472</td>
<td>6,446/31</td>
<td>121/1,782</td>
<td></td>
<td>Eudicot 3x + Solanaceae 3x</td>
</tr>
<tr>
<td>Thellungiella parvula</td>
<td>Thellungiella</td>
<td></td>
<td>2011</td>
<td>Nat Genet 43:913</td>
<td>WGS/S, 454, 1</td>
<td>7</td>
<td>160/137</td>
<td>8</td>
<td>30,419/??</td>
<td>8/5,290</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cucumis sativus</td>
<td>Cucumber</td>
<td>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></td>
<td></td>
<td>Eudicot 3x</td>
</tr>
<tr>
<td>Brassica rapa</td>
<td>Cabbage</td>
<td>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>41,174/27,788</td>
<td>39/1,971</td>
<td></td>
<td></td>
<td>Brassicaceae 2x + (2x)</td>
</tr>
<tr>
<td>Cajanus cajan</td>
<td>Pigeon pea</td>
<td>ICPL 87119</td>
<td>2011</td>
<td>Nat Biotech 30:83</td>
<td>WGS/S, 1</td>
<td>11</td>
<td>808/606</td>
<td>52</td>
<td>10-48</td>
<td>40,071/7815/23</td>
<td>380/516</td>
<td></td>
<td>Eudicot 3x + Legume 2x</td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>Medicago</td>
<td></td>
<td>2011</td>
<td>Nature 480:520</td>
<td>WGS/S, 454, 1</td>
<td>8</td>
<td>454/384</td>
<td>35-57</td>
<td>44,135/45,888</td>
<td>53/1270</td>
<td></td>
<td></td>
<td>Eudicot 3x + Legume 2x</td>
</tr>
<tr>
<td><strong>Species name</strong></td>
<td><strong>Common name</strong></td>
<td><strong>Genotype</strong></td>
<td><strong>Year</strong></td>
<td><strong>Publication</strong></td>
<td><strong>Technical method</strong></td>
<td><strong># Chrom</strong></td>
<td><strong>Est. genome size/assembled size (Mb)</strong></td>
<td><strong>Repeat content (%)</strong></td>
<td><strong>Chrom size range (Mb)</strong></td>
<td><strong># genes/transcripts</strong></td>
<td><strong>Contig N50/L50 (#/kb)</strong></td>
<td><strong>Scaffold N50/L50 (#/kb)</strong></td>
<td><strong>Genome duplication history</strong></td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------</td>
<td>--------------</td>
<td>----------</td>
<td>-----------------</td>
<td>----------------------</td>
<td>------------</td>
<td>-----------------------------------------</td>
<td>-----------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>---------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td><strong>Solanum lycopersicon</strong></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>Eudicot 3x + Solanaceae 3x</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Linum usitatissimum</strong></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>Eudicot 3x + (2x)</td>
<td></td>
</tr>
<tr>
<td><strong>Musa acuminata</strong></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>36/1</td>
<td>/1,311</td>
<td>Zingiberales 2x + (2x + 2x)</td>
</tr>
<tr>
<td><strong>Gossypium raimondii</strong></td>
<td>Cotton (B genome diploid)</td>
<td>2012</td>
<td>Nature 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><strong>Azadirachta indica</strong></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>13</td>
<td>20,169/?</td>
<td>??/0.7</td>
<td>??/452</td>
<td>Eudicot 3x + Gossypium 2x</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Prunus mume</strong></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>Eudicot 3x + Rosaceae 2x</td>
<td></td>
</tr>
<tr>
<td><strong>Pyrus bretschneideri</strong></td>
<td>Pear</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>
<td></td>
</tr>
<tr>
<td><strong>Citrullus lanatus</strong></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><strong>Morus notabilis</strong></td>
<td>Mulberry</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>Eudicot 3x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Physical Distance vs. Genetic Distance

Typical Chromosome
- High recombination on ends of chromosome
- Low recombination in the center (heterochromatic) region of the genome

Acrocentric Chromosome
- Heterochromatic repeat rich region at end of chromosome