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

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	2.5 milli	Lipid GWAS (Nat Genetics)
2007	Hundred		
2005	Hundred		
2003	Hundred		ap (Nature Genetics)
2002	Hundred		ap (Nature)
2001	Thousands	[illegible]	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	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, 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

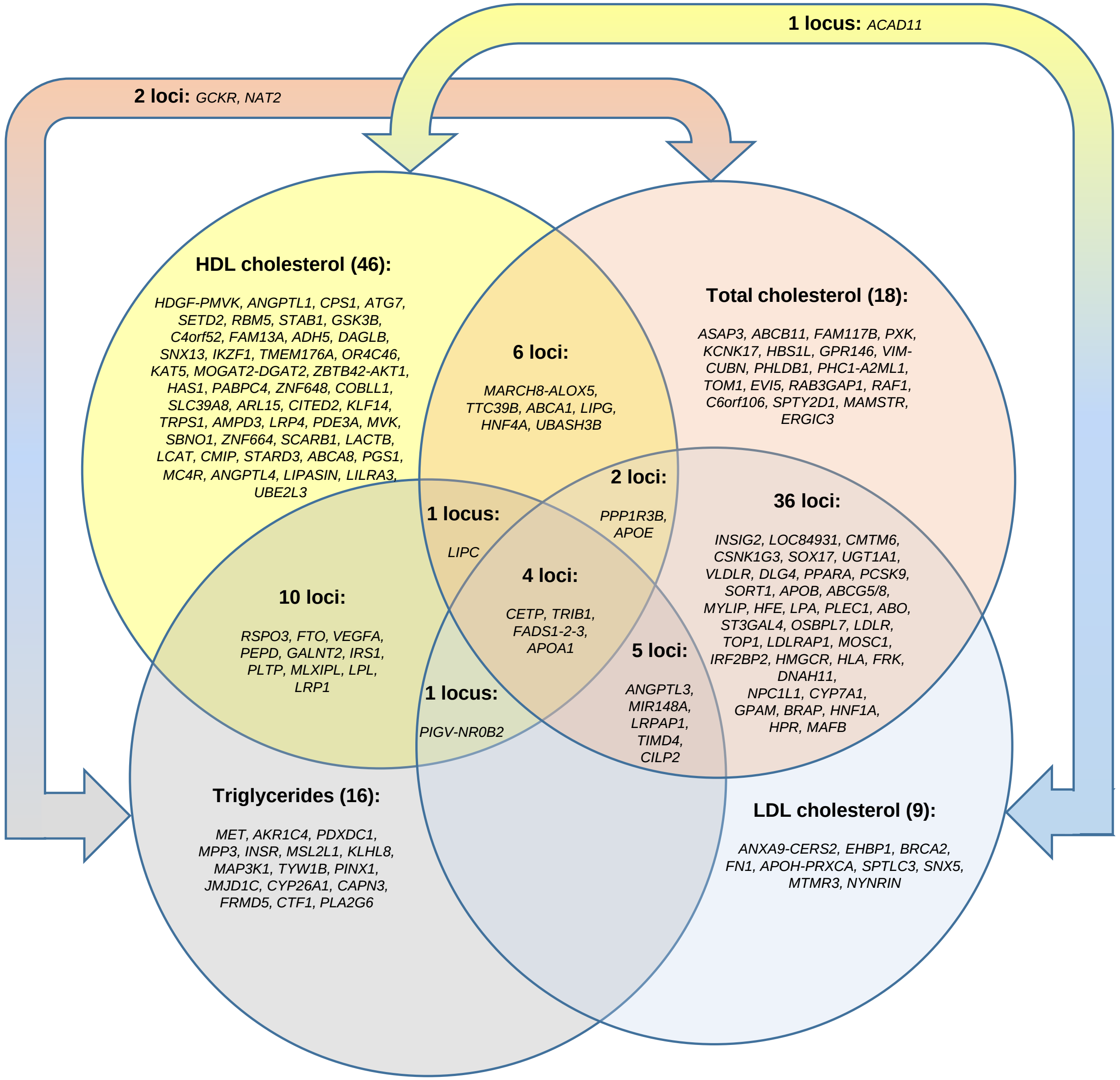
Sekar
Kathiresan

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



1 locus: *ACAD11*

2 loci: *GCKR, NAT2*

HDL cholesterol (46):

HDGF-PMVK, ANGPTL1, CPS1, ATG7, SETD2, RBM5, STAB1, GSK3B, C4orf52, FAM13A, ADH5, DAGLB, SNX13, IKZF1, TMEM176A, OR4C46, KAT5, MOGAT2-DGAT2, ZBTB42-AKT1, HAS1, PABPC4, ZNF648, COBLL1, SLC39A8, ARL15, CITED2, KLF14, TRPS1, AMPD3, LRP4, PDE3A, MVK, SBNO1, ZNF664, SCARB1, LACTB, LCAT, CMIP, STARD3, ABCA8, PGS1, MC4R, ANGPTL4, LIPASIN, LILRA3, UBE2L3

6 loci:

MARCH8-ALOX5, TTC39B, ABCA1, LIPG, HNF4A, UBASH3B

Total cholesterol (18):

ASAP3, ABCB11, FAM117B, PXK, KCNK17, HBS1L, GPR146, VIM-CUBN, PHLDB1, PHC1-A2ML1, TOM1, EVI5, RAB3GAP1, RAF1, C6orf106, SPTY2D1, MAMSTR, ERGIC3

2 loci:

PPP1R3B, APOE

36 loci:

INSIG2, LOC84931, CMTM6, CSNK1G3, SOX17, UGT1A1, VLDLR, DLG4, PPARA, PCSK9, SORT1, APOB, ABCG5/8, MYLIP, HFE, LPA, PLEC1, ABO, ST3GAL4, OSBPL7, LDLR, TOP1, LDLRAP1, MOSC1, IRF2BP2, HMGCR, HLA, FRK, DNAH11, NPC1L1, CYP7A1, GPAM, BRAP, HNF1A, HPR, MAFB

1 locus:

LIPC

4 loci:

CETP, TRIB1, FADS1-2-3, APOA1

10 loci:

RSPO3, FTO, VEGFA, PEPD, GALNT2, IRS1, PLTP, MLXIPL, LPL, LRP1

5 loci:

ANGPTL3, MIR148A, LRPAP1, TIMD4, CILP2

1 locus:

PIGV-NR0B2

Triglycerides (16):

MET, AKR1C4, PDXDC1, MPP3, INSR, MSL2L1, KLHL8, MAP3K1, TYW1B, PINX1, JMJD1C, CYP26A1, CAPN3, FRMD5, CTF1, PLA2G6

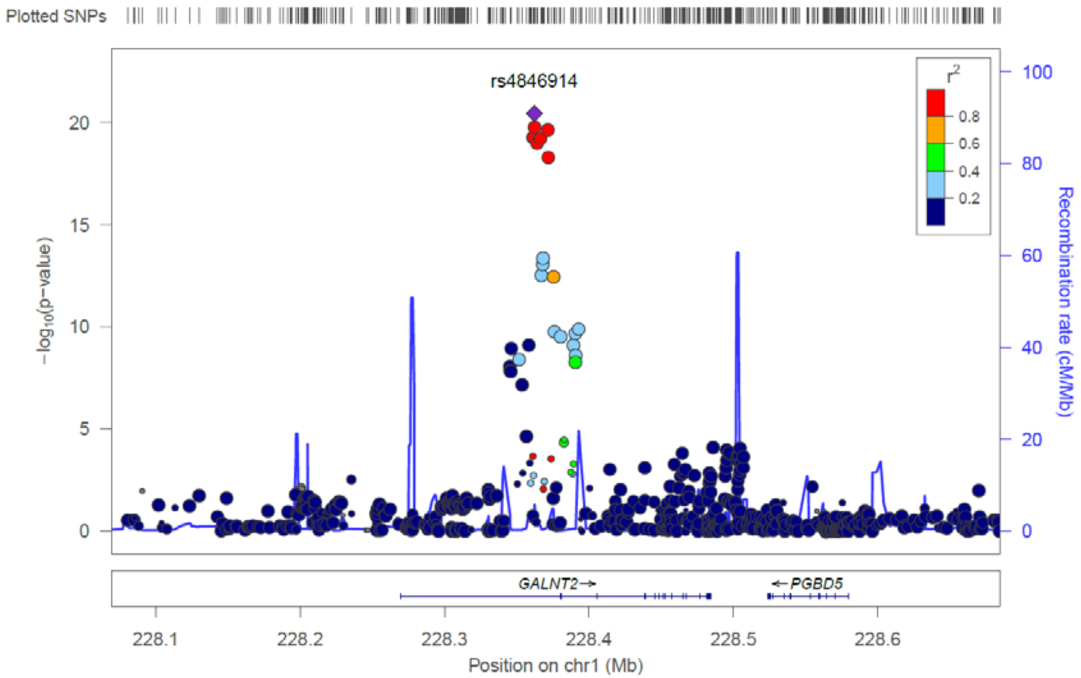
LDL cholesterol (9):

ANXA9-CERS2, EHBP1, BRCA2, FN1, APOH-PRXCA, SPTLC3, SNX5, MTMR3, NYNRIN

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)

Suggesting New Targets: GALNT2

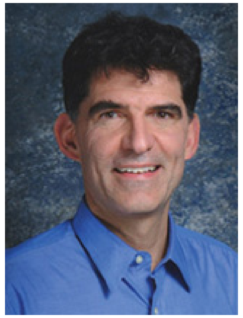


- GWAS allele with 40% frequency associated with ±1 mg/dl in HDL-C

- 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

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.

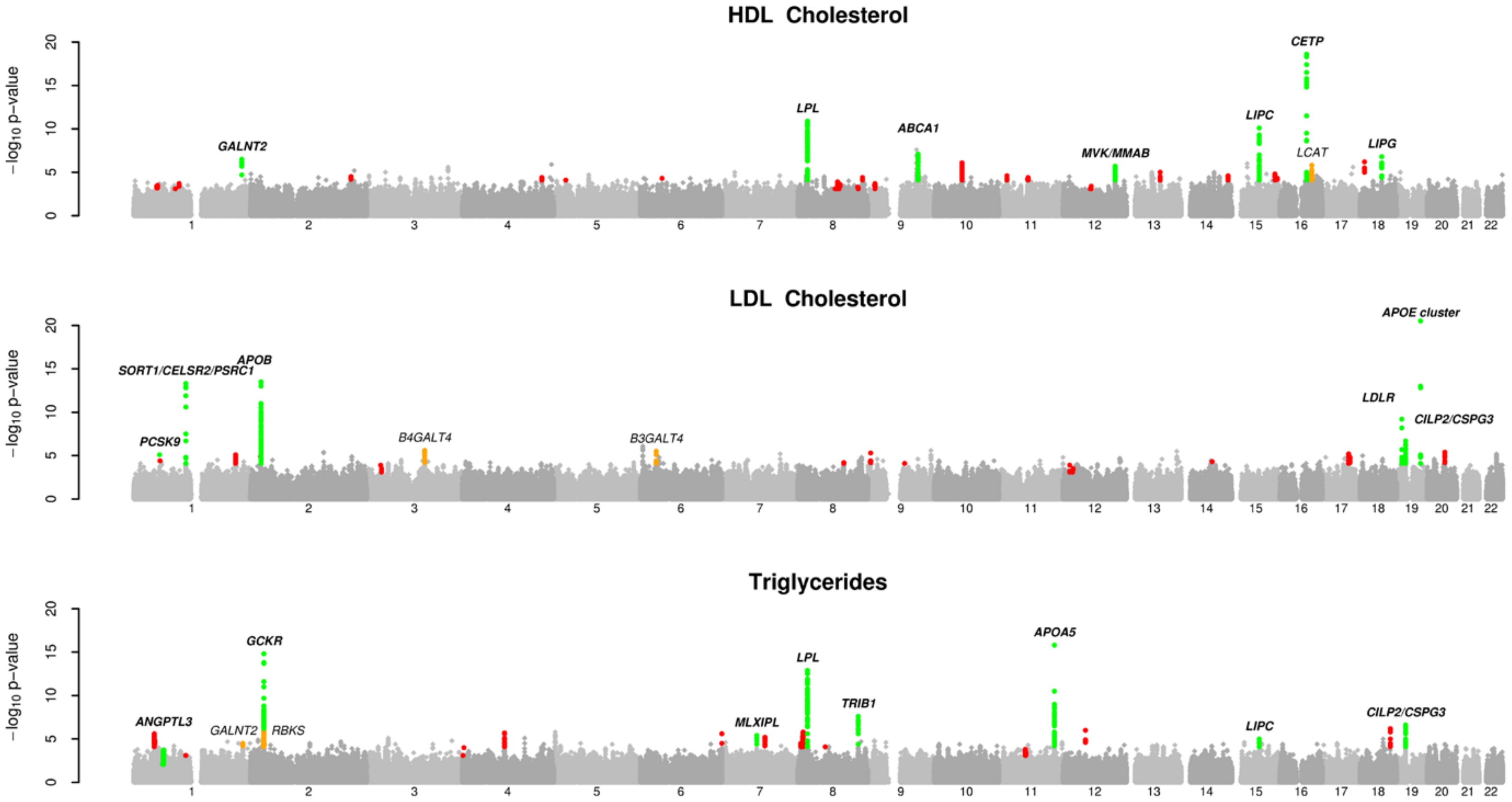
- Are there additional susceptibility loci to be found?
  - Only subset of functional elements include common variants …
  - Rare variants are more numerous and thus will point to additional loci

What Is the Total Contribution of Each Locus?

Evidence that

Multiple Variants Will be Important

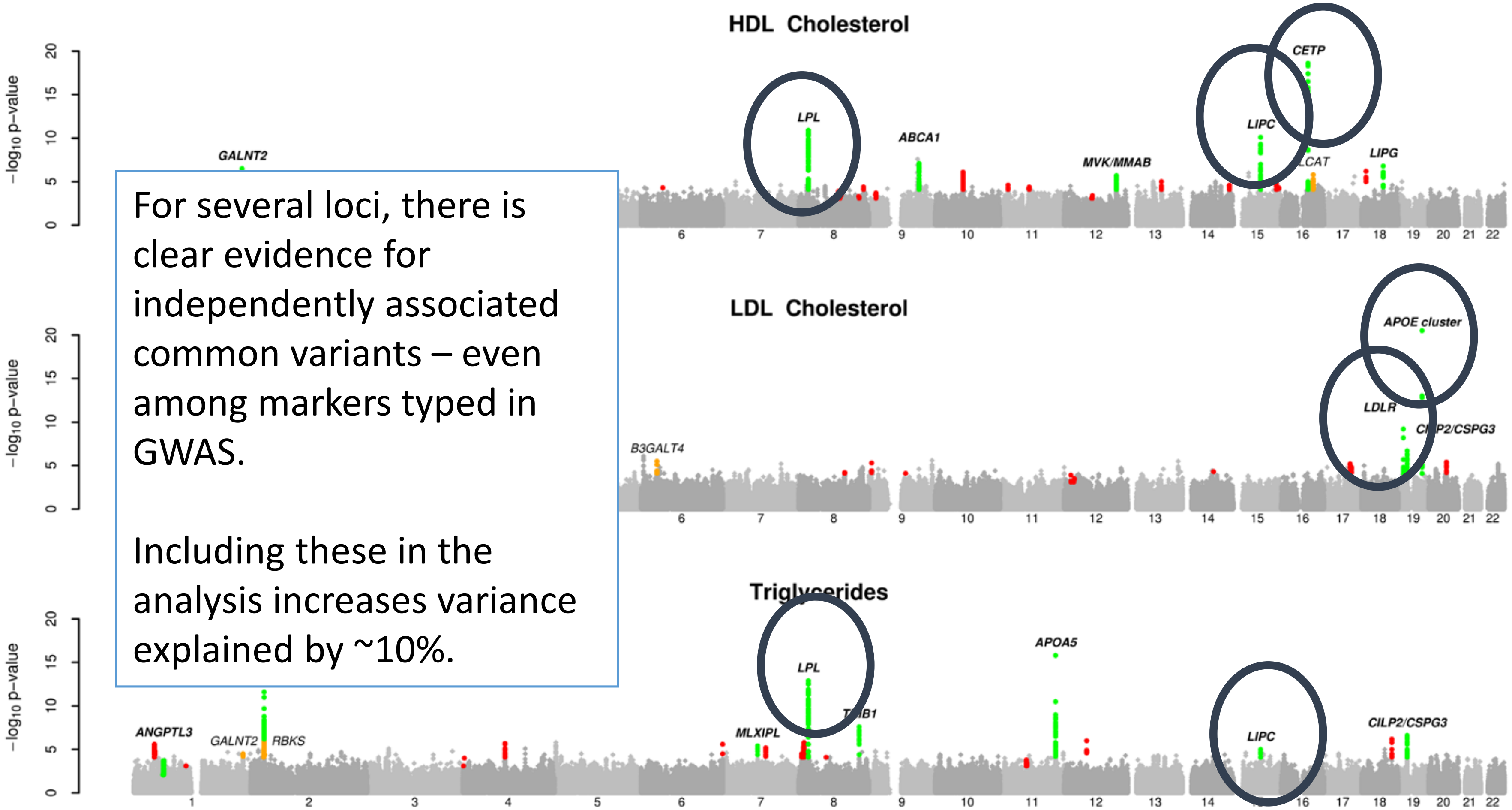
Evidence for Multiple Variants Per Locus Example from Lipid Biology



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%.

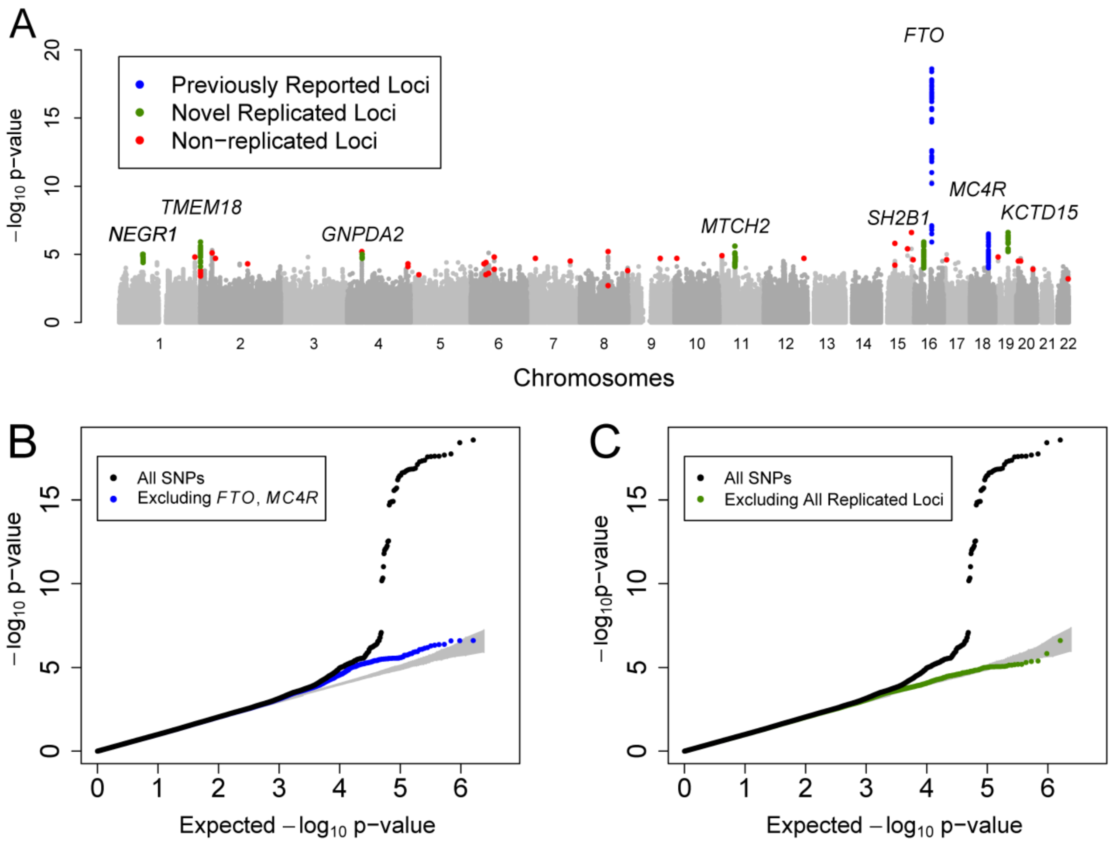
Willer et al, *Nat Genet*, 2008
Kathiresan et al, *Nat Genet,* 2008, 2009

What is The Contribution of Structural Variants?

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

but 100s of Structural Variants

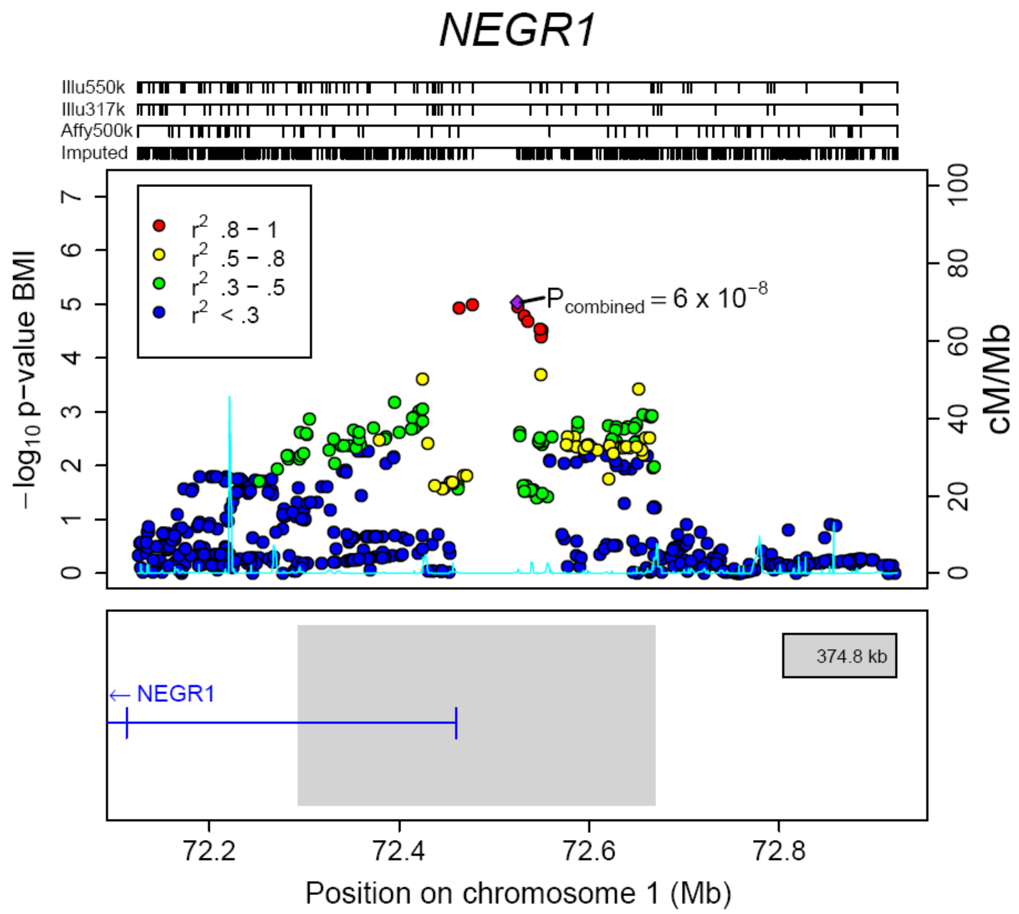
Evidence that Copy Number Variants Important Example from Genetics of Obesity



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

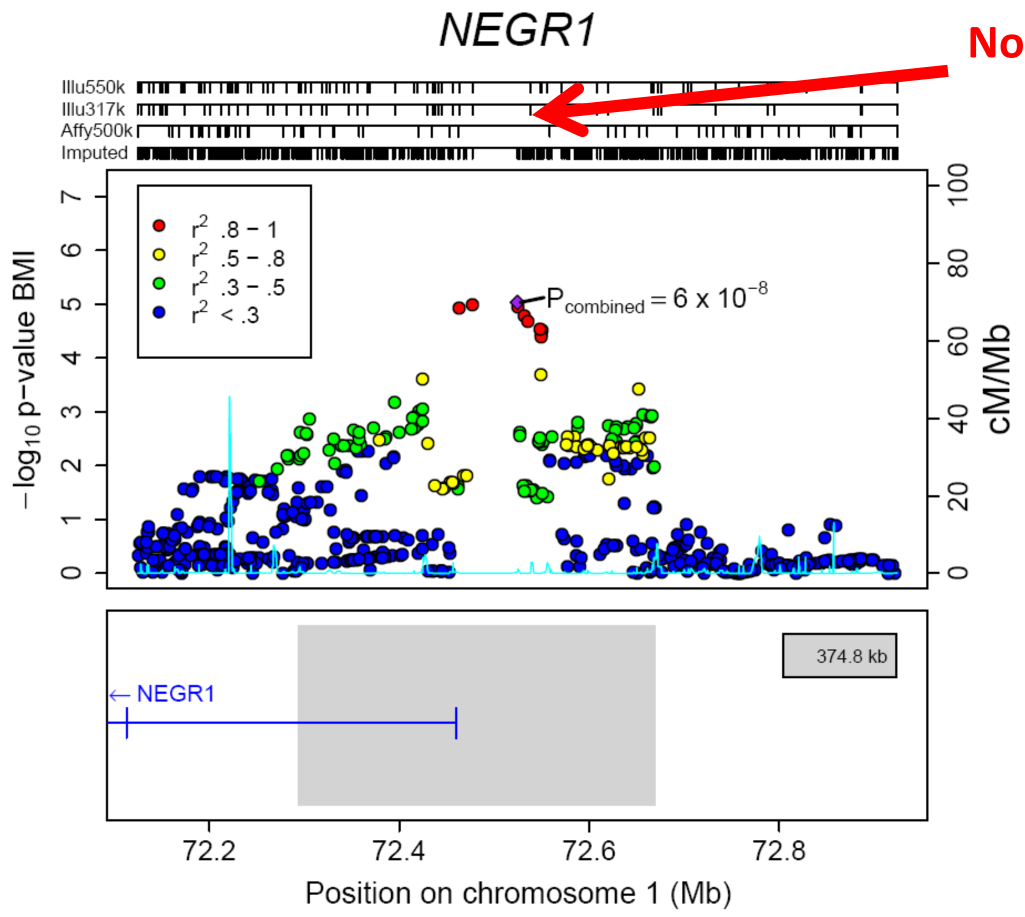
Willer et al, *Nature Genetics,* 2009

Evidence that Copy Number Variants Important Example from Genetics of Obesity



Willer et al, *Nature Genetics*, 2009

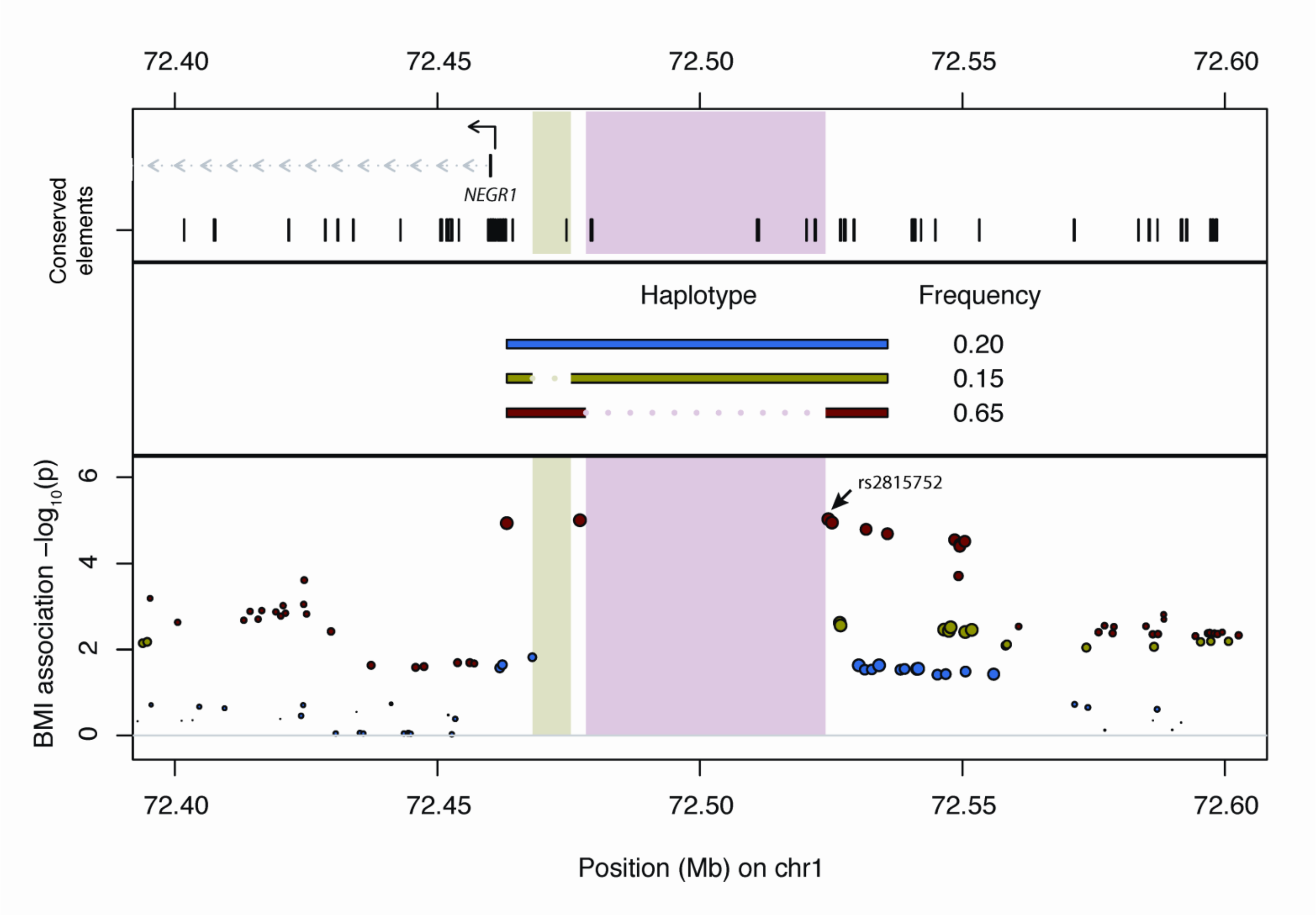
Evidence that Copy Number Variants Important Example from Genetics of Obesity



Note hole in marker panels….

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

ACTGGTCGATGCTAGCTGATAGCTAGCTA

GCTGATGAGCCCGATCGCTGCTAGCTCG

AGCTGATAGCTAGCTAGCTGATGAGCCCGA

GAGCCCGATCGCTGCTAGCTCGACG

- 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

Per Base Alignment Qualities

Short Read

GATAGCTAG**C**TAGCTGATGA GCCG

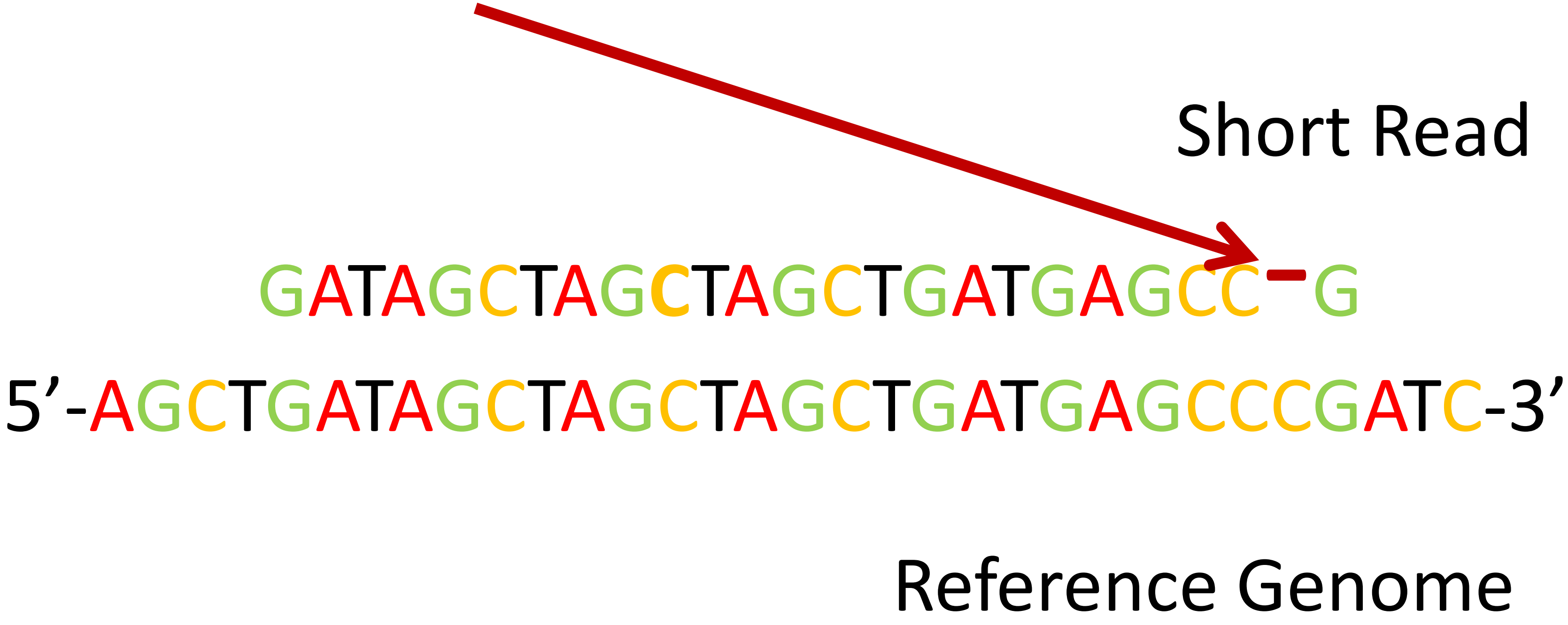
5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Heng Li

Per Base Alignment Qualities

Should we insert a gap?



Short Read

GATAGCTAGCTAGCTGATGAGCC-G

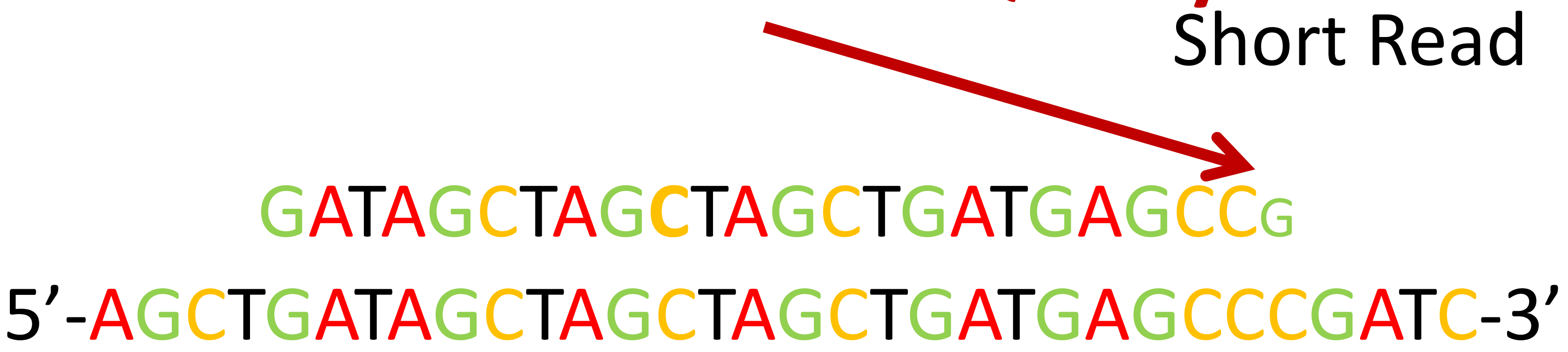
5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Heng Li

Per Base Alignment Qualities

Compensate for Alignment Uncertainty With Lower Base Quality



Short Read

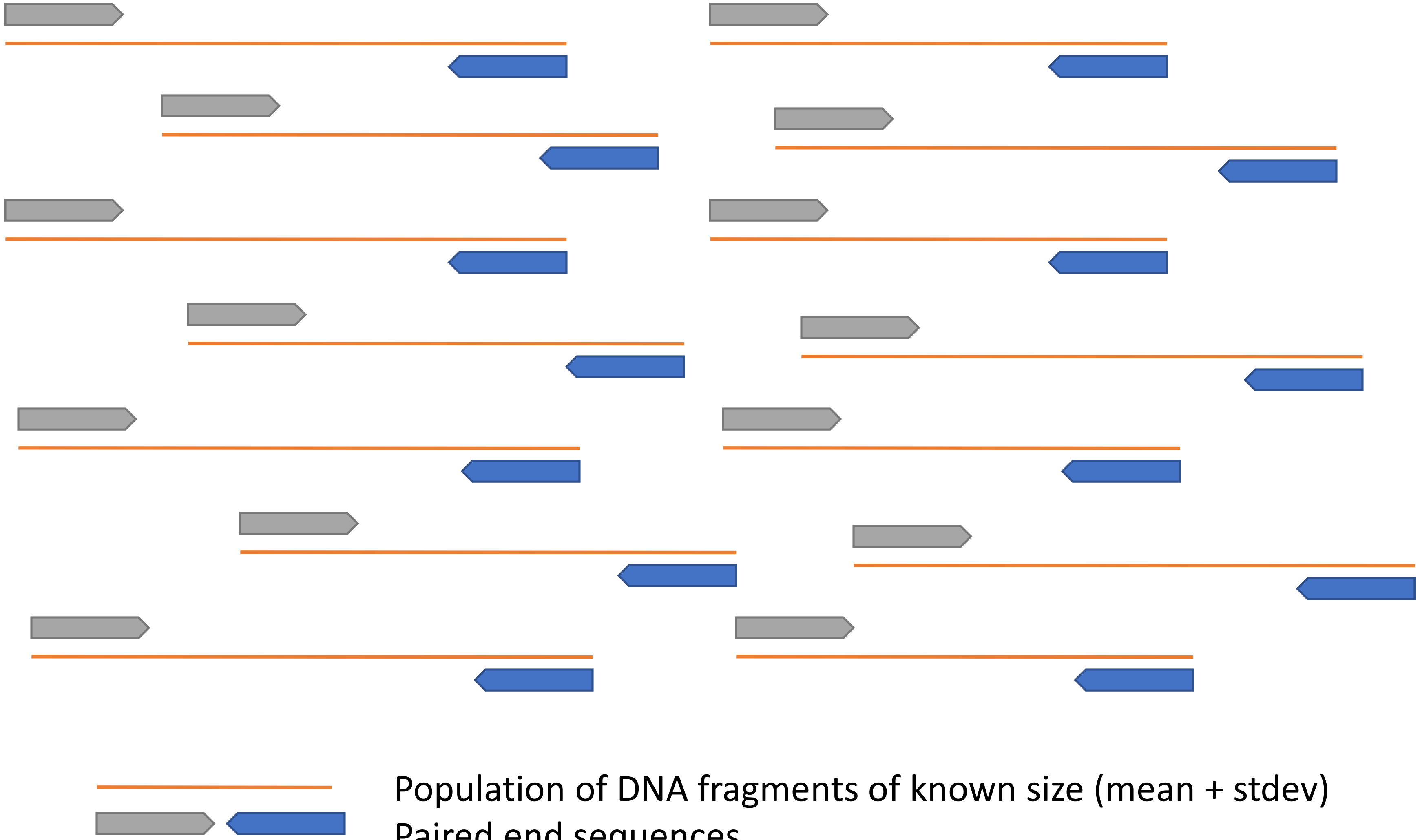
GATAGCTAGCTAGCTGATGAGCCG

5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Heng Li

Paired End Sequencing

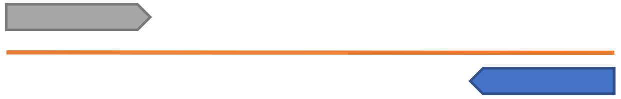


Population of DNA fragments of known size (mean + stdev)

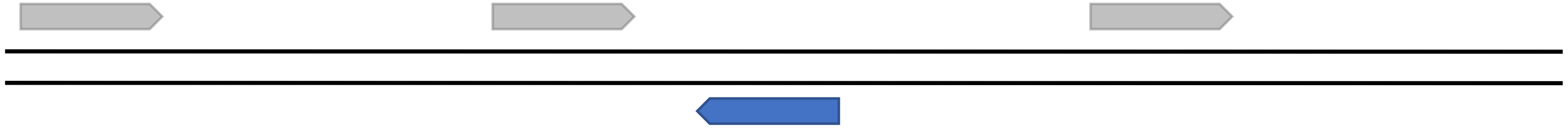
Paired end sequences

Paired End Sequencing

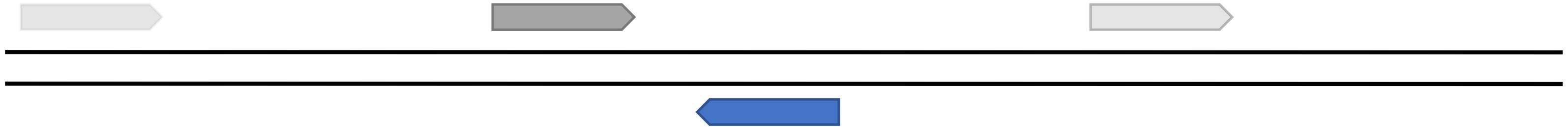
Paired Reads



Initial alignment to the reference genome

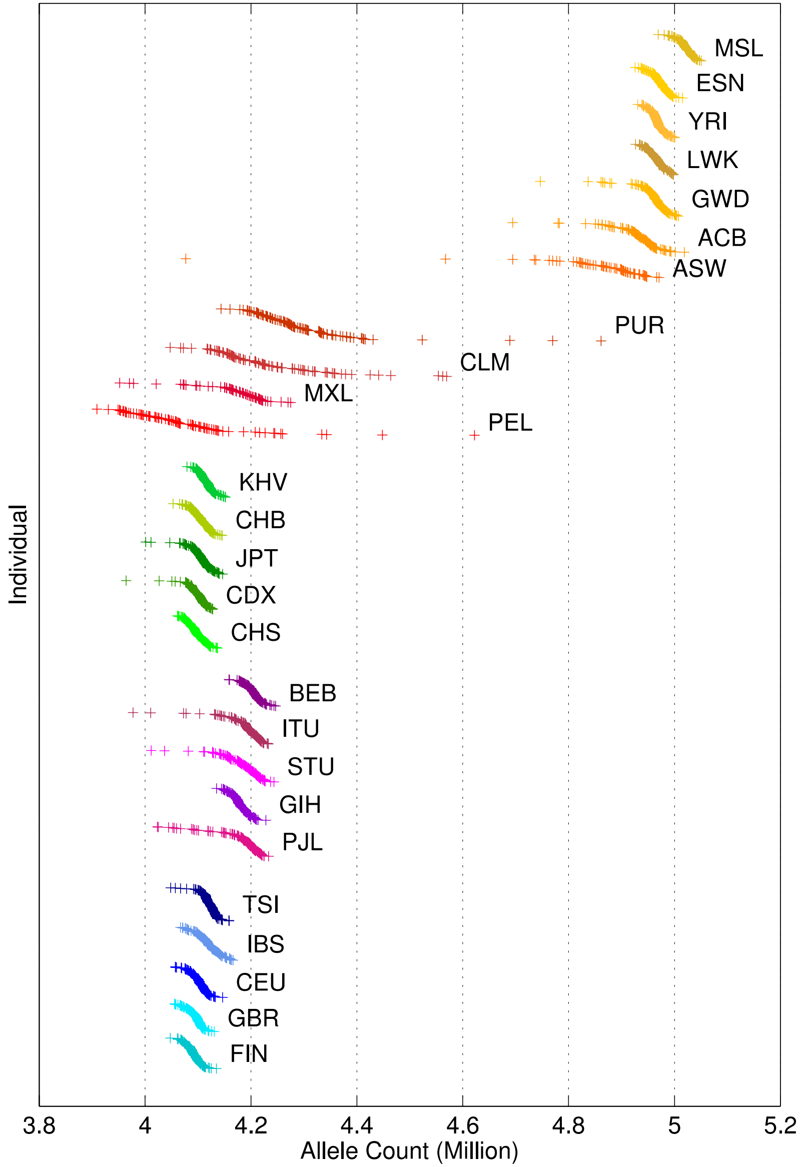


Paired end resolution



How much variation is there?

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



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

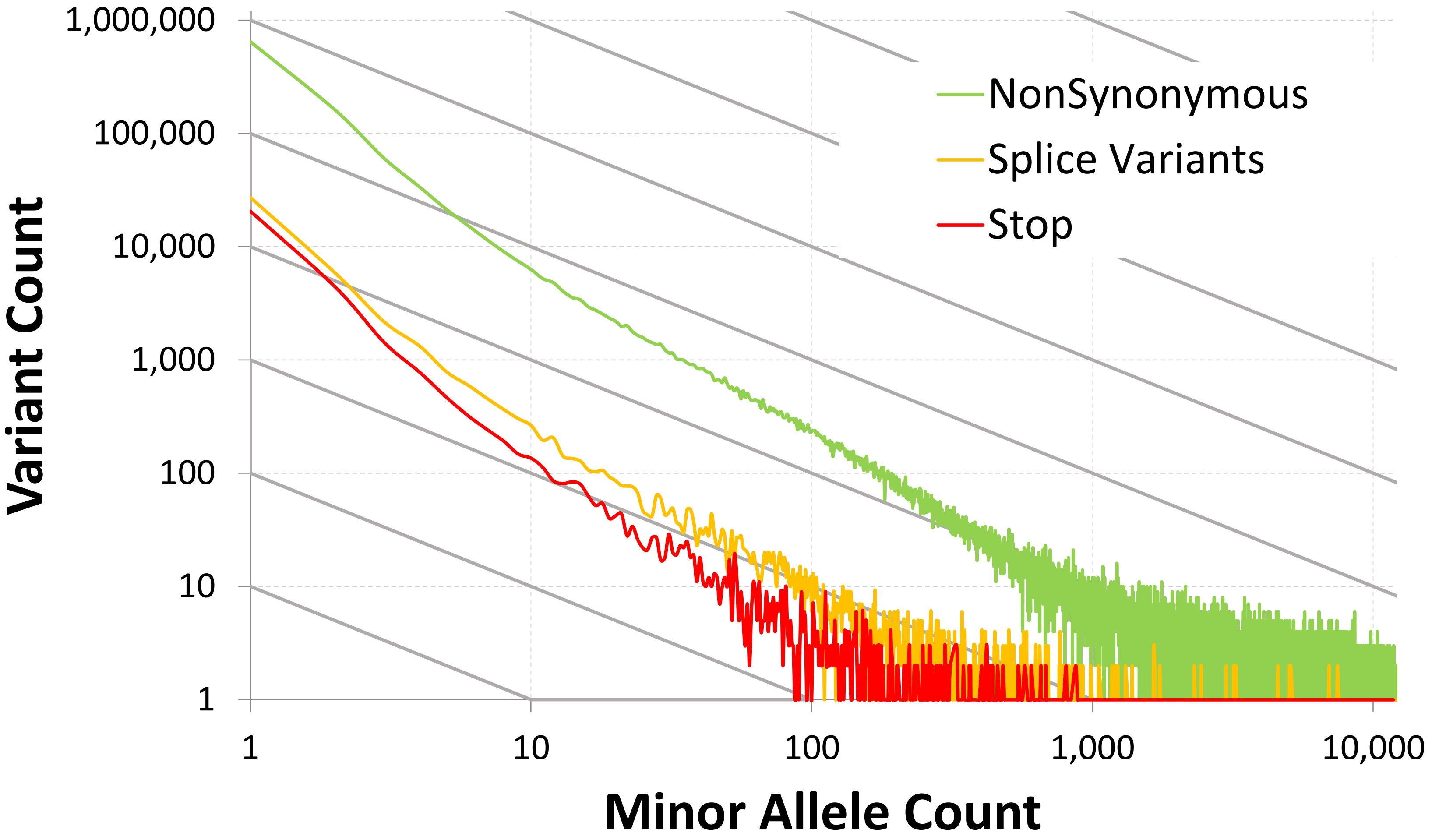
- Michigan caller combines …
  - Markov models to identify shared haplotypes,
  - Classifiers to distinguish true variants from error,
  - Strategies to distribute computation across cluster

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**

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



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

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

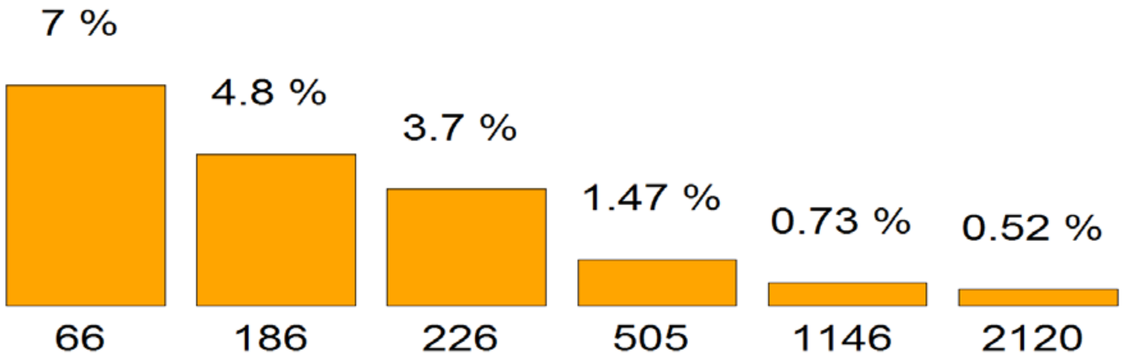
- In our own hands…
  - Mapped high quality bases
  - March 2010: ~5.0 Gb/lane
  - May 2010: ~7.5 Gb/lane
  - September 2010: ~8.6 Gb/lane
  - January 2011: ~16 Gb/lane
  - Summer 2011: ~45 Gb/lane

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

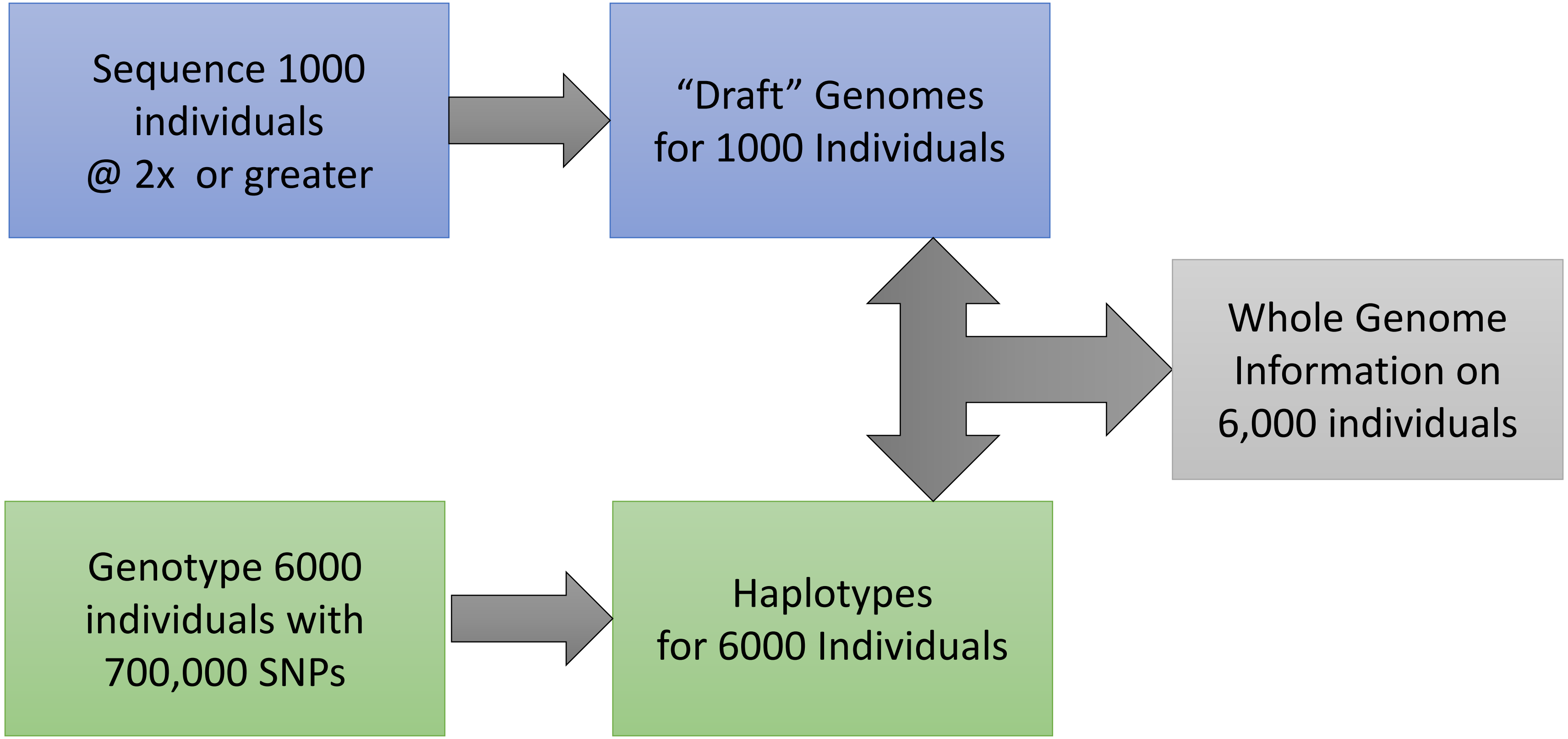
Fabio Busonero, Andrea Maschio

As more samples are sequenced, Accuracy increases

Heterozygous Mismatch Rate (in %)



Design



Sequence 1000 individuals @ 2x or greater

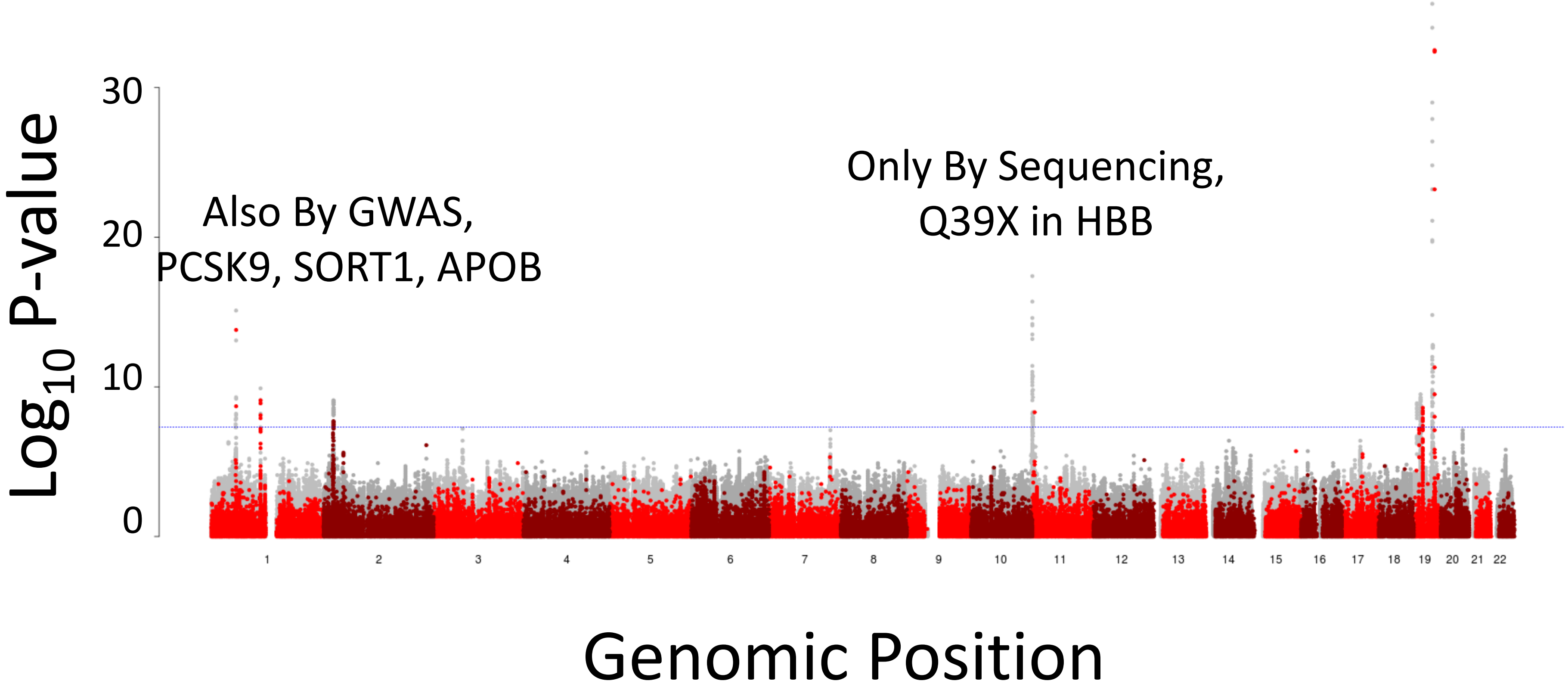
“Draft” Genomes for 1000 Individuals

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



Also By GWAS,
PCSK9, SORT1, APOB

Only By Sequencing,
Q39X in HBB

Also By GWAS,
LDLR, APOE

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

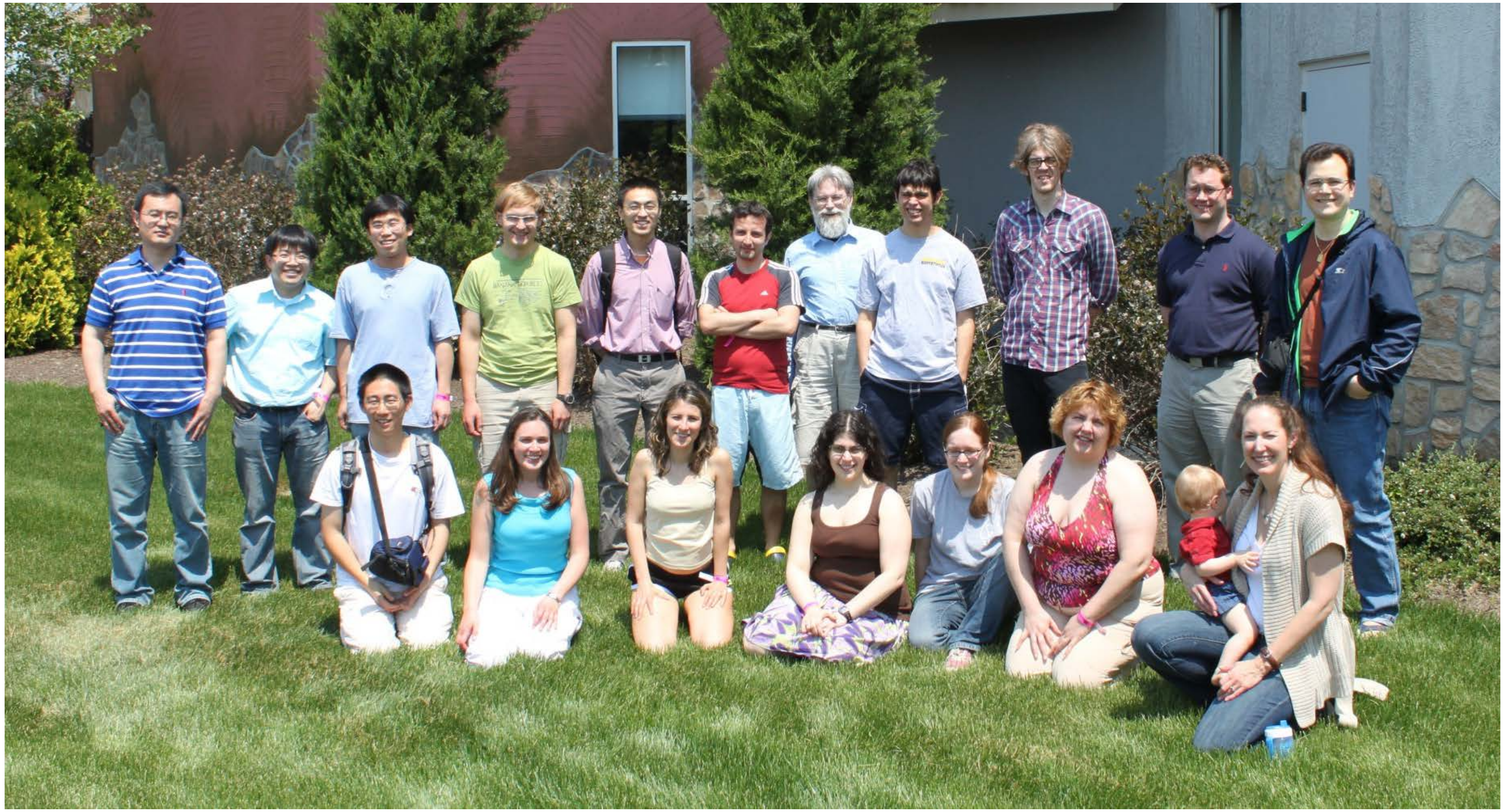
- Challenges and opportunities in genetic association studies.

- 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.