Bioinformatics and Sequencing
Relevant to SoICAP

C. Robin Buell
Department of Plant Biology
Michigan State University
East Lansing MI 48824

buell@msu.edu

An Overview of
DNA Sequencing
Prokaryotic DNA


Eukaryotic DNA

The two strands of a DNA molecule are held together by weak bonds (hydrogen bonds) between the nitrogenous bases, which are paired in the interior of the double helix.

The two strands of DNA are antiparallel; they run in opposite directions. The carbon atoms of the deoxyribose sugars are numbered for orientation.

The goal of sequencing DNA is to tell the order of the bases, or nucleotides, that form the inside of the double-helix molecule.

High throughput sequencing methods
- Sanger/Dideoxy
- Next Generation (NextGen)
Whole Genome Shotgun Sequencing

- Start with a whole genome
- Shear the DNA into many different, random segments.
- Sequence each of the random segments.
- Then, put the pieces back together again in their original order using a computer
Theory Behind Shotgun Sequencing

*Haemophilus influenzae* 1.83 Mb base

Coverage | unsequenced (%)
---|---
1X | 37%
2X | 13%
5X | 0.67%
6X | 0.25%
7X | 0.09%

For 1.83 Mb genome, 6X coverage is 10.98 Mb of sequence, or 22,000 sequencing reactions, 11000 clones (1.5-2.0 kb insert), 500 bp average read.
- Initial dideoxy sequencing involved use of radioactive dATP and 4 separate reactions (ddATP, ddTTP, ddCTP, ddGTP) & separation on 4 separate lanes on an acrylamide gel with detection through autoradiogram

- New technologies use 4 fluorescently labeled bases and separation on capillaries and detection through a CCD camera
Data Analysis

• An chromatogram is produced and the bases are called

• Software assign a quality value to each base
  • Phred & TraceTuner
    • Read DNA sequencer traces
    • Call bases
    • Assign base quality values
    • Write basecalls and quality values to output files.
454 Genome Sequencing System

- Library prep, amplification and sequencing: 2-4 days
- Single sample preparation from bacterial to human genomic DNA
- Single amplification per genome with no cloning or cloning artifacts
- Picoliter volume molecular biology
- 400 Mb per run (4-5 hr); less than $15,000 per run
- Read lengths 200-230 bases; new Titanium platform, 400 Mb per run, 400-500 bases per reads
- Massively parallel imaging, fluidics and data analysis
- Requires high genome coverage for good assembly
- Error rate of 1-2%
- Problem with homopolymers

454-Pyrosequencing

- Construct Single stranded adaptor
- Ligated DNA
- Perform emulsion PCR
- Depositing DNA Beads into the PicoTiter™Plate
- Sequencing by Synthesis:
  - Simultaneous sequencing of the entire genome in hundreds of thousands of picoliter-size wells
  - Pyrophosphate signal generation
Solexa/Illumina Sequencing

- Sequencing by synthesis (not chain termination)
- Generate up to 12 Gb per run
Other “Next Generation” Sequencing Technologies

- SoLiD by Applied Biosystems - short reads (~25-75 nucleotides)
- Helicos - short reads (<50 nucleotides)
- Pacific Biosystems - LONG reads (several kilobases)

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome Size (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potato</td>
<td>850 Mb</td>
</tr>
<tr>
<td>Wheat</td>
<td>16,000 Mb</td>
</tr>
<tr>
<td>Rice</td>
<td>430 Mb</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>130 Mb</td>
</tr>
<tr>
<td>John Doe</td>
<td>2,500 Mb</td>
</tr>
</tbody>
</table>

How much sequences are needed to assemble a eukaryotic genome?

- Depends on the genome size of the organism (genes plus repeats), ploidy level, heterozygosity, desired quality
Eukaryotic Genomes and Gene Structures

Gene | Intergenic Gene Region | Intergenic Gene Region

Expressed Sequence Tags (ESTs): Sampling the Transcriptome and Genic Regions

What is an EST?
- single pass sequence from cDNA
- specific tissue, stage, environment, etc.

- cDNA library in *E.coli*
- pick individual clones
- template prep

Multiple tissues, states..
- with enough sequences, can ask quantitative questions
Uses of EST sequencing:
- Gene discovery
- Digital northerns/insights into transcriptome
- Genome analyses, especially annotation of genomic DNA
- SNP discovery in genic regions

Issues with EST sequencing:
- Inherent low quality due to single pass nature
- Not 100% full length cDNA clones
- Redundant sequencing of abundant transcripts

Address through clustering/assembly to build consensus sequences:
- Gene Index
- Unigene Set
- Transcript Assembly

Locus/Gene
Gene models
Full length cDNAs
Expressed Sequence Tags
Types of Genomic/DNA-based Diagnostic Markers

1. Restriction Fragment Length Polymorphisms (RFLPs)
2. Random Amplification of Polymorphic DNA (RAPDs)
3. Cleaved Amplified Polymorphisms (CAPs)
4. Amplified Fragment Length Polymorphisms (AFLPs)
5. Simple Sequence Repeats (SSRs; microsatellites)
6. Single Nucleotide Polymorphisms (SNPs)

SSRs
-Specific primers that flank simple sequence repeat (mono-, di-, tri-, tetra-, etc) which has a higher likelihood of a polymorphism
-Amplify genomic DNA
-Separate on gel
-Look for size polymorphisms
SSRs
-Computational prediction of SSRs in potato transcriptome data
- http://solanaceae.plantbiology.msu.edu/analyses_ssr_query.php

SNPs
-Specific primers
-Amplify genomic DNA
-Detect mismatch (many methods for this)

http://omib bjmu.edu.cn/omibdata/snp/images/SNP.gif
**Potato SNPs:** Intra-varietal and inter-varietal
- Bulk of sequence data from ESTs (Sanger derived)
- Use computational methods to identify SNPs within existing potato ESTs
- [http://solanaceae.plantbiology.msu.edu/analyses_snp.php](http://solanaceae.plantbiology.msu.edu/analyses_snp.php)

**Illumina Paired End RNA-Seq**

- Potato Varieties: Atlantic, Premier, Snowden
- Two Paired End RNA-Seq runs were performed.
- Reads are 61bp long
- Insert sizes:
  - Atlantic: 350bp
  - Premier: 300bp
  - Snowden: 300bp
- Paired End Sequencing is carried out by an Illumina module that regenerates the clusters after the first run and sequences the clusters from the other end.
Velvet Assemblies of Potato Illumina Sequences

- With a read length of 31 and a minimum contig length of 150bp:
- Atlantic: 45214 contigs
- Premier: 54917 contigs
- Snowden: 58754 contigs

Sequence quality: Viewing a Atlantic potato contig from the Velvet assembly
Identify intra-varietal SNPs

<table>
<thead>
<tr>
<th>Query</th>
<th>Query SNPs</th>
<th>Filtered SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlantic Asm</td>
<td>224748</td>
<td>150669</td>
</tr>
<tr>
<td>Premier Asm</td>
<td>265673</td>
<td>181800</td>
</tr>
<tr>
<td>Snowden Asm</td>
<td>258872</td>
<td>166253</td>
</tr>
</tbody>
</table>

Hawkeye Viewer – Visualizing SNPs

A/C SNP

G/T SNP
**Analyses in progress**

**SNP Identification:**
- Identify inter-varietal SNPs using draft genome sequence from *S. phureja*
- Identify only biallelic SNPs
- Identify high confidence SNPs
- Identify SNPs that meet Infinium design requirements

**SNP Selection:**
- Annotate transcripts for gene function
- Identify candidate genes within SNP set