CS 294-81 (2 Units)

Current Topics in Computational Biology
Fall 2012, UC Berkeley

Instructor : Prof. Yun S. Song (yss@eecs.berkeley.edu, 642-2211)

Office hour : M 5–6, Th 1–2

Class Time : Tu 5:00–7:00

Place : 310 Soda Hall

Course web : http://www.eecs.berkeley.edu/~yss/courses/fa12-cs294-81/
Pre-requisite

There are no biology prerequisites for this course, but mathematical maturity will be assumed. Familiarity with the following concepts will be useful:

- Basic algorithms, probability, and statistics
- HMM
- EM algorithm
- Bayesian networks
- MCMC
- Gibbs Sampling

Course Description

This course will cover various algorithms and statistical models for detecting genetic variants (SNPs, indels, structural variants), phasing, and imputing missing data. The discussion will be rigorous, paying particular attention to the mathematical details.
Textbook

- There is no required textbook
- A reading list is available online

Grading

- 40% Paper Presentation (2 students per lecture)
- 60% Class Project (May form groups of 2-3 people)
Course Outline

1  8/28  Introduction to the course
2  9/4   Read Mapping / Alignment
3  9/11  SNP and Genotype Calling 1
4  9/18  Guest Lecture
5  9/25  SNP and Genotype Calling 2
6  10/2  Phasing 1
7  10/9  Phasing 2
8  10/16 Phasing 3
9  10/23 Imputation 1
10 10/30 Imputation 2
11 11/6  Imputation 3
12 11/13 Indel Detection
13 11/20 Structural Variant Detection 1
14 11/27 Structural Variant Detection 2

If you are registered, you should signup for a module
Paper Presentation

- Two students per module.
- Start with an overview and then go into the methodological details.
- In many cases you will need to cover appendices and Supplementary Material.
- You are strongly encouraged to use the whiteboard to discuss the mathematical details.
- Also talk about the accompanying software package, if available. Download and try running it.
Class Project

- Form groups of 2 or 3 people.
- You are strongly encouraged to identify your own project related to the course theme.
- Consult the instructor.
- Start early.
E-mail Policy

- Please do not send me e-mail about the course material if you think the answer is going to require more than two sentences. A question that takes a minute to ask may take tens of minutes to answer over e-mail. Make use of office hours.

- If you send us e-mail, please include “CS 294-81” in subject line. Otherwise, your e-mail may get ignored.
The Timetable of Major Events

Backward in time (in billion years)

Big Bang

Formation of the Earth and Solar system

First Life (Prokaryote)

Nucleated Cells (Eukaryote)

Multi-Cellular Eukaryotes

Present
Speciation

Phylogenetic tree

Mouse
Chimp
Human

150 100 50 Present

Backward in time (in million years)

It is estimated that there are more than 10 million living species on Earth today
The Universal Features of Cells on Earth

(Reference: Molecular Biology of the Cell, by Alberts et al.)

✦ Living organisms: “intricately organized chemical factories that take in matter from their surroundings and use these raw materials to generate copies of themselves.”

✦ The key common features of living organisms:
  
  • Reproduction.
  
  • Heredity: transmission of information specifying the characteristics the offspring shall have.
  
  • The heredity information is stored in DNA.

✦ A multi-cellular organism (e.g., a human body consists of about $10^{13}$ cells) arises from a single cell via cell divisions. All cells in a given individual contain the same heredity information (DNA), barring somatic mutations.
DNA

**Bases:**
- Adenine
- Cytosine
- Guanine
- Thymine

The human genome consists of $3.2 \times 10^9$ base-pairs (bp)

Source: NCBI MBR

Antiparallel strands
Eukaryotic Cell

Human nucleus:
~6µm

Source: NCBI MBR

Watch this video:
http://www.youtube.com/watch?v=OjPcT1uUZiE
Revolution of DNA Sequencing

- The Human Genome Project (started in 1990): Took 10–13 years and 3 billion USD.
- Now a single human genome can be sequenced in a few days, costing a few thousand USD.

Figure source?
Key Technology: Shotgun Short-Read Sequencing (Next-generation sequencing)

- Randomly fragment multiple copies of the genome into small pieces.
- Sequence in parallel tens to hundreds of millions of short DNA fragments.
- Typical “read” length (what gets sequenced from each fragment): 50 bp - 500 bp.
- Assemble the genome from which the fragments originated by stitching together the short reads.

Can’t read the genome from one end to the other end

We can obtain overlapping “short-reads” (50 bp - 500 bp)
Applications

- Large-scale population genomic whole-genome sequencing
- Exome sequencing
- RNA-seq
- SHAPE-seq
- Cancer genomics
- Ancient DNA analyses (e.g., Neanderthals, Denisovan)
What can we do with whole-genome sequence data

✦ Genome-wide association studies
✦ Study the contribution of each identified locus to a phenotype
✦ Demographic inference (population size changes, migration history, admixture, etc)
✦ Find genetic variants under natural selection
If a reference genome is available

1. Library preparation + sequencing
2. Base-calling
3. Read mapping / alignment
4. variant calling (may require targeted de novo assembly)
5. phasing
6. imputation

If we have a “reference genome,” we can use sequence alignment to infer where each read came from.
# Base-calling in the Illumina Platform

<table>
<thead>
<tr>
<th>Cycle</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
<tr>
<td>2</td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
</tr>
</tbody>
</table>

**Figure:** CCD images from cycles 1 and 2 of Illumina sequencing
Each bright spot corresponds to a cluster of 1000 or so identical molecules, and the goal is to infer the correct sequence (in \{A, C, G, T\}) for each cluster.

Extracting sequence information

Obtaining actual sequences involves two problems

1. **Image analysis**: CCD images $\rightarrow$ Intensities
2. **Base-calling**: Intensities $\rightarrow$ Sequences
Intensities for each cluster

- Obtained by *summing pixel values*
- Possible sources of complication include
  - Misalignment of images (within cycle and between cycles)
  - Overlapping clusters
Sources of complication for base-calling

- **Crosstalk**: response in channel $x$ due to fluorescence of a unit concentration of base $y$.
- **Phasing**: some templates may lag behind in synthesis.
- **Prephasing**: some templates may jump ahead in synthesis.
- **Intensity decay**: some templates stop joining in the reactions.
- **Residual effects**: background accumulation.
- **Stochastic fluctuation** in intensities.

The better we can model the actual stochastic sequencing process, the more accurate basecalls can be.

**Warning**: It is important to note that short-read data are not perfect. There are random and systematic errors.
Short-read data

FASTQ

a`X_\Va\J`KaYJHG~]b\a`BBBBBFFFFBBB
@FC42BF1AAXX:6:1:5:732#0/1
TGATTCTCTCGATATCCAGTCCTTAGTNCATAGN
+
a^_aaaa`aa`_aaa_aaa`__`__`VBBBBBBBB
@FC42BF1AAXX:6:1:5:492#0/1
AACAGTGGGAGGCAGGCAGCAGGATTNCTGAAN
+
ababb_abbbZbabaab`aaTaabbaBBB
@FC42BF1AAXX:6:1:5:480#0/1
ACCTCCTCAGAGTTCTCGAGCTCGAGAAANTCTGGN

--- quality score
--- read ID
--- read (bases)
Read Mapping / Alignment

- MAQ
- BWA
- Bowtie / Bowtie 2
- SOAP / SOAP 2
- Stampy
- Novoalign
- SNAP
- SeqAlto
- + many others
Single Nucleotide Polymorphism (SNP) and Genotype Calling

- On average two human genomes differ at one in a thousand bases.
- **SNP calling**: determine the positions in which polymorphisms exist
- **Genotype calling**: determine actual genetic variants at polymorphic sites
- Single sample calling vs. joint calling of multiple samples

Single Nucleotide Polymorphism (SNP) and Genotype Calling

- Alignment errors can lead to SNP/Genotype calling errors.

Figure from DePristo et al. (2011)
Sample one ball from each urn uniformly at random.
What is the conditional probability $\Pr(\text{ } | \text{ })$?

Sample one dumbbell from the urn uniformly at random.
What is the conditional probability $\Pr(\text{ } | \text{ })$?
Linkage Disequilibrium (LD):
Non-independence of alleles at different loci

Factors influencing LD:

- Biology
  - Mutation
  - Recombination
  - Natural Selection
- Population History
  - Random mating
  - Population expansions, contractions, bottlenecks
  - Subdivision and migration
LD patterns in a 100 kb region around the *Lactase* gene on Chromosome 2

**European Origin**

- CEU

**African**

- YRI

**Han Chinese and Japanese**

- CHB/JPT

Different populations exhibit different levels of LD.

<table>
<thead>
<tr>
<th></th>
<th>YRI</th>
<th>CEU</th>
<th>CHB/JPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of chromosomes</td>
<td>120</td>
<td>120</td>
<td>180</td>
</tr>
<tr>
<td>Number of unique haplotypes</td>
<td>34</td>
<td>18</td>
<td>35</td>
</tr>
<tr>
<td>Haplotype homozygosity</td>
<td>0.05</td>
<td>0.53</td>
<td>0.15</td>
</tr>
</tbody>
</table>

McVean (2007)
Recombination

- It is a **major evolutionary mechanism** responsible for generating genetic variation in a population.
- It **breaks down correlation** between two positions on the same chromosome. Because of recombination, different positions on the same chromosome may have different genealogies.
- It has far-reaching **implications for many computational problems**, including
  - disease-association mapping,
  - missing data imputation,
  - phasing genotype data into haplotype data,
  - inferring local ancestry of admixed populations,
  - detecting signatures of natural selection, and so on.
Figure from Sasaki, Lange, Keeney, Nat Rev Mol Cell Biol, 2010
Recombination backwards in time

- **Ancestral material**
- **Nonancestral material**

Crossover

Gene conversion
Crossover recombination and the coalescent

- As we scan along the sequence, recombination events cause different sites to have different genealogical histories.
- This has a profound effect on patterns of linkage disequilibrium.
Phasing (a.k.a. Haplotyping)

- Family information can help determine phase at many markers
- If family information is not available:
  - Utilize LD information in a statistical framework (Phase).
  - Assume that population consists of small number of distinct haplotypes (Clark’s algorithm).
  - Assumed that haplotypes are “derived” on parsimonious genealogies (e.g., Perfect Phylogeny Haplotyping).

<table>
<thead>
<tr>
<th>Unphased genotypes</th>
<th>Possible phasing A</th>
<th>Possible phasing B</th>
<th>Possible phasing C</th>
<th>Possible phasing D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/C</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>G/T</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td>A/T</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
</tr>
</tbody>
</table>

| Family haplotype frequency | 55% | 0% | 15% | 5% | 2% | 3% | 0% | 20% |
| Population frequency of unordered haplotype pair | 0% | 2 × (15% × 5%) = 1.5% | 2 × (2% × 3%) = 0.12% | 0% |
| Posterior probability of unordered haplotype pair | 0% | 1.5% / (1.5% + 0.12%) = 93% | 0.12% / (1.5% + 0.12%) = 7% | 0% |

Figure from Browning and Browning, Nature Reviews Genetics, 2011.
Imputing Missing Data

- Sequences from NGS platforms tend to contain a lot of missing data.
- **Imputation**: Use genotypes at a few markers to infer genotypes at other unobserved markers.
- Closely related individuals
  - Long segments of highly similar genetic content
- Distantly related individuals
  - Shorter segments of highly similar genetic content
- Missing genotypes are imputed with **uncertainty** and a probability distribution over the possible genotypes is produced. It is necessary to take account of this uncertainty in any downstream analysis of the imputed data.
Imputing Missing Data on a Pedigree
Box 1 | How genotype imputation works

b Testing association at typed SNPs may not lead to a clear signal

c Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel

d Reference set of haplotypes, for example, HapMap

f Testing association at imputed SNPs may boost the signal

e The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

Figure from Marchini and Howie, Nature Reviews Genetics, 2010
Structural Variation

Application of new technologies to the description of genomic variation suggests the existence of abundant structural variation (SV) in many organisms.

Figure from Alkan et al. (2011)
Structural Variation

- Recently it has emerged that such variations may also contribute substantially to phenotypic differences between individuals. (e.g., obesity: Willer et al, Nature Genetics, 2009)
- For example, it has been shown that SVs can disrupt gene expression and alter gene dosage.
- Further, it can also cause disease via micro deletion and duplication, or be a risk factor in complex trait diseases.
- Cancer cells are prone to undergo structural rearrangements.
- The ability to sequence a large sample of genomes with high-throughput sequencing technologies has started to offer a comprehensive view of SVs.
- Detecting SVs is harder than detecting SNPs.
- May require doing targeted de novo assembly.
**Presentation Sign-up**

1. 8/28  Introduction to the course
2. 9/4   Read Mapping / Alignment
3. 9/11  SNP and Genotype Calling 1
4. 9/18  Guest Lecture
5. 9/25  SNP and Genotype Calling 2
6. 10/2  Phasing 1
7. 10/9  Phasing 2
8. 10/16 Phasing 3
9. 10/23 Imputation 1
10. 10/30 Imputation 2
11. 11/6  Imputation 3
12. 11/13 Indel Detection
13. 11/20 Structural Variant Detection 1
14. 11/27 Structural Variant Detection 2

*If you are registered, you should signup for a module*