Whole Genome Sequencing
Low Pass Sequencing

Gonçalo Abecasis
Previous Lecture

- Introduction to Whole Genome Sequencing
  - What will we learn from whole genome sequencing?

- Challenges with Read Mapping

- Interpreting Mismatches: Variant or Error
  - Single individual analyses require deep sequencing
  - Multi-individual analyses can use shallower data

- Information contained in paired reads
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
Shotgun Sequence Data

TAGCTGATA TAGCTGATAGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGAT
ATAGCTAGA TAGCTGATAGCTGACGCTGCTGCTAGCTC
ATGCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCC
AGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

5′-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3′

Reference Genome

Sequence Reads

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

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

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

Combine these likelihoods with a prior incorporating information from other individuals and flanking sites to assign a genotype.
From Sequence to Genotype: Individual Based Prior

\begin{verbatim}
TAGCTGATAGCTAGATAGCTGATGAGCCCGAT
ATAGCTAGATAGCTGATAGCCCGATCGCTGCTAGCTC
ATGCTAGCTGATAGCTAGCTAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTG
AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAG
GCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTGATGAGCCCGATCGCTGCTAGCTG
\end{verbatim}

Sequence Reads

\begin{verbatim}
5'-ACTGGTCGATGCTAGCTGATGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTGACG-3'
\end{verbatim}

Reference Genome

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

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

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

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

\[
\begin{align*}
5' &- \text{ACTGGTCGATGCTAGCTAGCTAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG} - 3' \\
\text{Reference Genome} \\
\text{Sequence Reads}
\end{align*}
\]

\[
\begin{align*}
P(\text{reads} | A/A) &= 0.00000098 \quad \text{Prior}(A/A) = 0.04 \quad \text{Posterior}(A/A) = <.001 \\
P(\text{reads} | A/C) &= 0.03125 \quad \text{Prior}(A/C) = 0.32 \quad \text{Posterior}(A/C) = 0.999 \\
P(\text{reads} | C/C) &= 0.000097 \quad \text{Prior}(C/C) = 0.64 \quad \text{Posterior}(C/C) = <.001
\end{align*}
\]

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

• **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
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?
Current Genome Scale Approaches

• Deep whole genome sequencing
  – Can only be applied to limited numbers of samples
  – Most complete ascertainment of variation

• Exome capture and targeted sequencing
  – Can be applied to moderate numbers of samples
  – SNPs and indels in the most interesting 1% of the genome

• Low coverage whole genome sequencing
  – Can be applied to moderate numbers of samples
  – Very complete ascertainment of shared variation
  – Less complete ascertainment of rare variants
Current Genome Scale Approaches

• Deep whole genome sequencing
  – Can only be applied to limited numbers of samples
  – Most complete ascertainment of variation

• Exome capture and targeted sequencing
  – Can be applied to moderate numbers of samples
  – SNPs and indels in the most interesting 1% of the genome

• Low coverage whole genome sequencing
  – Can be applied to moderate numbers of samples
  – Very complete ascertainment of shared variation
  – Less complete ascertainment of rare variants

Our Focus For Today
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
Silly Cartoon View of Shot Gun Data

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th></th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>A</td>
<td></td>
<td>C</td>
<td>T</td>
<td></td>
<td></td>
<td>C</td>
<td>T</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td></td>
<td></td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>G</td>
<td></td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>A</td>
<td></td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>A</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Silly Cartoon View of Shot Gun Data
How Do We Update One Pair Of Haplotypes?

• Markov model similar to that for genotype imputation

• To carry out an update, select one individual
  – Let $X_i$ be observed bases overlapping position $i$ for individual

• Assume (temporarily) that current haplotype estimates for all other individuals are correct

• Model haplotypes for individual being updated as mosaic of the other available haplotypes
  – $S_i = (S_{i1}, S_{i2})$ denotes the pair of haplotypes being copied
Markov Model

Model is very similar to the one we previously used for imputation…
Likelihood

\[ L = \sum_{S_1} \sum_{S_2} \ldots \sum_{S_M} P(S_1) \prod_{i=2}^{M} P(S_i \mid S_{i-1}) \prod_{i=1}^{M} P(X_i \mid S_i) \]

- \( P(S_1) = 1 / H^2 \) where \( H \) is the number of template haplotypes
- \( P(S_i \mid S_{i-1}) \) depends on estimated population recombination rate
- \( P(X_i \mid S_i) \) are the genotype likelihoods
Simulation Results: Common Sites

• Detection and genotyping of Sites with MAF >5% (2116 simulated sites/Mb)

  – **Detected Polymorphic Sites: 2x coverage**
    – 100 people 2102 sites/Mb detected
    – 200 people 2115 sites/Mb detected
    – 400 people 2116 sites/Mb detected

  – **Error Rates at Detected Sites: 2x coverage**
    – 100 people 98.5% accurate, 90.6% at hets
    – 200 people 99.6% accurate, 99.4% at hets
    – 400 people 99.8% accurate, 99.7% at hets
Simulation Results: Rarer Sites

- Detection and genotyping of Sites with MAF 1-2% (425 simulated sites/Mb)

  - **Detected Polymorphic Sites: 2x coverage**
    - 100 people  139 sites/Mb detected
    - 200 people  213 sites/Mb detected
    - 400 people  343 sites/Mb detected

  - **Error Rates at Detected Sites: 2x coverage**
    - 100 people  98.6% accurate, 92.9% at hets
    - 200 people  99.4% accurate, 95.0% at hets
    - 400 people  99.6% accurate, 95.9% at hets

Yun Li
That’s The Theory ... Show Me The Data!

Results from 1000 Genomes Project
Project Goals

- >95% of accessible genetic variants with a frequency of >1% in each of multiple continental regions
- Extend discovery effort to lower frequency variants in coding regions of the genome
- Define haplotype structure in the genome
1000 Genomes Pilot Completed

- 2 deeply sequenced trios
- 179 whole genomes sequenced at low coverage
- 8,820 exons deeply sequenced in 697 individuals

- 15M SNPs, 1M indels, 20,000 structural variants
Accuracy of Low Pass Genotypes

Genotype accuracy for rare genotypes is lowest, but definition of rare changes as more samples are sequenced.

Hyun Min Kang
Does Haplotype Information Really Help?

**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 Michigan/EUR 60</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>August 2010 Michigan/EUR 186</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>October 2010 Michigan/EUR 280</td>
<td>13,276,643</td>
<td>50.1</td>
<td>1.8</td>
<td>2.20</td>
<td>97.91**</td>
</tr>
</tbody>
</table>

Accuracy of Low Pass Genotypes Generated by 1000 Genomes Project, When Analyzed Here At the University of Michigan
Some Important Notes

• The Markov model we described is one of several possible models for analysis of low pass data

• Alternative models, based on E-M algorithms or local clustering of individuals into small groups exist

• Currently, the best possible genotypes produced by running multiple methods and generating a consensus across analysis their results.
What Was Optimal Model for Analyzing Pilot Data?

<table>
<thead>
<tr>
<th>1000 Genomes Call Set (CEU)</th>
<th>Homozygous Reference Error</th>
<th>Heterozygote Error</th>
<th>Homozygous Non-Reference Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broad</td>
<td>0.66</td>
<td>4.29</td>
<td>3.80</td>
</tr>
<tr>
<td>Michigan</td>
<td>0.68</td>
<td>3.26</td>
<td>3.06</td>
</tr>
<tr>
<td>Sanger</td>
<td>1.27</td>
<td>3.43</td>
<td>2.60</td>
</tr>
<tr>
<td>Majority Consensus</td>
<td>0.45</td>
<td>2.05</td>
<td>2.21</td>
</tr>
</tbody>
</table>

- Pilot analyzed with different haplotype sharing models
  - Sanger (QCALL), Michigan (MaCH/Thunder), Broad (BEAGLE)
  - Consensus of the three callers clearly bested single callers
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 ($r^2$)</td>
<td>99.8%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Effective Sample Size ($n \cdot r^2$)</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·$r^2$)</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></td>
<td></td>
<td></td>
<td></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
Summary So Far

• 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.

• Intuition for why low pass analyses are attractive for complex disease association studies.
Design A Whole Genome Low Pass Sequencing Study

Gonçalo Abecasis
David Schlessinger
Francesco Cucca
SardiNIA Whole Genome Sequencing

• 6,148 Sardinians from 4 towns in the Lanusei Valley, Sardinia
  – Recruited among population of ~9,841 individuals
  – Sample includes >34,000 relative pairs

• Measured ~100 aging related quantitative traits

• Original plan:
  – Set out to sequence >1,000 individuals at 2x to obtain genomes
  – Genotype all individuals, impute sequences into relatives
Who To Sequence?
Assuming All Individuals Have Been Genotyped

0 Genomes Sequenced, 0 Genomes Analyzed
Who To Sequence?
Assuming All Individuals Have Been Genotyped

3 Genomes Sequenced, 9.5 Genomes Analyzed
Who To Sequence?
Assuming All Individuals Have Been Genotyped

5 Genomes Sequenced, 12.5 Genomes Analyzed
Who To Sequence?
Assuming All Individuals Have Been Genotyped

9 Genomes Sequenced, 17 Genomes Analyzed
Anything to Gain from Sequencing Trios?
Improved Accuracy at Heterozygous Sites

- Sequencing trios improves genotype call accuracy
  - At low coverage ...
  - Smaller gain w/deep coverage

- Leads to similar numbers of detected variants
  - At low coverage ...
  - No gain w/deep coverage

- Improved haplotype accuracy

Wei Chen and Bingshan Li
Assembling Sequences In Sardinia

Sardinian team led by Francesco Cucca, Serena Sanna, Chris Jones
How Is Sequencing Progressing?

• NHGRI estimates of sequencing capacity and cost …
  – Since 2006, for fixed cost …
  – … ~4x increase in sequencing output per year

• In our own hands…
  – Mapped high quality bases
  – March 2010: ~5.0 Gb/lane
  – May 2010: ~7.5 Gb/lane
  – September 2010: ~8.6 Gb/lane
  – January 2011: ~16 Gb/lane
  – Summer 2011: ~35 Gb/lane

• Discovered and genotyped >17M genetic variants so far.

Fabio Busonero, Hyun Min Kang, Bingshan Li
## Accuracy Of Variant Calls

<table>
<thead>
<tr>
<th>Sample Set</th>
<th>Genotype Class</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Homozygous Reference</td>
<td>Heterozygotes</td>
<td>Homozygous Non-Reference</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Analysis Ignoring Relatedness</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>66 Samples</td>
<td></td>
<td>2.1</td>
<td>8.7</td>
<td>3.2</td>
</tr>
<tr>
<td>226 Samples</td>
<td></td>
<td>1.0</td>
<td>5.5</td>
<td>1.9</td>
</tr>
<tr>
<td>508 Samples</td>
<td></td>
<td>0.2</td>
<td>1.3</td>
<td>0.4</td>
</tr>
<tr>
<td><strong>Trio-Aware Analysis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>66 Samples</td>
<td></td>
<td>1.0</td>
<td>5.4</td>
<td>1.5</td>
</tr>
<tr>
<td>226 Samples</td>
<td></td>
<td>0.6</td>
<td>3.6</td>
<td>1.1</td>
</tr>
<tr>
<td>508 Samples</td>
<td></td>
<td>0.2</td>
<td>0.6</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Carlo Sidore, Hyun Min Kang, Serena Sanna
Design

Sequence 1000 individuals @ 2x or greater

“Draft” Genomes for 1000 Individuals

Genotype 6000 individuals with 700,000 SNPs

Haplotypes for 6000 Individuals

Whole Genome Information on 6,000 individuals
Sardinian Haplotypes Are Great For Imputation In Sardinia

<table>
<thead>
<tr>
<th>Reference</th>
<th>Imputation Accuracy ($r^2$) IN SARDINIA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Panel</td>
</tr>
<tr>
<td>1000G (563)</td>
<td>20</td>
</tr>
<tr>
<td>Sardinia (508)</td>
<td>20</td>
</tr>
</tbody>
</table>

Data: Sardinia data set; chr20; Imputation-panel: Affy1M; Evaluation-panel: Metabochip
Sardinian Haplotypes Are Not Great for Imputation Outside Sardinia

<table>
<thead>
<tr>
<th>Reference</th>
<th>Imputation Accuracy ($r^2$) OUTSIDE SARDINIA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Panel Chr MAF 1-3% MAF 3-5% MAF &gt;5%</td>
</tr>
<tr>
<td></td>
<td>1000G Nov (563) 20 0.83 0.85 0.94</td>
</tr>
<tr>
<td></td>
<td>Sardinia (508) 20 0.77 0.83 0.92</td>
</tr>
</tbody>
</table>

Data: GAIN data set; chr20; Imputation-panel: Affy1M; Evaluation-panel: Perlegen Custom Array
What Do We See Genomewide?

LDL Cholesterol

Also By GWAS, LDLR, APOE

Only By Sequencing, Q39X in HBB

Genomic Position

Log_{10} P-value

Also By GWAS, PCSK9, SORT1, APOB
### LDL Genetics In Lanusei, Current Sequenced Based View

<table>
<thead>
<tr>
<th>Locus</th>
<th>Variants</th>
<th>MAF</th>
<th>Effect Size (SD)</th>
<th>$H^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBB</td>
<td>Q39X</td>
<td>.04</td>
<td>0.90</td>
<td>8.0%??</td>
</tr>
<tr>
<td>APOE</td>
<td>R176C, C130R</td>
<td>.04, .07</td>
<td>0.56, 0.26</td>
<td>3.3%</td>
</tr>
<tr>
<td>PCSK9</td>
<td>R46L, rs2479415</td>
<td>.04, .41</td>
<td>0.38, 0.08</td>
<td>1.2%</td>
</tr>
<tr>
<td>LDLR</td>
<td>rs73015013, V578R</td>
<td>.14, .005</td>
<td>0.16, 0.62</td>
<td>1.2%</td>
</tr>
<tr>
<td>SORT1</td>
<td>rs583104</td>
<td>.18</td>
<td>0.15</td>
<td>0.6%</td>
</tr>
<tr>
<td>APOB</td>
<td>rs547235</td>
<td>.19</td>
<td>0.19</td>
<td>0.5%</td>
</tr>
</tbody>
</table>

- Most of these variants are important across Europe, extensively studied.
- Q39X variant in HBB is especially enriched in Sardinia.
- V578R in LDLR is a Sardinia specific variant, particularly common in Lanusei.
Parting Thoughts ...

• Sequencing enables new genetic discoveries

• Achieving sufficient sample sizes is a challenge
  – Take advantage of efficient study designs
  – Take advantage of interesting sample sets

• Many challenges remain in analyzing data
  – At least as tough as generating it!
Recommended Reading


• Le SQ and Durbin R (2010) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Research* (in press)