From Sequence to Ancestry…

Rare Variant Meta-Analysis…

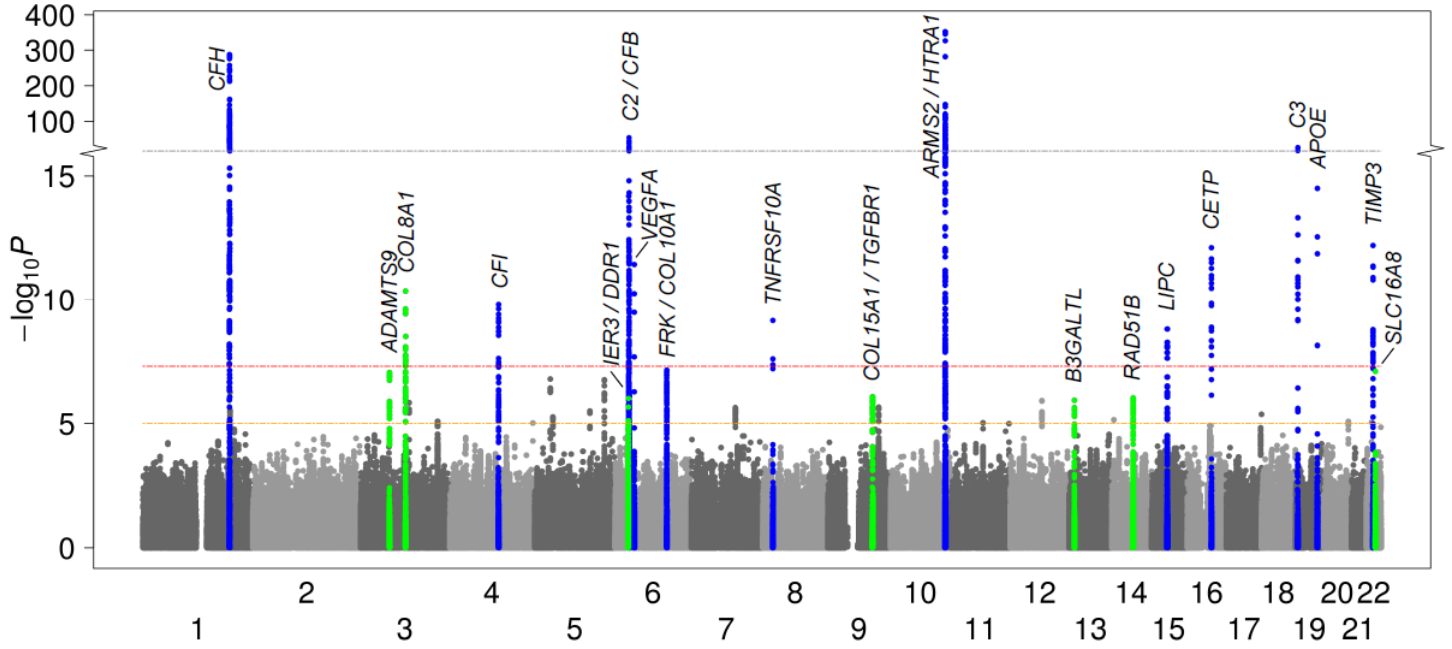
Gonçalo Abecasis

University of Michigan School of Public Health

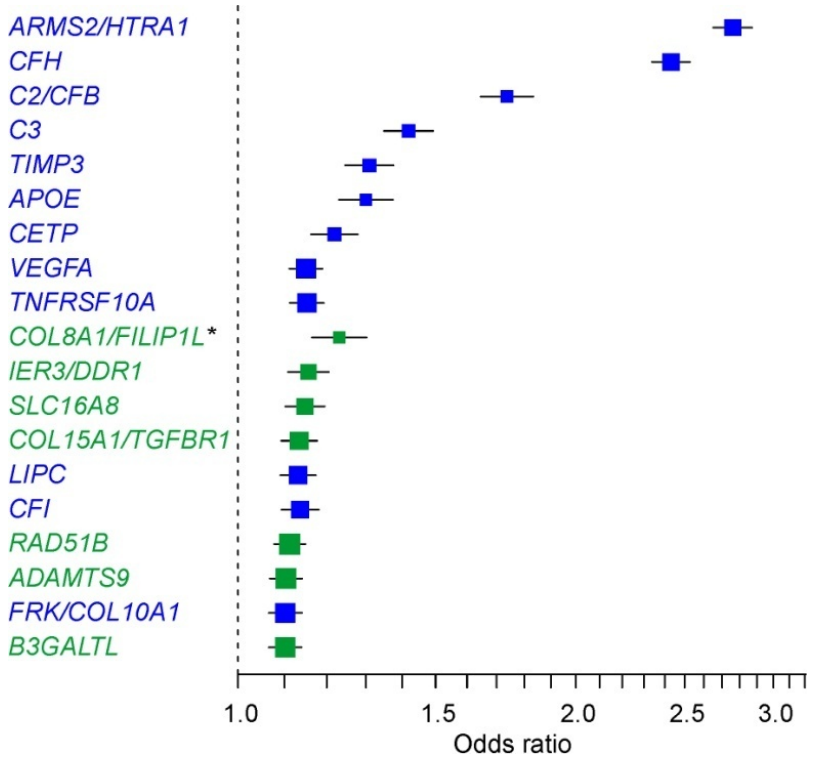
Estimates of Genetic Ancestry from Tiny Bits of Sequence Data

Age Related Macular Degeneration

- One of the first diseases with successful GWAS
  - Robustly associated loci
  - Insights into disease biology
- Largest studies now include 17,000 cases and 60,000 controls

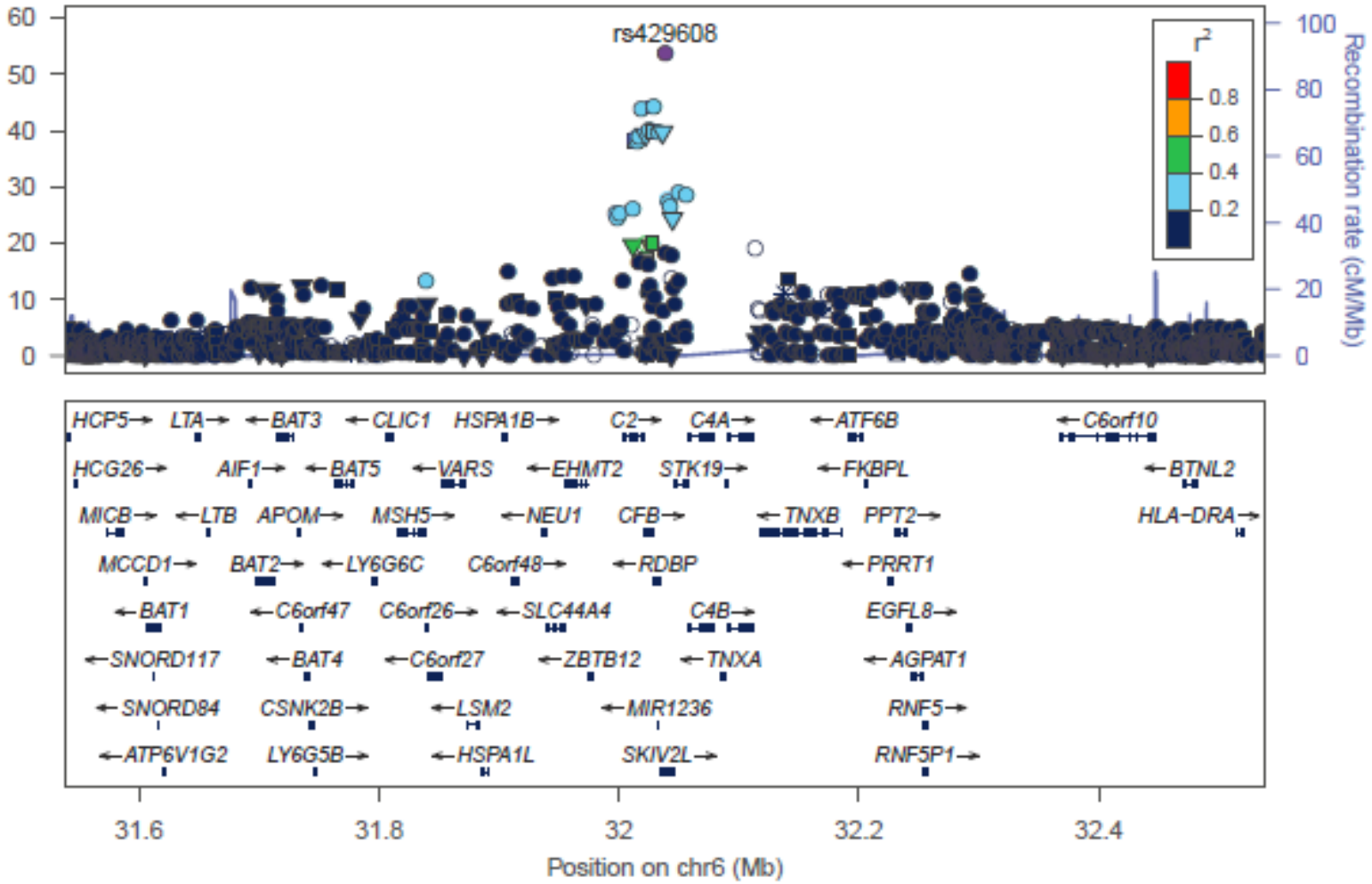


Common AMD Risk Alleles



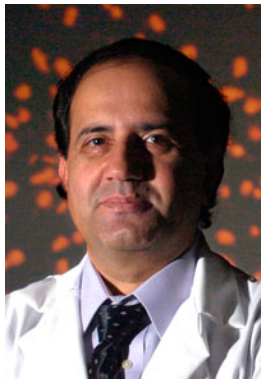
- Common alleles in 19 loci
- Pathways include excess of genes involved in:
  - Complement pathway
  - Angiogenesis
  - Lipid metabolism (HDL)
- 15-65% of genetic variance

Age Related Macular Degeneration:
Close-Up of Specific Region



Mingyao Li

Our First Detailed Look at CFH



Anand Swaroop

- Li et al (2006) *Nature Genetics* **38:**1049-1054.
- Examined 84 genetic variants near CFH.
- Found:
  - 2 common risk haplotypes (one without Y402H)
  - 2 common protective haplotypes
  - Rare haplotypes associated with disease risk

Rare Variants in CFH

- Raychauduri et al (2011) *Nature Genetics* **43:**1232-36

- Sequenced representatives of each haplotype
- Focused on carriers of a rare, high-risk haplotype
  - Frequency ~0.0004 in controls, ~ 0.007 in cases

- Showed R1210C variant strongly associated with AMD
  - Present in 40 of 2,423 cases
  - Present in 1 of 1,123 controls
  - Variant compromises CFH's C-terminal ligand binding

Targeted Sequencing of AMD Risk Loci

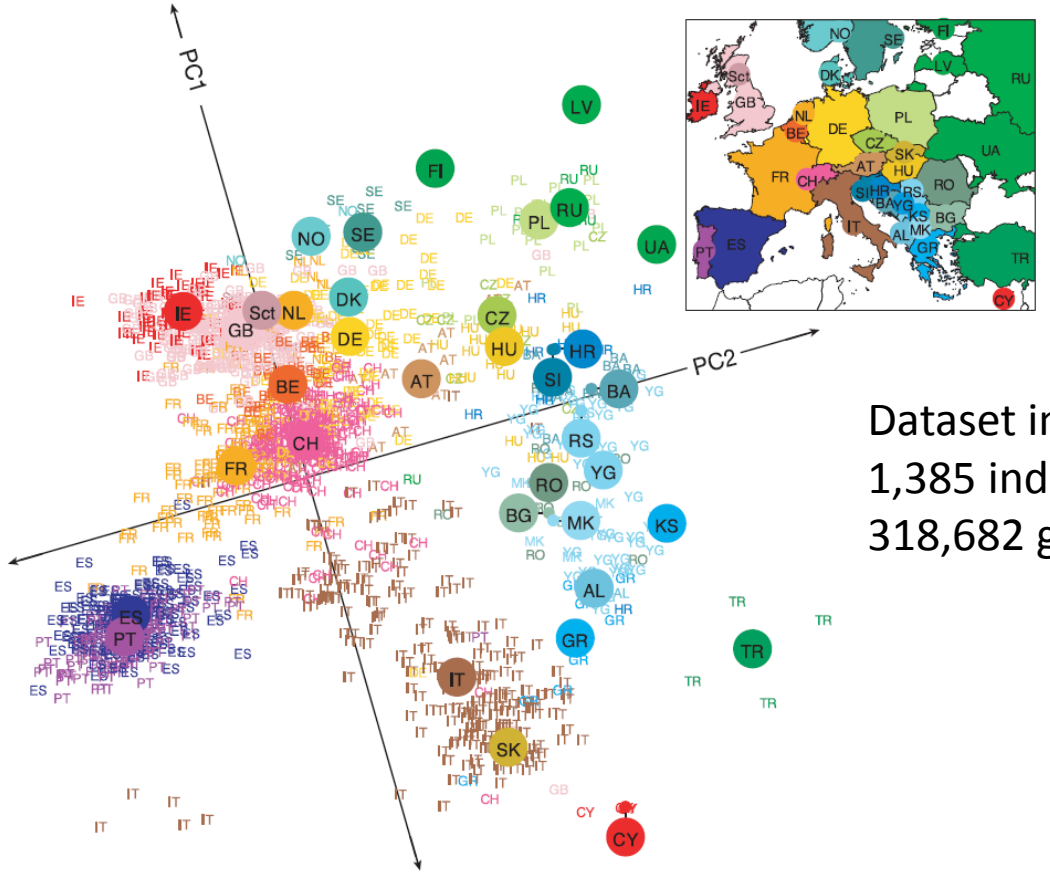
- Re-sequence GWAS loci
  - Search for additional high-risk variants that provide information about function
- Sequenced 2,348 AMD cases and 789 controls
  - Sequencing at **Washington University Genome Center**
  - R1210C variant seen in 23 cases, 0 controls (good!)
  - P-value is about .008 (middling!)
  - Variant present 2 of 12,000+ sequenced exomes (amazing!)
- Studying rare variants, requires very large sample sizes!

Expanding Our Experiment

- Can we identify additional well matched controls to augment our sequencing?

- Plan:
  - Place AMD samples in ancestry map of the world
  - Place other sequenced samples in the same map
  - Identify matched controls for each case …

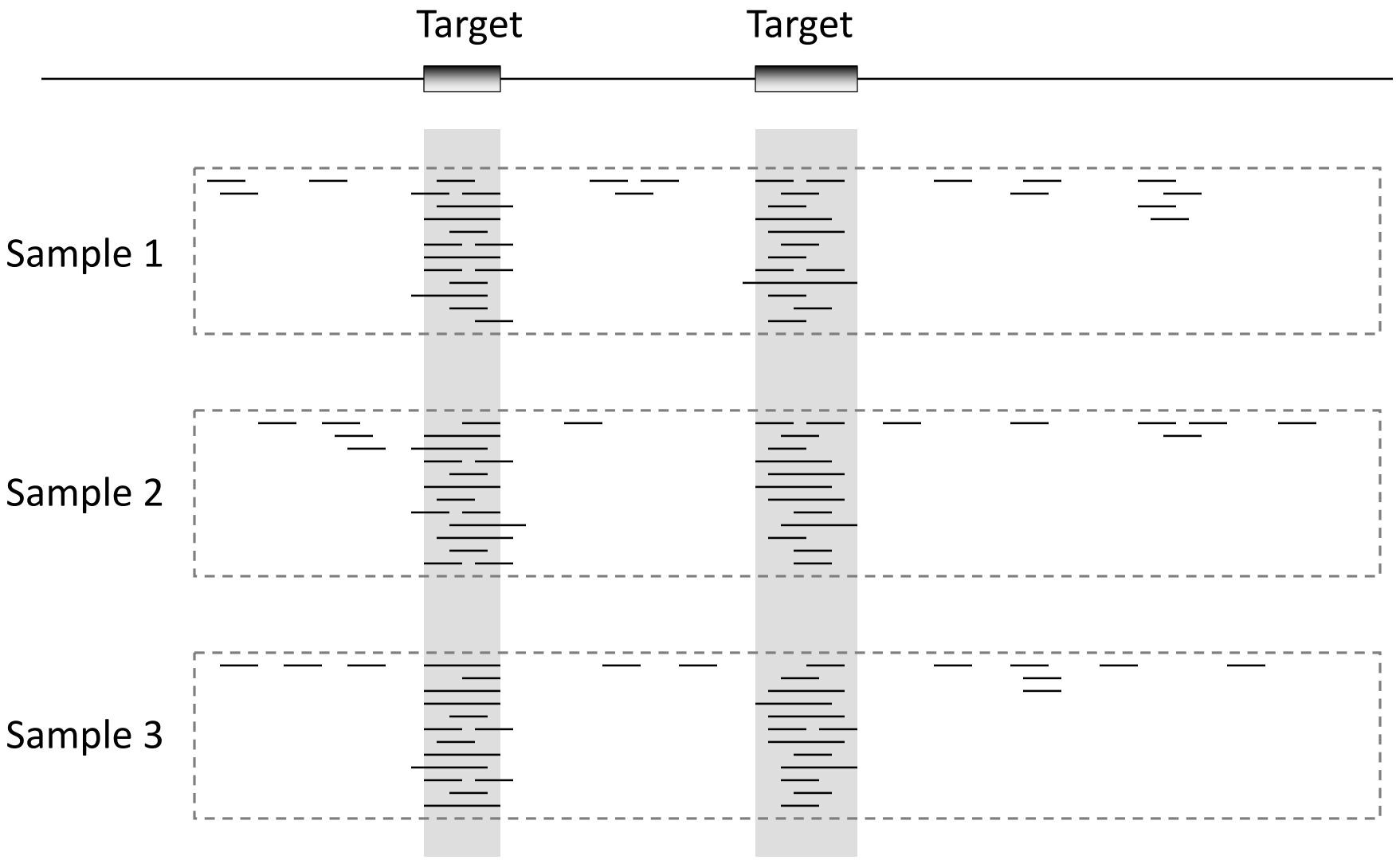
Principal Component Ancestry Map of Europe



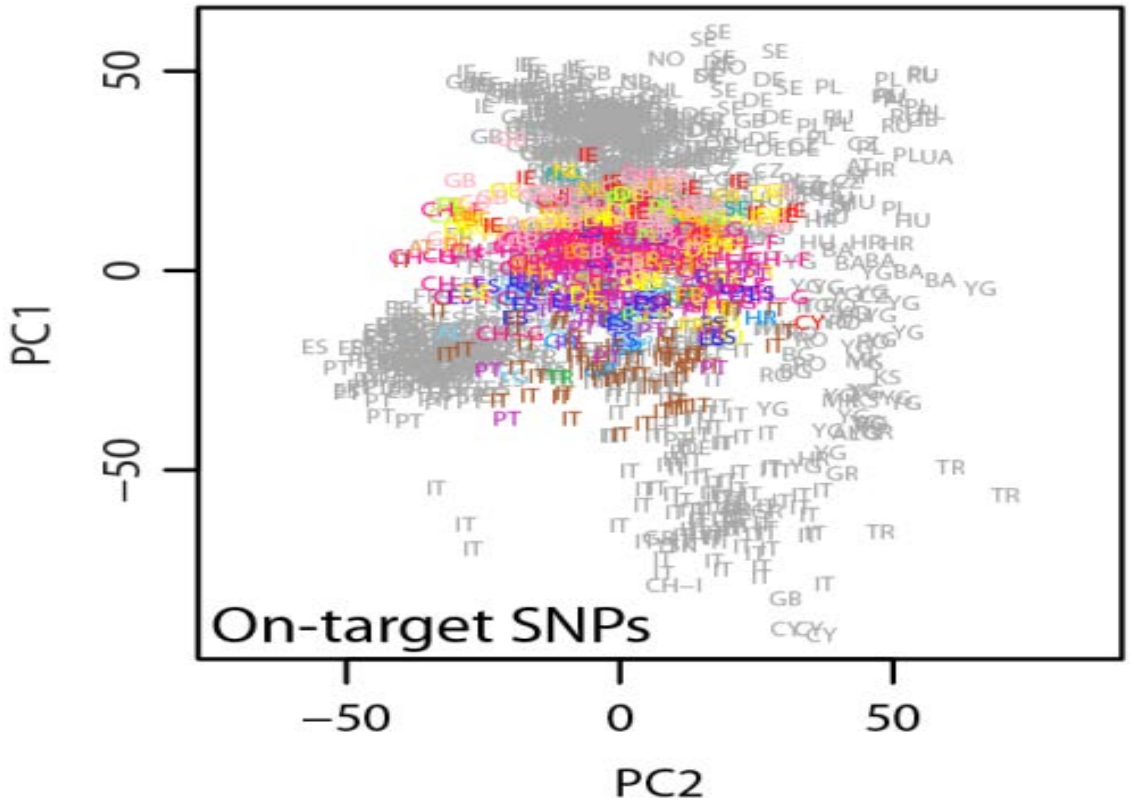
Dataset includes:
1,385 individuals of known ancestry
318,682 genetic markers passing filters

Novembre *et al.* (2008) *Nature*

Targeted sequencing data



What Happens When We Apply PCA Analysis to Targeted Sequence Data?



On-target genotypes don’t contain enough information to estimate the ancestry of a sample. The illustration is based on >80x deep whole exome data.

The Problem

How to place individuals on world wide ancestry map with very little sequence data?

Xiaowei Zhan

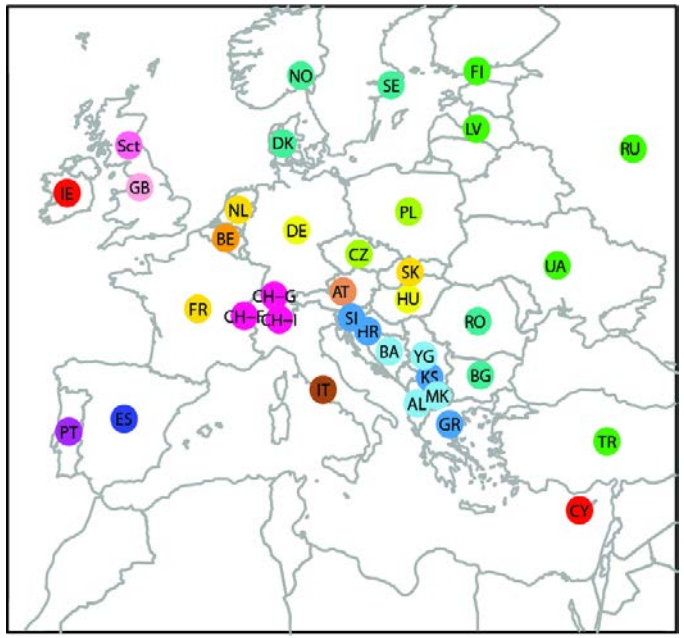
Chaolong Wang

Sebastian Zöllner

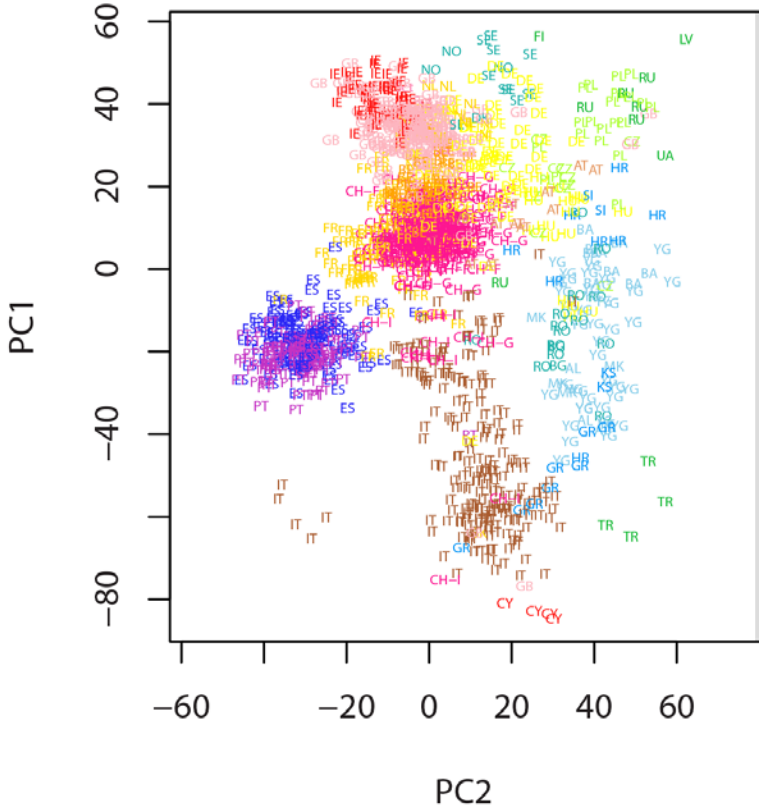
Step 1: Create Reference Map

Generate a reference map by applying PCA to SNP data for N reference individuals. (Map 0)

Geographic map



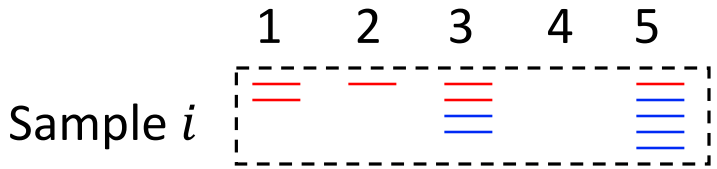
Map 0



37 populations, 1,385 individuals
318,682 autosomal SNPs
Novembre *et al.* (2008) *Nature*

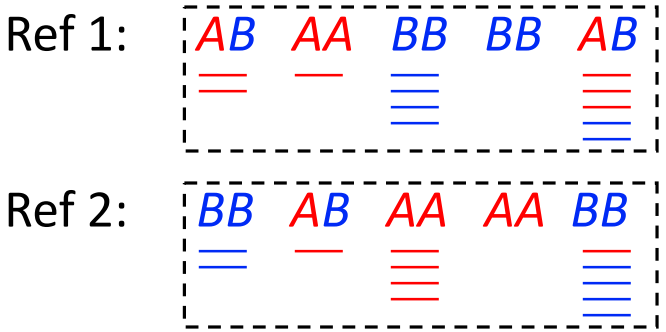
Step 2: Adjust Reference to Each Sample

Define C_{ij} as the coverage for sample *i* locus *j*

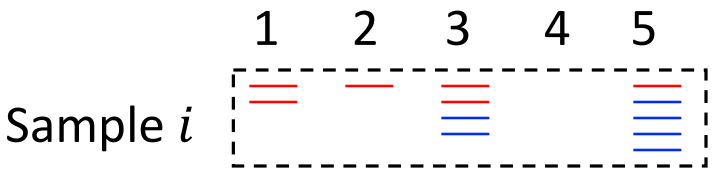


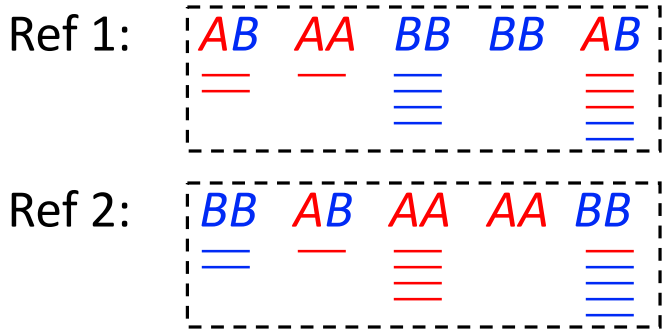
Simulate sequencing data for all reference individuals with coverage at each locus j equal to C_{ij}.

$$P(\text{drawing a read } A) = \begin{cases} 1-e & \text{if } g_{ij} = AA \\ 0.5 & \text{if } g_{ij} = AB \\ e & \text{if } g_{ij} = BB \end{cases}$$

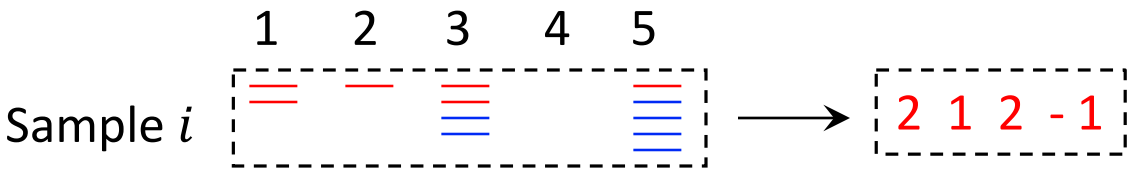


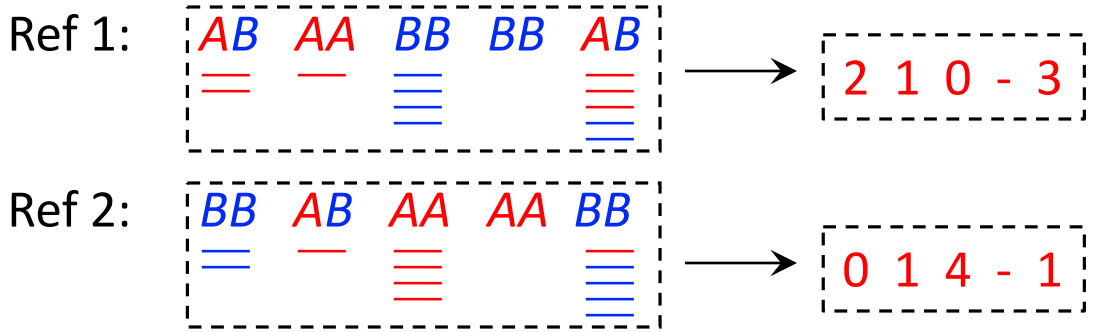
Step 2: Adjust Reference to Each Sample





Step 3: Count Variant Bases at Each Locus

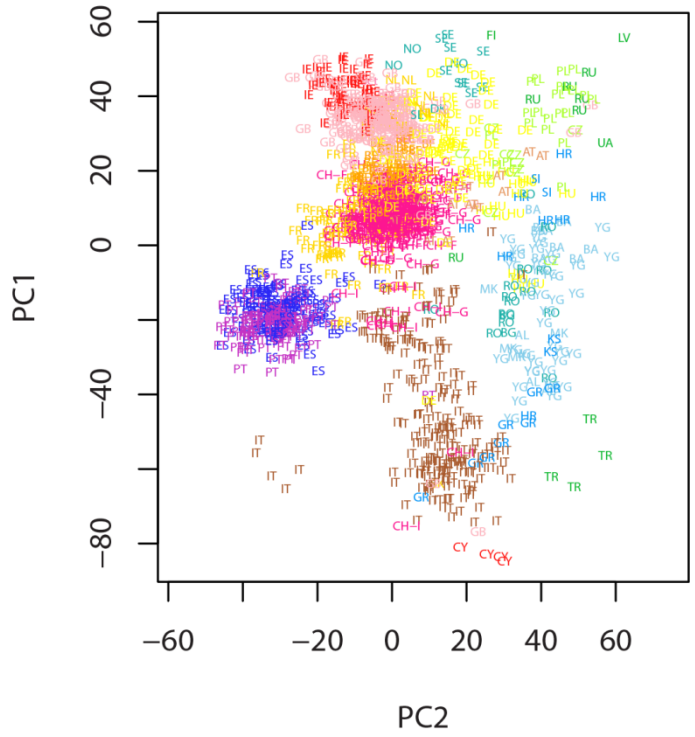




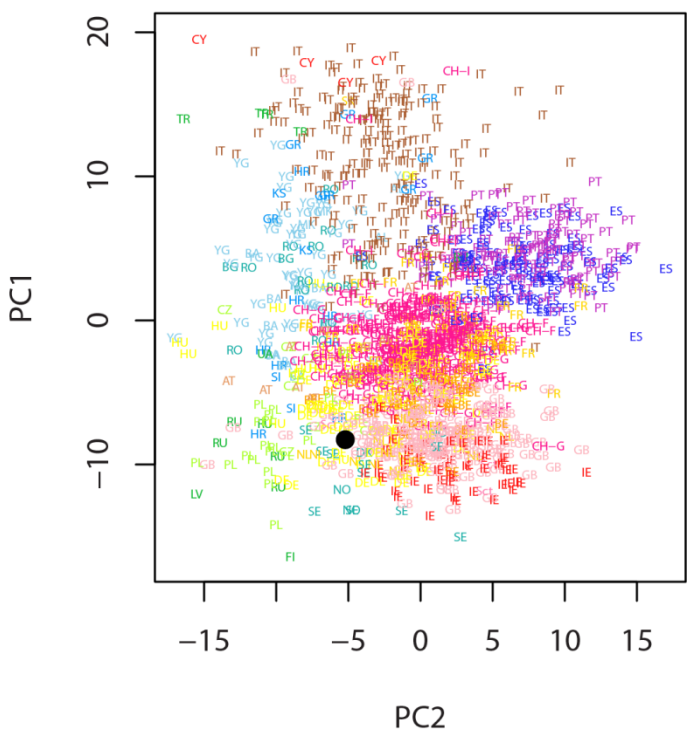
Step 4: Construct Sample Specific Map

Perform PCA on combined sequencing data of sample i and N reference individuals. (Map i)

Map 0

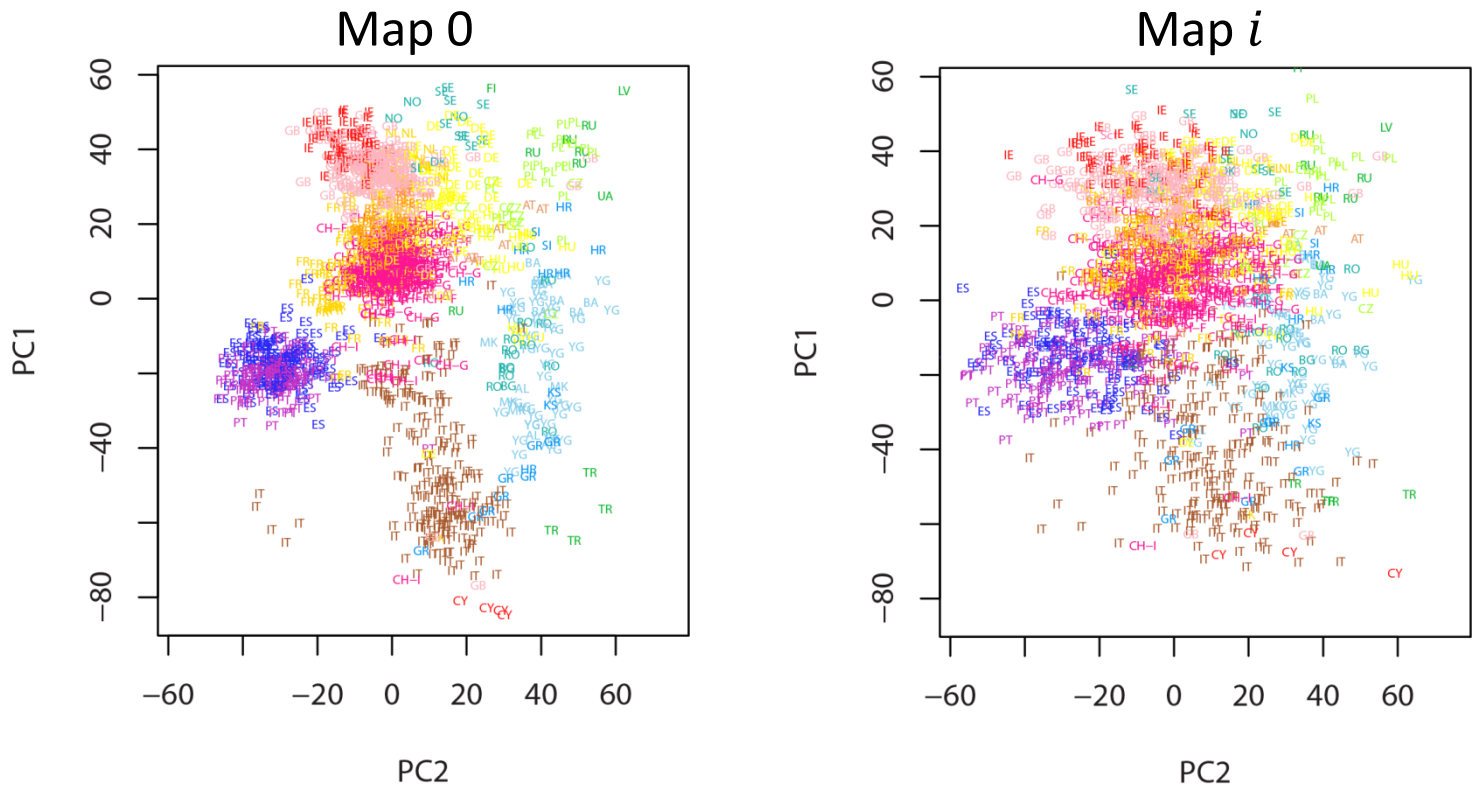


Map i



Step 5: Translate Between Map *i* and 0

Procrustes analysis: transform Map i to optimize the similarity to Map 0 based on the reference samples.

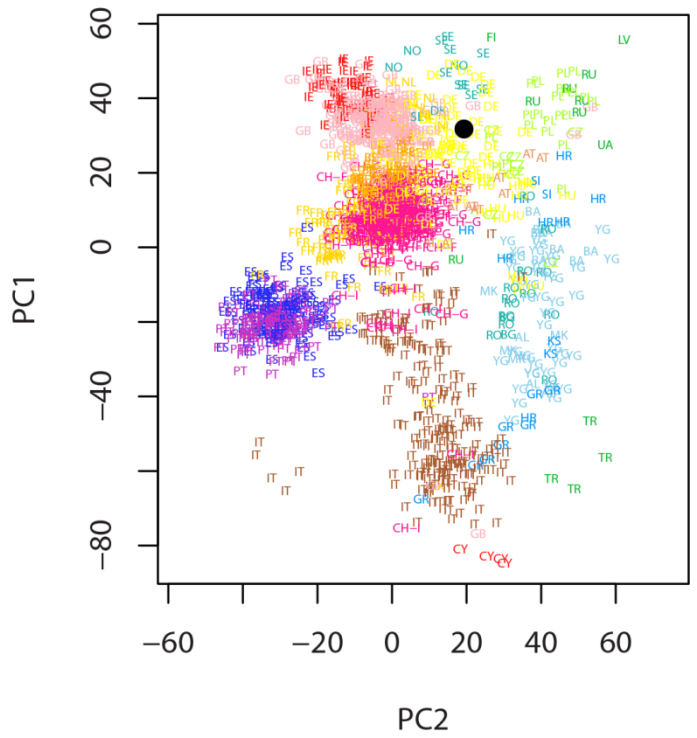


Wang *et al.* (2010) *Stat. Appl. Genet. Mol. Biol.*

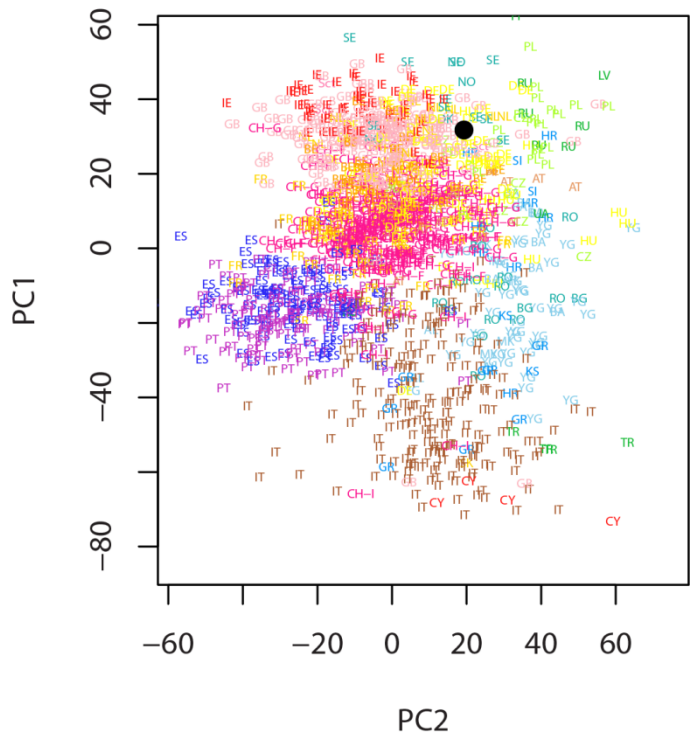
Step 6: Apply Translations

Apply the transformation to place sample i on the reference PCA map.

Map 0

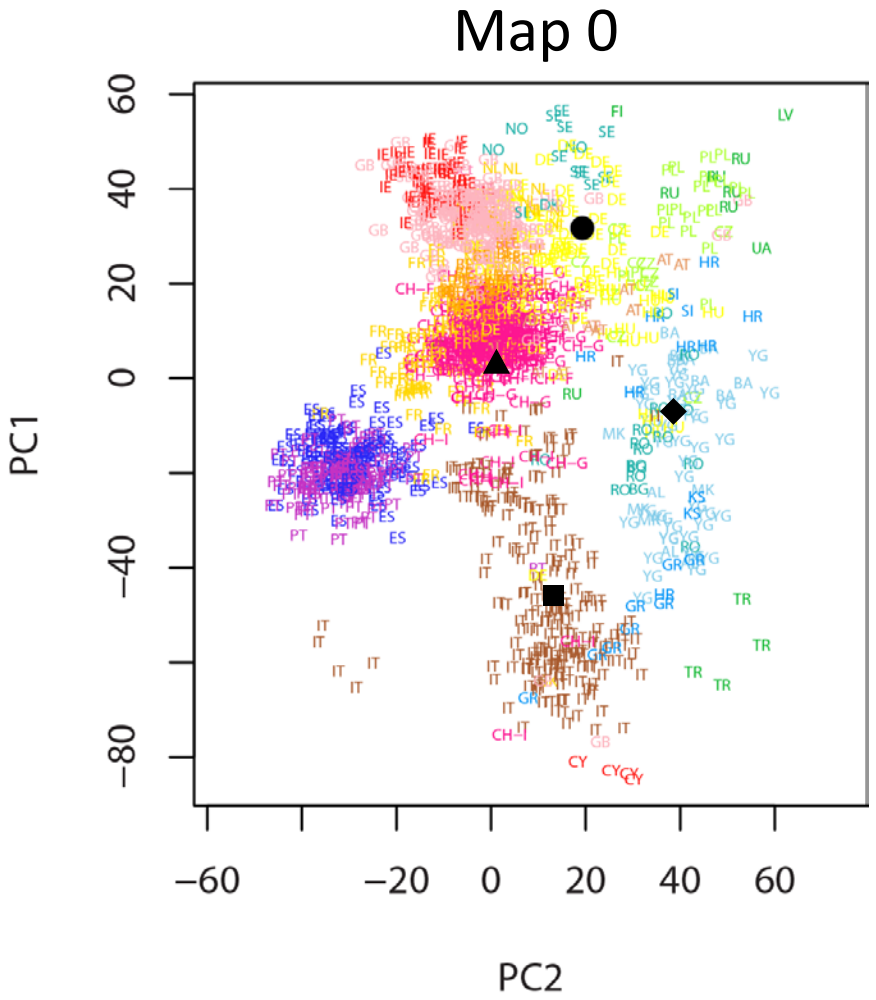


Map i



Step 7: Repeat!

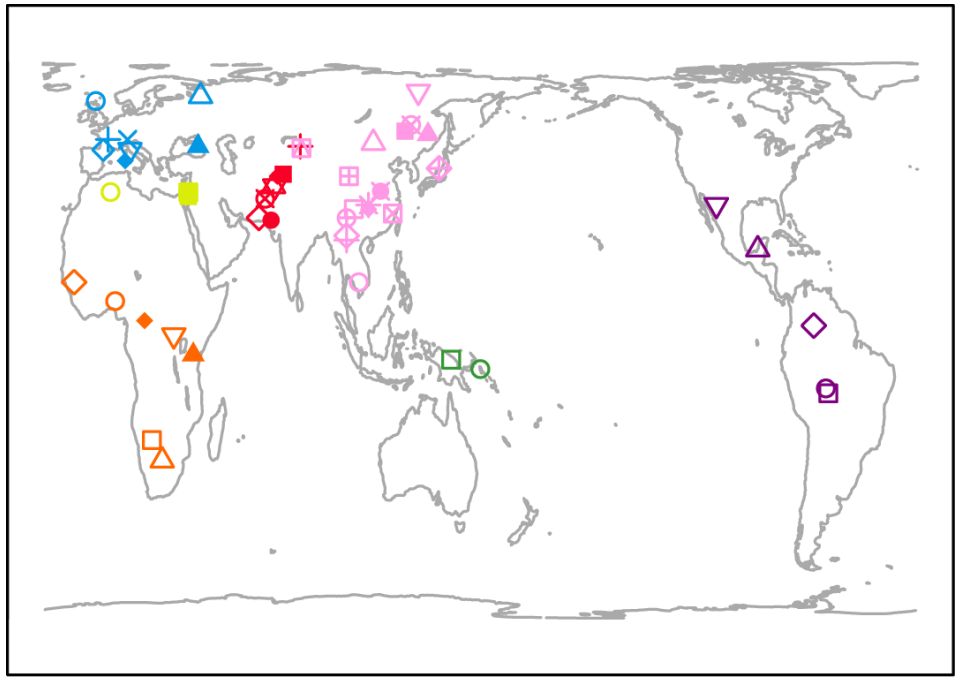
Repeat steps 2-6 for all sequenced samples.



Human Genome Diversity Panel

938 individuals, 632,958 markers

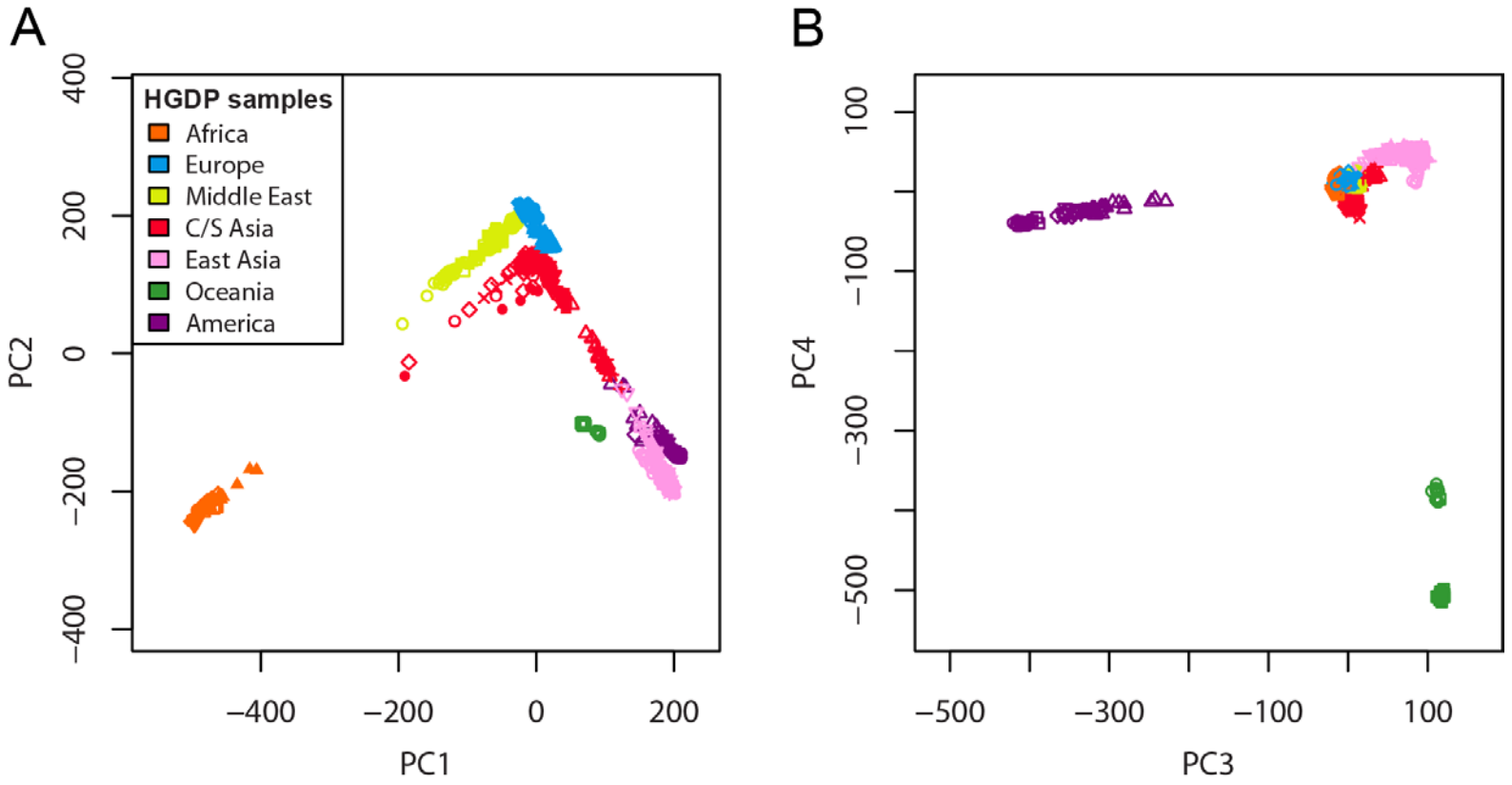
Li et al (Science, 2008)



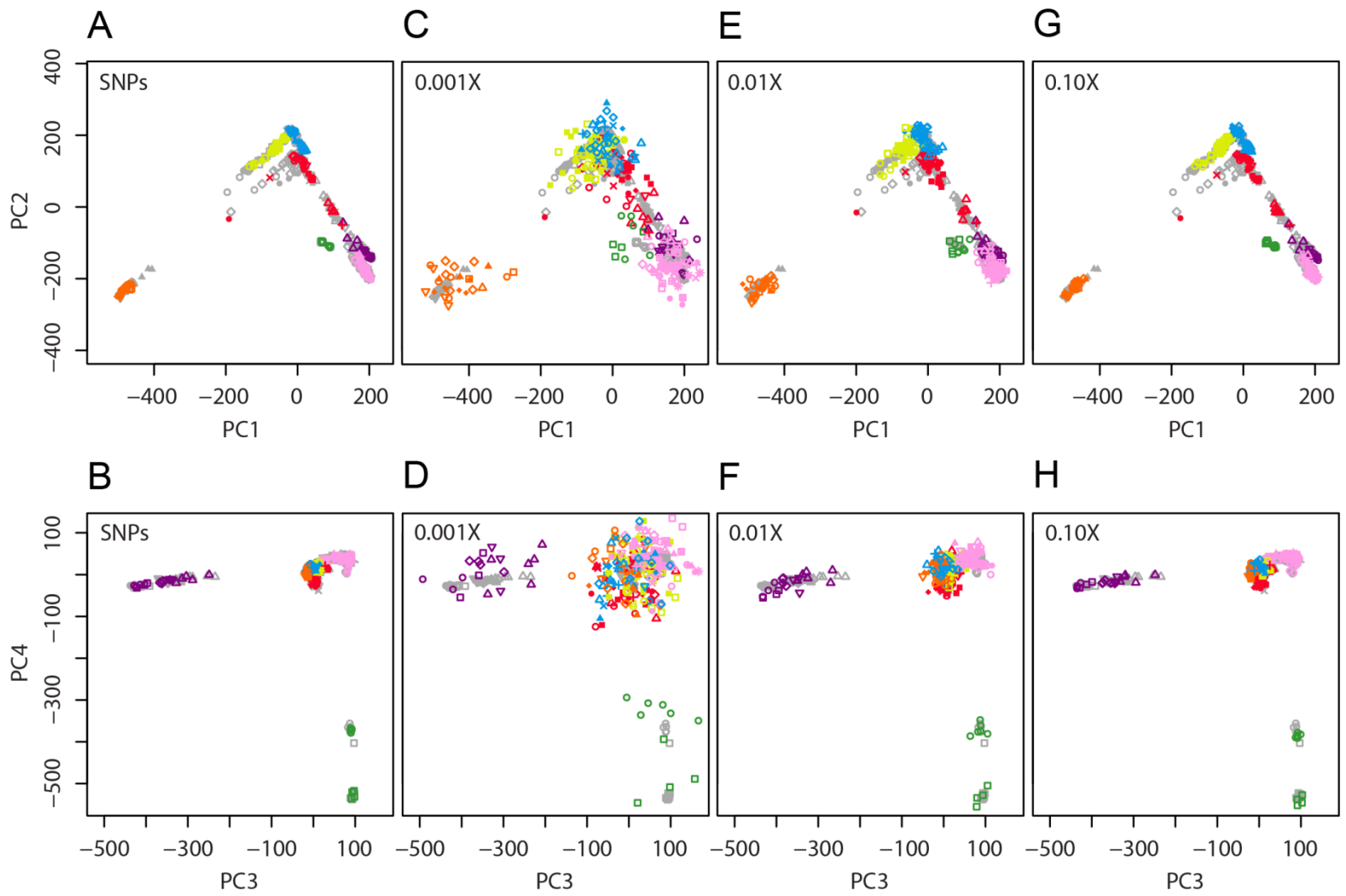
Africa
Europe
Middle East
C/S Asia
East Asia
Oceania
America

Adygei
Balochi
Bantu (Kenya)
Bantu (S. Africa)
Basque
Bedouin
Biaka Pygmy
Brahui
Burusho
Cambodian
Colombian
Dai
Daur
Druze
French
Han
Han (N. China)
Hazara
Hezhen
Italian
Japanese
Kalash
Karitiana
Lahu
Makrani
Mandenka
Maya
Mbuti Pygmy
Melanesian
Miao
Mongola
Mozabite
Naxi
Orcadian
Oroqen
Palestinian
Papuan
Pathan
Pima
Russian
San
Sardinian
She
Sindhi
Surui
Tu
Tujia
Tuscan
Uygur
Xibo
Yakut
Yi
Yoruba

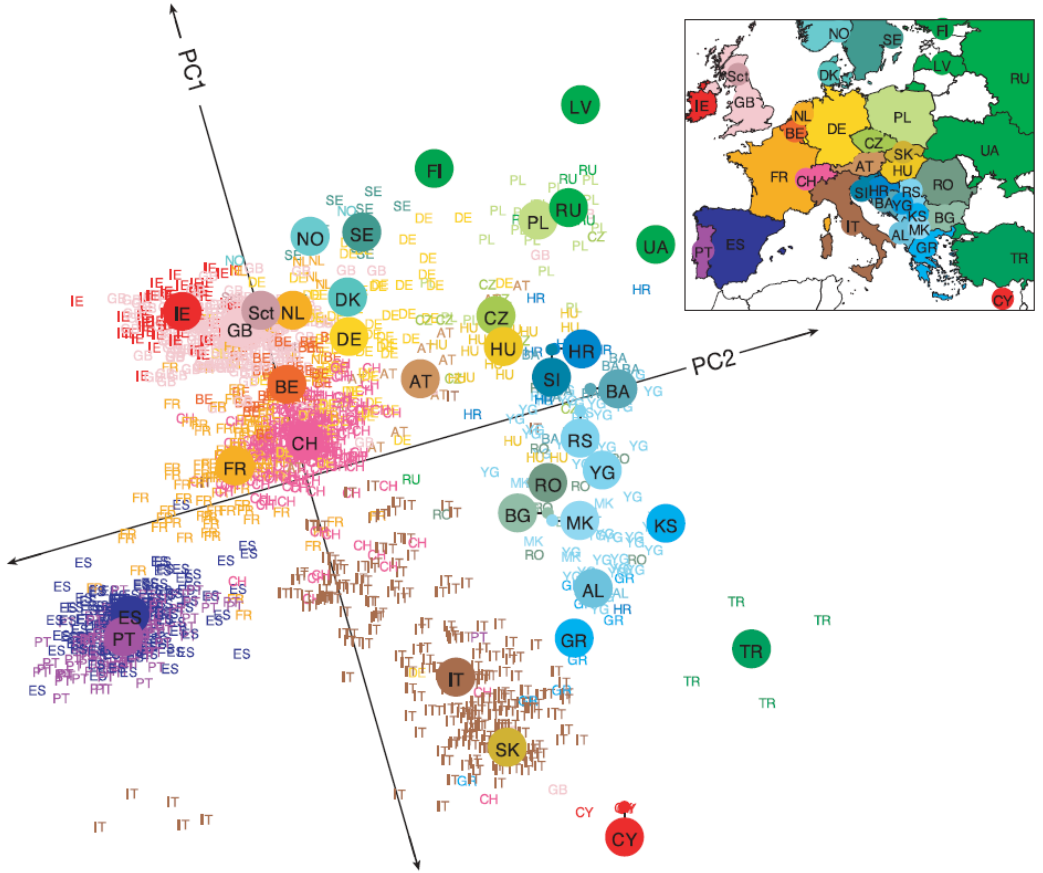
Worldwide Principal Component Ancestry Map



Placing Individuals on a Worldwide Ancestry Map

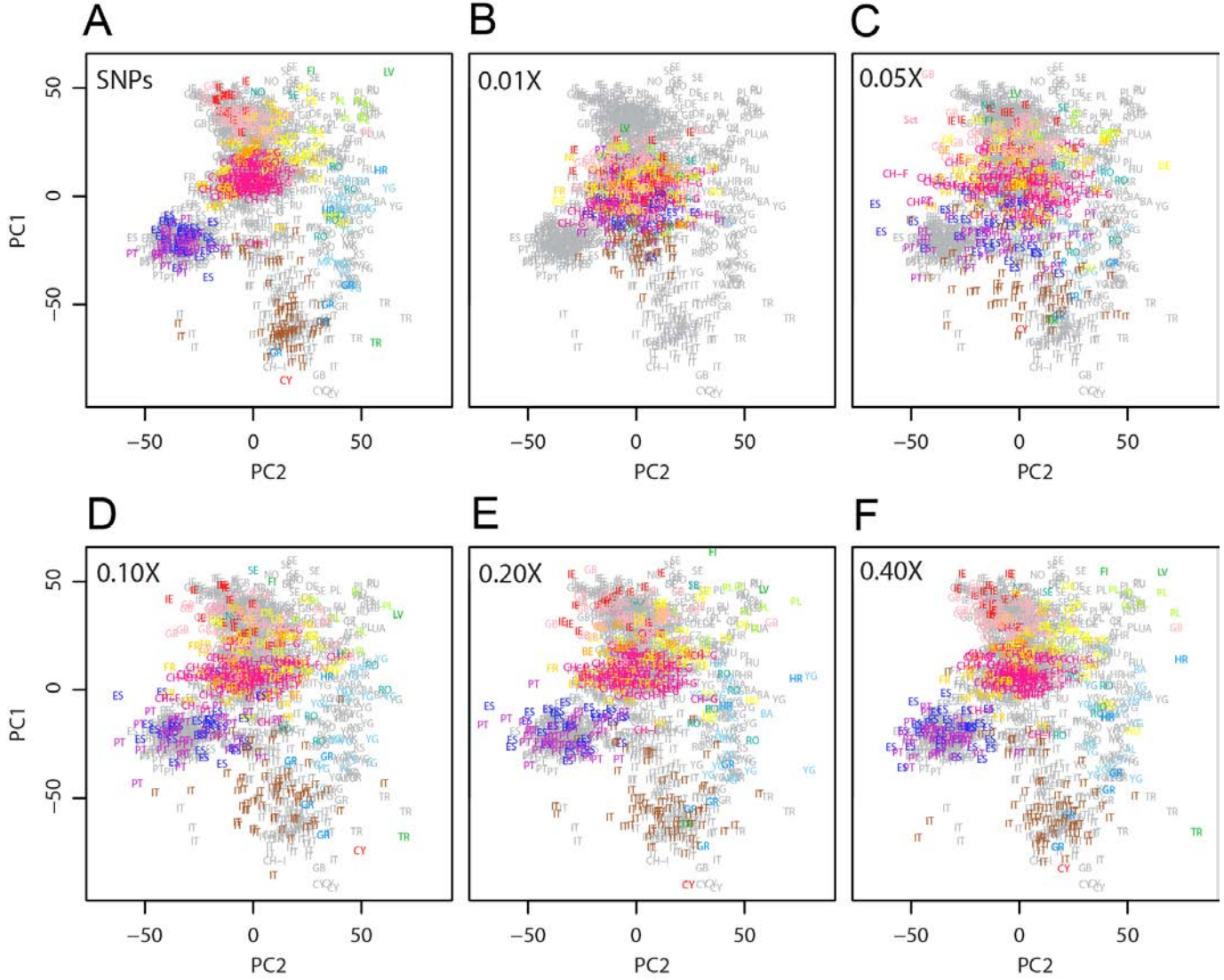


Principal Component Ancestry Map of Europe



Novembre *et al.* (2008) *Nature*

Placing Individuals on an European Ancestry Map

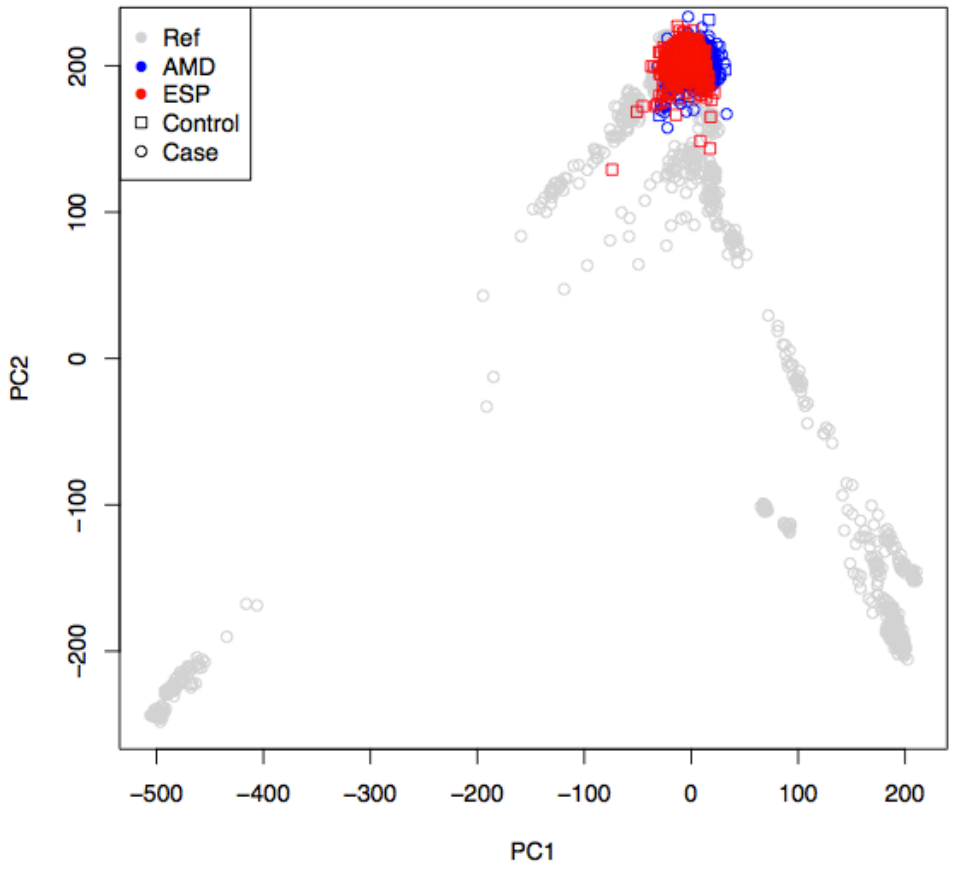


Simulation Results
European Ancestry Map

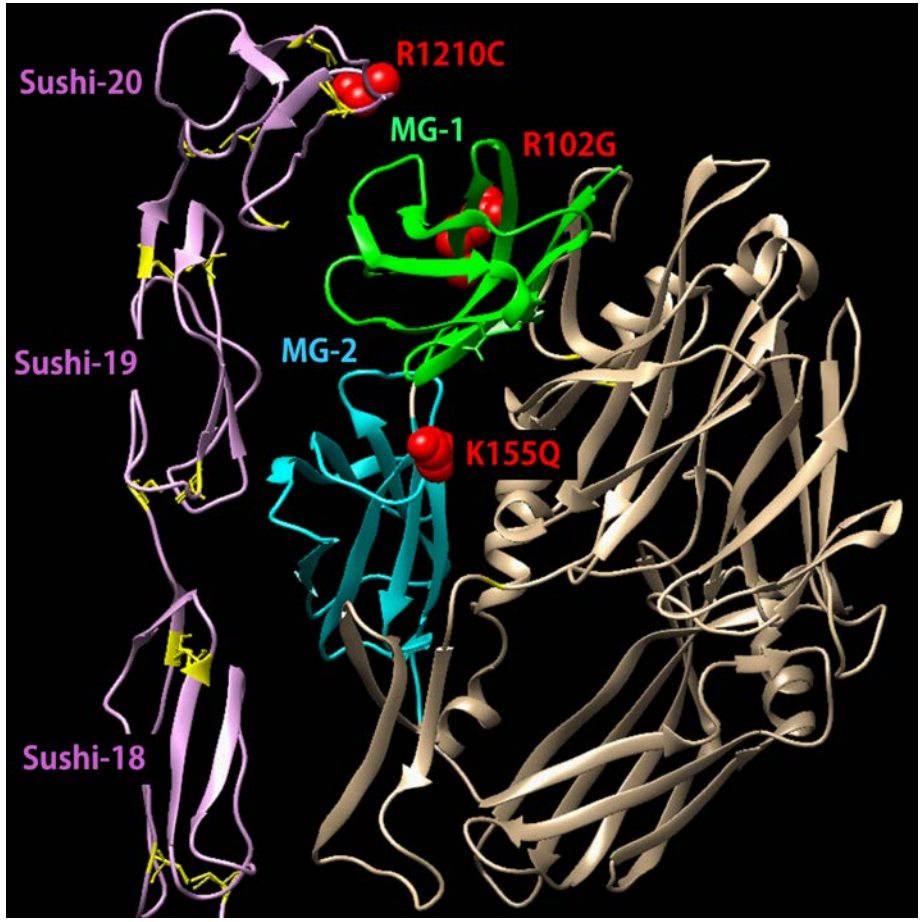
Simulated coverage λ	Loci with ≥ 1 reads	Sequence- vs. SNP-based coordinates	
		Pearson correlation of PC1	Pearson correlation of PC2
0.40	105,063	0.9927	0.9528
0.35	94,111	0.9933	0.9458
0.30	82,597	0.9906	0.9341
0.25	70,492	0.9898	0.9241
0.20	57,767	0.9868	0.8929
0.15	44,390	0.9825	0.8811
0.10	30,327	0.9752	0.8153
0.05	15,542	0.9408	0.5016
0.01	3,171	0.7541	0.1041

Matching Results

- Searched 6,800+ ESP samples for matches
- Built matched set
  - 2,268 AMD cases
  - 2,268 controls
  - Focused on sites with high depth
  - Excluded sites near indels
- R1210C variant now has $p<10^{-6}$
  - 23 cases
  - 1 control
- New rare variant signals under investigation, variant in C3 particularly interesting



AMD Risk Variants in CFH and C3 ….

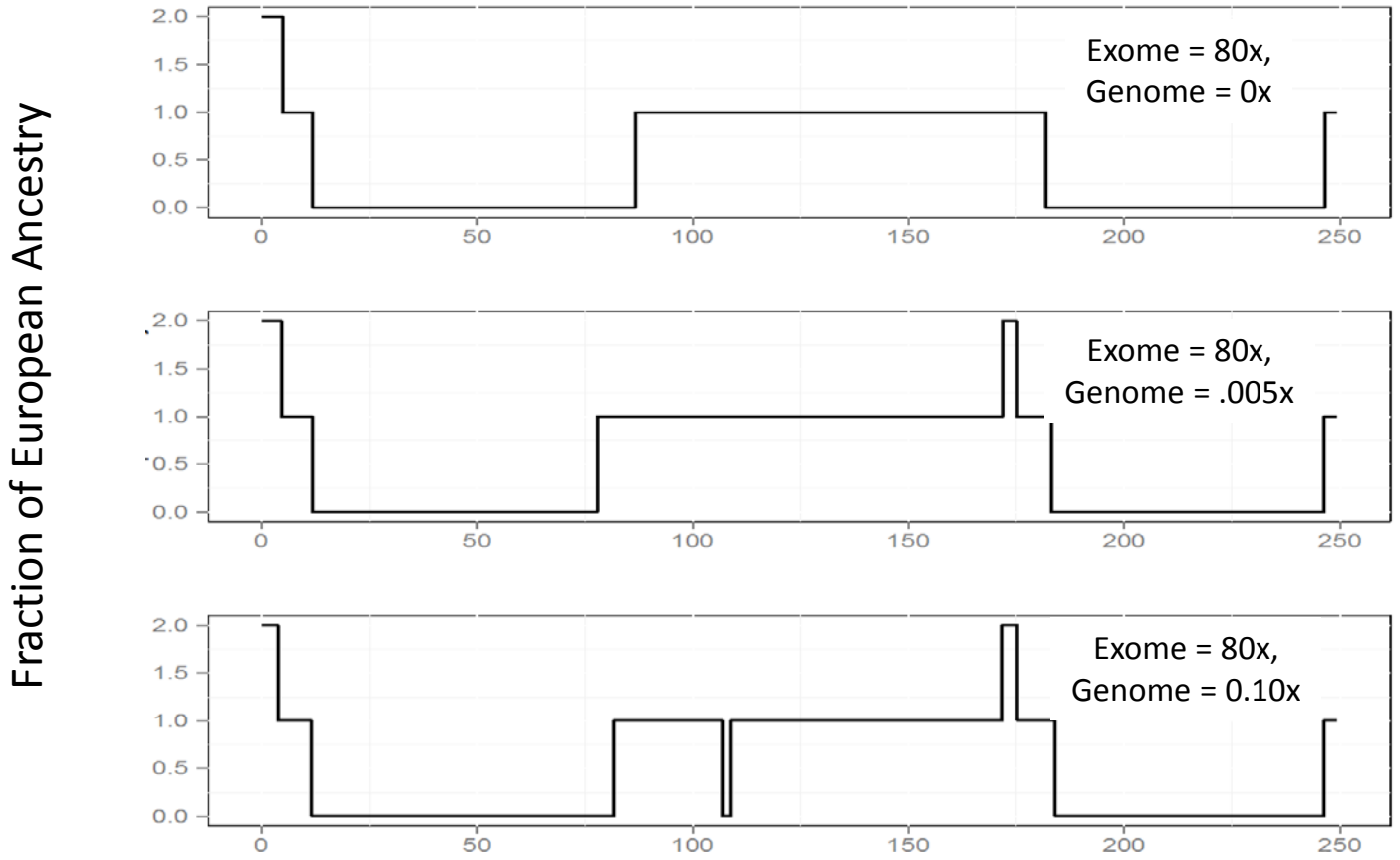


- CFH R1210, OR ~10
- C3 R102G, OR ~1.3
- C3 K155Q, OR ~3.0

- Variants appear to map in the region where C3 and CFH interact

- CFH inactivates C3 to downregulate alternate complement pathway

With Even Higher Depths, Possible To Estimate Local Ancestry



Truth

Increasing Depth

Youna Hu

Genotyping Arrays for Rare Variant Association Studies

Benjamin Neale

Gonçalo Abecasis

Further Expanding Rare Variant Analysis

- There are many interesting rare variants, but achieving large sample sizes is a challenge

- Consider CFH …
  - 4 stop variants in AMD cases, none in controls
  - If premature stops occur in 1,000 cases; no controls…

- To get to 10^{-6} significance might need to sequence…
  - ~21,000 cases and ~21,000 controls, or
  - ~7,000 cases and ~56,000 controls, or
  - Fewer cases if very large numbers of controls available

Motivation for an Exome Array

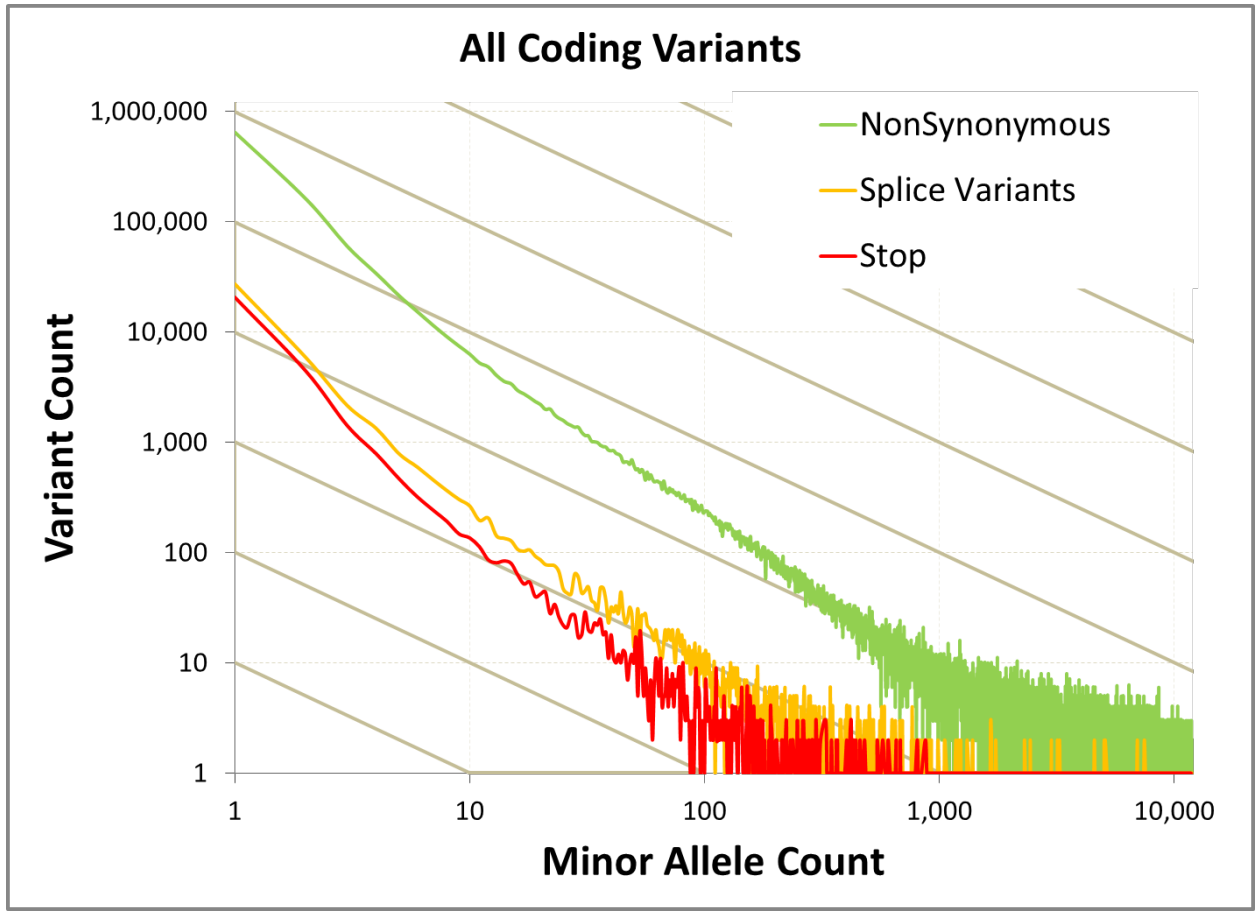
- Current sequencing studies are well powered to discover exome variants that contribute to disease (MAF > 0.1%)
- Sequencing studies may be underpowered to establish association of those variants to phenotype
  - Larger numbers of individuals must be examined to establish the effect of a variant than to discover it
- Genotyping is less expensive that exome sequencing, allowing larger sample sizes and, perhaps, power

Approach to designing the array

- Collate sites and counts from a “coalition of the willing” with data from exome or genome sequencing
  - Site lists constitute preliminary analyses of unpublished data
- Cover as much variation as possible while avoiding private mutations and technical artifacts
  - Quality filters, HWE checks
  - Nonsynonymous variants ≥3 times and in ≥ 2 studies
  - Splice and stop variants seen ≥ 2 times and in ≥ 2 studies
  - Relaxed frequency filter for ancestries with few samples

Ben Neale

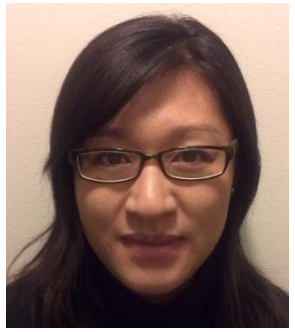
Coding Variants Ascertained By Sequencing 12,000+ Individuals



Gene Based Burden Tests

- Most coding variants are very rare
- Testing the effect of each variant may often be impossible
- Geneticists currently favor “gene burden tests”
- These tests evaluate the combined effect of rare variants in a gene
- Are there convenient ways to reach large sample sizes with these tests?

Rare Variant Analysis in Large Samples



Shuang Feng Dajiang Liu

- The insight …

- Simple burden tests can be calculated as a linear function of single variant score statistics

$$T \propto \vec{w}^T \vec{U}$$

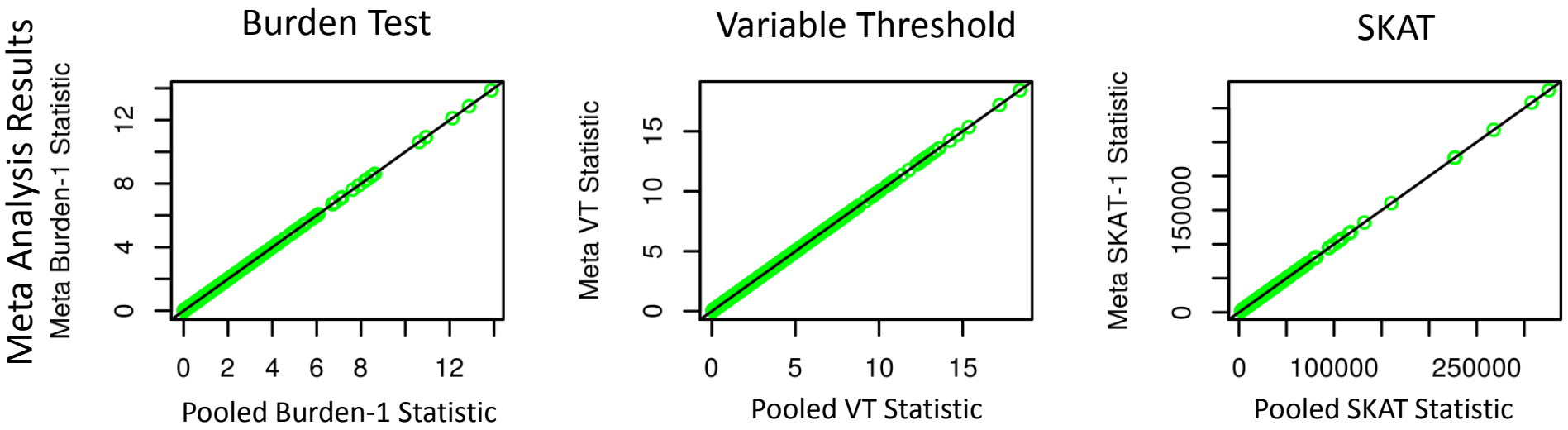
- More advanced burden tests (like SKAT) can be derived as quadratic function of single variant statistics

$$T \propto \vec{U}^T \mathbf{K} \vec{U}$$

The Solution

- For each study …
- Calculate components of single variant association score statistics
- Calculate variance covariance matrix for score statistics
- If we share these two pieces of information (single variant statistics, disequilibrium matrix) …
- … we can conveniently calculate many gene level tests across large samples

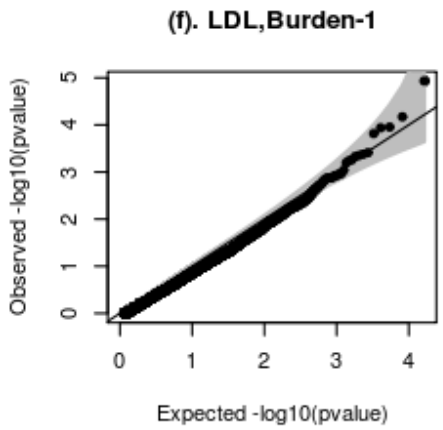
In the absence of heterogeneity, meta-analysis and pooled statistics match

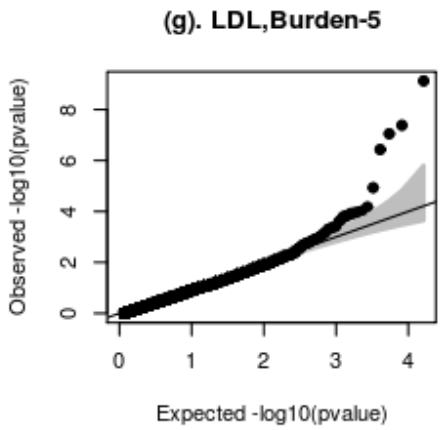


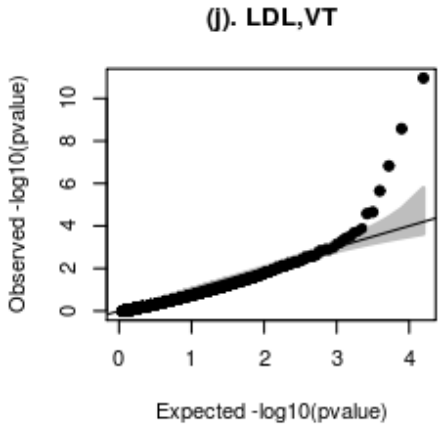
Pooled Data Results

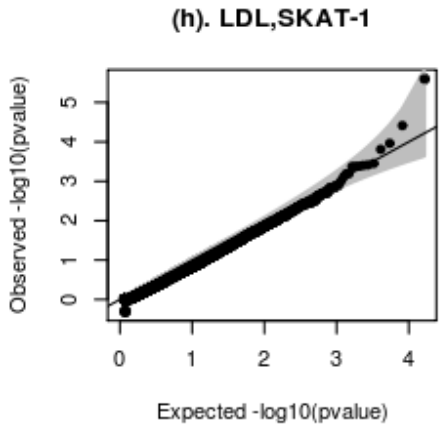
Rare Variant Meta-Analysis Example

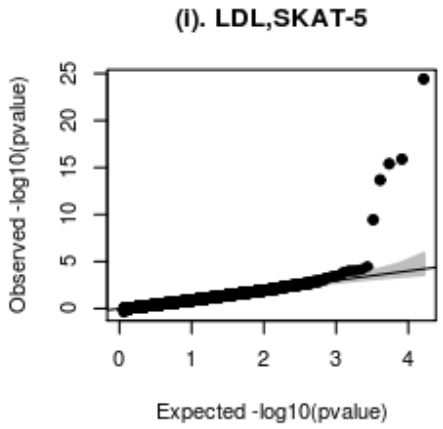
LDL cholesterol, 15,000 individuals











Analysis with Gina Peloso and Sekar Kathiresan

QQ-plots well behaved

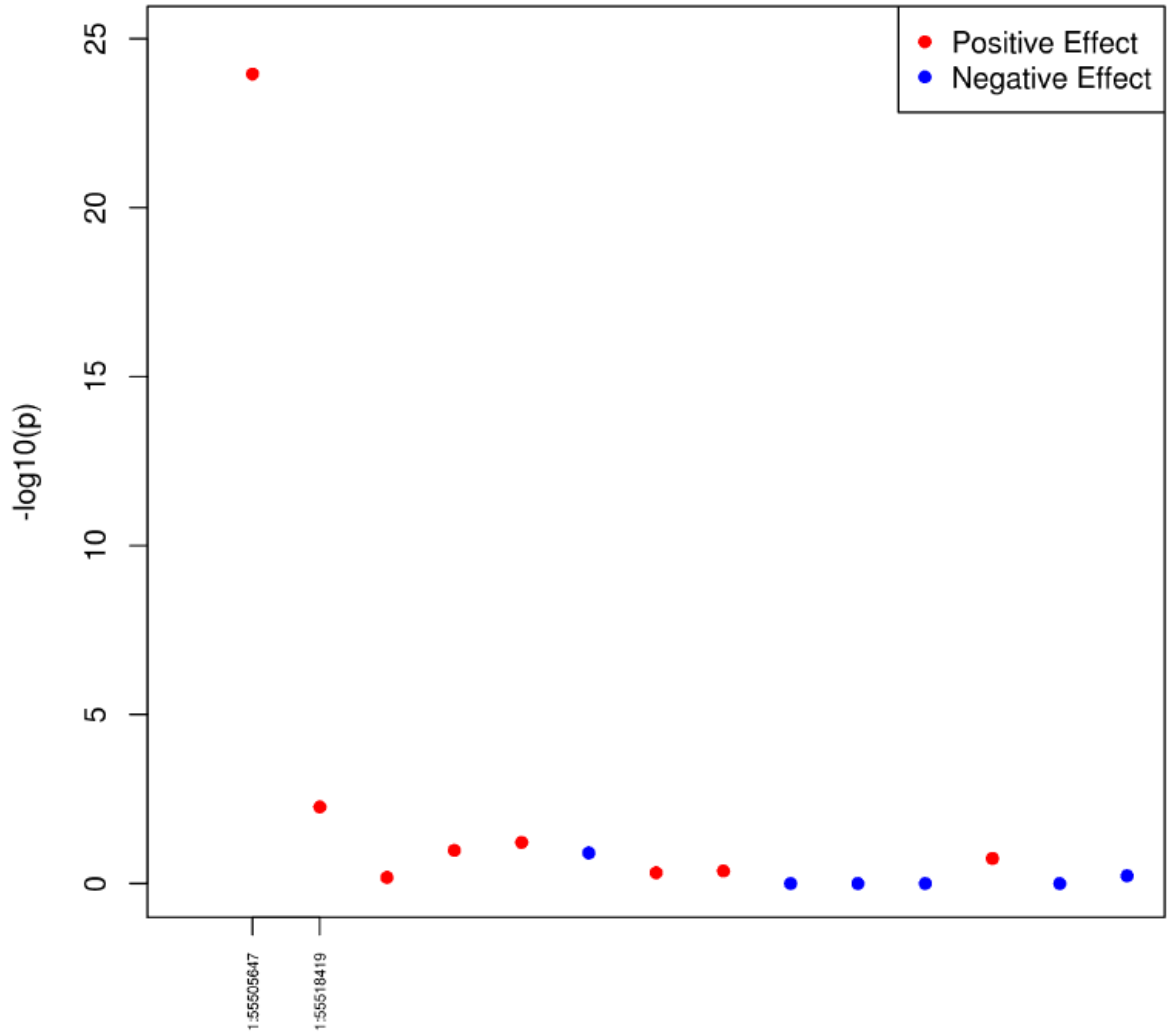
Rare Variant Meta-Analysis Example
LDL cholesterol, 15,000 individuals

Gene	Burden-5	SKAT-5	Variable Threshold	VT cut-off
PCSK9	**3×10^{-7}**	**7×10^{-25}**	**2×10^{-12}**	.015
APOB	**3×10^{-3}**	**2×10^{-14}**	**.046**	.041
LDLR	.071	**9×10^{-3}**	**2×10^{-5}**	.00074

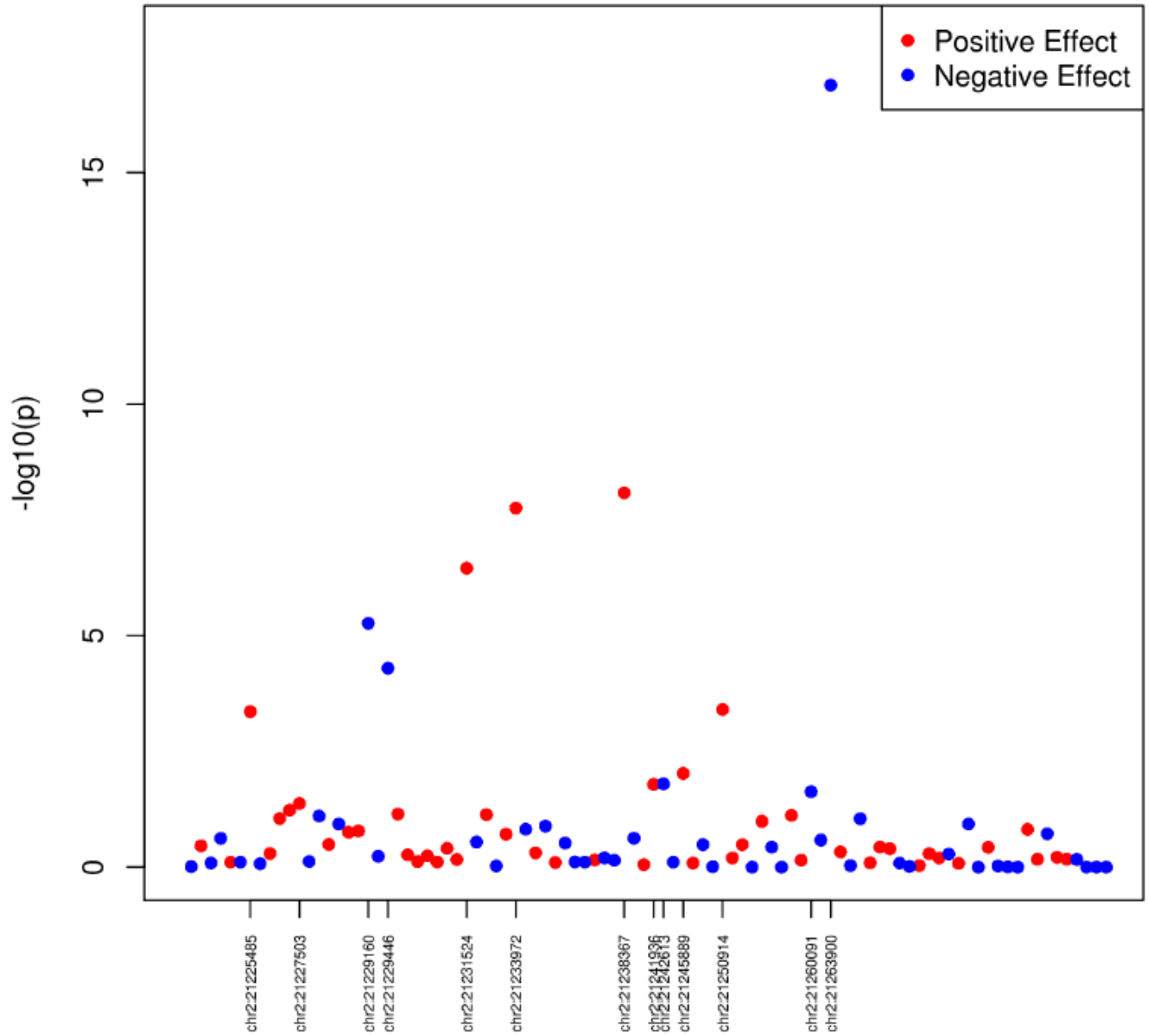
Analysis with Gina Peloso and Sekar Kathiresan

Results shown for Mendelian hypercholesterolemia genes with gene level p-value < .001 in 15,000 individuals (12 genes tested, 3 tests)

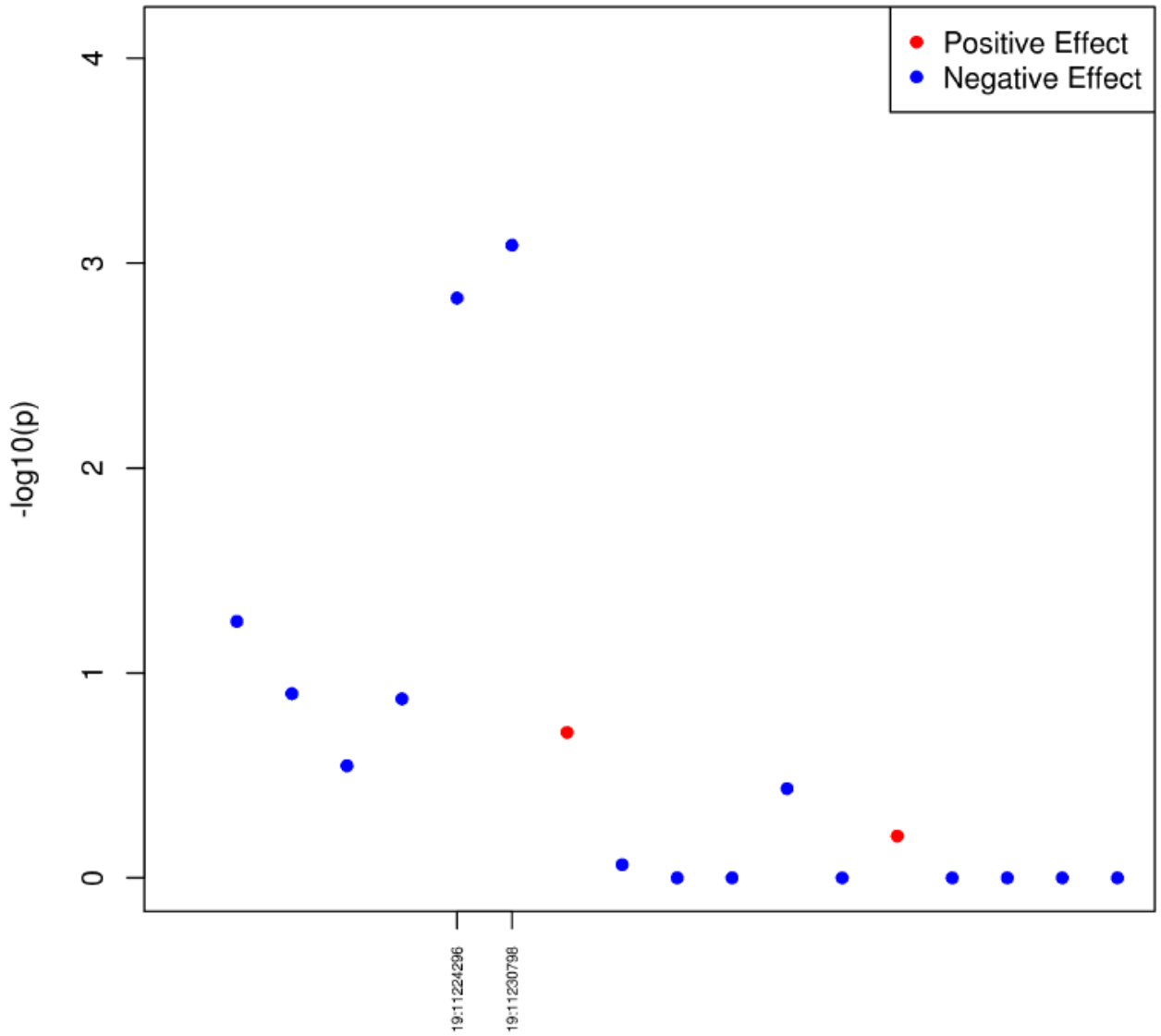
PCSK9: Marker-by-Marker



APOB: Marker-by-Marker



LDLR: Marker-by-Marker



State of Play

- Genomewide associations very effective at identifying disease susceptibility loci
  - Translating the subtle effects of these loci to function is challenging
- Sequencing studies can examine rare variants, including many with clear function
  - Achieving the sample sizes required to establish association is challenging
- There is a promising family of designs that are combine sequencing to discover rare variants and genotyping to reach very large sample sizes

Rare-Metal and Rare-Metal-Worker

- Tools for facilitating meta-analysis of rare variants
- Rare-Metal-Worker calculates per study summary statistics
- Rare-Metal combines these to calculate burden statistics

Exercise: Input Files

- Get a copy into your home directory

```
mkdir Rare-Metal-Example
cp /faculty/goncaloa/2013/Rare-Metal-Example/* Rare-Metal-Example
```

- For each study, we start with a:
  - Data File, listing traits and covariates
  - Pedigree File, listing phenotypes for each person
  - VCF File, listing genotypes for each person

- For the meta-analysis, we need:
  - A list of studies
  - A key for grouping variants into genes

Running RareMetalWorker …

- This will analyze each study and generate summaries

```
raremetalworker --dat sample1.dat --ped sample1.ped --vcf sample1.vcf.gz \
                --prefix sample1

raremetalworker --dat sample2.dat --ped sample2.ped --vcf sample2.vcf.gz \
                --prefix sample2
```

- Useful additional options include:
  - `--kinGeno`
  - `--inverseNormal`
  - `--useCovariates –makeResiduals`

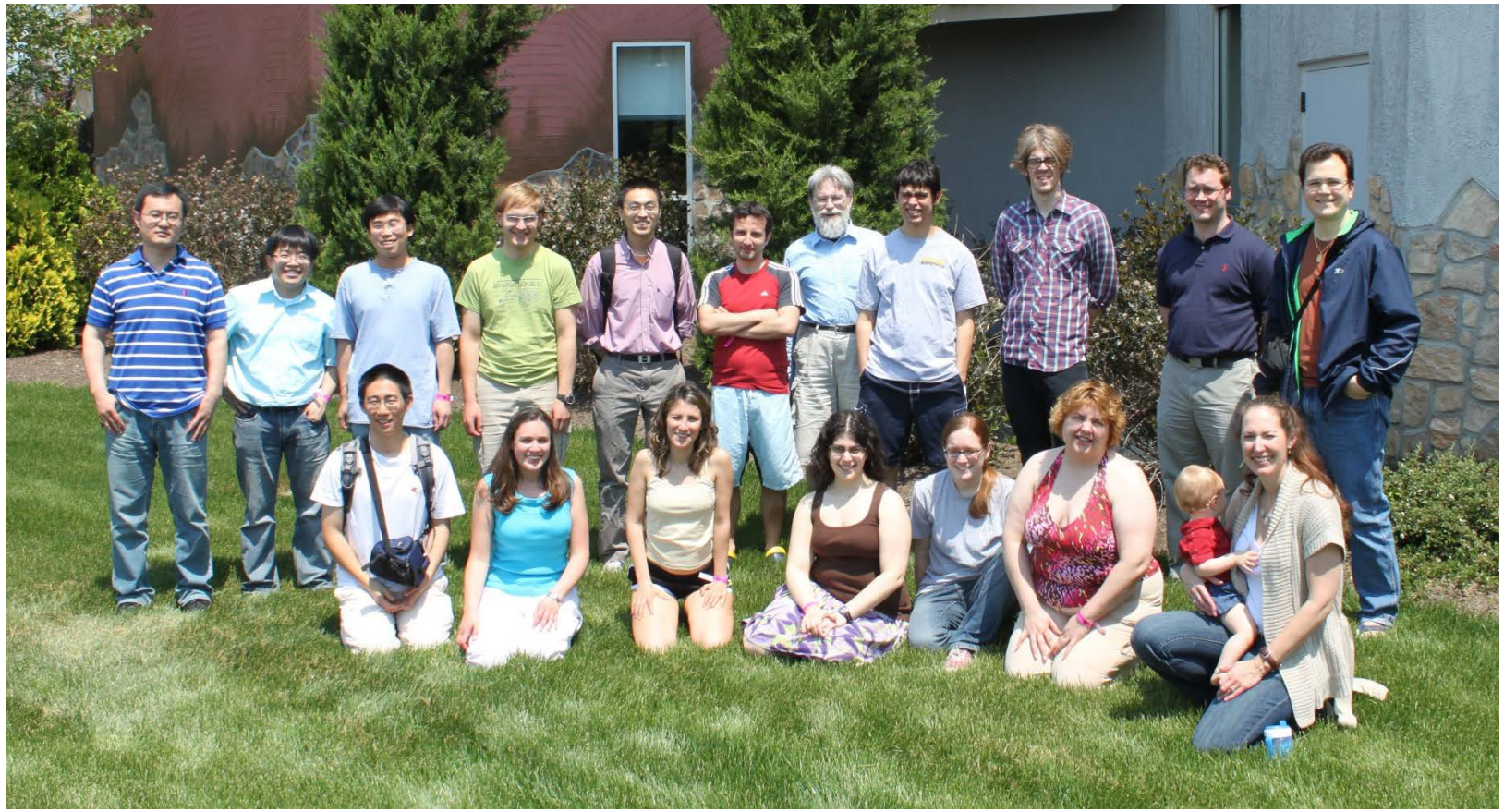
Running RareMetal

- This will combine studies to run a meta-analysis …

```
raremetal --study sample.lst --group burden_groupings.txt --SKAT --VT
```

- Useful additional options include:
  - `--burden --MB`
  - `--maf`
  - `--hwe`
  - `--callRate`

Acknowledgements



Thank you to the National Institutes of Health (NHGRI, NEI, NHLBI)
for supporting our work.