Computational Genetics
Spring 2011
Final Projects List

Eleazar Eskin

University of California, Los Angeles

Final Projects

As of Lecture 5.

April 11th, 2011

Final Project 1: Relatedness Estimator

- Given the genotypes of several individuals, how are all of them related?
- Parents transmit 1 chromosome to each child. Siblings share approximately 50% of their DNA. 1st cousins share about 25% of their DNA.
- A challenge is that some individuals may share DNA by chance.

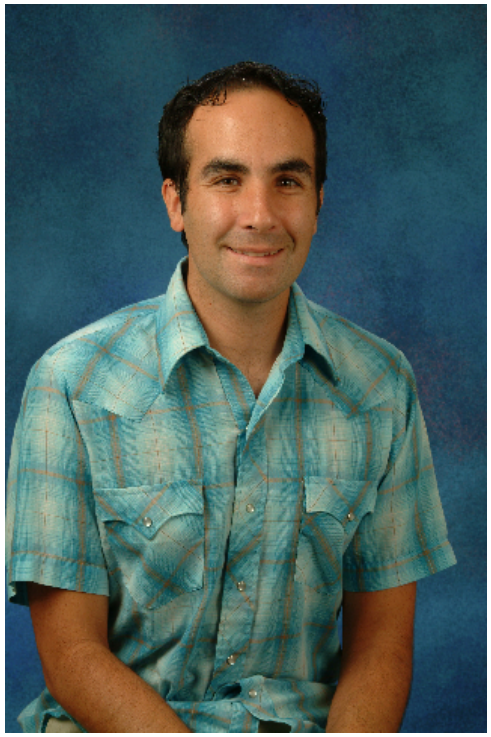
Final Project 1: Relatedness Estimator

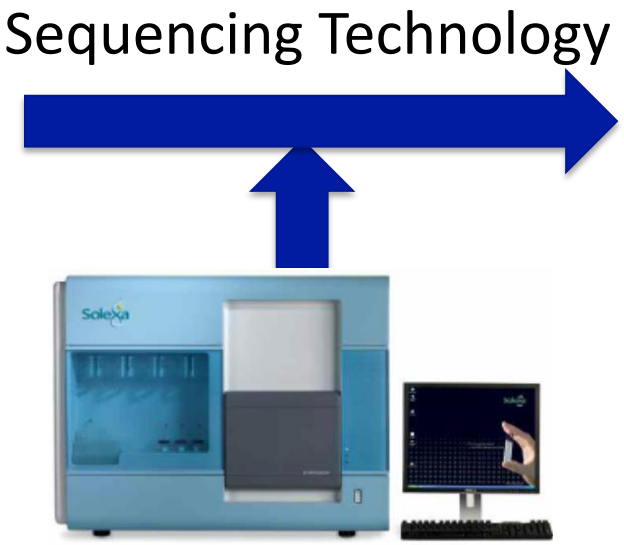
- Consider a SNP with minor allele frequency of .1.
- What is the probability of having the allele on both chromosomes?
- If your brother has the allele on both chromosomes, what is your probability of having the allele on both your chromosomes?
  - **Hint: It is much higher.**
- What about minor allele frequency of .01 or .4?
- What allele frequencies are most informative?
  - **Tradeoff between occurrence of alleles and information in sibling.**

Final Project 1: Relatedness Estimator

- Easy Project: Construct a method for determining whether 2 individuals are siblings.
- Medium Project: Construct a method for estimating how related any 2 individuals are. Take into account LD.
- Hard Project: Measure risk to related siblings given genotypes (ethical issue). Handle finite sample size of HapMap. Reconstruct family histories.

Sequencing Technology





Illumina / Solexa
Genetic Analyzer 1G
1000 Mb/run, 35bp reads

AGAGCAGTCGAC
AGGTATAGTCTA
CATGAGATCGAC
ATGAGATCGGTA
GAGCCGTGAGAT
CGACATGATAGC
CAGAGCAGTCGA
CAGGTATAGTCT
ACATGAGATCGA
CATGAGATCGGT
AGAGCCGTGAGA
TCGACATGATAG
CCAGAGCAGTCG
ACAGGTATAGTC
TACATGAGATCG
ACATGAGATCGG
TAGAGCCGTGAG
ATCGACATGATA
GCCAGAGCAGTC
GACAGGTATAGT
CTACATGAGATC
GACATGAGATCG
GTAGAGCCGTGA
GATCGACATGAT
AGCCAGAGCAGT
CGACAGGTATAG

- Next generation sequencing.
  - **Cheap sequencing.**
  - **“Short Reads”**

Short Read Sequencing

Full DNA Sequence

AGAGCAGTCGAC
AGGTATAGTCTA
CATGAGATCGAC
ATGAGATCGGTA
GAGCCGTGAGAT
CGACATGATAGC
CAGAGCAGTCGA
CAGGTATAGTCT
ACATGAGATCGA
CATGAGATCGGT
AGAGCCGTGAGA
TCGACATGATAG
CCAGAGCAGTCG
ACAGGTATAGTC
TACATGAGATCG
ACATGAGATCGG
TAGAGCCGTGAG
ATCGACATGATA
GCCAGAGCAGTC
GACAGGTATAGT
CTACATGAGATC
GACATGAGATCG
GTAGAGCCGTGA

- Short read sequencers generate random short substrings from the DNA sequence of a certain length.

ATGAGATCGGTAGAGCCGTGAGAT
GAGCAGTCGACAGGTATAGTCTAC
AGAGCAGTCGACAGGTATAGTCTA
TGAGATCGACATGATAGCCAGAGC
TAGCCAGAGCAGTCGACAGGTATA
GATAGCCAGAGCAGTCGACAGGTA
GAGATCGACATGATAGCCAGAGCA
GCAGTCGACAGGTATAGTCTACAT
AGCAGTCGACAGGTATAGTCTACA
TCGACATGAGATCGGTAGAGCCGT
CAGTCGACAGGTATAGTCTACATG
GAGATCGACATGATAGCCAGAGCA
GTAGAGCCGTGAGATCGACATGAT

Short Reads Difficulties

```
ATGAGATCGGTAGAGCCGTGAGAT
GAGCAGTCGACAGGTATAGTCTAC
AGAGCAGTCGACAGGTATAGTCTA
TGAGATCGACATGATAGCCAGAGC
TAGCCAGAGCAGTCGACAGGTATA
GATAGCCAGAGCAGTCGACAGGTA
GAGATCGACATGATAGCCAGAGCA
GCAGTCGACAGGTATAGTCTACAT
AGCAGTCGACAGGTATAGTCTACA
TCGACATGAGATCGGTAGAGCCGT
CAGTCGACAGGTATAGTCTACATG
GAGATCGACATGATAGCCAGAGCA
GTAGAGCCGTGAGATCGACATGAT
```

- We don't know where each read comes from!
- Can't identify where the mutations are!
- What do we do?

Key Idea: "Re"-Sequencing

We know that my genome is very close to the Human genome.

My Genome:

TACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATC

A Sequence Read:

TCGACATGAGATCGGTAGAGCCGT

The Human Genome:

Recovered Sequence:

TACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATC

"Re"-Sequencing Challenges (Why do we need Computer Science?)

- Sequences are long!
  - **Human Genome is 3,000,000,000 long.**
- Sequencers generate many reads!
  - **A single run generates over 300,000,000 reads.**
- We need efficient algorithms to "map" each read to its location in the genome.

There are other challenges which we are not mentioning.

Project 2: Mapping of reads

- Short reads need to be mapped to the genome for resequencing.
- Computer science problem:
  - **Given a string of length L=30, find where it matches a substring within D=2 mismatches in a length N=3,000,000,000 sequence.**
- Evaluate the quality based on:
  - **The speed of the mapping algorithm.**
  - **The memory use of the mapping algorithm.**
  - **The accuracy of the mapping algorithm (for approximate approaches).**

Project 2: Mapping of reads

- Easy: Build a small scale mapper that can map strings of length 30 to sequences of length 1,000,000.
- Medium: Build a mapper that can scale to sequences of length 3,000,000,000. It can be slow.
- Very Hard: Build a fast mapper.

Final Project 3: Ancestry Mapping

- For some populations, e.g. African Americans, each individual's genome comes from multiple populations.
- Goal of ancestry mapping is to identify which region originates from which population.
- Complications include:
  - **Correlation between SNPs**
  - **Similarity between ancestral populations**
  - **Unknown ancestral populations (Nature Americans)**

Final Project 3: Ancestry Mapping

- Four Versions of the Problem:
  - **Local Ancestry vs Global Ancestry**
  - **Known Populations vs Unknown Populations**
- Correlation Between SNPs
  - **Haplotype Structure of Populations**
  - **SNPs are not independent.**
- Evaluating Through Simulations
  - **We can simulate data using the HapMap**
  - **We can run our method over the simulated data to measure method accuracy.**

Final Project 3: Ancestry Mapping

- Easy: Global Ancestry Mapping with Known Populations
- Medium: Local Ancestry Mapping with Known Populations
- Hard: Global Ancestry Mapping with Unknown Populations
- Very Hard: Local Ancestry Mapping with Unknown Populations

Final Project 4: Disease Prediction

- Given the genome of an individual, can we predict how likely they are to get a disease?
- Key Ideas:
  - **If we know which SNPs cause disease, we can use those SNPs to predict disease.**
  - **If we know which individuals have a disease, we can see how related an individual is to individuals that have the disease (Family History).**

Final Project 4: Disease Prediction

- Easy: Assume a set of known disease SNPs. Create a method for predicting disease risk and evaluate how effective the method is depending on the relative risk and number of disease causing SNPs.
- Hard: Develop a disease prediction method given a set of genomes with known disease status, but unknown which SNPs cause the disease.

Project 5: SNP Detection from Sequencing

- Given mapped reads, we can identify SNPs between the reference and the sequenced genome.
- However, reads have errors.
- Not all mismatches are SNPs since some are errors.
- However, SNPs will occur in many reads, while errors will be in only one read.

More problems: Sequencing Errors

- Each sequence read can have some random errors.

My Genome:
TACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC

A Sequence Read:
TCGACATGAGATCGGTAGA**A**CCGT

The Human Genome:
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
TCGACATGAGATCGGTAGAACCGT

Recovered Sequence:
TACATGAGATC**G**ACATGAGATC**G**GTAGA**A**C**C**GTGAGATC

Sequencing Errors: Solution

- Collect redundant data.

My Genome:

TACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC

Sequence Reads:

TCGACATGAGATCGGTAGA**A**CCGT
GACA**A**GAGATCGGTAGAGCCGTGA
TGAGATCGG**T**AGAGCCGTGAGATC

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
TC**G**ACATGAGATC**G**GTAGA**A**C**C**GT
GACA**A**GAGATC**G**GTAGAGC**C**GTGA
TGAGATC**G**G**T**AGAGC**C**GTGAGATC

Recovered Sequence:

TACATGAGATC**G**ACATGAGATC**G**GTAGAAC**C**GTGAGATC

How much coverage do we need?

- If error rate is *e*, and we are going to predict the consensus sequence, what is the error rate if the coverage is 3.
- We will make a prediction with an error if two out of our three reads have an error in the same place.

$$e^3 + \binom{3}{2}(1-e)e^2$$

- This is approximately $3e^2$.

Diploid Sequencing

- Humans have 2 chromosomes.
- Each chromosome may have a different SNP.
- Some reads come from 1 chromosome, some come from other chromsome.
- Why does consensus method not work?
- How do we address this problem?

Project 5: SNP Detection from Sequencing

- Easy: Write a simple SNP caller from sequence data.
- Medium: Estimate how much coverage you need to get accurate SNP coverage.
- Hard: SNP Detection in copy number regions.

Final Project 6: Meta-Analysis

- If two different case/control studies have M and N individuals, intuitively, we can put the studies together to get a M+N individual study.
- There are some issues such as:
  - **Studies can be from different populations.**
  - **Studies can collect different SNPs.**
  - **Study phenotypes can be different.**

Final Project 6: Meta-Analysis

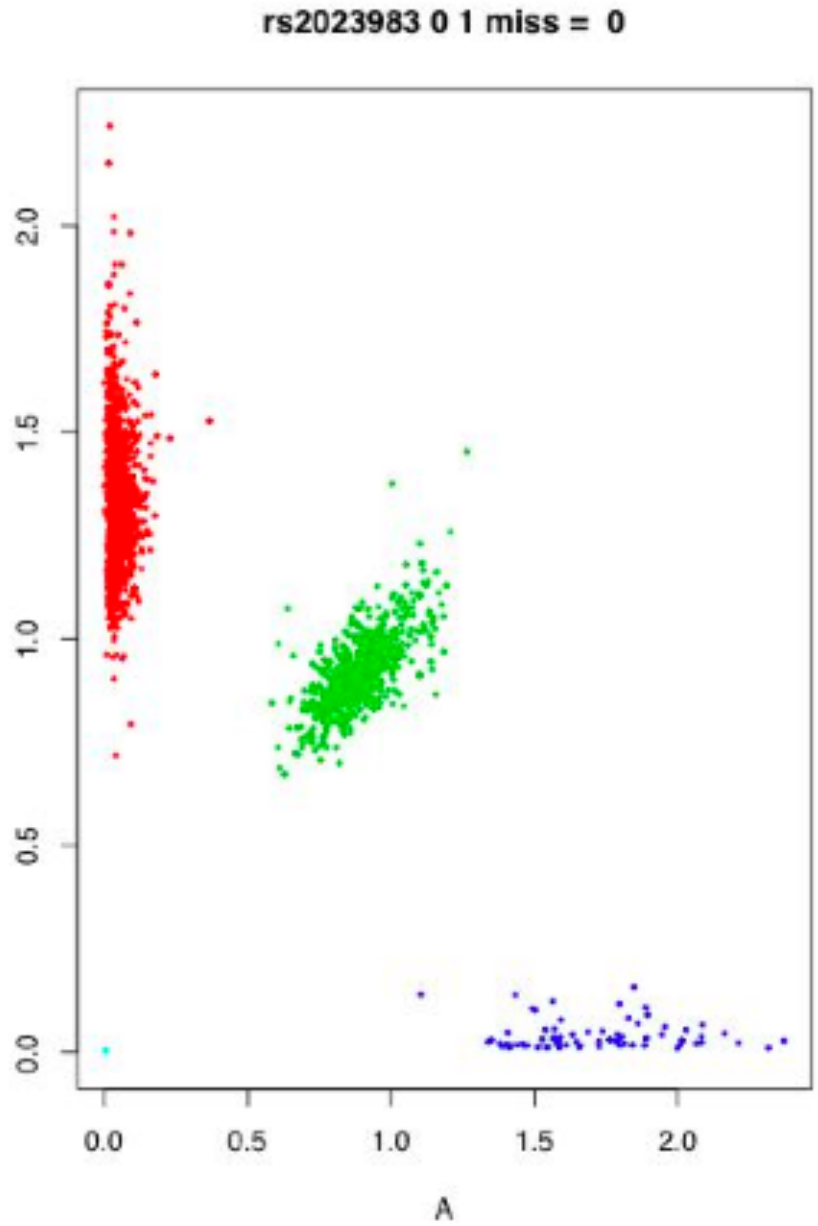
- In meta-analysis, due to differences in the studies, want to combine the statistics from the studies.
- Questions include:
  - **How do we combine the statistics of 2 studies?**
  - **What do we do if the markers are different?**
  - **What kinds of effects can create false positives?**

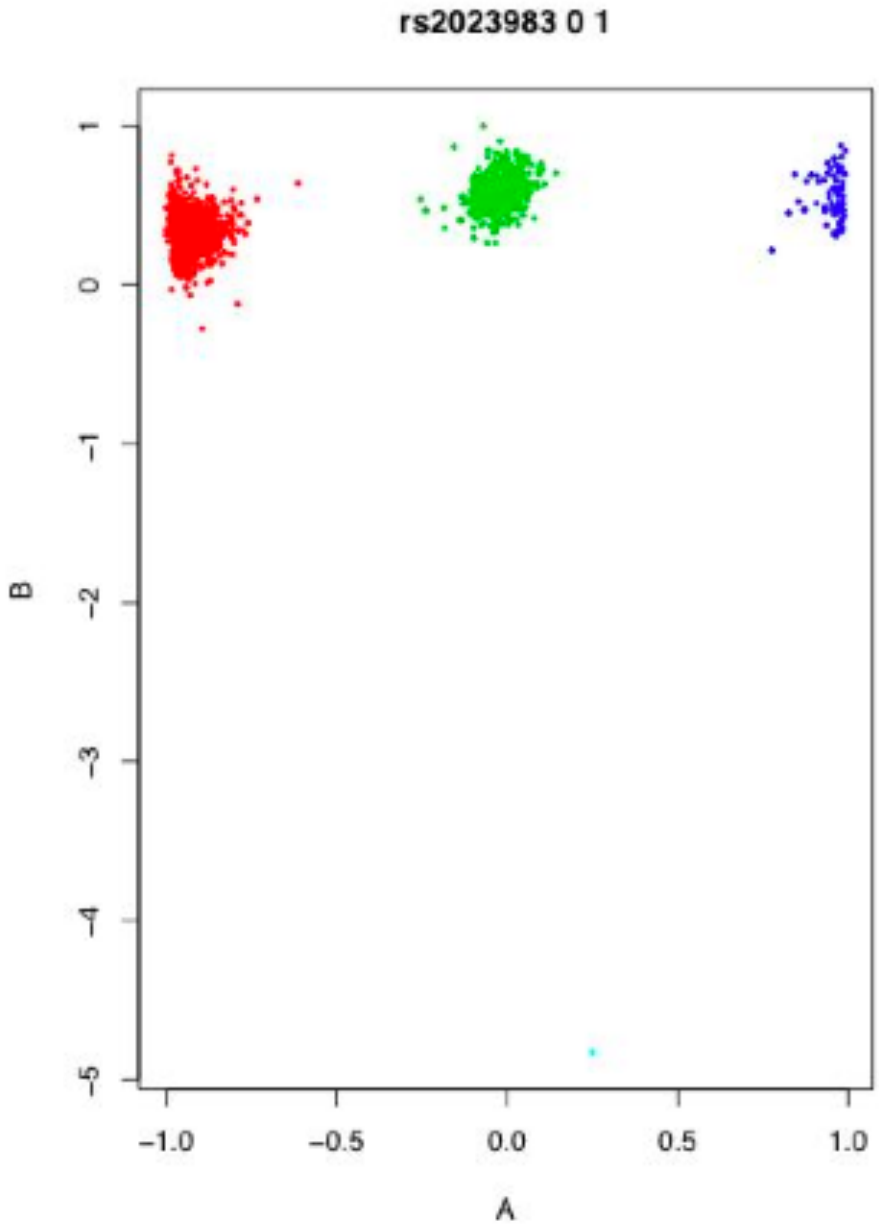
Final Project 6: Meta-Analysis

- Easy: Apply simple meta-analysis statistic to simulated data. How does the power compare to grouping individuals together.
- Medium: Develop statistics for combining studies on different marker sets.
- Very Hard: Develop statistics that are optimal with respect to power. Develop statistics for combining studies with cases and controls on different marker sets.

Project 7: Genotype Calling

- Genotype are obtained using a microarray technology.


rs2023983 0 1 miss = 0
0.0
0.5
1.0
1.5
2.0
A


rs2023983 0 1
B
1
0
−1
−2
−3
−4
−5
−1.0
−0.5
0.0
0.5
1.0
A

Project 7: Genotype Calling

- Easy: Make a simple genotype caller based on ratios of the probes.
- Medium: Identify clusters in the genotype plots and use the distance to the center of the cluster to make predictions. How does this compare to the other method?
- Hard: Make improvements to genotype calling including: Identify copy number variation from genotype calling. Identify outliers. Normalize across arrays.

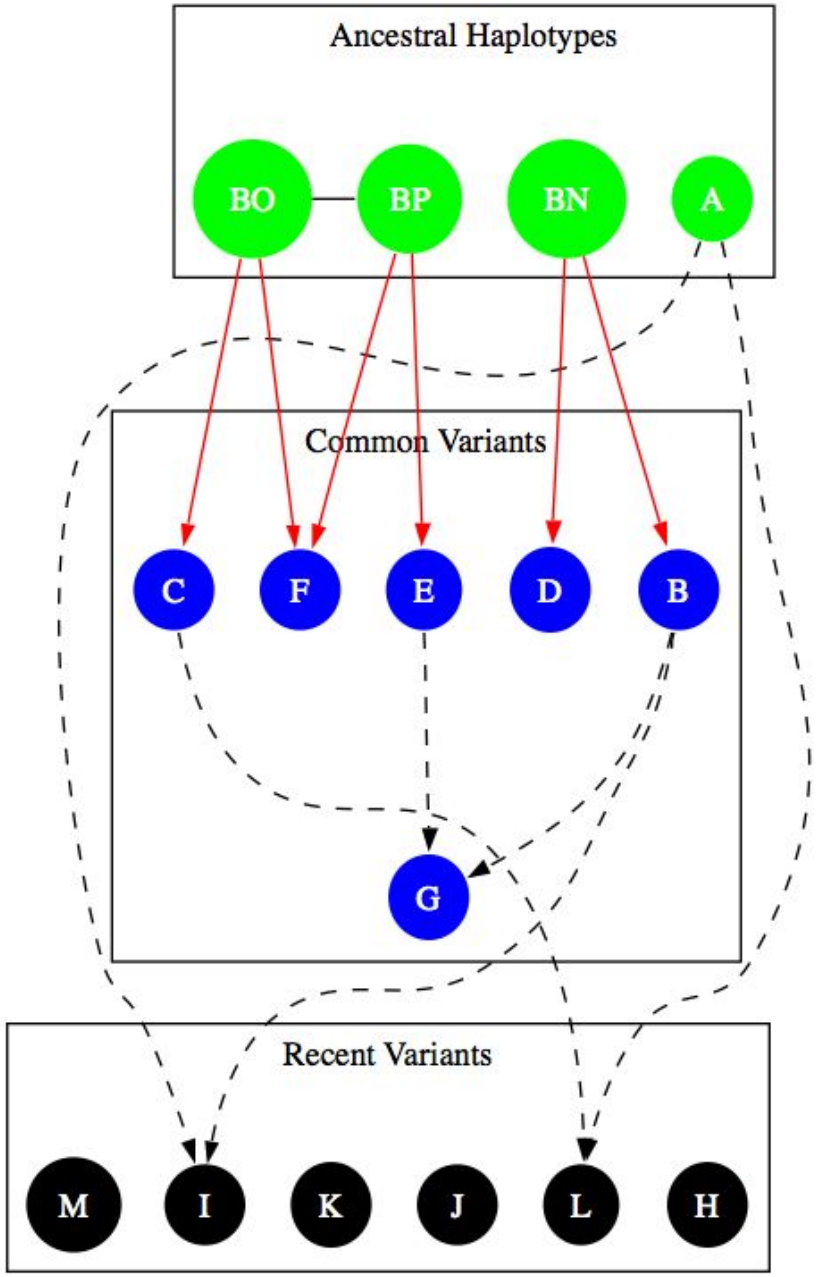
Project 8: Haplotype Phylogeny

- Human history is very unique.
- We left Africa and spread around the world very quickly.
- Most variation predates us leaving Africa
- Variation after that time is very rare in the population.

Ancestral Haplotypes		Num
A	CTTAAAGTTA	297
BN	CCTAAAAAT?	(387)
BO	C?AGGCATCT	(298)
BP	?CAGGCATCT	(215)

Common Variants		Num
B	CCTAAAAATA	227
C	CTAGGCATCT	217
D	CCTAAAAATT	140
E	ACAGGCATCT	135
F	CCAGGCATCT	72
G	ACTAAAAATA	67

Recent Variants		Num
H	ATAGGAAATT	51
I	CCTAAAGTTA	26
J	ACTAGAGTTA	26
K	ACAGGAAACT	16
L	CTTGGCATCT	15
M	CCTAGAGTTT	11



Project 8: Haplotype Phylogeny

- Easy: Implement a tool that takes in genotypes, applies a haplotype program and then visualizes the results using a haplotype phylogeny.
- Medium: Consider long haplotypes and recombination

“Re”-Sequencing: Insertions

My Genome:

TACATGAGATCCACAT**A**GAGATCTGTAGAGCTGTGAGATC

A Sequence Read:

CCACATAGAGATCTGTAGAGCTGT

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

CCACATAGAGATCTGTAGAGCTGT

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

CCACATAGAGATCTGTAGAGCTGT

How do we deal with this case?

“Re”-Sequencing: Insertions

My Genome:

TACATGAGATCCACATAGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:

CCACATAGAGATCTGTAGAGCTGT

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
CCACATAGAGATCTGTAGAGCTGT

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
CCACATAGAGATCTGTAGAGCTGT

Solution: Add Insertion to the Human Genome

TACATGAGATCCACAT-GAGATCTGTAGAGCTGTGAGATC
CCACATAGAGATCTGTAGAGCTGT

Difficulties for handling insertions

- Requires “Alignment” of reads to genome.
- Much more computational intensive

- Need to change assumptions for “sequence uniqueness” to use edit distance.

Project 9: Read Mapping with Insertions

- Medium: Develop a mapper than can map reads with up to 2 insertions in 1,000,000 length sequence.
- Hard: Handle 3,000,000,000 length sequences.
- Very Hard: Accurately identify insertions.

Project 9: Copy Number Variation

- Copy number variation between sequenced genome and target causes strange patterns of reads.
- If a sequenced genome has 2 copies of a region, both will generate reads and the coverage on the reference will appear 2x the coverage of the rest of the genome.
- If a sequence is deleted, there will be 0 coverage of a region.

Project 9: Copy Number Variation

- Easy: Find regions which have high copy number.
- Medium: Filter out repeated regions. Estimate the copy number.
- Very Hard: Find the boundaries of copy number variation.

Paired End Read

My Genome:

```
TACATGAGATCCACATGAGTGTAGAGCTGTGAGATC
```

A Sequence Read:

```
CCACATA------AGAGCTGT
```

The Human Genome:

```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
          CCACATA---------AGAGCTGT
```

What does the longer gap in the mapping mean?

Project 10: Inversions

- Regions of the chromosome may be inverted between individuals.
- “Paired end” sequencing gives information on structural variation.
  - **Provides a pair of reads a fixed distance in the sequenced genome.**
  - **Differences in distance between mapped reads on the reference suggests some differences.**
- Inversion leaves a distinct pattern of paired reads if the pair spans the inversion breakpoint.

Project 10: Inversions

- Medium: Identify potential inversions in mouse data.
- Hard: Design inversion detection method that allows for mapping problems.
- Very Hard: Recover inversion breakpoints.

Project 11: Haplotype Assembly

- Humans have 2 haplotypes while sequencing reads only contain information on one haplotype.
- Reads from the same region contain information on both haplotypes.
- Each read is coming from one haplotype, but we don't know which.
- By tiling reads together, we can reassemble into haplotypes.

Project 11: Haplotype Assembly

- Two related papers by Bafna group.
- Easy: Build a greedy algorithm for haplotype phasing from sequence reads.
- Very Hard: Find an optimal algorithm.

Project 12: Sequence Insertion Assembly

- Insertion of novel sequence into genomes is a type of structural variation.
- Reads from the sequence will not map to anywhere in the reference. However, one end of a paired end read mapping to the reference can give a clue to where the insertion is located.
- Overlapping unmapped reads can identify the sequence.

Project 12: Sequence Insertion Assembly

- Medium: Generate reads which simulate an insertion to the reference and identify the location of the insertion.
- Hard: Assume that the reference is non-repetitive, there are no sequencing errors and all non-insertion reads map somewhere in the reference. Design an algorithm that assembles the inserted sequence.
- Very Hard: Drop the assumptions above design an algorithm for insertion assembly.

Project 13: Multiple Phenotypes

- We often collect multiple phenotypes for each individual.
- If we are interested in a disease, we often collect “intermediate phenotypes” which affects the disease
- For example, in heart disease, an intermediate phenotype is the cholesterol level.

Project 13: Multiple Phenotypes

- Medium: Develop a technique for association for intermediate phenotypes. Apply to simulated data and measure power with and without intermediate phenotypes.
- Hard: Define a model for an intermediate phenotype. Identify under what cases can using an intermediate phenotype can increase power of an association study. What assumptions are necessary?

Project 14: Sequence Assembly and Reassembly

- Most sequencing applications assume that there is a reference sequence and they map reads to the reference.
- Alternate strategy is to use overlapping reads to create longer sequences until you reach the full genome.
  - **Difficult if there are repeats!**
- “Assembly” assumes that there is no reference.
- “Reassembly” assumes that there is a reference.

Project 14: Sequence Assembly and Reassembly

- Easy: Build an assembler assuming no repeated sequence
- Medium: Build an assembler for small sequences assuming some repeated sequence and use graph algorithms.
- Hard: Build a real genome assembler.

Project 15: Virus Assembly

- Sequencing can be applied to a sample of viruses.
  - **In the sample, there are many variants of the virus and each variant has a certain frequency and set of mutations.**
  - **Each read only covers a fraction of the virus.**
  - **By making some assumptions, we can predict the variants present in the sample and their frequency.**
- Key assumptions:
  - **Reads are uniformly distributed from the sample.**
  - **Small number of different variants in the sample.**

Project 15: Virus Assembly

- Easy: Assume that you know the different strains of the virus present in the sample. Simulate reads from this sample and estimate the frequency of each strain.
- Hard: Assume that you do not know the strains and predict the variants present in the sample and estimate their frequency.

Project 16: Meta Genomics

- There are more bacteria cells in us then our own cells.
- Meta-Genomics Sequencing of Bacteria Samples from people
- We want to know what types of bacterias and how common they are in people.

- Problem: Sequence reads come from one bacteria and we don't know which bacterias are in the samples.

Project 16: Meta Genomics

- Two versions of Meta-Genomics problems
- Easier: Given a set of bacteria in a sample and given reads from each individual, figure out how common each bacteria occurs in the sample.
- Harder: Given a set of reads from a sample, and a large set of possible bacterias in the sample, figure out which are there and how common they are.
- Very Hard: Given a set of reads from many samples, figure out which unknown bacterias are in the sample and how common they are.

Project 17: Isoform Assembly

- Popular application of sequencing is RNA-Seq, where sequencing is applied to RNA samples to obtain activity levels of genes.
- Each RNA is a combination of “exons” in a gene. Each such combination is called an isoform. The isoforms present in the sample are unknown.
- Each read only spans a portion of the isoform.
- Goal of Isoform assembly is to predict the isoforms and estimate their frequency.

Project 17: Isoform Assembly

- Easy: Assume you know the isoforms which are present in the sample. Simulate reads and estimate the frequency of each isoform.
- Medium: Assume you have paired end reads and predict isoforms and estimate their frequency.
- Hard: Use the reference genome to help predict isoforms from reads.

Project 18: Lethal Interactions

- Mutations often have big effects when they interact.
- Biological systems usually have backup systems (robustness).
- A lethal interaction is one where the presence of 2 mutations has a big effect. Typically, this is when a mutation breaks a process and its backup system.

Project 18: Lethal Interactions

- Medium: Define a model for lethal interactions where the relative risk is only greater than 1 for individuals that have 2 SNPs. What is the power to detect such SNPs?
- Hard: Define a method to discover lethal interactions by considering all pairs of mutations.

Project 19: Multiplexing Sequence Pools

- Sequencing a single sample has 2 costs
  1. **Sample Preparation**
  2. **Sequencing**
- Sequencing costs are decreasing, but sample preparation costs are constant.

- Idea: Sequence multiple samples with a single sample preparation step.

Project 19: Multiplexing Sequencing Pools

- Solution: Construct “pools” where samples are mixed together and entire sample is sequenced.

Project 19: Multiplexing Sequencing Pools

- Easy: Make scheme for saving factor of 5 assuming no errors and perfect sequencing for SNP occurring once in samples.
- Medium: Make a scheme for saving factor of 5 for SNPs that occur at frequency 2% assuming no errors and perfect sequencing.
- Hard: Make a scheme for saving factor of 10 allowing for errors.
- Very Hard: Make a scheme for saving factor of 10 for common and rare variants.