Dissecting the Genetics of Complex Phenotypes

Gonçalo Abecasis

University of Michigan School of Public Health

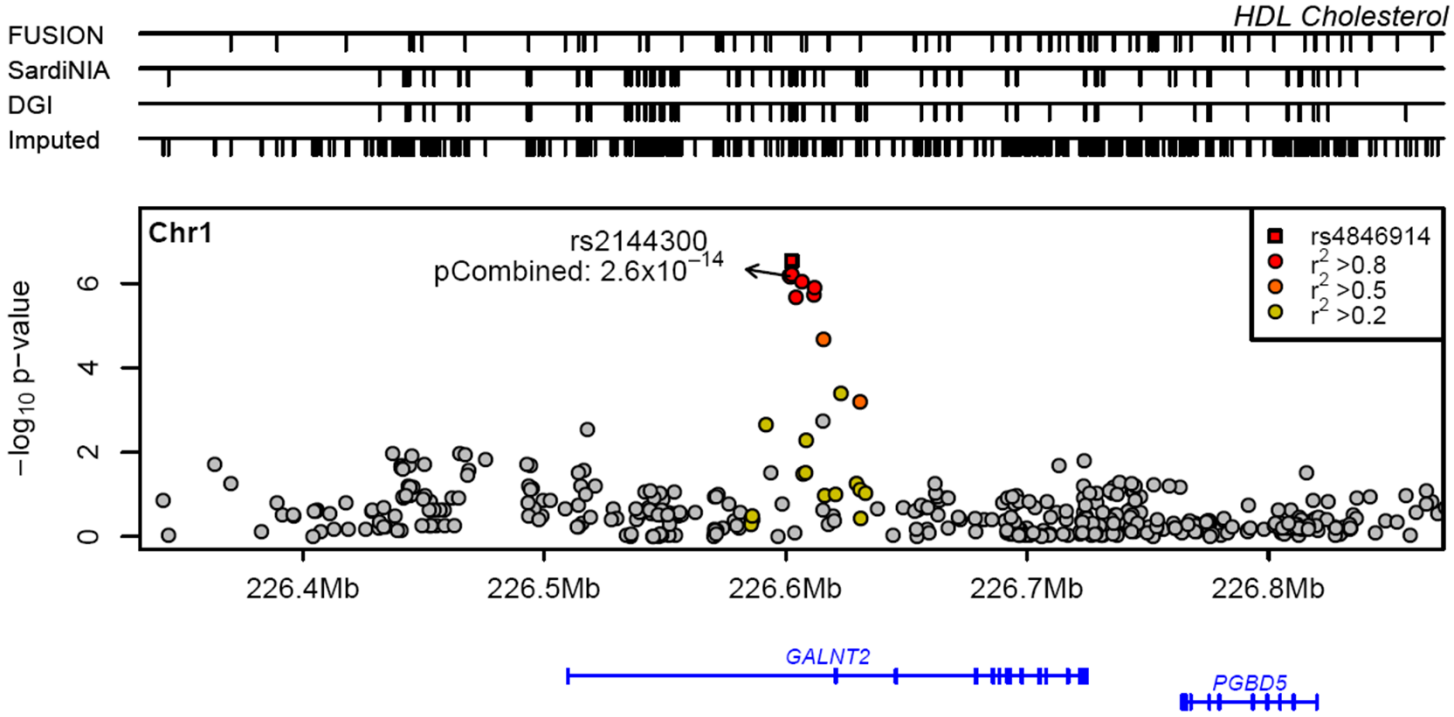
Current Association Studies

- Survey 100,000 – 1,000,000 SNPs in large samples
  - These markers represent a subset of ~10,000,000 common SNPs
- 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
  - Five years ago, 3 genetic loci clearly associated with type 2 diabetes
  - Currently, ~50 genetic loci associated with type 2 diabetes
- Genomewide approach reveals (unexpected) connections between traits

Examplar Genome Wide Study: 100,000 Sample GWAS for Blood Lipids

- 41 studies contributing a total of 100,151 samples of European ancestry
- Fasting lipid concentrations
  - Total cholesterol
  - Low-density lipoprotein (LDL) cholesterol
  - High-density lipoprotein (HDL) cholesterol
  - Triglycerides
- Individuals on lipid-lowering meds excluded
- ~2.6 million SNPs (typed or imputed), MAF ≥ 1%
- 95 loci associated with blood lipid levels

GWAS Leads to Key Biological Switches



- 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%

Teslovich et al, *Nature*, 2010

Can Rare Variants Replace Model Systems? Example from Type 1 Diabetes

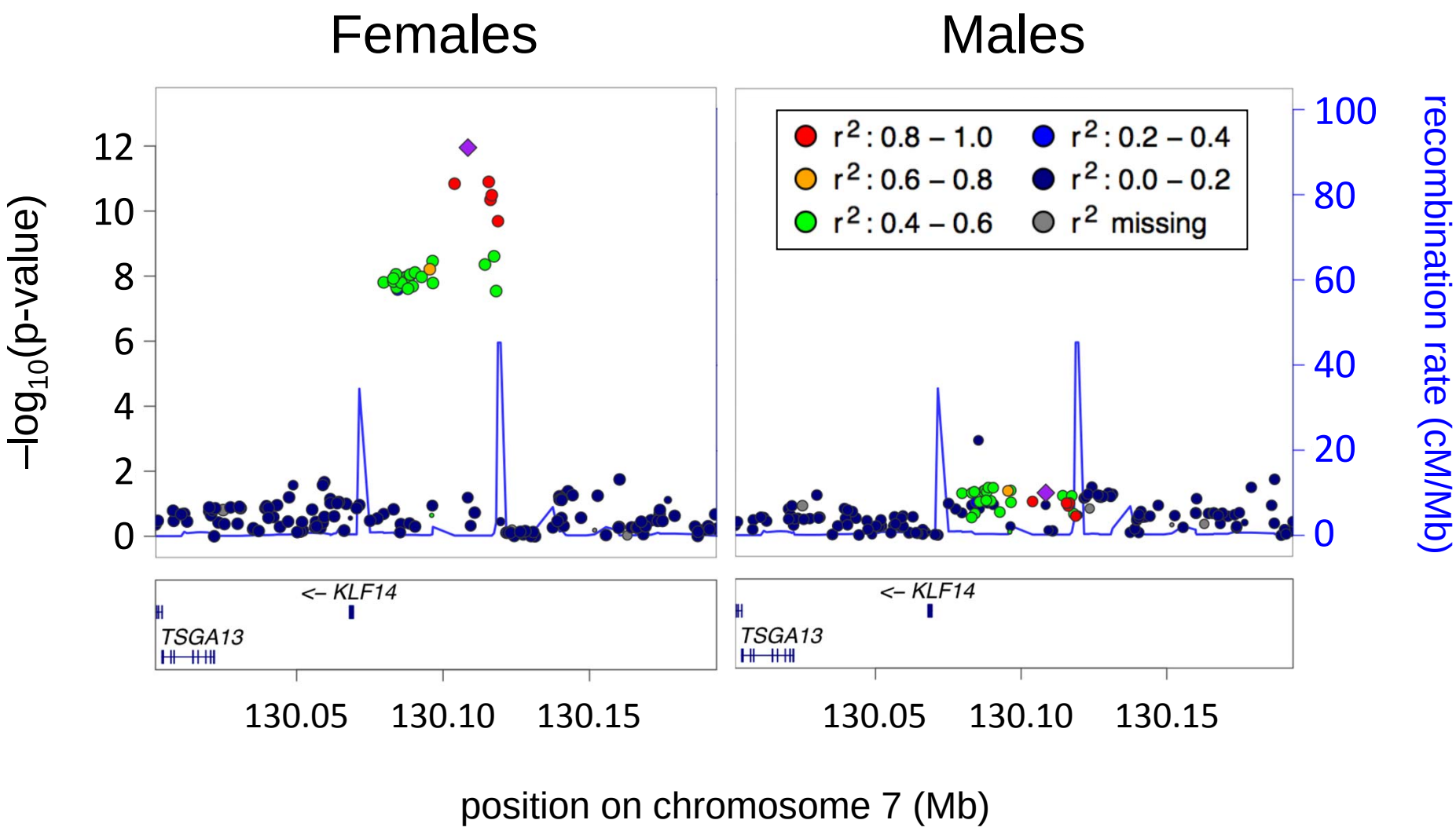
- Nejentsev, Walker, Riches, Egholm, Todd (2009)
  IFIH1, gene implicated in anti-viral responses, protects against T1D
  Science ***324:****387-389*

- Common variants in IFIH1 previously associated with type 1 diabetes

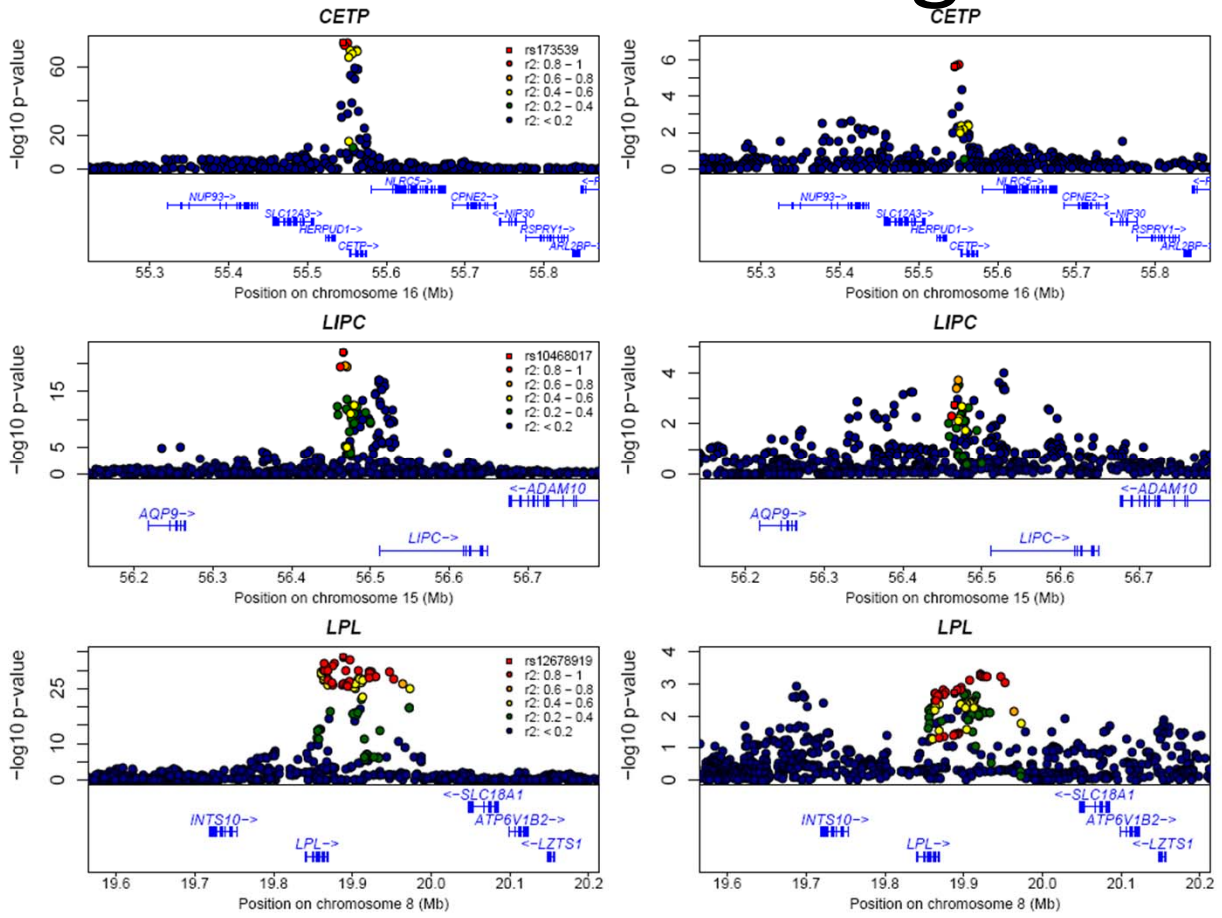
- Sequenced IFIH1 in ~480 cases and ~480 controls
- Followed-up of identified variants in >30,000 individuals

- 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

Triglyceride association: *KLF14*
Sex-specific effect



Overlapping Susceptibility Loci for HDL-C and Macular Degeneration



Top Three HDL Associated Loci All Appear Associated With Macular Degeneration

Probability of Equal or Strong Overlap is $P < 6 \times 10^{-9}$

Chen et al, PNAS, 2010

Some Lessons About Study Design from GWAS

- Ideal to study well phenotyped samples, not just case control collections
- Combining information across samples adds great value
- Results of association study are only a beginning point for much additional research
- The genetics doesn’t always lead where you expect

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 Can We Hope For?
Hints From a Smaller Experiment…

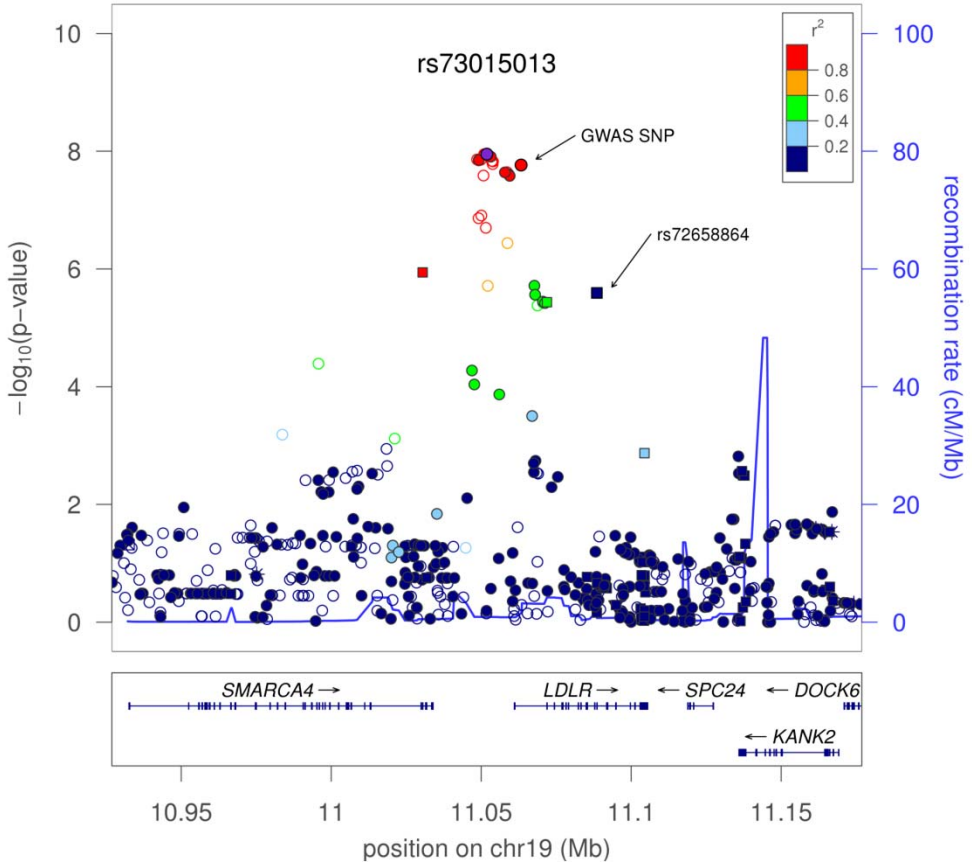
- Re-sequenced coding regions of LDL-C associated genes in first generation GWAS
  - *APOB, APOC1, APOC2, APOE, LDLR, PCSK9, SORT1, B3GALT4*, B4GALT4**

- 256 SardiNIA samples at extremes of LDL-C
  - Also sequenced 120 HapMap CEU, YRI founders
  - Required sequenced samples to be unrelated
  - Low samples all had LDL-C < 102 mg/dl
  - High samples all had LDL-C > 186 mg/dl

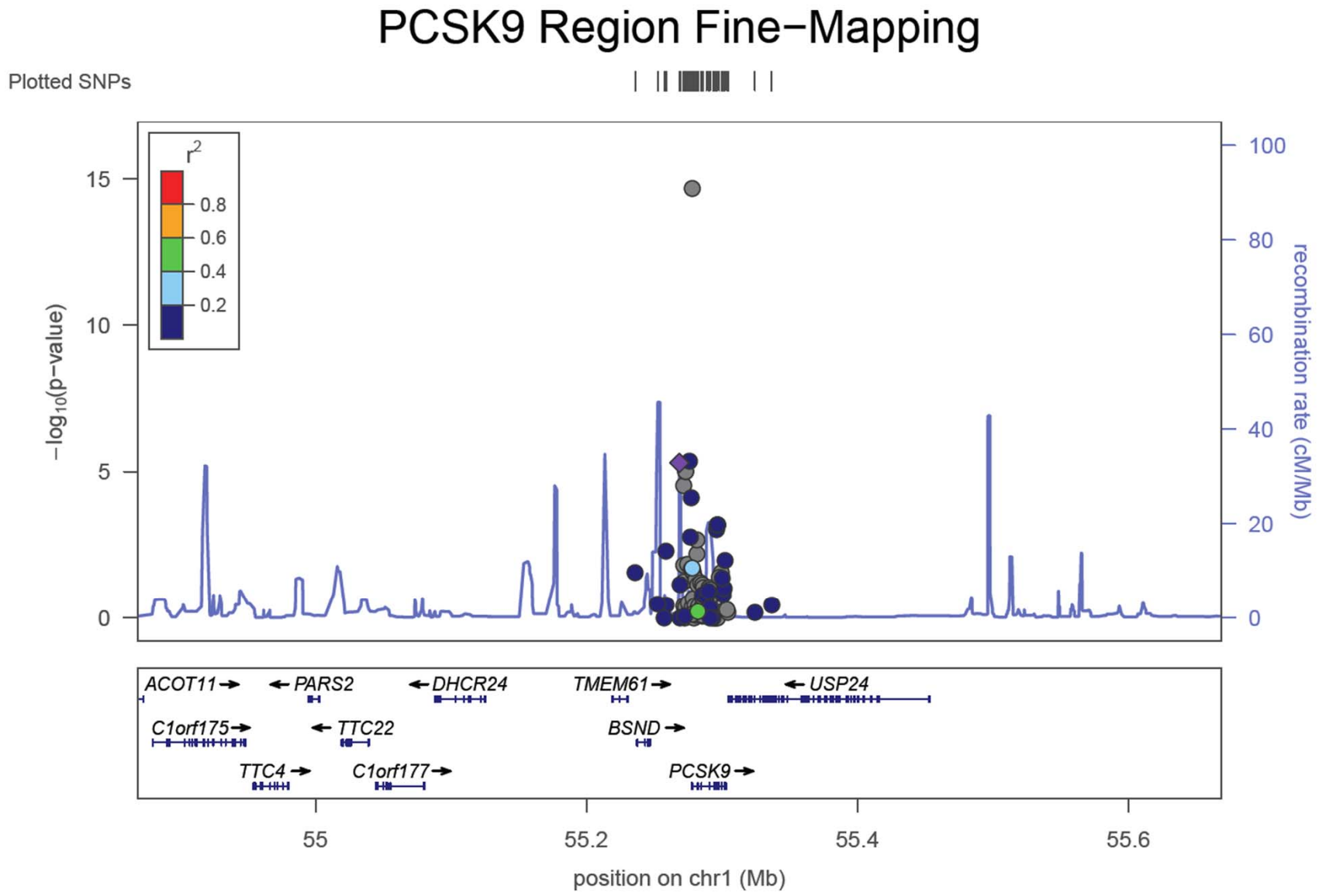
- Discovered 783 SNPs
  - Sequencing done at University of Washington
  - Thanks, Debbie Nickerson and Mark Rieder!
  - Thanks, NHLBI RS&G program!

- Discovered SNPs, 1000G SNPs genotyped in full Sardinia cohort

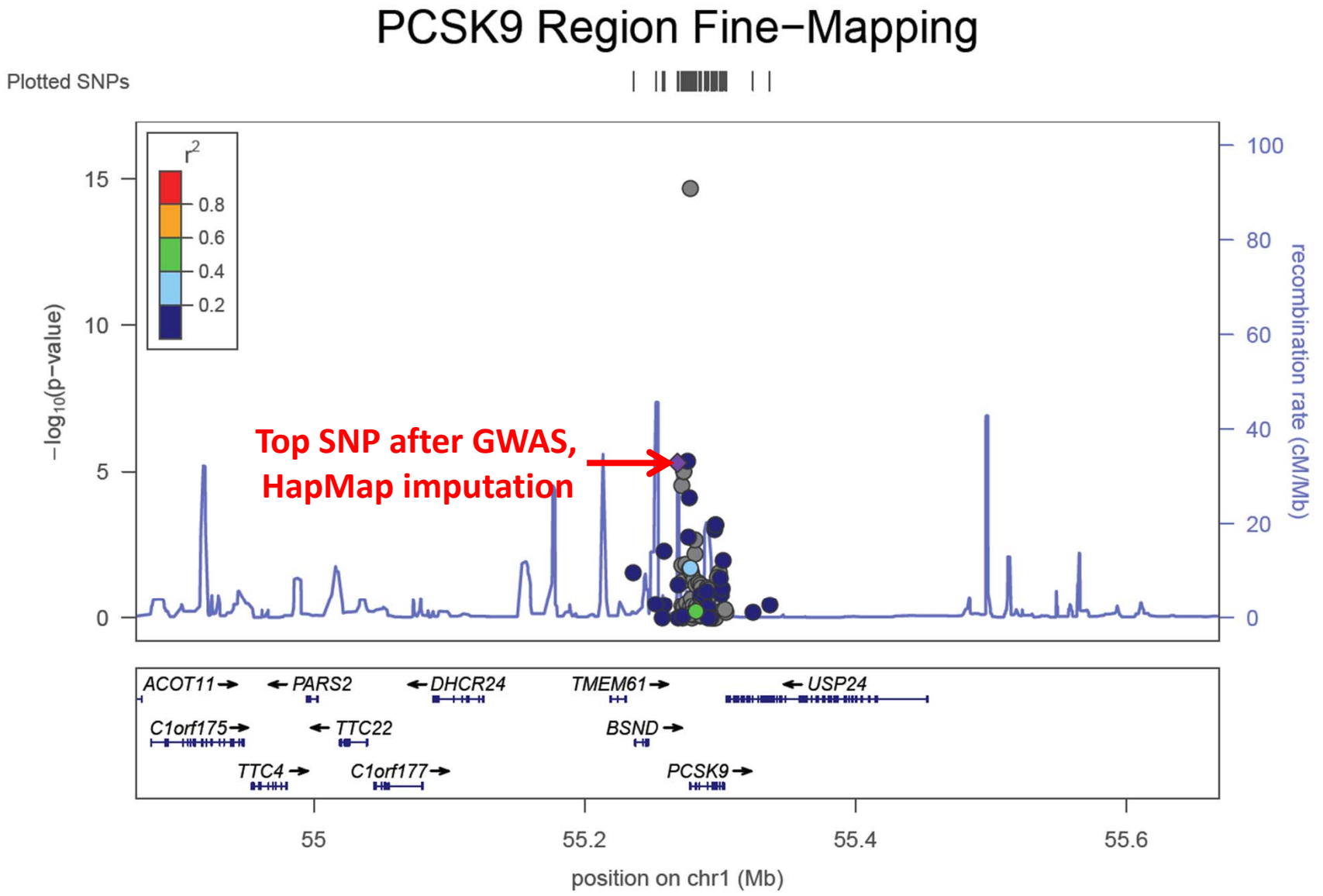
LDLR Region Fine-Mapping



Serena Sanna

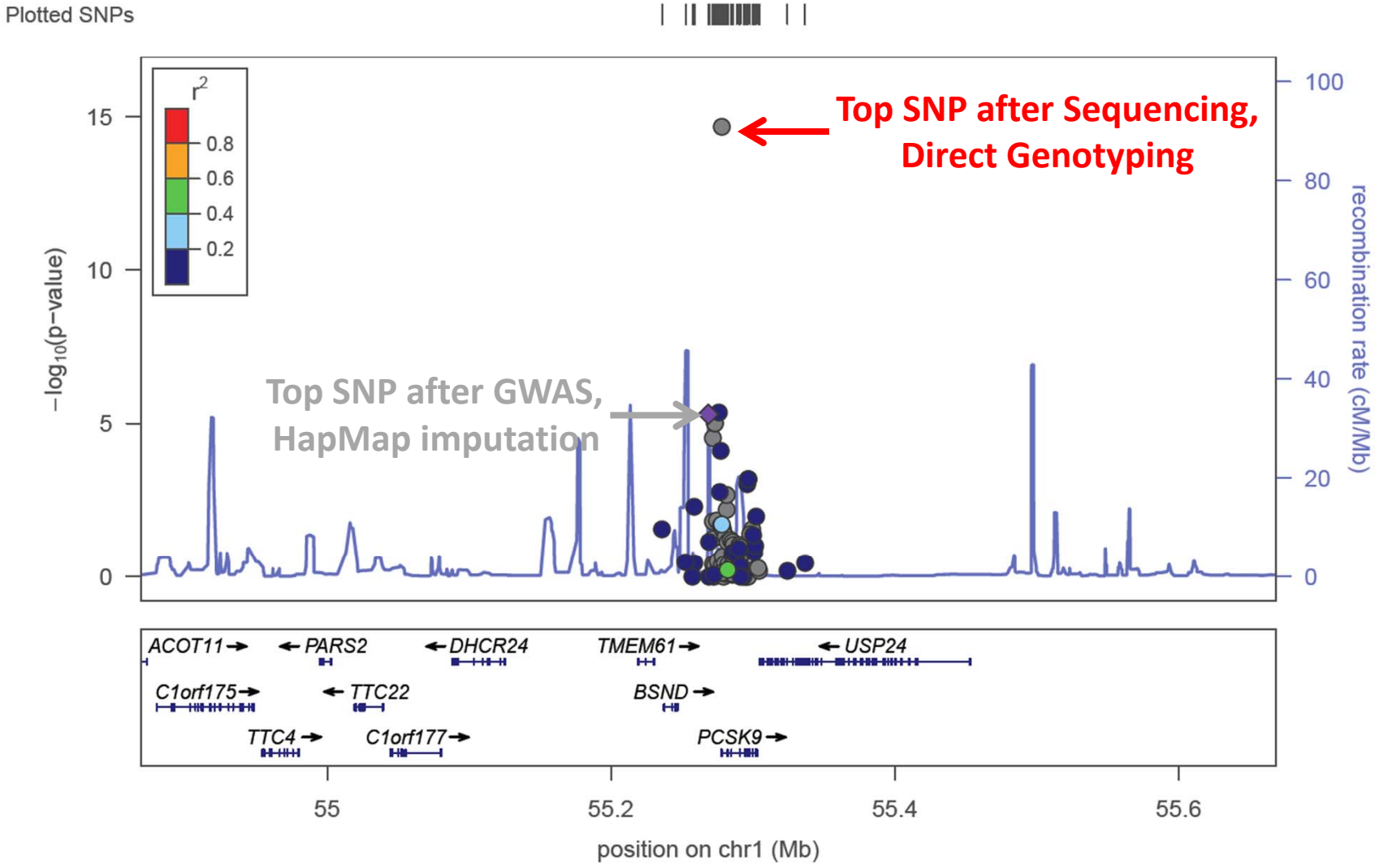


Serena Sanna



Serena Sanna

PCSK9 Region Fine−Mapping



Serena Sanna

Scorecard

- Top signal was in strong LD with GWAS variant
  - *SORT1 (r^2 = 0.90), APOB (r^2 = 0.98), LDLR (r^2 = 1.00)*
- Top signal was distinct from GWAS variant
  - *PCSK9* (variant explains GWAS hit, *r^2 = 0.10)*
  - *APOC1/C2/APOE* (variant independent of GWAS hit, *r^2 = 0.00)*
- Loci with multiple associated variants
  - *PCSK9*
  - *APOB*
  - *LDLR* (second variant appears to be specific to Sardinia)
  - *APOC1/C2/E4*
- Bottom line for LDL-C:
  - GWAS variants (1 per locus) together explain 3.1% of variance
  - Fine-mapping variants (1 or 2 per locus) together explain 6.9% of variance

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?**

How To Use Sequencing Capacity in Complex Trait Studies?

Shallow or Deep?

Families or Unrelateds?

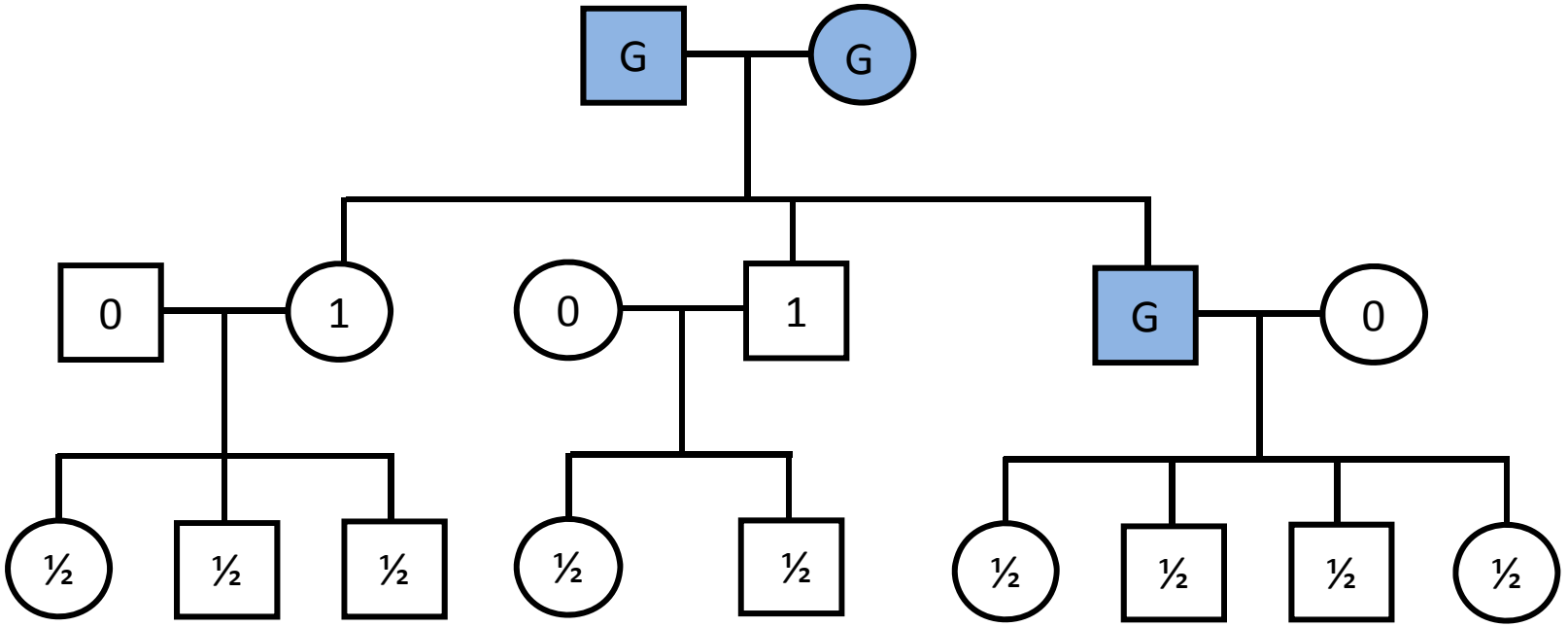
What Sort of Population Samples?

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

Who To Sequence?

Assuming All Individuals Have Been Genotyped

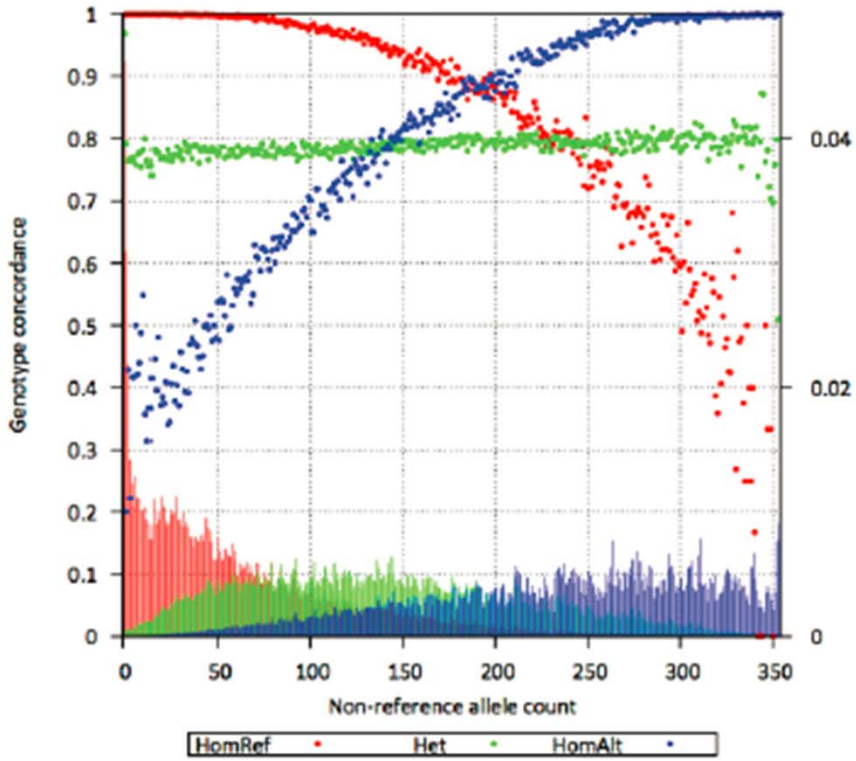


3 Genomes Sequenced, 9.5 Genomes Analyzed

Does Haplotype Information Really Help?

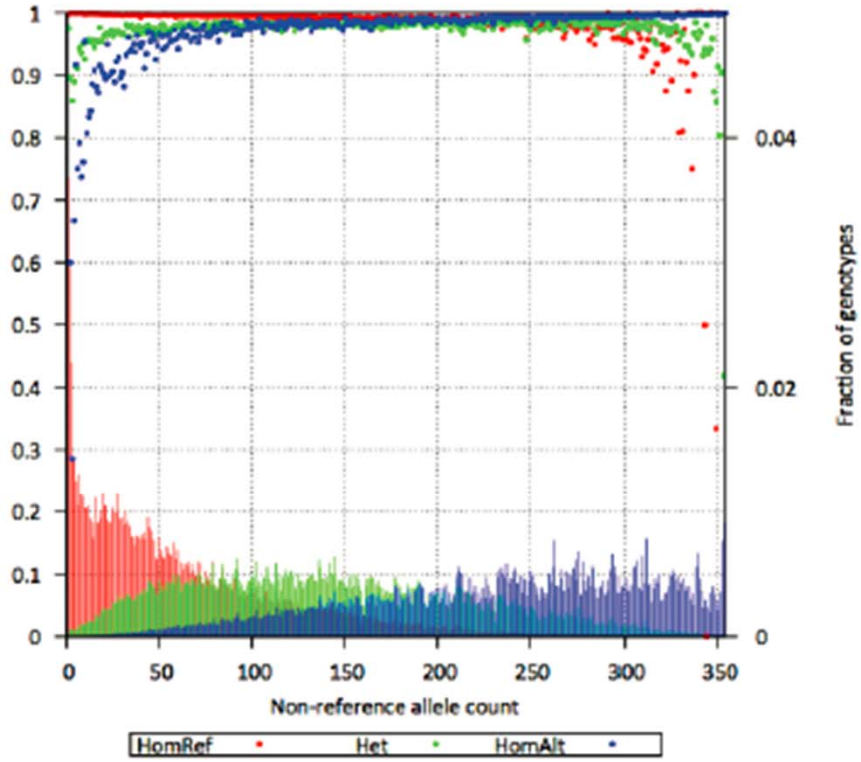
Single Site Analysis

– 21.4% HET errors



Haplotype Aware Analysis

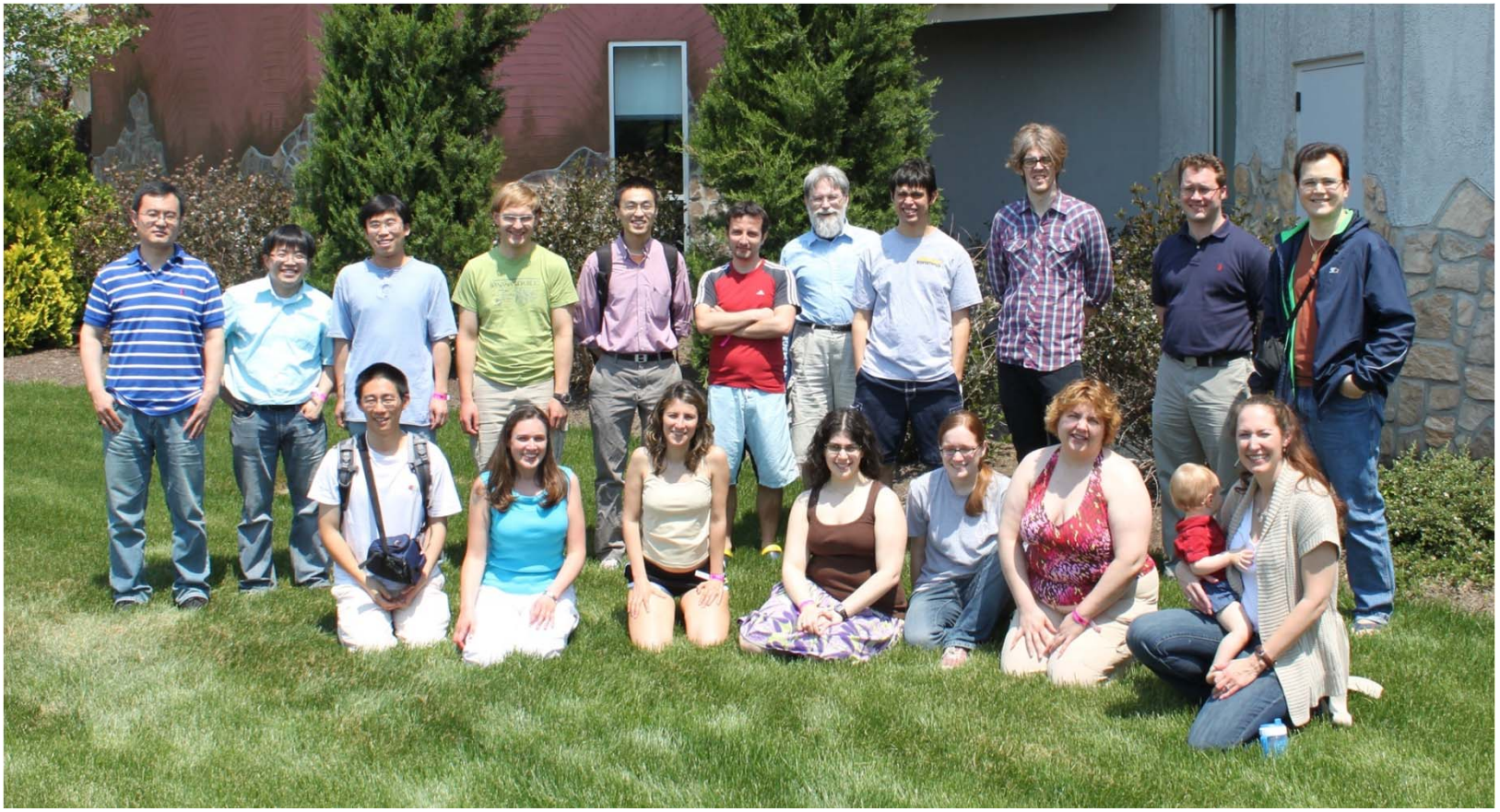
– 2.0% HET errors



Challenges

- Sequencing provides much greater detail
  - Remains relatively slow
  - High 1000s of genotyped samples / month / tech
  - Low 100s of sequenced samples / month / tech
- Success of complex trait studies is greatly limited by sample size
  - Small studies rarely lead to new findings
- Genotyping remains useful complement to sequencing assays

Acknowledgements



Sardinia Study Collaborators led by David Schlessinger, Francesco Cucca, Manuela Uda