Whole Genome Sequencing

Biostatistics 666
Genomewide Association Studies

• Survey 500,000 SNPs in a large sample

• An effective way to skim the genome and ...
• ... find common variants associated with a trait of interest

• Rapid increase in number of known complex disease loci
  – For example, ~50 genes now identified for type 2 diabetes.

• Techniques for genetic analysis are changing rapidly
  – What are some of the potential benefits and challenges for replacing genotyping with sequencing in complex trait studies?
Questions that Might Be Answered With Complete Sequence Data...

- What is the contribution of each identified locus to a trait?
  - Likely that multiple variants, common and rare, will contribute

- What is the mechanism? What happens when we knockout a gene?
  - Most often, the causal variant will not have been examined directly
  - Rare coding variants will provide important insights into mechanisms

- What is the contribution of structural variation to disease?
  - These are hard to interrogate using current genotyping arrays.

- Are there additional susceptibility loci to be found?
  - Only subset of functional elements include common variants ...
  - Rare variants are more numerous and thus will point to additional loci
What Is the Total Contribution of Each Locus?

Evidence that
Multiple Variants Will be Important
Evidence for Multiple Variants Per Locus
Example from Lipid Biology

For several loci, there is clear evidence for independently associated common variants – even among markers typed in GWAS.

Including these in the analysis increases variance explained by ~10%.

Evidence for Multiple Variants Per Locus
Example from Lipid Biology

Private mutations in PCSK9 change LDL by >100 mg/dl
(Abifadel et al, 2003)

Rare variants (MAF 1%) in PCSK9 can change LDL by ~16 mg/dl
(Cohen et al, 2005)

Common variants (MAF 20%) in PCSK9 change LDL by ~3 mg/dl
(Willer et al, 2008)

What is The Contribution of Structural Variants?

Current Arrays Interrogate 1,000,000s of SNPs, but 100s of Structural Variants
Evidence that Copy Number Variants Important
Example from Genetics of Obesity

Seven of eight confirmed BMI loci show strongest expression in the brain...

Evidence that Copy Number Variants Important

Example from Genetics of Obesity

Willer et al, Nature Genetics, 2009
Evidence that Copy Number Variants Important
Example from Genetics of Obesity

Willer et al, Nature Genetics, 2009
Associated Haplotype Carries Deletion

What is the Mechanism?
What Can We Learn From Rare Knockouts?

What We’d Like to Know
Recent Example from John Todd’s Group
Can Rare Variants Replace Model Systems? Example from Type 1 Diabetes

- Nejentsev, Walker, Riches, Egholm, Todd (2009)
  IFIH1, gene implicated in anti-viral responses, protects against T1D
  *Science* **324**:387-389

- Common variants in IFIH1 previously associated with type 1 diabetes

- Sequenced IFIH1 in ~480 cases and ~480 controls
- Followed-up of identified variants in >30,000 individuals

- Identified 4 variants associated with type 1 diabetes including:
  - 1 nonsense variant associated with reduced risk
  - 2 variants in conserved splice donor sites associated with reduced risk
  - Result suggests disabling the gene protects against type 1 diabetes
HDL-C Associated Locus

- GWAS allele with 40% frequency associated with ±1 mg/dl in HDL-C

- **GALNT2** expression in mouse liver (Edmonson, Kathiresan, Rader)
  - Overexpression of **GALNT2** or **Galnt2** decreases HDL-C ~20%
  - Knockdown of **Galnt2** increases HDL-C by ~30%
The Challenge

- Whole genome sequence data will greatly increase our understanding of complex traits

- Although a handful of genomes have been sequenced, this remains a relatively expensive enterprise

- Dissecting complex traits will require whole genome sequencing of 1,000s of individuals

- How to sequence 1,000s of individuals cost-effectively?
Next Generation Sequencing
Massive Throughput Sequencing

• Tools to generate sequence data evolving rapidly

• Commercial platforms produce gigabases of sequence rapidly and inexpensively
  – Illumina is currently the dominant technology (by far)

• Sequence data consist of millions or billions of short sequence reads with moderate accuracy
  – 0.5 – 1.0% error rates per base may be typical
21st Century Sequencing Costs

http://genome.gov/sequencingcostsdata
Shotgun Sequence Reads

- Typical short read might be <50-150 bp long and not very informative on its own
- Reads must be arranged (aligned) relative to each other to reconstruct longer sequences
Base Qualities

Each base is typically associated with a quality value

Measured on a “Phred” scale, which was introduced by Phil Green for his Phred sequence analysis tool

\[
BQ = -\log_{10}(\epsilon), \text{ where } \epsilon \text{ is the probability of an error}
\]
Read Alignment

The first step in analysis of human short read data is to align each read to genome, typically using a hash table based indexing procedure.

This process can now handle tens of millions of reads per hour ...

Analyzing these data without a reference human genome would require much longer reads or result in very fragmented assemblies.
Read Alignment – Food for Thought

• Typically, all the words present in the genome are indexed to facilitate read mapping ...
  – What are the benefits of using short words?
  – What are the benefits of using long words?

• How matches do you expect, on average, for a 10-base word?
  – Do you expect large deviations from this average?
Mapping Quality

• Measures the confidence in an alignment, which depends on:
  – Size and repeat structure of the genome
  – Sequence content and quality of the read
  – Number of alternate alignments with few mismatches

• The mapping quality is usually also measured on a “Phred” scale

• Idea introduced by Li, Ruan and Durbin (2008) Genome Research 18:1851-1858
Mapping Quality Definition

• Given a particular alignment \( A \), we can calculate

\[
P(S|A, Q) = \prod_{i} P(S_i|A, Q)
\]

\[
= \prod_{i} \left( \frac{1}{3} 10^{-Q_i/10} \right)^{i(S_i \text{ mismatch}|A)} \{1 - 10^{-Q_i/10}\}^{i(S_i \text{ match}|A)}
\]

• Then, the mapping quality is:

\[
MQ(S|A_{\text{best}}, Q) = \frac{P(S|A_{\text{best}}, Q)}{\sum_i P(S|A_i, Q)}
\]

• In practice, summing over all possible alignments is too costly and this quantity is approximated (for example, by summing over the most likely alignments).
Refinements to Mapping Quality

• In their simplest form, mapping qualities apply to the entire read

• However, in gapped alignments, uncertainty in alignment can differ for different portions of the read
  – For example, it has been noted that many wrong variant calls are supported by bases near the edges of a read

• Per base alignment qualities were introduced to summarize local uncertainty in the alignment
Per Base Alignment Qualities

Short Read

5’-AGCTGATAGCTAGCTGATGAGCCCGATC-3’

Reference Genome
Per Base Alignment Qualities

Should we insert a gap?

Short Read

GATAGCTAGCTAGCTGATGAGCCG

Reference Genome

5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGGATC-3’
Per Base Alignment Qualities

Compensate for Alignment Uncertainty With Lower Base Quality

Short Read

GATAGCTAGCTAGCTAGCTGATGAGCCG

5’-AGCTGATAGCTAGCTAGCTAGCTGATGAGCCCCGATC-3’

Reference Genome

Heng Li
Calling Consensus Genotype - Details

• Each aligned read provides a small amount of evidence about the underlying genotype
  – Read may be consistent with a particular genotype …
  – Read may be less consistent with other genotypes …
  – A single read is never definitive

• This evidence is cumulated gradually, until we reach a point where the genotype can be called confidently

• Let’s outline a simple approach …
Shotgun Sequence Data

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTGAGCCCAGA

Predicted Genotype

A/C
Shotgun Sequence Data

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

Sequence Reads

P(reads|A/A, read mapped)= 1.0

P(reads|A/C, read mapped)= 1.0

P(reads|C/C, read mapped)= 1.0

Possible Genotypes
Shotgun Sequence Data

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

GCTAGCTGATAGCTAGC

TAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Possible Genotypes

P(reads|A/A, read mapped) = P(C observed|A/A, read mapped)

P(reads|A/C, read mapped) = P(C observed|A/C, read mapped)

P(reads|C/C, read mapped) = P(C observed|C/C, read mapped)
Shotgun Sequence Data

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTGCTCGACG-3'

Possible Genotypes

$P(\text{reads} | \text{A/A, read mapped}) = 0.01$

$P(\text{reads} | \text{A/C, read mapped}) = 0.50$

$P(\text{reads} | \text{C/C, read mapped}) = 0.99$
Shotgun Sequence Data

AGCTGATAGCTAG\textcolor{red}{C}TAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAG\textcolor{red}{C}TAGCTGATGAGCCCGA

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Sequence Reads

Reference Genome

\textcolor{red}{C}TAGCTGATGAGCCCGA

Possible Genotypes

P(\text{reads} | A/A, \text{read mapped}) = 0.0001

P(\text{reads} | A/C, \text{read mapped}) = 0.25

P(\text{reads} | C/C, \text{read mapped}) = 0.98
Shotgun Sequence Data

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Refence Genome

AGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTTCGATGCTAGCTGATAGCTAGCTAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTGACG-3'

P(reads | A/A, read mapped) = 0.000001

P(reads | A/C, read mapped) = 0.125

P(reads | C/C, read mapped) = 0.97

Possible Genotypes
Shotgun Sequence Data

\[
\begin{align*}
\text{Reference Genome} & : \\
& \quad \text{ATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG} \\
& \quad \text{ATGCTAGCTGATAGCTAGC} \\
& \quad \text{AGCTGATAGCTAGC} \\
& \quad \text{GCTAGCTGATAGCTAGC} \tag{Sequence Reads}
\end{align*}
\]

\[
\begin{align*}
\text{5’-ActGGTagctgATGCTAGCTGATAGCTAGC} & \tag{Reference Genome} \\
& \text{TAGCTAGCTGATGAGCCCGATCGCTGCTAGCTC} \\
& \text{ACTGGTCGATGCTAGCTAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG} - 3’
\end{align*}
\]

\[
\begin{align*}
\text{Possible Genotypes} & : \\
P(\text{reads} | A/A \text{, read mapped}) & = 0.00000099 \\
P(\text{reads} | A/C \text{, read mapped}) & = 0.0625 \\
P(\text{reads} | C/C \text{, read mapped}) & = 0.0097
\end{align*}
\]
Shotgun Sequence Data

 TAGCTGATAGCTAGA
 TAGCTGATAGCTAGA
 TAGCTGATAGCTAGA
 TAGCTGATAGCTAGA
 TAGCTGATAGCTAGA

 Sequence Reads

 5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

 Reference Genome

 ATGCTAGCTGATAGCTAGC
 ATGCTAGCTGATAGCTAGC
 ATGCTAGCTGATAGCTAGC
 ATGCTAGCTGATAGCTAGC
 ATGCTAGCTGATAGCTAGC

 Possible Genotypes

 \[
 \begin{align*}
 P(\text{reads}|A/A, \text{read mapped}) &= 0.00000098 \\
 P(\text{reads}|A/C, \text{read mapped}) &= 0.03125 \\
 P(\text{reads}|C/C, \text{read mapped}) &= 0.000097
 \end{align*}
 \]
Shotgun Sequence Data

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome
GCTAGCTGATAGCTAG
C
TAGCTGATGAGCCCGA
AGCTGATAGCTAG
C
TAGCTGATGAGCC
ATAGCTAG
A
TAGCTGATGAGCCCGATCGCTG
ATAGCTAGCTGATAGCTAG
C
TAGCTGATGAGCC
AGCTGATAGCTAG
C
TAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAG
C
TAGCTGATGAGCC
ATAGCTAG
A
TAGCTGATGAGCCCGA
Reference Genome

Sequence Reads
5’-ACTGGTCGATGCTAGCTGATAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

\[ P(\text{reads}|A/A, \text{read mapped}) = 0.00000098 \]

\[ P(\text{reads}|A/C, \text{read mapped}) = 0.03125 \]

\[ P(\text{reads}|C/C, \text{read mapped}) = 0.000097 \]

Combine these likelihoods with a prior incorporating information from other individuals and flanking sites to assign a genotype.
Shotgun Sequence Data

$\text{TAGCTGATAGCTAGA}\text{TAGCTGATAGCTAGCTAGCTGCTGCTAGCTCGACG}$

Reference Genome

$\text{GCTAGCTGATAGCTAG}$

$\text{ATAGCTGATAGCTAG}$

$\text{ATAGCTGATAGCTAG}$

$\text{ATAGCTGATAGCTAG}$

Sequence Reads

$\text{GCTAGCTGATAGCTAG}$

$\text{GCTAGCTGATAGCTAG}$

$\text{GCTAGCTGATAGCTAG}$

$\text{GCTAGCTGATAGCTAG}$

$\text{GCTAGCTGATAGCTAG}$

$\text{GCTAGCTGATAGCTAG}$

$\text{GCTAGCTGATAGCTAG}$

$\text{GCTAGCTGATAGCTAG}$

$\text{GCTAGCTGATAGCTAG}$

$\text{5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTC'-3'}$

$\text{Reference Genome}$

$$P(\text{Genotype|reads}) = \frac{P(\text{reads|Genotype}) Prior(\text{Genotype})}{\sum_G P(\text{reads|G}) Prior(G)}$$

Combine these likelihoods with a prior incorporating information from other individuals and flanking sites to assign a genotype.
Ingredients That Go Into Prior

• Most sites don’t vary
  – $P(\text{non-reference base}) \sim 0.001$

• When a site does vary, it is usually heterozygous
  – $P(\text{non-reference heterozygote}) \sim 0.001 \times \frac{2}{3}$
  – $P(\text{non-reference homozygote}) \sim 0.001 \times \frac{1}{3}$

• Mutation model
  – Transitions account for most variants ($C \leftrightarrow T$ or $A \leftrightarrow G$)
  – Transversions account for minority of variants
From Sequence to Genotype:
Individual Based Prior

\[
\begin{align*}
\text{TAGCTGATAGCTAG} &\quad \text{A} \quad \text{TAGCTGATAGCTAGCTAGCTAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'} \\
\text{ATAGCTAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTC} &
\end{align*}
\]

\[
\begin{align*}
\text{ATGCTAGCTGATAGCTAGCTAGCTGATGAGCCG} &
\end{align*}
\]

\[
\begin{align*}
\text{AGCTGATAGCTAGCTGATGAGCCCGATCGCTG} &
\end{align*}
\]

\[
\begin{align*}
\text{GCTAGCTGATAGCTAGCTGATGAGCCC} &
\end{align*}
\]

Sequence Reads

\[
\begin{align*}
\text{5'-ACTGCTGATGCTAGCTGATAGCTAGCTGATGAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'} \\
\text{Reference Genome}
\end{align*}
\]

\[
\begin{align*}
\text{P(reads|A/A)} = 0.00000098 &\quad \text{Prior(A/A)} = 0.00034 &\quad \text{Posterior(A/A)} = <.001 \\
\text{P(reads|A/C)} = 0.03125 &\quad \text{Prior(A/C)} = 0.00066 &\quad \text{Posterior(A/C)} = 0.175 \\
\text{P(reads|C/C)} = 0.000097 &\quad \text{Prior(C/C)} = 0.99900 &\quad \text{Posterior(C/C)} = 0.825
\end{align*}
\]

Individual Based Prior: Every site has 1/1000 probability of varying.
From Sequence to Genotype:
Individual Based Prior

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome
GCTAGCTGATAGCTAG
C
TAGCTGATGAGCCCGA
AGCTGATAGCTAGC
TAGCTGATGAGCC
ATAGCTGATAGCTAGC
TAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAGC
TAGCTGATGAGCC
Sequence Reads
5’-ACTGGTCGATGCTAGCTGATAGCTAGC
TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’
Reference Genome

P(reads|A/A) = 0.00000098  Prior(A/A) = 0.00034  Posterior(A/A) = <.001
P(reads|A/C) = 0.03125  Prior(A/C) = 0.00066  Posterior(A/C) = 0.175
P(reads|C/C) = 0.000097  Prior(C/C) = 0.99900  Posterior(C/C) = 0.825

Individual Based Prior: Every site has 1/1000 probability of varying.
Sequence Based Genotype Calls

- **Individual Based Prior**
  - Assumes all sites have an equal probability of showing polymorphism
  - Specifically, assumption is that about 1/1000 bases differ from reference
  - If reads where error free and sampling Poisson ...
  - ... 14x coverage would allow for 99.8% genotype accuracy
  - ... 30x coverage of the genome needed to allow for errors and clustering

- **Population Based Prior**
  - Uses frequency information obtained from examining other individuals
  - Calling very rare polymorphisms still requires 20-30x coverage of the genome
  - Calling common polymorphisms requires much less data

- **Haplotype Based Prior or Imputation Based Analysis**
  - Compares individuals with similar flanking haplotypes
  - Calling very rare polymorphisms still requires 20-30x coverage of the genome
  - Can make accurate genotype calls with 2-4x coverage of the genome
  - Accuracy improves as more individuals are sequenced
From Sequence to Genotype:
Population Based Prior

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’
Reference Genome

AGCTGATAGCTAGCTGATGAGCCCGA

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’
Reference Genome

\[
P(\text{reads}|A/A) = 0.00000098 \quad \text{Prior}(A/A) = 0.04 \quad \text{Posterior}(A/A) = <.001
\]

\[
P(\text{reads}|A/C) = 0.03125 \quad \text{Prior}(A/C) = 0.32 \quad \text{Posterior}(A/C) = 0.999
\]

\[
P(\text{reads}|C/C) = 0.000097 \quad \text{Prior}(C/C) = 0.64 \quad \text{Posterior}(C/C) = <.001
\]

Population Based Prior: Use frequency information from examining others at the same site.

*In the example above, we estimated \(P(A) = 0.20\)*
From Sequence To Genotype: Population Based Prior

Population Based Prior: Use frequency information from examining others at the same site.

*In the example above, we estimated \( P(A) = 0.20 \)
Sequence Based Genotype Calls

- **Individual Based Prior**
  - Assumes all sites have an equal probability of showing polymorphism
  - Specifically, assumption is that about 1/1000 bases differ from reference
  - If reads where error free and sampling Poisson ...
  - ... 14x coverage would allow for 99.8% genotype accuracy
  - ... 30x coverage of the genome needed to allow for errors and clustering

- **Population Based Prior**
  - Uses frequency information obtained from examining other individuals
  - Calling very rare polymorphisms still requires 20-30x coverage of the genome
  - Calling common polymorphisms requires much less data
Shotgun Sequence Data

Haplotype Based Prior

Haplotype Based Prior: Examine other chromosomes that are similar at locus of interest.

In the example above, we estimated that 90% of similar chromosomes carry allele A.
Shotgun Sequence Data

Haplotype Based Prior

* 

TAGCTGATAGCTAG\textcolor{red}{A}\textcolor{blue}{T}AGCTGATGAGCCCGAT
ATAGCTAG\textcolor{red}{A}\textcolor{blue}{T}AGCTGATGAGCCCGATCGCTGCTAGCTC
ATGCTAGCTAGCTAG\textcolor{blue}{C}\textcolor{red}{T}AGCTGATGAGCC
AGCTGATAGCTAG\textcolor{blue}{C}\textcolor{red}{T}AGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAG\textcolor{blue}{C}\textcolor{red}{T}AGCTGATGAGCCCGA

Sequence Reads

5’-\textcolor{red}{A}CTGG\textcolor{blue}{G}T\textcolor{red}{G}\textcolor{blue}{A}\textcolor{red}{T}GCTAGCTGATGCTAGCTAG\textcolor{blue}{C}\textcolor{red}{T}\textcolor{blue}{A}\textcolor{red}{G}\textcolor{blue}{C}\textcolor{red}{T}GATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

\textcolor{red}{P(\text{reads}|A/A)} = 0.00000098 \quad \text{Prior}(A/A) = 0.81 \quad \text{Posterior}(A/A) = <.001$

\textcolor{blue}{P(\text{reads}|A/C)} = 0.03125 \quad \text{Prior}(A/C) = 0.18 \quad \text{Posterior}(A/C) = 0.999$

\textcolor{blue}{P(\text{reads}|C/C)} = 0.000097 \quad \text{Prior}(C/C) = 0.01 \quad \text{Posterior}(C/C) = <.001$

**Haplotype Based Prior:** Examine other chromosomes that are similar at locus of interest.

*In the example above, we estimated that 90% of similar chromosomes carry allele A.*
Sequence Based Genotype Calls

• **Individual Based Prior**
  – Assumes all sites have an equal probability of showing polymorphism
  – Specifically, assumption is that about 1/1000 bases differ from reference
  – If reads where error free and sampling Poisson ...
  – ... 14x coverage would allow for 99.8% genotype accuracy
  – ... 30x coverage of the genome needed to allow for errors and clustering

• **Population Based Prior**
  – Uses frequency information obtained from examining other individuals
  – Calling very rare polymorphisms still requires 20-30x coverage of the genome
  – Calling common polymorphisms requires much less data

• **Haplotype Based Prior or Imputation Based Analysis**
  – Compares individuals with similar flanking haplotypes
  – Calling very rare polymorphisms still requires 20-30x coverage of the genome
  – Can make accurate genotype calls with 2-4x coverage of the genome
  – Accuracy improves as more individuals are sequenced
Challenges with the basic approach ...

5'–ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG–3'
Challenges with the basic approach...

ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG

5′-ACTGGTCGATGCTAGCTGATAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3′
Challenges with the basic approach ...

CTAGATGATGAGCCCGATCGCTGCTAGCTC
AGATGATGAGCCCGATCGCTGCTAGCTGAG
GATGATGAGCCCGATCGCTGCTAGCTAGCTA
AGATGATGAGCCCGATCGCTGCTAGCTGACG
GATGATGAGCCCGATCGCTGCTAGCTCGACG
AGATGATGAGCCCGATCGCTGCTAGCTGCA
AGATGATGAGCCCGATCGCTGCTAGCTGACG
GCTAGCTAGCTGATGAGCCCGATCGCTGCT
GATGCTAGCTAGCTGATGAGCCCGACTGC
AGCTAGCTAGCTGATGAGCCCGATCGCTAGC
CTAGCTAGCTGATGAGCCCGATCGCTGCTAGC
GCTGATAGCTAGCTAGCTGATGAGCCCGAT
GATGCTAGCTAGCTAGCTGATGAGCTAGTA
GTCGATGCTAGCTAGCTAGCTAGCTAGTA
TAGCTAGCTAGCTAGCTAGCTAGCTGACG

5’-ACTGGTCGATGCTAGCTGATGCTAGCTGATGAGCCCGATCGCTGCTAGCTGACG-3’
Challenges with the basic approach ...
Variant Filtering

• Modern callers start with a candidate list of sites and annotate these ...
  – Likely good sites: variants in HapMap or Omni 2.5M arrays
  – Likely problematic sites: variants that deviate from HWE or don’t segregate in multiple families

• Then, build a model that separates likely good sites from likely bad ones ...
  – SVM, VQSR, self-organizing maps, ....

• Possible features ...
  – What is the mapping quality of reads with the variant?
  – How many other differences in reads with the variant?
  – How many individuals are heterozygotes and homozygotes?
  – How many reads with the variant are on the forward and reverse strand?
  – What fraction of reads have the variant in heterozygotes?
  – ...

Paired End Sequencing

Population of DNA fragments of known size (mean + stdev)
Paired end sequences
Paired End Sequencing

Paired Reads

Initial alignment to the reference genome

Paired end resolution
Detecting Structural Variation

- Read depth
  - Regions where depth is different from expected
    - Expectation defined by comparing to rest of genome ...
    - ... or, even better, by comparing to other individuals

- Split reads
  - If reads are longer, it may be possible to find reads that span the structural variation

- Discrepant pairs
  - If we find pairs of reads that appear to map significantly closer or further apart than expected, could indicate an insertion or deletion
    - For this approach, “physical coverage” which is the sum of read length and insert size is key

- De Novo Assembly
How Much Variation is There?

• An average genome includes:
  – About 4M SNPs
  – About 500K indels
  – Hundreds or thousands of larger deletions

• Numbers are probably underestimates ...

• ... some variants are hard to call with short reads

## Variants per genome

(1000 Genomes Project)

<table>
<thead>
<tr>
<th>Type</th>
<th>Variant sites / genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPs</td>
<td>~3,800,000</td>
</tr>
<tr>
<td>Indels</td>
<td>~570,000</td>
</tr>
<tr>
<td>Mobile Element Insertions</td>
<td>~1000</td>
</tr>
<tr>
<td>Large Deletions</td>
<td>~1000</td>
</tr>
<tr>
<td>CNVs</td>
<td>~150</td>
</tr>
<tr>
<td>Inversions</td>
<td>~11</td>
</tr>
</tbody>
</table>
Allele Frequency Spectrum
(After Sequencing 12,000+ Individuals)

http://genome.sph.umich.edu/wiki/Exome_Chip_Design
## How Much Variation is There? (TOPMed 65K)

<table>
<thead>
<tr>
<th>Variant Type</th>
<th>Category</th>
<th># PASS</th>
<th># FAIL</th>
<th>% dbSNP (PASS)</th>
<th>Known/Novel Ts/Tv (PASS)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SNP</strong></td>
<td>All</td>
<td>438M</td>
<td>85M</td>
<td>22.9%</td>
<td>1.93 / 1.69</td>
</tr>
<tr>
<td></td>
<td>Singleton</td>
<td>202M</td>
<td>24M</td>
<td>8.5%</td>
<td>1.23 / 1.54</td>
</tr>
<tr>
<td></td>
<td>Doubleton</td>
<td>69M</td>
<td>8.8M</td>
<td>12.6%</td>
<td>1.61 / 1.74</td>
</tr>
<tr>
<td></td>
<td>Tripleton ~ 0.1%</td>
<td>142M</td>
<td>24M</td>
<td>34.9%</td>
<td>2.23 / 1.99</td>
</tr>
<tr>
<td></td>
<td>0.1% ~ 1%</td>
<td>13M</td>
<td>4.5M</td>
<td>98.2%</td>
<td>2.17 / 1.79</td>
</tr>
<tr>
<td></td>
<td>1 ~ 10%</td>
<td>6.5M</td>
<td>2.9M</td>
<td>99.6%</td>
<td>1.82 / 1.75</td>
</tr>
<tr>
<td></td>
<td>&gt;10%</td>
<td>5.3M</td>
<td>2.0M</td>
<td>99.8%</td>
<td>2.11 / 1.88</td>
</tr>
<tr>
<td><strong>Indels</strong></td>
<td>All</td>
<td>33.4M</td>
<td>26.2M</td>
<td>20.1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Singleton</td>
<td>15.7M</td>
<td>4.7M</td>
<td>10.1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Doubleton</td>
<td>5.3M</td>
<td>1.8M</td>
<td>12.6%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tripleton ~ 0.1%</td>
<td>10.7M</td>
<td>8.0M</td>
<td>26.7%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1% ~ 1%</td>
<td>2.8M</td>
<td>968K</td>
<td>88.9%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 ~ 10%</td>
<td>432K</td>
<td>2.3M</td>
<td>98.5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;10%</td>
<td>298K</td>
<td>1.4M</td>
<td>99.6%</td>
<td></td>
</tr>
</tbody>
</table>
# How Much Variation is There?
(TOPMed 65K – Coding Variation)

<table>
<thead>
<tr>
<th>Type</th>
<th>Category</th>
<th>PASS Variants</th>
<th>% AC = 1</th>
<th>% AC ≤ 2</th>
<th>AF &lt; 0.1%</th>
<th>AF &lt; 1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP</td>
<td>All</td>
<td>438M</td>
<td>46.1%</td>
<td>61.9%</td>
<td>94.2%</td>
<td>98.7%</td>
</tr>
<tr>
<td></td>
<td>Synonymous</td>
<td>1.62M</td>
<td>42.9%</td>
<td>58.7%</td>
<td>94.5%</td>
<td>97.6%</td>
</tr>
<tr>
<td></td>
<td>Missense</td>
<td>3.44M</td>
<td>47.7%</td>
<td>64.1%</td>
<td>96.8%</td>
<td>98.8%</td>
</tr>
<tr>
<td></td>
<td>Stop Gain</td>
<td>103K</td>
<td>54.4%</td>
<td>71.3%</td>
<td>98.4%</td>
<td>99.5%</td>
</tr>
<tr>
<td></td>
<td>Essential Splice</td>
<td>111K</td>
<td>54.2%</td>
<td>70.3%</td>
<td>96.8%</td>
<td>98.6%</td>
</tr>
<tr>
<td>Indels</td>
<td>All</td>
<td>33.4M</td>
<td>47.0%</td>
<td>62.8%</td>
<td>94.9%</td>
<td>98.8%</td>
</tr>
<tr>
<td></td>
<td>Frameshift</td>
<td>97.0K</td>
<td>59.9%</td>
<td>76.0%</td>
<td>98.7%</td>
<td>99.6%</td>
</tr>
<tr>
<td></td>
<td>Inframe</td>
<td>65.6K</td>
<td>48.6%</td>
<td>65.3%</td>
<td>97.5%</td>
<td>99.3%</td>
</tr>
<tr>
<td></td>
<td>Ess. Splice &amp; Others</td>
<td>12.7K</td>
<td>52.7%</td>
<td>68.8%</td>
<td>97.0%</td>
<td>98.8%</td>
</tr>
</tbody>
</table>
Summary

• Introduction to whole genome sequencing
  – Read mapping
  – Genotype calling
  – Analysis of structural variation

• Sequencing and the genetics of complex traits
  – Advantages and disadvantages versus genotyping
  – What sorts of things might we learn?
Recommended Reading