The Role of Rare Variants in Complex Disease

Gonçalo Abecasis

University of Michigan School of Public Health

Human Genetics, Sample Sizes over My Time

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

Human Genetics, Sample Sizes over My Time

Year	No. of Samples	No. of Markers	Publication
Ongoing	~33,000	50 million	Haplotype Reference Consortium (Talk #176)
Ongoing	~40,000	12 million	Macular Degeneration Study (Talk #384/387)
2012	1,092	40 million	The 1000 Genomes Project (Nature)
2010	Hundreds	16 million	The 1000 Genomes Project (Nature)
2010	~100,000	2.5 million	Lipid GWAS (Nature)
2008	~9,000	2.5 million	Lipid GWAS (Nature Genetics)
2007	Hundreds		
2005	Hundreds		
2003	Hundreds		p (Nature Genetics)
2002	Hundreds		p (Nature)
2001	Thousands		Three Region Variation Map (Am J Hum Genet)
2000	Hundreds	26	T-cell receptor variation (Hum Mol Genet)

Early studies looked at a few genetic variants, picked based on intuition and prejudice.

New discoveries were few and far between.

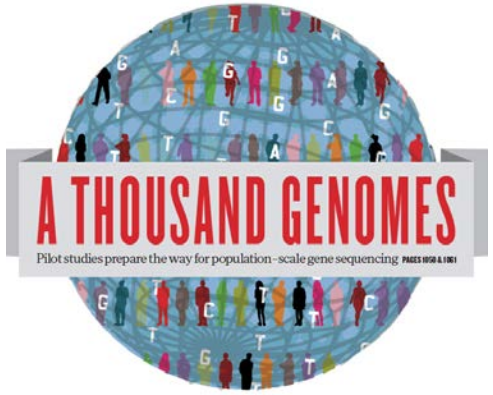
Human Genetics, Sample Sizes over My Time

Year	No. of Samples	No. of Markers	Publication
Ongoing	~33,000	50 million	Haplotype Reference Consortium (Talk #176)
Ongoing	~40,000	[illegible]	Macular Degeneration Study (Talk #384/387)
2012	1,092	[illegible]	oject (Nature)
2010	Hundreds	[illegible]	oject (Nature)
2010	~100,000	[illegible]	
2008	~9,000	[illegible]	enetics)
2007	Hundreds	[illegible]	HapMap (Nature)
2005	Hundreds	1 million	HapMap (Nature)
2003	Hundreds	10,000	Chr. 19 Variation Map (Nature Genetics)
2002	Hundreds	1,500	Chr. 22 Variation Map (Nature)
2001	Thousands	127	Three Region Variation Map (Am J Hum Genet)
2000	Hundreds	26	T-cell receptor variation (Hum Mol Genet)

Modern studies are more comprehensive and systematic.

New discoveries accumulate fast.
Much potential for secondary uses of data.

The 1000 Genomes Project



Gil McVean David Altshuler Richard Durbin

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)

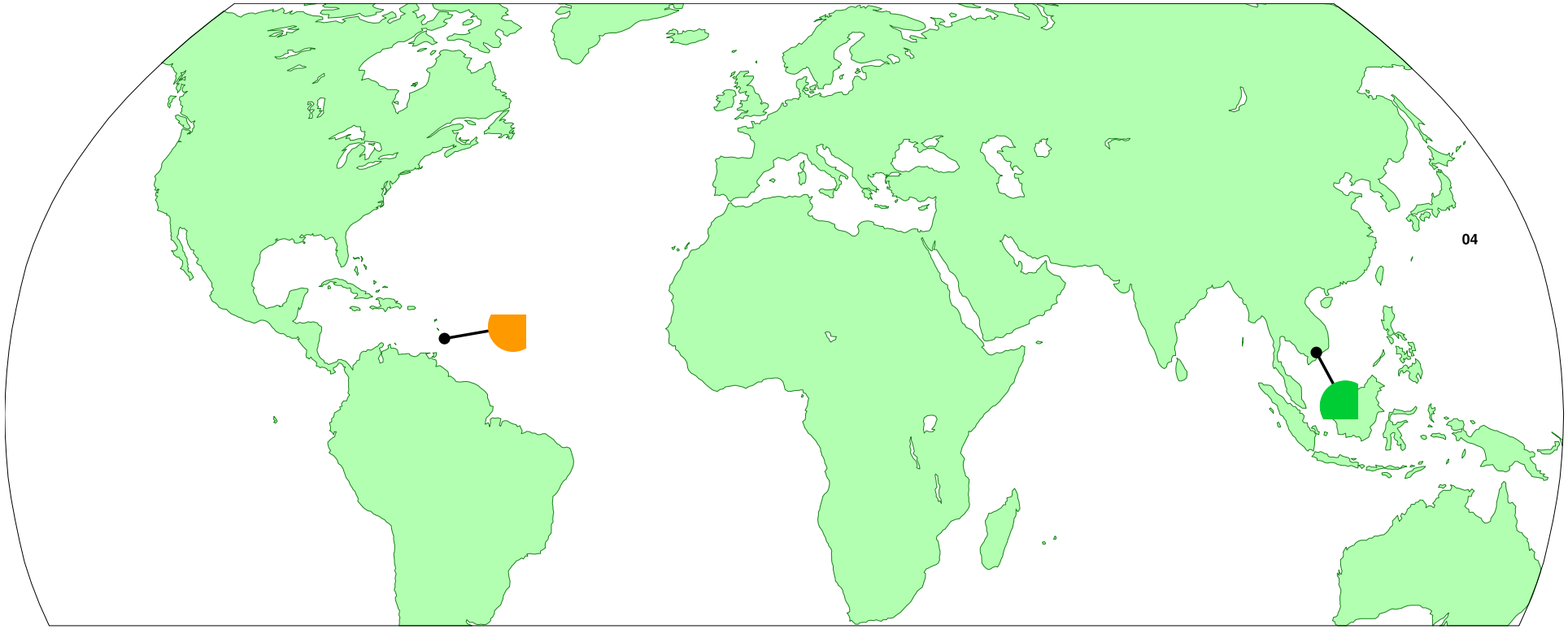
- >38.5 million SNP
  - 8.5M sites discovered before project (dbSNP 129)
  - 30M sites newly discovered
  - 98.9% of HapMap III sites rediscovered
  - Transition/transversion ratio of 2.16 vs 2.04 in pilot

- ~1.5M insertion deletion polymorphisms

- ftp://ftp.1000genomes.ebi.ac.uk
- ftp://ftp.ncbi.nlm.nih.gov/1000genomes/

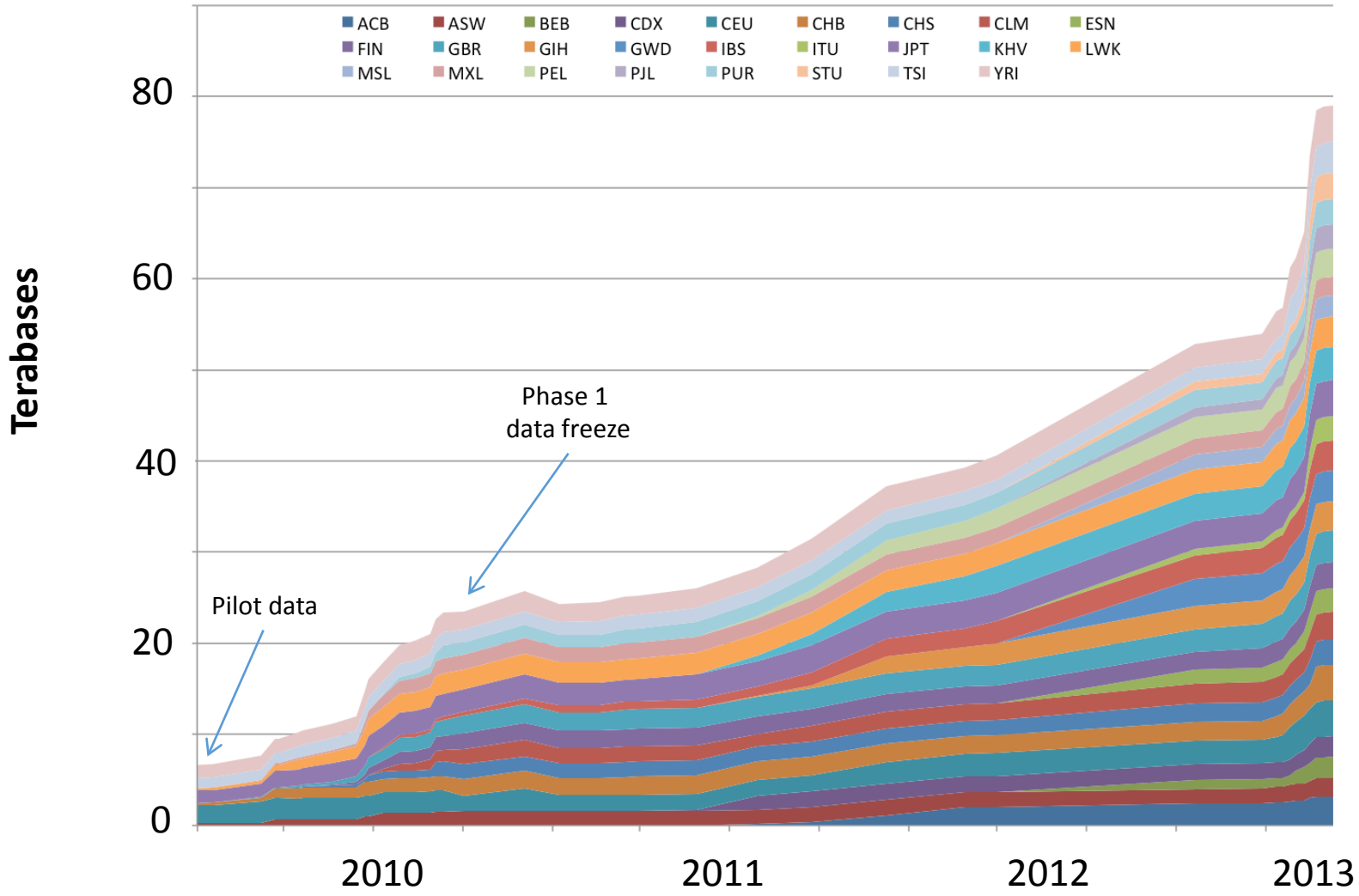
The 1000 Genomes Project (Nature, 2012)

Samples in the final phase



Bubble size = sample size

1000 Genomes data generation

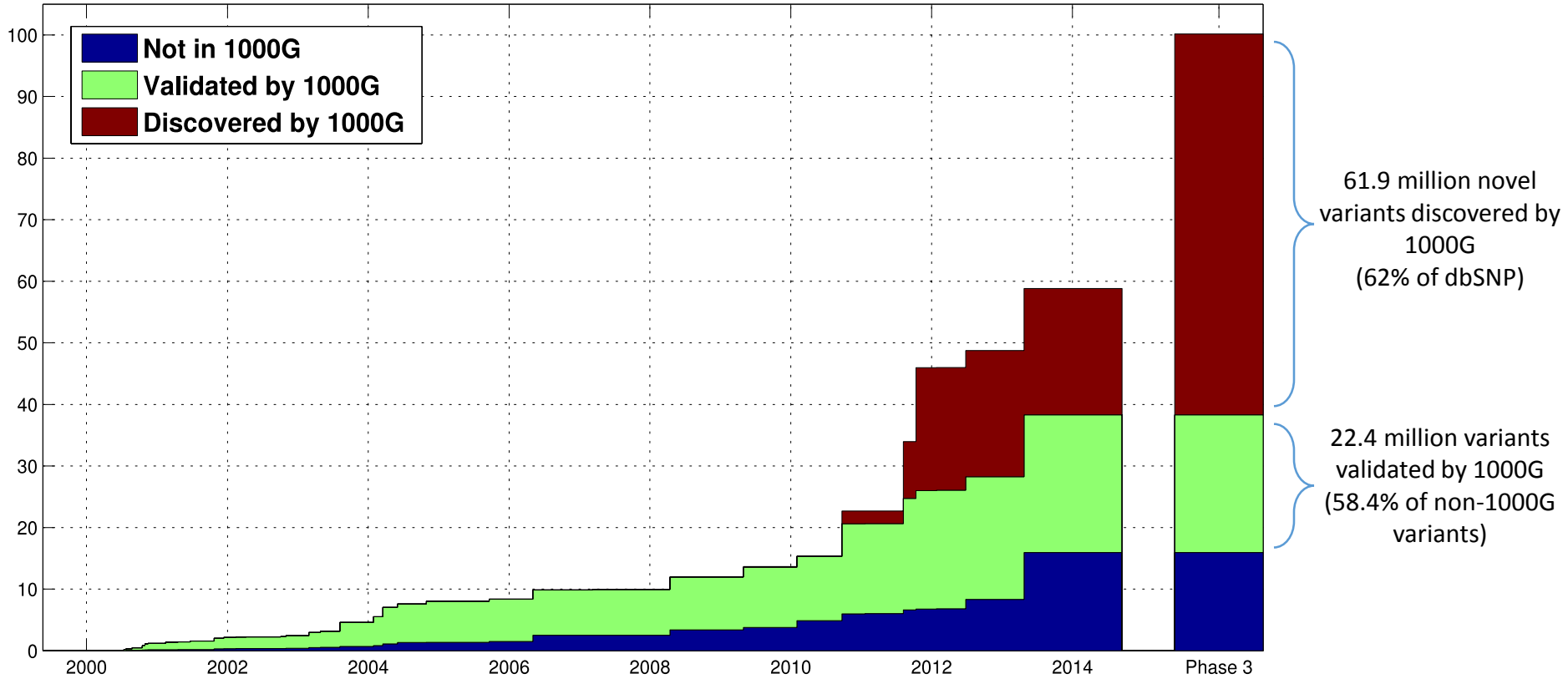


1000 Genomes Data

Total Dataset:
84 TB of BAM Files

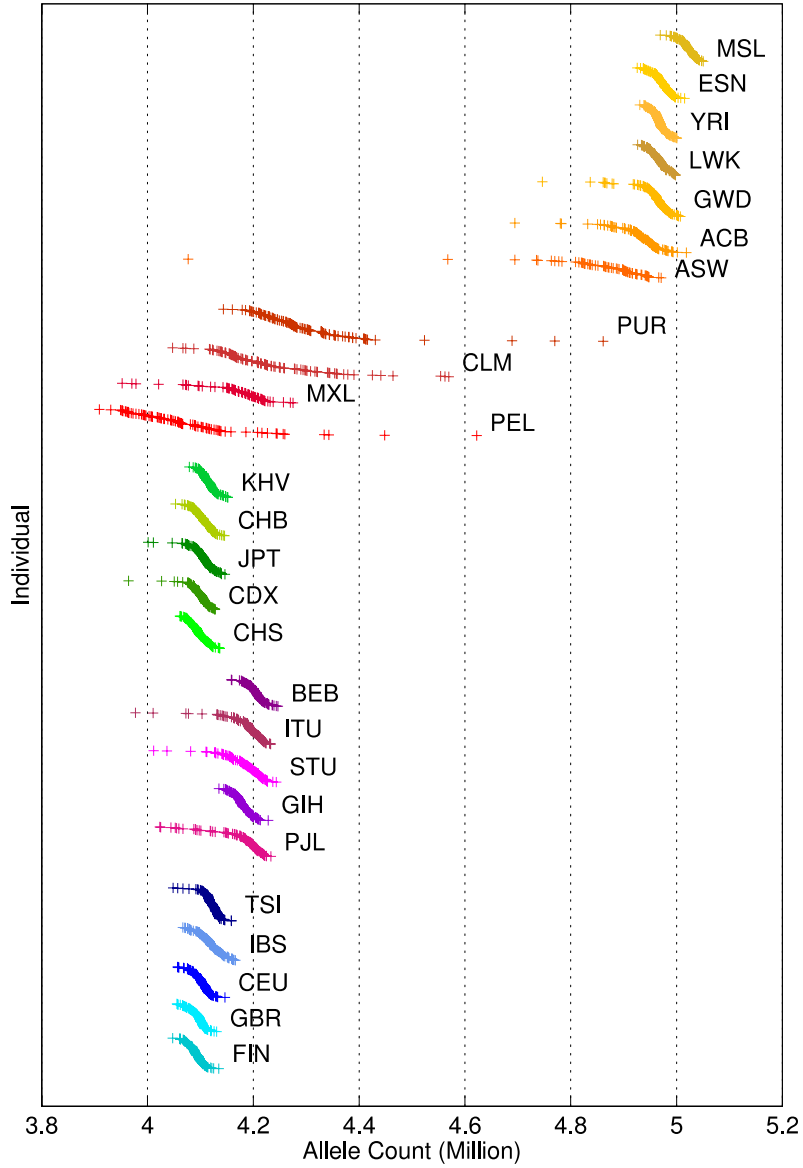
Data Generation Complete:
May 2013

Contribution of the 1000G to dbSNP



Variants per genome

Type	Variant sites / genome
SNPs	$3.8 * 10^6$
Indels	$5.7 * 10^5$
Mobile Element Insertions	~1000
Large Deletions	~1000
CNVs	~150
Inversions	~11



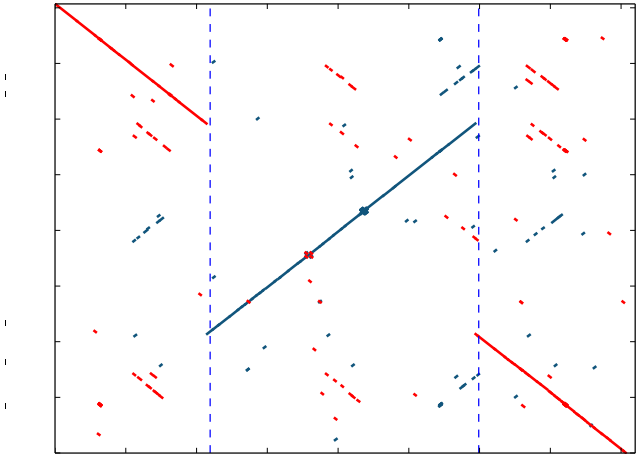
Quality Control of Short Variants

- 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%.

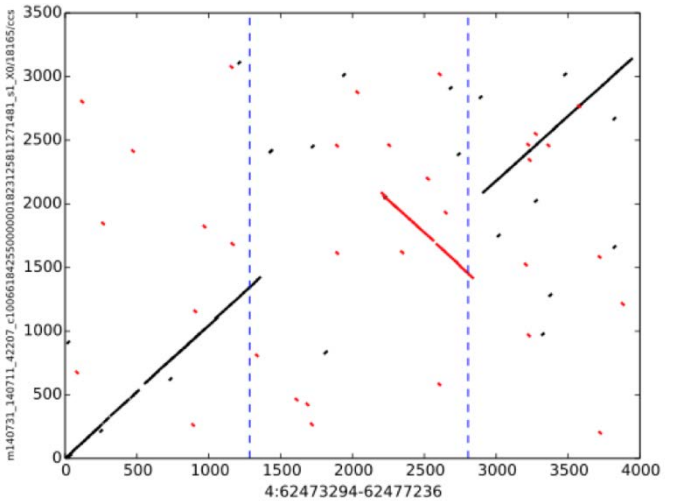
	Haplotype scaffold		MVNcall variants	
Variant Type	Bi-allelic SNPs	Bi-allelic Indels	Multi-allelic SNPs	Multi-allelic indels
Per-allele FDR	4.07%	0.59%	4.91%	4.95%

Verification & further characterization of inversions by PacBio sequencing

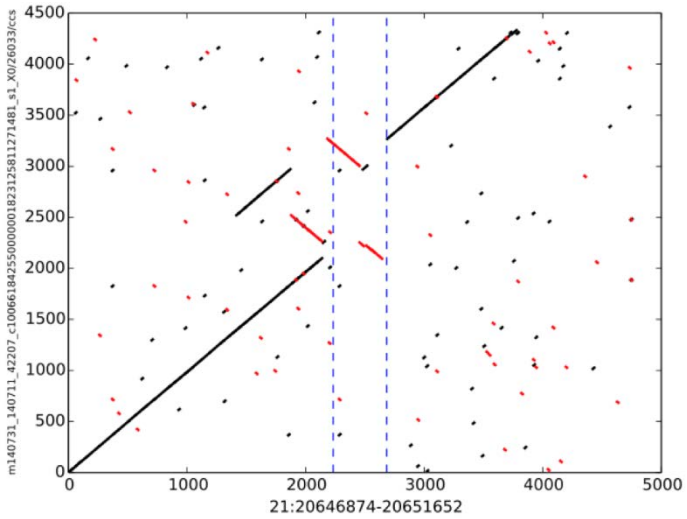
Regular (“simple”) inversion



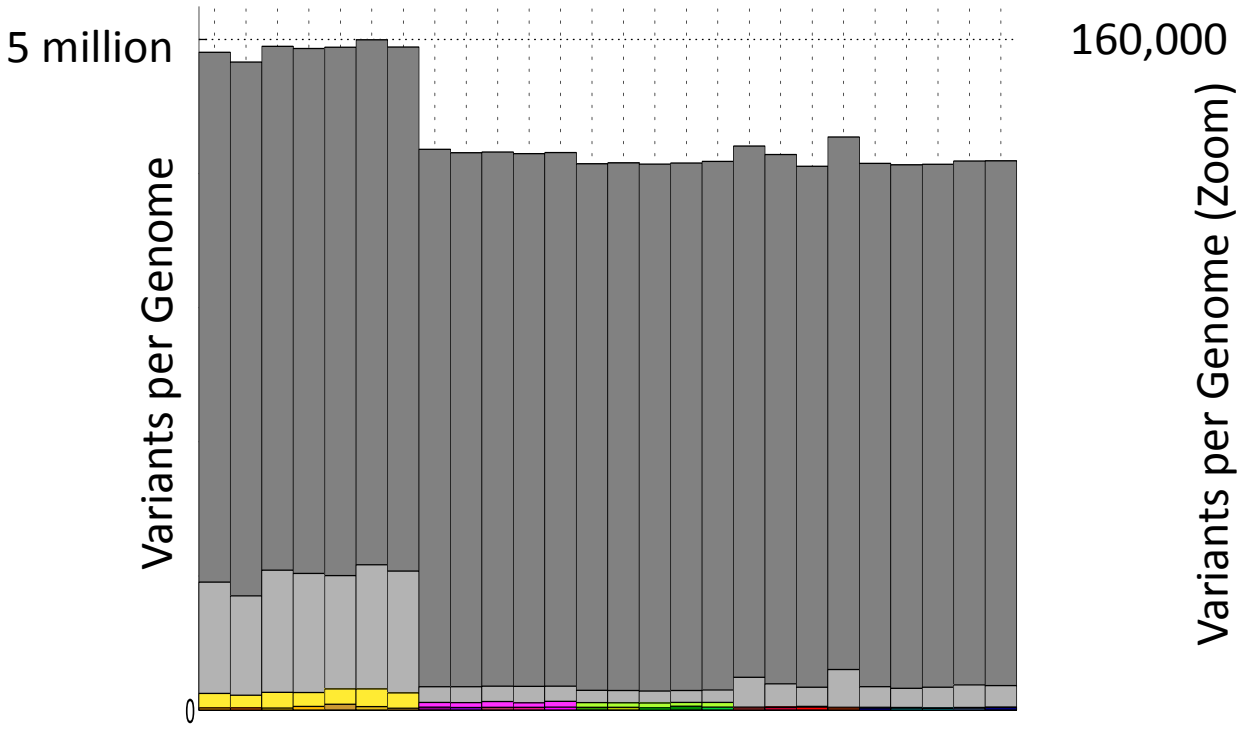
Inversion with flanking deletion



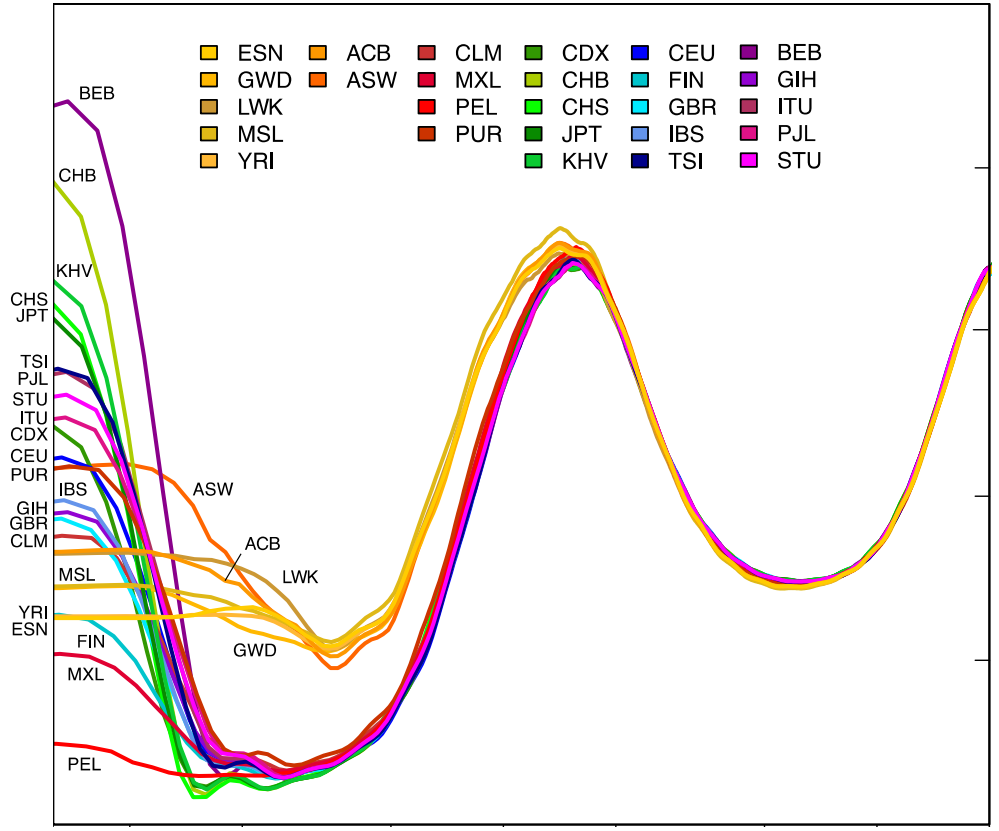
Complex SVs with inverted sequences



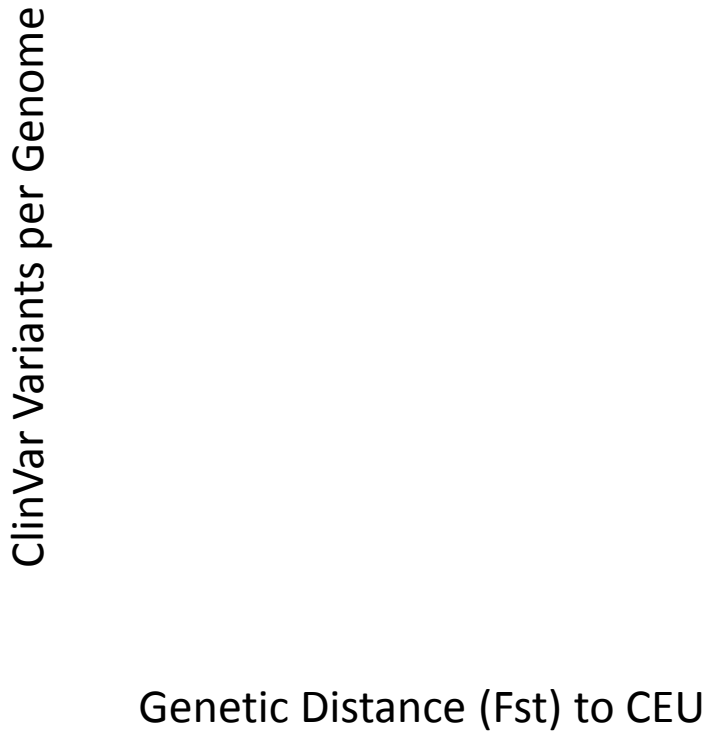
Private vs. Shared Variation (Individual View)



Population histories



Biases in Variation Databases?



Optimal Model for Analyzing 1000 Genomes?

1000 Genomes Call Set (CEU)	Homozygous Reference Error	Heterozygote Error	Homozygous Non-Reference Error
Broad	0.66	4.29	3.80
Michigan	0.68	3.26	3.06
Sanger	1.27	3.43	2.60

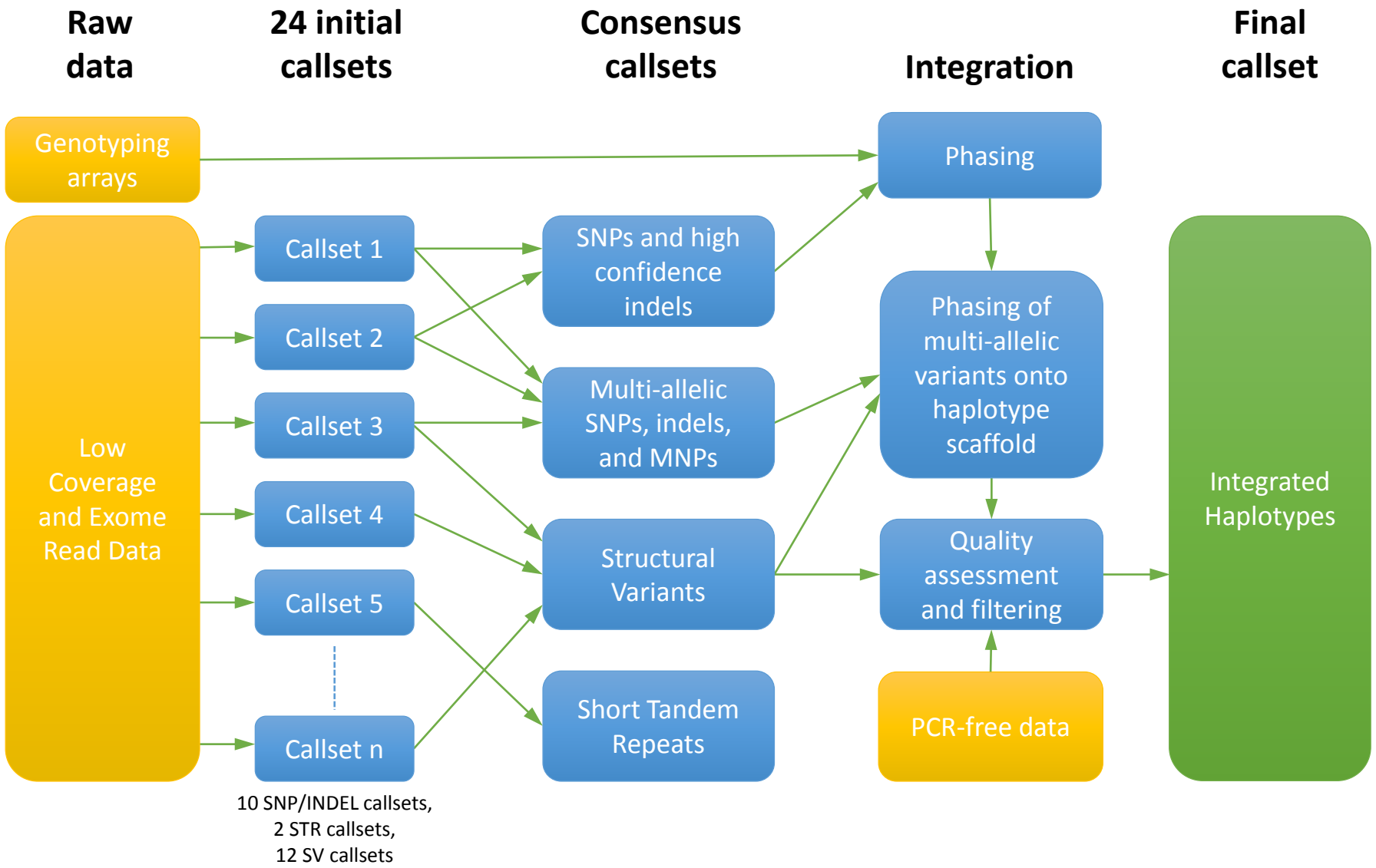
- Michigan caller combines …
  - Markov models to identify shared haplotypes,
  - Classifiers to distinguish true variants from error,
  - Strategies to distribute computation across cluster

Optimal Model for Analyzing 1000 Genomes?

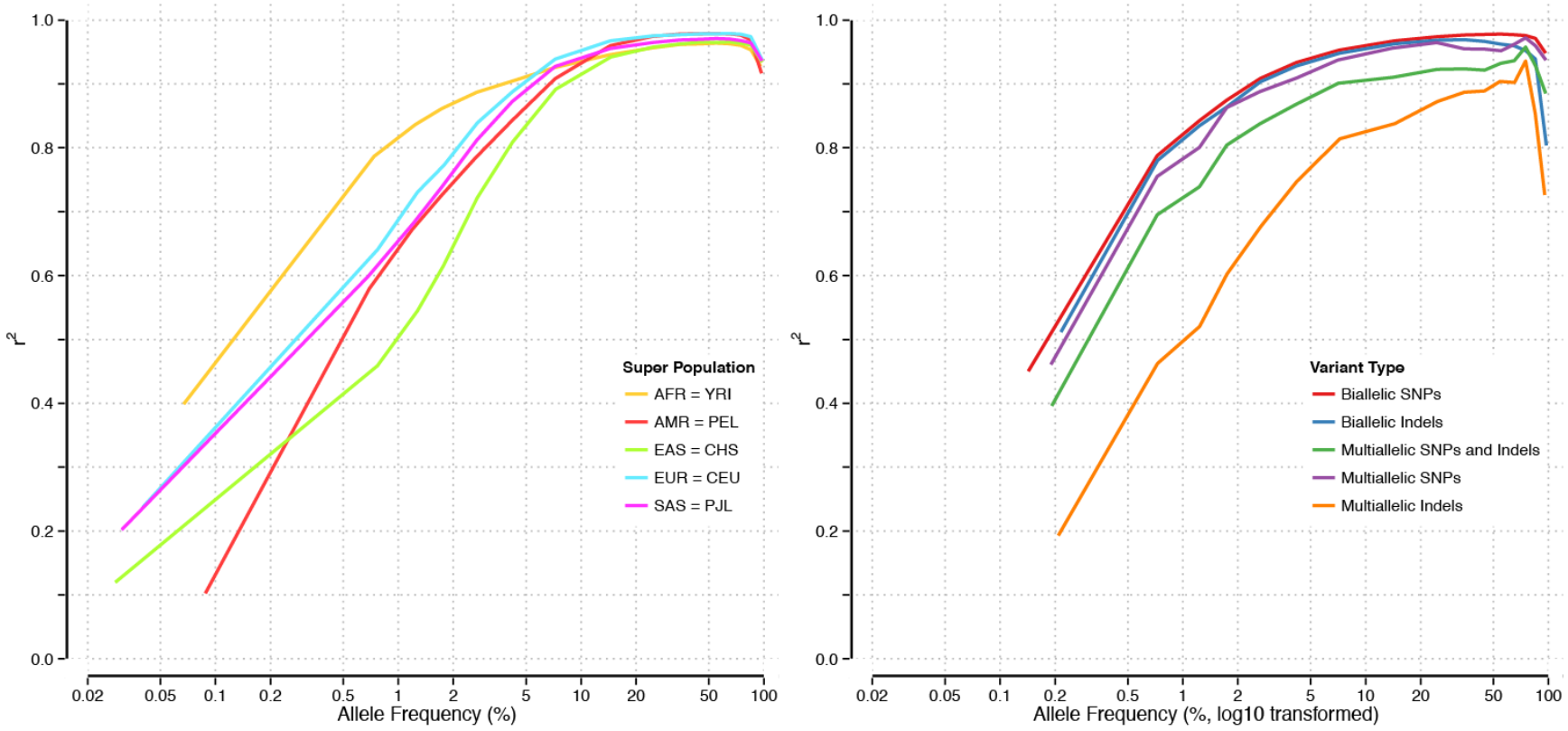
1000 Genomes Call Set (CEU)	Homozygous Reference Error	Heterozygote Error	Homozygous Non-Reference Error
Broad	0.66	4.29	3.80
Michigan	0.68	3.26	3.06
Sanger	1.27	3.43	2.60
Majority Consensus	0.45	2.05	2.21

- Common to see **“ensemble” methods outperform the best single method**

Current 1000 Genomes Analysis Pipeline

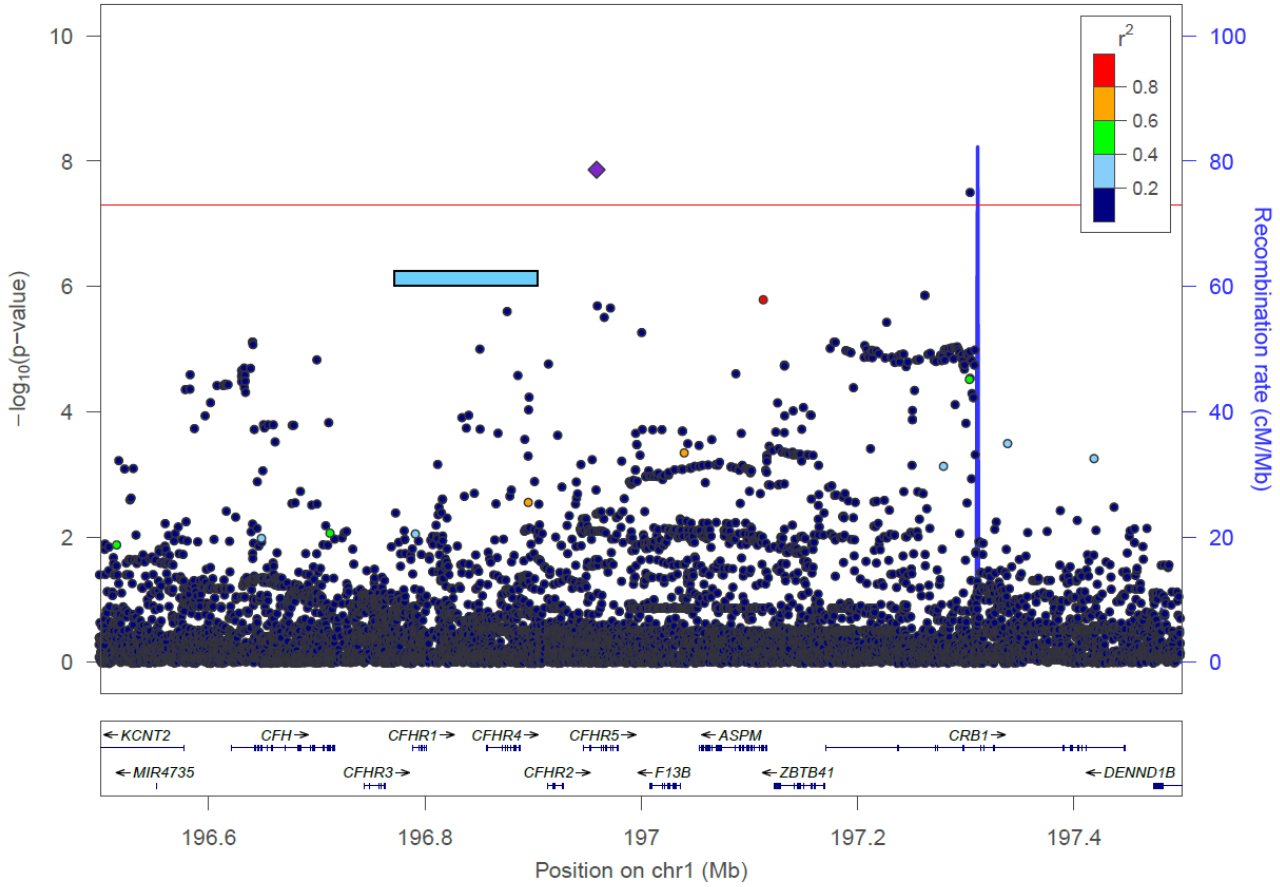


Imputation Accuracy

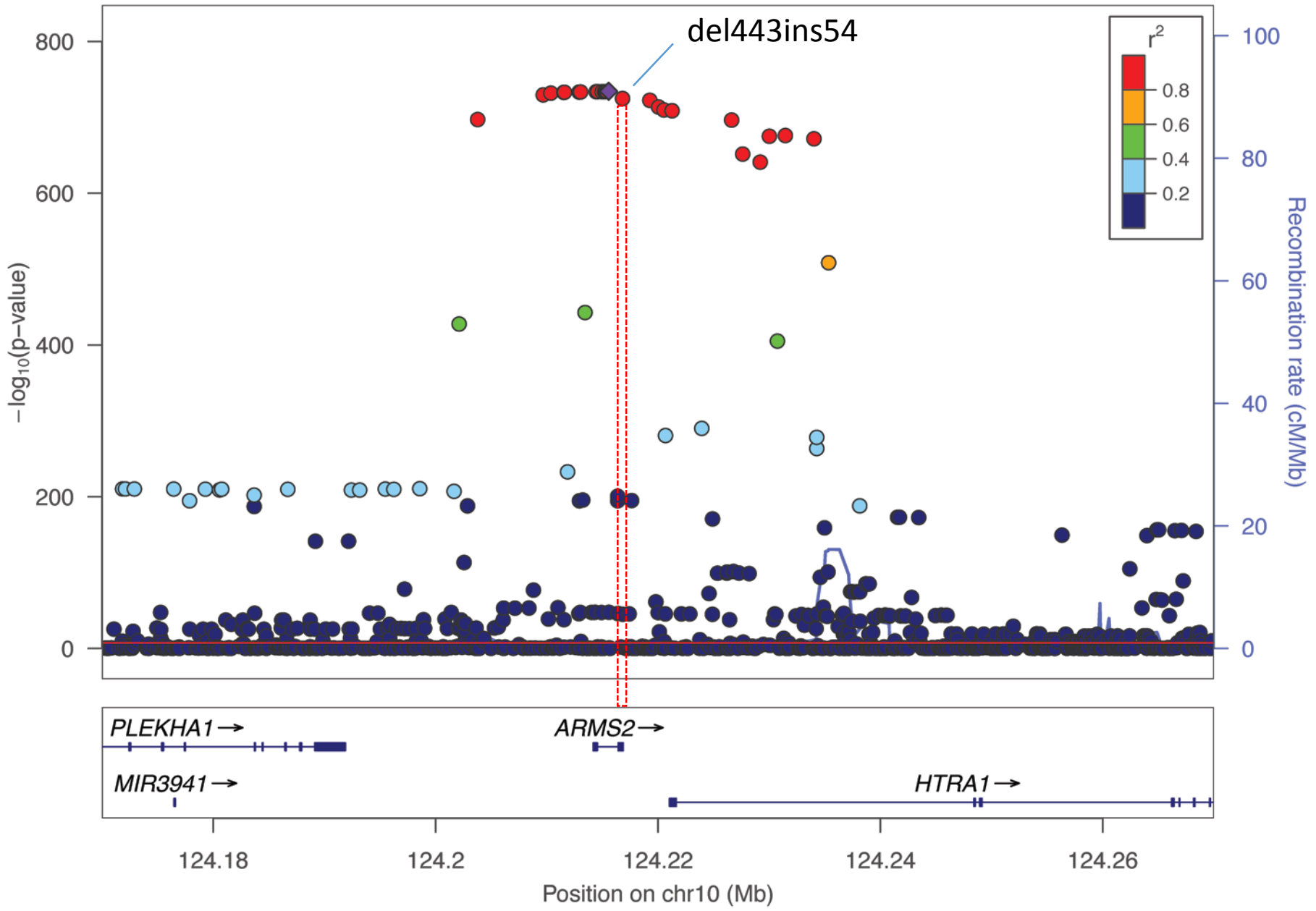


TODO: Multiallelic SNPs and indels to be renamed

AMD Imputation Example #1



Imputation
Example #2



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.

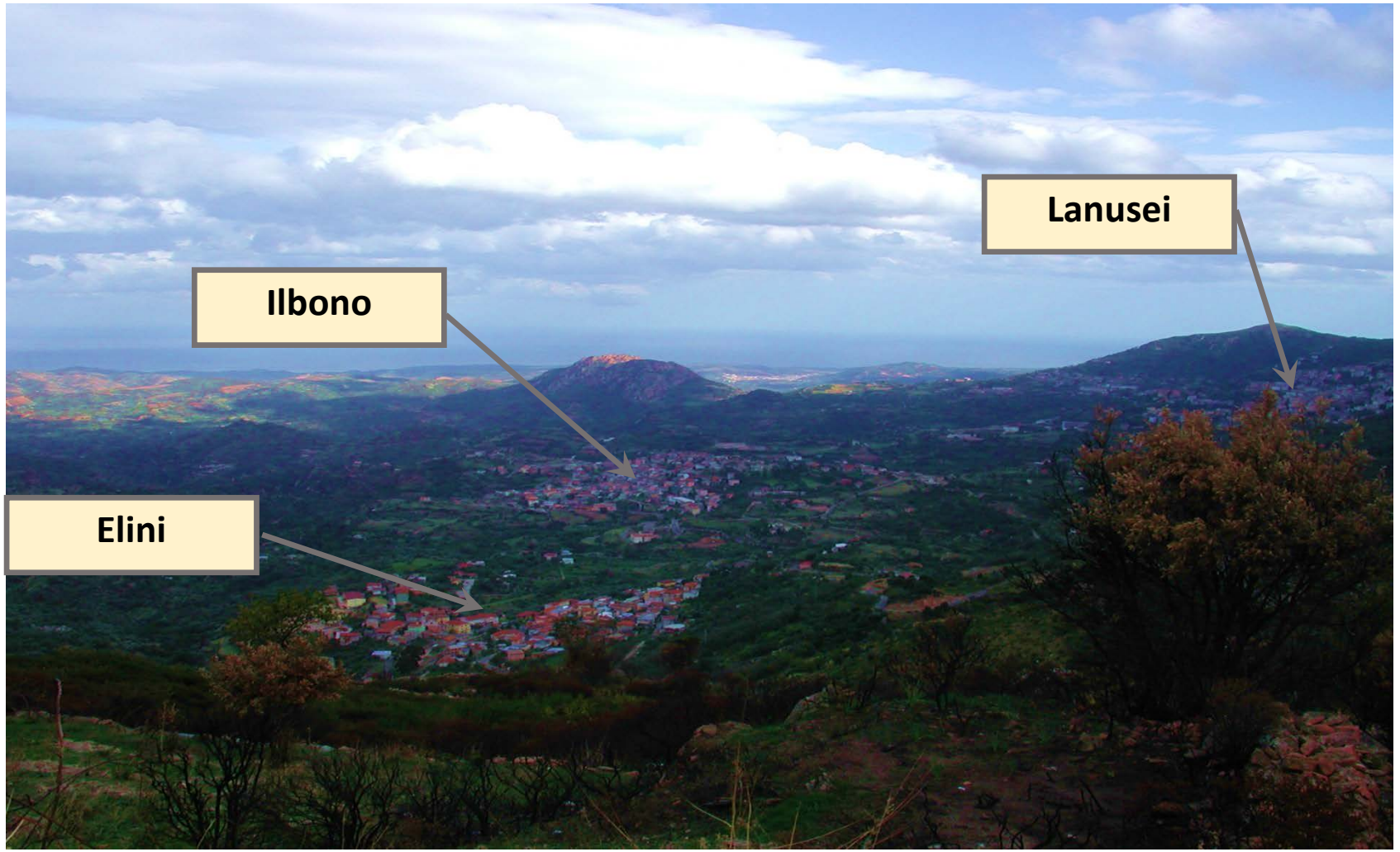
Whole Genome Study in Sardinia

Gonçalo Abecasis

David Schlessinger

Francesco Cucca

Lanusei, Ilbono, and Elini viewed from Arzana



SardiNIA Whole Genome Sequencing

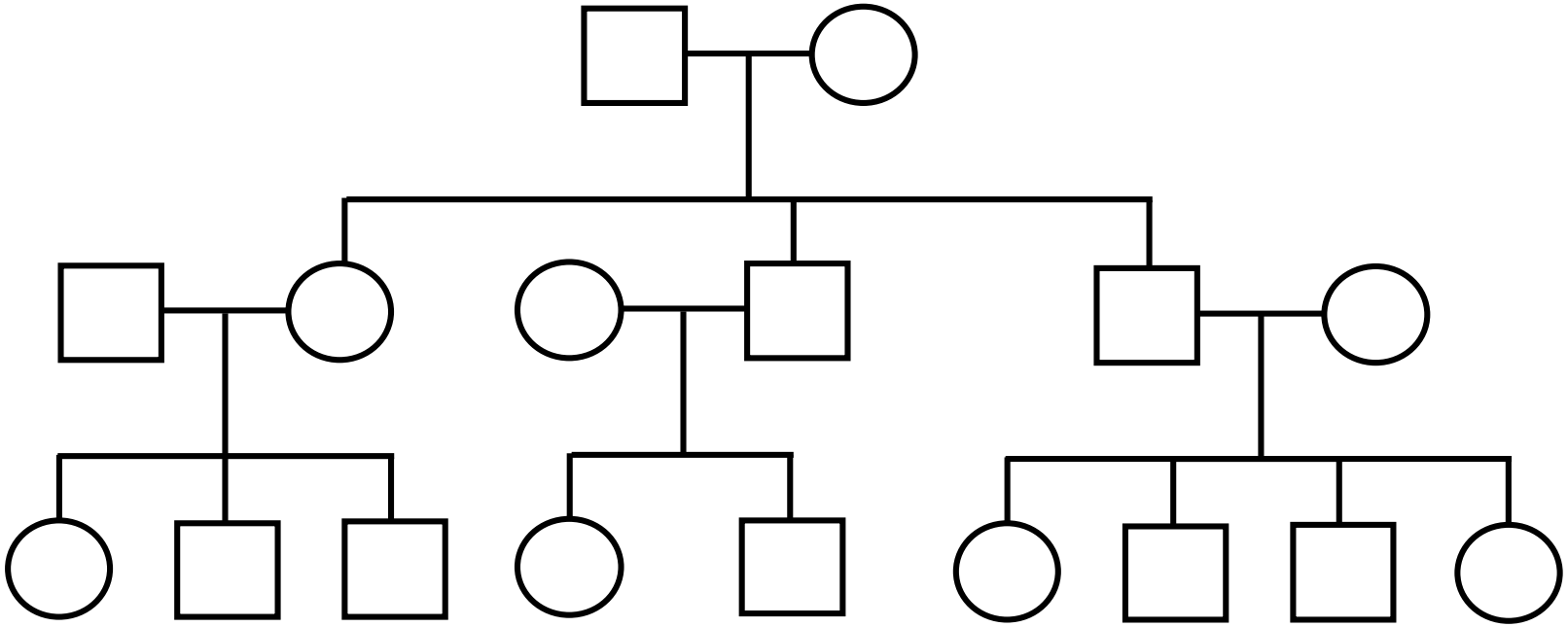
- 6,148 Sardinians from 4 towns in the Lanusei Valley, Sardinia, Italy
  - Recruited among population of ~9,841 individuals
  - Sample includes many close relatives (siblings, cousins, etc.)

- 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?

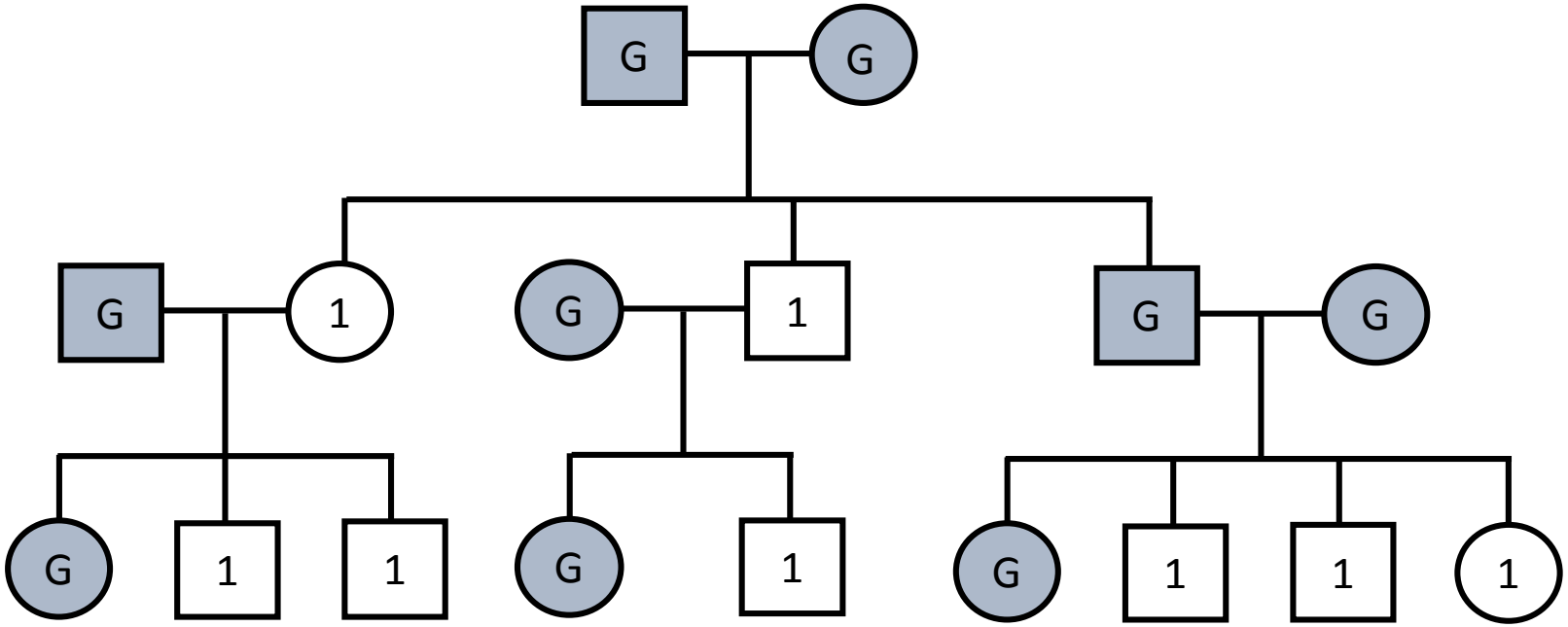
Assuming All Individuals Have Been Genotyped



0 Genomes Sequenced, 0 Genomes Analyzed

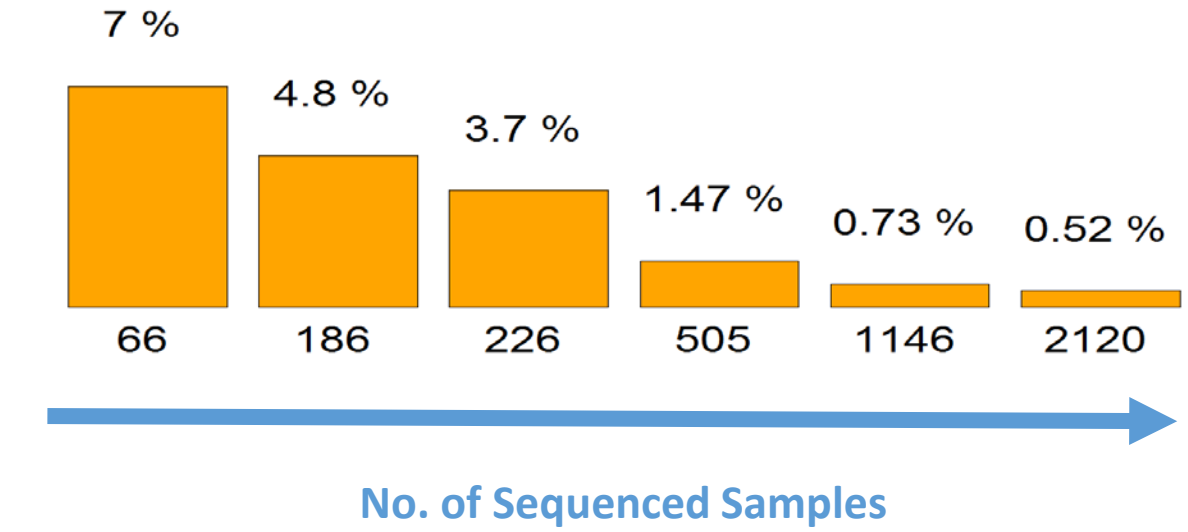
Who To Sequence?

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



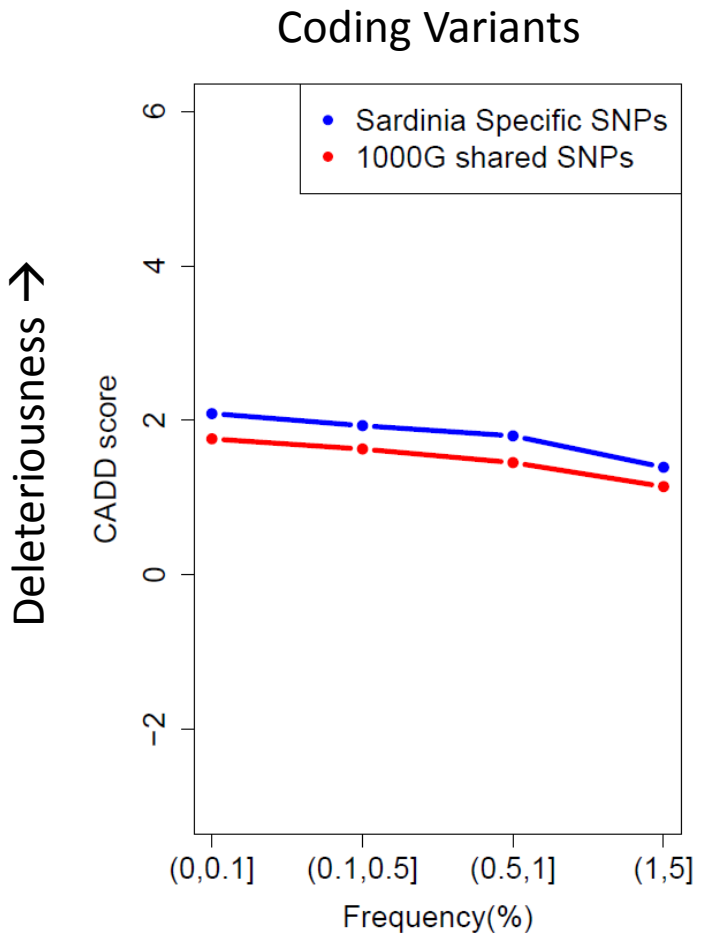
Results of Sequence Analysis

- 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.

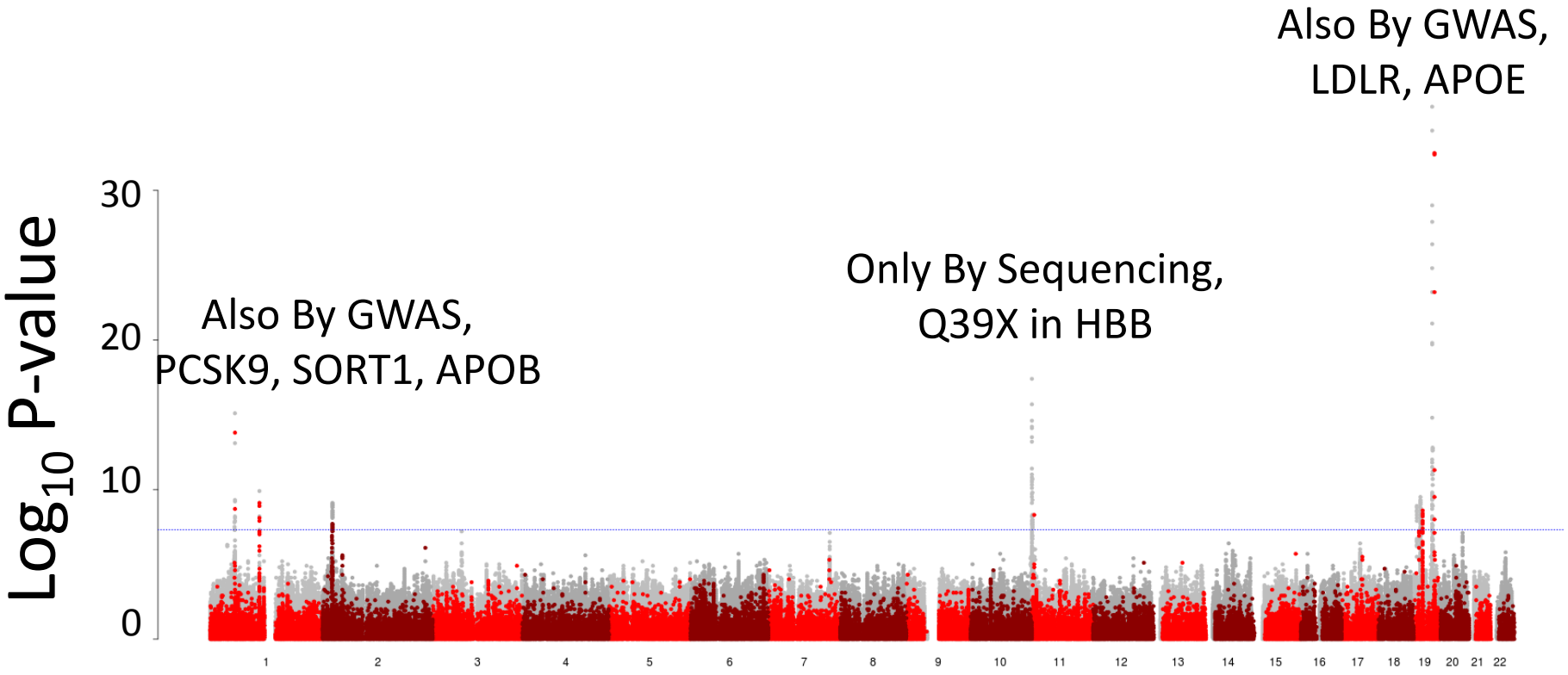
Sardinian variants appear more deleterious



- 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

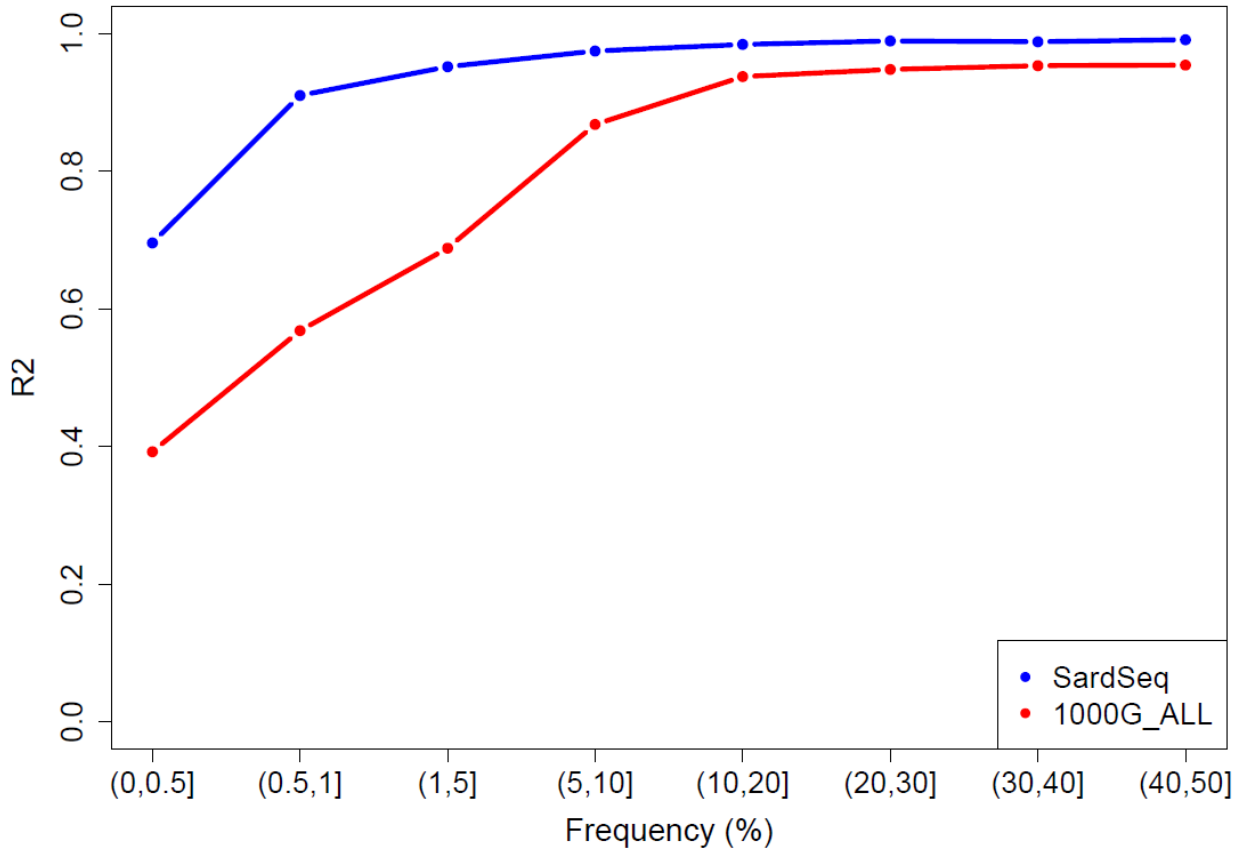


LDL Genetics In Lanusei Valley, Sardinia, Current Sequenced Based View

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

- 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 …



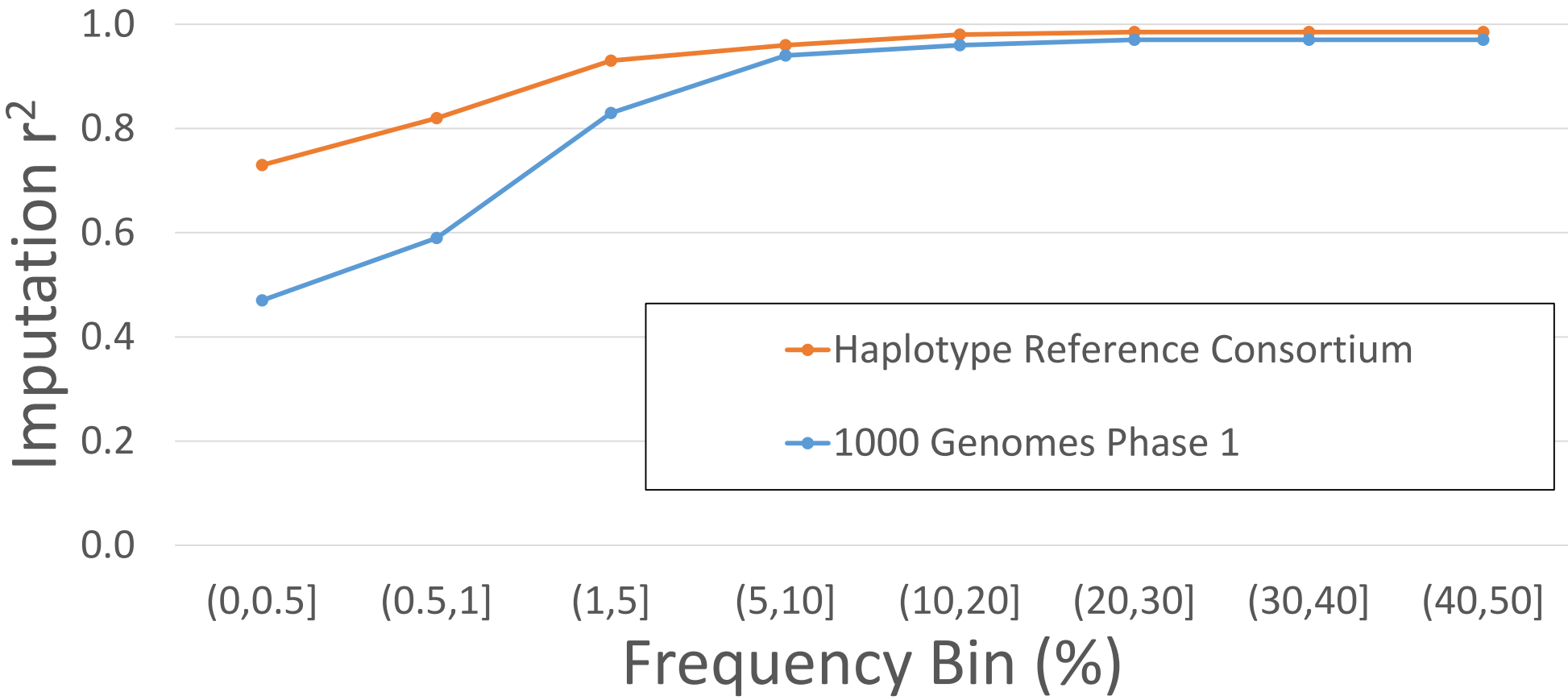
Rare variant imputation in all of Europe?

- 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?

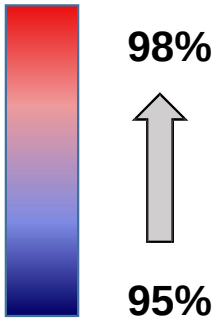
- Haplotype Reference Consortium,
  - with Jonathan Marchini, Richard Durbin, Goncalo Abecasis
  - http://imputationserver.sph.umich.edu/
  - http://haplotype-reference-consortium.org/

Imputation Accuracy using Haplotype Consortium: Preliminary Results

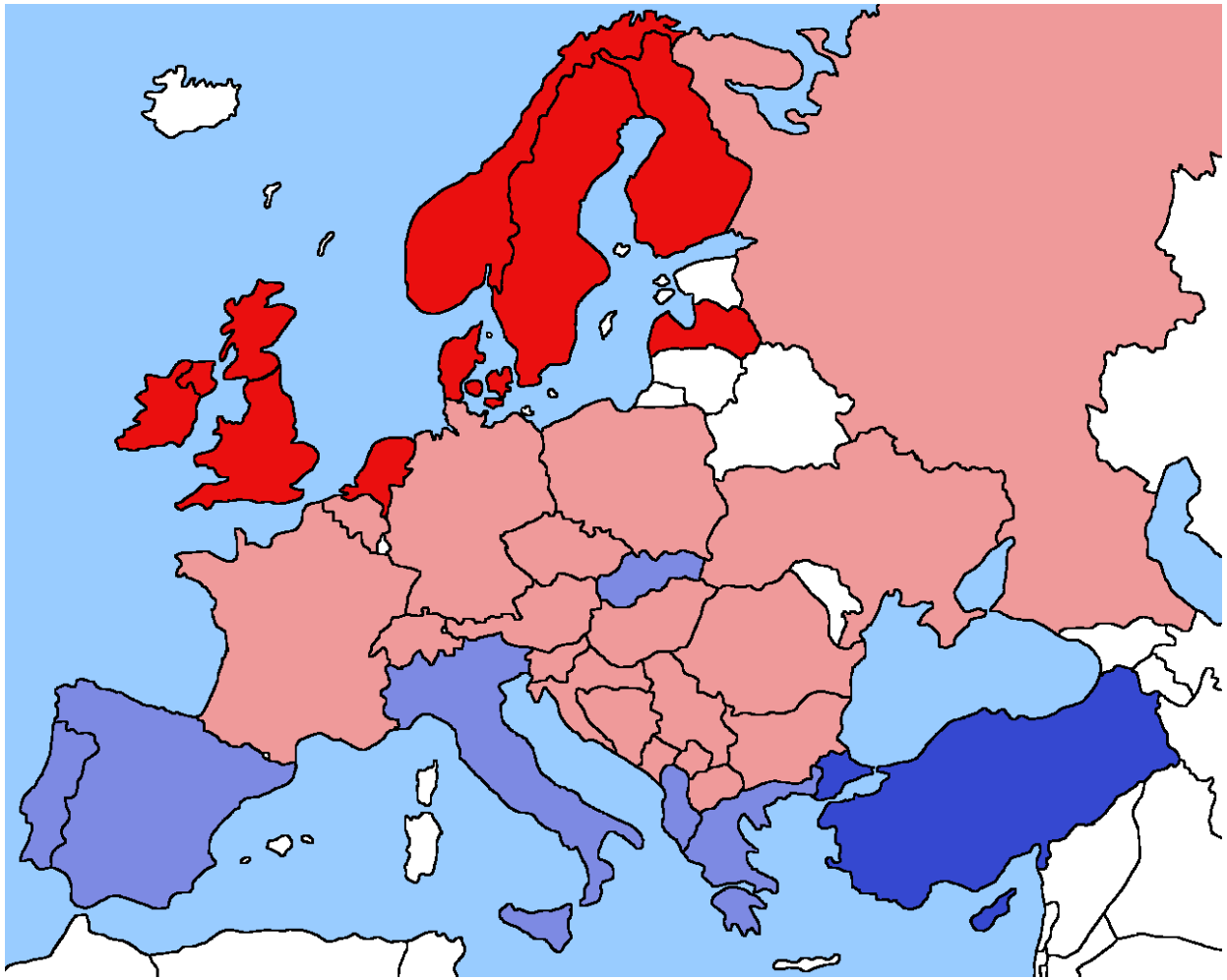


http://www.haplotype-reference-consortium.org

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.

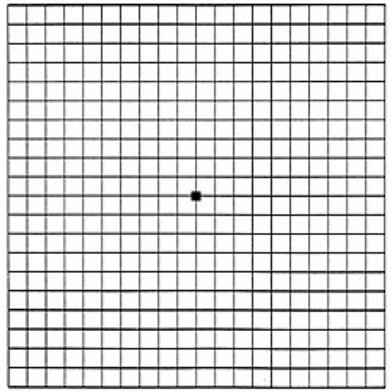
Targeted Sequencing and Genotyping to Study Macular Degeneration

International Age-Related Macular Degeneration Genomics Consortium

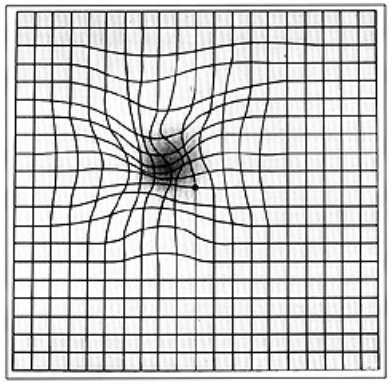
Lars Fritsche, Anand Swaroop, Emily Chew, Dwight Stambolian

Age-Related Macular Degeneration

- Common cause of blindness among the elderly
- Affects >2 million individuals in the United States
- Prevalence increases with old age:
  - ~4% at age 75
  - ~12% at age 80

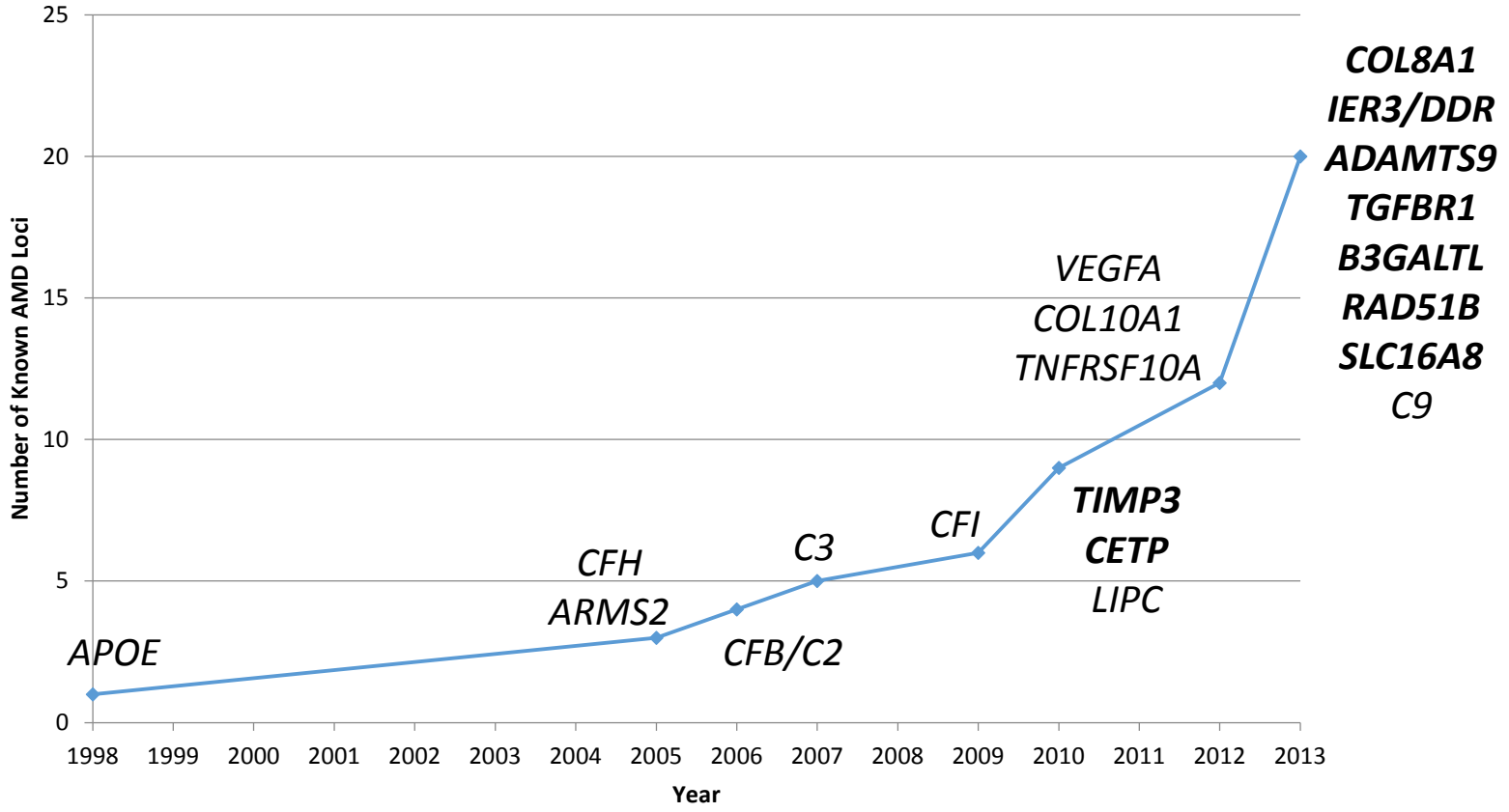


Normal Vision



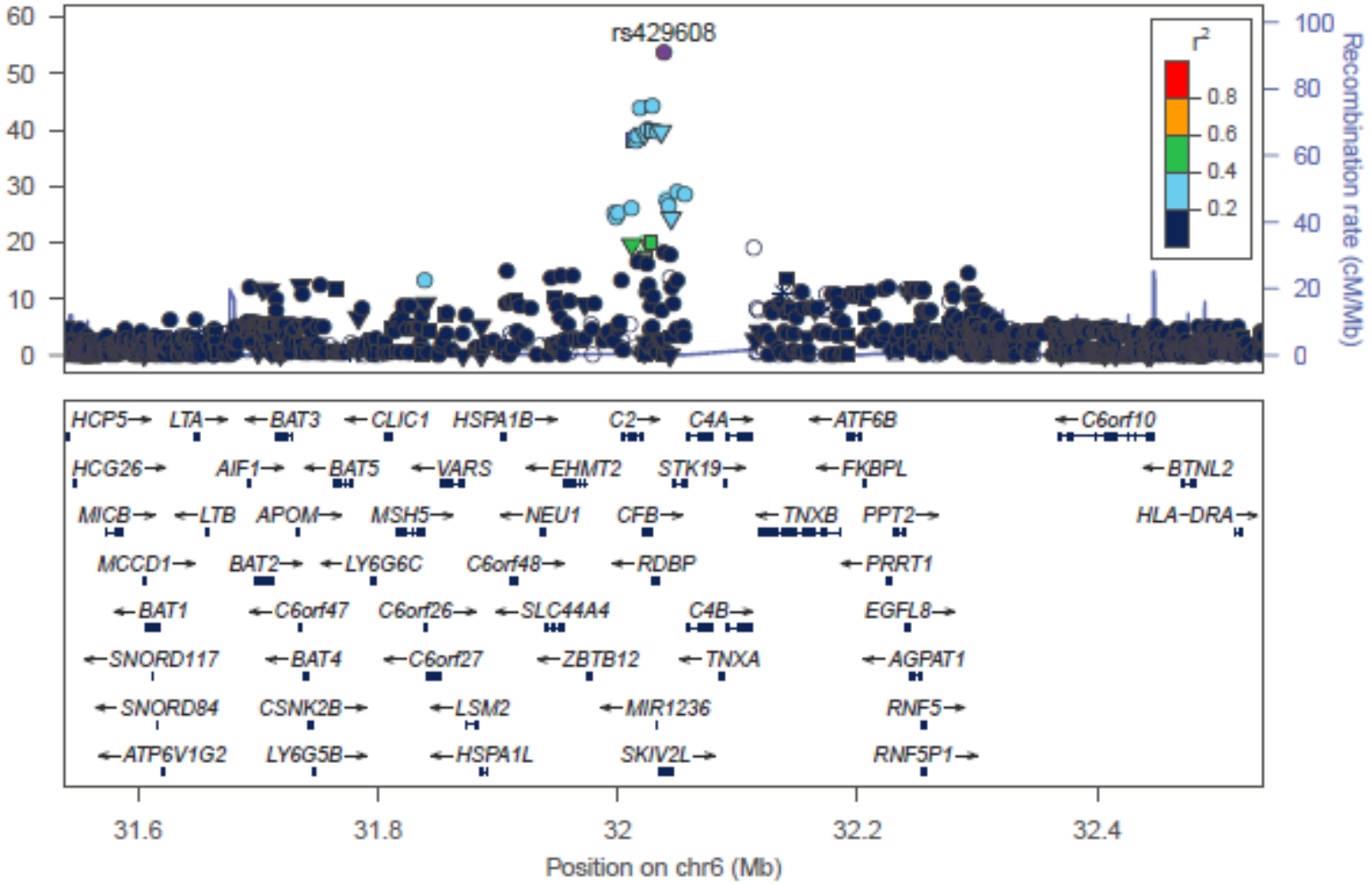
Macular Degeneration

Genetic Risk Factors for Macular Degeneration (1998 – 2013)



Recent updates in Fritsche et al (Nature Genetics, 2013) and Zhan et al (Nature Genetics, 2013).

Age Related Macular Degeneration:
Close-Up of Specific Region



Targeted Sequencing of All Known Risk Loci

- Examine rare variants in known loci to obtain clues about function
  - Cost to carry out search genomewide outside our budget
  - Set out to examine previously identified risk loci

- 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!)

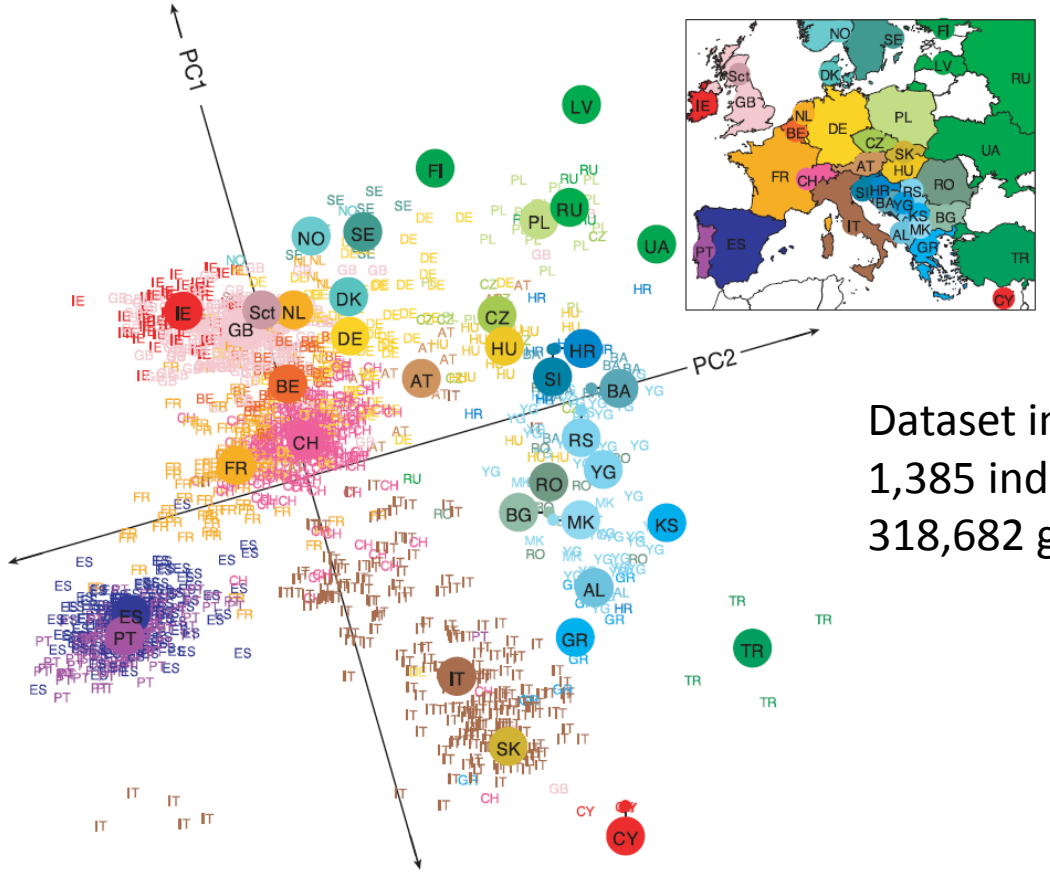
- Studying rare variants, requires very large sample sizes!

Expanding Our Experiment

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

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

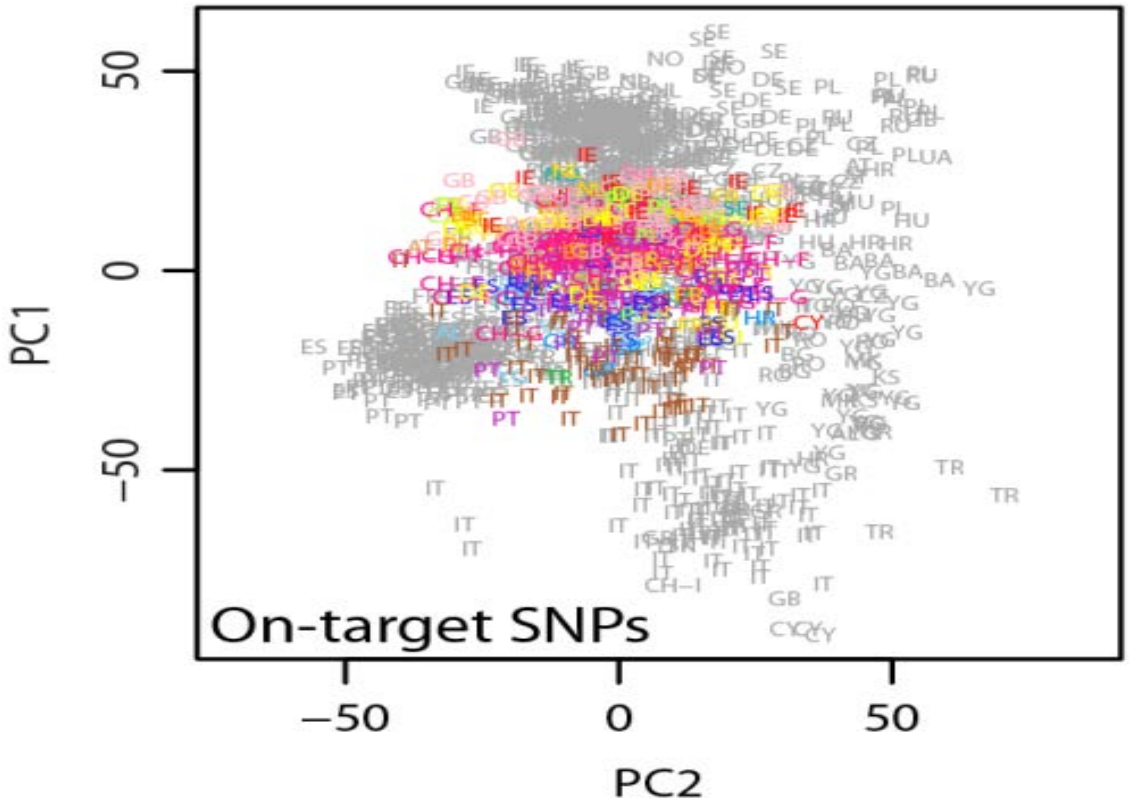
Principal Component Ancestry Map of Europe



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

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

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



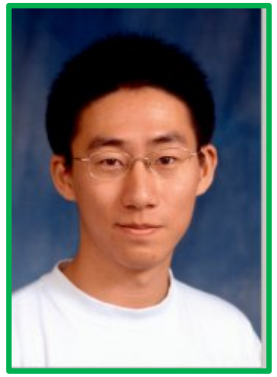
On-target genotypes don’t contain enough information to estimate the ancestry of a sample. The illustration is based on >80x deep whole exome data.

The Problem

- We would like to place individuals on worldwide ancestry map, but …

- Very little information about the genotype of each individual
  - Principal components are weighted sum of genotype
  - Must reflect how well we can reconstruct each genotype
  - Must reflect information about ancestry from each marker
  - Will vary by individual!

- Fortunately, some very smart colleagues helped us develop a solution to this problem.
  - Wang et al (Nature Genetics, 2014) describe a new method for estimating ancestry from sequence data.



Xiaowei Zhan

Chaolong Wang

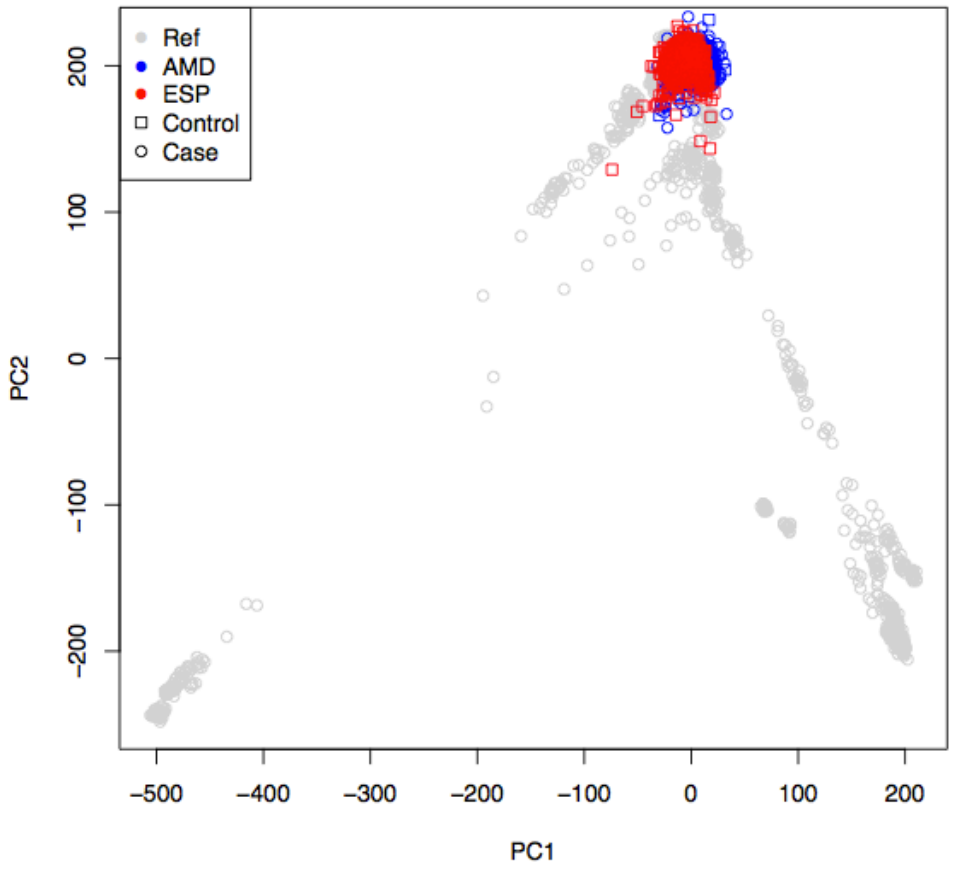
Sebastian Zöllner

Using Ancestry Estimates in Genetic Analysis

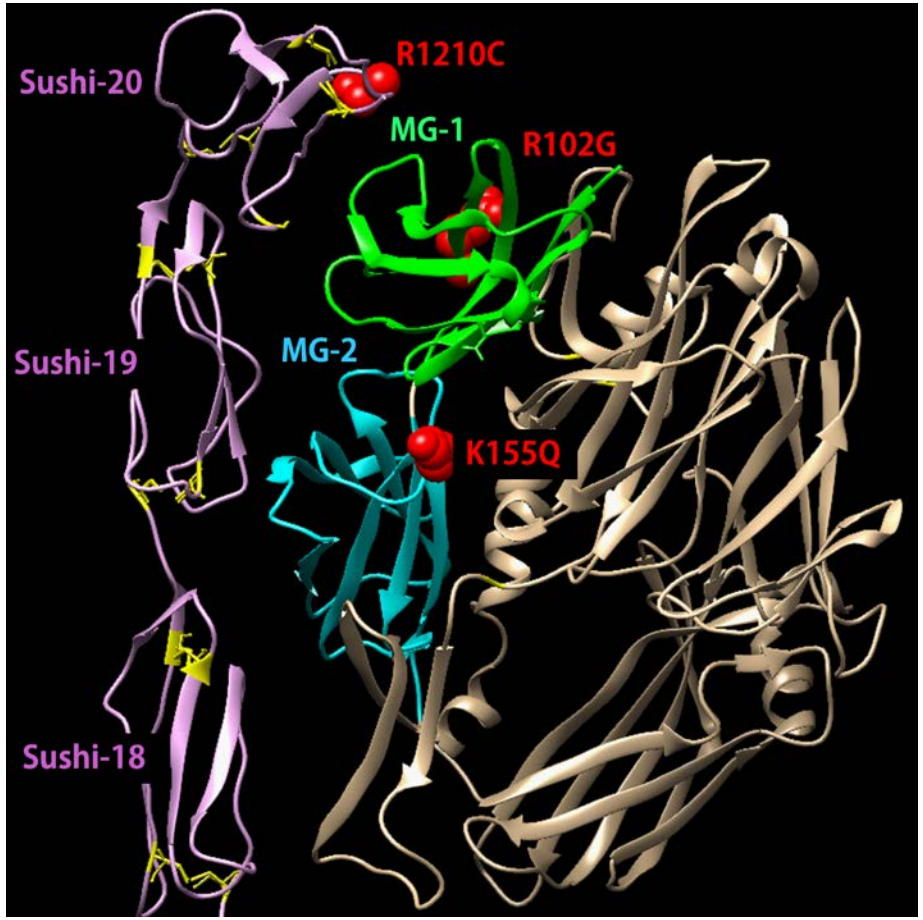
- How to use ancestry estimates in genetic association study?
- Explored possibilities using simulation…
- We recommend using ancestry estimates to find well-matched controls.
  - Overall, better than using ancestry estimates as covariates in analysis.
- As very large numbers of genomes are sequenced, we expect many opportunities to combine information across studies.

Matching Results in our AMD Study

- 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 signal at K155Q in C3 confirmed, reaches $p < 10^{-15}$ after follow-up



AMD Risk Variants in CFH and C3 ….



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

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

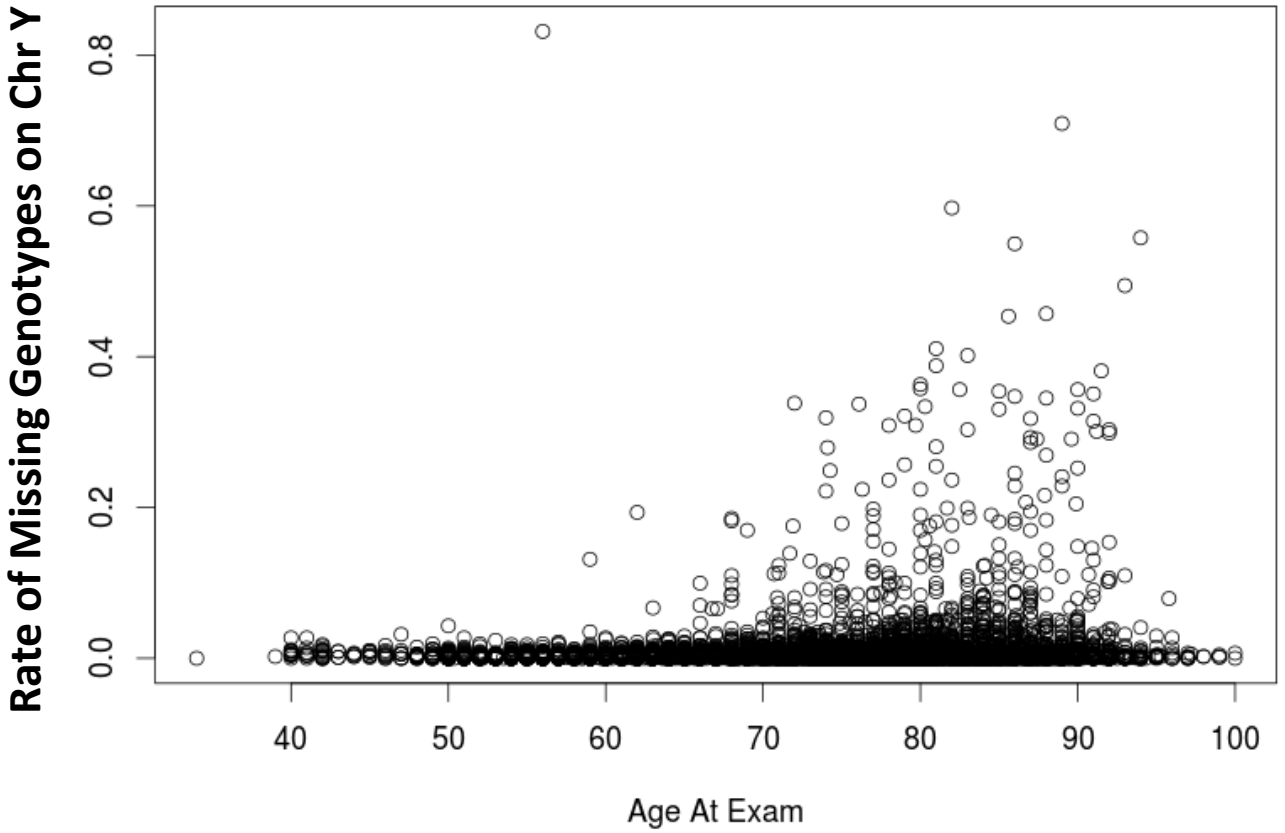
- CFH inactivates C3 to downregulate alternate complement pathway

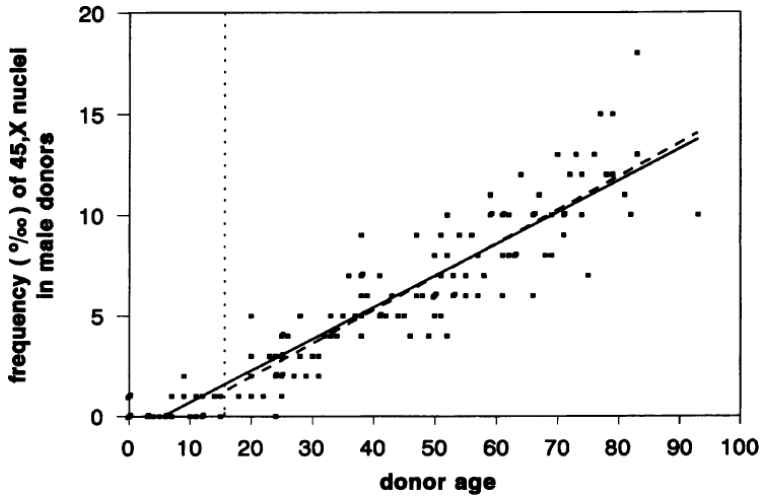
Poor Man's Sequencing …

- We have been using exome arrays to further study the role of rare variation in age-related macular degeneration

- We have genotyped >16,000 advanced cases of macular degeneration and >17,000 controls

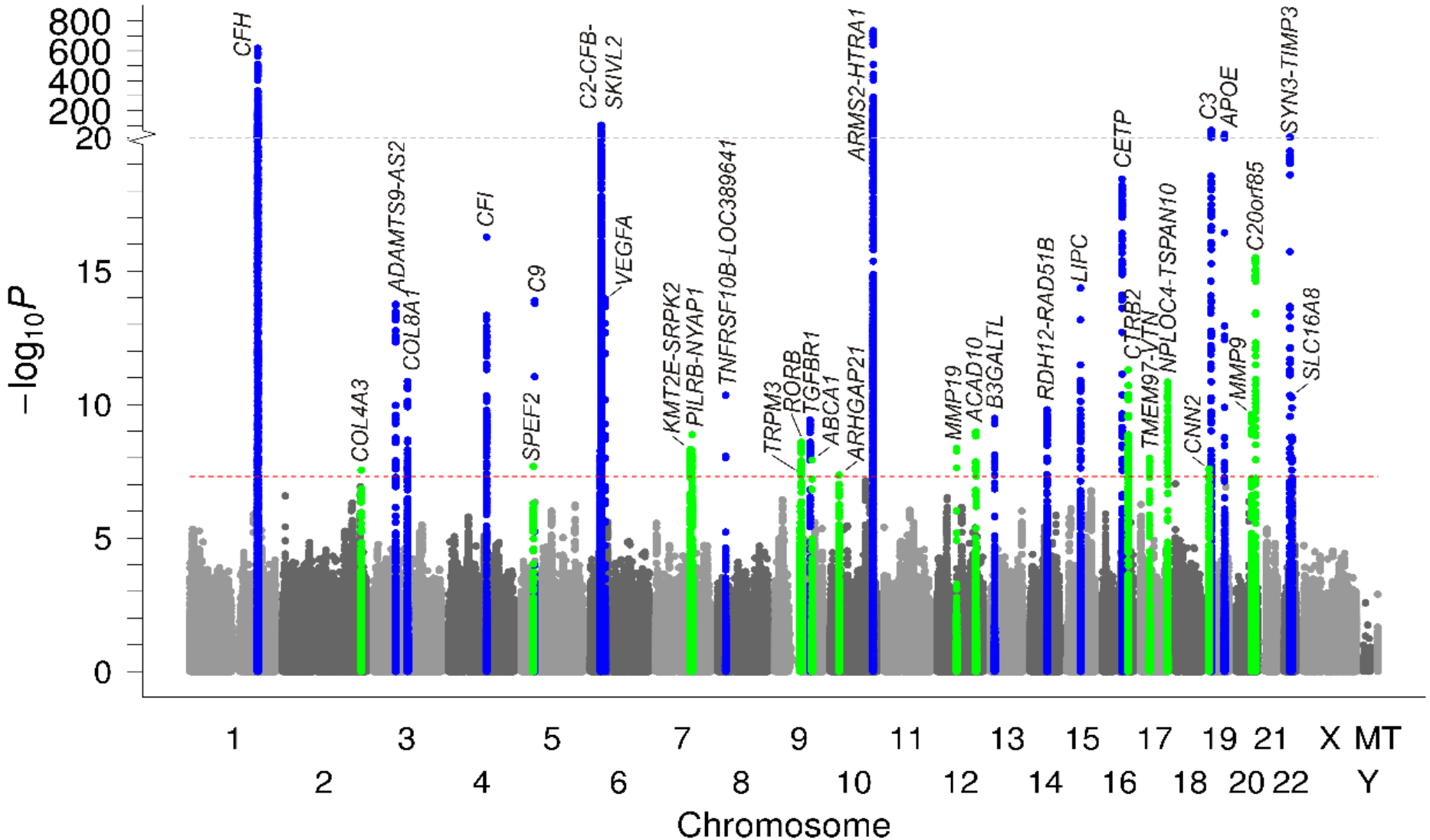
Second Step QC: Age-dependent Y-Chromosome Loss



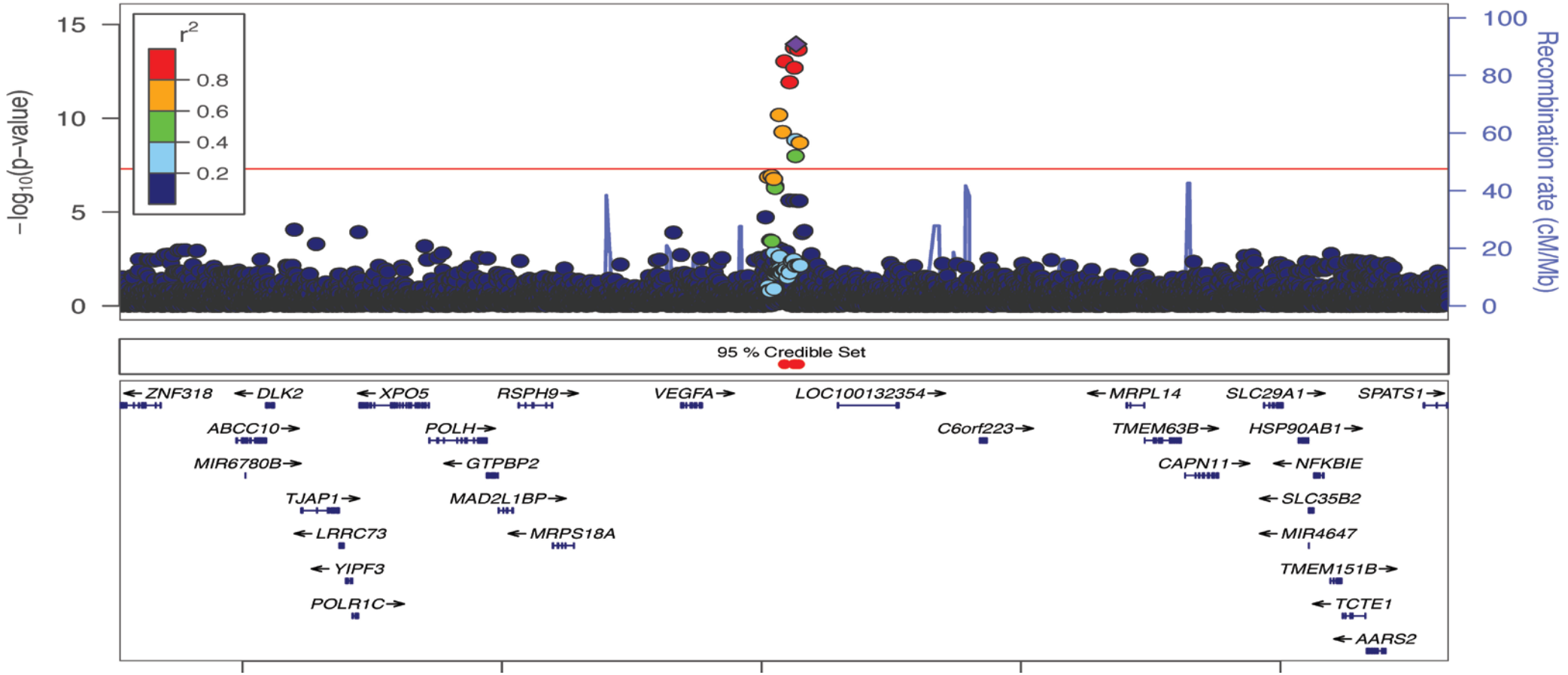


Guttenbach et al., Sex chromosome loss and aging: in situ hybridization studies on human interphase nuclei. Am J Hum Genet. 1995 Nov;57(5):1143-50. PubMed PMID: 7485166

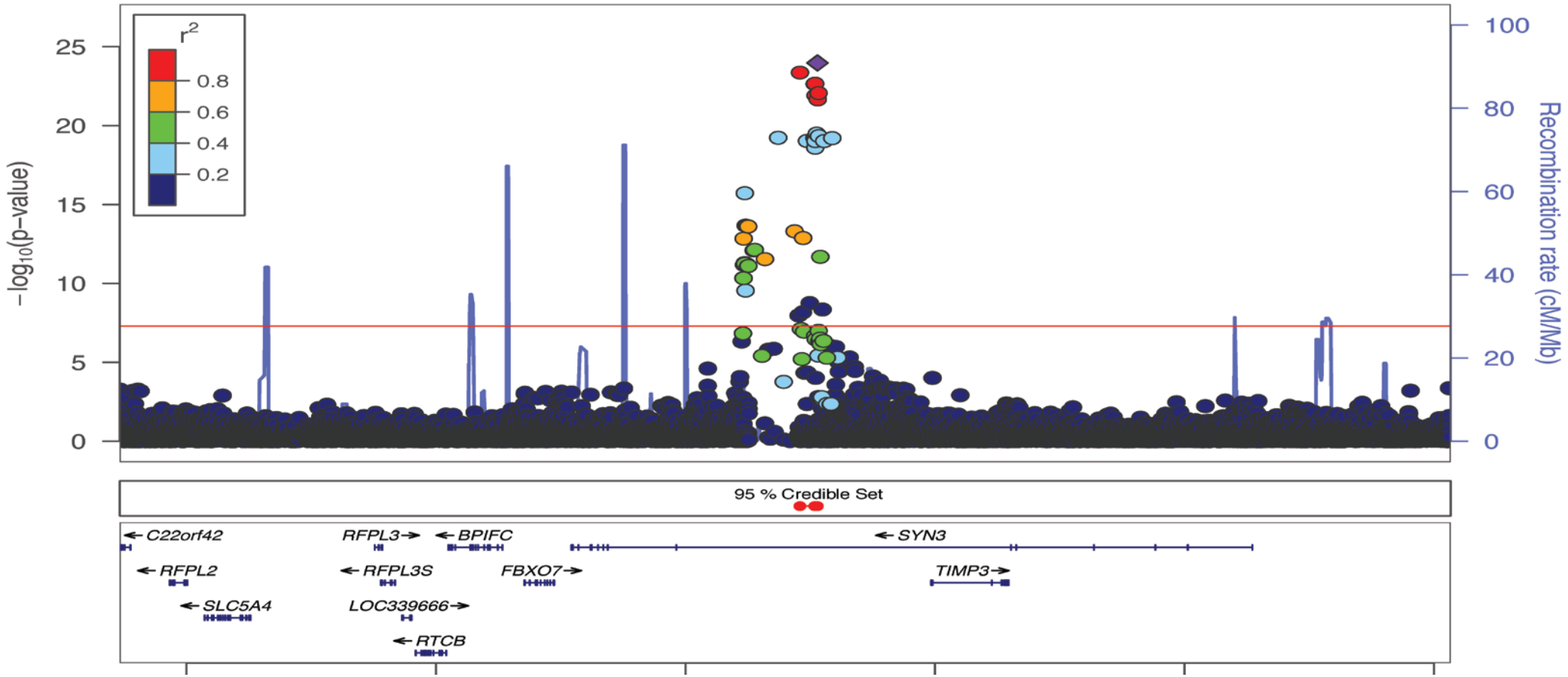
Macular Degeneration, Comparison of Case and Control Genomes



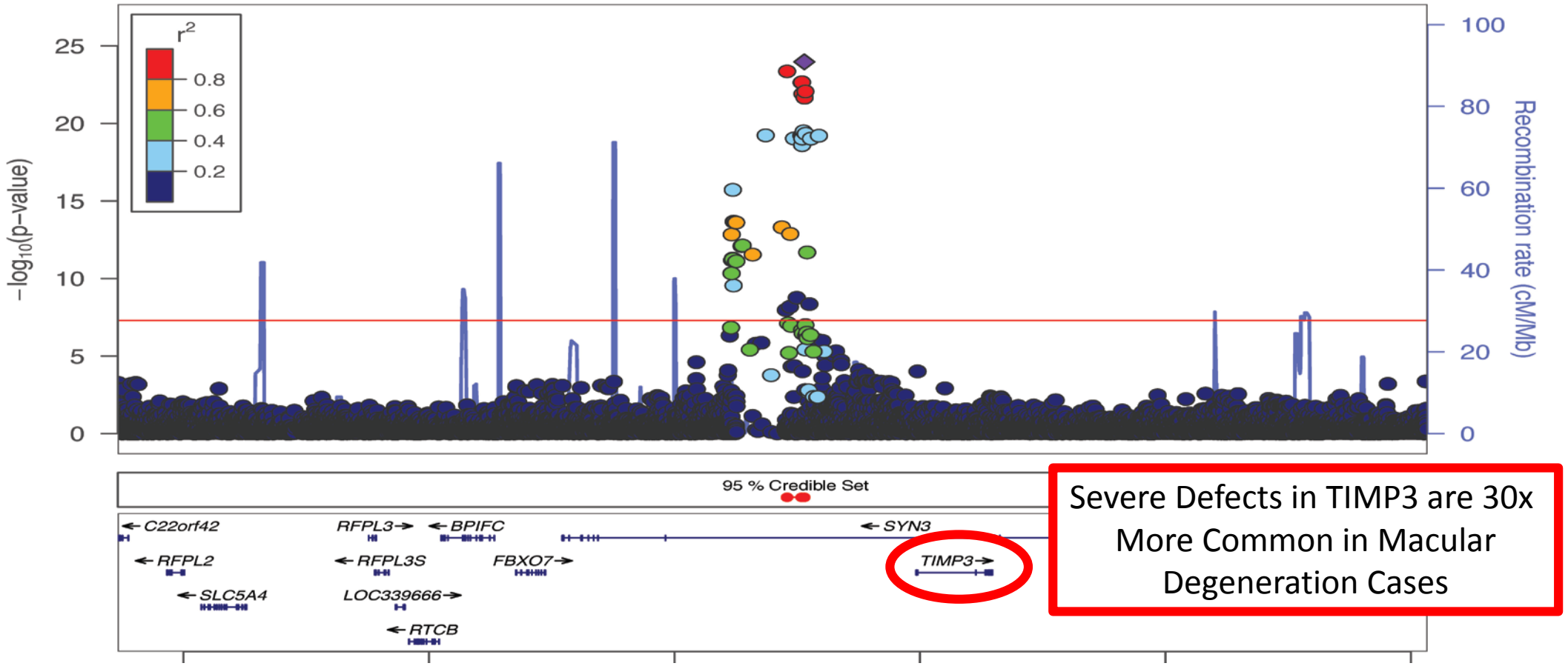
Comparison around VEGFA gene



Comparison in a region of chromosome 22



Comparison in a region of chromosome 22



Severe Defects in TIMP3 are 30x More Common in Macular Degeneration Cases

Rare TIMP3 variants and AMD

Amino Acid	Allele Count		Design
	AMD N = 16,144	Controls N=17,832	
Ser38Cys	14	0	Cystine Disrupting Variant
Gly58Cys	1	0	
Tyr109Cys	1	0	
Arg132Cys	2	0	
Gly173Cys	0	1	
Glu162Lys	1	0	Reported Mendelian Variant
His181Arg	5	0	
Ser204Cys	4	0	
	28	1	

OR = 30

$p = 10^{-8}$

Rare TIMP3 variants and AMD

Amino Acid	Allele Count		Design
	AMD N = 16,144	Controls N=17,832	
Ser38Cys	14	0	Cystine Disrupting Variant
Gly58Cys	1	0	
Tyr109Cys	1	0	
Arg132Cys	2	0	
Gly173Cys	0	1	
Glu162Lys	1	0	Reported Mendelian Variant
His181Arg	5	0	
Ser204Cys	4	0	
	28	1	

OR = 30

$p = 10^{-8}$

Across loci, most trait associated rare variants have frequency <0.1% …

Rare TIMP3 variants and AMD

Amino Acid	Allele Count		Design
	AMD N = 16,144	Controls N=17,832	
Ser38Cys	14	0	
Gly58C			ystine
Tyr109C			srupting
Arg132C			ariant
Gly173C			
Glu162Lys	1	0	Reported
His181Arg	5	0	Mendelian
Ser204Cys	4	0	Variant
	28	1	

Coding variation is well understood.

How will we interpret and analyze the other 99% of rare variants?

OR = 30

$p = 10^{-8}$

Poor Man's Sequencing …

- We have been using exome arrays to further study the role of rare variation in age-related macular degeneration
- We have genotyped >16,000 advanced cases of macular degeneration and >17,000 controls
- What do we see?
  - 45 independent common variant signals (with frequency >1%)
  - 7 independent rare variant signals (with frequency <1%)
  - Three genes with excess burden of rare variation among cases …
    - In all of these, disease associated rare variants each have frequency <0.1%
- Common variants explain 30% of disease risk, rare variants explain 1% of disease risk

Notes …

- Studies of rare variants may often require even larger sample sizes than studies of common variation

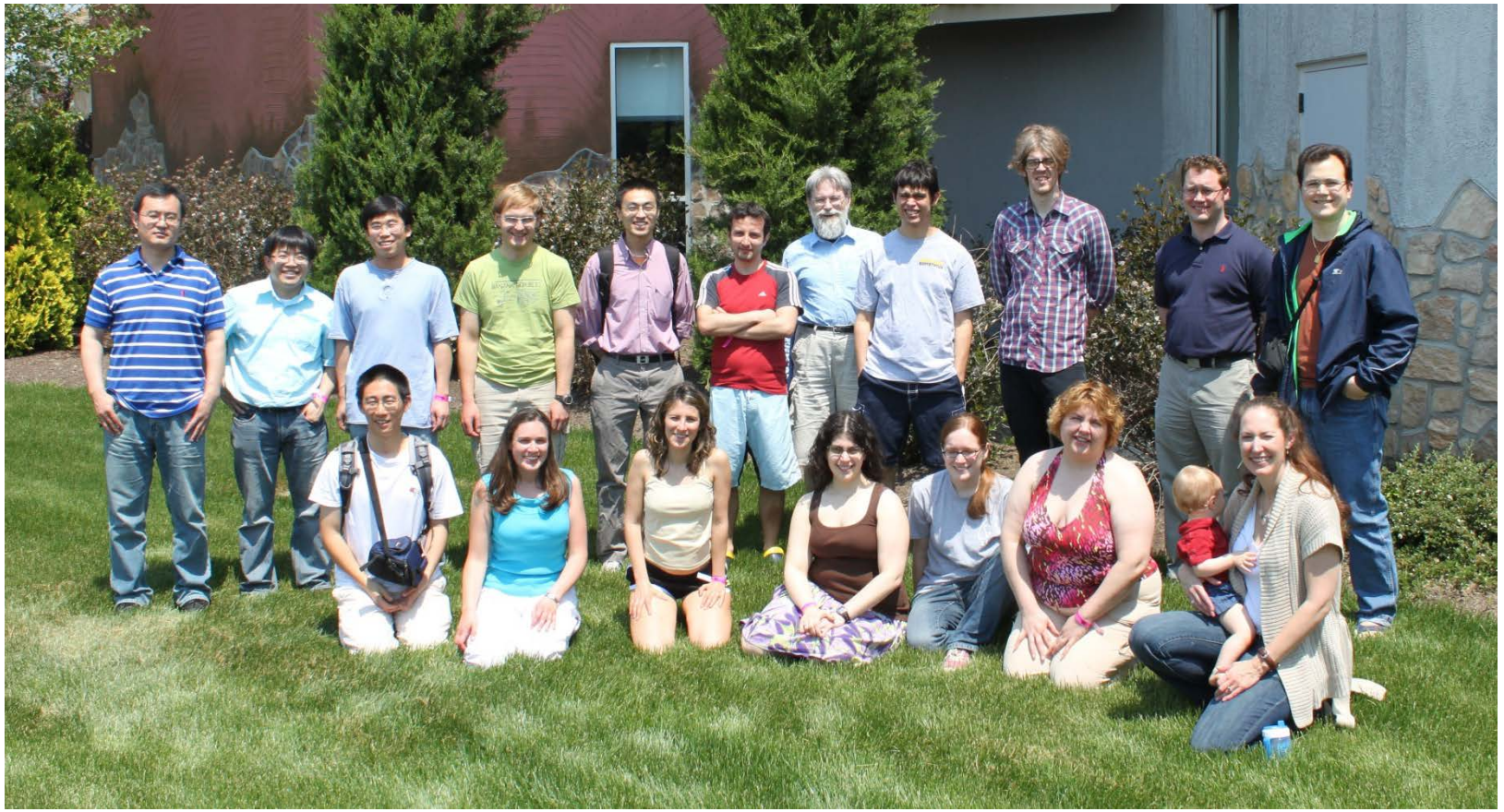
- In our experience, rare variants don’t account for much missing heritability…
- … but they can clarify disease biology and mechanisms.

- Combining sequencing information and results across studies can help reach the sample sizes necessary for new discoveries

- Creative uses of array genotyping technologies can also be extremely powerful.

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