The Role of Rare Variants in Complex Disease

Gonçalo Abecasis
University of Michigan School of Public Health
## Human Genetics, Sample Sizes over My Time

<table>
<thead>
<tr>
<th>Year</th>
<th>No. of Samples</th>
<th>No. of Markers</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ongoing</td>
<td>~33,000</td>
<td>50 million</td>
<td>Haplotype Reference Consortium (ASHG Talk #176)</td>
</tr>
<tr>
<td>Ongoing</td>
<td>~40,000</td>
<td>12 million</td>
<td>Macular Degeneration Study (ASHG #384/387)</td>
</tr>
<tr>
<td>2012</td>
<td>1,092</td>
<td>40 million</td>
<td>The 1000 Genomes Project (Nature)</td>
</tr>
<tr>
<td>2010</td>
<td>Hundreds</td>
<td>16 million</td>
<td>The 1000 Genomes Project (Nature)</td>
</tr>
<tr>
<td>2010</td>
<td>~100,000</td>
<td>2.5 million</td>
<td>Lipid GWAS (Nature)</td>
</tr>
<tr>
<td>2008</td>
<td>~9,000</td>
<td>2.5 million</td>
<td>Lipid GWAS (Nature Genetics)</td>
</tr>
<tr>
<td>2007</td>
<td>Hundreds</td>
<td>3.1 million</td>
<td>HapMap (Nature)</td>
</tr>
<tr>
<td>2005</td>
<td>Hundreds</td>
<td>1 million</td>
<td>HapMap (Nature)</td>
</tr>
<tr>
<td>2003</td>
<td>Hundreds</td>
<td>10,000</td>
<td>Chr. 19 Variation Map (Nature Genetics)</td>
</tr>
<tr>
<td>2002</td>
<td>Hundreds</td>
<td>1,500</td>
<td>Chr. 22 Variation Map (Nature)</td>
</tr>
<tr>
<td>2001</td>
<td>Thousands</td>
<td>127</td>
<td>Three Region Variation Map (Am J Hum Genet)</td>
</tr>
<tr>
<td>2000</td>
<td>Hundreds</td>
<td>26</td>
<td>T-cell receptor variation (Hum Mol Genet)</td>
</tr>
</tbody>
</table>
Project Goals (2008)

• >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
Pilot Projects (2010)

- 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
Phase I (2012)

- More diverse set of populations sequenced
  - Total >1,092 individuals (EUR, ASN, AFR, AMR groupings)

- >38.5 million SNP
  - 8.5M sites discovered before project (dbSNP 129)
  - 30M sites newly discovered
  - 98.9% of HapMap III sites rediscovered
  - Transition/transversion ratio of 2.16 vs 2.04 in pilot

- ~1.5M insertion deletion polymorphisms

- [ftp://ftp.1000genomes.ebi.ac.uk](ftp://ftp.1000genomes.ebi.ac.uk)

The 1000 Genomes Project (Nature, 2012)
Samples in the final phase

Bubble size = sample size
1000 Genomes data generation

Total Dataset: 84 TB of BAM Files

Data Generation Complete: May 2013
Contribution of the 1000G to dbSNP

- 61.9 million novel variants discovered by 1000G (62% of dbSNP)
- 22.4 million variants validated by 1000G (58.4% of non-1000G variants)
## Variants per genome

<table>
<thead>
<tr>
<th>Type</th>
<th>Variant sites / genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPs</td>
<td>$3.8 \times 10^6$</td>
</tr>
<tr>
<td>Indels</td>
<td>$5.7 \times 10^5$</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>
Quality Control of Short Variants

- For short variants, the high coverage PCR-free data from 26 individuals was used to assess the false discovery rate for each variant type.

- An allele is considered ‘validated’ if multiple supporting reads can be identified in PCR-free data.

- Sites included in the Phase 3 haplotypes have been selected to control the allele False Discovery Rate at 5%.

<table>
<thead>
<tr>
<th>Haplotype scaffold</th>
<th>MVNcall variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variant Type</td>
<td>Bi-allelic SNPs</td>
</tr>
<tr>
<td>Per-allele FDR</td>
<td>4.07%</td>
</tr>
</tbody>
</table>
Verification & further characterization of inversions by PacBio sequencing

Regular ("simple") inversion

Inversion with flanking deletion

Complex SVs with inverted sequences
Private vs. Shared Variation (Individual View)

Variants per Genome

5 million

160,000

Variants per Genome (Zoom)
Population histories
Biases in Variation Databases?

ClinVar $r^2 = 0.4059$
Optimal Model for Analyzing 1000 Genomes?

<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>
</tbody>
</table>

- Michigan caller combines ...
  - Markov models to identify shared haplotypes,
  - Classifiers to distinguish true variants from error,
  - Strategies to distribute computation across cluster
Optimal Model for Analyzing 1000 Genomes?

<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>

• Common to see “ensemble” methods outperform the best single method
Current 1000 Genomes Analysis Pipeline

Raw data
- Genotyping arrays
- Low Coverage and Exome Read Data

24 initial callsets
- Callset 1
- Callset 2
- Callset 3
- Callset 4
- Callset 5
- Callset n

Consensus callsets
- SNPs and high confidence indels
- Multi-allelic SNPs, indels, and MNPs
- Structural Variants
- Short Tandem Repeats

Integration
- Phasing
- Phasing of multi-allelic variants onto haplotype scaffold
- Quality assessment and filtering

Final callset
- PCR-free data

Integrated Haplotypes

10 SNP/INDEL callsets, 2 STR callsets, 12 SV callsets
Imputation Accuracy

TODO: Multiallelic SNPs and indels to be renamed
AMD Imputation Example #1
Imputation Example #2

- del443ins54

- PLEKHA1
- ARMS2
- MIR3941
- HTRA1

Position on chr10 (Mb)
1000 G: Parting Thoughts

• Variation is extremely rare
  • In any one genome, nearly all variation is shared ...
  • But almost all variants are unique to a population or continent

• Great benefits to integrated analyses
  • But analyses still requires time comparable to data generation

• Major improvements in genome coverage, variant quality and integration

• Advances can be transferred to disease studies through imputation
Current State of Genetic Association Studies

- Surveying common variation across 10,000s - 100,000s of individuals is now routine, using genotyping arrays

- Many common alleles have been associated with a variety of human complex traits

- The functional consequences of these alleles are often subtle, and translating the results into mechanistic insights remains challenging

- Sequencing studies are starting to allow studies to extend to rare variants, which can lead to easier to understand biology
Current Challenges and Opportunities

• The major challenge for common disease genetics is translating the large number of association signals into biology.

• Studies of rare variants with clear functional outcome provide a systematic approach for advancing human genetics.

• Will require collaboration between clinical experts, biologists, geneticists.
  • Ensure that we focus on the most important outcomes.
  • Ensure that efficient and powerful study designs are used.
  • Ensure that we translate findings into biological insights.
Whole Genome Study in Sardinia

Gonçalo Abecasis
David Schlessinger
Francesco Cucca
Lanusei, Ilbono, and Elini viewed from Arzana
SardiNIA Whole Genome Sequencing

• 6,148 Sardinians from 4 towns in the Lanusei Valley, Sardinia, Italy
  • Recruited among population of ~9,841 individuals
  • Sample includes many close relatives (siblings, cousins, etc.)

• Participants have all been measured for ~100 cardiovascular and blood traits, here we focus on LDL-cholesterol

• The experiment
  • Genotype all individuals so we can identify shared haplotypes
  • Sequence ~2,000 selected individuals at 4x to obtain draft whole genomes
  • Propagate information from sequenced individuals to other shared haplotypes
Who To Sequence?
Assuming All Individuals Have Been Genotyped

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

9 Genomes Sequenced, 17 Genomes Analyzed
Our analysis examines all sequence information jointly; As more samples are sequenced, accuracy increases.
Results of Sequence Analysis

• 17.6 M discovered variants (48% newly discovered)

• 172,997 variants (0.98%) overlap protein coding sequences
  • 84,312 non-synonymous variants (59% newly discovered)
  • 2,504 variants in essential splice sites (53% newly discovered)
  • 2,013 variants introduce a stop codon (70% newly discovered)

• Half of the variants we see not observed (or studied!) anywhere else…
  • … this fraction is even higher for variants that change protein sequences.
Sardinian variants appear more deleterious

- Used CADD scores to assess deleteriousness of Sardinia specific variants
  - Combines conservation and structural modeling.
  - Average variant has a score of 0.
  - 2.5% of variants have scores >2.

- General patterns:
  - Coding variants are more deleterious.
  - Rare variants are also more deleterious.
  - Sardinian specific variants are more deleterious.
What Do We See Genomewide?
LDL Cholesterol

Also By GWAS, LDLR, APOE

Also By GWAS, PCSK9, SORT1, APOB

Only By Sequencing, Q39X in HBB

Genomic Position
### LDL Genetics In Lanusei Valley, Sardinia, Current Sequenced Based View

<table>
<thead>
<tr>
<th>Locus</th>
<th>Variants</th>
<th>MAF</th>
<th>Effect Size (SD)</th>
<th>H²</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.
Our island specific panel increased imputation accuracy ...
Rare variant imputation in all of Europe?

• We combined information from ~33,000 sequenced human genomes
  • Through collaboration with 20 large ongoing complex disease studies
  • This includes ~50 million variants seen in 5+ individuals

• Generating the largest panel of sequenced haplotypes across Europe
  • First version should be complete in Fall 2014
  • Will enable systematic rare variant imputation, perhaps as good as Sardinia?

• Haplotype Reference Consortium,
  • with Jonathan Marchini, Richard Durbin, Goncalo Abecasis
  • http://imputationserver.sph.umich.edu/
  • http://haplotype-reference-consortium.org/
Imputation Accuracy using Haplotype Consortium: Preliminary Results

http://www.haplotype-reference-consortium.org
The HRC Panel – POPRES data

Per Sample accuracy using HRC Panel
Notes ...

• Demonstrated that, in Sardinia, loss-of-function variants in HBB gene greatly reduce LDL-cholesterol levels.
  • Potentially, through increased turnover of red blood cells.

• Creative uses of sequencing technology enabled us to sequence the genomes of thousands of individuals in a cost effective manner...
  • Much of the variation we discovered was population specific.

• We were able to further increase sample size through imputation...
  • Upcoming resources, like the Haplotype Reference Consortium panel, will enable improved rare variant imputation across much of Europe.
The secret of success ...
Acknowledgements

Thank you to the National Institutes of Health (NEI, NHGRI, NHLBI), GlaxoSmithKline and the University of Michigan for funding our work.

Key thanks:

**Sardinia Sequencing:**
- Carlo Sidore
- Serena Sanna
- Fabio Busonero
- Andrea Maschio

**Haplotype Consortium:**
- Sayantan Das
- HRC Collaborators

**AMD Sequencing:**
- Chaolong Wang
- Xiaowei Zhan

**AMD Genotyping:**
- Lars Fritsche
- IAMDGC Consortium