Adventures in Human Genetics

Goncalo Abecasis

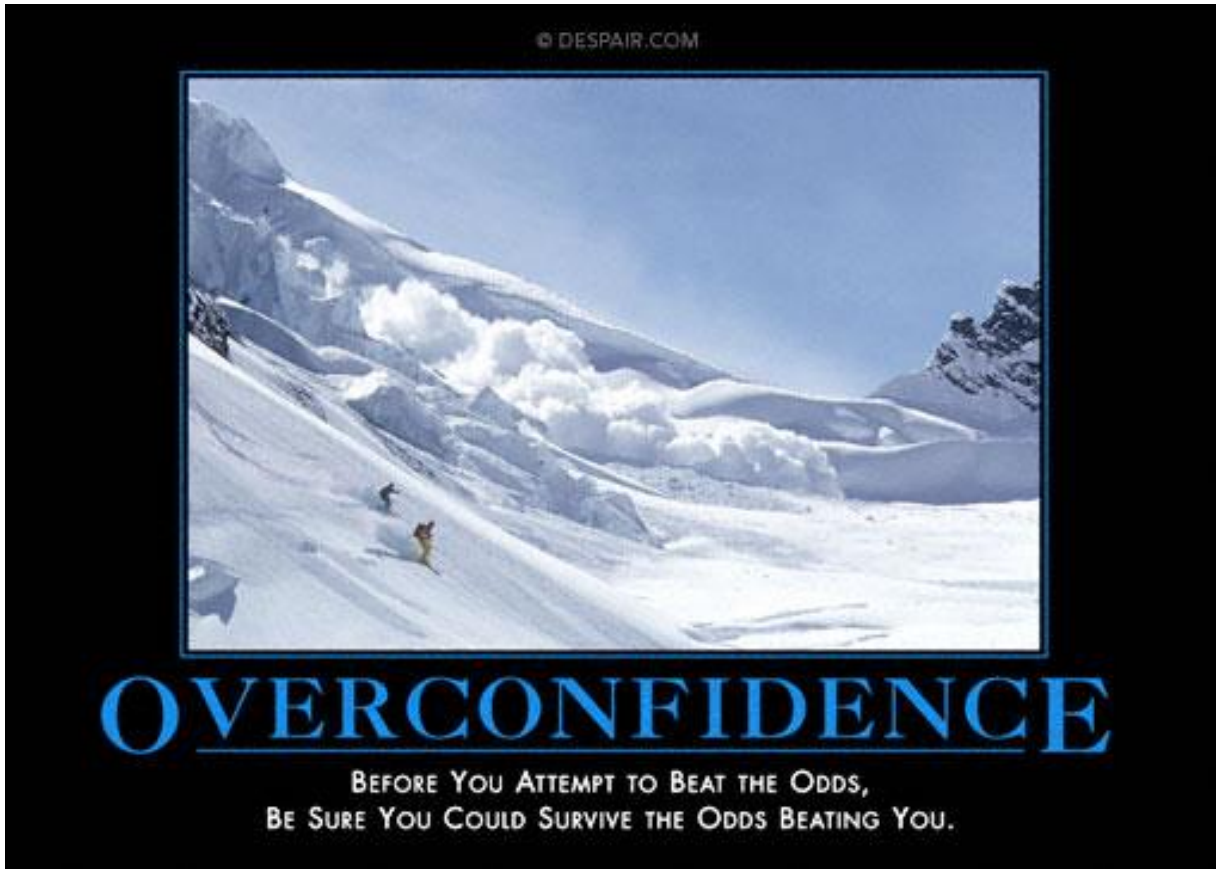
University of Michigan School of Public Health

@gabecasis

A motivational talk?

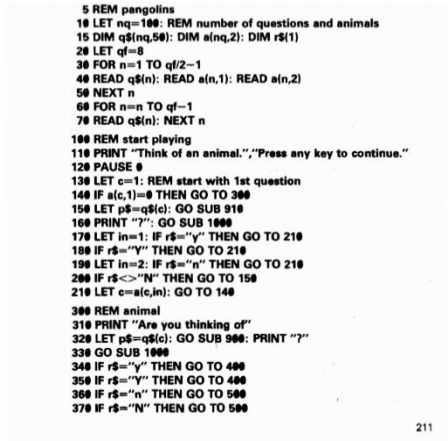
- Many opportunities for computational biology …
- 10,000s of sequenced human genomes.
- Bigger datasets than we have ever handled before.

A humorous talk?



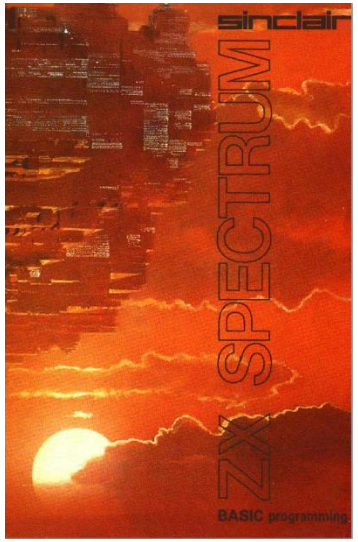
It is a larger dataset than we have ever handled…
But we can do it!

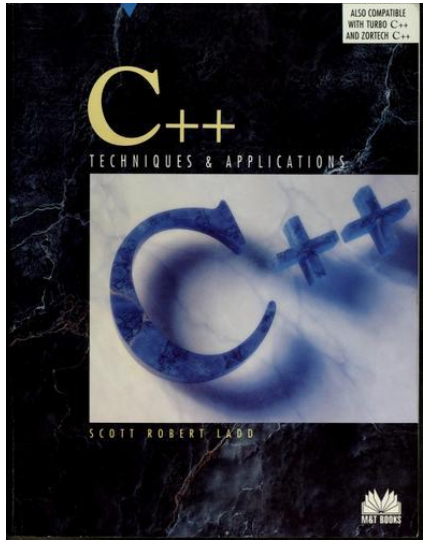
Should we start from the beginning?

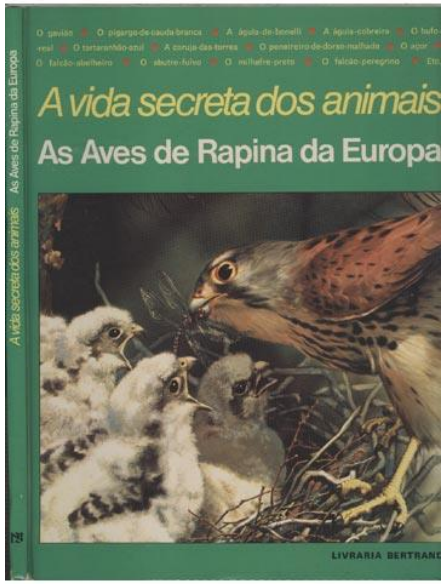


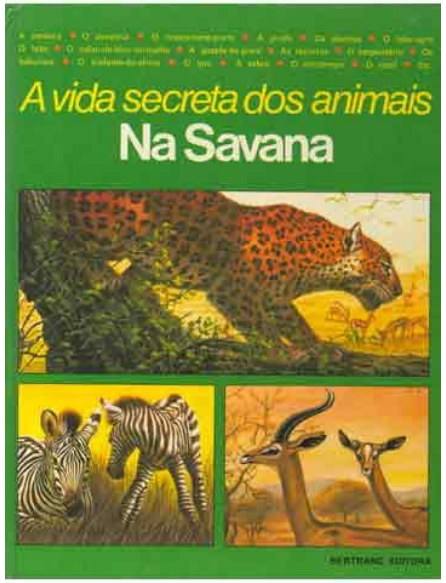
```
 5 REM pangolins
10 LET nq=100: REM number of questions and animals
15 DIM q$(nq,50): DIM a(nq,2): DIM r$(1)
20 LET qf=8
30 FOR n=1 TO qf/2-1
40 READ q$(n): READ a(n,1): READ a(n,2)
50 NEXT n
60 FOR n=n TO qf-1
70 READ q$(n): NEXT n

100 REM start playing
110 PRINT "Think of an animal.","Press any key to continue."
120 PAUSE 0
130 LET c=1: REM start with 1st question
140 IF a(c,1)=0 THEN GO TO 300
150 LET p$=q$(c): GO SUB 910
160 PRINT "?": GO SUB 1000
170 LET in=1: IF r$="y" THEN GO TO 210
180 IF r$="Y" THEN GO TO 210
190 LET in=2: IF r$="n" THEN GO TO 210
200 IF r$<>"N" THEN GO TO 150
210 LET c=a(c,in): GO TO 140

300 REM animal
310 PRINT "Are you thinking of"
320 LET p$=q$(c): GO SUB 900: PRINT "?"
330 GO SUB 1000
340 IF r$="y" THEN GO TO 400
350 IF r$="Y" THEN GO TO 400
360 IF r$="n" THEN GO TO 500
370 IF r$="N" THEN GO TO 500
```

211

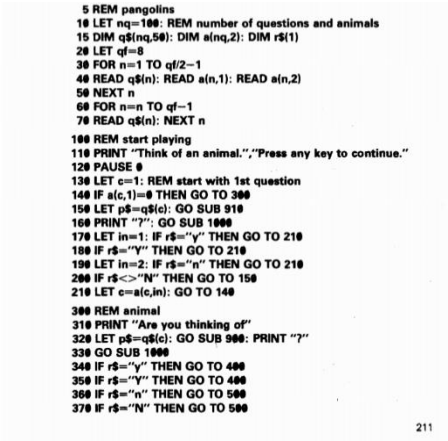






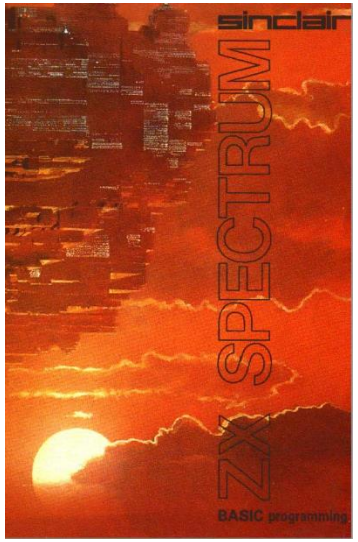


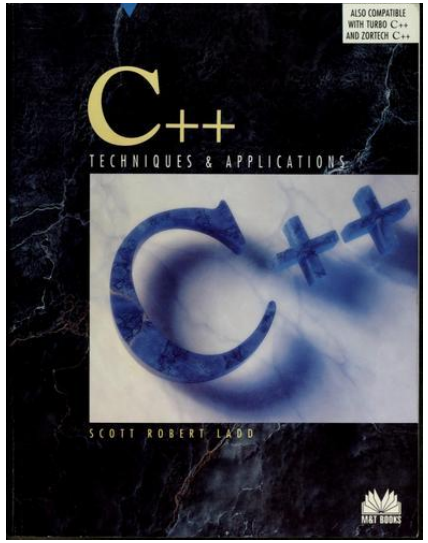
Should we start from the beginning?

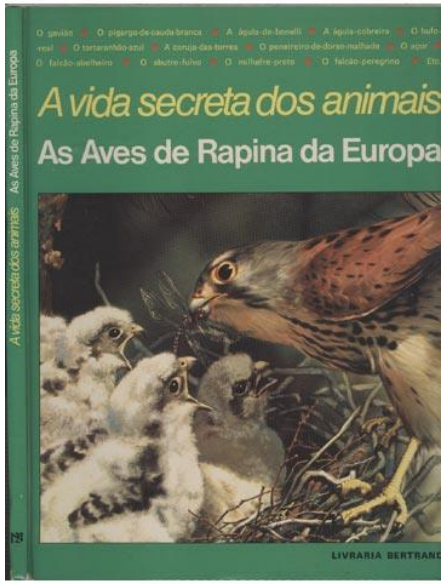


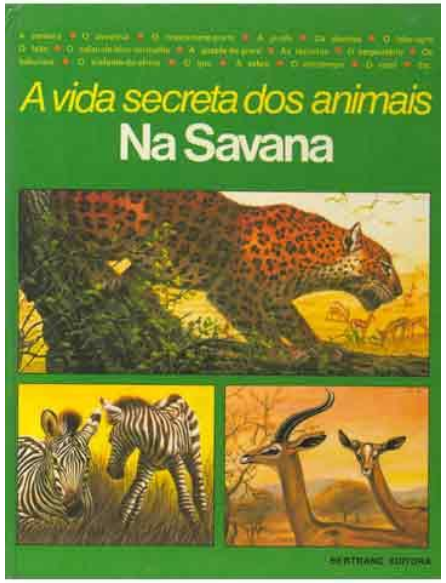
```
 5 REM pangolins
10 LET nq=100: REM number of questions and animals
15 DIM q$(nq,50): DIM a(nq,2): DIM r$(1)
20 LET qf=8
30 FOR n=1 TO qf/2-1
40 READ q$(n): READ a(n,1): READ a(n,2)
50 NEXT n
60 FOR n=n TO qf-1
70 READ q$(n): NEXT n

100 REM start playing
110 PRINT "Think of an animal.","Press any key to continue."
120 PAUSE 0
130 LET c=1: REM start with 1st question
140 IF a(c,1)=0 THEN GO TO 300
150 LET p$=q$(c): GO SUB 910
160 PRINT "?": GO SUB 1000
170 LET in=1: IF r$="y" THEN GO TO 210
180 IF r$="Y" THEN GO TO 210
190 LET in=2: IF r$="n" THEN GO TO 210
200 IF r$<>"N" THEN GO TO 150
210 LET c=a(c,in): GO TO 140

300 REM animal
310 PRINT "Are you thinking of"
320 LET p$=q$(c): GO SUB 900: PRINT "?"
330 GO SUB 1000
340 IF r$="y" THEN GO TO 400
350 IF r$="Y" THEN GO TO 400
360 IF r$="n" THEN GO TO 500
370 IF r$="N" THEN GO TO 500
```

211





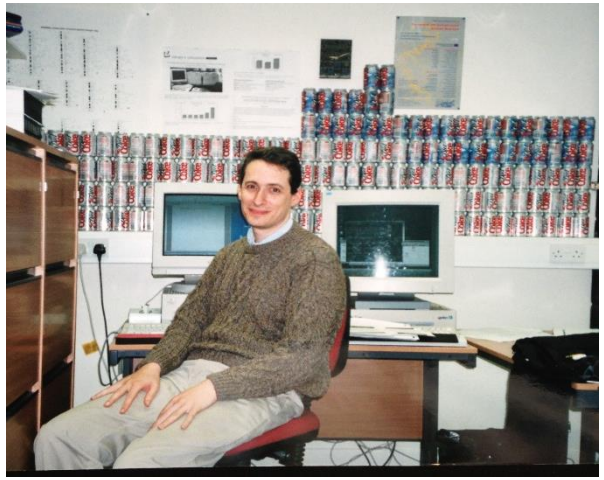




Perhaps we don’t need to go quite this far back!

My start in human genetics …

- Wellcome Trust Center for Human Genetics (1997-2001)
- Developing and applying early SNP discovery and genotyping technologies to genetic studies of asthma
- Complex trait studies were shifting in focus from linkage to association mapping
- A big question concerned move from family samples, which are ideal for linkage analysis, to unrelated samples, which are better suited for association mapping
- Working with William Cookson and Lon Cardon



1997 - 2001

Association Mapping in Families…

Am. J. Hum. Genet. 66:279–292, 2000

A General Test of Association for Quantitative Traits in Nuclear Families

G. R. Abecasis, L. R. Cardon, and W. O. C. Cookson

The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford

Summary

High-resolution mapping is an important step in the identification of complex disease genes. In outbred populations, linkage disequilibrium is expected to operate over short distances and could provide a powerful fine-mapping tool. Here we build on recently developed methods for linkage-disequilibrium mapping of quantitative traits to construct a general approach that can

Introduction

Increasingly large numbers of single-nucleotide polymorphisms are available in public and private databases (Collins et al. 1997). The emergence of high-throughput methods for their analysis holds promise for saturation mapping of human complex-disease loci (Risch and Merikangas 1996; Chakravarti 1998; Lander 1999). Whereas allele-sharing methods of linkage analysis can

European Journal of Human Genetics (2000) 8, 545–551

www.nature.com/ejhg

ARTICLE

Pedigree tests of transmission disequilibrium

Gonçalo R Abecasis, William OC Cookson and Lon R Cardon

Wellcome Trust Center for Human Genetics, University of Oxford, UK

High-resolution mapping is essential for the positional cloning of complex disease genes. In outbred populations, linkage disequilibrium is expected to extend for short distances and could provide a powerful fine-mapping tool. Current family-based association tests use nuclear family members to define allelic transmission and controls, but ignore other types of relatives. Here we construct a general approach for scoring allelic transmission that accommodates families of any size and uses all available genotypic information. Family data allows for the construction of an expected genotype for every non-founder, and orthogonal deviates from this expectation are a measure of allelic transmission. These allelic transmission scores can be used to extend previously described tests of linkage disequilibrium for dichotomous or quantitative traits. Some of these tests are illustrated, together with a permutation framework for estimating exact significance levels. Simulation studies are used to investigate power and error rates of the

Association Analysis in a Variance Components Framework

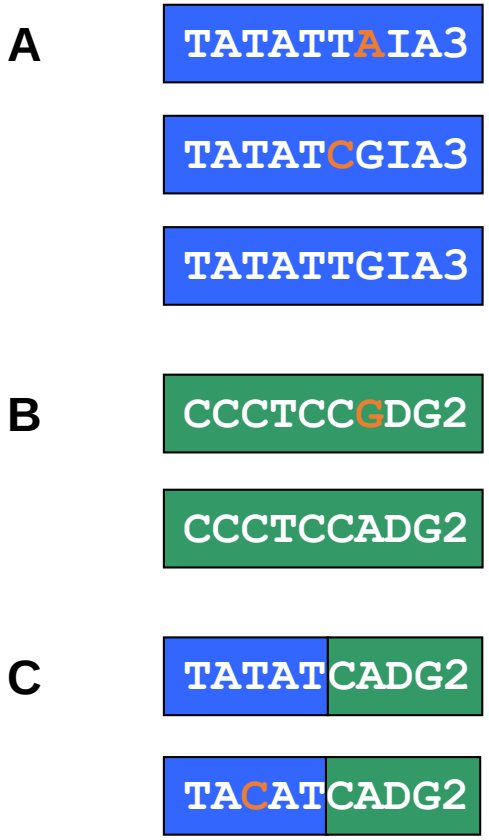
Gonçalo R. Abecasis, Lon R. Cardon, William O.C. Cookson, Pak C. Sham, and Stacey S. Cherny

Wellcome Trust Centre for Human Genetics (G.R.A., L.R.C., W.O.C.C., S.S.C.), University of Oxford, Oxford; Social, Genetic and Developmental Psychiatry Research Center and Department of Psychiatry (P.C.S.), Institute of Psychiatry, London, United Kingdom

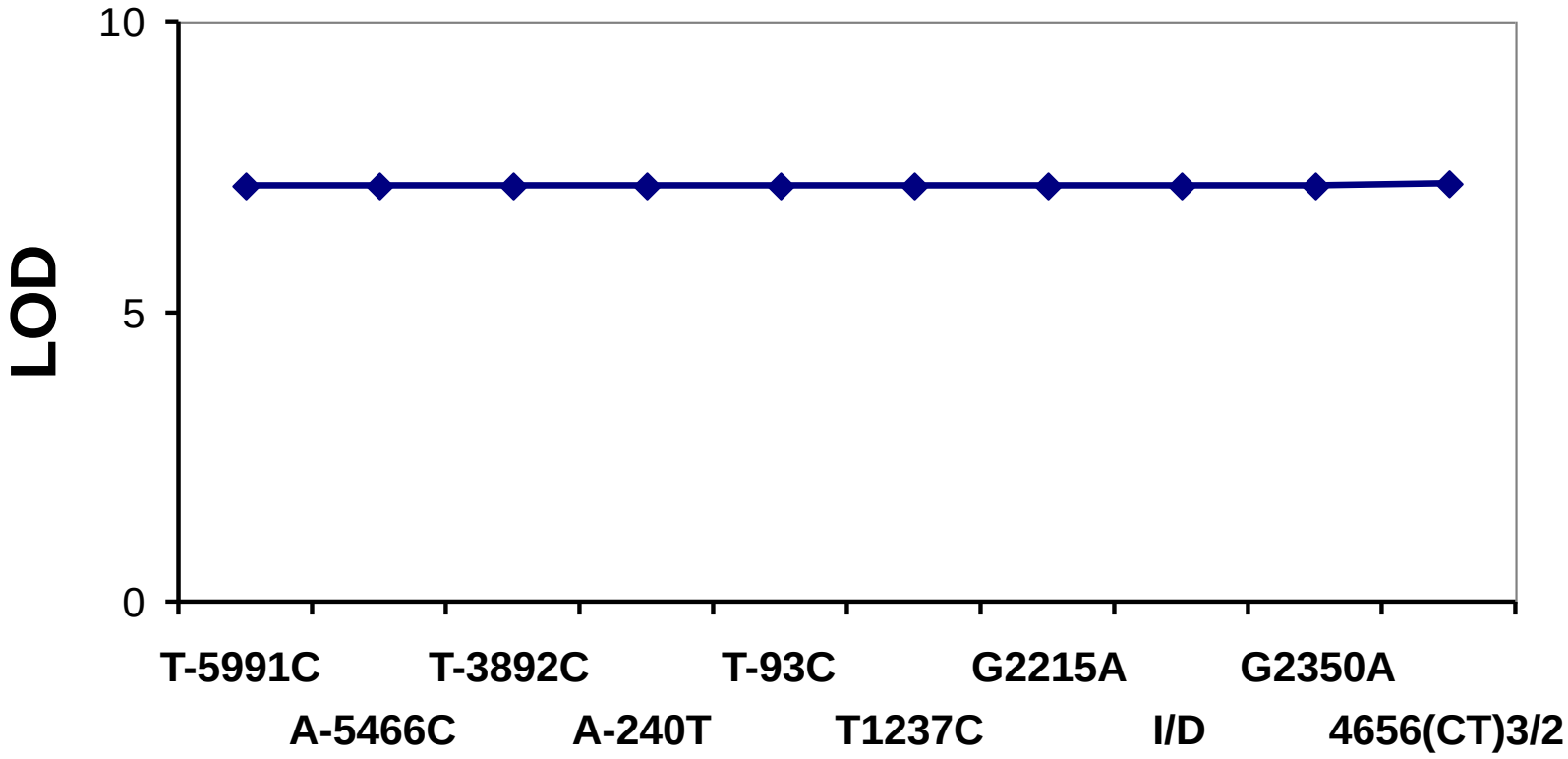
"…association at genomewide significance levels (that is P < 5x10-8 corresponding to 1,000,000 independent tests)…"

The Angiotensin Converting Enzyme…

- Data collected by Bernard Keavney and Colin McKenzie
- ACE levels and genotypes for 10 SNPs in a collection of families
- Broadly speaking, the 10 SNPs are organized into 3 common haplotypes
- The first true genetic association I saw!

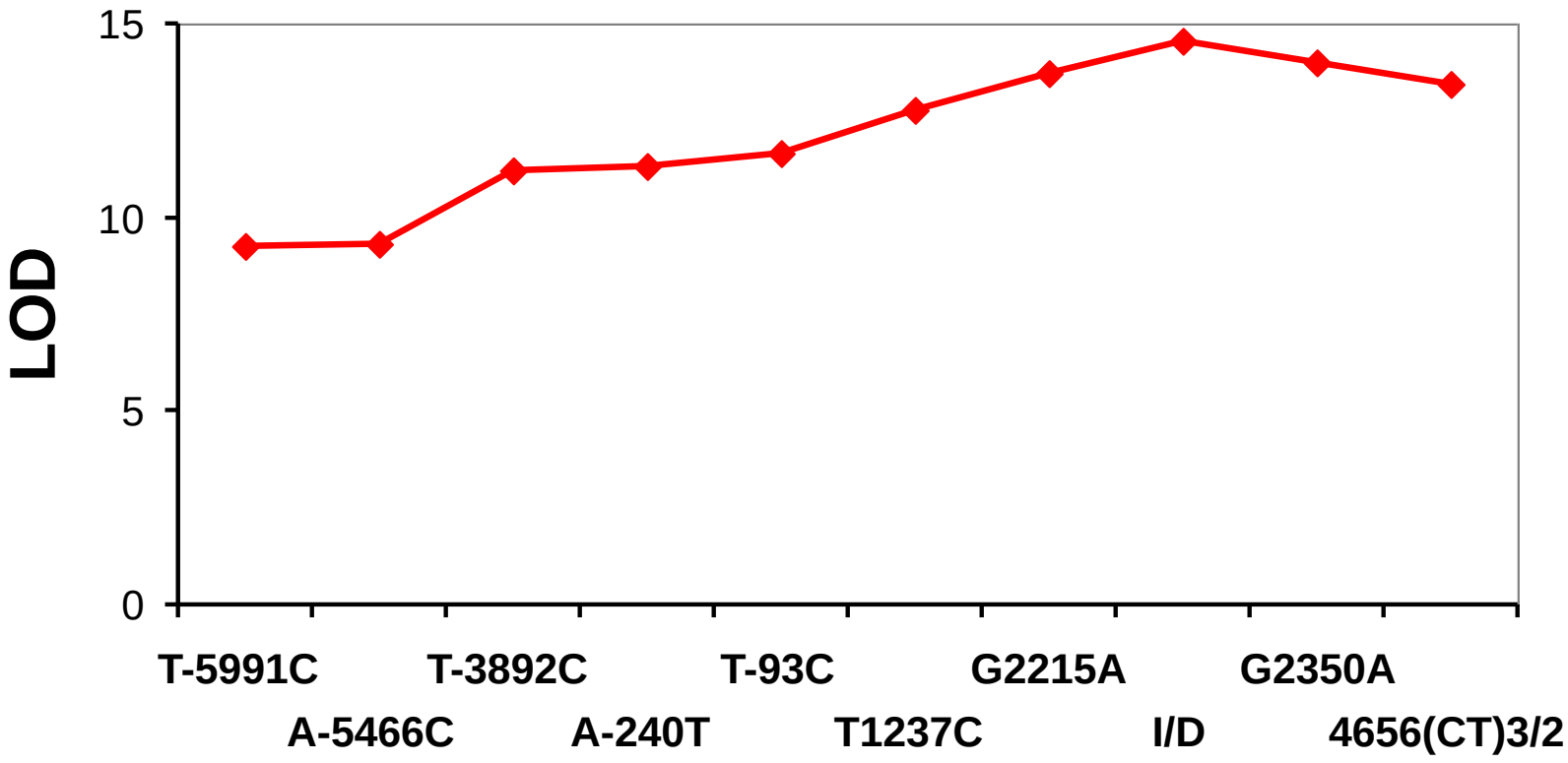


Linkage: *ACE* gene and ACE levels



$$H_0 : (\mu, \sigma_g^2, \sigma_e^2) \qquad H_1 : (\mu, \sigma_g^2, \sigma_e^2, \sigma_a^2)$$

Association: *ACE* gene and ACE levels



$$H_0 : (\mu, \sigma_g^2, \sigma_a^2, \sigma_e^2, \beta_b) \qquad H_1 : (\mu, \sigma_g^2, \sigma_a^2, \sigma_e^2, \beta_b, \beta_w)$$

March/April 2002 · Vol. 4 · No. 2

review

A comprehensive review of genetic association studies

Joel N. Hirschhorn, MD, PhD[1–3], Kirk Lohmueller[1], Edward Byrne[1], and Kurt Hirschhorn, MD[4]

"… of the 166 associations which have been studied 3 or more times, only six have been consistently replicated."

Hirschhorn et al (2002)

Patterns of Linkage Disequilibrium in the Genome

Abecasis et al (Bioinformatics, 2000)

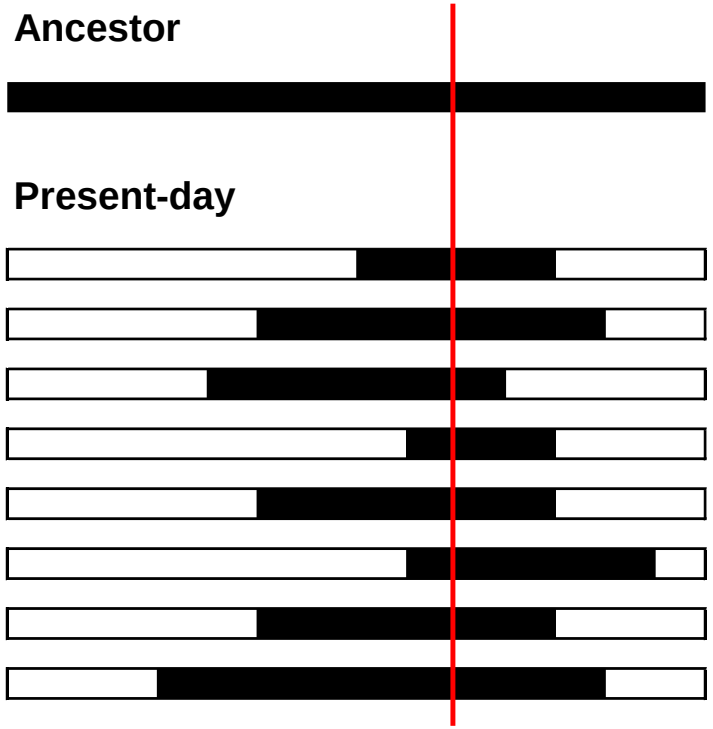
Abecasis et al (Am J Hum Genet, 2001)

Dawson et al (Nature, 2002)

The HapMap Consortium Days

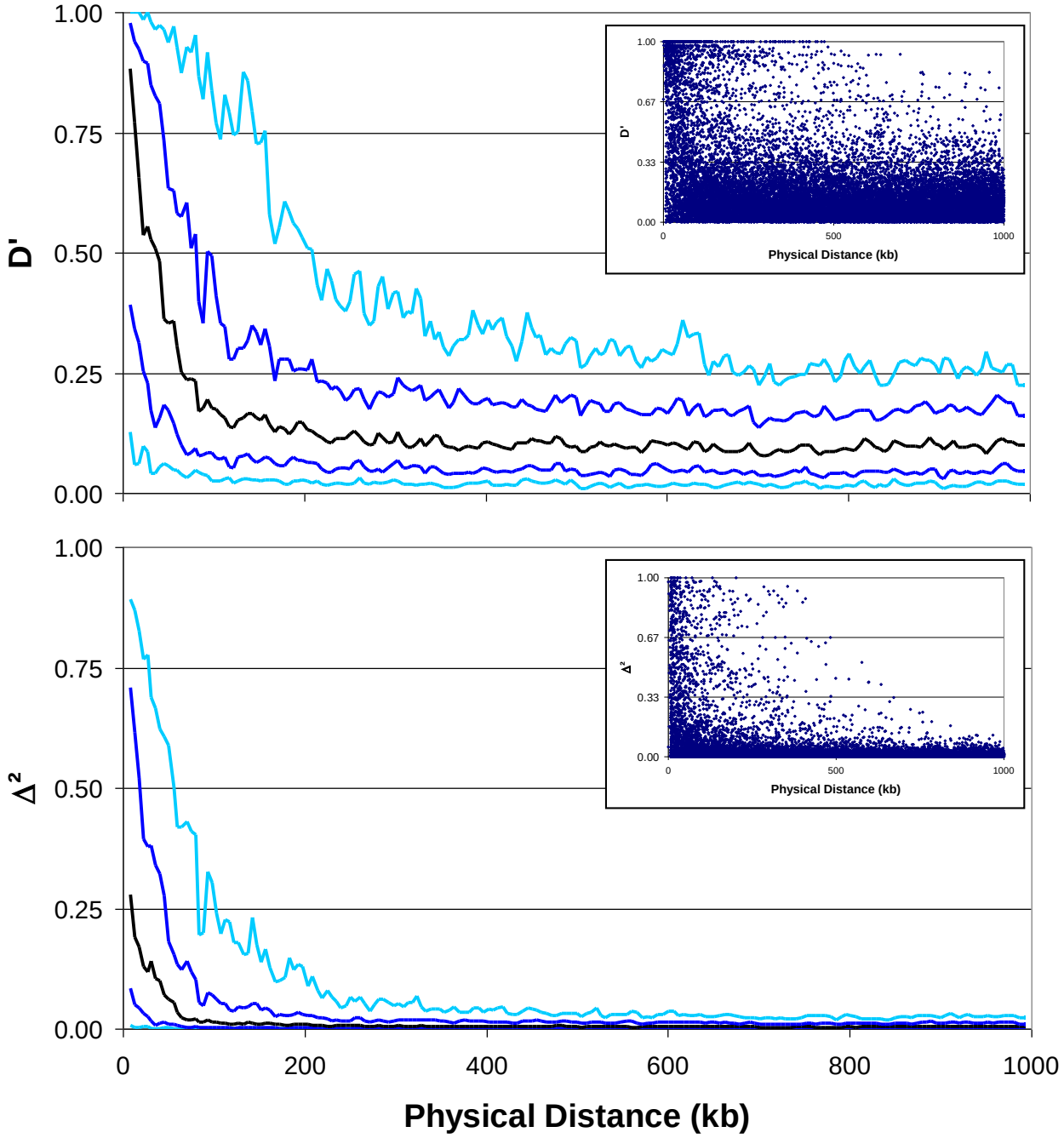
Linkage Disequilibrium

- Chromosomes are mosaics
- Tightly linked markers
  - Alleles not randomly associated
  - Reflect ancestral haplotypes
- Recombination, Mutation, Drift

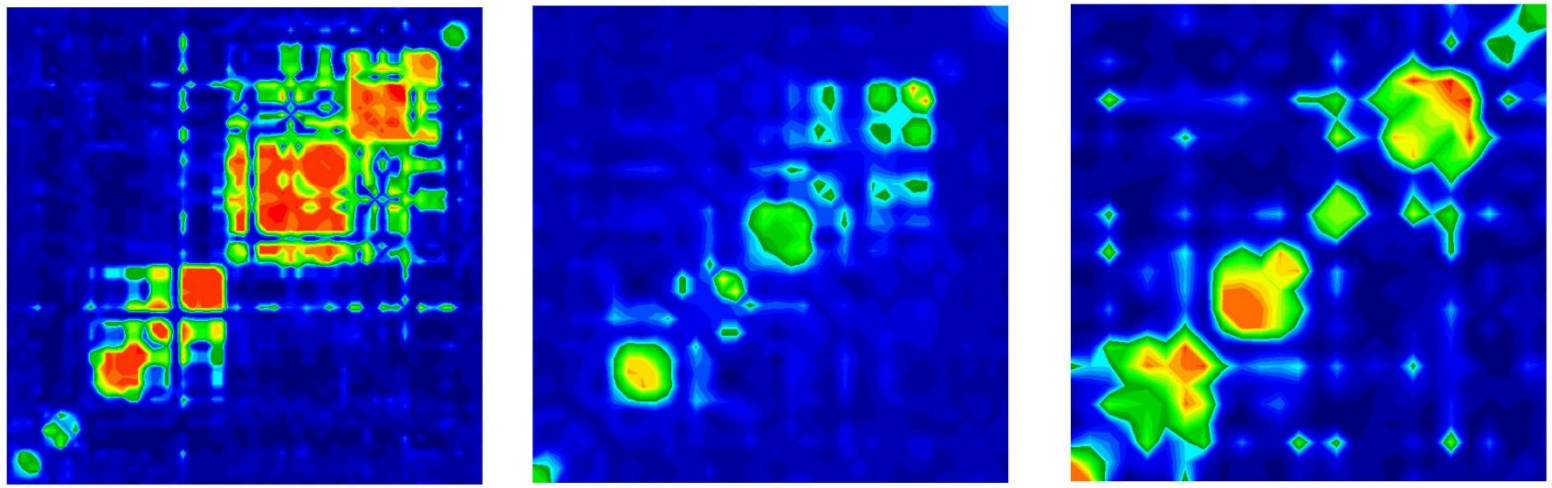


Variability Of Pair-Wise LD

Median
Quartiles
Deciles



GOLD: Graphical Overviews of Linkage Disequilibrium



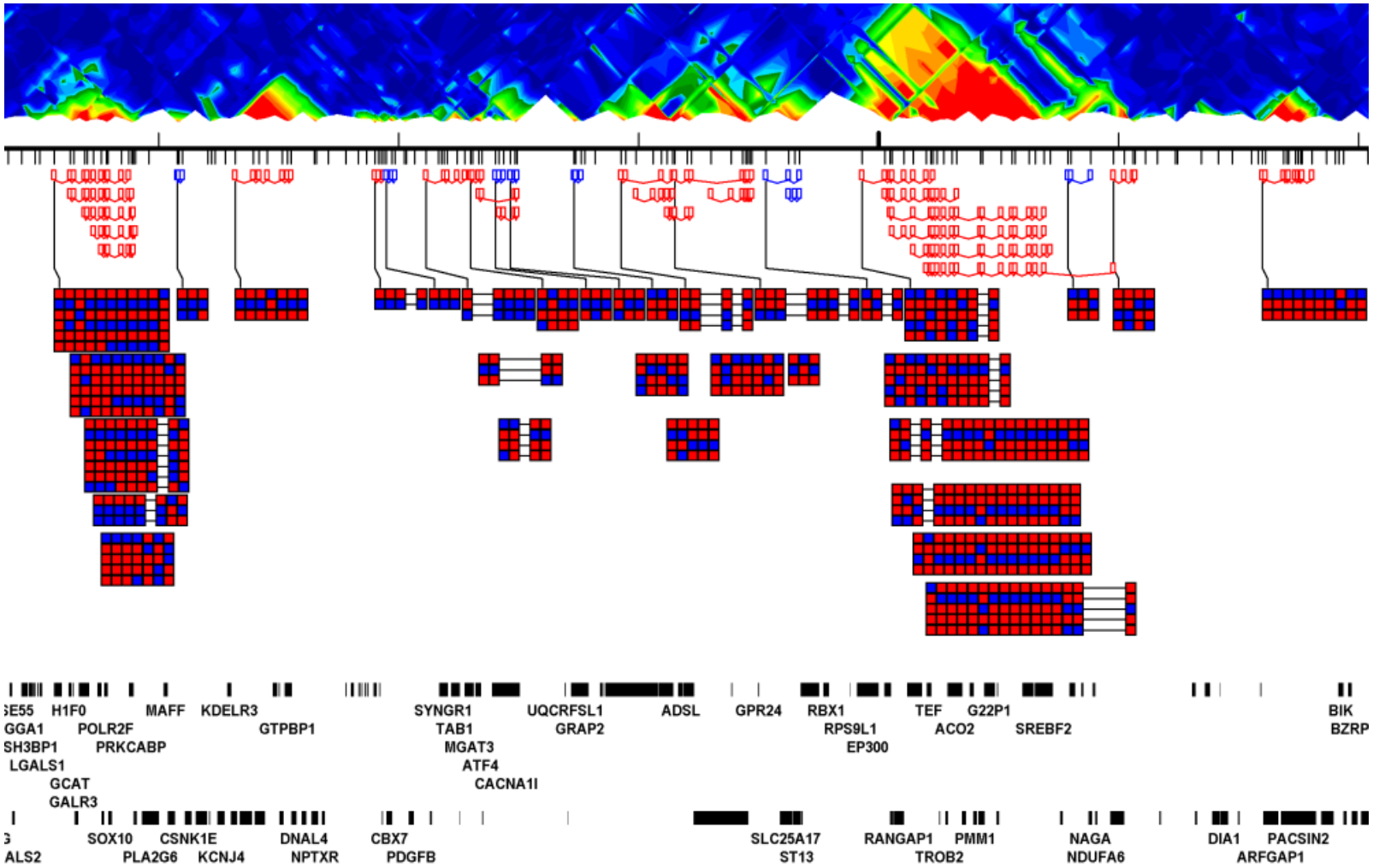
2q13
(63 markers)

13q13
(38 markers)

14q11
(26 markers)

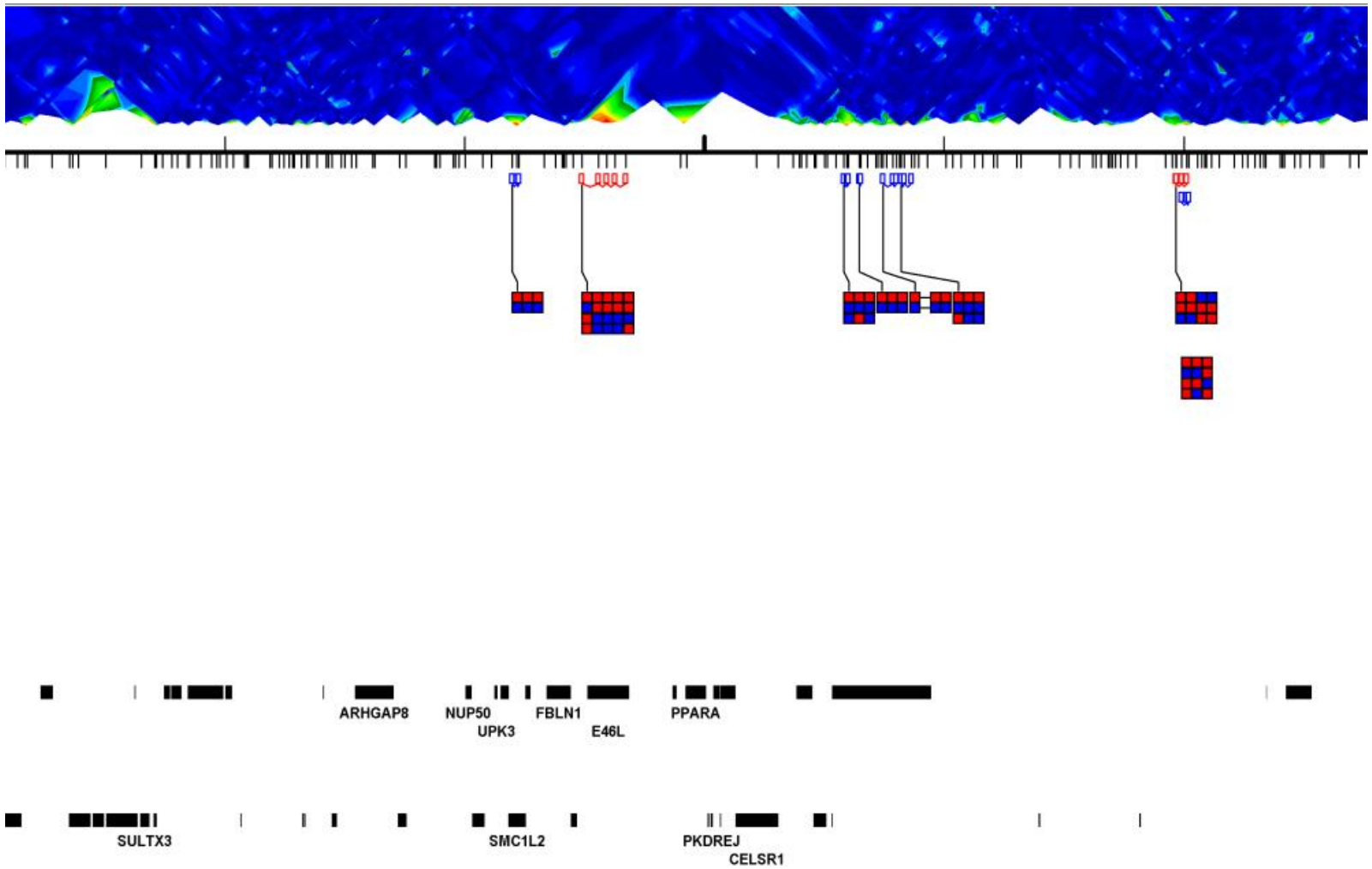
Abecasis et al, *Bioinformatics*, 2001
Abecasis et al, *Am J Hum Genet*, 2001

Chr22 High LD: 22-27 Mb



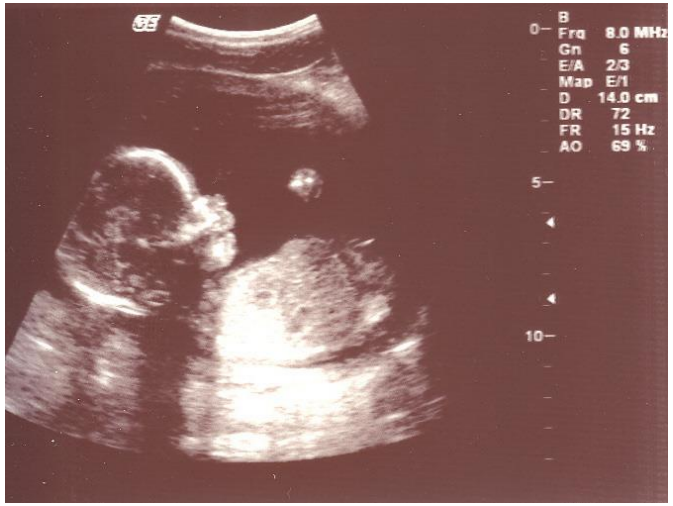
Dawson et al, *Nature*, 2002

Chr22 Low LD: 27-32 Mb

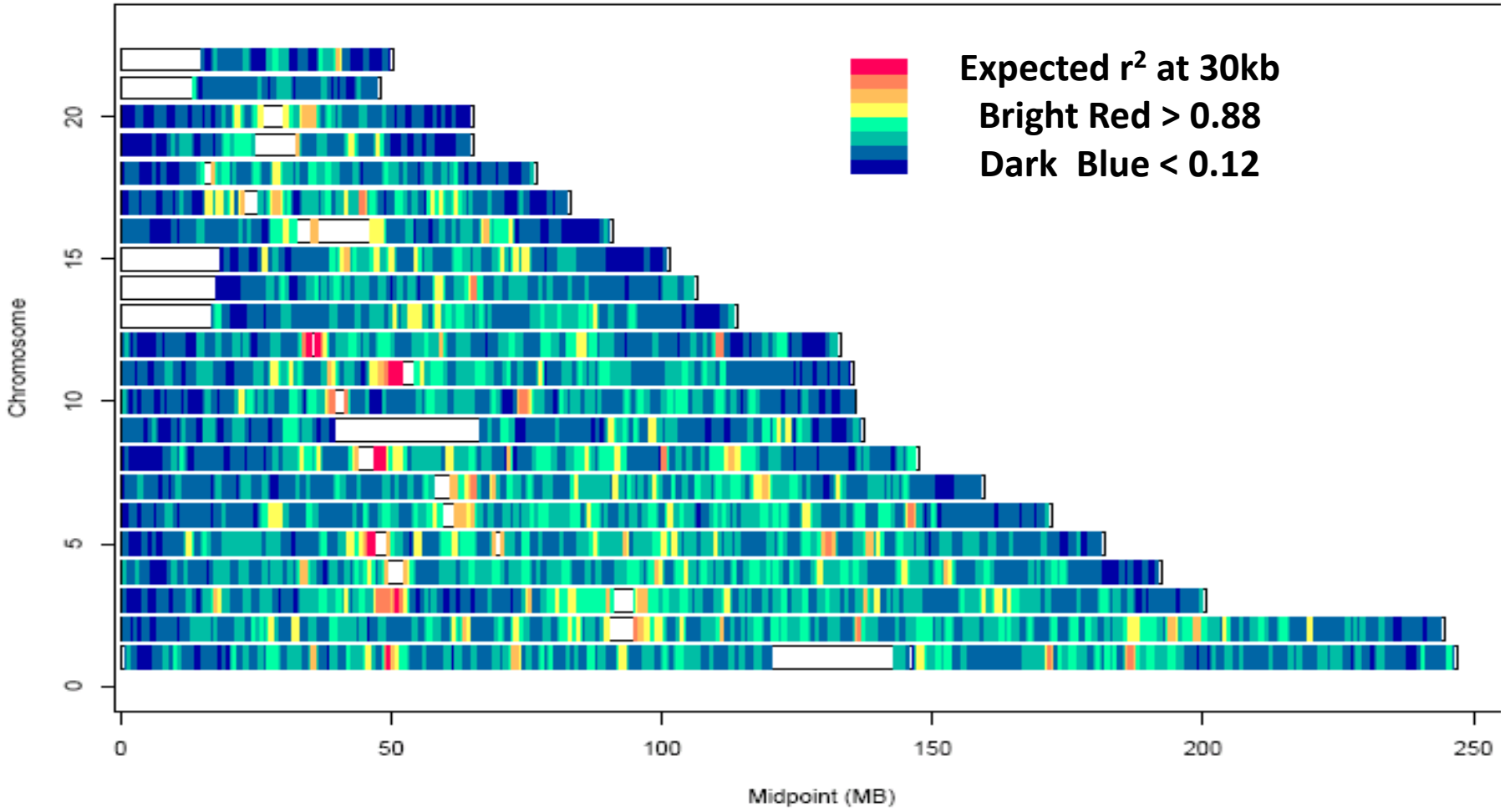


Dawson et al, *Nature*, 2002

2003 - 2005



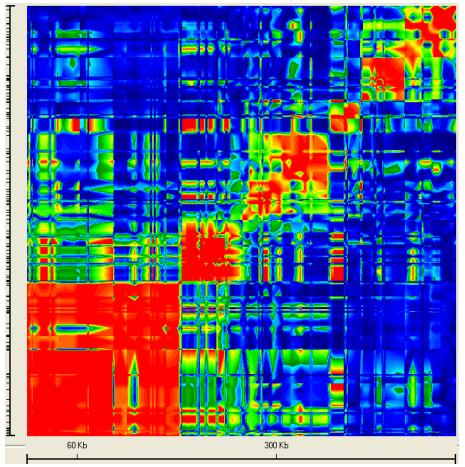
Genomic Variation in Disequilibrium (CEPH)



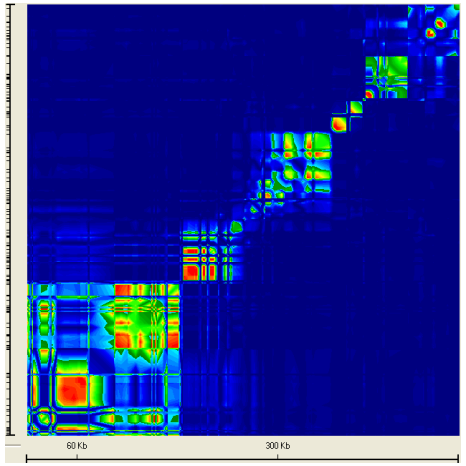
Dense Region 1

- Chromosome 7
  - 157 markers / 520 kb
  - 27.0 – 27.5 Mb
  - Average LD region

- SNP picking (33/157 = 21%)
  - 12 unique SNPs
  - 21 tagging SNPs
  - Others, average $r^2 = 0.73$

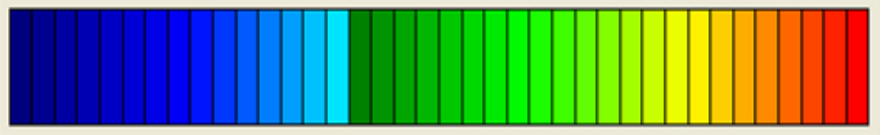


D’



R^2

0.0

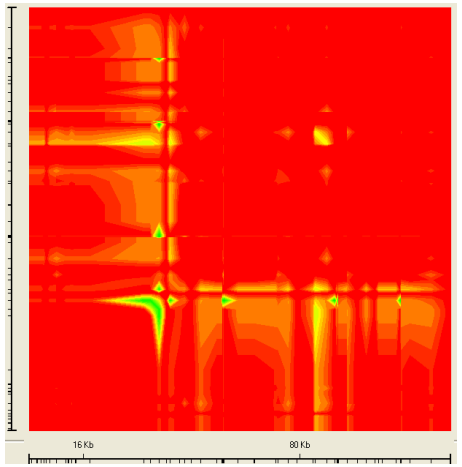


1.0

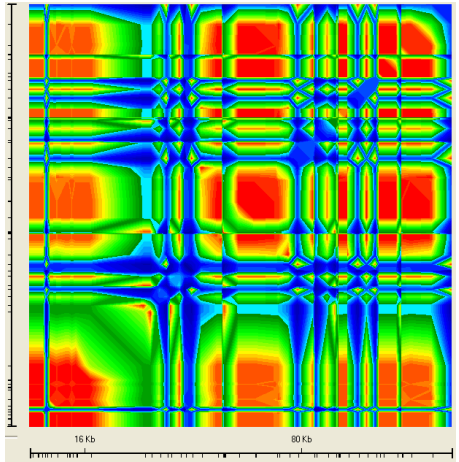
Dense Region 2

- Chromosome 21
  - 57 markers / 130 kb
  - 37.37 – 37.50 Mb
  - High LD region

- SNP picking (8/57 = 14%)
  - 5 unique SNPs
  - 3 tagging SNPs
  - Others, average $r^2 = 0.94$

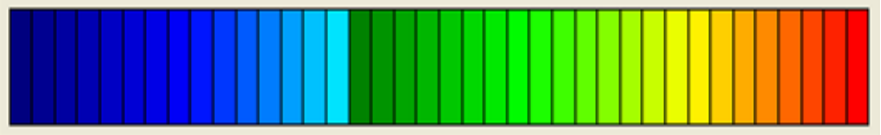


D’



R^2

0.0



1.0

HapMap Analysis Committee

David Altshuler
Aravinda Chakravarti
Peter Donnelly

- Andrew Morris
- Lon Cardon
- David Cutler
- Mark Daly
- Gil McVean
- Bruce Weir
- Simon Myers
- Jonathan Marchini
- Paul de Bakker
- Itsik Pe'er
- Steve Schaffner

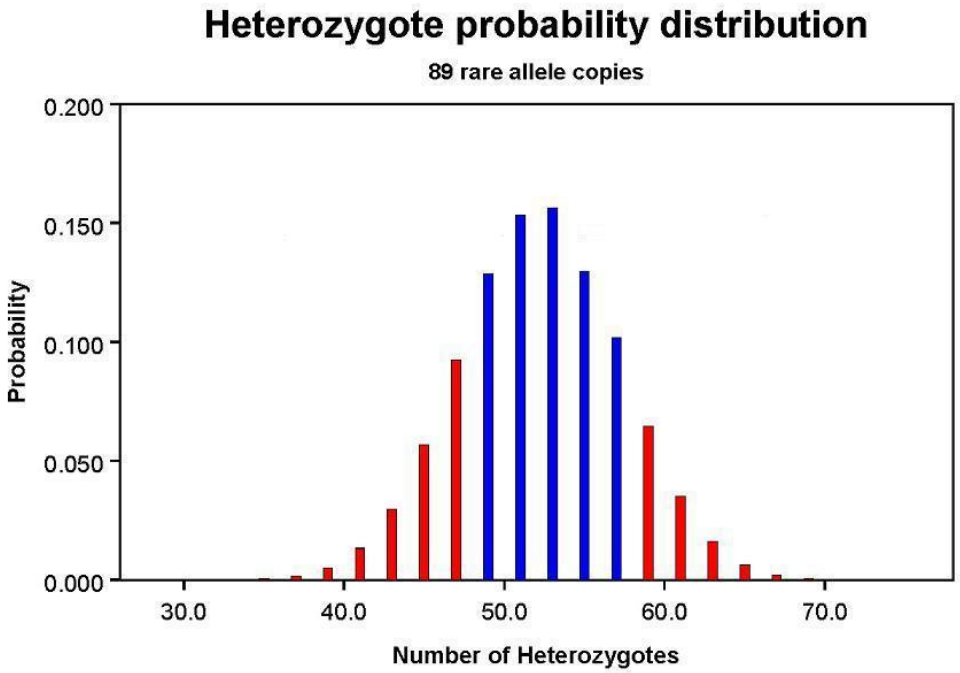
HapMap Analysis Committee… my role!

- My main assigned role in the HapMap project was to…
  - Evaluate quality control metrics for generated data

- This required lots of political finagling…

- And some interesting exact algorithms for rapidly evaluating the likelihood of a particular genotype configuration…

Wigginton et al (2005)



An accident along the way!…

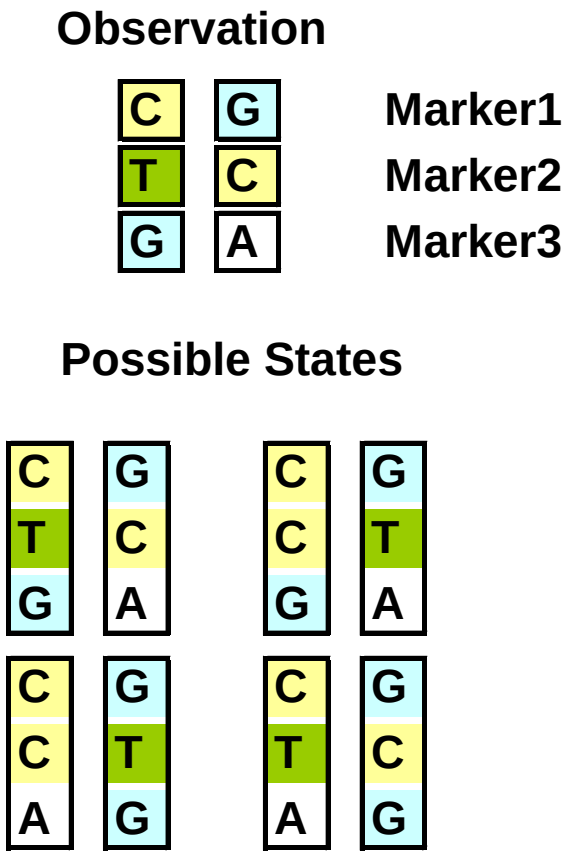
- Our early linkage disequilibrium studies typically focused on small families, where it was computationally simple to estimate haplotypes

- However, due to an mistake in tracking meta-data at CEPH and Coriell, we genotyped an interconnected 24-member superfamily…

- … analyzing a few dozen SNPs in this sort of pedigree was beyond the capabilities of analytical methods at the time.

Typical Genotype Data

- Two alleles for each individual
  - Unknown Phase
- Maternal and paternal origin unknown
- Genetic markers provide imperfect information on gene flow



The Haplotyping Problem in Family Data

- For each person
  - 2 meioses, each with 2 possible outcomes
  - $2n$ meioses in pedigree with n non-founders
- For each genetic locus
  - One location for each of m genetic markers
  - Distinct, non-independent meiotic outcomes
- Up to 4^{nm} distinct outcomes

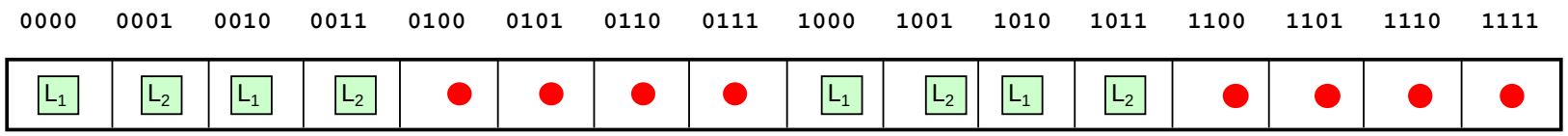
- $O(4^{mn})$ with a naïve solution

MERLIN

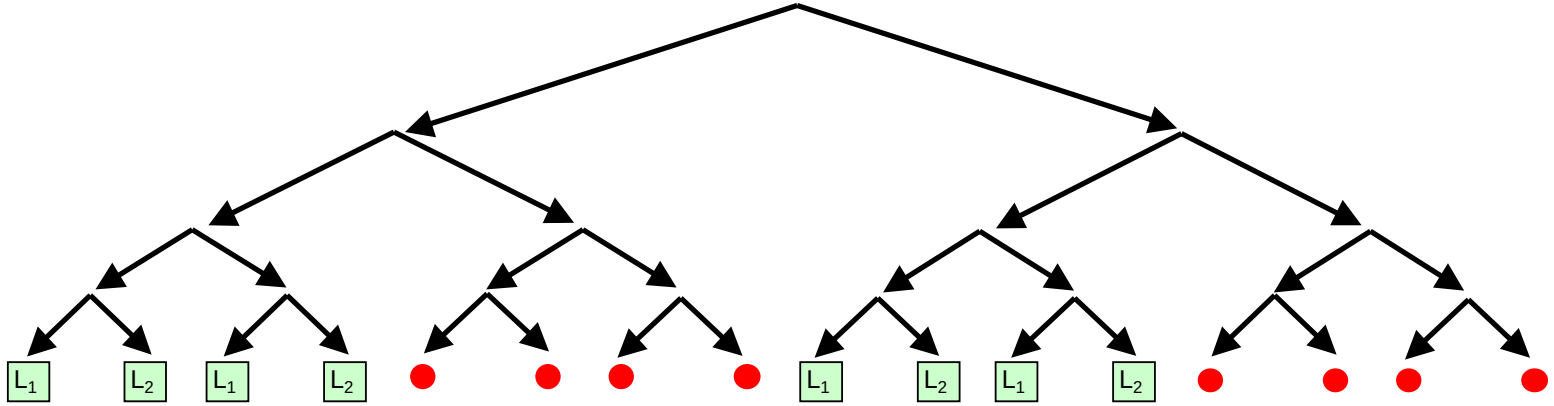
Multipoint Engine for Rapid Likelihood Inference

- Linkage analysis
- Haplotyping
- Error detection
- Simulation

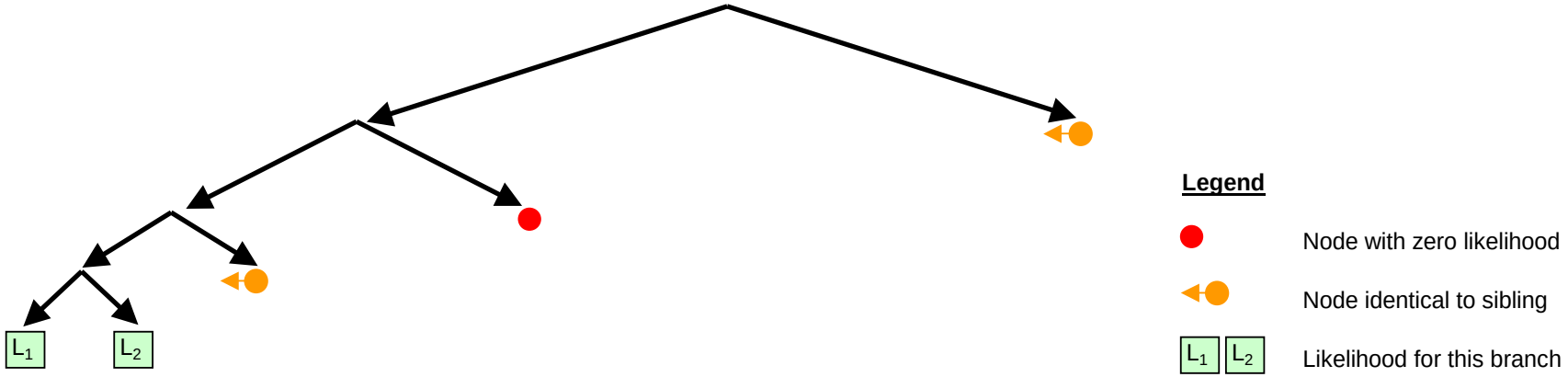
a) **bit-indexed array**



b) **packed tree**



c) **sparse tree**

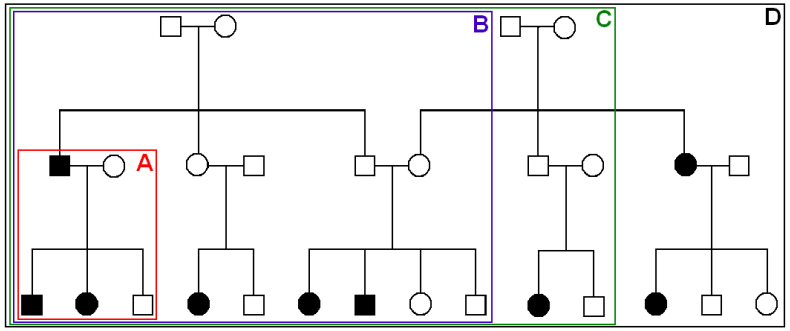


Tree Complexity: 28 person pedigree

Missing Genotypes	Info	Total Nodes			Leaf Nodes
		Mean	Median	95% C.I.	
2-allele marker with equifrequent alleles					
-	0.42	706.0	151	57 – 5447	66.9
5%	0.39	1299.8	225	57 – 8443	159.6
10%	0.36	2157.7	329	61 – 15361	148.9
20%	0.31	8595.9	872	64 – 42592	1293.9
50%	0.14	55639.1	4477	135 – 383407	9173.5

(Simulated pedigree with 28 individuals, 40 meioses, requiring 2^{32} = ~4 billion likelihood evaluations using conventional schemes)

Merlin is fast…



	Time	Memory
Exact	40s	100 MB
No recombination	<1s	4 MB
≤1 recombinant	2s	17 MB
≤2 recombinants	15s	54 MB
Genehunter 2.1	16min	1024MB

Keavney et al (1998) ACE data, 10 SNPs within gene,
4-18 individuals per family

My Research Team (2006)

- 4 students (MS and PhD)
- 3 postdocs
- 1 programmer

- Collaborators
  - Mike Boehnke, Noah Rosenberg, Laura Scott, Steve Qin (Biostatistics)
  - Other collaborators at the Medical School, Kellogg Eye Center, Rockefeller University and National Institute on Aging (NIH)

The First Genomewide Association Studies

Joint Analysis

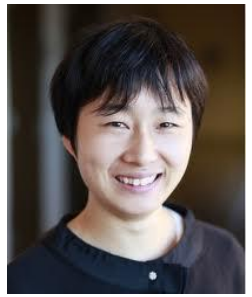
Imputation

More Imputation



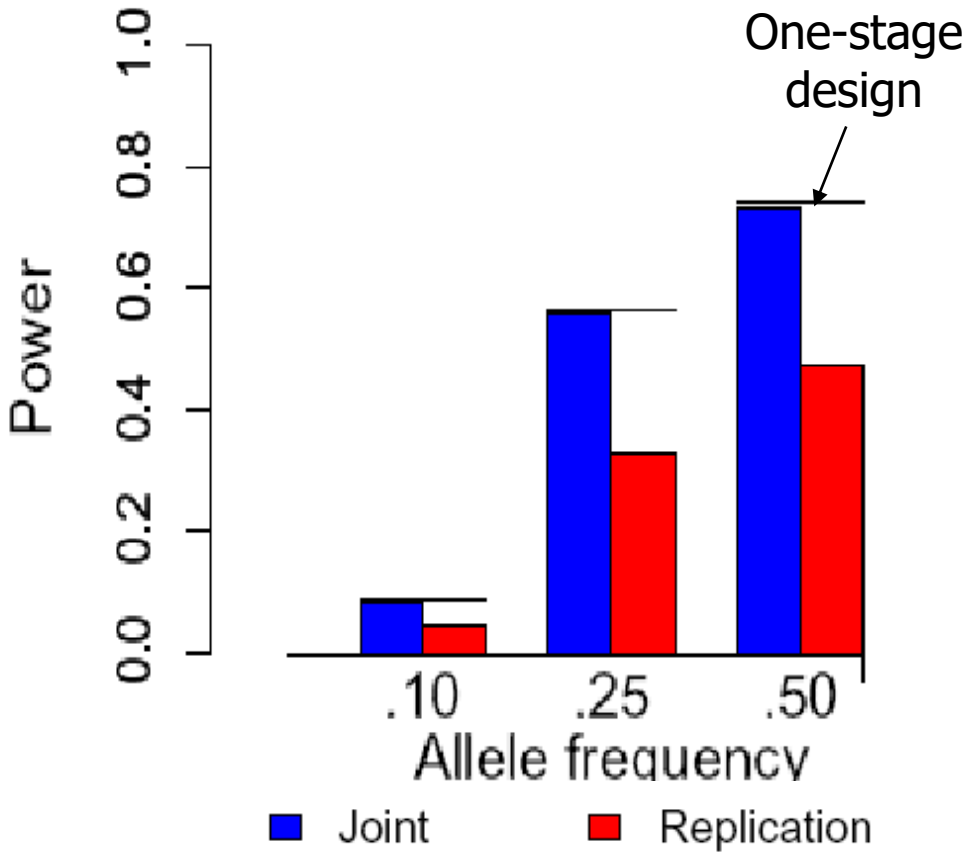




Joint Analysis far outperforms Replication

50% of samples in discovery sample, 1% of markers in follow up



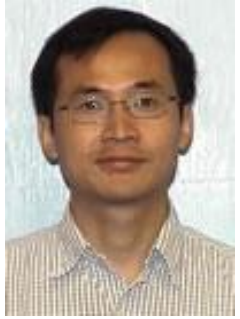
- With the HapMap catalog, …
- Improved genotyping arrays…
- Genomewide association studies became possible…
- … my experience with QC of HapMap data proved timely!
- Started to explore issues related to study design in Skol et al (Nature Genetics, 2006).

Incorporating Family Information in Genome Wide Studies

- Family members will share large segments of chromosomes

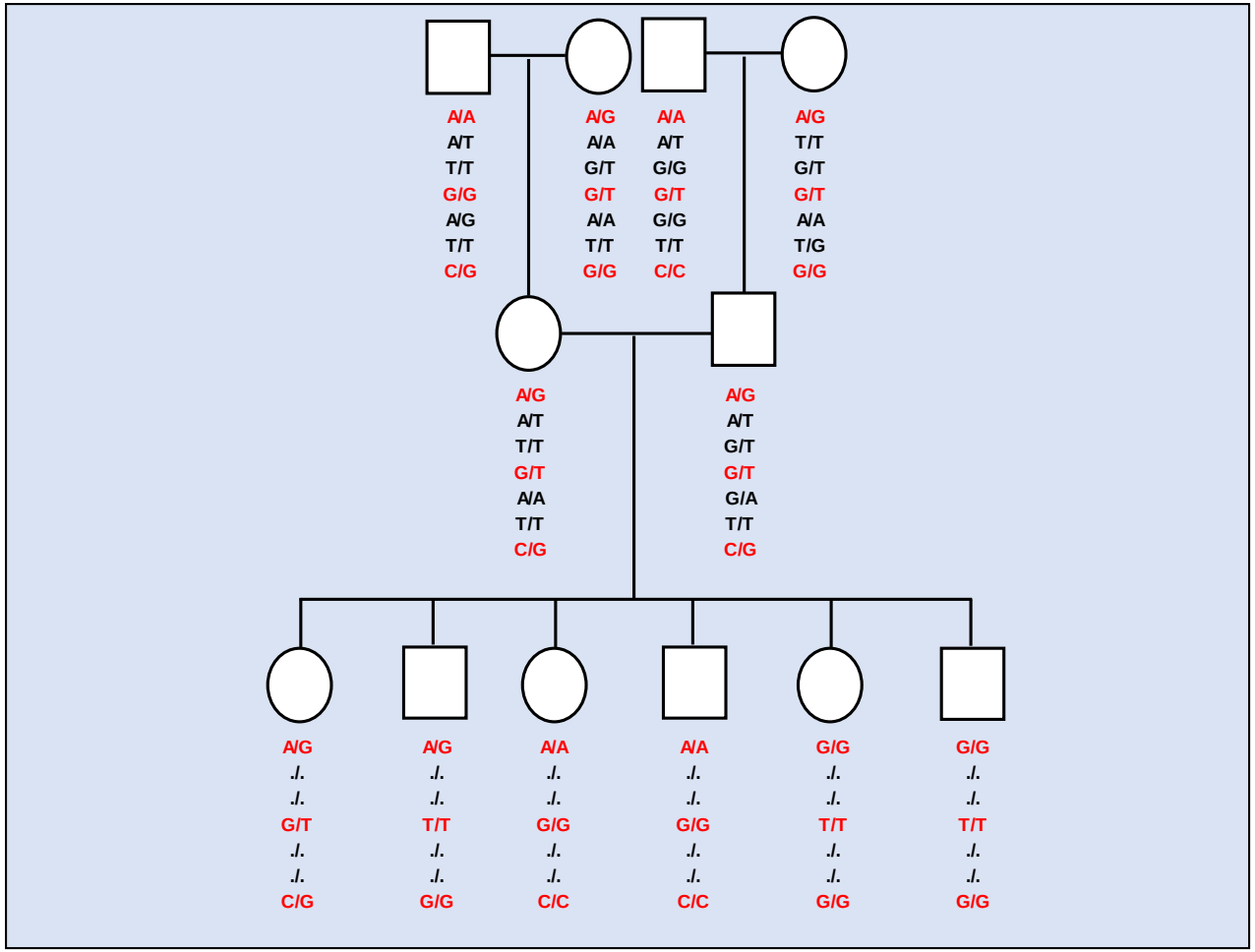
- If we genotype many related individuals, we will effectively be genotyping a few chromosomes many times

- In fact, we can:
  - genotype a few markers on all individuals
  - use high-density panel to genotype a few individuals
  - infer shared segments and then estimate the missing genotypes

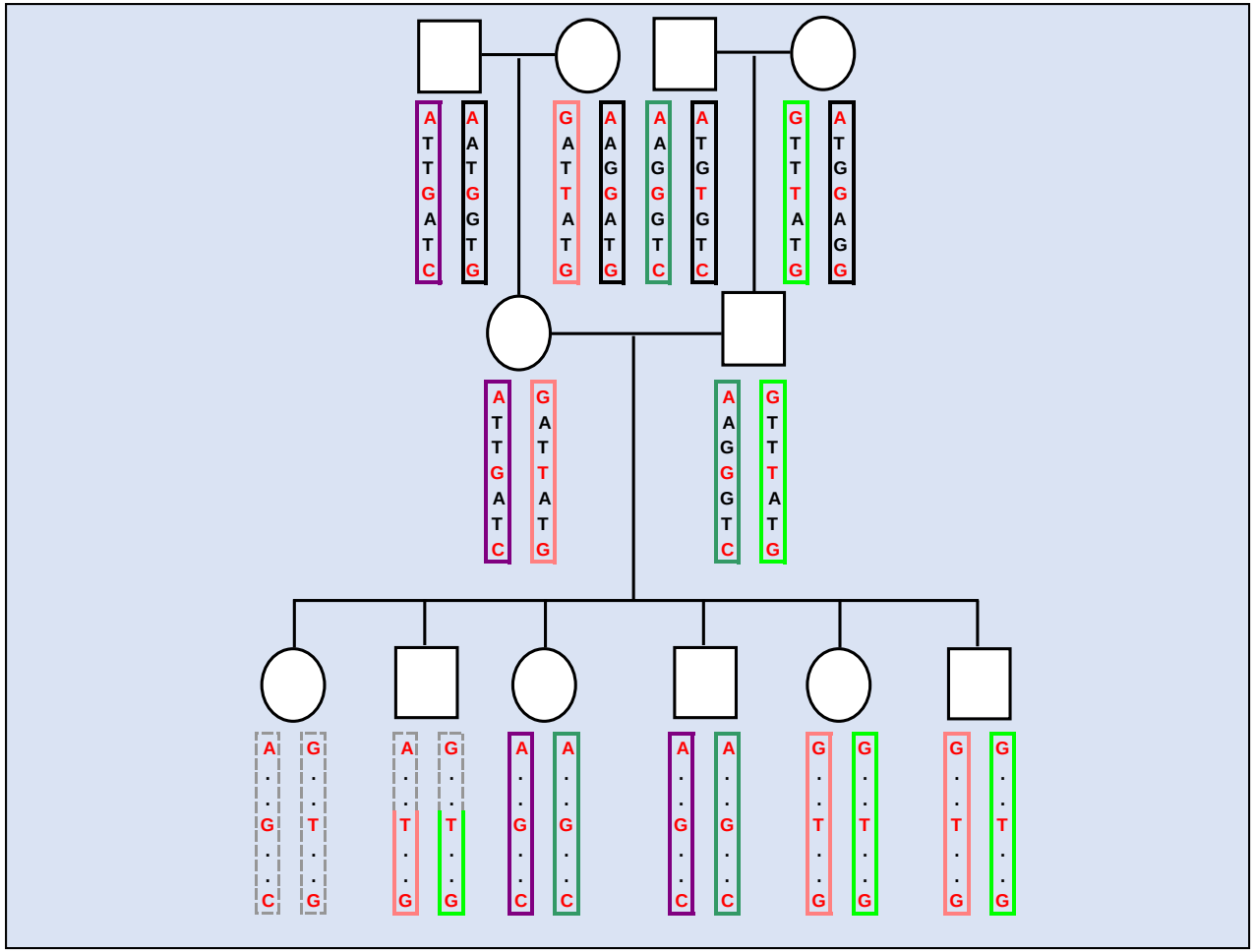


Burdick et al, Nat Genet, 2006
Chen et al, Am J Hum Genet, 2007

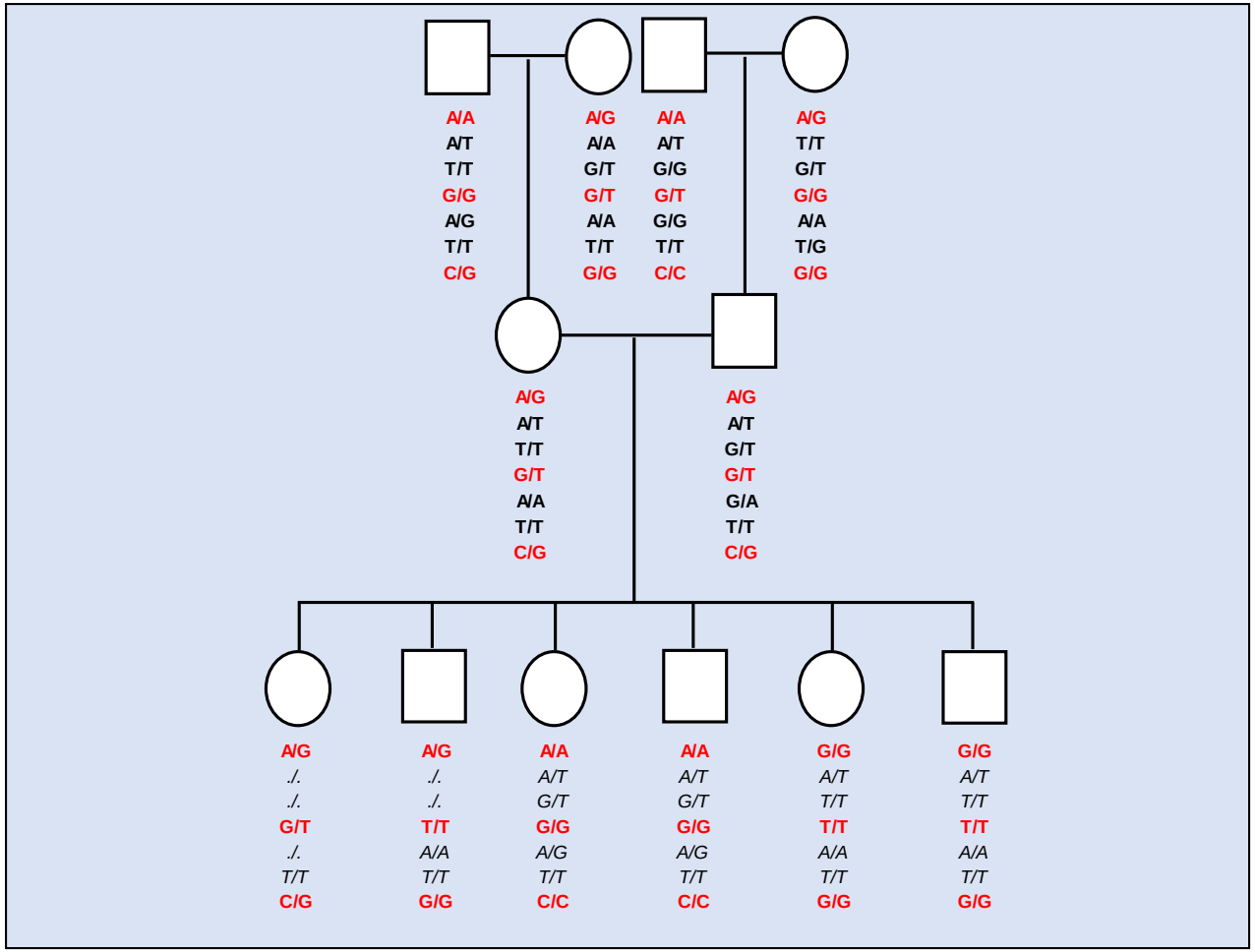
Genotype Inference
Part 1 – Observed Genotype Data



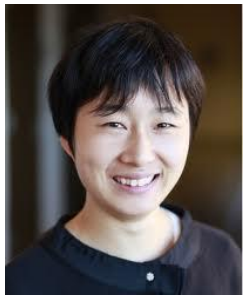
Genotype Inference
Part 2 – Inferring Allele Sharing



Genotype Inference
Part 3 – Imputing Missing Genotypes



In Silico Genotyping For Unrelated Individuals




- In families, long stretches of shared chromosome
- In unrelated individuals, shared stretches are much shorter
- The plan is still to identify stretches of shared chromosome between individuals…
- … we then infer intervening genotypes by contrasting samples typed at a few sites with those with denser genotypes

Scott et al, Science, 2007
Li et al, Annual Review of Genetics and Human Genomics, 2009
Li et al, Gen Epid, 2010

1. Imputation setting

Observed GWAS Genotypes

```
. . . . A . . . . . . . A . . . . A . . .
. . . . G . . . . . . . C . . . . A . . .
```

Reference Haplotypes (e.g. 1000G)

```
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
```

2. Identify match among reference

Observed GWAS Genotypes

```
. . . . A . . . . . . . A . . . . A . . .
. . . . G . . . . . . . C . . . . A . . .
```

Reference Haplotypes (e.g. 1000G)

```
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
```

3. Impute

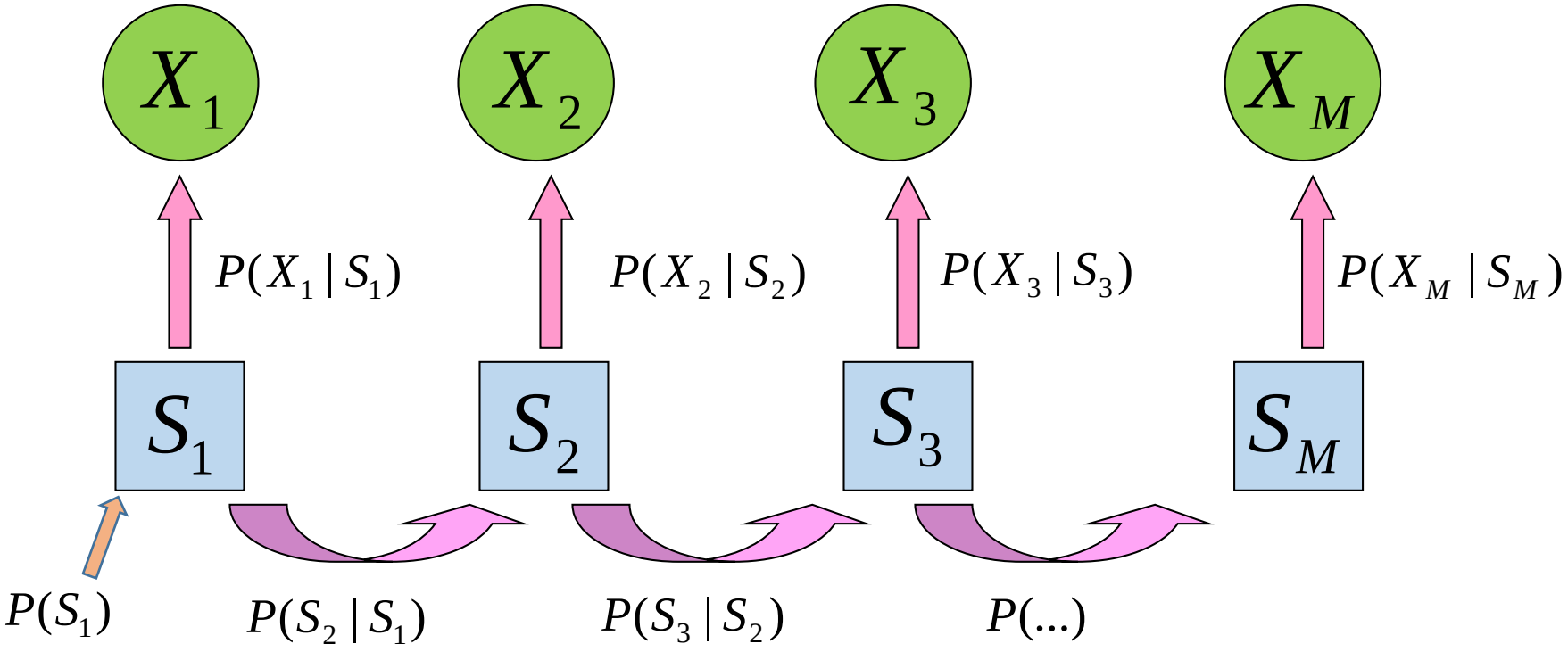
Observed GWAS Genotypes

c	g	a	g	A	t	c	t	c	c	c	g	A	c	c	t	c	A	t	g	g
c	g	a	a	G	c	t	c	t	t	t	t	C	t	t	t	c	A	t	g	g

Reference Haplotypes (e.g. 1000G)

C	G	A	G	A	T	C	T	C	C	T	T	C	T	T	C	T	G	T	G	C
C	G	A	G	A	T	C	T	C	C	C	G	A	C	C	T	C	A	T	G	G
C	C	A	A	G	C	T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
C	G	A	A	G	C	T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
C	G	A	G	A	C	T	C	T	C	C	G	A	C	C	T	T	A	T	G	C
T	G	G	G	A	T	C	T	C	C	C	G	A	C	C	T	C	A	T	G	G
C	G	A	G	A	T	C	T	C	C	C	G	A	C	C	T	T	G	T	G	C
C	G	A	G	A	C	T	C	T	T	T	T	C	T	T	T	T	G	T	A	C
C	G	A	G	A	C	T	C	T	C	C	G	A	C	C	T	C	G	T	G	C
C	G	A	A	G	C	T	C	T	T	T	T	C	T	T	C	T	G	T	G	C

Markov Model



Number of states to be considered increases exponentially with panel size …

Does This Really Work?

- Used about ~300,000 SNPs from Illumina HumanHap300 to impute 2.1M HapMap SNPs in 2500 individuals from a study of type II diabetes

- Compared imputed genotypes with actual experimental genotypes in a candidate region on chromosome 14
  - 1190 individuals, 521 markers not on Illumina chip

- Errors are concentrated on a few markers
  - 14.8% error for 1% of SNPs with the worst predicted imputation quality
  - 11.1% error for next 1% of SNPs (1st – 2nd percentile)
  - 5.9% error for next 1% of SNPs (2nd – 3rd percentile)
  - 1.1% error for top 95% of SNPs

Scott et al, *Science*, 2007

Impact of HapMap Imputation on Power

	Power	
Disease SNP MAF	**tagSNPs**	**Imputation**
2.5%	24.4%	56.2%
5%	55.8%	73.8%
10%	77.4%	87.2%
20%	85.6%	92.0%
50%	93.0%	96.0%

Power for Simulated Case Control Studies.
Simulations Ensure Equal Power for Directly Genotype SNPs.

Simulated studies used a tag SNP panel that captures
80% of common variants with pairwise $r^2 > 0.80$.

Can we do even better?

- Ask a better statistician?

- Collect more data?
  - 60 individuals in reference, 1.78% error rate per allele
  - 100 individuals in reference, 1.03% error rate
  - 200 individuals in reference, 0.78% error rate
  - 500 individuals in reference, 0.41% error rate

  - Maybe we could use a larger HapMap?

How long does it take to impute one genome?

- Depends on the reference size …

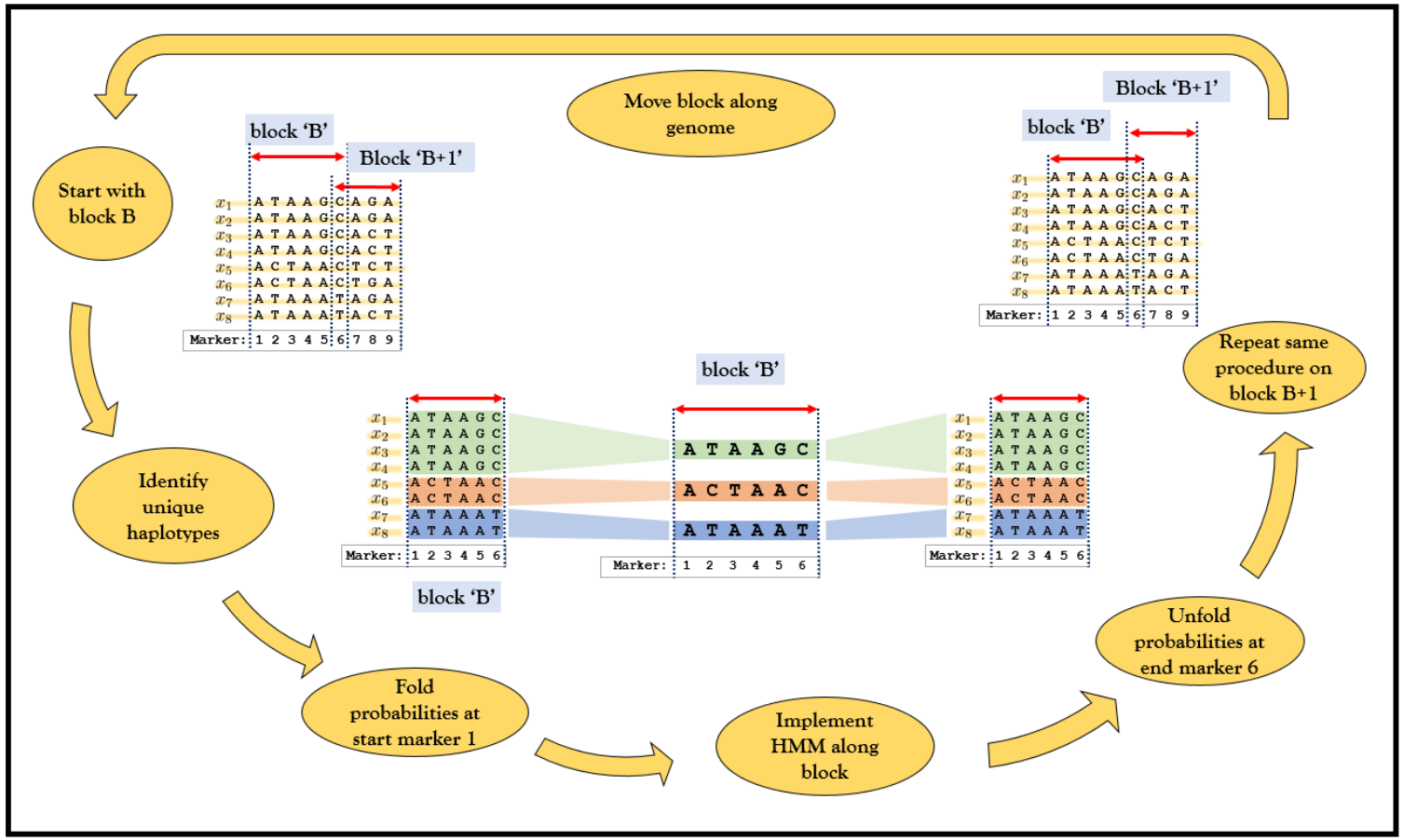
• 2007:	60 samples, 2.5M SNPs	14 min
• 2009:	60 samples, 7.3M SNPs	41 min
• 2011:	283 samples, 11.6M SNPs	1,287 min
• 2012:	381 samples, 37.4M SNPs	7,800 min
• 2015:	33,000 samples, 37M SNPs	63,000,000 min (estimated)

How long does it take to impute one genome?

- Depends on the computational methods …

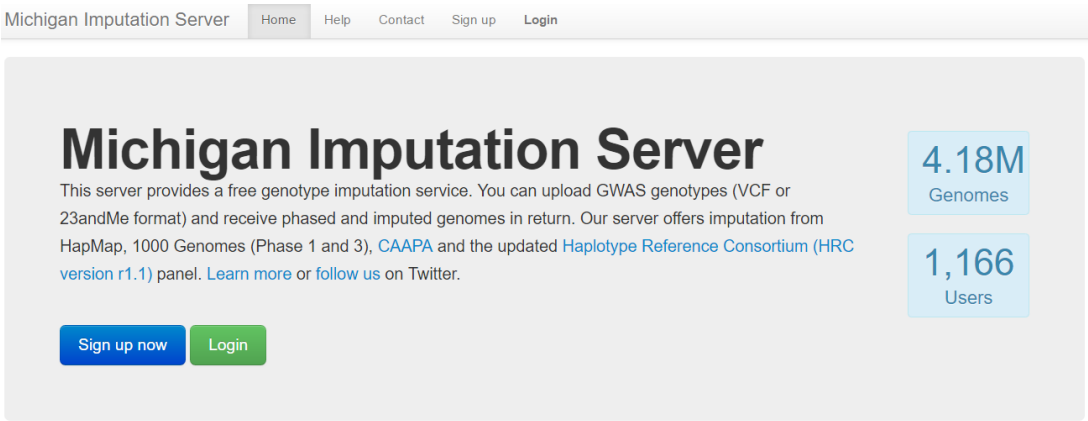
• 2007 software:	381 samples, 37.4M SNPs	7,800 min	$O(MH^2)$
• 2010 software:	381 samples, 37.4M SNPs	512 min	$O(MH)$
• 2012 software:	381 samples, 37.4M SNPs	24 min	$O(MH)$
• 2015 software:	381 samples, 37.4M SNPs	1 min	$<O(MH)$
• 2016 software:	381 samples, 37.4M SNPs	<5 secs	$<O(MH)$

Most Recent Imputation Improvements Minimac3



Imputation Servers

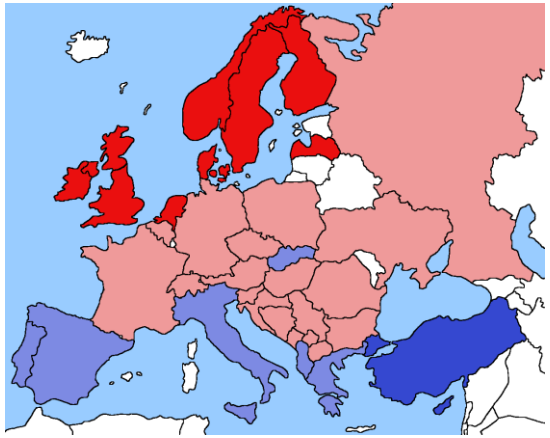
https://imputationserver.sph.umich.edu



Upload your genotypes to our server located in Michigan. All interactions with the server are **secured**.

Choose a reference panel. We will take care of pre-phasing and imputation.

Download the results. All results are encrypted with a one-time password. After 7 days, all results are deleted from our server.



Studies of Lipid Genetics (2006-)

Global Lipids Genetics Consortium

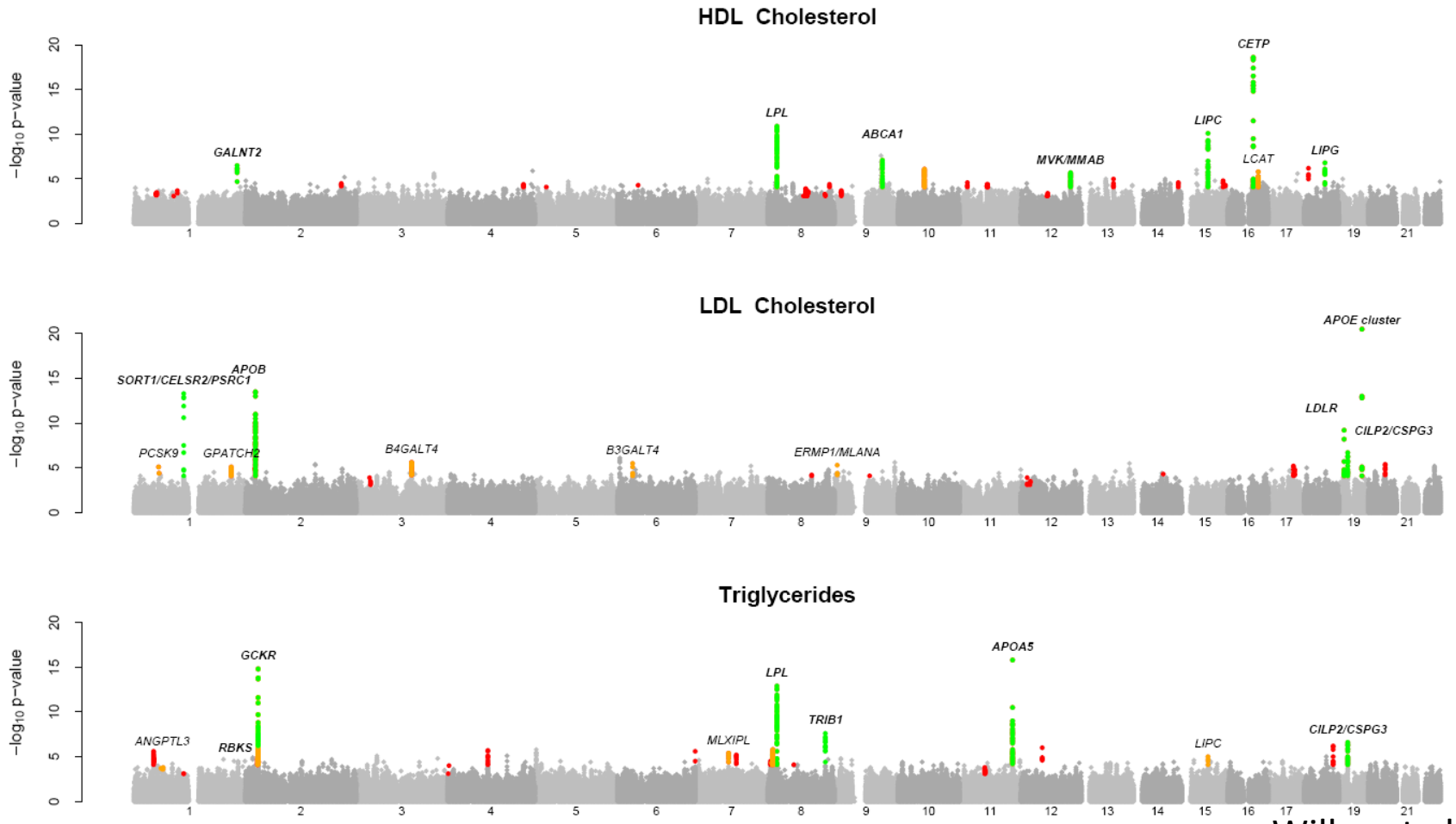
Sekar Kathiresan

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

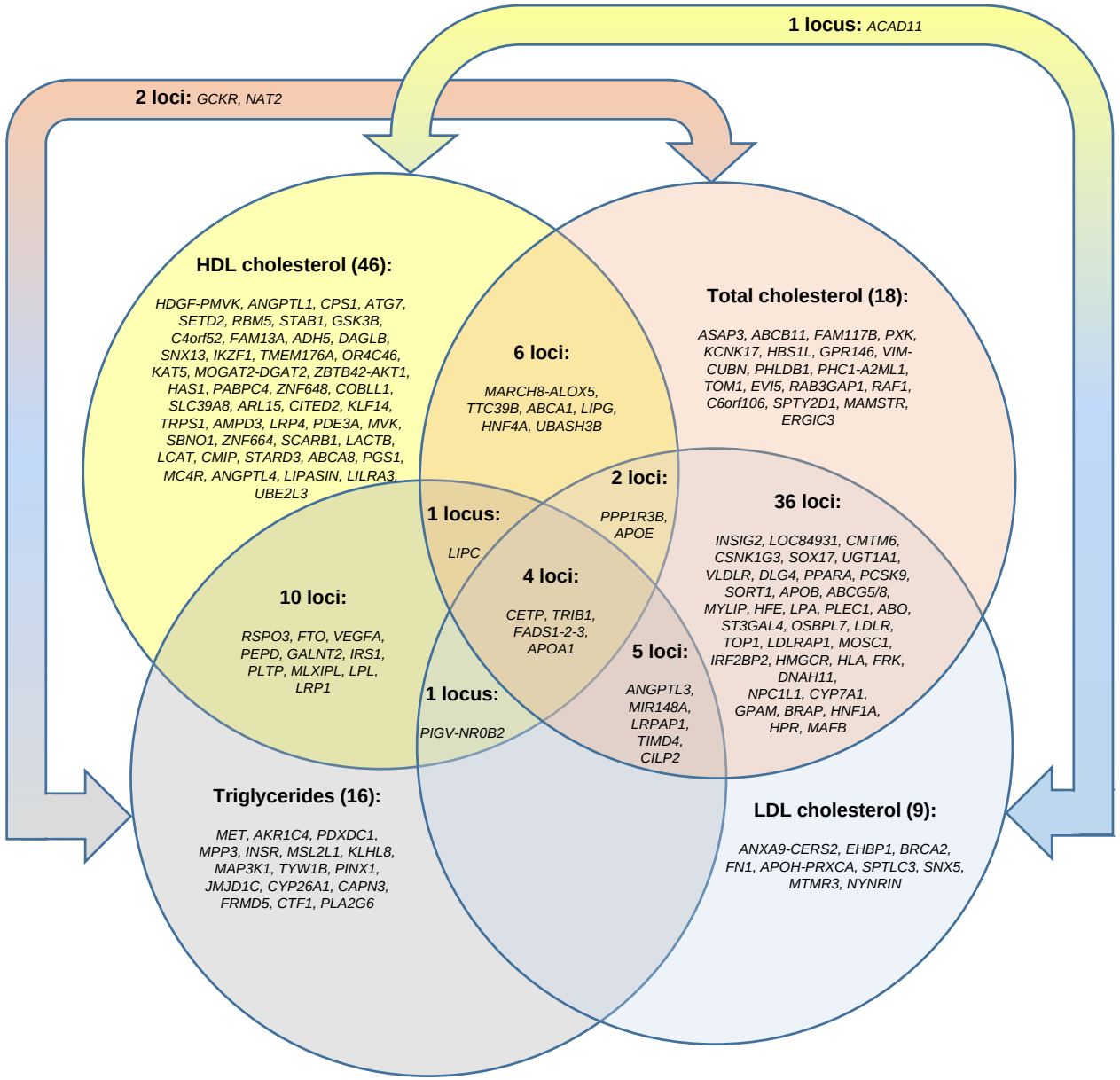
Willer et al, Nat Genet, 2008; Teslovich et al, Nature, 2010; Willer et al, in press

First Meta-Analysis Using Imputation… Seventeen Hits by Combining 3 Almost "Null" Studies

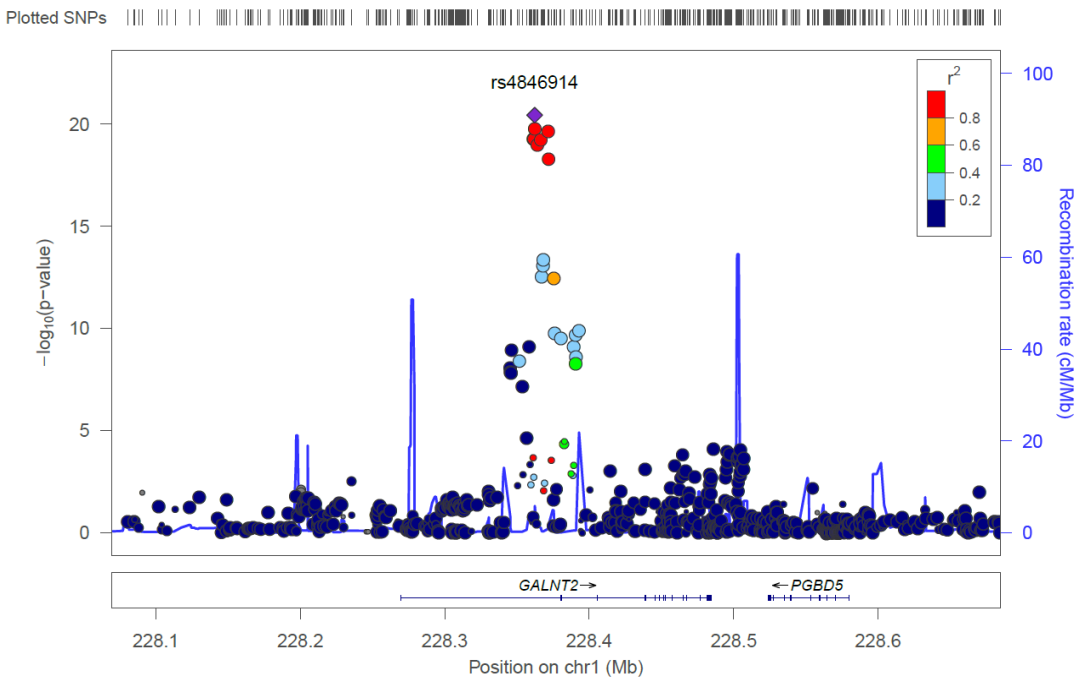


Willer et al, Nat Genet, 2008
Willer et al, Bioinformatics, 2010
Pruim et al, Bioinformatics, 2010

A SNAPSHOT OF LIPID GENETICS



Suggesting New Targets: GALNT2

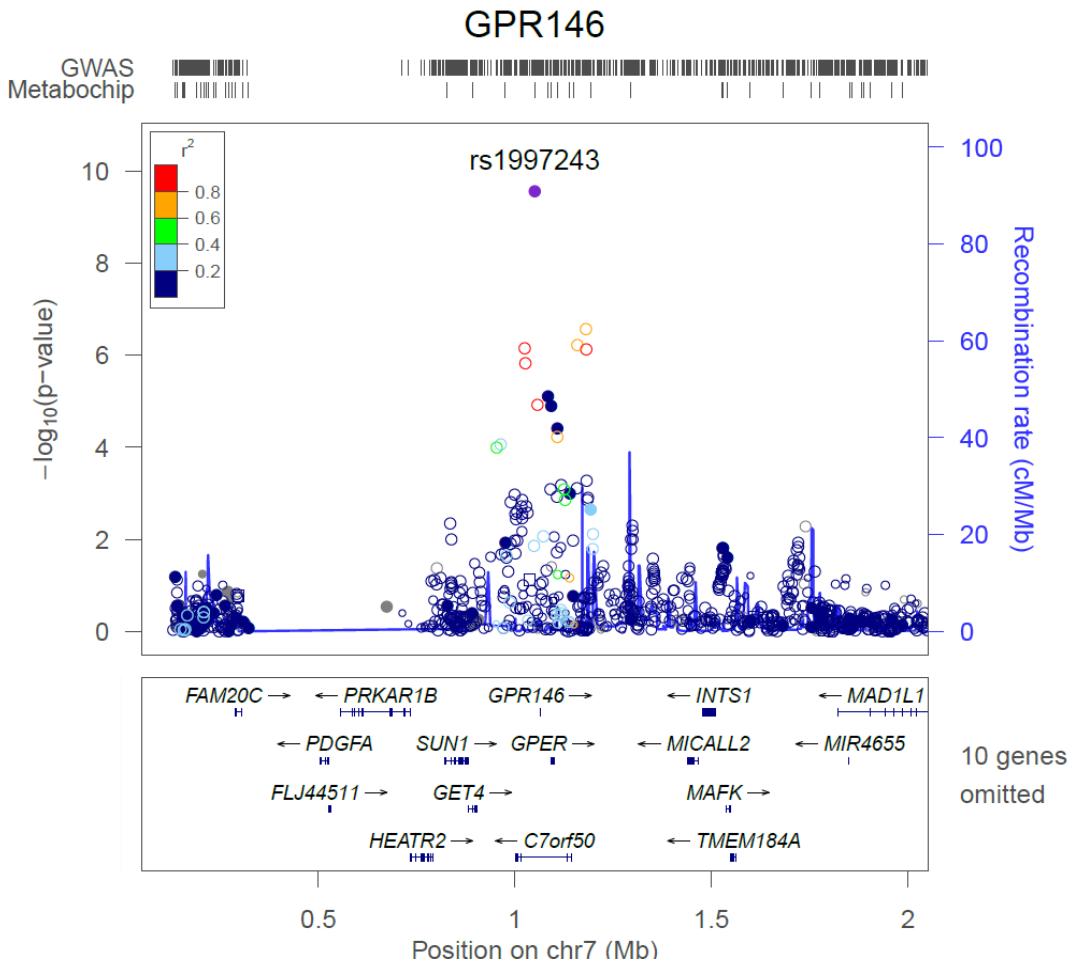


Dan Rader

- 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%

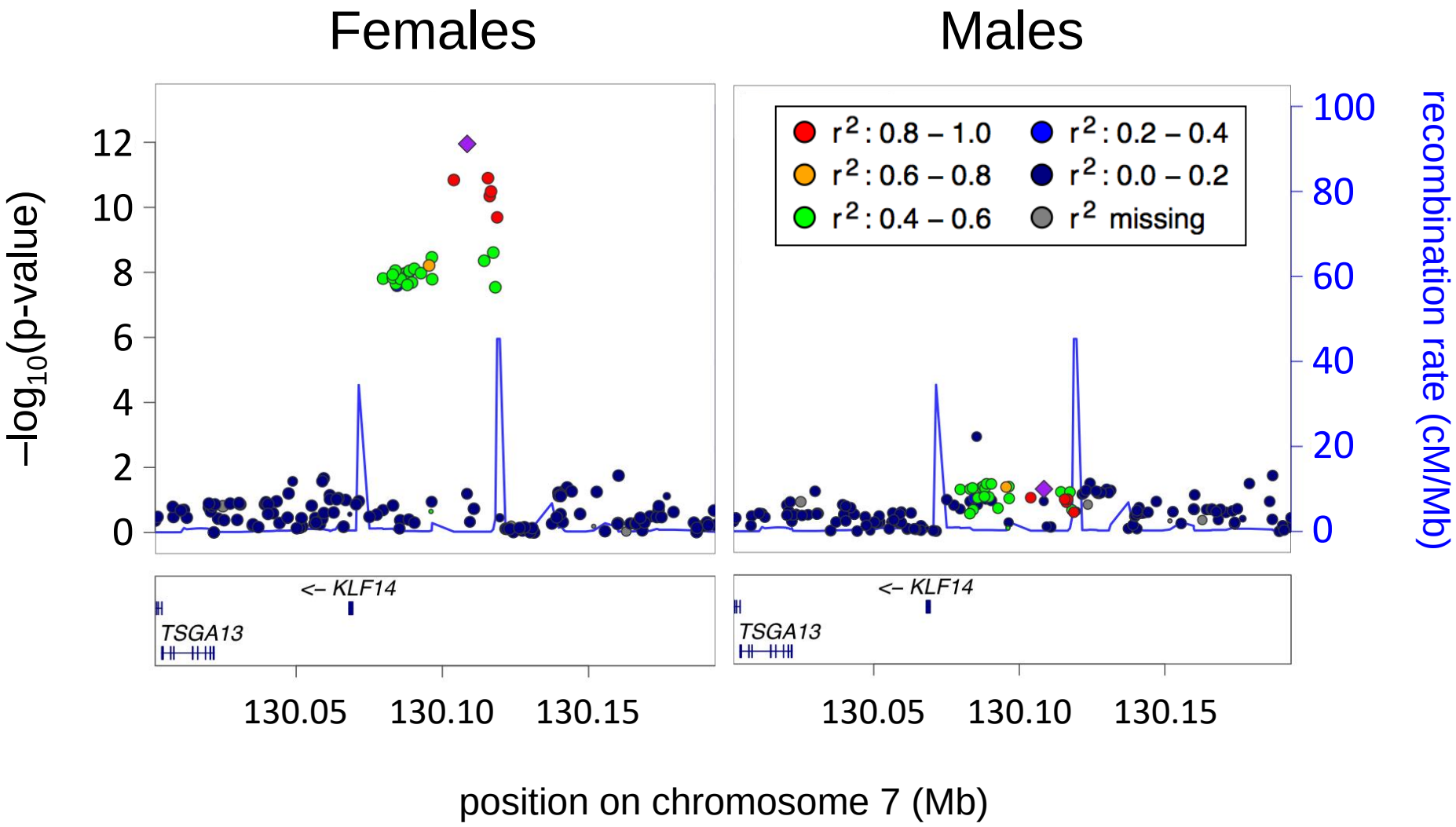
Teslovich et al, Nature, 2012

Supporting Previous Leads: GPR146

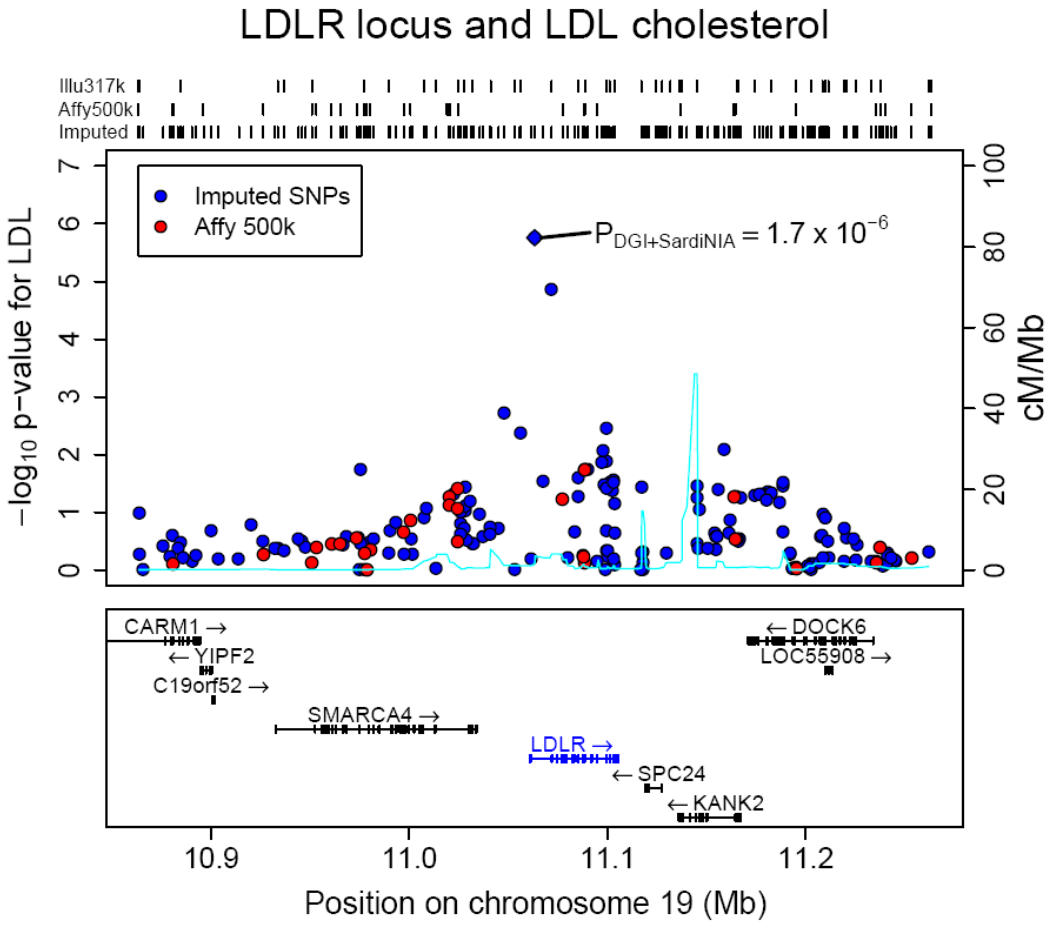


- Our work shows that variants near GPR146 are associated with total cholesterol
- U. S. Patent Application #20,090,036,394 discloses that, in mice, targeting GPR146 lowers cholesterol
- Together, the two pieces of evidence could encourage human trials

Triglyceride association: *KLF14*
Sex-specific effect



Imputation Helps
LDLR and LDL example



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 also suggests a causal role for LDL-C associated alleles (but not for HDL-C)

Current State of GWAS

- 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

A Key Goal of Sequence Based Association Studies

UNDERSTAND FUNCTION

LINKING EACH LOCUS TO DISEASE

What happens in gene knockouts?

- Use sequencing to find rare human “knockout” alleles
- Why? Results of animal studies an *in vitro* studies often murky
- The challenge? Natural knockouts are extremely rare

How Can We Cost Effectively Sequence 1,000s of Genomes?

Whole Genome Sequencing (2009-)

How Do Sequence Reads Get Transformed Into Genotypes?

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

?

Predicted Genotype

From Sequence To Genotype: Calculate Likelihoods for Each Possibility

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A , read mapped)= 0.00000098

P(reads|A/C , read mapped)= 0.03125

P(reads|C/C , read mapped)= 0.000097

Possible Genotypes

From Sequence to Genotype: Agnostic Prior

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.00000098 **Prior(A/A) =** 0.00034 **Posterior(A/A) =** <.001

P(reads|A/C)= 0.03125 **Prior(A/C) =** 0.00066 **Posterior(A/C) =** 0.175

P(reads|C/C)= 0.000097 **Prior(C/C) =** 0.99900 **Posterior(C/C) =** 0.825

Individual Based Prior: Every site has 1/1000 probability of varying.

From Sequence to Genotype: Population Based Prior

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.00000098	**Prior(A/A) =** 0.04	**Posterior(A/A) =** <.001
P(reads|A/C)= 0.03125	**Prior(A/C) =** 0.32	**Posterior(A/C) =** 0.999
P(reads|C/C)= 0.000097	**Prior(C/C) =** 0.64	**Posterior(C/C) =** <.001

Population Based Prior: Use frequency information from examining others at the same site.

In the example above, we estimated P(A) = 0.20

Sequence Based Genotype Calls

- **Individual Based Prior**
  - Assumes all sites have an equal probability of showing polymorphism
  - Specifically, assumption is that about 1/1000 bases differ from reference
  - If reads where error free and sampling Poisson …
  - … 14x coverage would allow for 99.8% genotype accuracy
  - … 30x coverage of the genome needed to allow for errors and clustering

- **Population Based Prior**
  - Uses frequency information obtained from examining other individuals
  - Calling very rare polymorphisms still requires 20-30x coverage of the genome
  - Calling common polymorphisms requires much less data

- **Haplotype Based Prior or Imputation Based Analysis**
  - Compares individuals with similar flanking haplotypes
  - Calling very rare polymorphisms still requires 20-30x coverage of the genome
  - Can make accurate genotype calls with 2-4x coverage of the genome
  - Accuracy improves as more individuals are sequenced

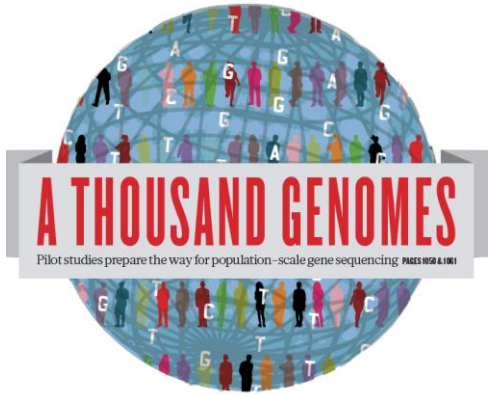
Recipe: Genotypes for Shotgun Sequence Data

- Start with some plausible configuration for each individual
- Use Markov model to update one individual conditional on all others
- Repeat previous step many times
- Generate a consensus set of genotypes and haplotypes for each individual

Genotypes with Shotgun Sequence Data

- Sequence 400 individuals at 2x depth
  - Assume error rate is of about 0.5%
- If we analyze a single individual, almost impossible to call genotypes
  - False positives due to error, 1 in every 100 bases
  - Allele of interest not sampled, 1 in every two heterozygous sites
- If we do an imputation based analysis
  - Expect to call genotypes with 99.7% accuracy for sites with frequency >1%

The 1000 Genomes Project



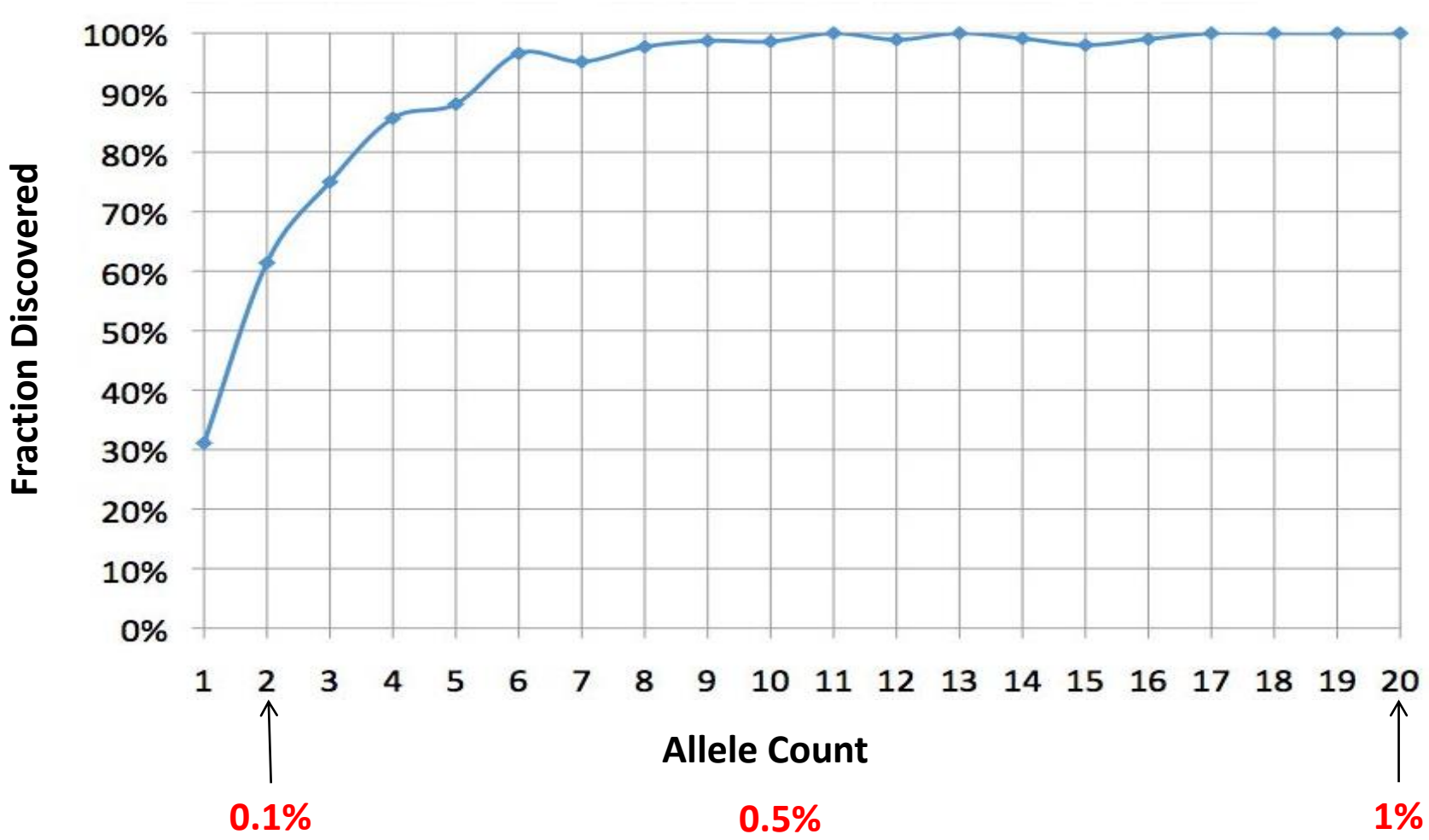
Gil McVean

David Altshuler

Richard Durbin

Empirical Variant Discovery Power

1000 Genomes Project, 4x Sequencing



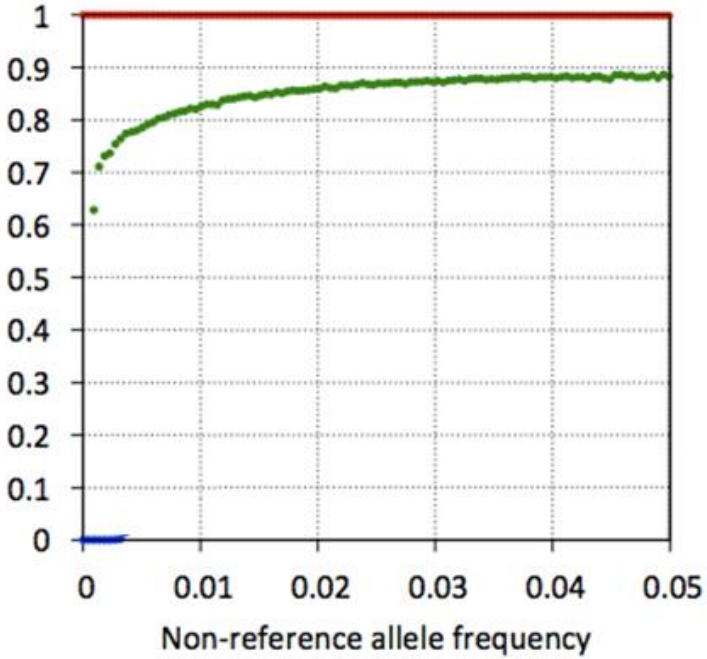
Fraction of variants discovered in low pass sequencing, estimated by comparison with External data.

Hyun Min Kang

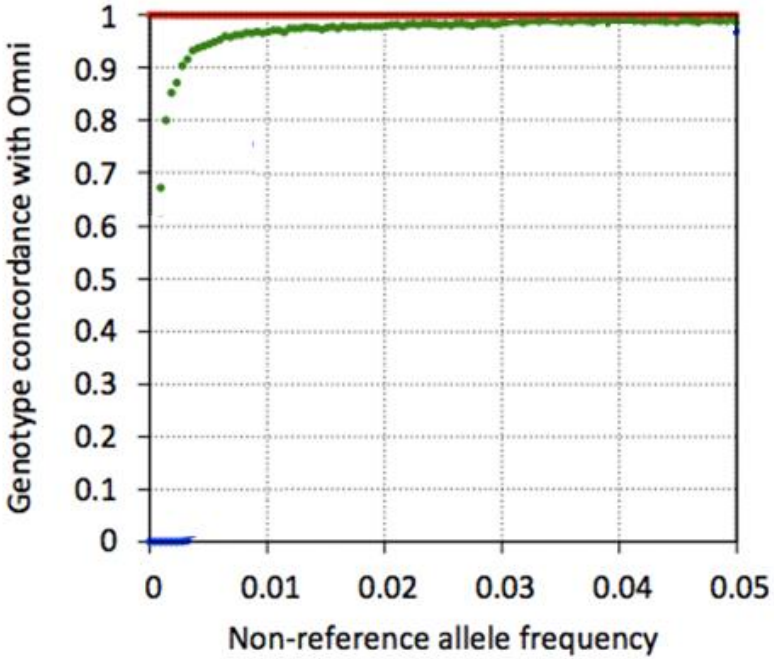
Empirical Evaluation of Haplotype Callers

1000 Genomes Project, 4x Sequencing

Without Haplotype Information



Using Haplotype Information



Homozygote Sites, **Heterozygote Sites**

Optimal Model for Analyzing 1000 Genomes?

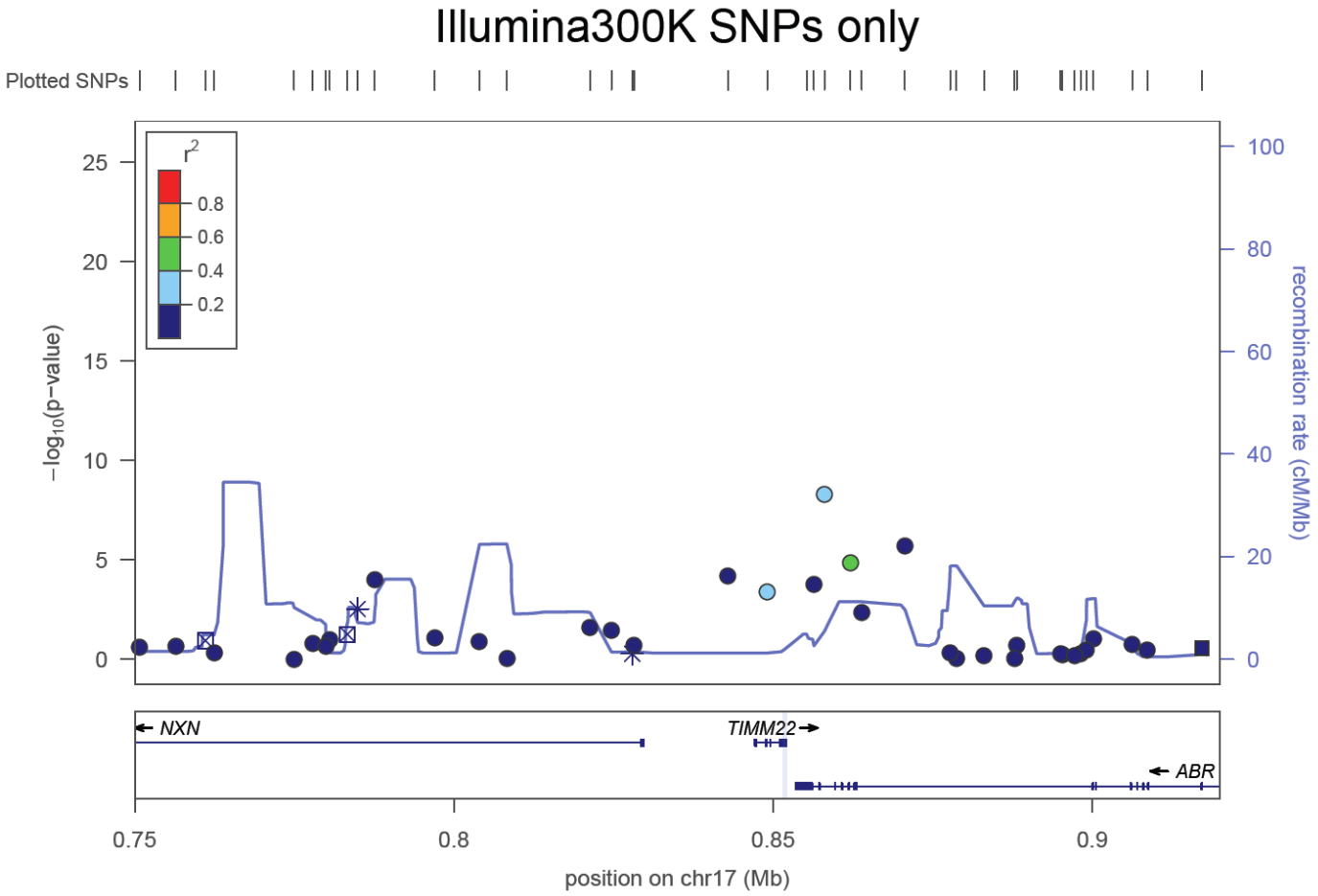
1000 Genomes Call Set (CEU)	Reference Errors	Heterozygote Errors	Homozygous Non-Reference Errors
Broad	0.66	4.29	3.80
Michigan	0.68	3.26	3.06
Sanger	1.27	3.43	2.60

Optimal Model for Analyzing 1000 Genomes?

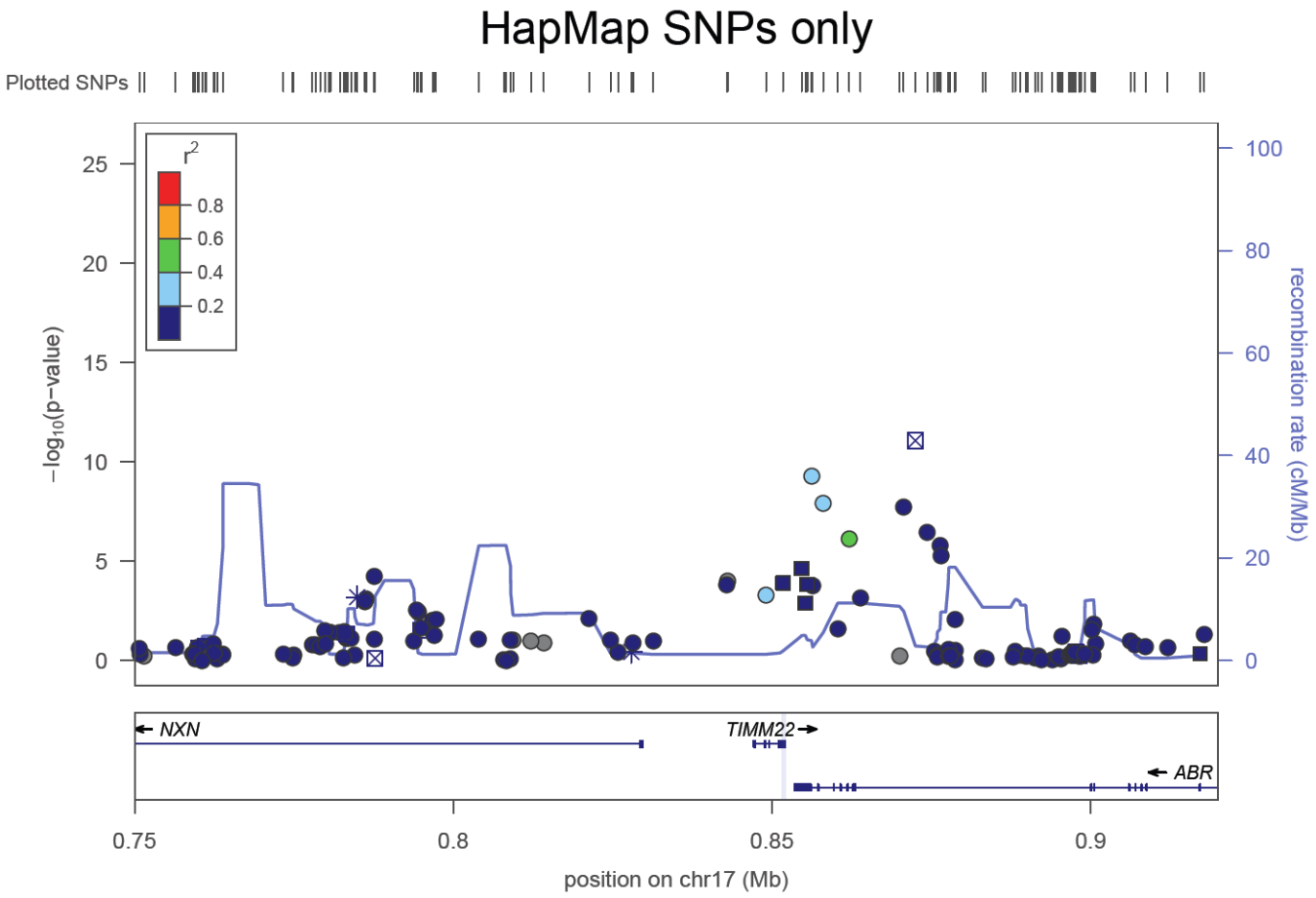
1000 Genomes Call Set (CEU)	Reference Errors	Heterozygote Errors	Homozygous Non-Reference Error
Broad	0.66	4.29	3.80
Michigan	0.68	3.26	3.06
Sanger	1.27	3.43	2.60
Majority Consensus	**0.45**	**2.05**	**2.21**

- “Ensemble” outperforms the best method

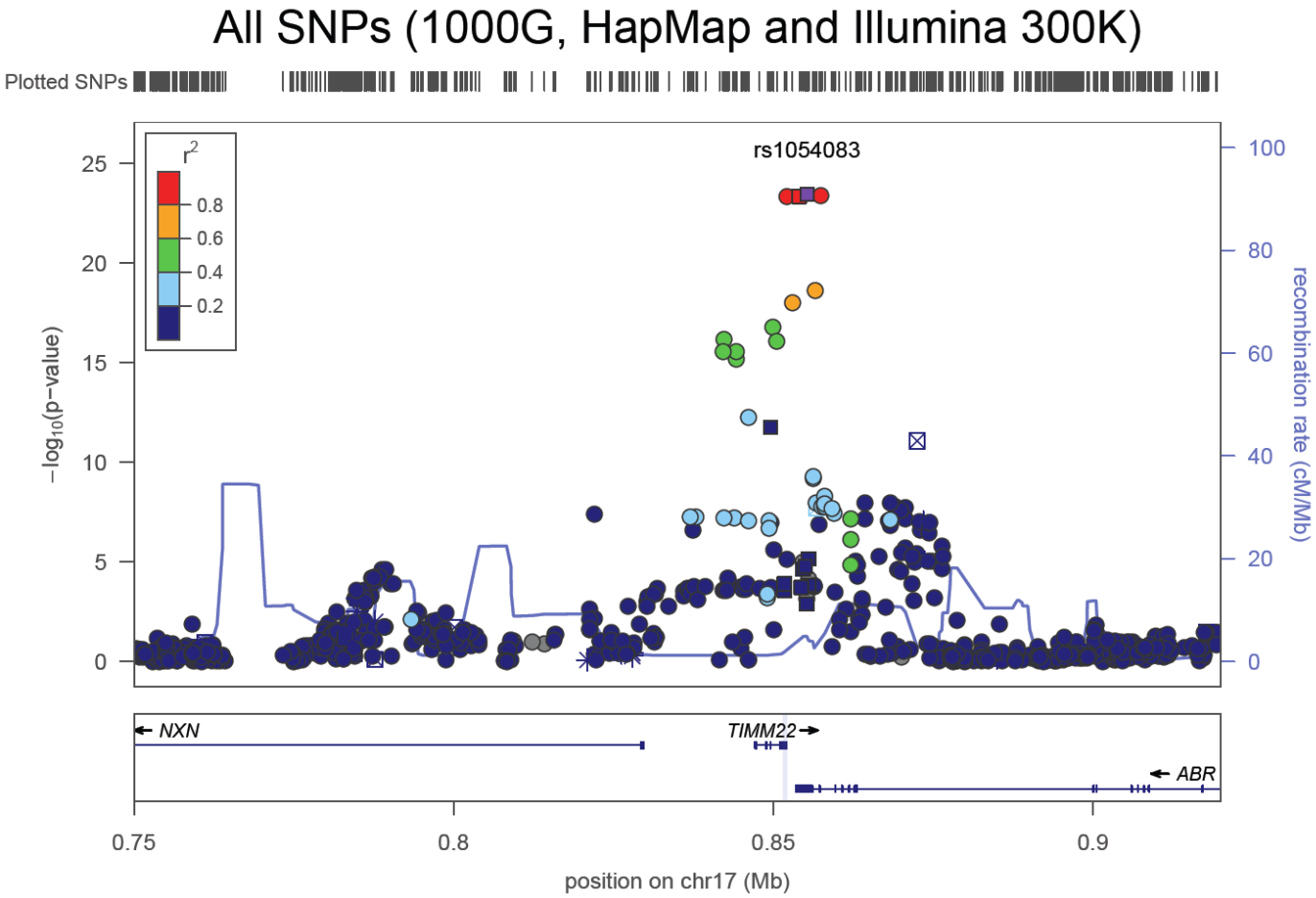
Enhance Association Studies: eQTL Imputation Example



Enhance Association Studies: eQTL Imputation Example



Enhance Association Studies: eQTL Imputation Example



Challenges with the basic approach …

ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG

5′-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3′

Challenges with the basic approach …

ACTGGTCGATGCTAGCTGATAGCTAGCTAG**A**TGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAG**A**TGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAG**A**TGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAG**A**TGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAG**A**TGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAG**A**TGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAG**A**TGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAG**A**TGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG

5′-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3′

Challenges with the basic approach …

```
                                  CTAGATGATGAGCCCGATCGCTGCTAGCTC
                                    AGATGATGAGCCCGATCGCTGCTAGCTCGA
                                     GATGATGAGCCCGATCGCTGTTAGCTCGAC
                                    AGATGATGAGCCCGATCGCTGCTAGCTCGA
                                      ATGATGAGCCCGATCGCTGCTAGCTCGACG
                                     GATGATGAGCCCGATCGCTGCTAGCTCGAC
                                    AGATGATGAGCCCGATCGCTGCTAGCTCGA
                                     GATGATGAGCCCGATCGCTGCTAGCTCGAC
                              GCTAGCTAGCTGATGAGCCCGATCGCTGCT
                         GATAGCTAGCTAGCTGATGAGCCCGCTCGC
                                AGCTAGCTGATGAGCCCGATCGCTGCTAGC
                                  CTAGCTGATGAGCCCGATCGCTGCTAGCTC
                     GCTGATAGCTAGCTAGCTGATGAGCCCGAT
                GATGCTAGCTGATAGCTAGCTAGCTGATGA
            GTCGATGCTAGCTGATAGCTAGCTAGCTGA
                           TAGCTAGCTAGCTGATGAGCCCGATCGCTG

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'
```

Challenges with the basic approach …

ACT**A**GTCGATGCT**G**GCTGATAGCTAGCTAG**A**TGATGAGCCCG**T**TCGCT**C**CTAGCTCGACG
ACT**A**GTCGATGCT**G**GCTGATAGCTAGCTAG**A**TGATGAGCCCG**T**TCGCT**C**CTAGCTCGACG
ACT**A**GTCGATGCT**G**GCTGATAGCTAGCTAG**A**TGATGAGCCCG**T**TCGCTGCTAGCTCGACG
ACT**A**GTCGATGCT**G**GCTGATAGCTAGCTAG**A**TGATGAGCCCG**T**TCGCT**C**CTAGCTCGACG
ACT**A**GTCGATGCT**G**GCTGATAGCTAGCTAG**A**TGATGAGCCCG**T**TCGCT**C**CTAGCTCGACG
ACT**A**GTCGATGCT**G**GCTGATAGCTAGCTAG**A**TGATGAGCCCGATCGCTGCTAGCTCGACG
ACT**A**GTCGATGCT**G**GCTGATAGCTAGCTAG**A**TGATGAGCCCG**T**TCGCT**C**CTAGCTCGACG
ACT**A**GTCGATGCTAGCTGATAGCTAGCTAG**A**TGATGAGCCCG**T**TCGCT**C**CTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG

5′-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3′

Design A Whole Genome Sequencing Study in Sardinia

Gonçalo Abecasis

David Schlessinger

Francesco Cucca

Given Fixed Capacity, Should We Sequence Deep or Shallow?

	.5 – 1%	1 – 2%	2-5%
400 Deep Genomes (30x)			
Discovery Rate	100%	100%	100%
Het. Accuracy	100%	100%	100%
Effective N	400	400	400
3000 Shallow Genomes (4x)			
Discovery Rate	100%	100%	100%
Het. Accuracy	90.4%	97.3%	98.8%
Effective N	2406	2758	2873

Li et al, *Genome Research,* 2011

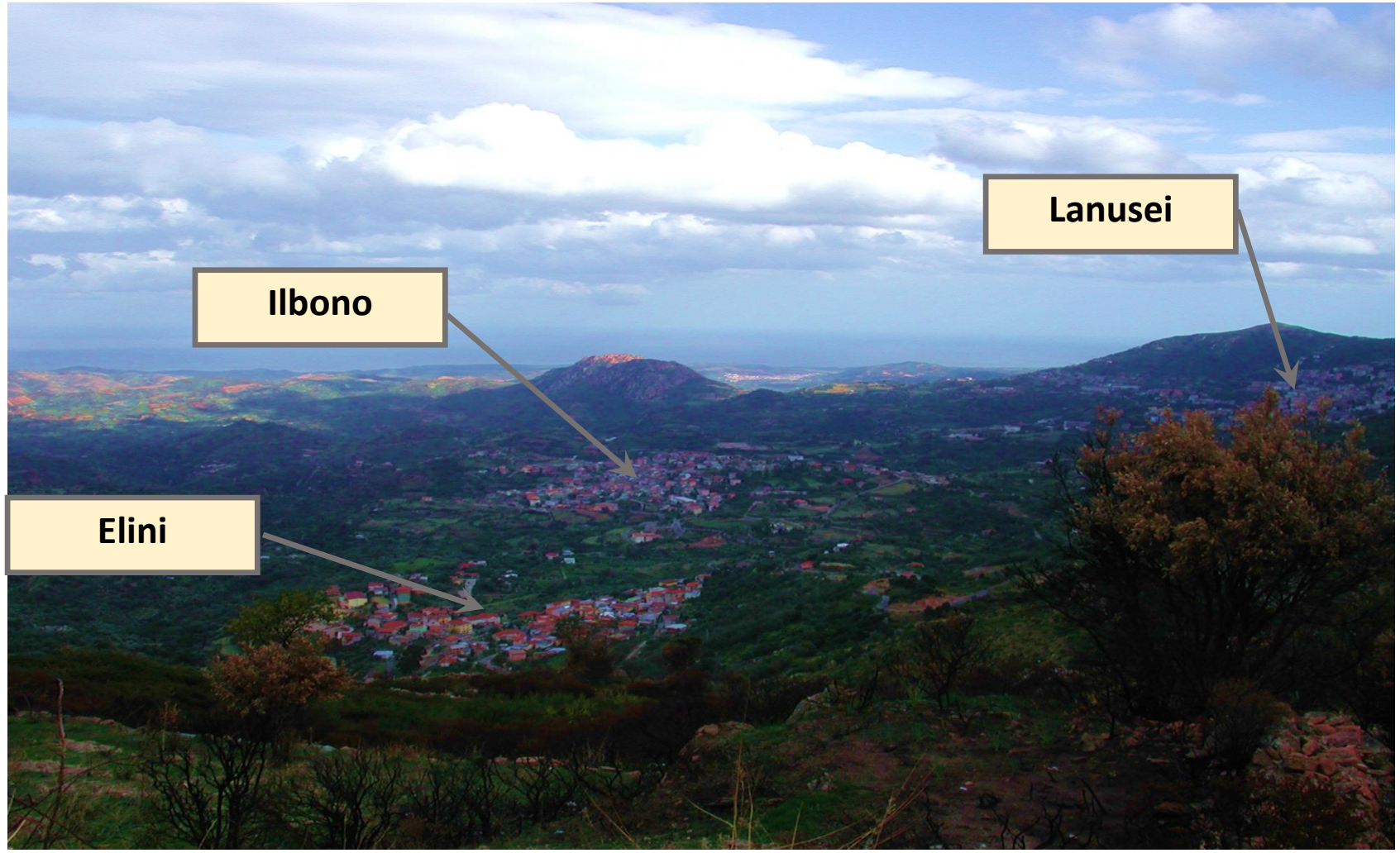
SardiNIA Whole Genome Sequencing

- 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

Lanusei, Ilbono, and Elini viewed from Arzana

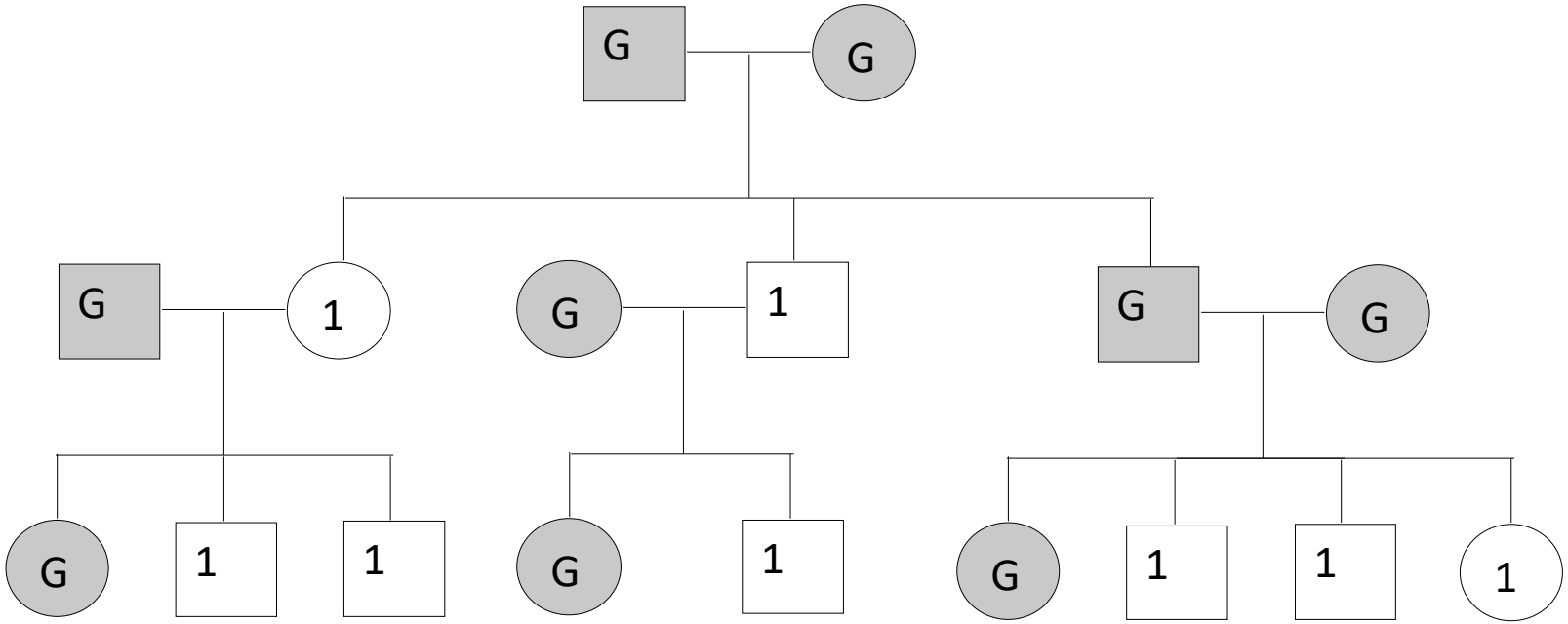


Assembling Sequences In Sardinia

Sardinian team led by Francesco Cucca, Serena Sanna, Chris Jones

Who To Sequence?

Assuming All Individuals Have Been Genotyped



9 Genomes sequenced, 17 Genomes analyzed

How Is Sequencing Progressing?

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

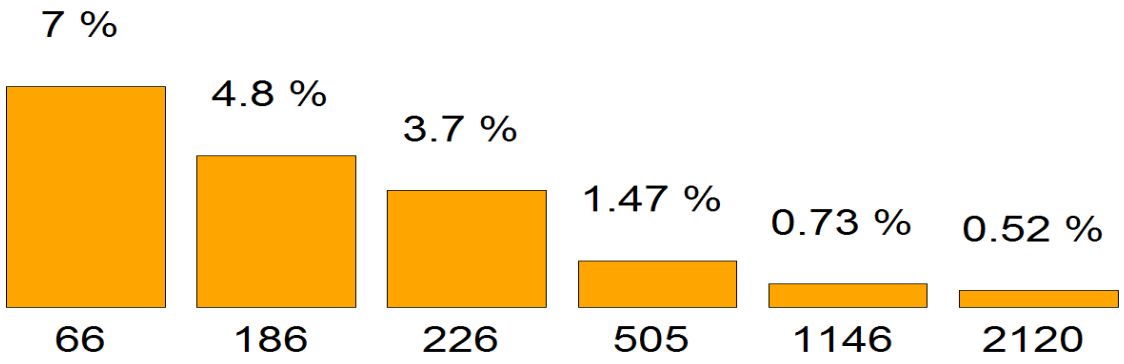
- In our own hands…
  - Mapped high quality bases
  - March 2010: ~5.0 Gb/lane
  - May 2010: ~7.5 Gb/lane
  - September 2010: ~8.6 Gb/lane
  - January 2011: ~16 Gb/lane
  - Summer 2011: ~45 Gb/lane

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

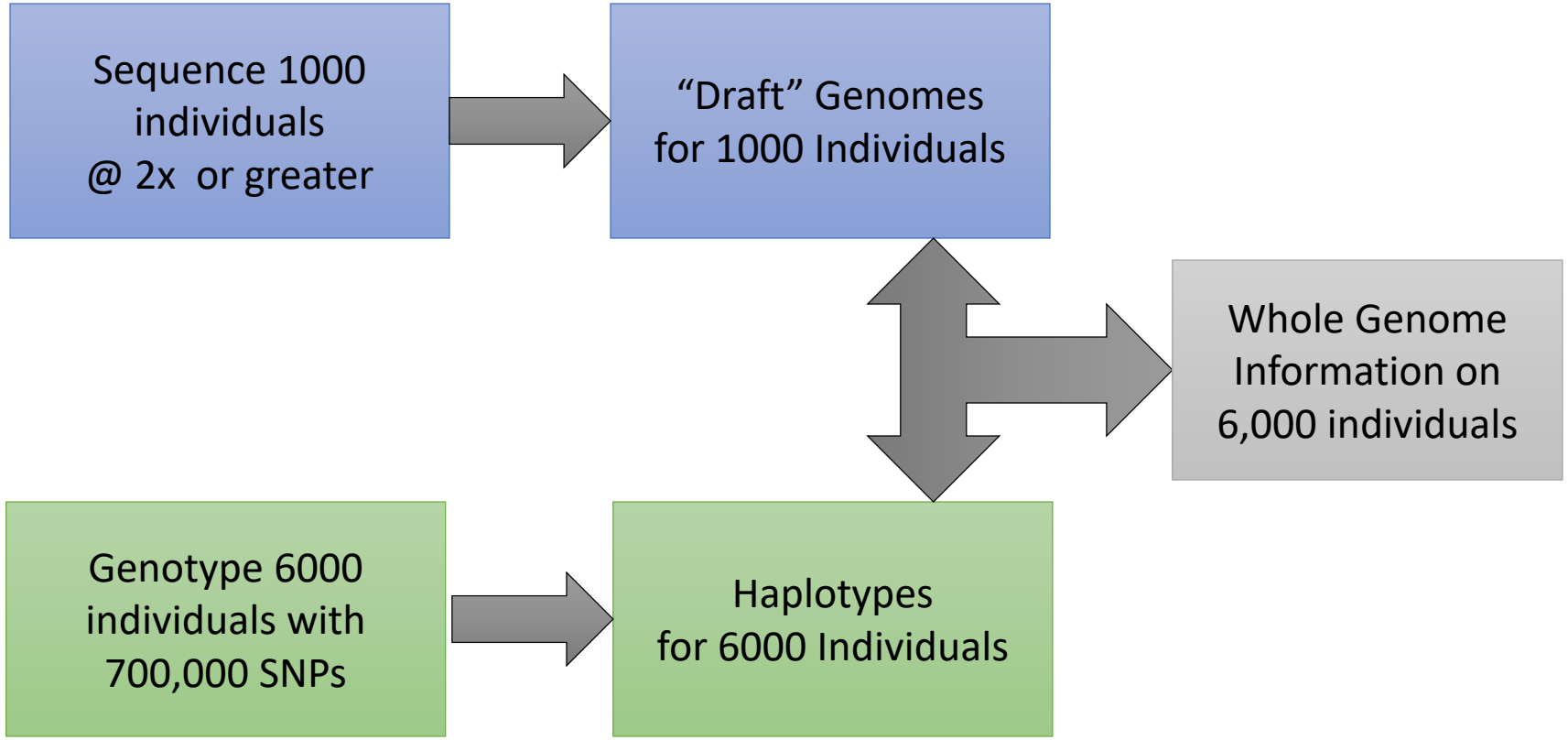
Fabio Busonero, Andrea Maschio

As more samples are sequenced, Accuracy increases

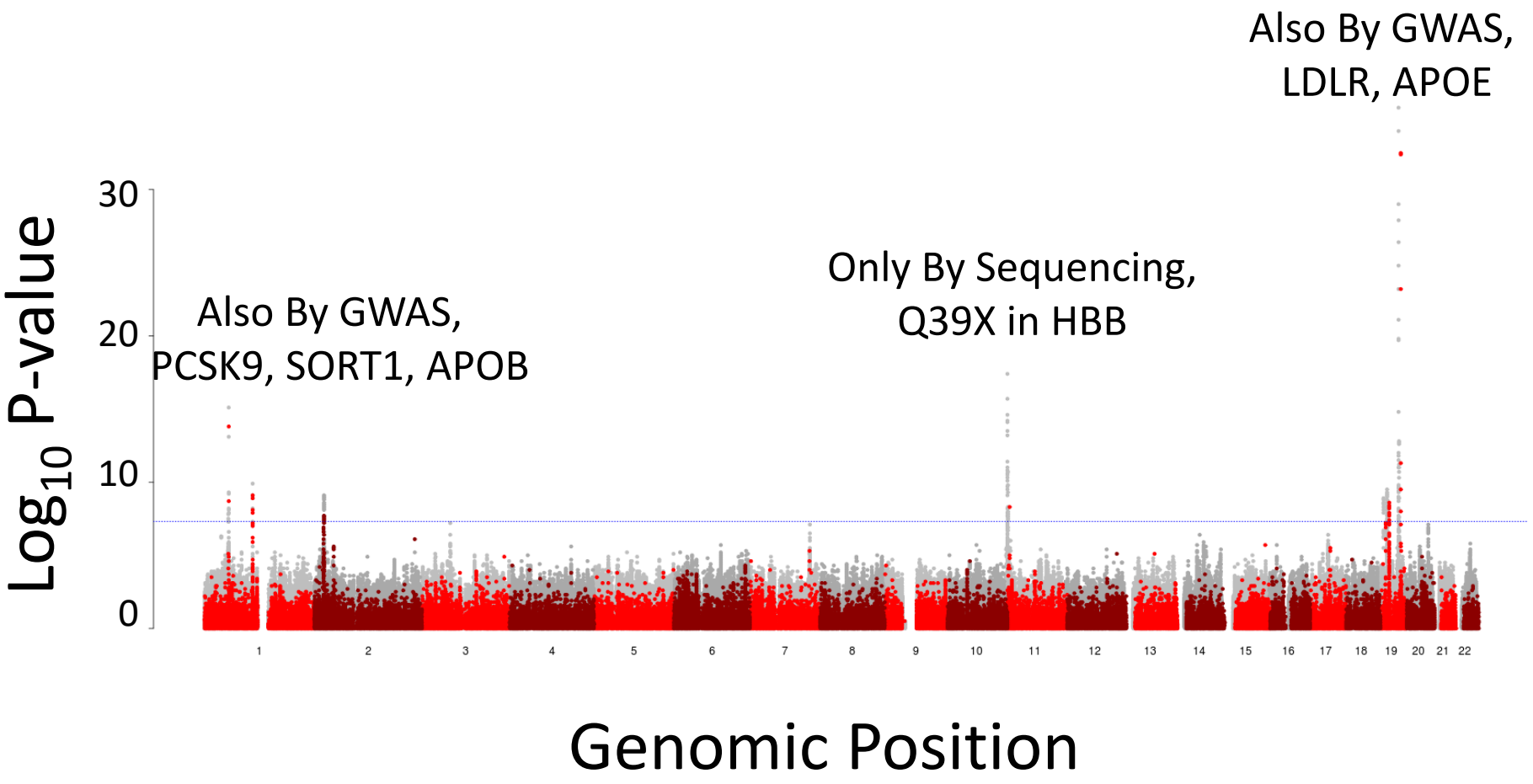
Heterozygous Mismatch Rate (in %)

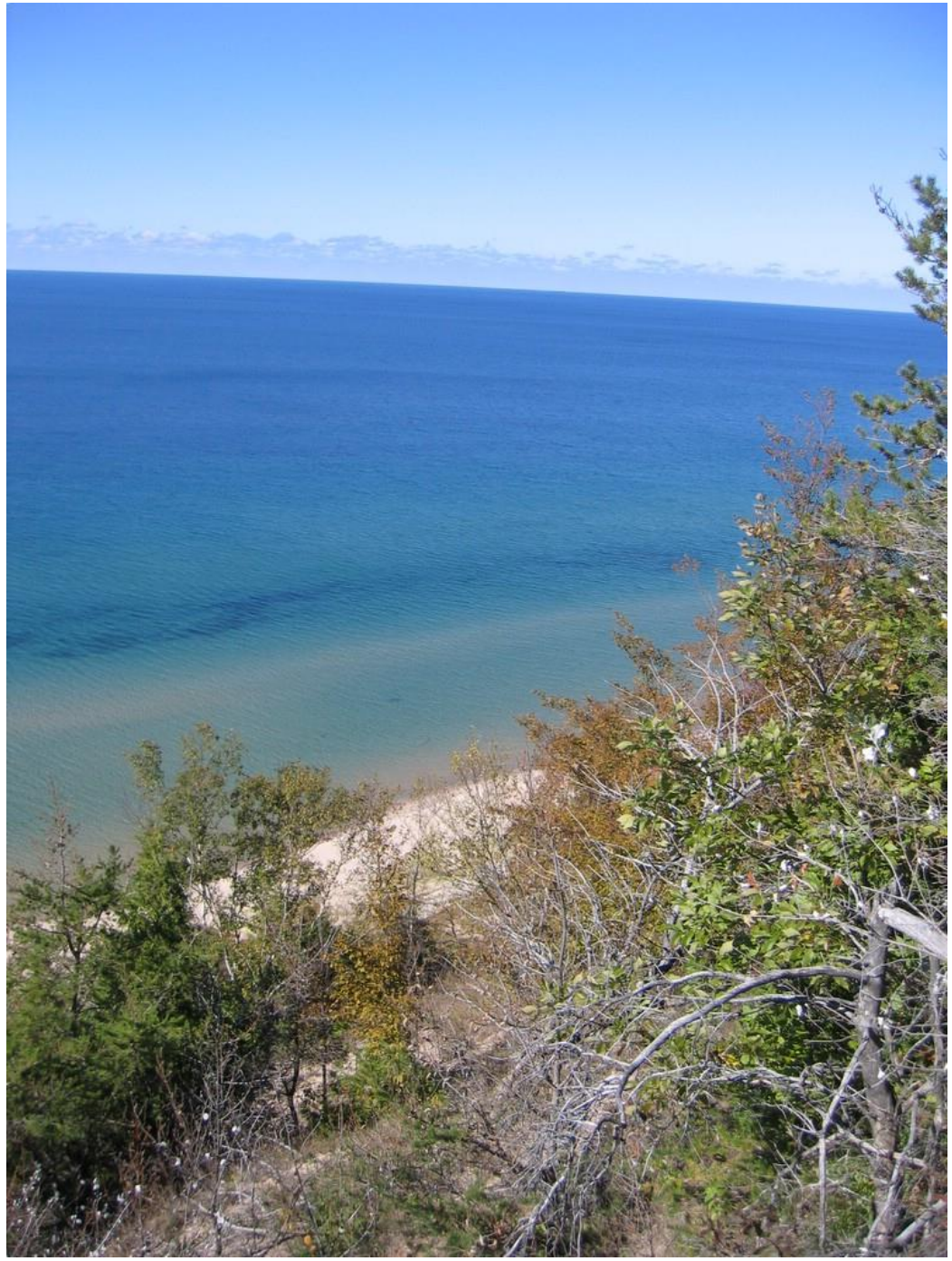


Design


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





The Future is Now

Human Genetics, Sample Sizes over My Time

Year	No. of Samples	No. of Markers	Publication
Ongoing	120,000	600 million	NHLBI Precision Medicine Cohorts / TopMed
2016	32,488	40 million	Haplotype Reference Consortium (Nature Genetics)
2015	2,500	80 million	The 1000 Genomes Project (Nature)
2012	1,092	40 million	The 1000 Genomes Project (Nature)
2010	179	16 million	The 1000 Genomes Project (Nature)
2010	100,184	2.5 million	Lipid GWAS (Nature)
2008	8,816	2.5 million	Lipid GWAS (Nature Genetics)
2007	270	3.1 million	HapMap (Nature)
2005	270	1 million	HapMap (Nature)
2003	80	10,000	Chr. 19 Variation Map (Nature Genetics)
2002	218	1,500	Chr. 22 Variation Map (Nature)
2001	800	127	Three Region Variation Map (Am J Hum Genet)
2000	820	26	T-cell receptor variation (Hum Mol Genet)

Challenges

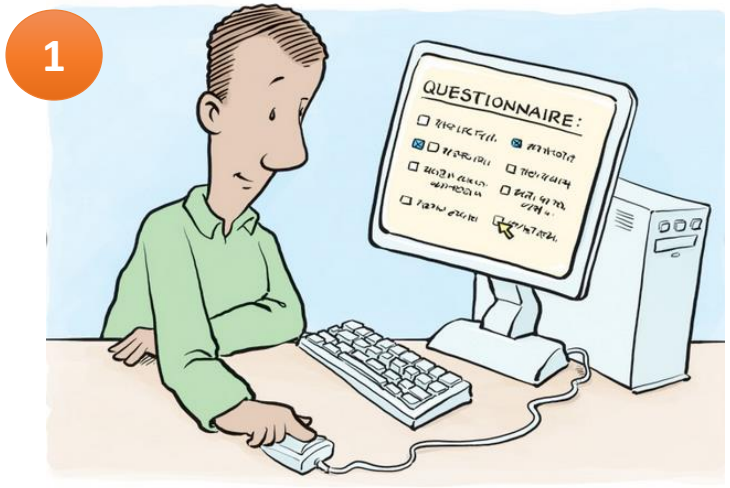
- How do we move faster from cataloguing loci to advancing biology?
- Engaging populations at the scale of 10,000s of individuals
- Sequencing at the scale of 10,000s of genomes
- Explore new technologies that accelerate functional analyses
- Make sure we don’t get bogged down with basics
  - Simplify processes for running analyses we are good at
  - Simplify processes for trying new ideas on data

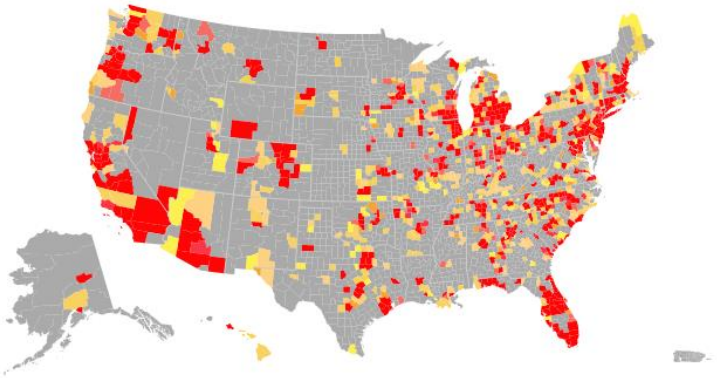
How Can We Engage 10,000s of Research Participants?

Part I – Genes for Good

GENES for GOOD

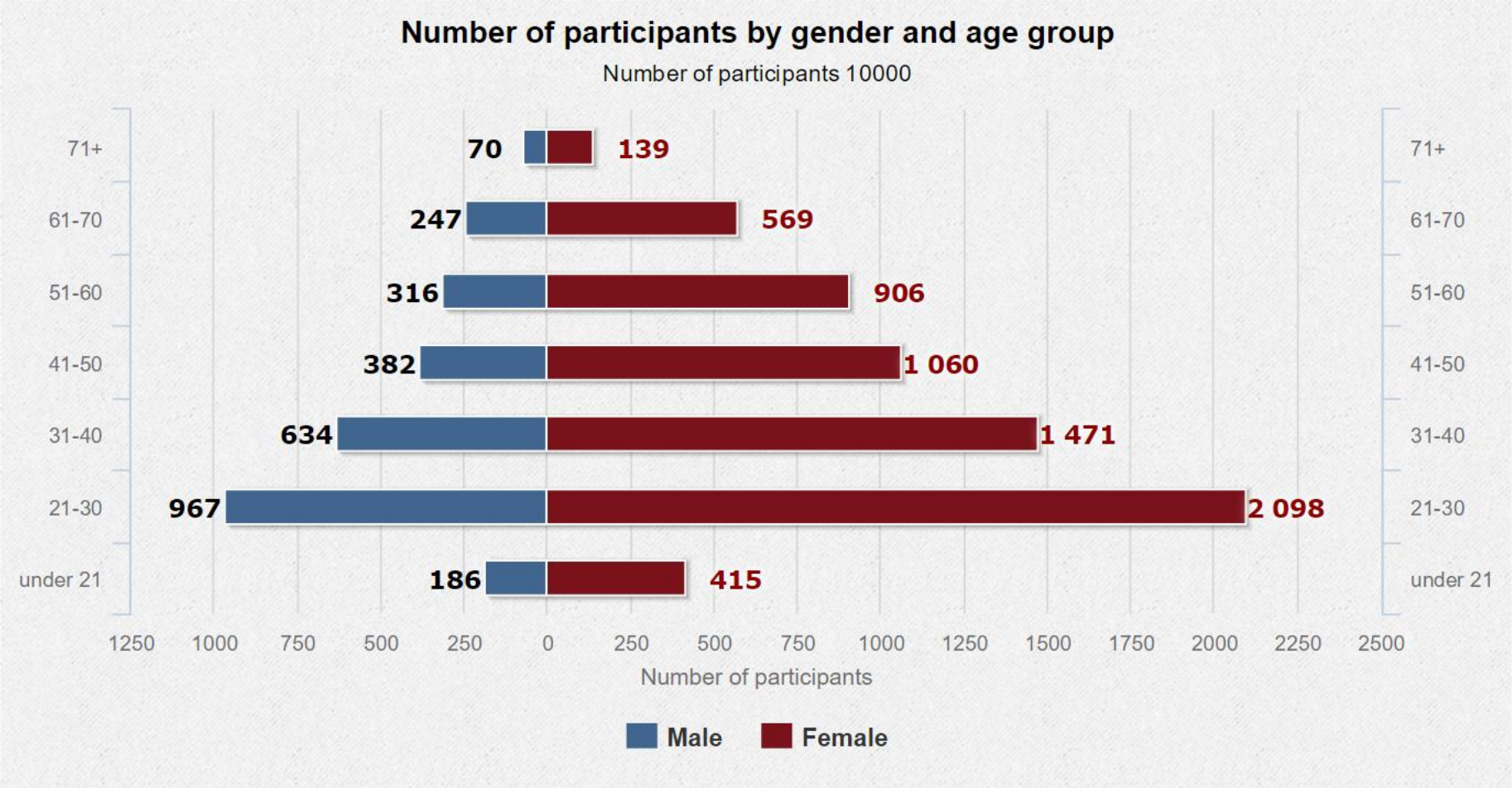
Scott Vrieze



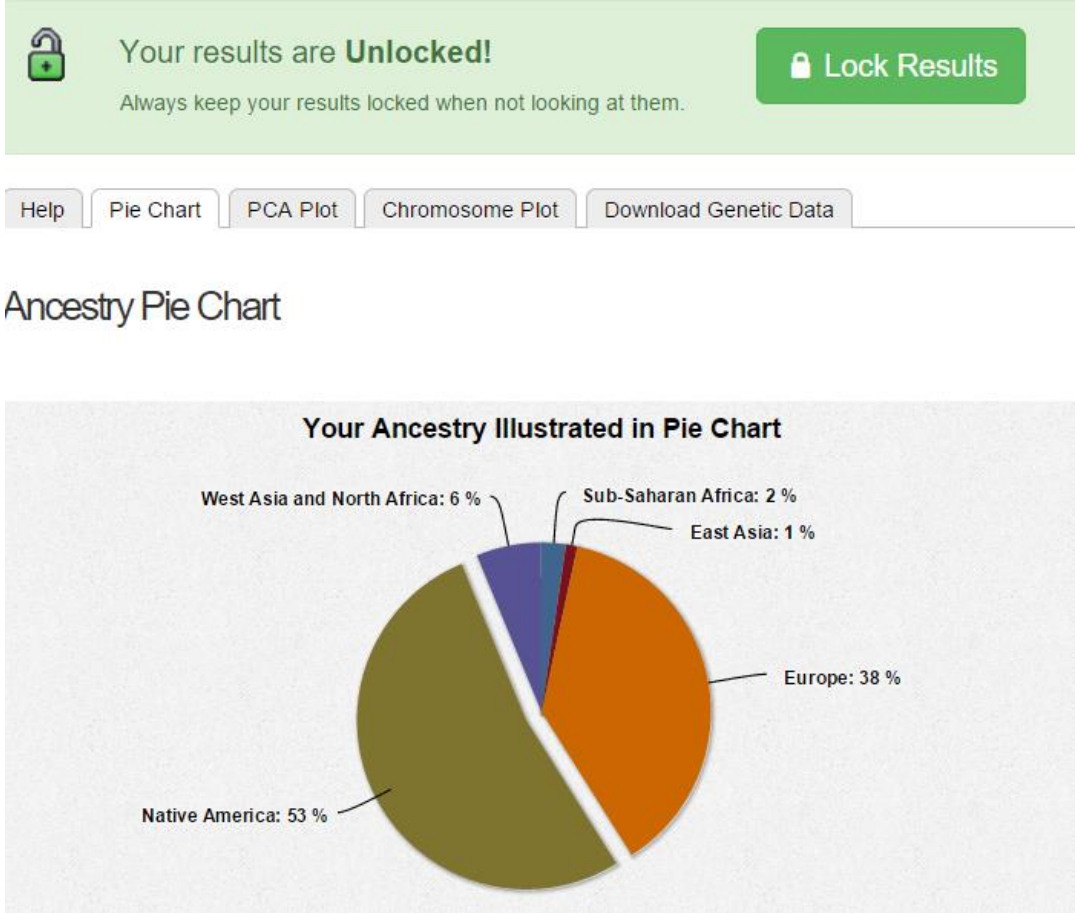


- Exploring new ways to engage populations in research
- Continuous Engagement, Web, Mobile Devices
- Currently, >25,000 participants
- www.genesforgood.org

10,000 Participants…

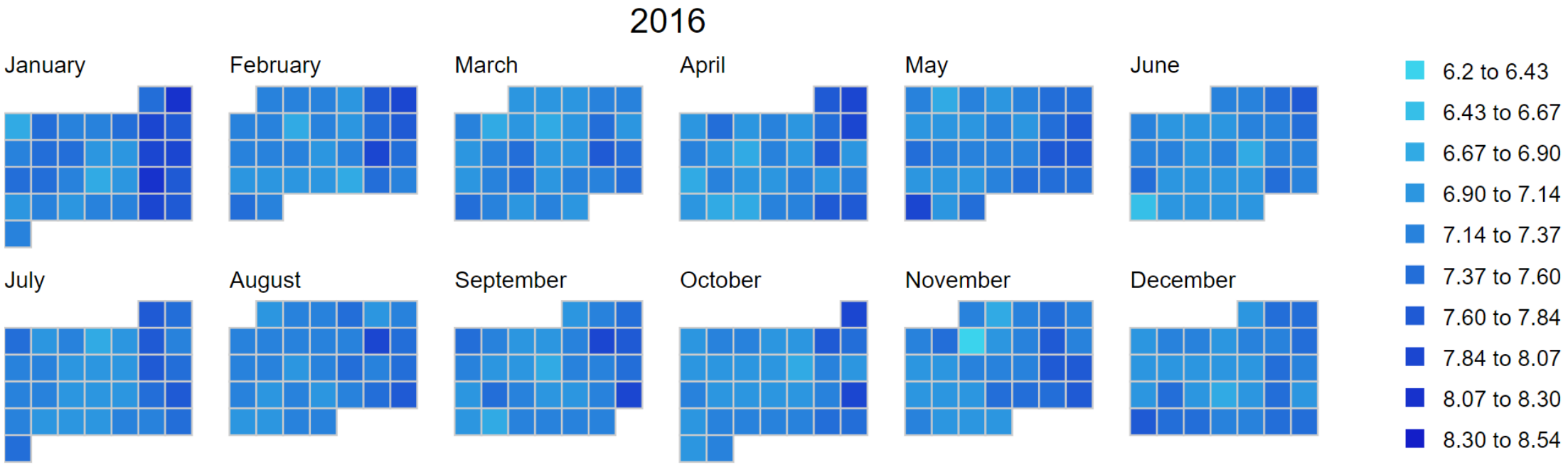


Return of Results

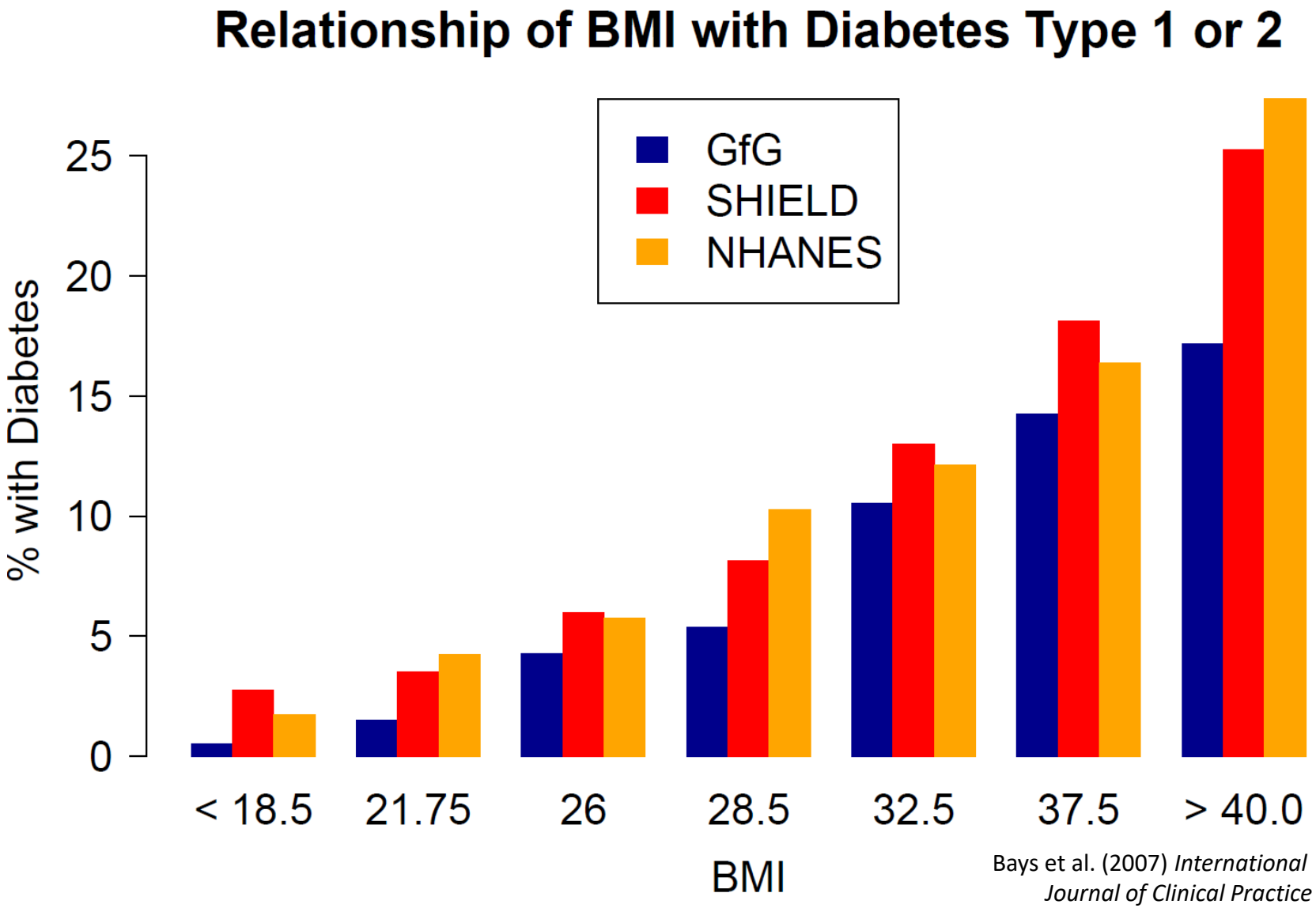


Average Reported Sleep Hours Over a Year

Anita Pandit



BMI, Age & Diabetes



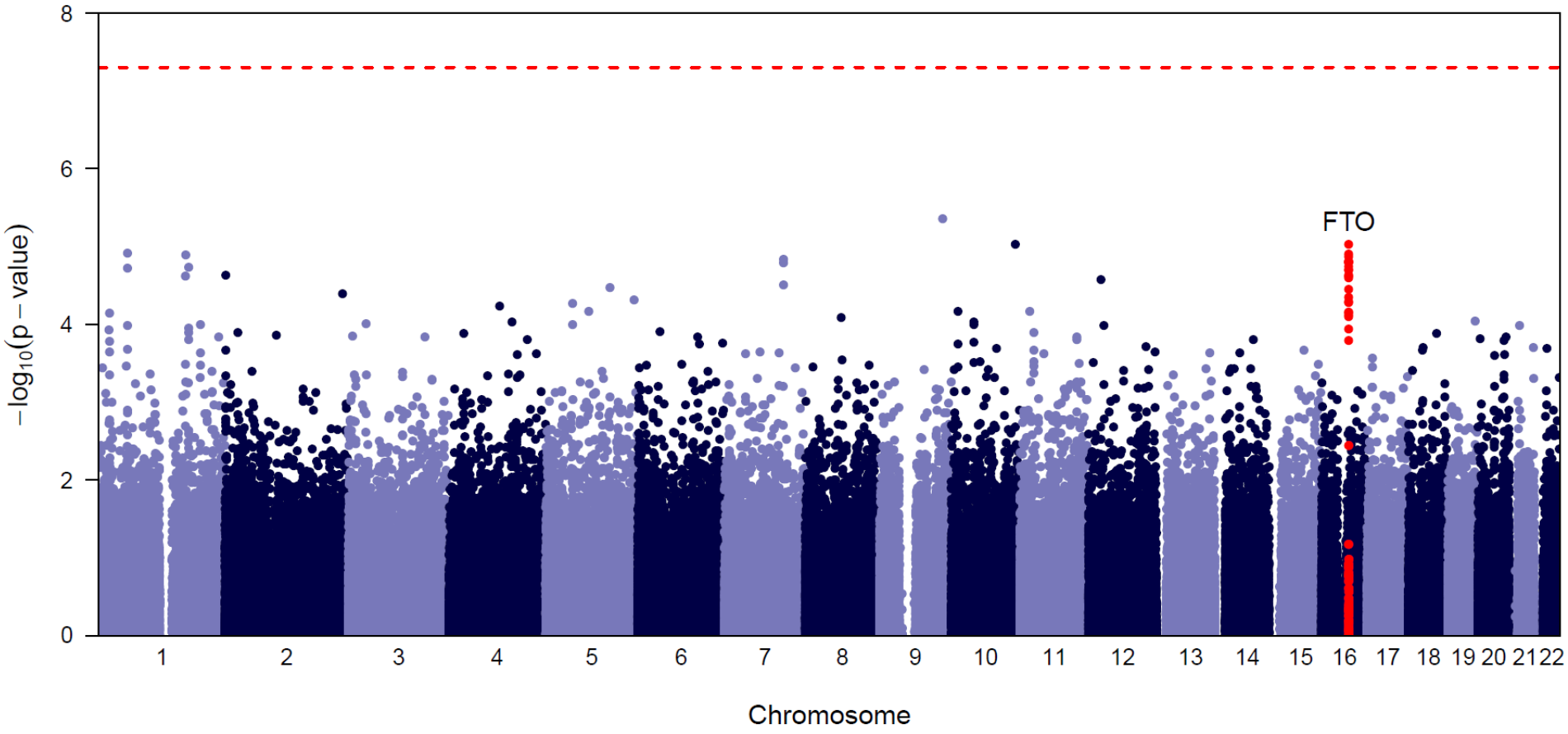
Bays et al. (2007) *International Journal of Clinical Practice*

Kate Brieger

Results: BMI GWAS

Greg Zajac

Pheno	n	Chr:Pos	SNP	Gene	Our P	Other P*
BMI	2,851	16:53803574	rs1558902	*FTO*	5×10^{-5}	5×10^{-120}



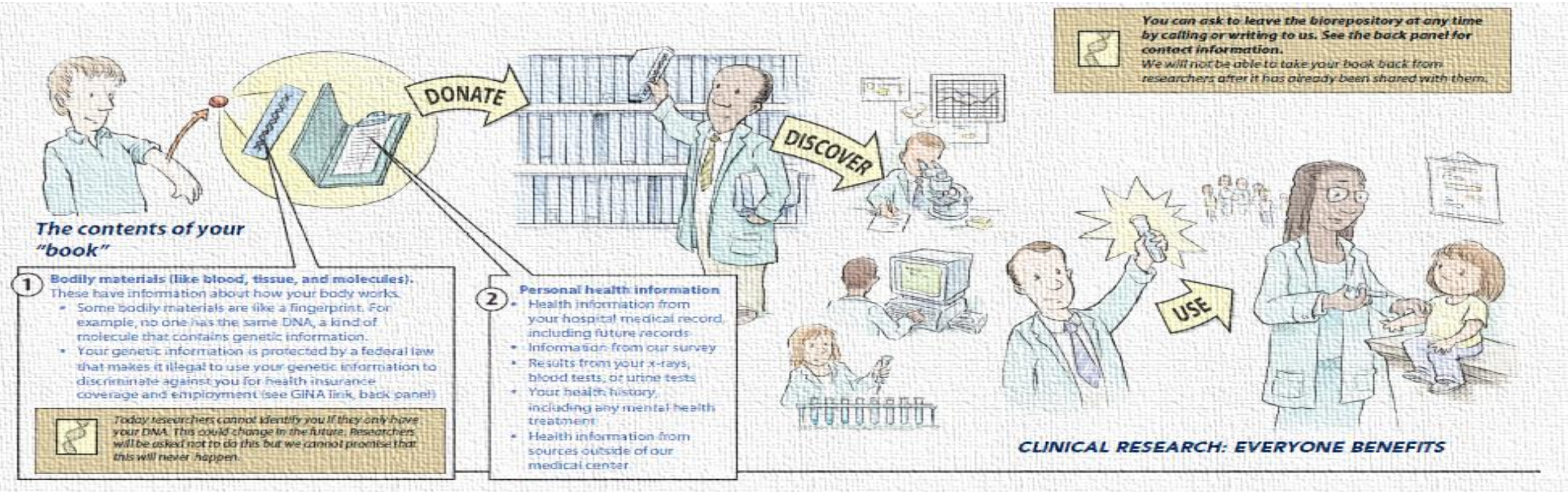
*Speliotes et al. (2010) *Nature Genetics*

How Can We Engage 10,000s of Research Participants?

Part II – Michigan Genomics Initiative

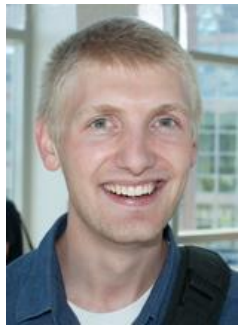
Michigan Genomics Initiative

- Combine genetic and electronic health information on 40,000+ patients — 50 new participants per day
- Use genetic information study many traits and diseases — Diverse traits – 40% w/cancer
- Build catalog of naturally occurring human knockouts — Speed and improve translation
- Clear, easy to understand consent – full participant buy-in. — Key for long term success
- **Team effort: Schmidt (Analysis), Ketherpal (Electronic Health Records), Brummett (Recruitment)**

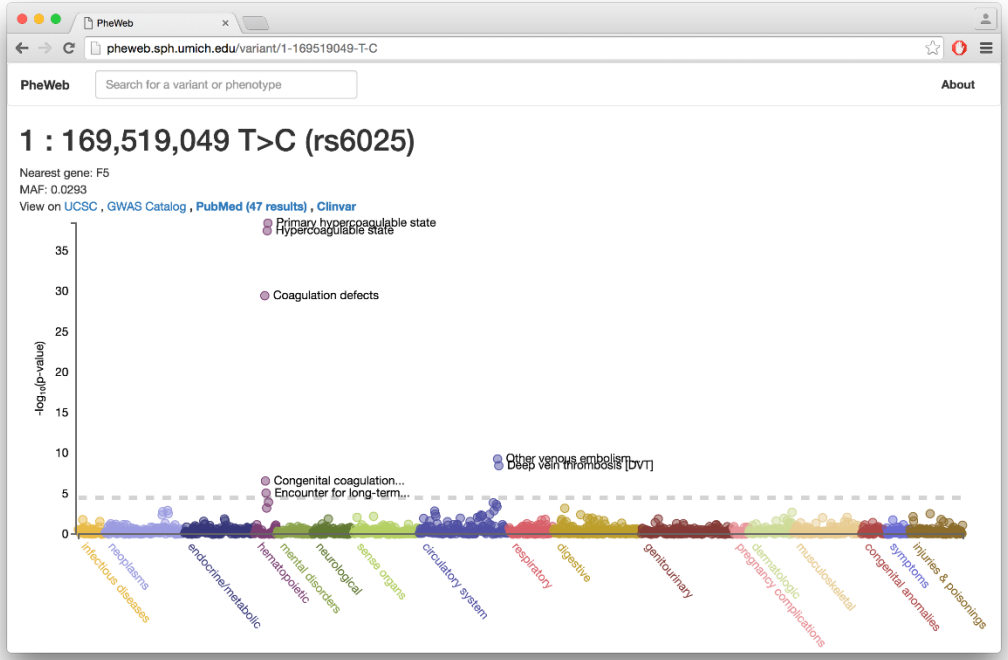


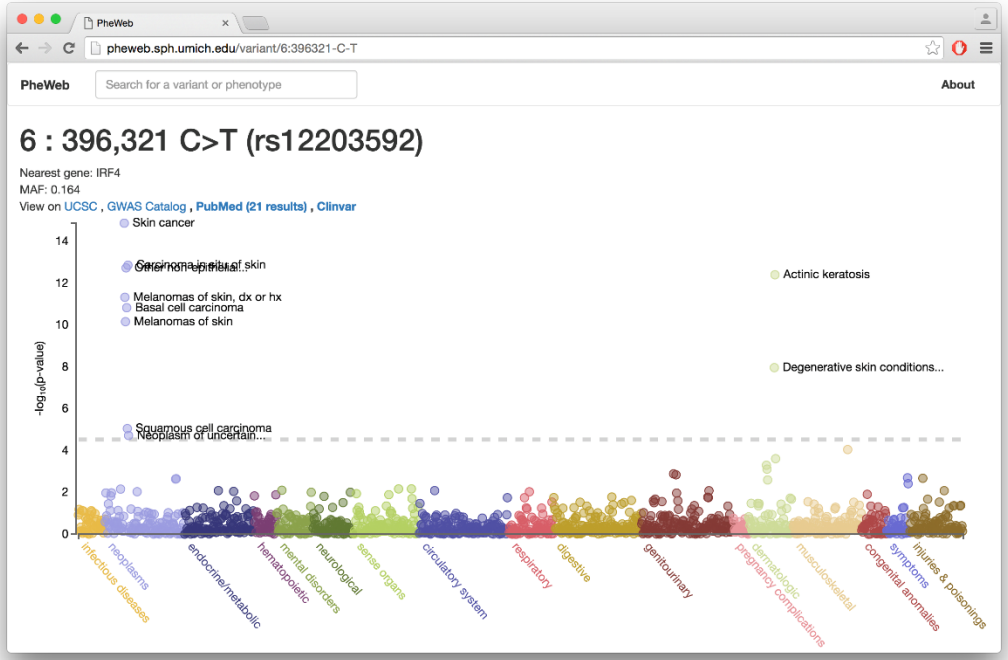
Michigan Genomics Initiative (Freeze 1)
20,000 individuals
7.5 million variants x 1,500 phenotypes

Ellen Schmidt



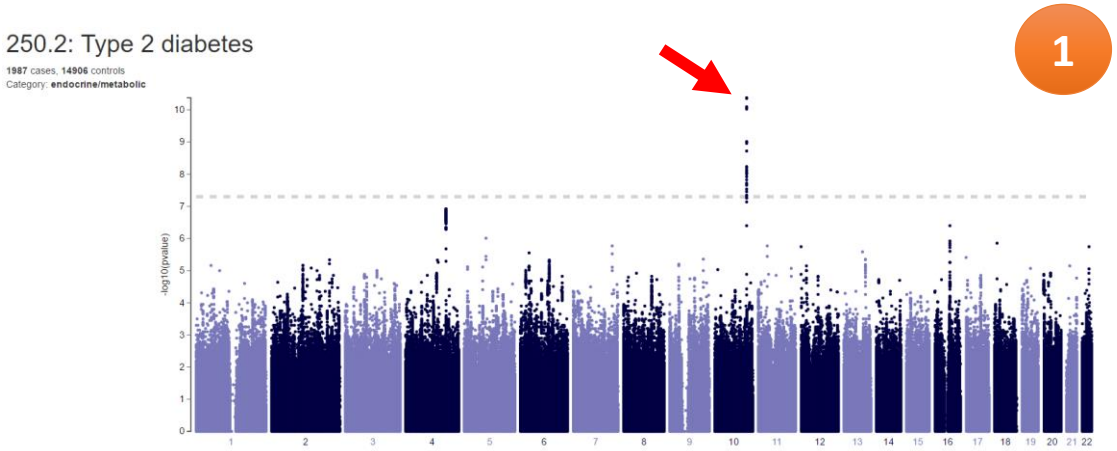
Peter VandeHaar

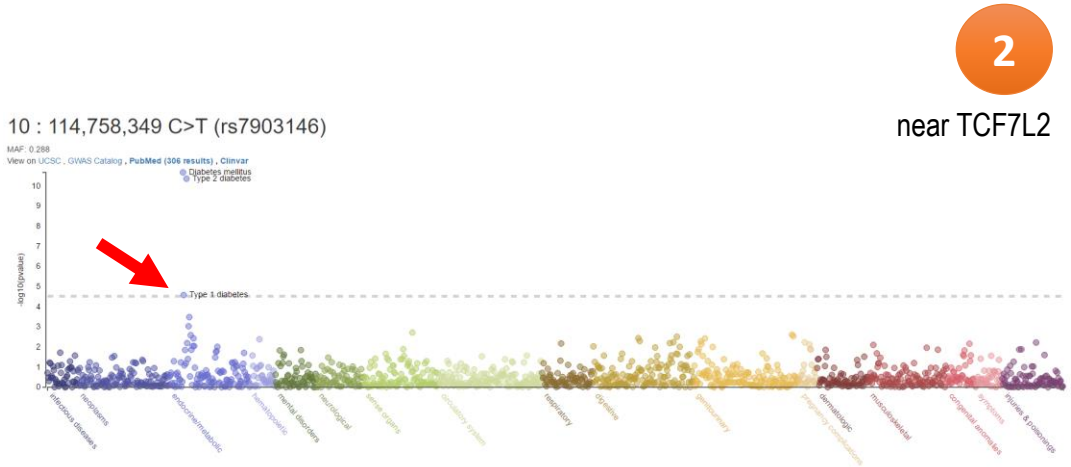


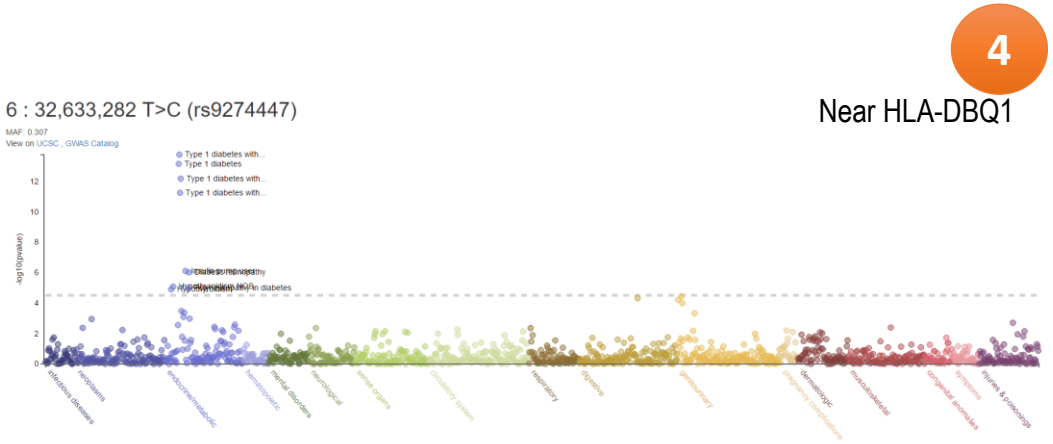


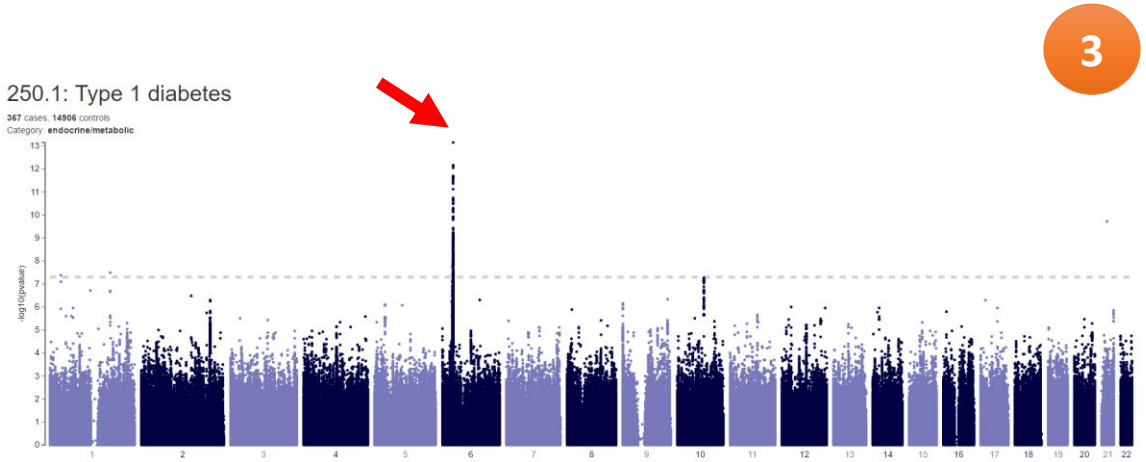
Michigan Genomics Initiative Association Statistics

http://pheweb.sph.umich.edu

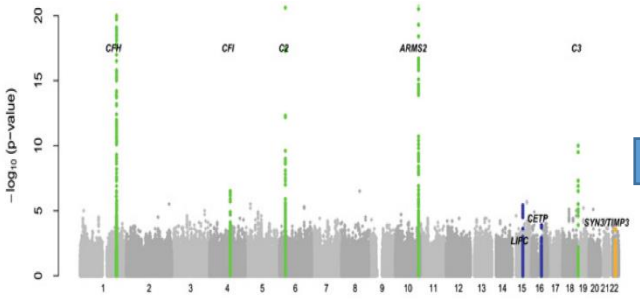


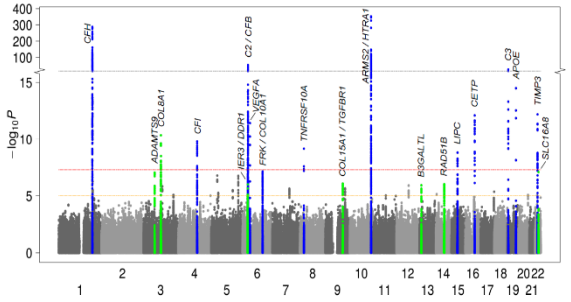


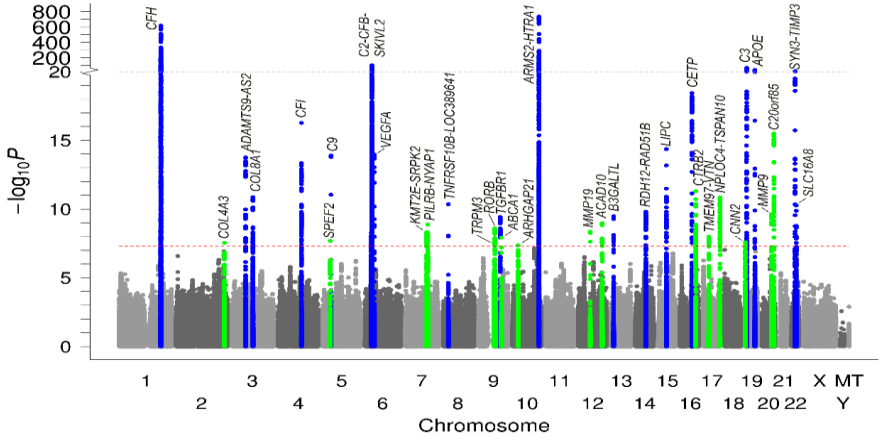




Federate!







Wouldn’t it be nice to combine analysis without data use agreements and exchanging individual level data?

PheWeb Goals

- Enable researchers to easily federate data

- Enable remixing interesting analysis without accessing individual data
  - Compute a novel association statistic
  - Retrieve association results for all variants in a set
  - Compute a new burden test for a gene or coding element
  - Carry out a Mendelian randomization analysis

- How?
  - Enable APIs to deliver intermediate algebra results that go into analyses

Sequencing At Scale

The NHLBI TOPMed Program

- Trans-Omics for Precision Medicine
- Advance knowledge of heart, lung and blood disorders
- Add high-quality ‘omics’ data to high-priority studies
  - Whole genome sequencing currently executed at scale
  - Gene expression and metabolomics in pilot phases
- Data deposited in national databases, available for others to analyze

TOPMed Sequencing as of February 15, 2017

http://nhlbi.sph.umich.edu/

- 68,503 genomes
  - 67,317 pass quality checks (98.3%)
  - 823 flagged for low coverage (1.2%)
  - 358 fail quality checks (0.5%)

- Mean depth: 38.3x
- Genome covered: 98.7%
- Contamination: 0.29%

- 9 x 10^{15} sequenced bases

9×10^{15} sequenced bases

Number of snowflakes covering ~9 square miles in a 10-inch deep snowstorm.
100x more data than the 1000 Genomes Project.

9×10^{15} sequenced bases

US corn production in 2014: 1.3×10^{15} kernels

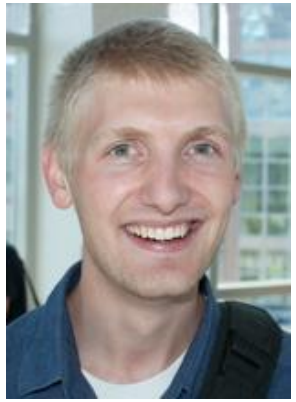
Image: Patrick Porter @ Smug Mug

1.6M Coding Variants

Category	Count	Singletons
All SNPs	191M	43.5%
-- Missense SNPs	1.5M	47.9%
-- LoF SNPs	39K	55.5%
All Indels	10.1M	43.2%
-- Inframe Coding Indels	21K	49.3%
-- Frameshift Indels	31K	59.2%

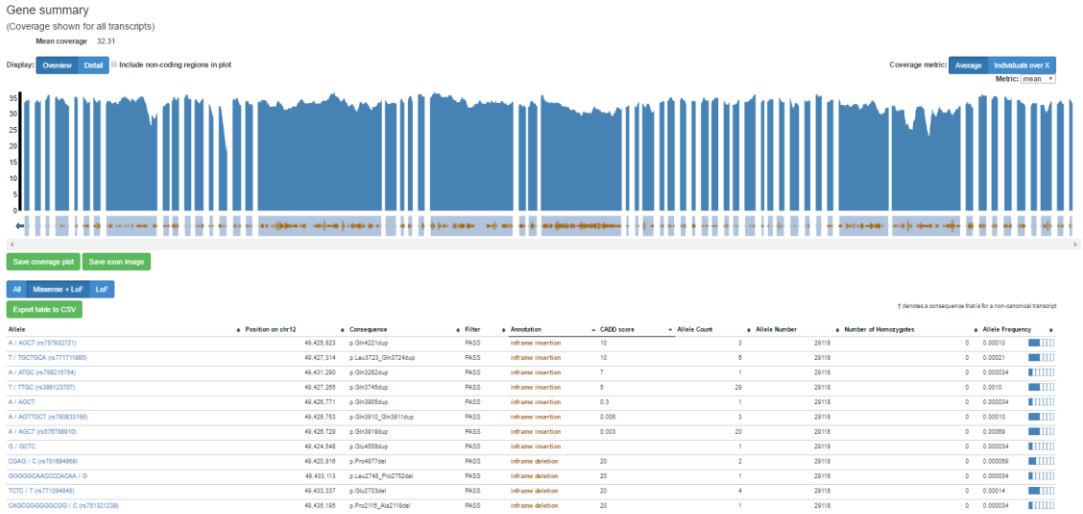
Browse All Variations Online

http://bravo.sph.umich.edu



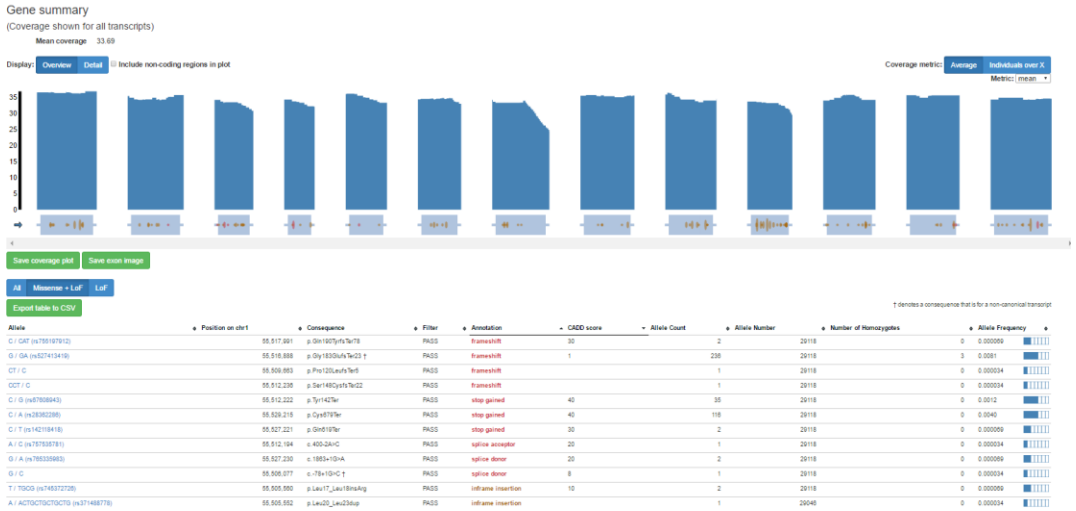
Peter VandeHaar

KMT2D



496 missense, 26 inframe indels, 0 stop or frameshifts

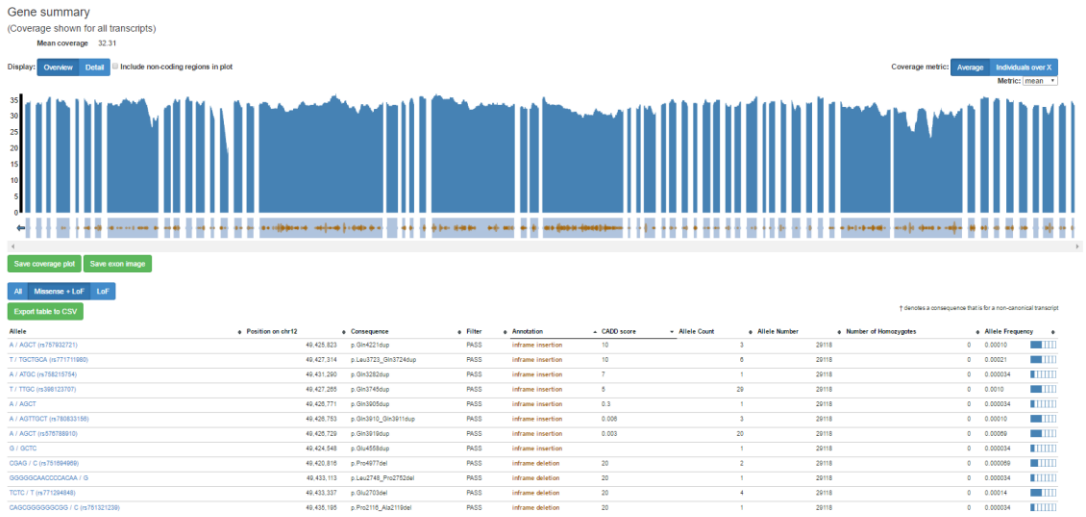
PCSK9



91 missense, 4 inframe indels, 7 stop or frameshifts

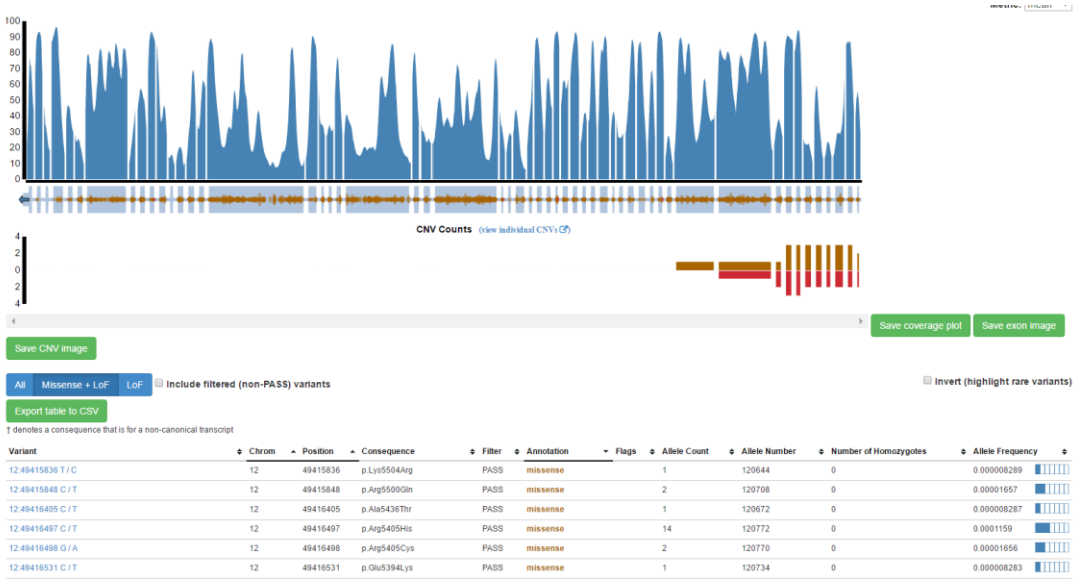
Federate!

KMT2D - BRAVO



496 missense, 26 inframe indels, 0 stop or frameshifts

KMT2D - ExAC



1842 missense, 23 inframe indels, 11 stop or frameshifts

The future?

- We will sequence millions of genomes.
- We will discover better computational methods and strategies.

- However, data is not understanding and tools are not analyses (unfortunately!)

- We must enable scientists to interact with data, understand it, choose powerful study designs, and answer the important questions that drive them.

- Need continued, and probably increased, focus on communication across fields and subfields.

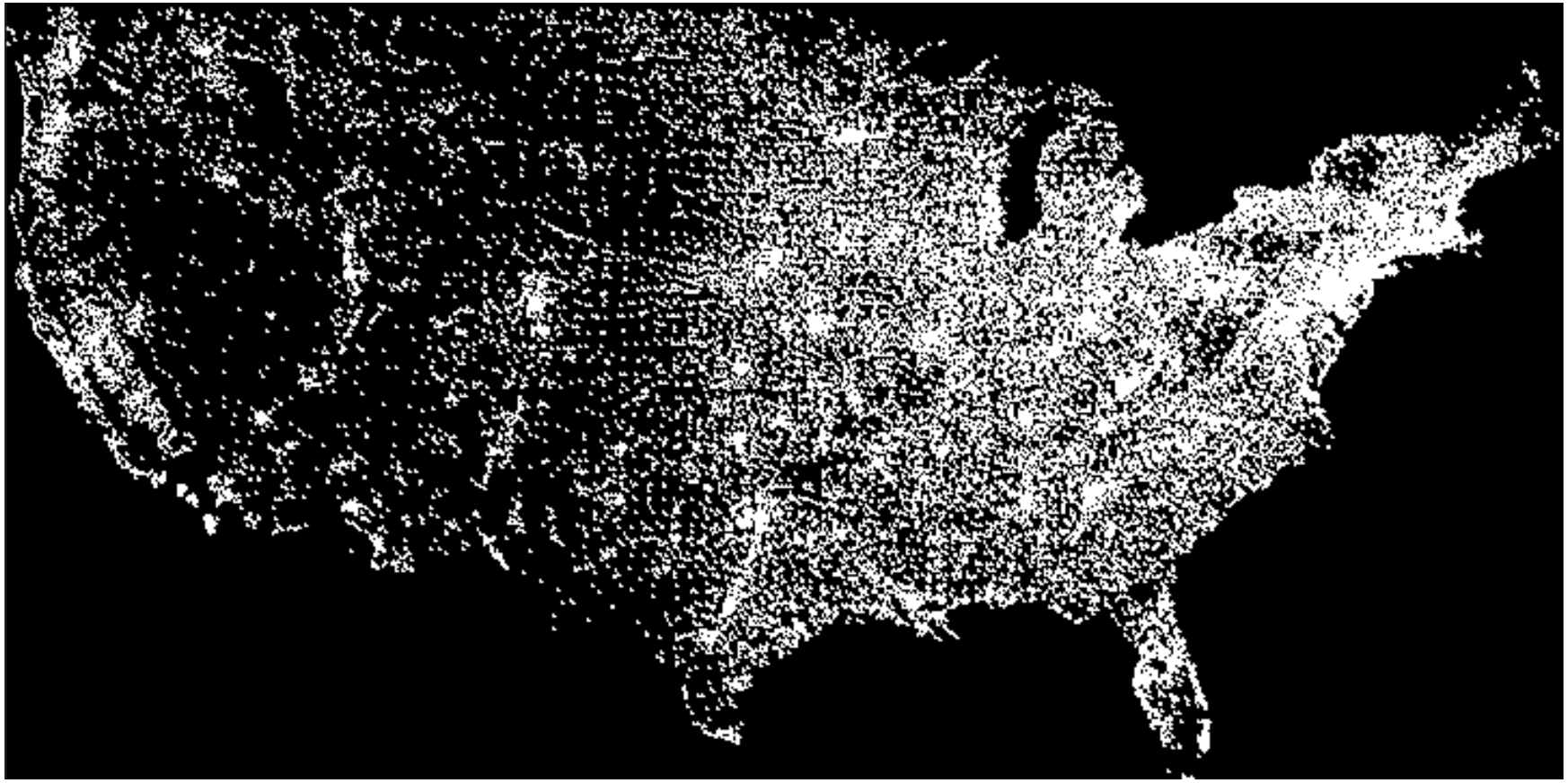
How Great Analysts Contribute …

- Carry out top-notch analyses that point biology in the right direction
- New analysis tools and methods that scale, add value and meaning to data
- Enable new paradigms for collecting and sharing research data
- Expose data and analysis tools to broad community, including non-experts
  - Infuse high-quality health and genetic data in all research

Making Data & Results Available …

- Non-technical users …
  - Need to access and understand results of large scale genomic studies
  - Interact with analysis and results on demand, often don’t need individual data
- Statisticians & method developers
  - Need to compute over data without becoming big data computer scientists!
  - Use APIs to remix interesting analyses based on sufficient statistics
- Hard core method developers
  - Will want to access raw sequence data and develop low level computational methods
  - Need access to individual level data
- Three different needs and expectations
  - We should try to make some of these uses as close to zero friction as possible

A Lattice of Sequenced Genomes



Lessons learned…

- One person and a good idea can make a difference.
- The best students, postdocs, collaborators know something you don’t.
- Take the time to be amazed. Drop everything and explore a new idea.
- Keep learning. There a so many great ideas out there.
- The most valuable tools and algorithms are often extremely simple.

Acknowledgements

- National Institutes of Health
  - NHLBI
  - NHGRI
  - NEI
  - NIMH
- Pew Charitable Trusts
- GlaxoSmithKline
- Wellcome Trust
- University of Michigan

Thank you!
Michigan Team

Work, family, leisure. It’s all life.