SeqShop Day 2

Variant Calling & Filtering: SNPs

Hyun Min Kang
Department of Biostatistics
University of Michigan
In this Lecture...

- Calling single nucleotide polymorphisms (SNPs)
- Improving quality of SNP calls
- Design of whole genome sequencing studies
- More details about sequence analysis pipeline
  - Detecting DNA contamination
  - Duplication removal & Base quality recalibration
**Shotgun Sequence Reads**

- Typical short read might be 70-150 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
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 now takes no more than a few hours per million reads. Analyzing these data without a reference human genome would require much longer reads or result in very fragmented assemblies.
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

Sequence Reads

5’-ACTGGTCACTGCTAGCTAGCCTAGCTAGGAGCCCCTCAGTGCCTGCTAGCTGACG-3’

Reference Genome

A/C

Predicted Genotype

A/C

SEQUENCE ANALYSIS WORKSHOP
SHOTGUN SEQUENCE DATA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

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**

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-A<sup>CTGGTCGATGCTAGCTGATAGCTAGCTGATGAGCCCGATCGCTGCTAGCTGCTGACG-3'</sup>

Reference Genome

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

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

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

Possible Genotypes
SHOTGUN SEQUENCE DATA

Sequence Reads

5' - ACTGGTCACTGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTAGC

Reference Genome

Possible Genotypes

P(reads|A/A, read mapped)= 0.01
P(reads|A/C, read mapped)= 0.50
P(reads|C/C, read mapped)= 0.99
SHOTGUN SEQUENCE DATA

AGCTGATAGCTAGCTAGCTAGCTAGGAGCCCGATCGCTG
GCTAGCTGATAGCTAGCTAGCTAGCTAGGAGCCCGA

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTAGCTAGCTAGCTAGCTCCGACG-3’

Reference Genome

Sequence Reads

P(reads|A/A, read mapped) = 0.0001
P(reads|A/C, read mapped) = 0.25
P(reads|C/C, read mapped) = 0.98

Possible Genotypes
**Shotgun Sequence Data**

ATGCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCC
AGCTGATAGCTAGCTAGCTAGCTGATGAGCCCCGATCGCTG
GCTAGCTGATAGCTAGCTAGCTGATGAGCCCCGA

Sequence Reads

5’-ACTGGTCTGATGCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

\[ P(\text{reads}|\text{A/A, read mapped}) = 0.000001 \]

\[ P(\text{reads}|\text{A/C, read mapped}) = 0.125 \]

\[ P(\text{reads}|\text{C/C, read mapped}) = 0.97 \]

Possible Genotypes
**SHOTGUN SEQUENCE DATA**

```
ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC
ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC
AGCTGATAGCTAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA
```

Sequence Reads

```
5’-ACTGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’
```

Reference Genome

\[
P(\text{reads}|A/A, \text{read mapped}) = 0.00000099 \\
P(\text{reads}|A/C, \text{read mapped}) = 0.0625 \\
P(\text{reads}|C/C, \text{read mapped}) = 0.0097
\]

Possible Genotypes
SHOTGUN SEQUENCE DATA

TAGCTGATAGCTAGA TAGCTGATGAGCCCGAT
ATAGCTAGA TAGCTGATGAGCCCGATCGCTGCTAGCTC
ATGCTAGCTGATAGCTAGC TAGCTGATGAGCC
AGCTGATAGCTAGC TAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAGC TAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGC TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

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

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

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

TAGCTGATAGCTAG\textcolor{red}{A}TAGCTGATGAGCCCGAT
\textcolor{red}{A}TAGCTGATAGCTGATGAGCCCGATCGCTGCTAGCTC
ATGCTAGCTGATAGCT\textcolor{red}{C}TAGCTGATGAGCC
AGCTGATAGCTAG\textcolor{red}{C}TAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAG\textcolor{red}{G}TAGCTGATGAGCCCGA
5’-ACTGGTCGATGCTAGCTGATAGCTAG\textcolor{red}{C}TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Sequence Reads

Reference Genome

P(reads|A/A, read mapped) = 0.00000098

P(reads|A/C, read mapped) = 0.03125

P(reads|C/C, 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**

\[
\Pr(G|R) = \frac{\Pr(R|G) \pi(G)}{\sum_g \Pr(R|g) \pi(g)}
\]

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
  - $P(\text{non-reference base}) \approx 0.001$

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

- 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

```
TAGCTGATAGCTAGA TAGCTGATGAGCCCGAT
ATAGCTAGA TAGCTGATGAGCCCGATCGCTGCTAGCTC
ATGCTAGCTGATAGCTAGC TAGCTGATGAGCC
AGCTGATAGCTAGC TAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAGC TAGCTGATGAGCCCGA
```

Sequence Reads

```
5'-ACTGGTGATGCTAGCTGATAGCTAGC TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'
```

Reference Genome

\[
P(\text{reads}|A/A) = 0.00000098 \quad \text{Prior}(A/A) = 0.00034 \quad \text{Posterior}(A/A) = <.001
\]

\[
P(\text{reads}|A/C) = 0.03125 \quad \text{Prior}(A/C) = 0.00066 \quad \text{Posterior}(A/C) = 0.175
\]

\[
P(\text{reads}|C/C) = 0.000097 \quad \text{Prior}(C/C) = 0.99900 \quad \text{Posterior}(C/C) = 0.825
\]

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

Sequence Reads

5’-ACTGGTGCATGCTAGCTGATAGCTAGCCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

TAGCTGATAGCTAGATAAGCTGATGAGCCCGAT
ATAGCTAGATAAGCTGATGAGCCCGATCGCTGCTAGCTC
AGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGAT
GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

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

\[ P(\text{reads}|A/A) = 0.00000098 \quad \text{Prior}(A/A) = 0.00034 \quad \text{Posterior}(A/A) = <.001 \]

\[ P(\text{reads}|A/C) = 0.03125 \quad \text{Prior}(A/C) = 0.00066 \quad \text{Posterior}(A/C) = 0.175 \]

\[ P(\text{reads}|C/C) = 0.000097 \quad \text{Prior}(C/C) = 0.99900 \quad \text{Posterior}(C/C) = 0.825 \]
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 were 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
**From Sequence to Genotype: Population Based Prior**

```
TAGCTGATAGCTAGATAGCTGATGAGCCCGAT
ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC
ATGCTAGCTGATAGCTAGCTAGCTAGCTAGGAGCC
AGCTGATAGCTAGCTAGCTAGCTAGCTAGCTAGCCC
GCTAGCTGATAGCTAGCTAGCTAGCTAGCTAGCCC
```

Sequence Reads  
```
5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTAGCTAGCTAGCTAGCTGCTAGCTAGCTCGACG-3'
```

Reference Genome

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

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

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

**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**

```
TAGCTGATAGCTAGATAGCTGATGAGCCCGAT
ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC
ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC
AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA
```

Sequence Reads

```
5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'
```

Reference Genome

<table>
<thead>
<tr>
<th>Event</th>
<th>Probability</th>
<th>Prior Probability</th>
<th>Posterior Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>(reads</td>
<td>A/A)</td>
<td>0.000000098</td>
<td>0.04</td>
</tr>
<tr>
<td>(reads</td>
<td>A/C)</td>
<td>0.03125</td>
<td>0.32</td>
</tr>
<tr>
<td>(reads</td>
<td>C/C)</td>
<td>0.000097</td>
<td>0.64</td>
</tr>
</tbody>
</table>

**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**

**HAPLOTYPING BASED PRIOR**

```
TAGCTGATAGCTAGATAGCTGATGAGCCCGAT
ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC
ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC
AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA
```

Sequence Reads

5’-ACTGGTCAATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTGACG-3’

Reference Genome

\[
P(\text{reads}|A/A) = 0.00000098 \quad \text{Prior}(A/A) = 0.81 
\]

\[
P(\text{reads}|A/C) = 0.03125 \quad \text{Prior}(A/C) = 0.18 
\]

\[
P(\text{reads}|C/C) = 0.000097 \quad \text{Prior}(C/C) = 0.01 
\]

**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
HAPLOTYPt TYPE BASED PRIOR

TAGCTGATAGCTAGACTAGCTGATGAGCCCGAT
ATAGCTAGACTAGCTGATGAGCCCGATCGCTGCTAGCTC
ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC
AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads
5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'
Reference Genome

P(reads|A/A) = 0.00000098  Prior(A/A) = 0.81  Posterior(A/A) = <.001
P(reads|A/C) = 0.03125   Prior(A/C) = 0.18  Posterior(A/C) = 0.999
P(reads|C/C) = 0.000097  Prior(C/C) = 0.01  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.
SUMMARY: 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
**False Positive / Negative SNP calls**

- In an ideal world (no errors in alignment or base quality)
  - Posterior probability quantifies statistical evidence
  - Determine posterior probability cutoff to calibrate between false positive vs. false negative

- In the real world..
  - Alignment errors create false positive SNPs
    - Segmental duplication mapped to the same position
    - Alignment can be ambiguous at the end of the read
  - Low depth limits the power to detect rare SNPs
    - If the SNP is not shared across individuals, insufficient sequence depth limits the SNP detection power
**Per Base Alignment Qualities**

Short Read

GATAGCTAGCTAGCTGATGA GCCG
5'–AGCTGATAGCTAGCTAGCTAGCTGATGAGCCCAGATC-3'

Reference Genome
PER BASE ALIGNMENT QUALITIES

Should we insert a gap?

Short Read

GATAGCTAGCTAGCTGATGAGCC-G

5’-AGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome
Compensate for Alignment Uncertainty With Lower Base Quality

**Per Base Alignment Qualities**

Reference Genome

Short Read

$$\text{GATAGCTAGCTAGCTGATGAGCCC}_G$$

5’-AGCTGATAGCTAGCTAGCTAGCTGATGAGCCCAGATC-3’
How to Tell Good from Bad: Example

Reference: 

Sample 1

We expect 50:50 read distribution for HET sites (Allelic imbalance is indicator of alignment artifacts)
**How to Tell Good from Bad:** Example

Reference:

**Sample 1**

<table>
<thead>
<tr>
<th>Reference:</th>
<th>... AGGTCTAA ...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>... C ...</td>
</tr>
<tr>
<td></td>
<td>... T ...</td>
</tr>
<tr>
<td></td>
<td>... C</td>
</tr>
<tr>
<td></td>
<td>... T ...</td>
</tr>
<tr>
<td></td>
<td>... T ... 0.6</td>
</tr>
</tbody>
</table>

**Good**

<table>
<thead>
<tr>
<th>Reference:</th>
<th>... GAATTACA ...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>... C ...</td>
</tr>
<tr>
<td></td>
<td>... T ...</td>
</tr>
<tr>
<td></td>
<td>... T ...</td>
</tr>
<tr>
<td></td>
<td>... T ... 0.8</td>
</tr>
</tbody>
</table>

**Bad**
How to Tell Good from Bad: Example

Hard to tell whether it’s random deviation or not on a single sample.
MULTI-SAMPLE FILTERING IS INFORMATIVE

Reference: ... AGGTCTAA ...

Sample 1

   ... C T C T T T 0.6

Sample 2

   ... T C C C T T 0.4

Sample N

   ... C C T T T T 0.67

Overall Balance: 0.56

... GAATTACA ...

Sample 1

   ... C T T T T T 0.8

Sample 2

   ... T T T T C T 0.8

Sample N

   ... C T T T T C 0.67

Overall Balance: 0.75
# Filtering Criteria Examples

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td>Overall depth across samples</td>
</tr>
<tr>
<td>QUAL</td>
<td>Overall genotype confidence</td>
</tr>
<tr>
<td>Allele Balance</td>
<td>$(# \text{REF})/(# \text{ALT})$ in HET sites</td>
</tr>
<tr>
<td>Strand Bias</td>
<td>Correlation of ALT allele with $+/-$ strand</td>
</tr>
<tr>
<td>Cycle Bias</td>
<td>Correlation of ALT allele with read cycle</td>
</tr>
<tr>
<td>Mapping Quality</td>
<td>Average quality of aligned sequences</td>
</tr>
<tr>
<td>HWE</td>
<td>Hardy-Weinberg Equilibrium Test Statistics</td>
</tr>
</tbody>
</table>
**Hard Filtering by Individual Thresholds**

- **Problems**
  - False negative increases with number of filters
  - Too many knobs to turn (thresholds)
Filtering by Supervised Learning

- Use features to train a support vector machine (SVM)
  - Can be trained using suspected positive/negative examples
  - Provides single score from all features combined

- Training
  - Positive examples
    - Known polymorphic sites
  - Negative examples
    - Filtered out by multiple hard filters

- Input
  - All individual features collected for each site
FILTERING BY SUPERVISED LEARNING

• Use features to train a support vector machine (SVM)
  – Can be trained using suspected positive/negative examples
  – Provides single score from all features combined

• Training
  – Positive examples
    • Known polymorphic sites
  – Negative examples
    • Filtered out by multiple hard filters

• Input
  • All individual features collected for each site
A >20 dimensional feature set was used for final filtering under nonlinear kernel space.
**Improved Sensitivity by SVM**

SNP Discovery Sensitivity on Low-coverage Data

- **GotCloud-Raw**
- **GotCloud-SVM**
- **GATK-Raw**
- **GATK-VQSR**

Jun et al. (2015) Genome Research
HOW CAN WE KNOW THE QUALITY OF A SNP CALL SET?

- Overlap with existing SNPs (dbSNP, HapMap)
- Ts/Tv ratio is an effective surrogate for SNP quality

**Genome-wide**
In a true SNP, expected $Ts/Tv \sim 2$

In a (random) false SNP expected $Ts/Tv \sim 0.5$

**In exomes (coding sequences)**
True SNP has expected $Ts/Tv \sim 3$

Nonsynonymous SNPs: $Ts/Tv \sim 2$
Synonymous SNPs: $Ts/Tv > 5$
(Calculated from codon frequency)
## Summarizing Quality of SNP Calls

### 1000 Genomes Phase 3 call sets using GotCloud

<table>
<thead>
<tr>
<th>Call set</th>
<th># SNPs (# WGS)</th>
<th>in dbSNP (129)</th>
<th>Known Ts/Tv</th>
<th>Novel Ts/Tv</th>
<th>HapMap3 Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Call Set</td>
<td>1,787,005 (79M)</td>
<td>11.5%</td>
<td>2.29</td>
<td>2.15</td>
<td>98.8%</td>
</tr>
<tr>
<td>Filtered (PASS) Call Set</td>
<td>1,541,273 (68M)</td>
<td>12.4%</td>
<td>2.35</td>
<td>2.37</td>
<td>98.5%</td>
</tr>
<tr>
<td>Filtered-out (FAIL) Call Set</td>
<td>245,732 (11M)</td>
<td>5.7%</td>
<td>1.68</td>
<td>1.29</td>
<td>0.3%</td>
</tr>
</tbody>
</table>
Results in Exome Sequencing Project

<table>
<thead>
<tr>
<th></th>
<th>Unfiltered</th>
<th>Filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ts/Tv</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ts/Tv</td>
<td>2.78</td>
<td></td>
</tr>
</tbody>
</table>

In Variants with AF > 1%

- Not in 1000G
- In 1000G
**SUMMARY : VARIANT FILTERING**

- Initial SNP calls typically contain false positive variants

- Ts/Tv, overlap with external resources are good surrogates of SNP quality

- Various features from sequence data can be used to extract high quality variants

- Combining information across multiple samples using SVM further improves the performance of variant filtering
Low-coverage or Deep Sequencing?

• 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?
Recipe For Imputation With Shotgun Sequence Data

• Start with some plausible configuration for each individual

• Use Markov model to update one individual conditional on all others

• Repeat previous step many times

• Generate a consensus set of genotypes and haplotypes for each individual
**Haplotype-aware Genotype Refinement**

- People share ‘blocks’ of genotypes
- Haplotype-phasing improves genotype accuracy by correcting unlikely genotypes and filling in missing genotypes

![Diagram showing haplotype-phasing process]
Haplotype-Aware Analysis Improves Genotype Accuracy

Single Site Analysis
- 21.4% HET errors

Haplotype Aware Analysis
- 2.0% HET errors
As More Samples Are Sequenced, Low Pass Genotypes Improve

<table>
<thead>
<tr>
<th>Analysis</th>
<th>#SNPs</th>
<th>dbSNP%</th>
<th>Missing HapMap %</th>
<th>Ts/Tv</th>
<th>Accuracy at Hets*</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 2010</td>
<td>9,158,226</td>
<td>63.5</td>
<td>7.0</td>
<td>1.91</td>
<td>96.74</td>
</tr>
<tr>
<td>Michigan/EUR 60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>August 2010</td>
<td>10,537,718</td>
<td>52.5</td>
<td>5.6</td>
<td>2.04</td>
<td>97.56</td>
</tr>
<tr>
<td>Michigan/EUR 186</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>October 2010</td>
<td>13,276,643</td>
<td>50.1</td>
<td>1.8</td>
<td>2.20</td>
<td>97.91**</td>
</tr>
<tr>
<td>Michigan/EUR 280</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Accuracy of Low Pass Genotypes Generated by 1000 Genomes Project, When Analyzed Here At the University of Michigan
IMPLICATIONS FOR
WHOLE GENOME SEQUENCING STUDIES

• Suppose we could afford 2,000x data (6,000 GB)
• We could sequence 67 individuals at 30x

<table>
<thead>
<tr>
<th>Minor Allele Frequency</th>
<th>0.5 – 1.0%</th>
<th>1.0 – 2.0%</th>
<th>2.0 – 5.0%</th>
<th>&gt;5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of Detected Sites</td>
<td>59.3%</td>
<td>90.1%</td>
<td>96.9%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Genotyping Accuracy</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>.... Heterozygous Sites Only</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Correlation with Truth (r²)</td>
<td>99.8%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Effective Sample Size (n·r²)</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>67</td>
</tr>
</tbody>
</table>
### Implications for Whole Genome Sequencing Studies

- Suppose we could afford 2,000x data (6,000 GB)
- We could sequence 1000 individuals at 2x

<table>
<thead>
<tr>
<th>Minor Allele Frequency</th>
<th>0.5 – 1.0%</th>
<th>1.0 – 2.0%</th>
<th>2.0 – 5.0%</th>
<th>&gt;5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of Detected Sites</td>
<td>79.6%</td>
<td>98.8%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Genotyping Accuracy</td>
<td>99.6%</td>
<td>99.5%</td>
<td>99.5%</td>
<td>99.8%</td>
</tr>
<tr>
<td>.... Heterozygous Sites Only</td>
<td>78.8%</td>
<td>89.5%</td>
<td>95.9%</td>
<td>99.8%</td>
</tr>
<tr>
<td>Correlation with Truth ($r^2$)</td>
<td>56.7%</td>
<td>76.1%</td>
<td>88.2%</td>
<td>97.8%</td>
</tr>
<tr>
<td>Effective Sample Size ($n \cdot r^2$)</td>
<td>567</td>
<td>761</td>
<td>882</td>
<td>978</td>
</tr>
</tbody>
</table>
**Given Fixed Capacity, Should We Sequence Deep or Shallow?**

<table>
<thead>
<tr>
<th></th>
<th>.5 – 1%</th>
<th>1 – 2%</th>
<th>2-5%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>400 Deep Genomes (30x)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discovery Rate</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Het. Accuracy</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Effective N</td>
<td>400</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td><strong>3000 Shallow Genomes (4x)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discovery Rate</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Het. Accuracy</td>
<td>90.4%</td>
<td>97.3%</td>
<td>98.8%</td>
</tr>
<tr>
<td>Effective N</td>
<td>2406</td>
<td>2758</td>
<td>2873</td>
</tr>
</tbody>
</table>

Li et al, *Genome Research*, 2011
Low Pass Sequencing detects common variants very well.
Genotypes are more accurate for common variants.
## Overall genotype accuracy (Genomewide, 1000G Phase 3)

<table>
<thead>
<tr>
<th>Gold</th>
<th>Eval</th>
<th>Type</th>
<th>Alleles</th>
<th>ALL</th>
<th>NonRef</th>
<th>Het</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGI</td>
<td>Phase3</td>
<td>SNP</td>
<td>Biallelic</td>
<td>0.9994</td>
<td>0.9923</td>
<td>0.9930</td>
</tr>
<tr>
<td>CGI</td>
<td>Phase3</td>
<td>SNP</td>
<td>Multiallelic</td>
<td>0.9979</td>
<td>0.9877</td>
<td>0.9892</td>
</tr>
<tr>
<td>CGI</td>
<td>Phase3</td>
<td>Indel</td>
<td>Biallelic</td>
<td>0.9983</td>
<td>0.9815</td>
<td>0.9889</td>
</tr>
<tr>
<td>CGI</td>
<td>Phase3</td>
<td>Indel</td>
<td>Multiallelic</td>
<td>0.9867</td>
<td>0.9453</td>
<td>0.9722</td>
</tr>
</tbody>
</table>
Higher Depth, Higher Sensitivity
HIGHER DEPTH, SMALLER GENOTYPING ERRORS

% Het Errors with CG vs. Sequencing Depth

POP
- LWK
- YRI
- PEL
- CHS
- KHV
- CEU
- PJL
DEEPER SEQUENCING DETECTS MORE SINGLETONS
**PER-SAMPLE VARIANT COUNT VARIES BY POPULATION**

- AFR
- AMR
- EAS
- EUR
- SAS

The graph shows the distribution of total number of variants per sample across different sequencing depths for various populations. The population distribution varies significantly, with some populations showing higher densities of variants at certain sequencing depths, while others have a more even distribution. The legend on the right indicates different sub-populations within each broader category.
How Much Variation In a Genome? (Based on 1000G Phase 3)

• An average genome includes (varies by populations):
  – 3.55M ~ 4.31 SNPs
  – 546K ~ 625K indels
  – 939 ~ 1,100 large deletions

• Numbers are probably underestimates ...
• ... some variants are hard to call with short reads
How Much Variation In a Genome?
(Based on 1000G Phase 3)

• An average genome includes (varies by populations):
  – 3.55M ~ 4.31 SNPs
  – 546K ~ 625K indels
  – 939 ~ 1,100 large deletions
  – 153 ~ 170 large duplications
  – 845 ~ 1,030 Alu repeats

• Numbers are probably underestimates ...
• ... some variants are hard to call with short reads
How Much Variation In an Exome? (Based on 1000G Phase 3)

- An average exome includes (varies by populations):
  - 11.4k ~ 13.8k synonymous variants
  - 10.2k ~ 12.2k non-synonymous variants
  - 149 ~ 182 loss-of-function variants
  - 16 ~ 20 HGMD disease-causing mutations
  - 24 ~ 50 Clinvar variants

- Number of GWAS (not just in exome) variants are 1990 ~ 2080.
Allele Frequency Spectrum (After Sequencing 12,000+ Individuals)

http://genome.sph.umich.edu/wiki/Exome_Chip_Design
**Summary: Design of Sequencing Studies**

- Analysis of Low Pass Sequence Data
  - Single sample analyses produce poor quality variants.
  - Single site analyses produce poor quality genotypes.
  - Multi-sample, multi-site analyses can work quite well.
  - Haplotype-aware analysis works even better

- Low pass analyses are attractive for complex disease association studies.

- To study rarer variants better, what is the optimal coverage?
DNA Sample Contamination

*Picture from D. Figarelli, National Forensic Science Tech. Center*
Unexpected Results in SNP Calls

Per-sample genotype count

<table>
<thead>
<tr>
<th></th>
<th>Ref</th>
<th>Dad</th>
<th>Mom</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
</tbody>
</table>

Expected Het:Hom is 2:1 (or less)

Something changed?
**Reference-Aligned Sequence Reads**

Reference: 5′ - **AGCTGATAGCTAGCTACCTGACGAGCCCAGATC** - 3′

Sample:
- AGCTGATAGCTGGCTA
- AGCTGATAGCTGGCTAACTG
- GCTGATAGCTAGCTAACTGACGAG
- CTGATAGCTAGCTAACTGACGAGC
- TGATAGCTGGCTAACTGACGAGCC
- ATAGCTAGCTAACTGACGAGCCC
**Single-Nucleotide Polymorphism (SNP)**

**Reference**

5’ - AGCTGATA GCTAGCTA CCTGACGAGCCCAGATC - 3’

**Sample**

- AGCTGATA GCTGGCTA
- AGCTGATA GCTGGCTA AACTG
- GCTGATA GCTAGCTA AACTGACGAG
- CTGATA GCTAGCTA AACTGACGAGC
- TGATAGCTGGCTA AACTGACGAGC
- ATAGCTAGCTA AACTGACGAGCCCG

**Genotype:**

- GA (Heterozygote)
- AA (Homozygote)
**Base Distribution in Two Samples**

**Reference**

5’ - AGCTGATAAGCTAGCTATCTGACGAGCCCGATC - 3’

**Sample 1**

AGCTGATAAGCTGGCTAGCTG
GCTGATAAGCTAGCTAGCTGACGAG
CTGATAAGCTGGCTAGCTGACGAGC
ATAGCTAGCTAGCTGACGAGCCCG

**Sample 2**

AGCTGATAAGCTGGCTATCTG
GCTGACAGCTGGCTATCTGACGAG
CTGACAGCTGGCTATCTGACGAGC
ATAGCTGGCTATCTGACGAGCCCG
**Base Distribution in Two Samples**

Reference 5’ - AGCTGATAGCTAGCTATCTGACGAGCCCCGATC - 3’

Sample 1

- AGCTGATAGCTGGCTAGCTG
- GCTGATAGCTAGCTAGCTGAG
- CTGATAGCTGGCTAGCTGAGC
- ATAGCTAGCTAGCTGAGCCCG

Sample 2

- AGCTGATAGCTGGCTATCTG
- GCTGACAGCTGGCTATCTGACGAG
- CTGACAGCTGGCTATCTGACGAC
- ATAGCTGGCTATCTGACGAGCCCG

**Heterozygous**  
**Homozygous ALT**
**Contamination: Mixture of Samples**

Reference

\[5' - \textcolor{red}{AGCTGATAGCTAGCTATCTGACGA}GCCC\textcolor{blue}{GATC} - 3'\]

Sample 1+2

\[
\begin{align*}
&\textcolor{red}{AGCTGATAGCTGGCTAGCTAGCTG} \\
&\textcolor{green}{GCTGATAGCTAGCTAGCTGACGAG} \\
&\textcolor{purple}{CTGATAGCTGGCTAGCTGACGAGC} \\
&\textcolor{blue}{ATAGCTAGCTAGCTGACGA}GCCC \\
&\textcolor{red}{AGCTGATAGCTGGCTATCTG} \\
&\textcolor{green}{GCTGACAGCTGGCTATCTGACGAG} \\
&\textcolor{purple}{CTGACAGCTGGCTATCTGACGAGC} \\
&\textcolor{blue}{ATAGCTGGCTATCTGACGAGCCC} \\
\end{align*}
\]


Contamination: Changes Base Distributions

Reference
5’-AGCTGATAGCTAGCTATCTGACGAGCCCGATC-3’

Sample 1+2
AGCTGATAGCTGGCTAGCTG
GCTGATAGCTAGCTAGCTGACGAG
CTGATAGCTGGCTAGCTGACGAC
ATAGCTAGCTAGCTGACGAGCCC
AGCTGATAGCTGGCTATCTG
GCTGACAGCTGGCTATCTGACGAG
CTGACAGCTGGCTATCTGACGAC
ATAGCTGGCTATCTGACGAGCCC

More heterozygote SNPs with biased distribution
Using Mixture Model to Detect and Estimate Contamination

# of HETs

#HETs/#HOMs

Estimated % contamination

Sequencing Date

Something changed
DNA Contamination can be precisely estimated

- Simulated contamination from real sequence data
  - Can accurately detect as low as 1% contamination
  - Works with or without known genotype data
**Summary: DNA Contamination**

- DNA contamination is a practically important problem in the analysis of sequence data.

- Detecting and estimating contamination from sequence data is possible via mixture models.

- Is it also possible to correct for the contamination effect if exists (not shown today).
SEQUENCE ALIGNMENT PIPELINE

FASTQ → \texttt{bwa} / \texttt{mosaik} → Raw BAM → \texttt{rgMergeBam} → Merged BAM → Processed BAM → qplot & verifyBamID → QC metrics

Each FASTQ pair tags different readGroup
Merge BAMs with same library before dedupping

GotCloud integrates dedup/recalibration steps together
dedup & recab
SEQUENCE ALIGNMENT PIPELINE

FASTQ → Raw BAM

Each FASTQ pair tags different readGroup

RAW TEXT START

bwa / mosaik → Raw BAM

merge & sort → Merged BAM

Merge BAMs with same library before dedupping

dedup → Processed BAM

QC Cloud integrates dup/re-calibration steps together

QC metrics

RAW TEXT END
DUPLICATE READS - EXAMPLES
REMOVING DUPLICATES – WHY?

• PCR duplicates are NOT independent observations
• Sequencing errors in PCR duplicates will look like true variants, and creates false positive (rare) variants.

![Graph showing Ts/Tv ratio for Known and Novel variants with and without ClipOverlap for All Variants and Singletons.]
Removing Duplicates – How?

• Key Idea
  – If there are too many reads mapped at the same position (beyond random chance), they are likely duplicate reads
  – Keep only one among many possible copies

• For single-end reads
  – Remove if the read is mapped at the same position.
  – This is probably too conservative (how much?)

• For paired-end reads
  – Remove if both ends are mapped at the same positions.
  – Random chance of collision is very low (how much?)

Technically, we compare ‘unclipped’ positions of reads
BASE QUALITY RECALIBRATION – WHY?

Empirical vs reported Phred score

- dedup
- recalNoOQ1
**Base Quality Recalibration – How?**

- Consider only mapped & unique (non-dup) reads
- Ignore all known variant positions in dbSNP
  - The remaining sites should have very small number of variants (e.g. $10^{-5}$ / bp), so most of them should match to the reference genome
- For each possible combinations of covariates,
  - such as (1) original base quality (2) cycle (3) strand (4) neighboring nucleotides, and (5) readgroups
  - calculate the number of matches and mismatches to the reference
- Convert the probability of mismatch as recalibrated score.
Summary: Dedup & Recalibration

- Removing duplicated reads are important to improve the quality of variant calls
  - By avoiding multiple correlated errors
  - Especially helpful for reducing false rare variants

- Base quality recalibration improves the accuracy of the base quality
  - By empirically investigating the mismatch to reference genome

- How much important are these procedures for deep sequence data?