Human Genetic Studies: Challenges and Opportunities

Goncalo Abecasis

Ann Arbor, MI

Goal of Human Genetic Studies

Find biological processes that,
when changed, alter disease course

Understand Disease:
Enable new treatments

Predict disease:
Enable early prevention and early decision making

How human genetic studies work …

- DNA is our instruction manual
- We are all built mostly to the same plan…
  - Any two human DNA molecules are ~99.9% the same
- We each have our manual, with small variations from the typical plan
  - Some variations are common and typically have small consequences
  - Many variations are rare and these can have more severe consequences

Human Genetics, Study Sizes over My Time

Year	No. of Samples	No. of Markers	Publication
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, Study Sizes over My Time

Year	No. of Samples	No. of Markers	Publication
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		Genetics)
2007	Hundred		
2005	Hundred		
2003	Hundred		ap (Nature Genetics)
2002	Hundred		ap (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, Study Sizes over My Time

Year	No. of Samples	No. of Markers	Publication
2012	1,092	40 million	The 1000 Genomes Project (Nature)
2010	Hundreds	[illegible]	The 1000 Genomes Project (Nature)
2010	~100,000		
2008	~9,000		enetics)
2007	Hundreds		
2005	Hundreds		
2003	Hundreds	[illegible]	[illegible] (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, but understanding their implications is challenging.

Current State of Genetic Association Studies

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

Global Lipids Genetics Consortium

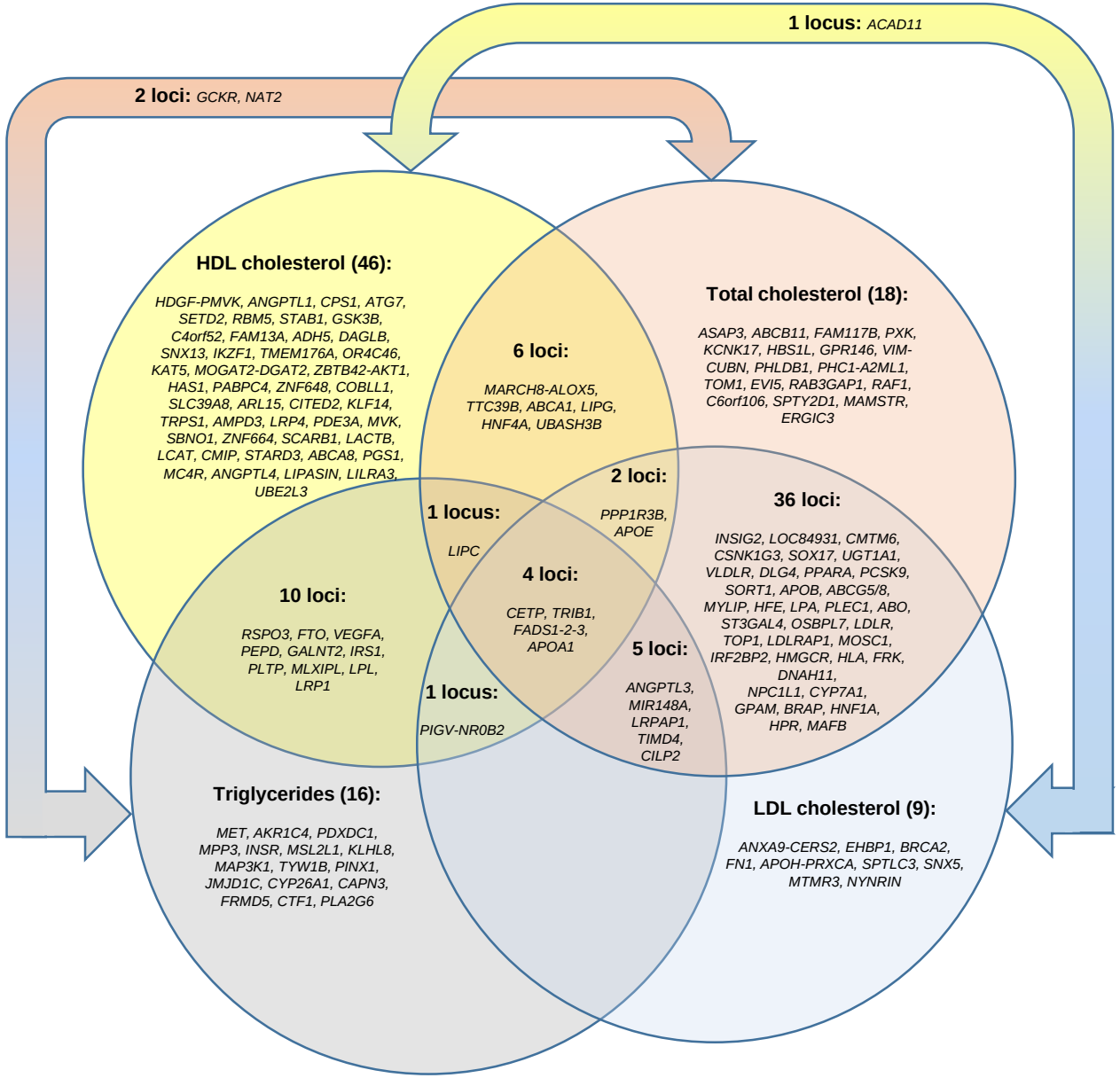
Sekar Kathiresan

Cristen Willer

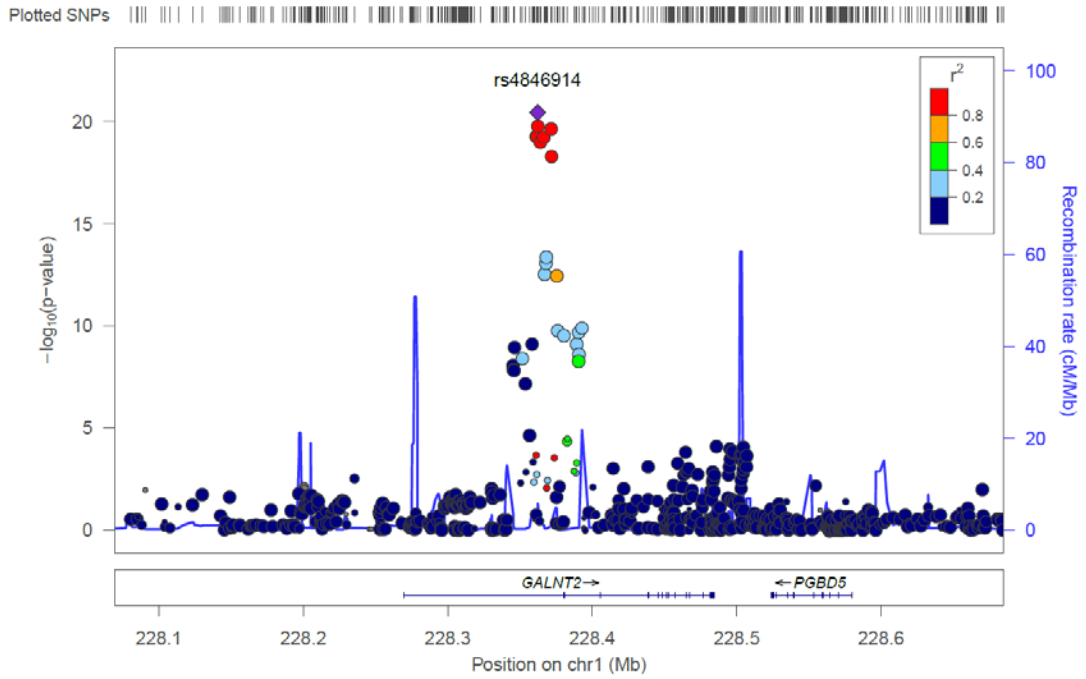
- An example of the current standard for genetic association studies
- Most recent analysis includes 188,578 individuals and identifies 157 loci associated with blood lipid levels
- Associated loci can:
  - Suggest new targets for therapy
  - Confirm suspected targets or known biology
  - Provide insights on the relationship between lipids and other phenotypes

Willer et al, Nat Genet, 2008; Teslovich et al, Nature, 2010; Willer et al, Nat Genet, 2013; Do et al, Nat Genet, 2013

A SNAPSHOT OF LIPID GENETICS



Suggesting New Targets: GALNT2

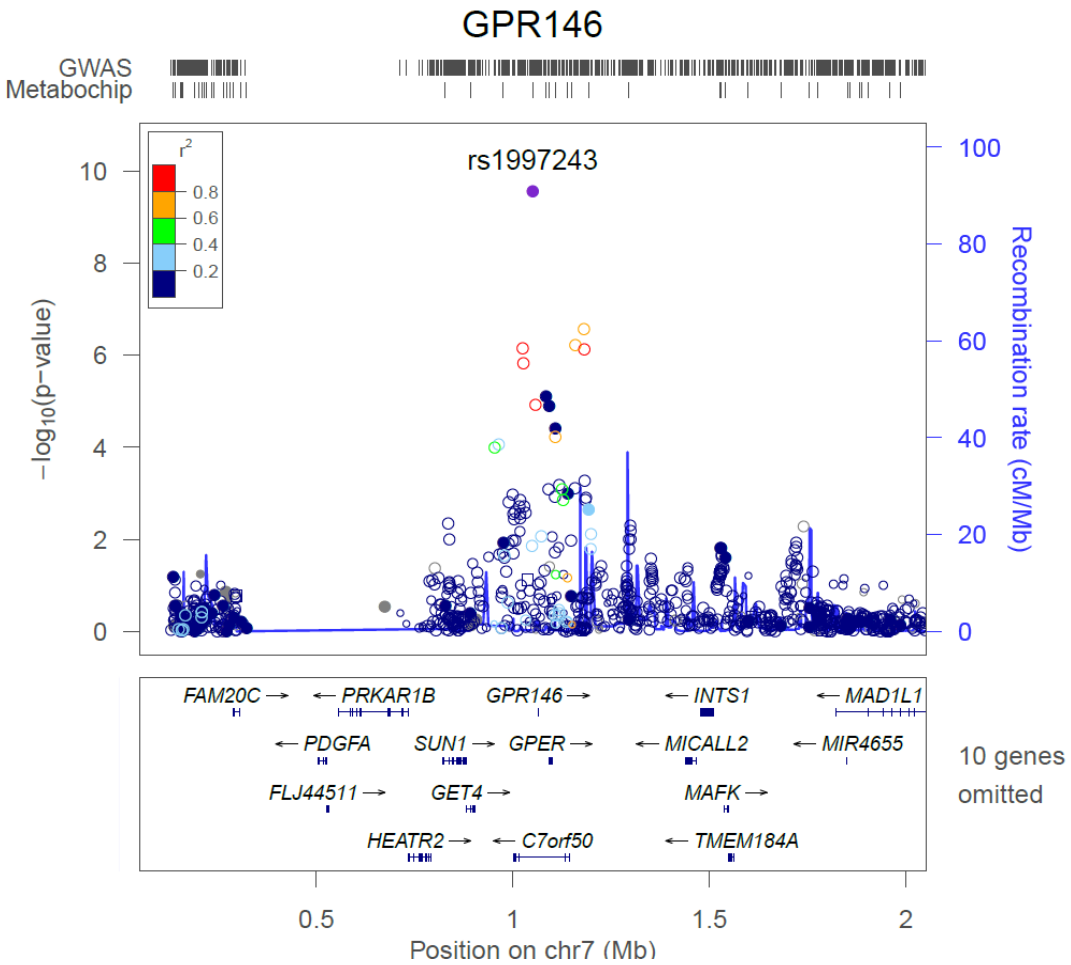


- GWAS allele with 40% frequency associated with ±1 mg/dl in HDL-C
- Explored consequences of modifying GALNT2 expression in mouse liver…
- Overexpression of *GALNT2* or *Galnt2* decreases HDL-C ~20%
- Knockdown of *Galnt2* increases HDL-C by ~30%

Dan Rader

Teslovich et al, Nature, 2012

Supporting Previous Leads: GPR146



- Our work shows that variants near GPR146 are associated with total cholesterol
- U. S. Patent Application #20,090,036,394 discloses that, in mice, targeting GPR146 lowers cholesterol
- Together, the two pieces of evidence could encourage human trials

Insights about biology …

- In our first lipid GWAS, we showed that every allele that increased LDL-C was also associated with increased coronary heart disease risk…
- Later, we showed that alleles with the largest impact on HDL-C in blood, also modify the risk of age related macular degeneration
- Our most recent analysis show that the impact of an allele on triglyceride levels predicts heart disease risk
  - Even after controlling for its association with HDL-C and LDL-C
  - Analysis continues to support causal role for LDL-C (but not for HDL-C)

Challenges and Opportunities

- Discovery of variants with clear functional consequence is now easier.
  - Almost every gene will be severely defective in at least a few individuals.

- Studies of these variants present new challenges and opportunities.
  - With loss-of-function variants, much easier path from association to biology.
  - Most loss-of-function variants are individually very rare.

- To increase power, it is useful to consider cost-effective strategies:
  - Analytical strategies that economize sequencing effort.
  - Opportunities to aggregate information across studies and variants.

- Need more collaboration about clinical experts, biologists and human genetics.
  - Ensure that we focus on the most important outcomes.
  - Ensure that we translate findings into biological insights.

Questions that Might Be Answered With Complete Sequence Data…

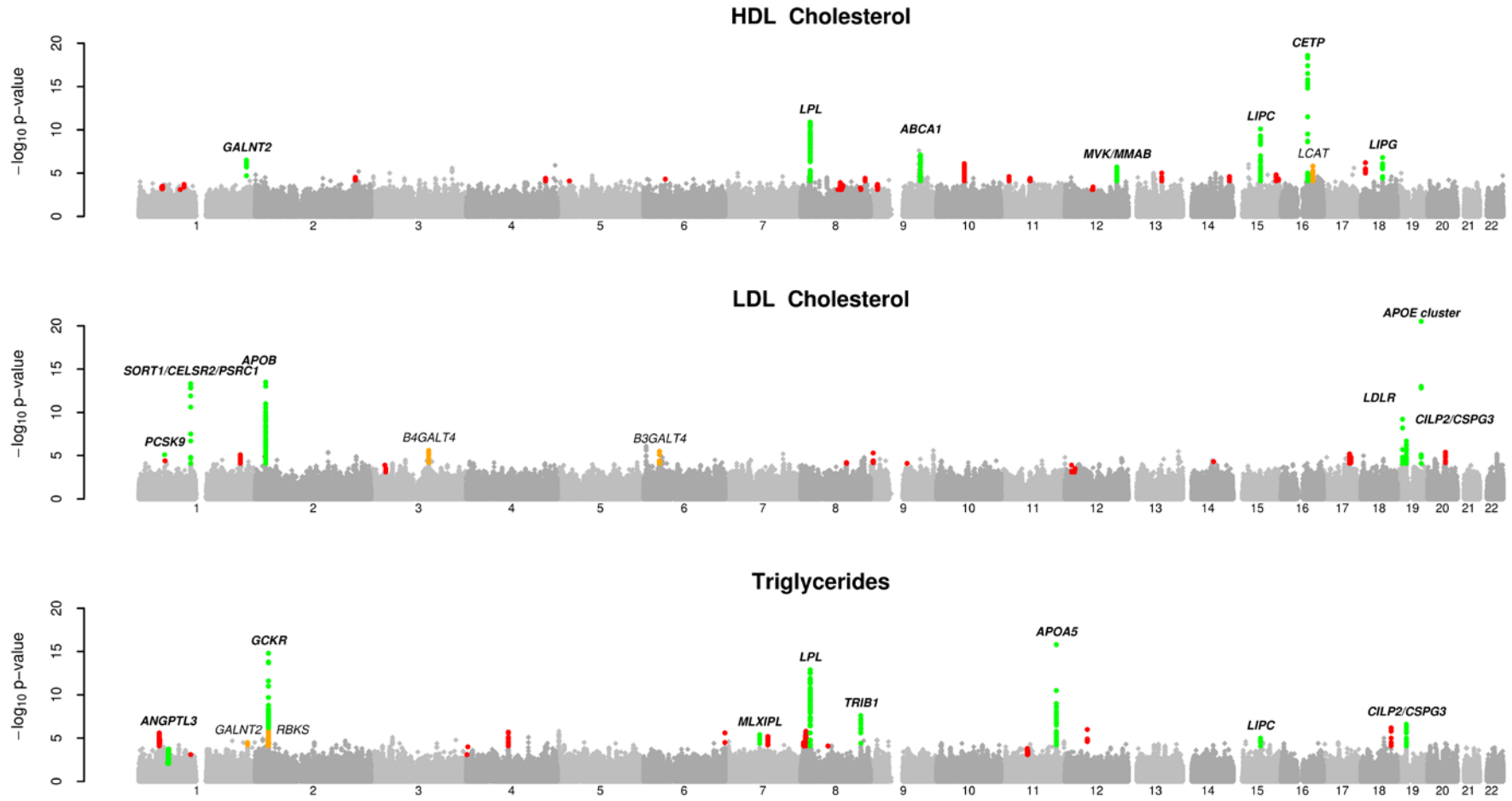
- What is the contribution of each identified locus to a trait?
  - Likely that multiple variants, common and rare, will contribute
- What is the mechanism? What happens when we knockout a gene?
  - Most often, the causal variant will not have been examined directly
  - Rare coding variants will provide 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 Is the Total Contribution of Each Locus?

Evidence that

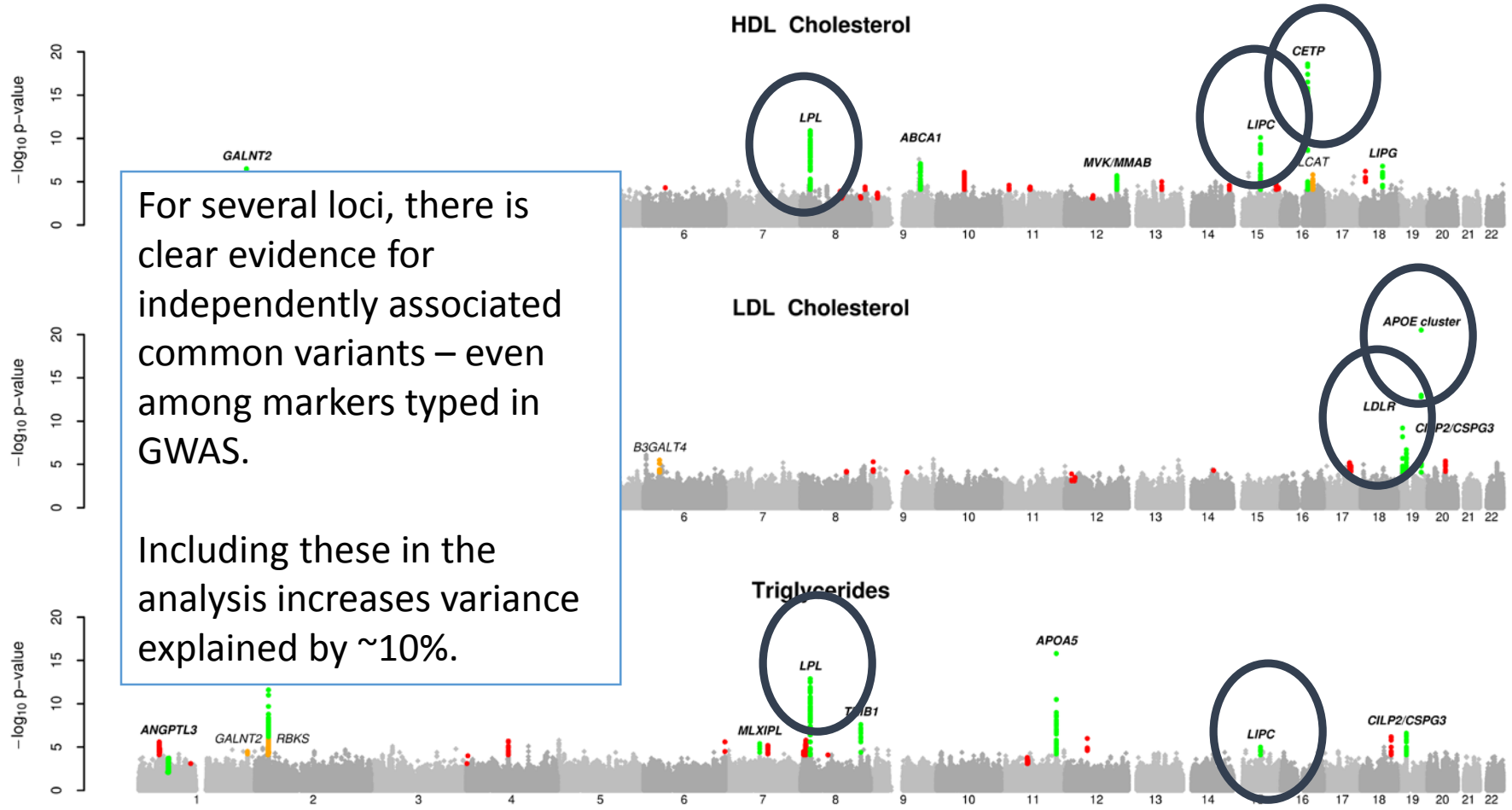
Multiple Variants Will be Important

Evidence for Multiple Variants Per Locus Example from Lipid Biology



Willer et al, *Nat Genet*, 2008
Kathiresan et al, *Nat Genet,* 2008, 2009

Evidence for Multiple Variants Per Locus Example from Lipid Biology



For several loci, there is clear evidence for independently associated common variants – even among markers typed in GWAS.

Including these in the analysis increases variance explained by ~10%.

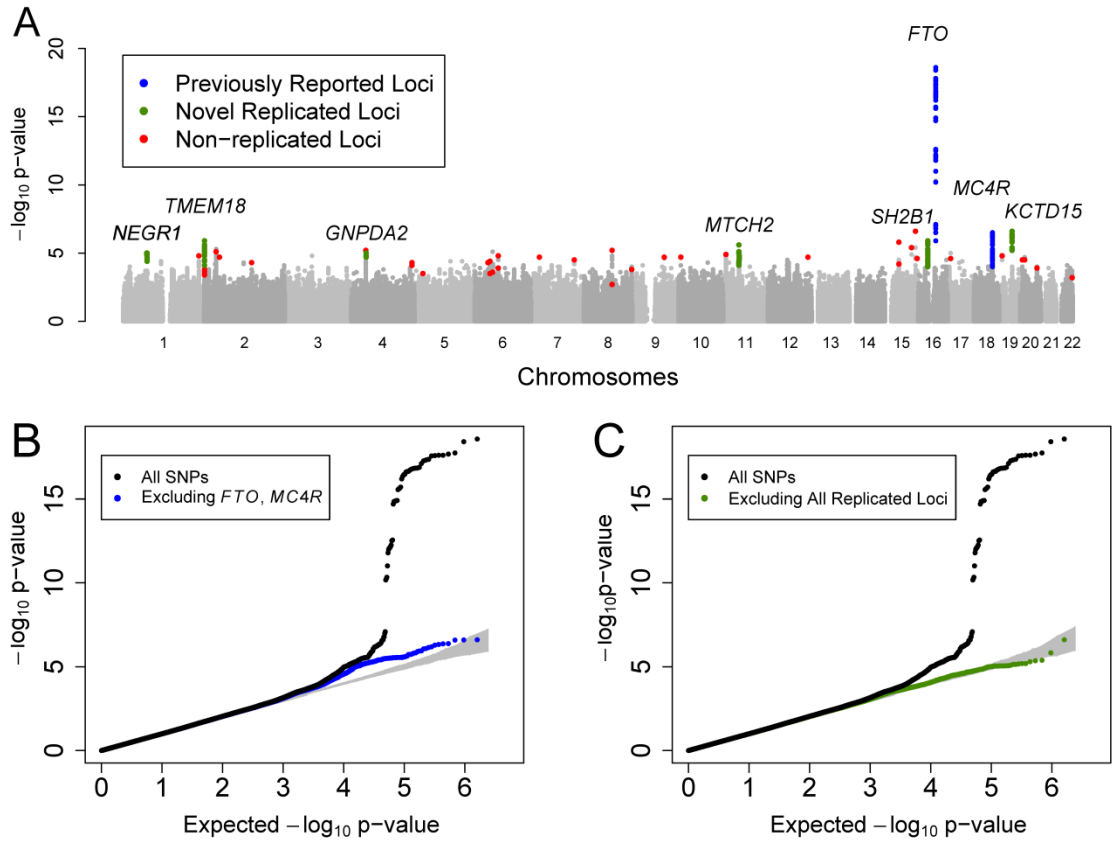
Willer et al, *Nat Genet*, 2008
Kathiresan et al, *Nat Genet,* 2008, 2009

What is The Contribution of Structural Variants?

Current Arrays Interrogate 1,000,000s of SNPs,

but 100s of Structural Variants

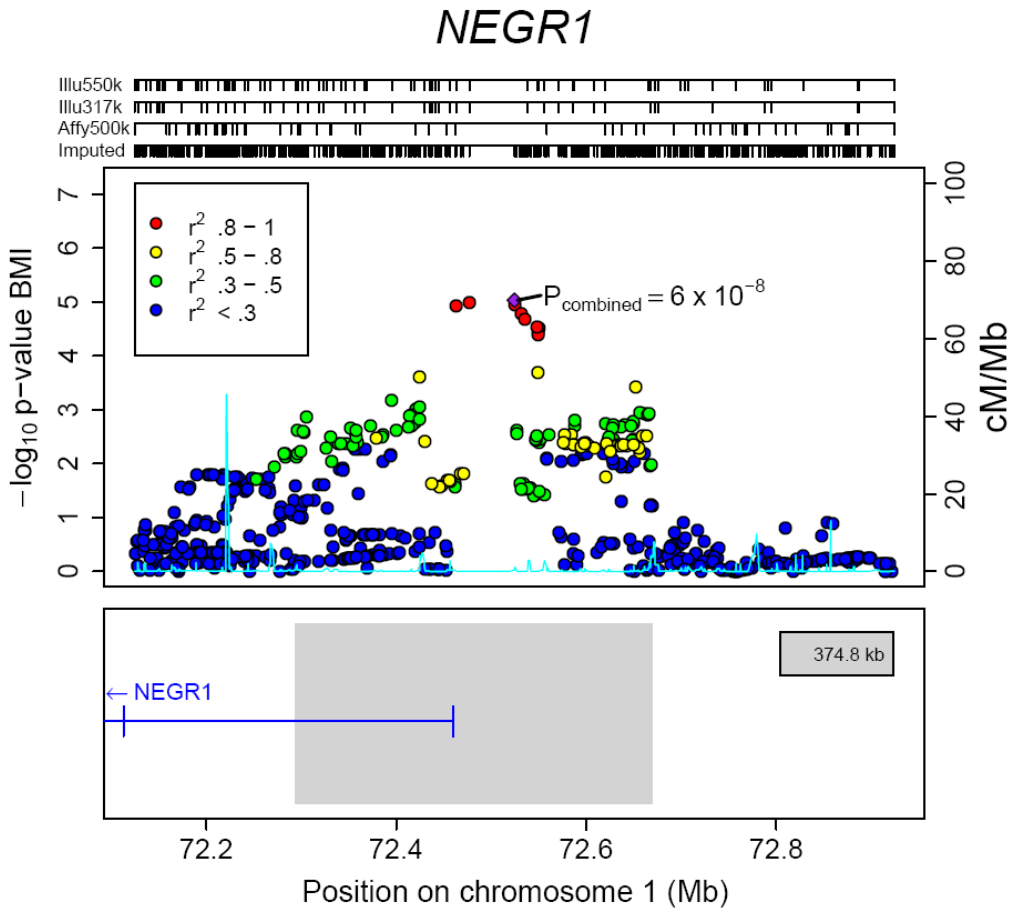
Evidence that Copy Number Variants Important Example from Genetics of Obesity



Seven of eight confirmed BMI loci show strongest expression in the brain…

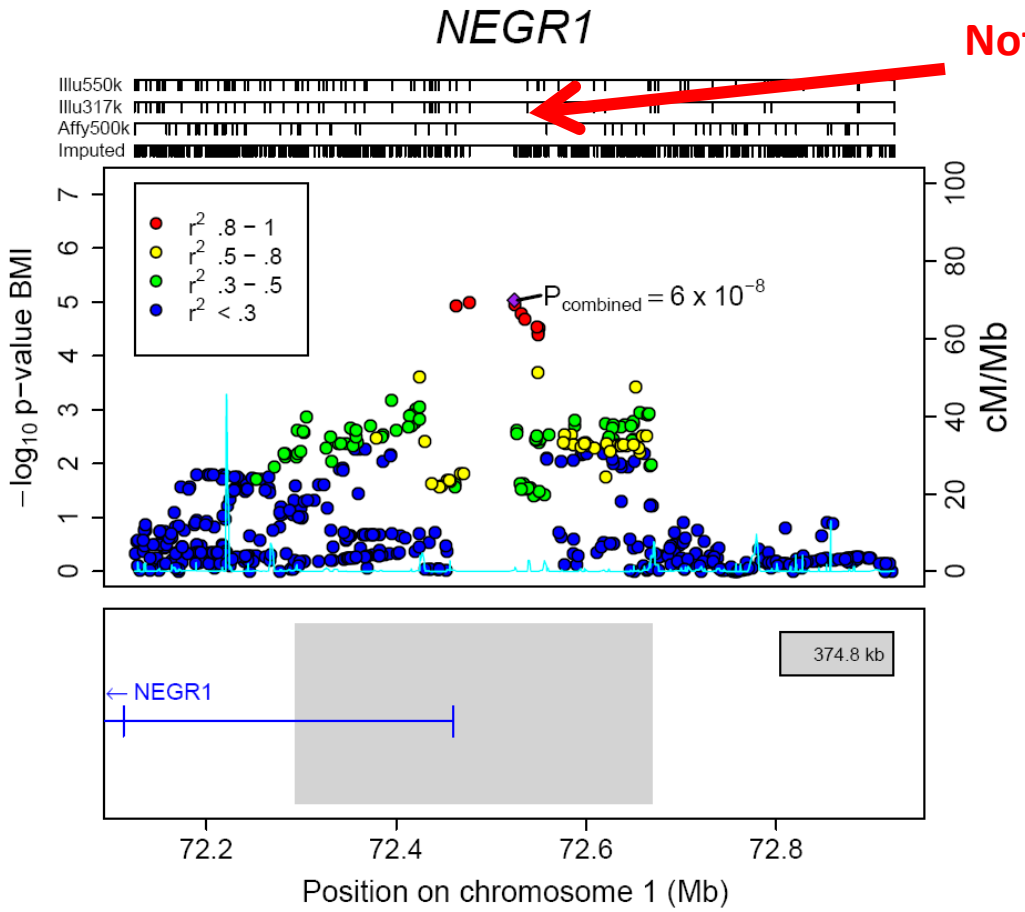
Willer et al, *Nature Genetics,* 2009

Evidence that Copy Number Variants Important Example from Genetics of Obesity



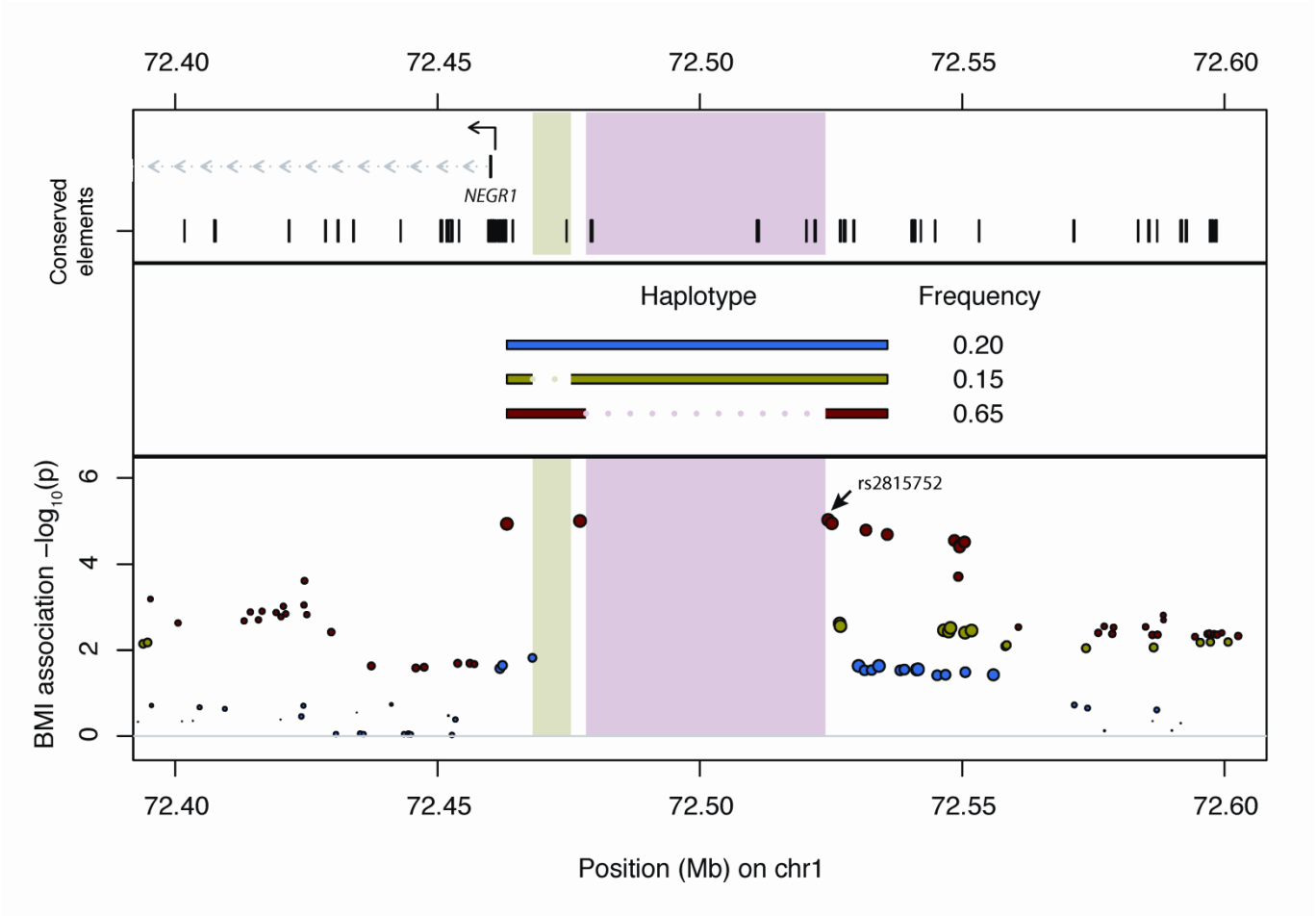
Willer et al, *Nature Genetics,* 2009

Evidence that Copy Number Variants Important Example from Genetics of Obesity



Willer et al, *Nature Genetics,* 2009

Associated Haplotype Carries Deletion



Willer et al, *Nature Genetics,* 2009

What is the Mechanism? What Can We Learn From Rare Knockouts?

Early Example from Type 1 Diabetes

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

Next Generation Sequencing

Massive Throughput Sequencing

- Tools to generate sequence data evolving rapidly
- Commercial platforms produce gigabases of sequence rapidly and inexpensively
  - ABI SOLiD, Illumina Solexa, Roche 454, Complete Genomics, Ion Torrent, and others…
- Sequence data consist of thousands or millions of short sequence reads with moderate accuracy
  - 0.5 – 1.0% error rates per base may be typical

Shotgun Sequence Reads

- Typical short read might be <25-100 bp long and not very informative on its own

- Reads must be arranged (*aligned*) relative to each other to reconstruct longer sequences

Base Qualities

Short Read Sequence

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Short Read Base Qualities

30.30.28.28.29.27.30.29.28.25.24.26.27.24.24.23.20.21.22.10.25.25.20.20.18.17.16.15.14.14.13.12.10

- Each base is typically associated with a quality value
- Measured on a “Phred” scale, which was introduced by Phil Green for his Phred sequence analysis tool

$$BQ = -\log_{10}(\epsilon)\,, where\ \epsilon\ is\ the\ probability\ of\ an\ error$$

Read Alignment

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Short Read (30-100 bp)

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome (3,000,000,000 bp)

- The first step in analysis of human short read data is to align each read to genome, typically using a hash table based indexing procedure
- This process now takes no more than a few hours per million reads …
- Analyzing these data without a reference human genome would require much longer reads or result in very fragmented assemblies

Read Alignment – Food for Thought

- Typically, all the words present in the genome are indexed to facilitate read mapping …
  - What are the benefits of using short words?
  - What are the benefits of using long words?

- How matches do you expect, on average, for a 10-base word?
  - Do you expect large deviations from this average?

Mapping Quality

- Measures the confidence in an alignment, which depends on:
  - Size and repeat structure of the genome
  - Sequence content and quality of the read
  - Number of alternate alignments with few mismatches

- The mapping quality is usually also measured on a “Phred” scale

- Idea introduced by Li, Ruan and Durbin (2008) *Genome Research* **18:**1851-1858

Refinements to Mapping Quality

- In their simplest form, mapping qualities apply to the entire read
- However, in gapped alignments, uncertainty in alignment can differ for different portions of the read
  - For example, it has been noted that many wrong variant calls are supported by bases near the edges of a read
- Per base alignment qualities were introduced to summarize local uncertainty in the alignment

Per Base Alignment Qualities

Short Read

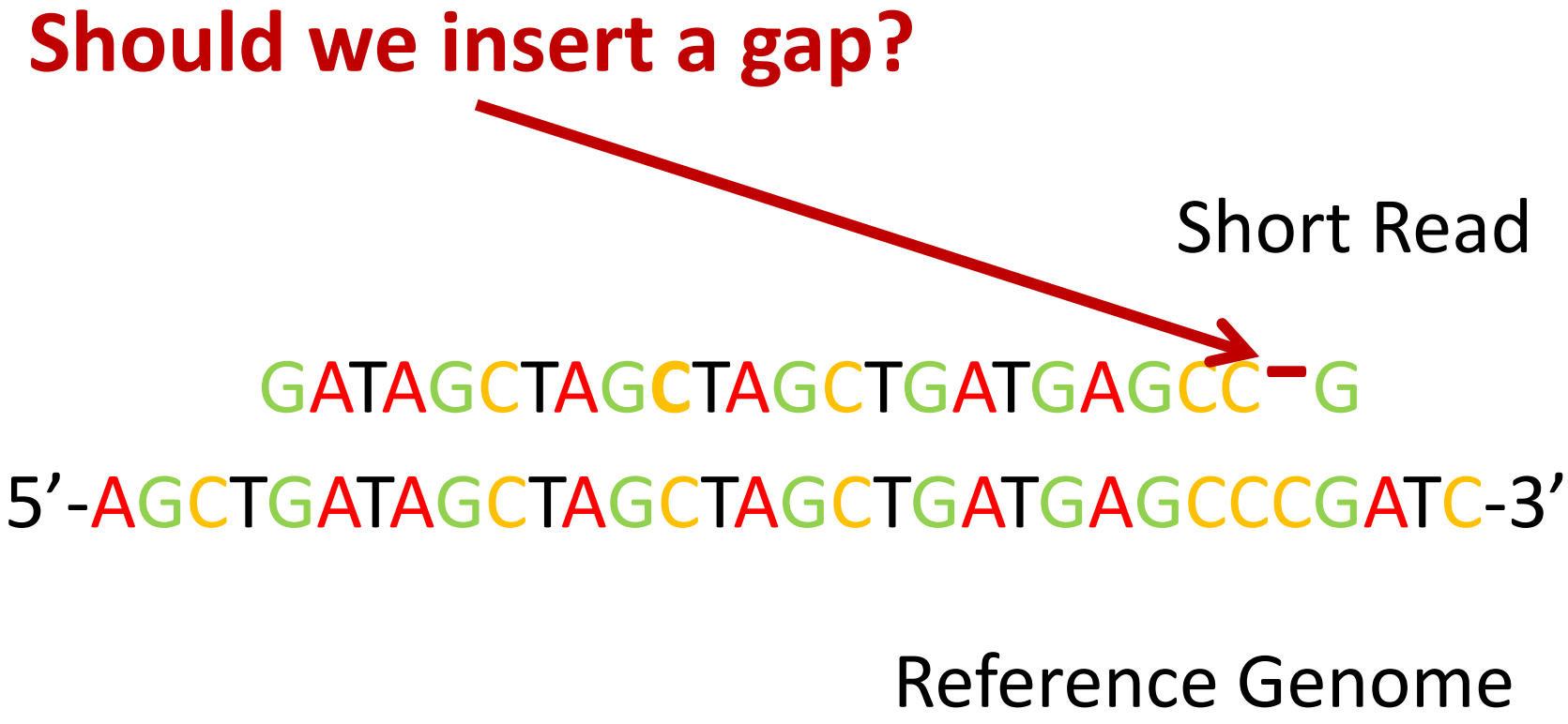
GATAGCTAGCTAGCTGATGA GCCG

5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

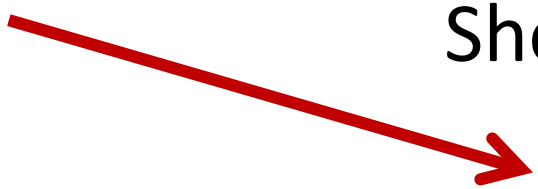
Heng Li

Per Base Alignment Qualities



Per Base Alignment Qualities

Compensate for Alignment Uncertainty
With Lower Base Quality



Short Read

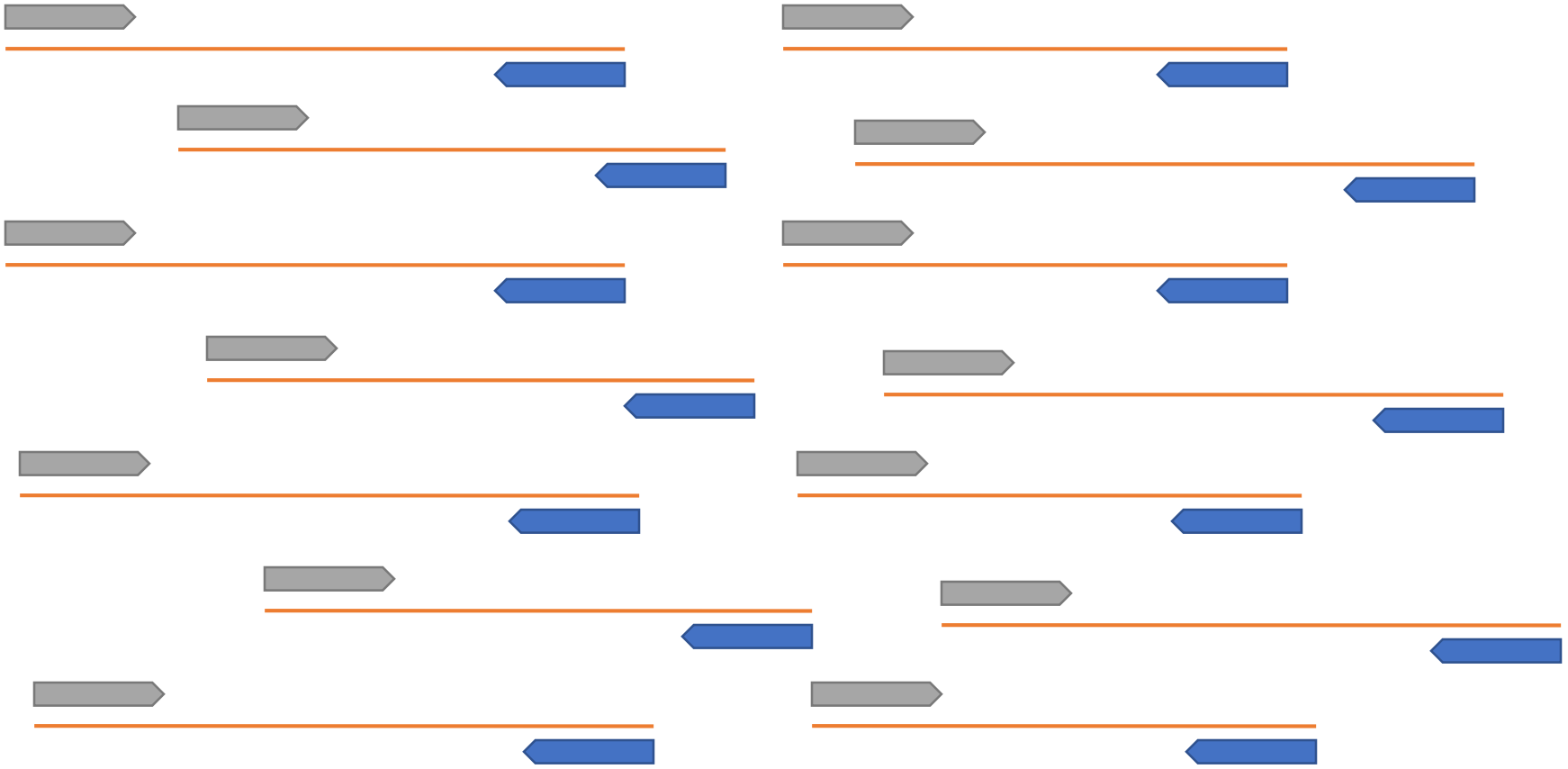
GATAGCTAGCTAGCTGATGAGCCG

5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Heng Li

Paired End Sequencing



Population of DNA fragments of known size (mean + stdev)
Paired end sequences

Paired End Sequencing

Paired Reads

Initial alignment to the reference genome

Paired end resolution

Detecting Structural Variation

- Read depth
  - Regions where depth is different from expected
    - Expectation defined by comparing to rest of genome …
    - … or, even better, by comparing to other individuals
- Split reads
  - If reads are longer, it may be possible to find reads that span the structural variation
- Discrepant pairs
  - If we find pairs of reads that appear to map significantly closer or further apart than expected, could indicate an insertion or deletion
  - For this approach, “physical coverage” which is the sum of read length and insert size is key
- De Novo Assembly

How Much Variation is There?

- An average genome includes:
  - 3.6M SNPs
  - 350K indels
  - 700 large deletions

- Numbers are probably underestimates …
- … some variants are hard to call with short reads

- 1000 Genomes Project (2012) *Nature* **491**:56-65

How Much Variation is There?

SNPs Per Individual in Gene Regions

Primarily European Ancestry

European Ancestry	# SNP	# HET	# ALT	# Singletons	Ts/Tv
SILENT	10127	6174	3953	38.2	5.10
MISSENSE	8541	5184	3357	72.2	2.16
NONSENSE	86	57	29	2.1	1.70

Primarily African Ancestry

African Ancestry	# SNP	# HET	# ALT	# Singletons	Ts/Tv
SILENT	12028	8038	3990	53.2	5.19
MISSENSE	9870	6502	3367	94.2	2.16
NONSENSE	92	57	35	2.4	1.57

NHLBI Exome Sequencing Project

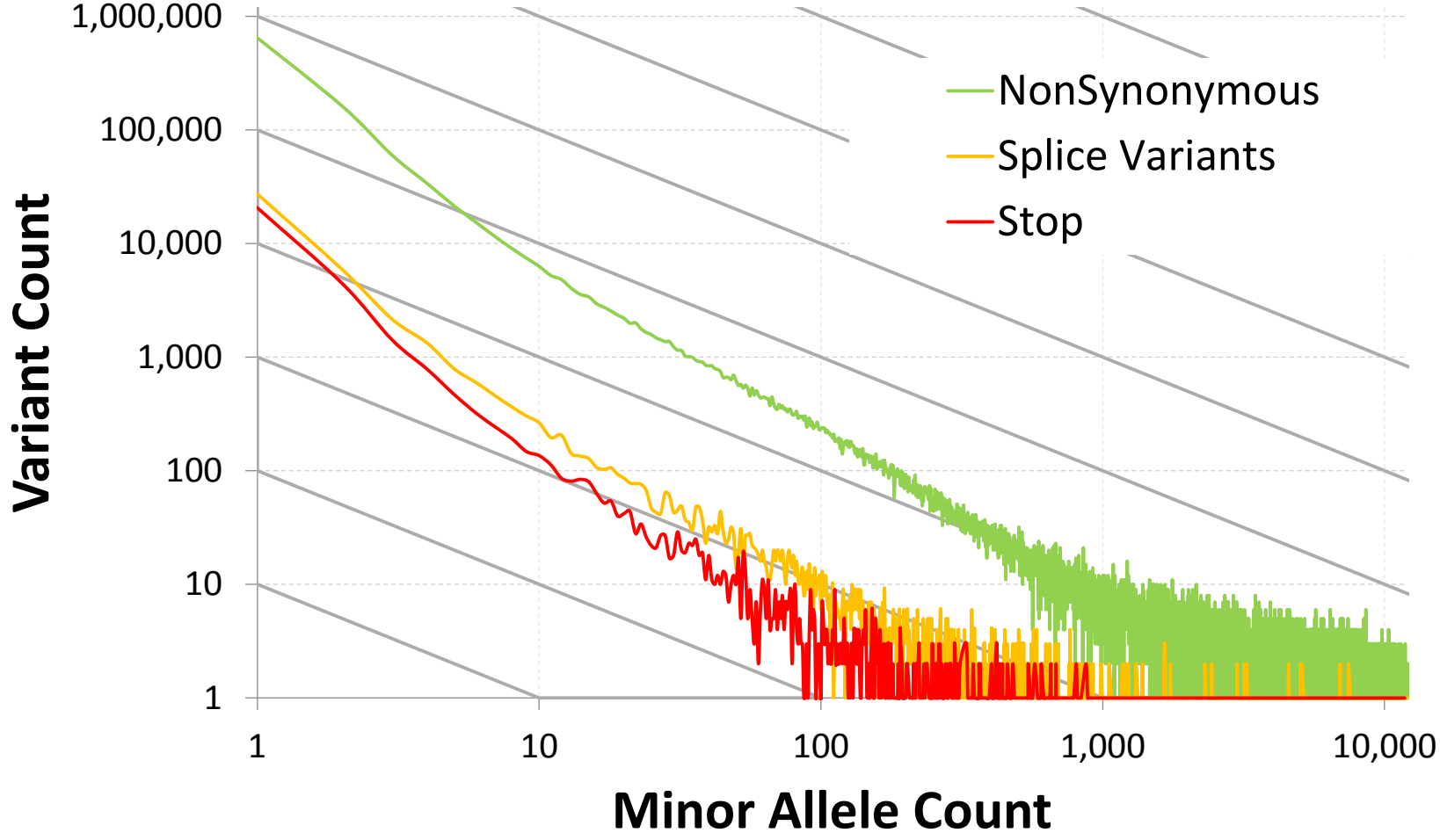
Lots of Rare Functional Variants to Discover

SET	# SNPs	Singletons	Doubletons	Tripletons	>3 Occurrences
Synonymous	270,263	128,319 (47%)	29,340 (11%)	13,129 (5%)	99,475 (37%)
Nonsynonymous	410,956	234,633 (57%)	46,740 (11%)	19,274 (5%)	110,309 (27%)
Nonsense	8,913	6,196 (70%)	926 (10%)	326 (4%)	1,465 (16%)
Non-Syn / Syn Ratio		1.8 to 1	1.6 to 1	1.4 to 1	1.1 to 1

There is a very large reservoir of extremely rare, likely functional, coding variants.
(Results above correspond to approximately 5,000 individuals)

NHLBI Exome Sequencing Project

Allele Frequency Spectrum (After Sequencing 12,000+ Individuals)

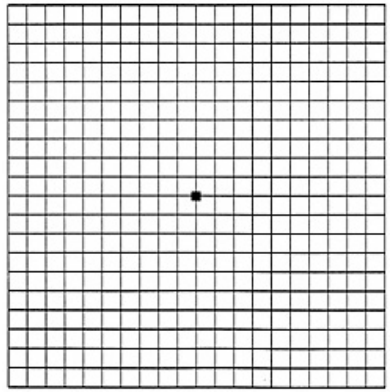


http://genome.sph.umich.edu/wiki/Exome_Chip_Design

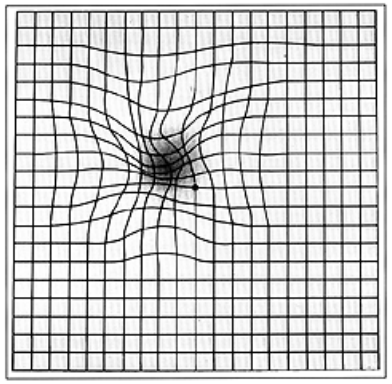
Estimates of Genetic Ancestry from Tiny Bits of Sequence Data

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

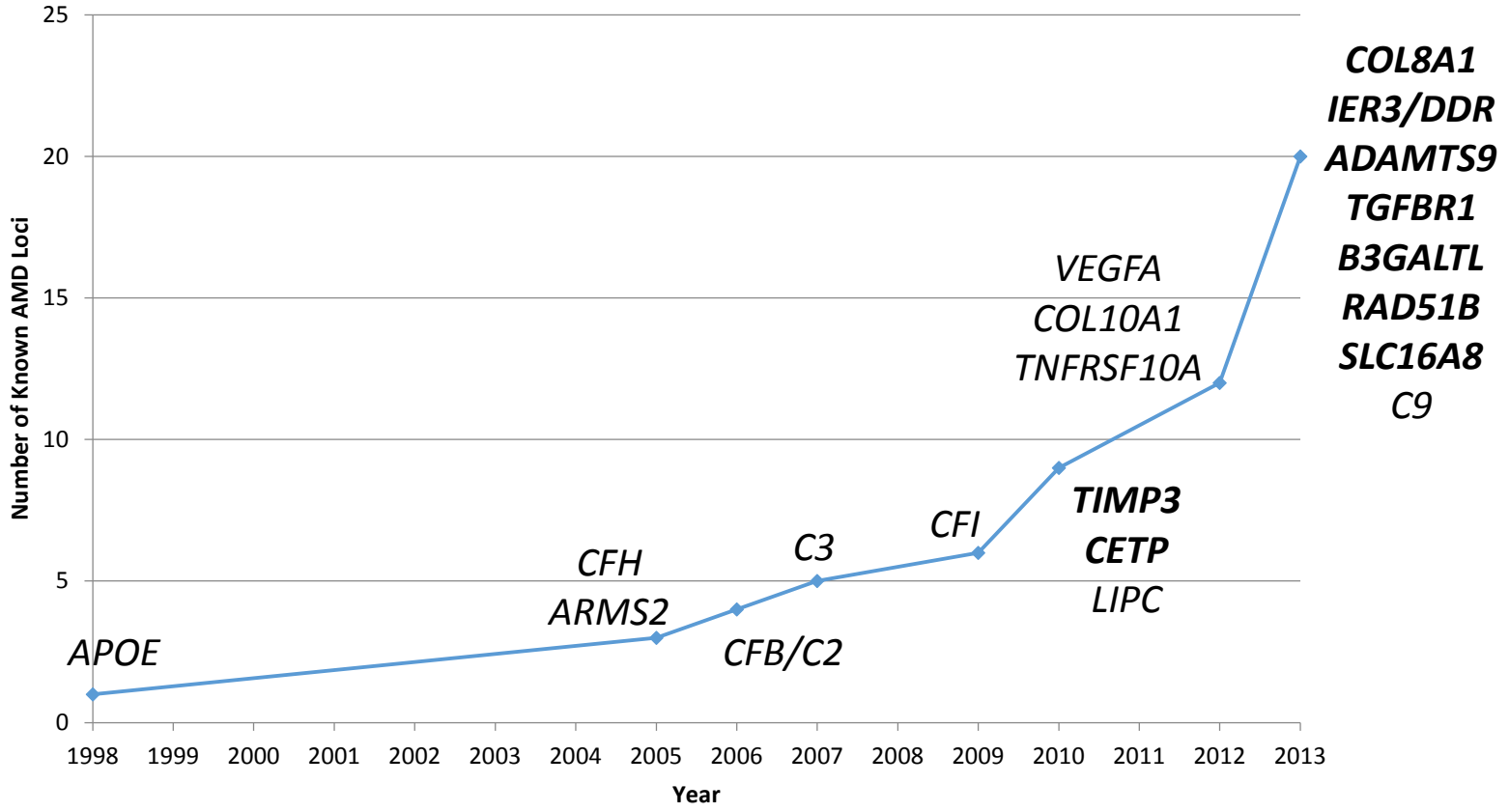


Normal Vision



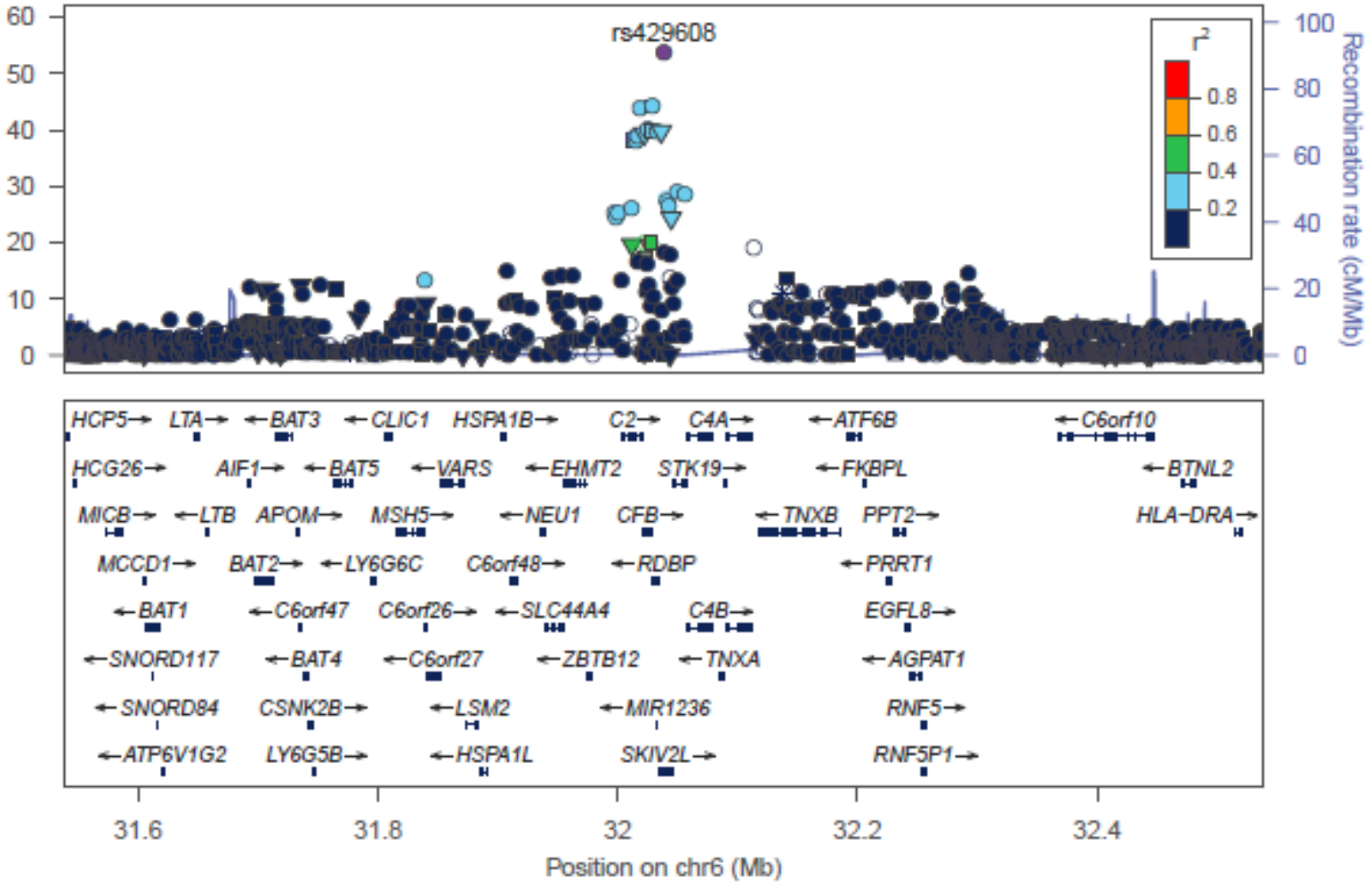
Macular Degeneration

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

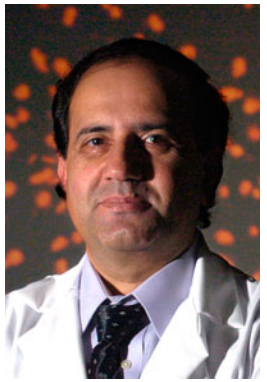


Evidence for Association is… “Circumstantial!”

- In any one region, many alleles will typically be associated
  - These all appear in the ancestral fragment carrying the “causal” variant(s)
- In any one region, many causal genes and mechanisms can often be postulated
  - Need a strategy to systematically adjudicate between options
- Identifying causal mechanisms requires…
  - Exhaustively examining all variants
  - Additional experiments
  - Guesswork
- Identification of causal mechanisms can be helped by …
  - Studies in different populations, with different haplotype structure
  - Identification of independently associated variants

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

- Examine rare variants in known loci to obtain clues about functional mechanisms
  - 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!)

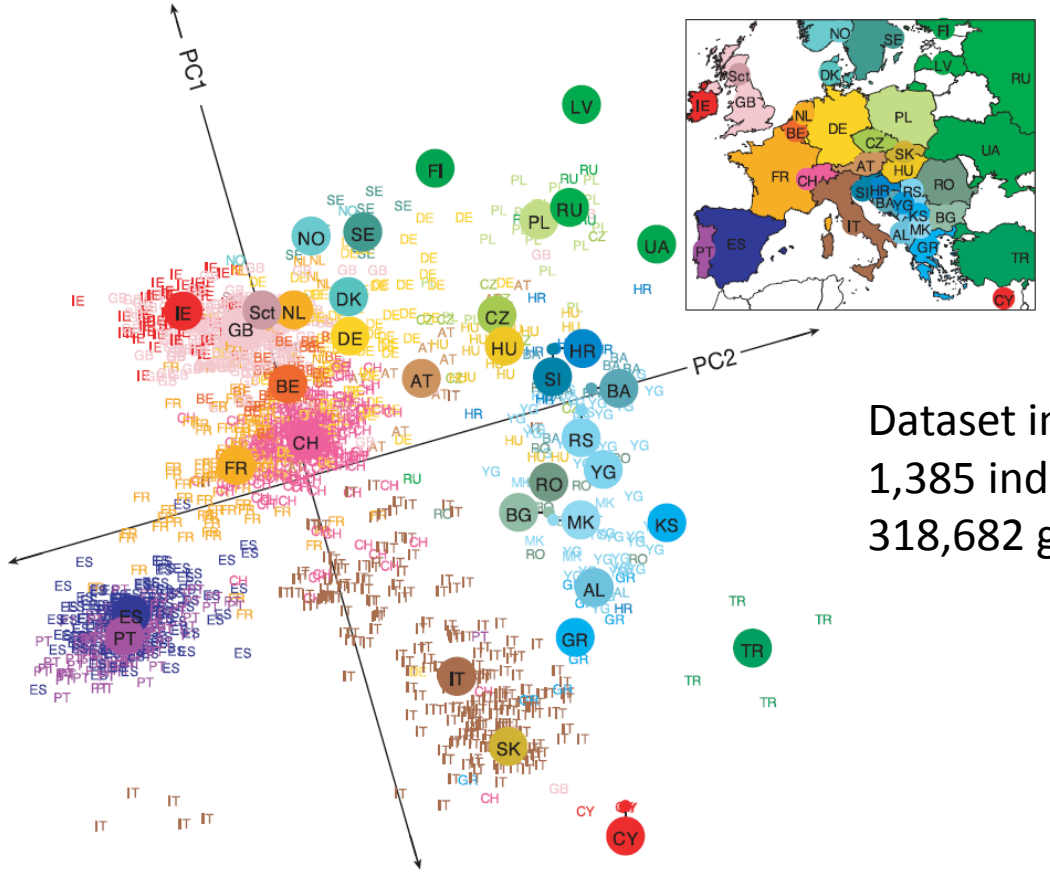
- Studying rare variants, requires very large sample sizes!

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 …

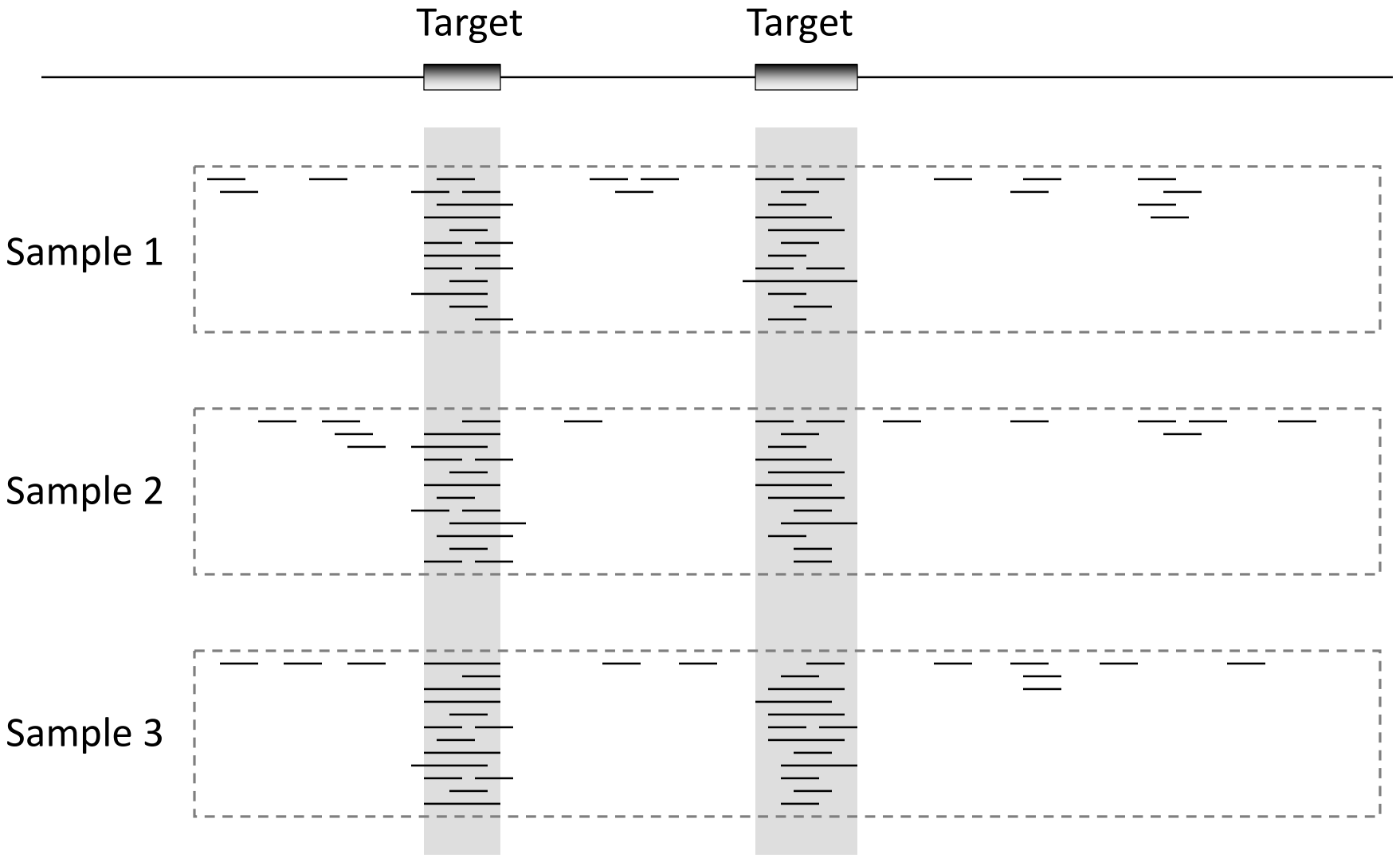
Principal Component Ancestry Map of Europe



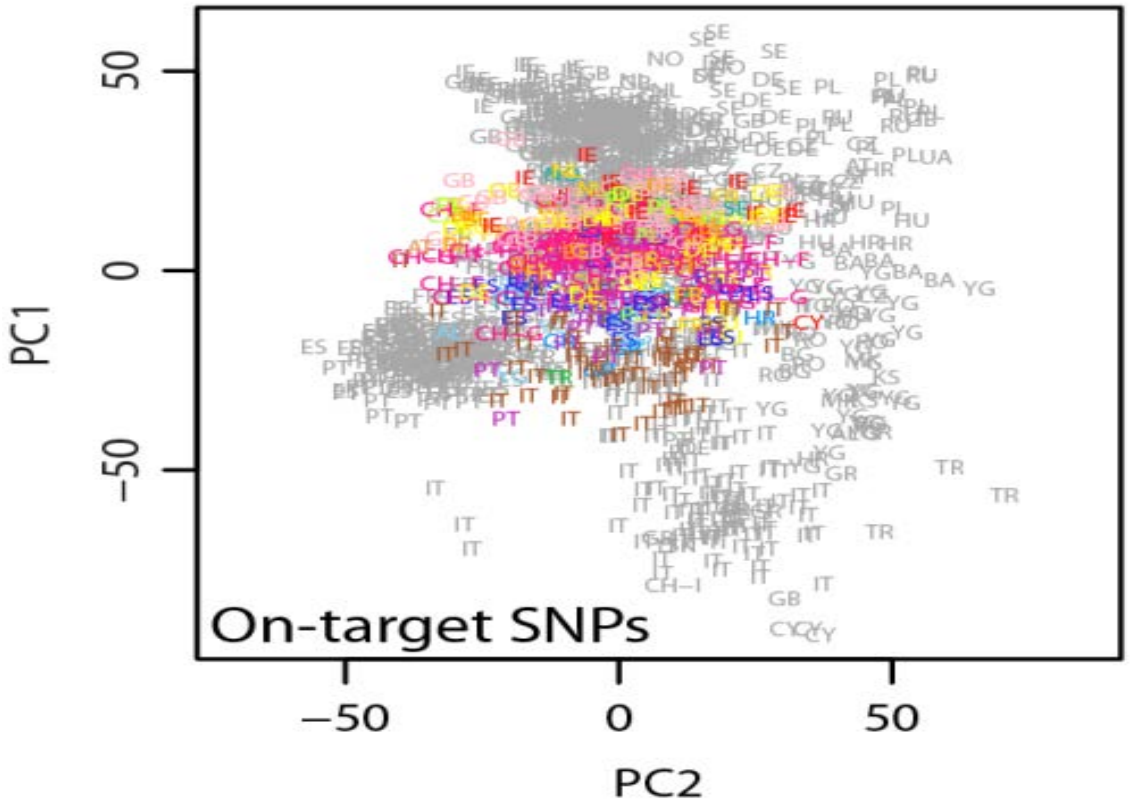
Dataset includes:
1,385 individuals of known ancestry
318,682 genetic markers passing filters

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

Targeted sequencing data



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



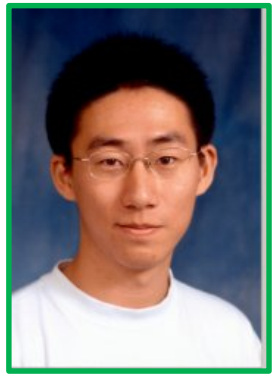
On-target genotypes don’t contain enough information to estimate the ancestry of a sample. The illustration is based on >80x deep whole exome data.

The Problem

- 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!

- Very little overlap between any pair of individuals…
  - Need to build a reference coordinate space



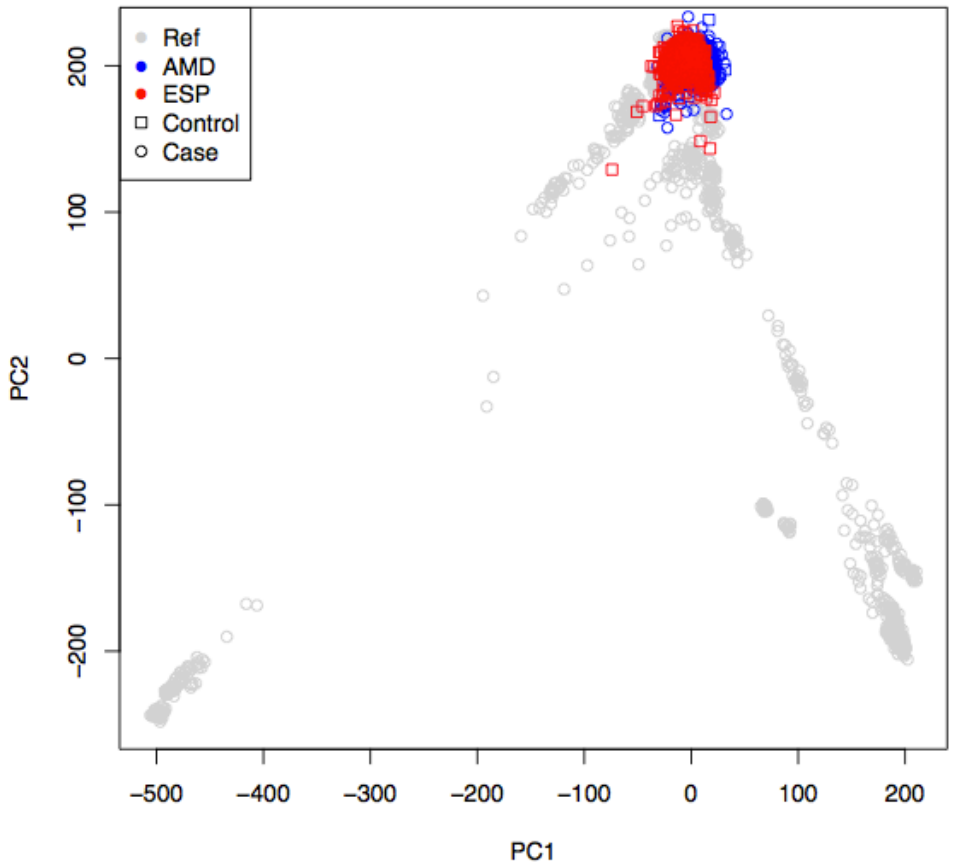
Xiaowei Zhan

Chaolong Wang

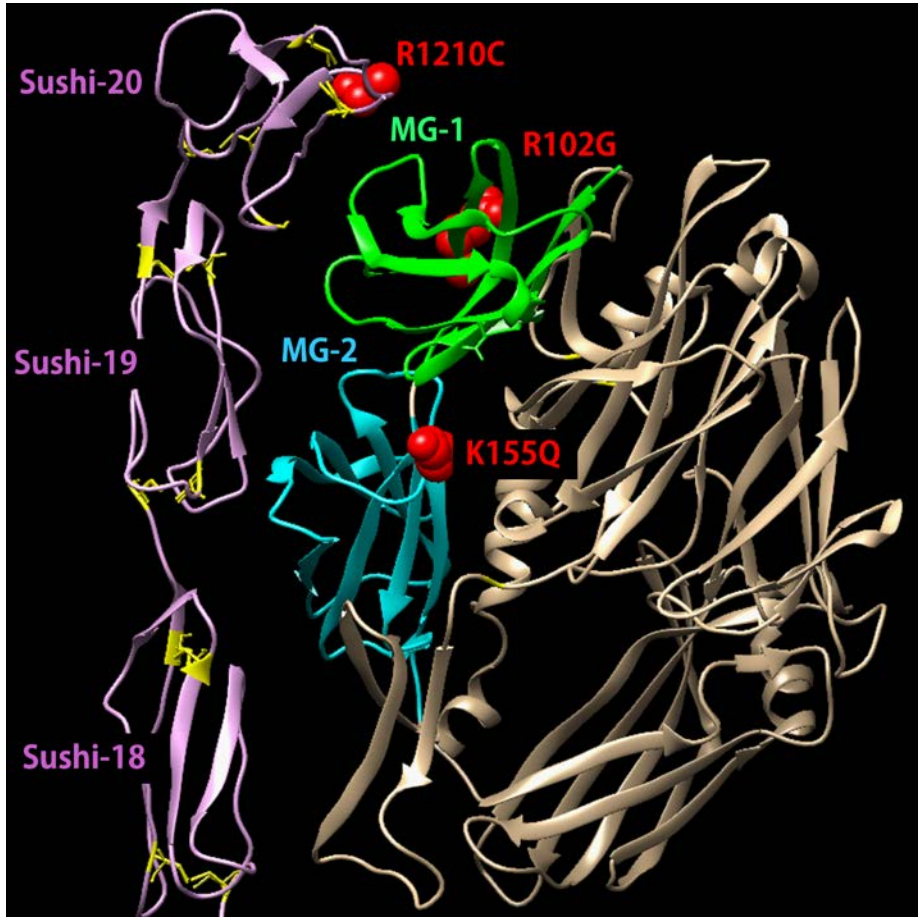
Sebastian Zöllner

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 signal at K155Q in C3 gene looks promising, reaching 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

Design A Whole Genome Sequencing Study in Sardinia

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:
  - Sequence >1,000 individuals at 2x to obtain draft sequences
  - Genotype all individuals, impute sequences into relatives

How Is Sequencing Progressing?

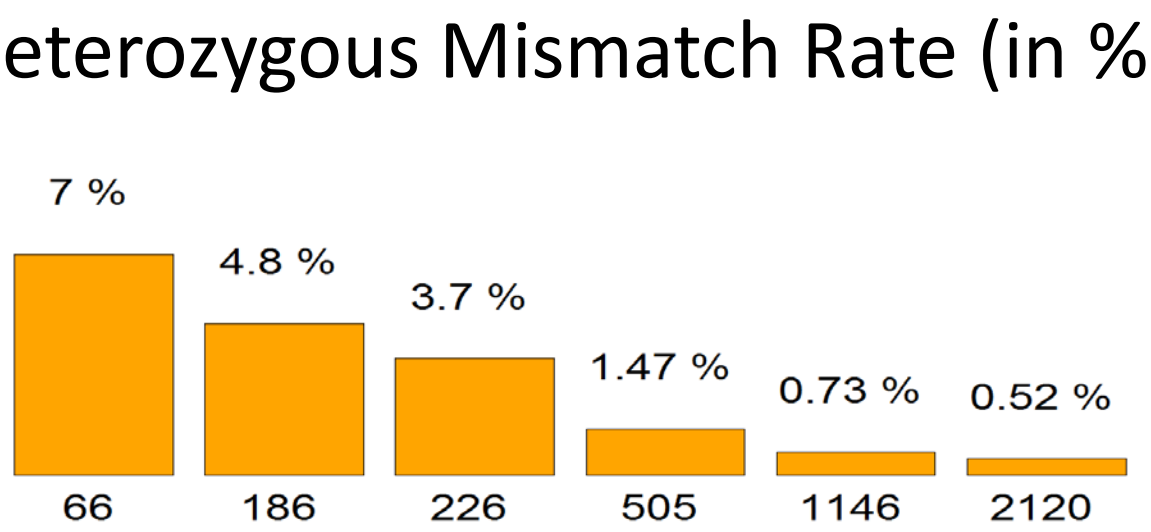
- NHGRI estimates of sequencing capacity and cost …
  - Since 2006, for fixed cost …
  - … ~4x increase in sequencing output per year

- 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: ~45 Gb/lane

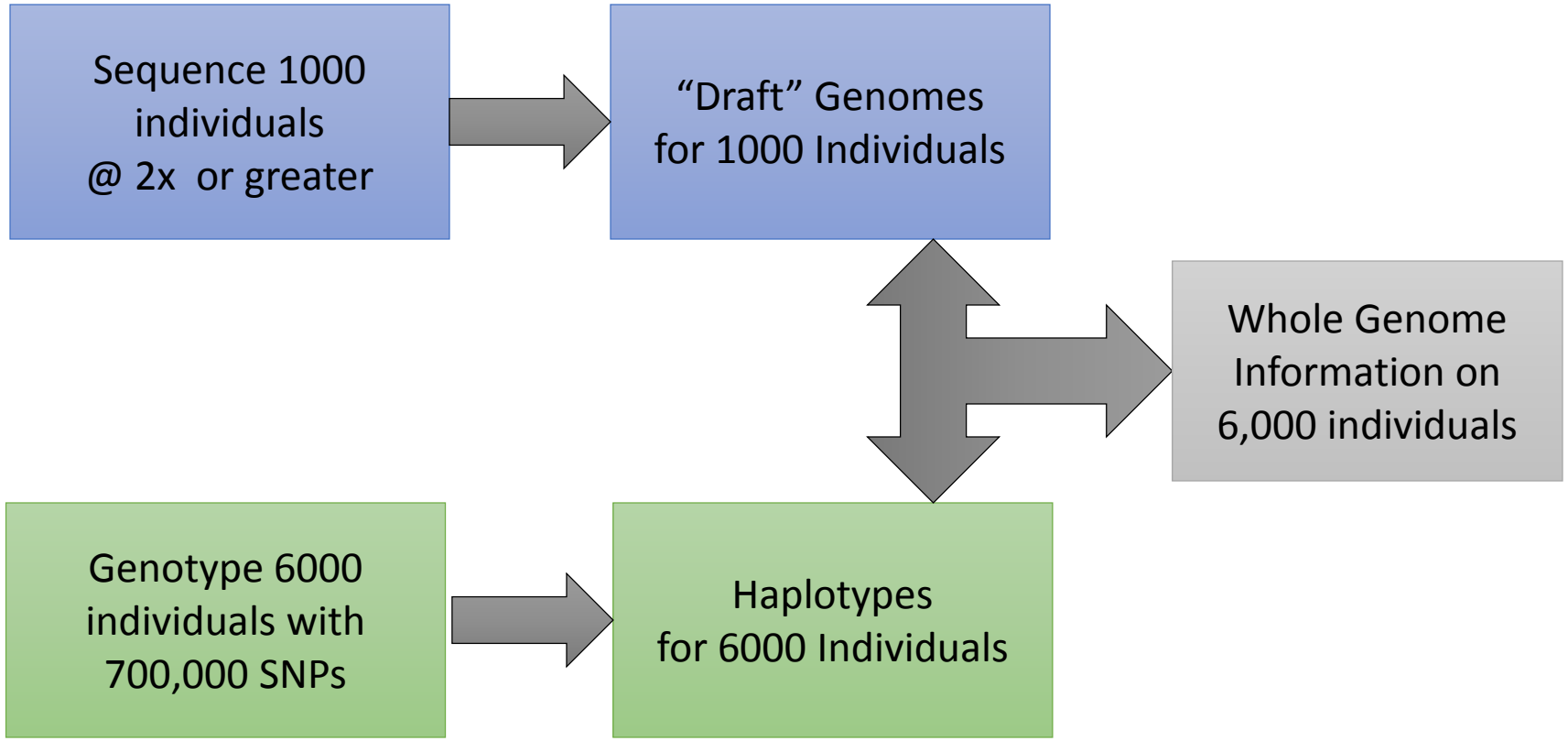
- Other small improvements
  - No PCR libraries increase genome coverage, reduce duplicate rates

Fabio Busonero, Andrea Maschio

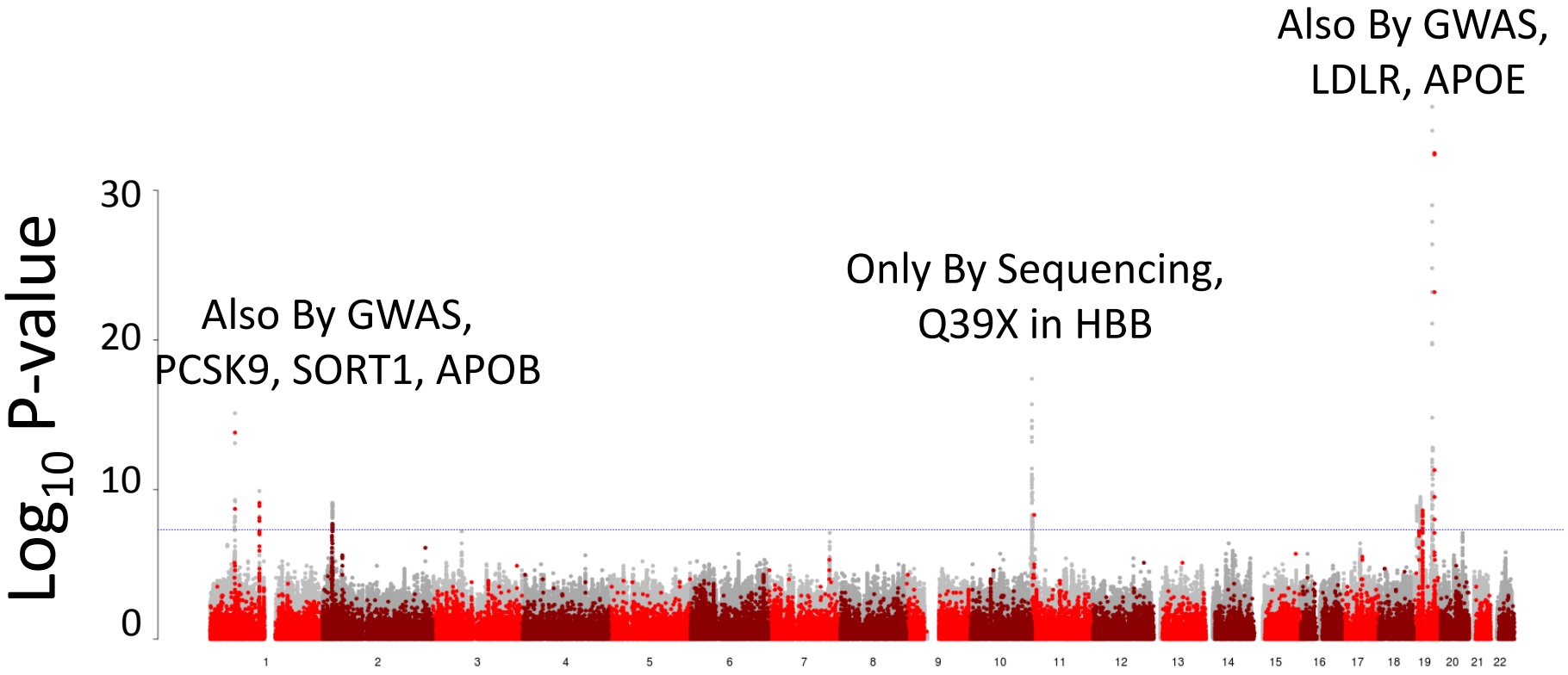
As more samples are sequenced, Accuracy increases



Design


Sequence 1000 individuals @ 2x or greater
"Draft" Genomes for 1000 Individuals
Whole Genome Information on 6,000 individuals
Genotype 6000 individuals with 700,000 SNPs
Haplotypes for 6000 Individuals

What Do We See Genomewide?
LDL Cholesterol



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

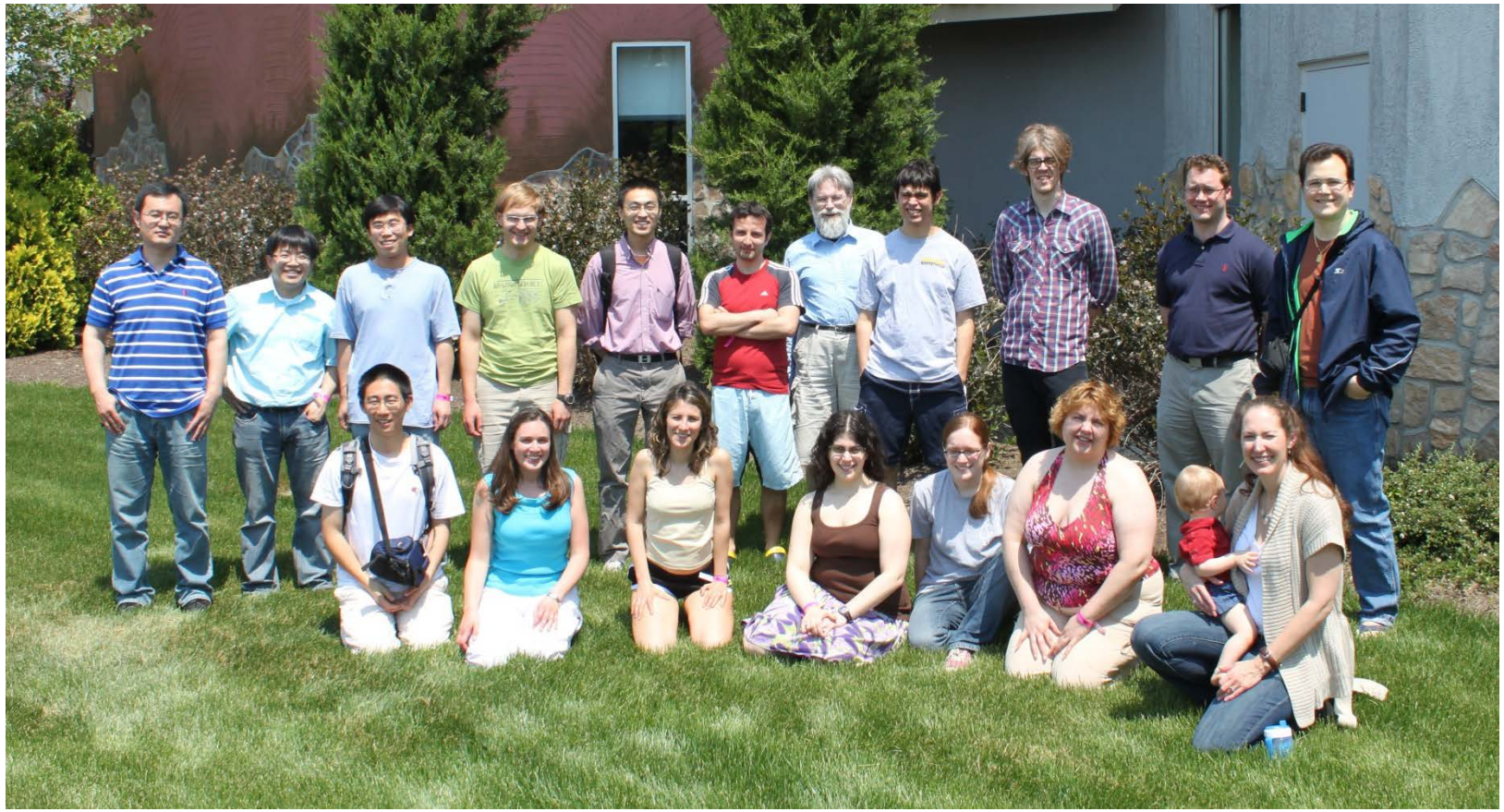
Locus	Variants	MAF	Effect Size (SD)	H^2
HBB	**Q39X**	.04	0.90	8.0%??
APOE	R176C, C130R	.04, .07	0.56, 0.26	3.3%
PCSK9	R46L, rs2479415	.04, .41	0.38, 0.08	1.2%
LDLR	rs73015013, **V578R**	.14, .005	0.16, 0.62	1.2%
SORT1	rs583104	.18	0.15	0.6%
APOB	rs547235	.19	0.19	0.5%

- 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.

Summary

- Challenges and opportunities in genetic association studies.
- Great need for statistical and computational method development.
- In a specific examples, we …
  - Designed method to combine sequence information across samples.
  - Applied the method to sequence an interesting population in Sardinia.
  - Designed method to infer ancestry from small amounts of sequence.
  - Applied the method to identify additional controls for sequencing study.

Acknowledgements



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