Course Overview

02-715 Advanced Topics in Computational Genomics
Course Overview

- Instructor: Seyoung Kim (Lane Center for Computational Biology, CMU)
- Course Website: [www.cs.cmu.edu/~sssykim/teaching/s13/s13.html](http://www.cs.cmu.edu/~sssykim/teaching/s13/s13.html)
- Location: DH 2105
- Time: Monday, Wednesday, & Friday: 3:30-4:20pm
- Office hours: Friday 4:30-5:30pm
Grading

• Write-ups for required reading (30%)
  – Starting the 2nd week
  – Summary of contributions, critique (strengths and weaknesses).
  – Under 300 words for each paper.
  – Submit to blackboard by midnight the day before the class.
    • Late submission policy: 70% before the class, 0% afterwards.

• Class participation (20%)

• Paper presentation (30%)

• Final project (20%)
  – One-page project proposal: due March 18 in class.
  – Project presentation: the last week of the course.
  – Final project report: due May 10th.
Overview

• Next-generation sequencing technology

• Genetic polymorphisms

• Population genetics review
  – Haplotype inference, recombination rate estimation, linkage disequilibrium, tag SNPs

• From Human Genome Sequencing Project to HapMap Project to 1000 Genome Project
Decline in Sequencing Costs

Cost and Growth of Bases

Billions of bases

Cost (\$)

SOURCE: NCBI

Cost per million base pairs of sequence (log scale)

$10,000

$1,000

$100

$10

$1

GenBank

2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
DNA sequencing – vectors

DNA

Shake

DNA fragments

Vector
Circular genome (bacterium, plasmid)

+  =  

Known location
(restriction site)

Adopted from http://www.cs.utoronto.ca/~brudno/csc2431w10/2431_lec1.ppt
Method to sequence longer regions

- genomic segment
- cut many times at random (Shotgun)
- Get two reads from each segment
- ~500 bp

Adopted from http://www.cs.utoronto.ca/~brudno/csc2431w10/2431_lec1.ppt
Reconstructing the Sequence (Fragment Assembly)

Cover region with ~7-fold redundancy (7X)

Overlap reads and extend to reconstruct the original genomic region

Adopted from http://www.cs.utoronto.ca/~brudno/csc2431w10/2431_lec1.ppt
Definition of Coverage

Length of genomic segment: \( L \)
Number of reads: \( n \)
Length of each read: \( l \)

**Definition:** Coverage \( C = \frac{n \cdot l}{L} \)

How much coverage is enough?

**Lander-Waterman model:**
Assuming uniform distribution of reads, \( C=10 \) results in 1 gapped region /1,000,000 nucleotides

Adopted from http://www.cs.utoronto.ca/~brudno/csc2431w10/2431_lec1.ppt
Depth of Coverage and Physical Coverage

- Single-end sequencing
- Paired-end sequencing
- Paired-end sequencing
<table>
<thead>
<tr>
<th>Platform</th>
<th>Library/template preparation</th>
<th>NGS chemistry</th>
<th>Read length (bases)</th>
<th>Run time (days)</th>
<th>Gb per run</th>
<th>Machine cost (US$)</th>
<th>Pros</th>
<th>Cons</th>
<th>Biological applications</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche/454’s GS FLX Titanium</td>
<td>Frag, MP/emPCR</td>
<td>PS</td>
<td>330*</td>
<td>0.35</td>
<td>0.45</td>
<td>500,000</td>
<td>Longer reads improve mapping in repetitive regions; fast run times</td>
<td>High reagent cost; high error rates in homopolymer repeats</td>
<td>Bacterial and insect genome de novo assemblies; medium scale (&lt;3 Mb) exome capture; 16S in metagenomics</td>
<td>D. Muzny, pers. comm.</td>
</tr>
<tr>
<td>Illumina/Solexa’s GA</td>
<td>Frag, MP/solid-phase</td>
<td>RTs</td>
<td>75 or 100</td>
<td>47, 95</td>
<td>185, 355</td>
<td>540,000</td>
<td>Currently the most widely used platform in the field</td>
<td>Low multiplexing capability of samples</td>
<td>Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics</td>
<td>D. Muzny, pers. comm.</td>
</tr>
<tr>
<td>Life/AGP’s SOLiD 3</td>
<td>Frag, MP/emPCR</td>
<td>Cleavable probe SBL</td>
<td>50</td>
<td>71, 145</td>
<td>30, 505</td>
<td>595,000</td>
<td>Two-base encoding provides inherent error correction</td>
<td>Long run times</td>
<td>Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics</td>
<td>D. Muzny, pers. comm.</td>
</tr>
<tr>
<td>Polonator G.007</td>
<td>MP only/emPCR</td>
<td>Non-cleavable probe SBL</td>
<td>26</td>
<td>53</td>
<td>123</td>
<td>170,000</td>
<td>Least expensive platform; open source to adapt alternative NGS chemistries</td>
<td>Users are required to maintain and quality control reagents; shortest NGS read lengths</td>
<td>Bacterial genome resequencing for variant discovery</td>
<td>J. Edwards, pers. comm.</td>
</tr>
<tr>
<td>Helicos BioSciences HeliScope</td>
<td>Frag, MP/single molecule</td>
<td>RTs</td>
<td>32*</td>
<td>81</td>
<td>375</td>
<td>999,000</td>
<td>Non-bias representation of templates for genome and seq-based applications</td>
<td>High error rates compared with other reversible terminator chemistries</td>
<td>Seq-based methods</td>
<td>91</td>
</tr>
<tr>
<td>Pacific Biosciences (target release: 2010)</td>
<td>Frag only/single molecule</td>
<td>Real-time</td>
<td>964*</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Has the greatest potential for reads exceeding 1 kb</td>
<td>Highest error rates compared with other NGS chemistries</td>
<td>Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks</td>
<td>S. Turner, pers. comm.</td>
</tr>
</tbody>
</table>
Next Generation Sequencing (NGS) based methods

- RNA-Seq: methods for determining mRNA abundance and sequence content
  - Rare transcripts discovery
  - Alternative splicing event detection
  - Transcript sequence variation detection
Next Generation Sequencing (NGS) based methods

- ChIP-Seq: methods for measuring genome-wide profiles of immunoprecipitated DNA-protein complexes
Overview

• Next-generation sequencing technology

• Genetic polymorphisms

• From Human Genome Sequencing Project to HapMap Project to 1000 Genome Project
Why Genetic Variations?

• Genetic variations can be
  – Used to find signatures of evolution, positive selection.
  – Giving insights on population structure.
  – Causal variations that influence phenotypes such as disease susceptibility, drug response: finding them can be the first key steps to cures in medicine.
Genetic Variations

• Types of genetic variations
  – Single nucleotide polymorphisms (SNPs)
    • Widely used as genetic markers
    • Highly abundant in genomes
  – Structural variants: insertions/deletions, duplications, copy number variations
Other Genetic Variations

- **Copy Number Variation**
  - DNA segment whose numbers differ in different genomes
    - Kilobases to megabases in size
  - Usually two copies of all autosomal regions, one per chromosome
  - Variation due to deletion or duplication
Variant Frequencies from 1000 Genome Pilot Project

![Graph showing variant frequencies from the 1000 Genome Pilot Project. The x-axis represents log10(size) with various size markers such as -100 kb, -10 kb, -100 bp, 10 bp, 1 kb, 10 kb, and 100 kb. The y-axis represents log10(number of variants) ranging from 0 to 10. The graph includes bars for Deletions, SNPs, and Insertions, with different colors for LINE and Alu elements. The purple line represents the proportion of variants that are novel.]
Terminology

- **Allele**: different forms of genetic variations at a given gene or genetic locus
- **Genotype**: specific allelic make-up of an individual’s genome
- **Heterozygous/Homozygous**
Terminology

- Haplotype: A collection of alleles derived from the same chromosome

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Haplotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 13</td>
<td>2 13</td>
</tr>
<tr>
<td>1 6</td>
<td>6 1</td>
</tr>
<tr>
<td>9 15</td>
<td>9 15</td>
</tr>
<tr>
<td>4 17</td>
<td>4 17</td>
</tr>
<tr>
<td>1 9</td>
<td>1 9</td>
</tr>
<tr>
<td>2 6</td>
<td>2 6</td>
</tr>
<tr>
<td>9 17</td>
<td>9 17</td>
</tr>
<tr>
<td>2 12</td>
<td>2 12</td>
</tr>
<tr>
<td>7 12</td>
<td>7 12</td>
</tr>
<tr>
<td>6 14</td>
<td>6 14</td>
</tr>
<tr>
<td>1 7</td>
<td>1 7</td>
</tr>
<tr>
<td>18 18</td>
<td>18 18</td>
</tr>
<tr>
<td>1 4</td>
<td>1 4</td>
</tr>
<tr>
<td>10 10</td>
<td>10 10</td>
</tr>
</tbody>
</table>

Chromosome phase is unknown

Chromosome phase is known
Working with SNP Data in Practice

• At each locus, SNPs are represented as 0 or 1.
  – A/T/C/G letters are converted to 0 or 1 for minor/major alleles
  – Genotypes at each locus of each individual are coded as
    • 0 : minor allele homozygous
    • 1: heterozygous
    • 2: major allele homozygous

• Given genotype data for $N$ individuals
  • (Minor allele frequency) = (the number of individuals with minor alleles)/(total number of individuals)
Detecting Genome Alterations with SNP Arrays (Affymetrix GeneChip Probe Array)
Detecting Genome Alterations with Next Generation Sequencing Technology
Sequencing vs. SNP Genotyping

• Sequencing a whole genome is much more costly than genotyping a small number of genetic loci for SNPs
Linkage Disequilibrium in HapMap Data

• $r^2$ in HapMap Data
Using Reference Datasets for Genotype Imputation

- Reference data: dense SNP data from HapMap III
- New data: SNP data for individuals in a given study
- Data after imputation
Using Reference Datasets for Genotype Imputation

- Reference data: sequence data from 1000 genome project
- New data: SNP data for individuals in a given study
- Data after imputation
Genotype Imputation

PHASE can be used for imputation!
Overview

• Next-generation sequencing technology

• Genetic polymorphisms

• From Human Genome Sequencing Project to HapMap Project to 1000 Genome Project
A Little Bit of History

• 2001: A draft of human genome sequence become available

• 2001: The International SNP Map Working Group publishes a SNP Map of 1.42 million SNPs that contained all SNPs identified so far

• 2005: HapMap Phase I
  – Genotype at least one common SNP (MAF>5%) every 5kb across 270 individuals
  – Geographic diversity
    • 30 trios from Yoruba in Ibadan, Nigeria (YRI)
    • 30 trios of European ancestry living in Utah (CEPH)
    • 45 unrelated Han Chinese in Beijing (CHB)
    • 45 unrelated Japanese (JPT)
  – 1.3 million SNPs
A Little Bit of History

• 2007: HapMap Phase II
  – Genotype additional 2.1 million SNPs for the same individuals
  – SNP density about 1 per kb
  – Estimated to contain 25-35% of all 9-10 million common SNPs in assembled human genome.

• 2010: HapMap Phase III
  – 1184 individuals from 11 populations, including HapMap Phase I, II samples
  – Rare variants (MAF=0.05-0.5%), low frequency variants (MAF=0.5%-5%)
  – Copy number variations, resequencing of selected regions

• 2010 : 1000 Genome Pilot Project
  – A more complete characterization of human genetic variations
Common Variants vs. Rare Variants

• First-generation genome-wide association study (GWAS): common variant common disease hypothesis

• Common variants with minor allele frequency (MAF)>5%
  – dbGap: ~11 million SNPs
  – HapMap: 3.5 million SNPs
  – A successful GWAS requires a more complete catalogue of genetic variations

• Rare variants (MAF<0.5%), low-frequency variants (MAF:0.5%~5%)
  – Captured by sequencing with next-generation sequencing technology
  – Possibly significant contributors to the genetic architecture of disease
    • Causal variants are subject to negative selection
1000 Genome Project  
(The 1000 Genome Project Consortium, Nature 2010)

The goal is to characterize over 95% of variants that are in genomic regions accessible to current high-throughput sequencing technologies and that have allele frequency of 1% or higher (the classical definition of polymorphism) in each of five major population groups (populations in or with ancestry from Europe, East Asia, South Asia, West Africa and the Americas).

Pilot project:  
- 179 individuals from four populations  
  (low coverage: 2-6x)  
- 6 individuals in two trios  
  (deep sequencing: average 42x)  
- 697 individuals from seven populations  
  (exon sequencing of 8,140 exons: average 50x)

Main project: sequence 2500 genomes at 4x coverage
Catalogue of Genetic Variants from 1000 Genome Pilot Project

- 15 million SNPs
- 1 million short insertions/deletions
- 20,000 structural variants

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Low coverage</th>
<th>Trios</th>
<th>Exon (total)</th>
<th>Union across projects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CEU</td>
<td>YRI</td>
<td>CHB+JPT</td>
<td>Total</td>
</tr>
<tr>
<td>Samples</td>
<td>60</td>
<td>59</td>
<td>60</td>
<td>179</td>
</tr>
<tr>
<td>Total raw bases (Gb)</td>
<td>1,402</td>
<td>874</td>
<td>596</td>
<td>2,872</td>
</tr>
<tr>
<td>Total mapped bases (Gb)</td>
<td>817</td>
<td>596</td>
<td>468</td>
<td>1,881</td>
</tr>
<tr>
<td>Mean mapped depth (×)</td>
<td>4.62</td>
<td>3.42</td>
<td>2.65</td>
<td>3.56</td>
</tr>
<tr>
<td>Bases accessed (% of genome)</td>
<td>2.43 Gb</td>
<td>2.39 Gb</td>
<td>2.41 Gb</td>
<td>2.42 Gb</td>
</tr>
<tr>
<td></td>
<td>(86%)</td>
<td>(85%)</td>
<td>(85%)</td>
<td>(86.0%)</td>
</tr>
<tr>
<td>No. of SNPs (% novel)</td>
<td>7,943,827</td>
<td>10,938,130</td>
<td>6,273,441</td>
<td>14,894,361</td>
</tr>
<tr>
<td>Mean variant SNP sites per individual</td>
<td>(33%)</td>
<td>(47%)</td>
<td>(28%)</td>
<td>(54%)</td>
</tr>
<tr>
<td>No. of indels (% novel)</td>
<td>2,918,623</td>
<td>3,335,795</td>
<td>2,810,573</td>
<td>3,019,909</td>
</tr>
<tr>
<td>Mean variant indel sites per individual</td>
<td>(39%)</td>
<td>(52%)</td>
<td>(39%)</td>
<td>(57%)</td>
</tr>
<tr>
<td>No. of deletions (% novel)</td>
<td>354,767</td>
<td>383,200</td>
<td>347,400</td>
<td>361,669</td>
</tr>
<tr>
<td>No. of genotyped deletions (% novel)</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>15,893</td>
</tr>
<tr>
<td>No. of duplications (% novel)</td>
<td>259</td>
<td>320</td>
<td>280</td>
<td>407</td>
</tr>
<tr>
<td>No. of mobile element insertions (% novel)</td>
<td>3,202</td>
<td>3,105</td>
<td>1,952</td>
<td>4,775</td>
</tr>
<tr>
<td>No. of novel sequence insertions (% novel)</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
</tbody>
</table>
1000 Genome Projects: Known vs. Novel Variants
Summary

• Next generation sequencing technology

• Genetics study designs evolve as the technology evolves

• Genetic polymorphisms: SNPs, structural variants