Human Genetics Zeitgeist

Goncalo Abecasis

University of Michigan School of Public Health

@gabecasis

Goal of Human Genetic Studies

Find biological processes that,
when changed, alter disease course

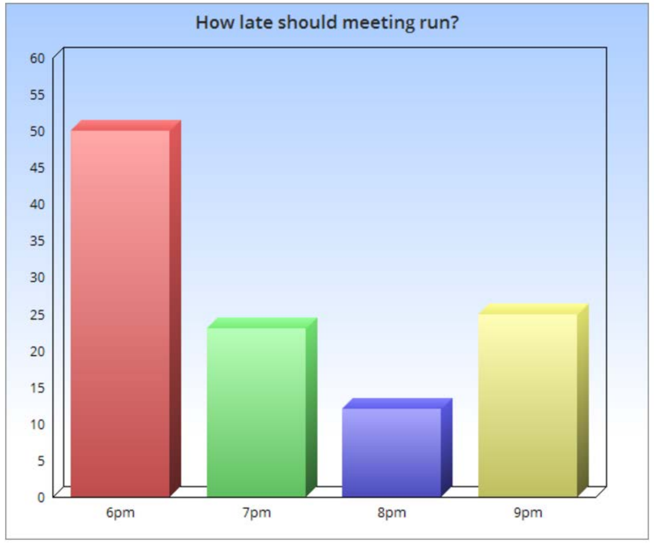
Understand Disease:
Enable new treatments

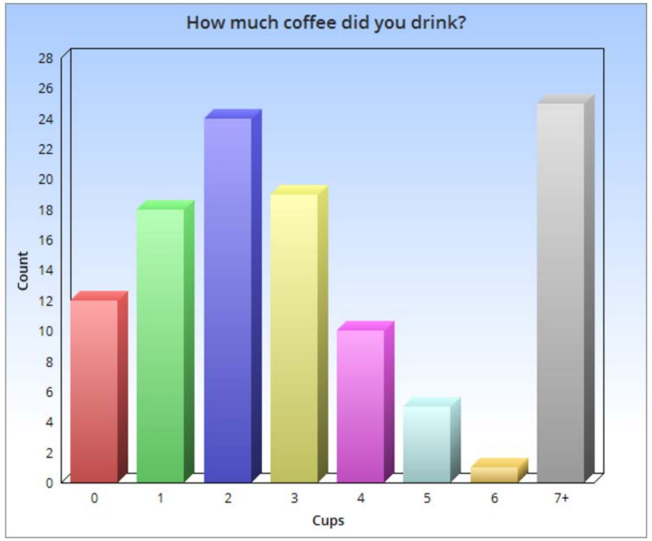
Predict disease:
Enable early prevention, decision making

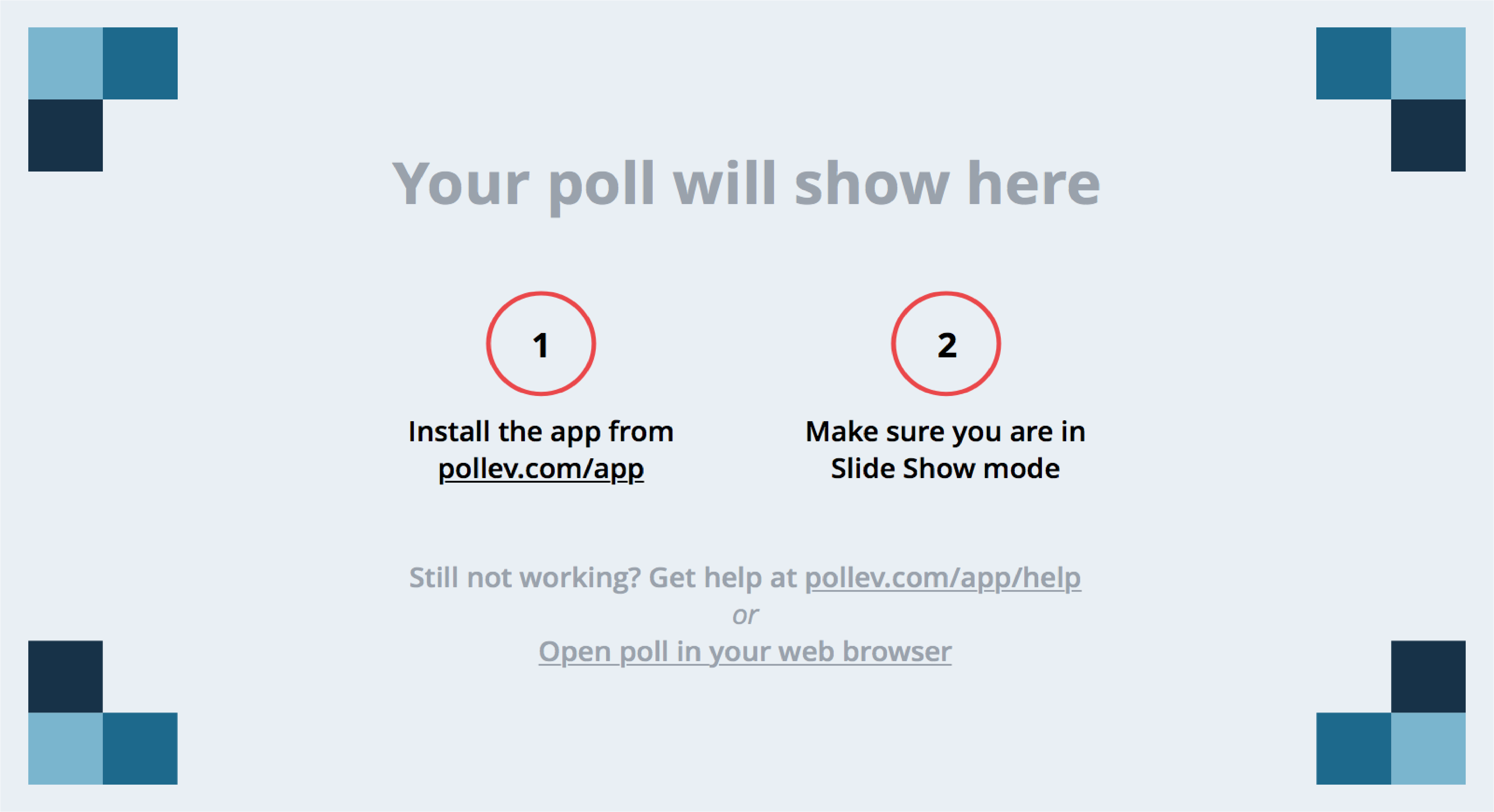
Something to keep you awake.

- For the next bit of the talk, you will see some interspersed polls.

- To participate, go to:
  - On a browser: Go to **pollev.com/topmed**
  - On a phone: Text **topmed to 22333**







Your poll will show here

1

Install the app from pollev.com/app

2

Make sure you are in Slide Show mode

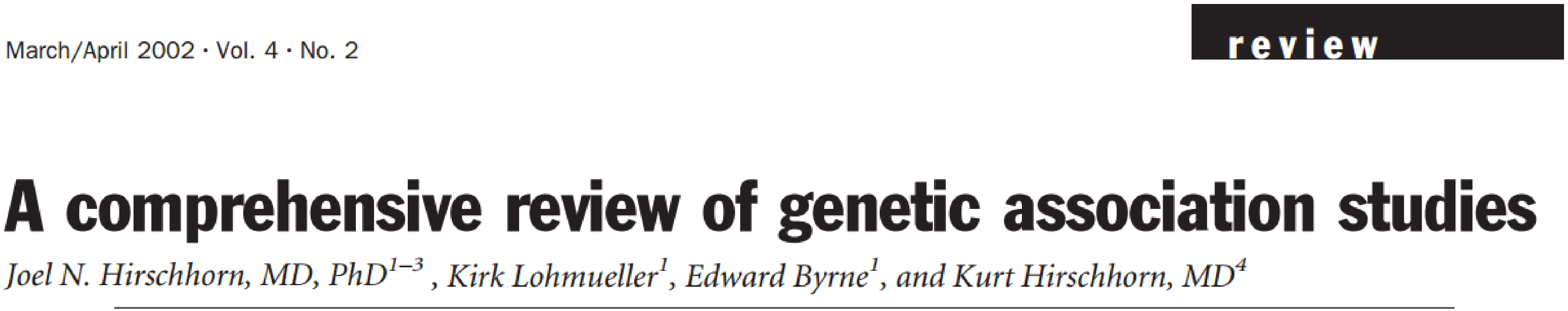
Still not working? Get help at pollev.com/app/help

or

Open poll in your web browser

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)



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)

Your poll will show here

1

Install the app from
pollev.com/app

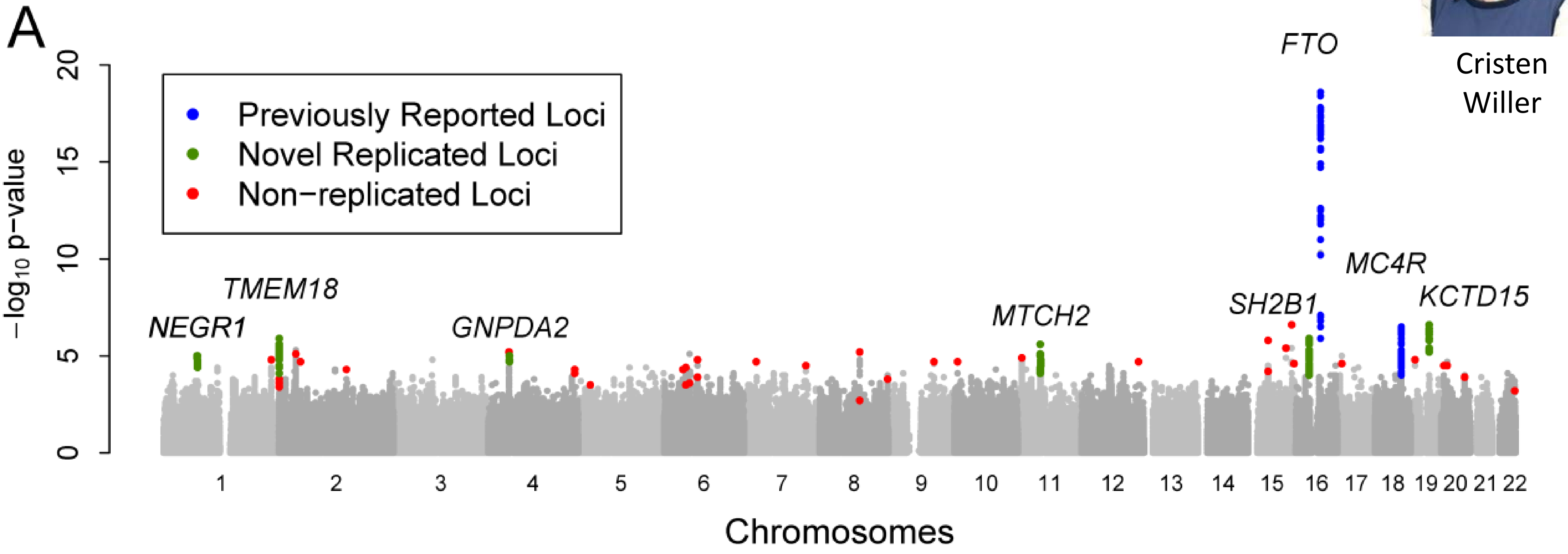
2

Make sure you are in
Slide Show mode

Still not working? Get help at pollev.com/app/help
or
Open poll in your web browser

Search for Genetic Variants Influencing Body Mass

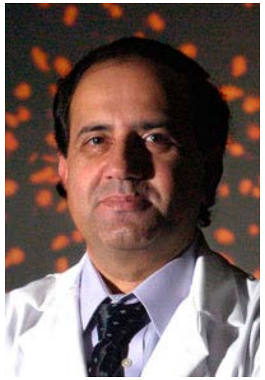
Cristen Willer

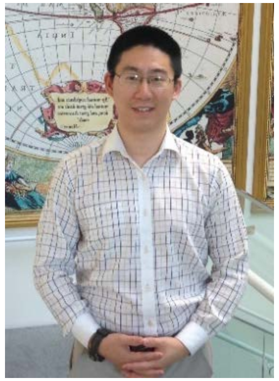


Seven of eight confirmed BMI loci show strongest expression in the brain…

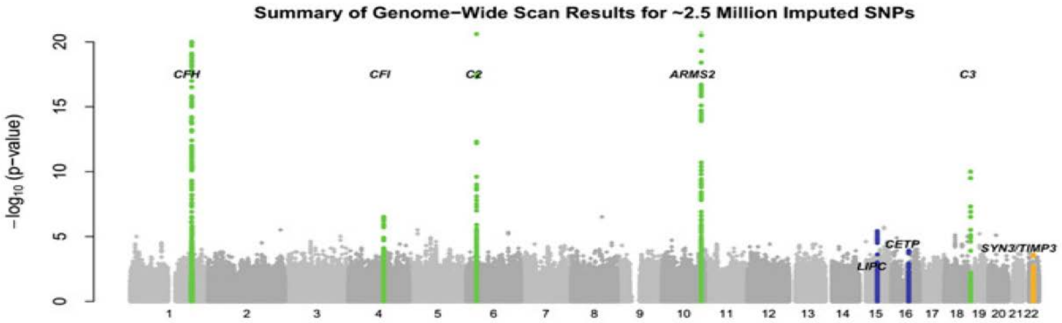
Willer et al, *Nature Genetics,* 2009

Macular Degeneration, 2010 - 2015

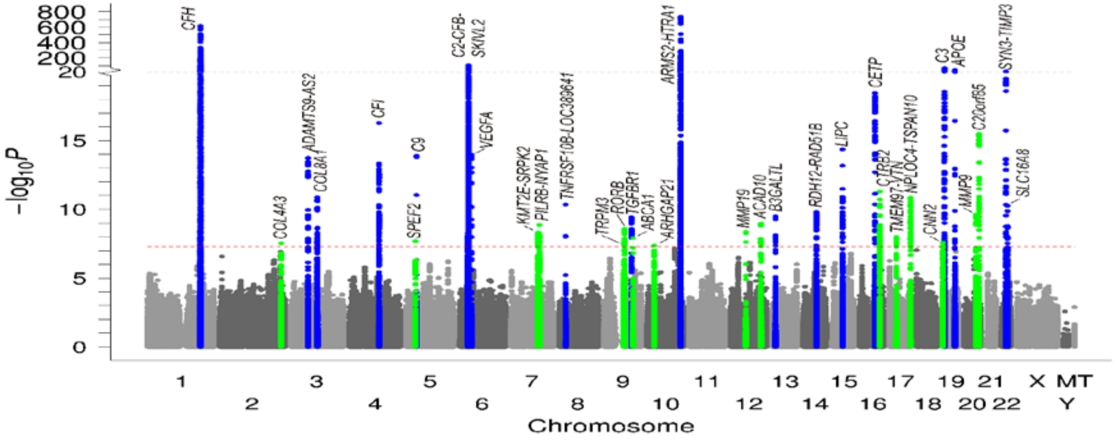




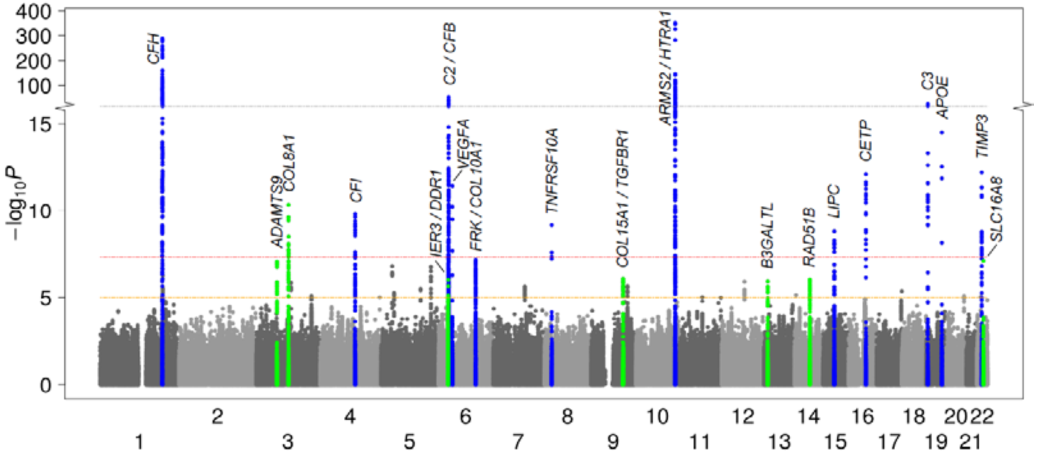
2010



2015



2013

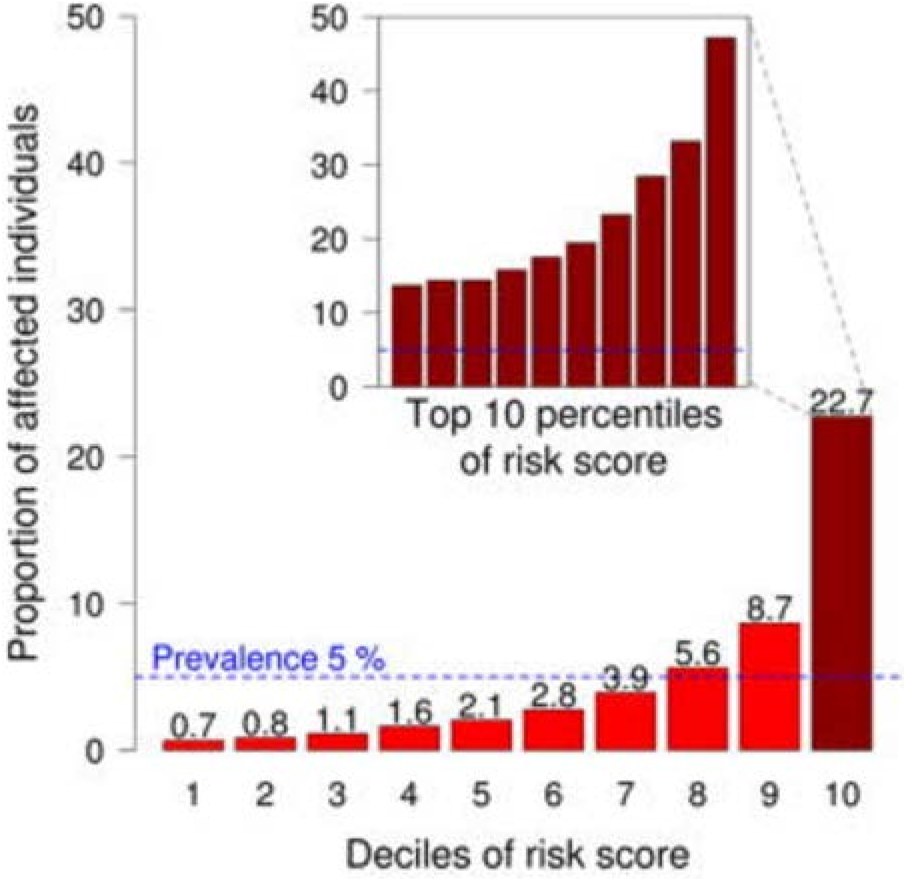


As sample sizes grow, we have more loci.

How do we enable more scientists to explore data and explore big questions?

By default, easy to spend a lot of energy on the basics of data aggregation and initial analysis.

Combined Effects of Many Alleles Strongly Predict Risk (2015)



Low risk → High risk

Your poll will show here

1

Install the app from
pollev.com/app

2

Make sure you are in
Slide Show mode

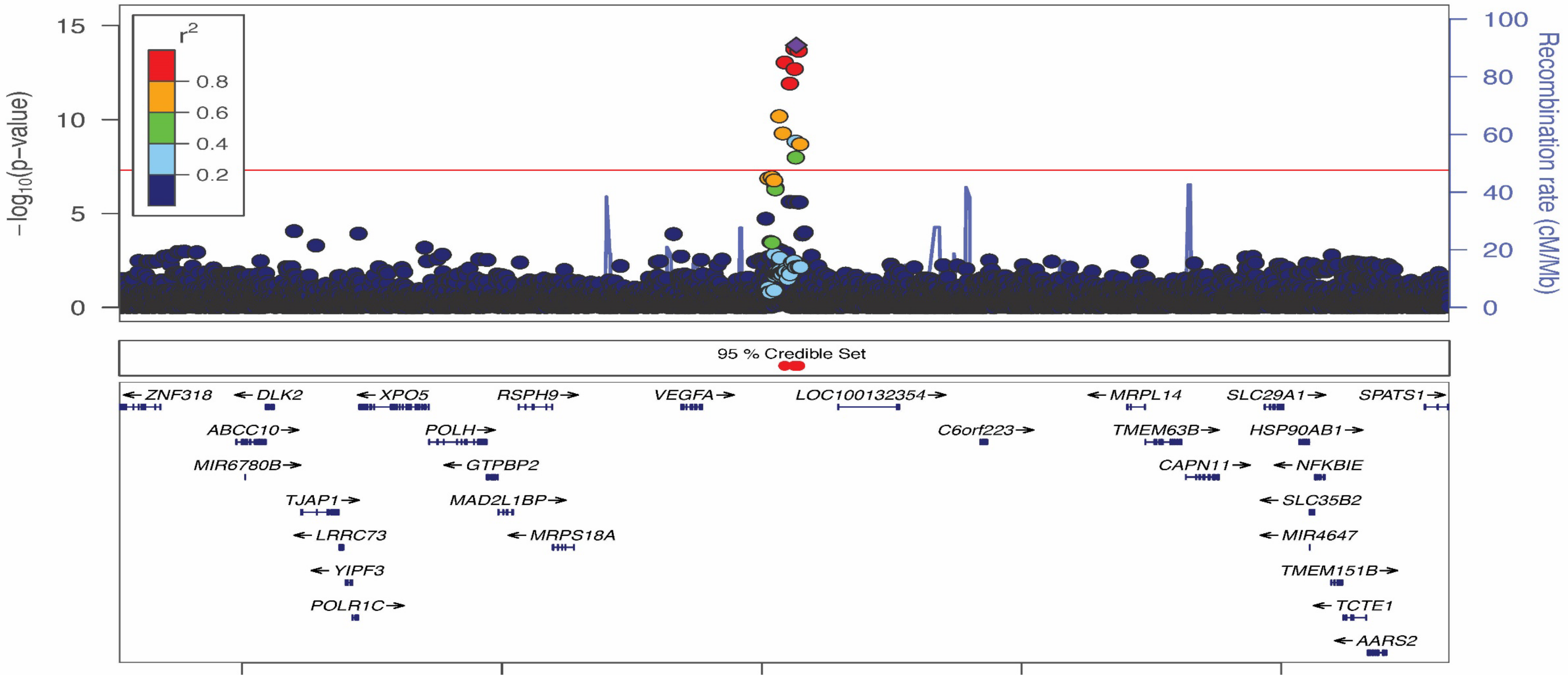
Still not working? Get help at pollev.com/app/help

or

Open poll in your web browser

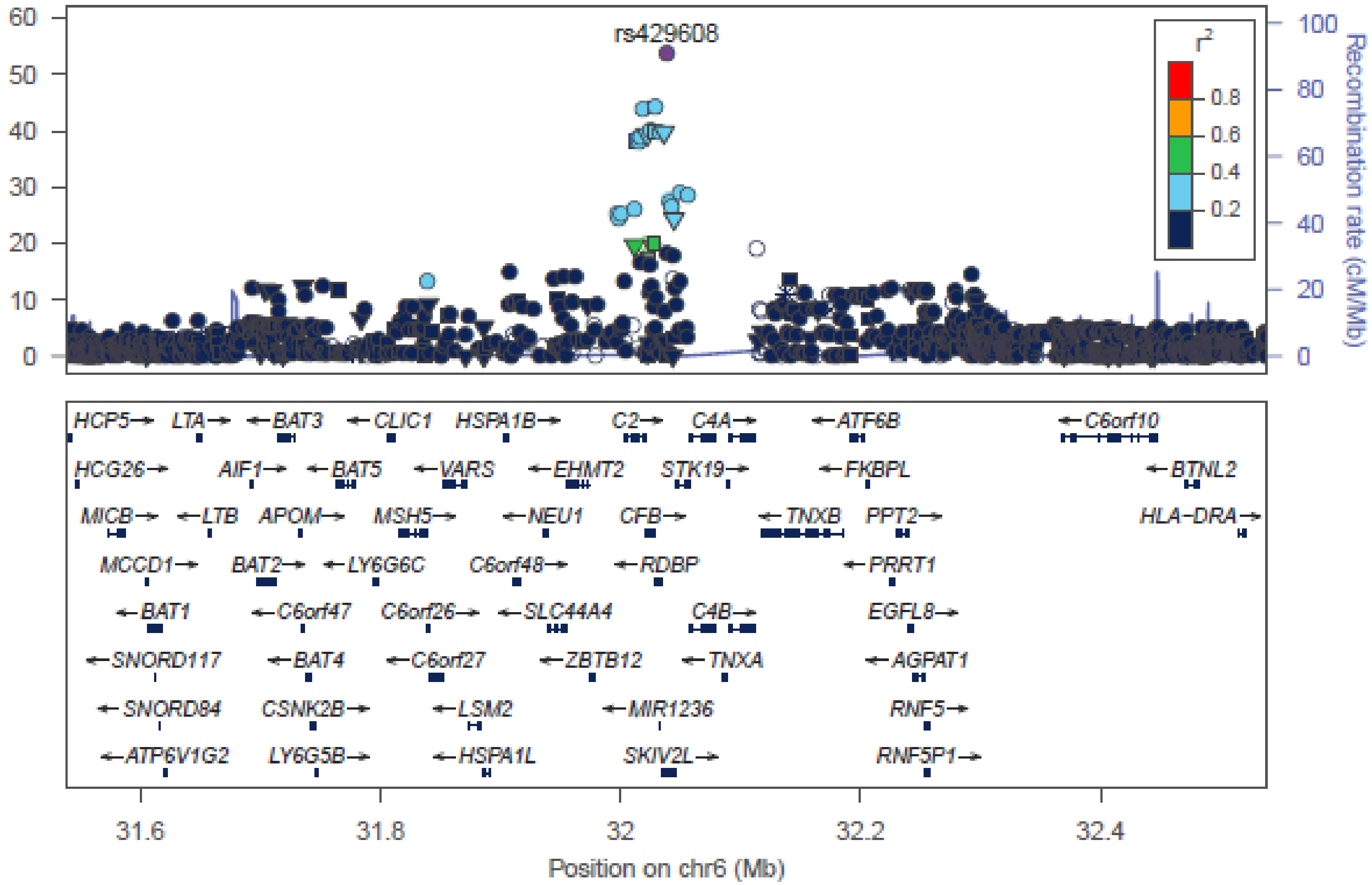
Age-Related Macular Degeneration:

Close-up near *VEGFA*



Age Related Macular Degeneration:

Close-Up of Specific Region

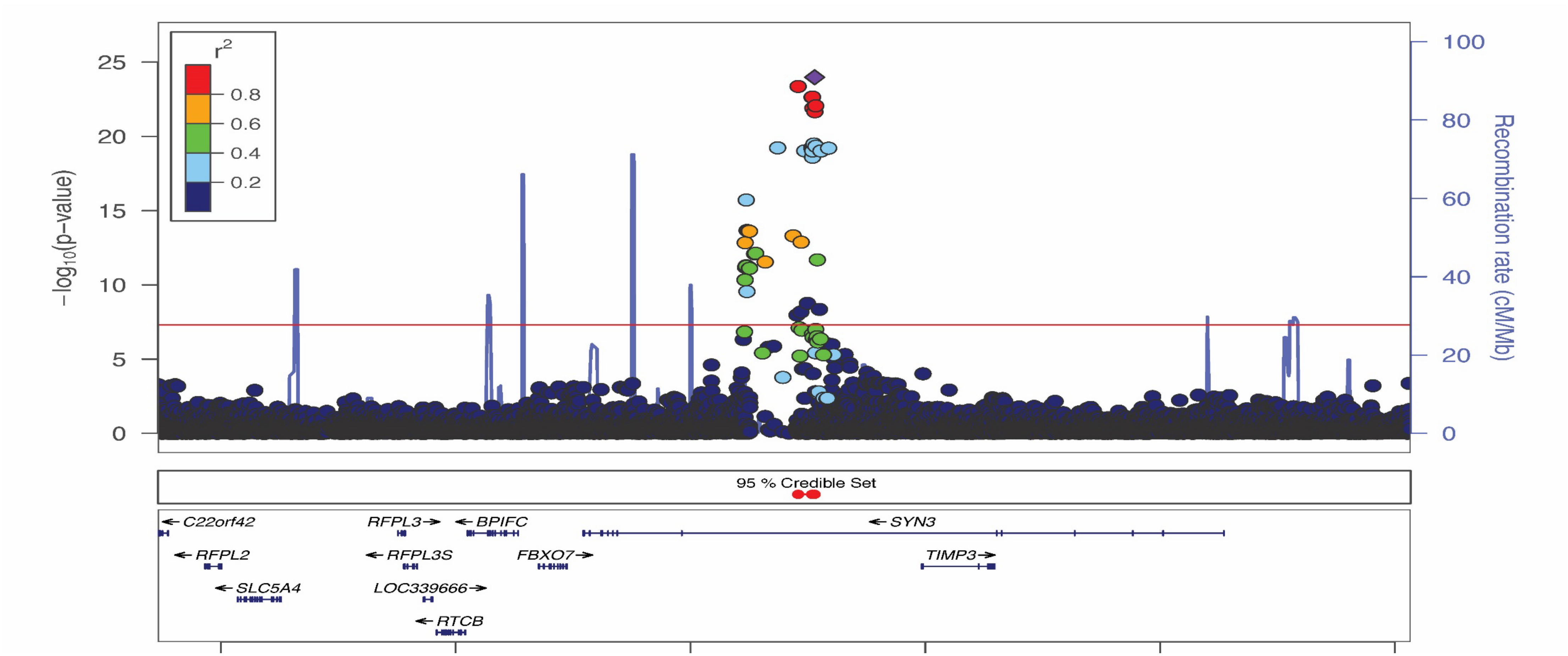


What do AMD Associated Variants Have in Common?

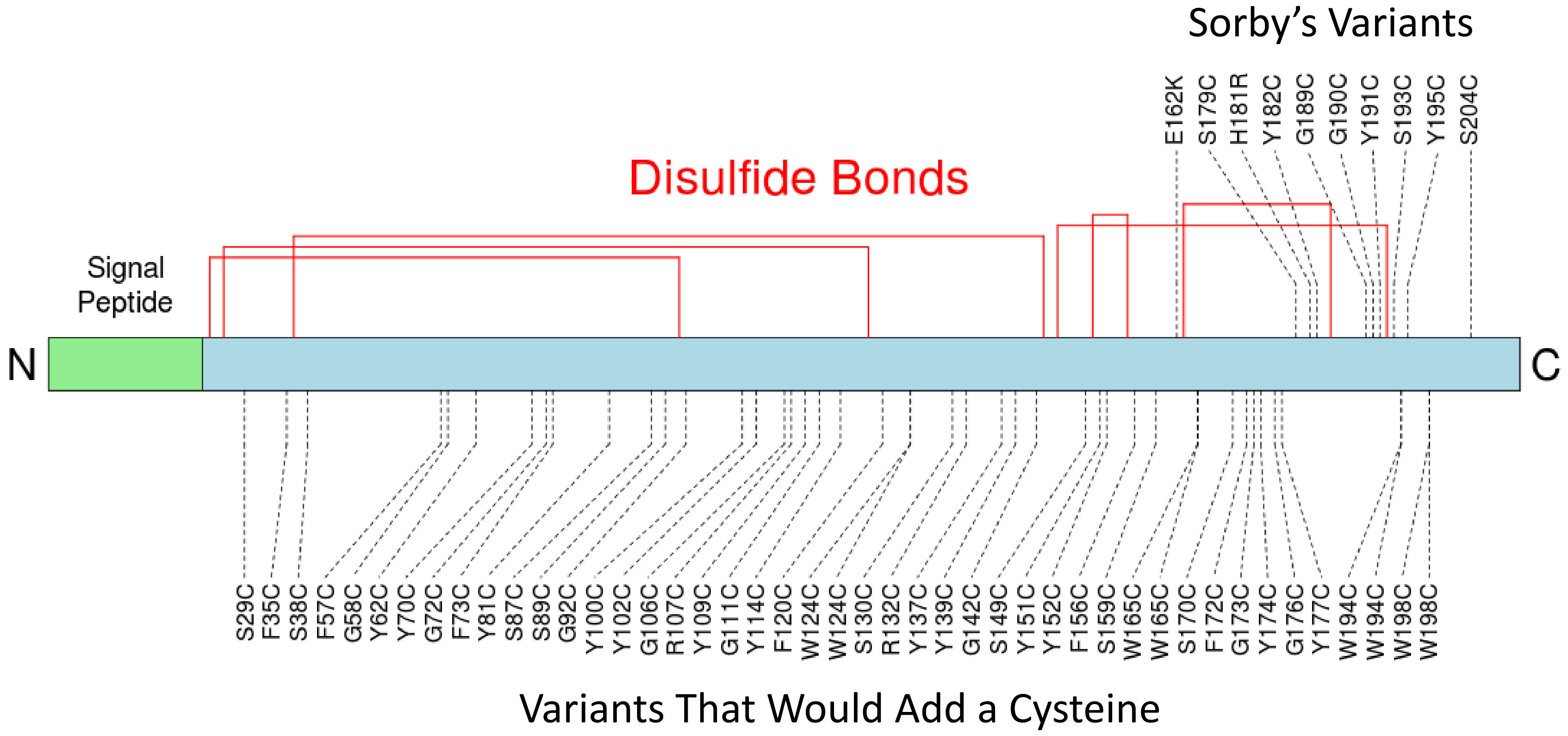
	Enrichment
Protein Coding	21x
Near Complement Gene	183x
Deleterious (CADD >20)	255x

Age-Related Macular Degeneration:

Another Close-up, Chromosome 22



TIMP3 Variants on AMD Array



8 of 10 Sorby’s mutations cause unpaired cysteine residues

→ Polymerization of TIMP3 protein

→ Accumulation in extracellular matrix

Your poll will show here

1

Install the app from pollev.com/app

2

Make sure you are in Slide Show mode

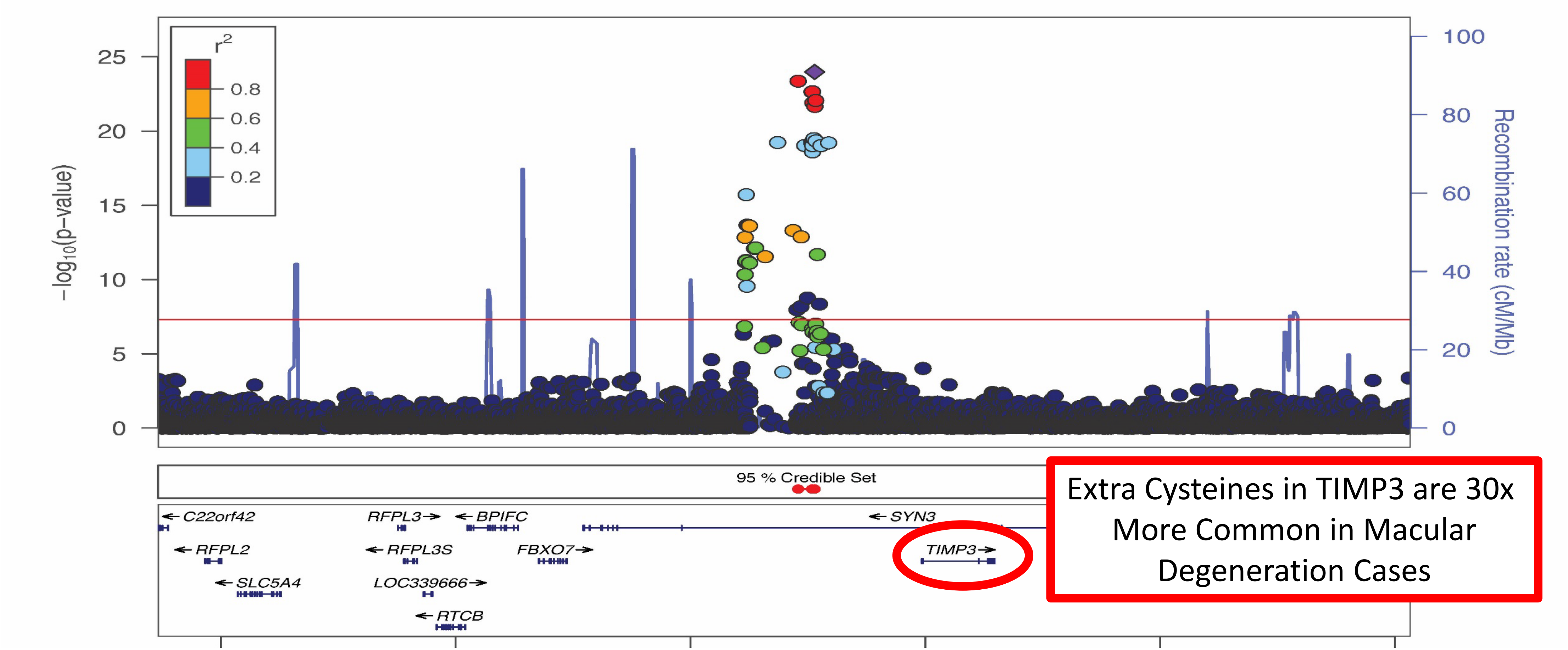
Still not working? Get help at pollev.com/app/help

or

Open poll in your web browser

Age-Related Macular Degeneration:

Another Close-up, Chromosome 22



Extra Cysteines in TIMP3 are 30x More Common in Macular Degeneration Cases

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 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 Available: Three Use Cases

- 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

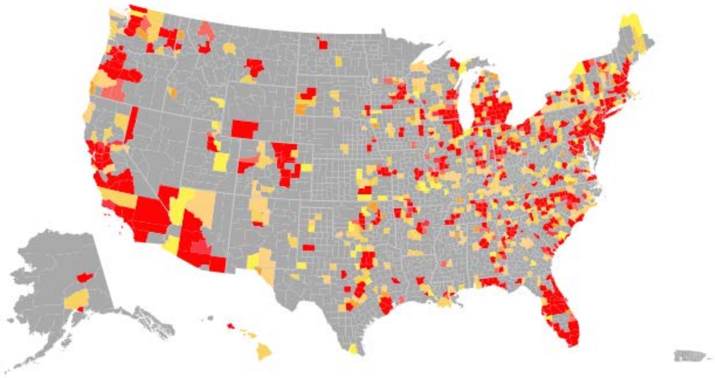
- Three different needs and expectations
  - We should try to make some of these uses as close to zero friction as possible

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

Part I – Genes for Good

GENES for GOOD

1
QUESTIONNAIRE:
2
3
PROTECTIONS
DNA SAMPLE #273610
4
DAILY NEWS
ONLINE STUDY LEADS TO BIG BREAKTHROUGH MEDICAL FINDING

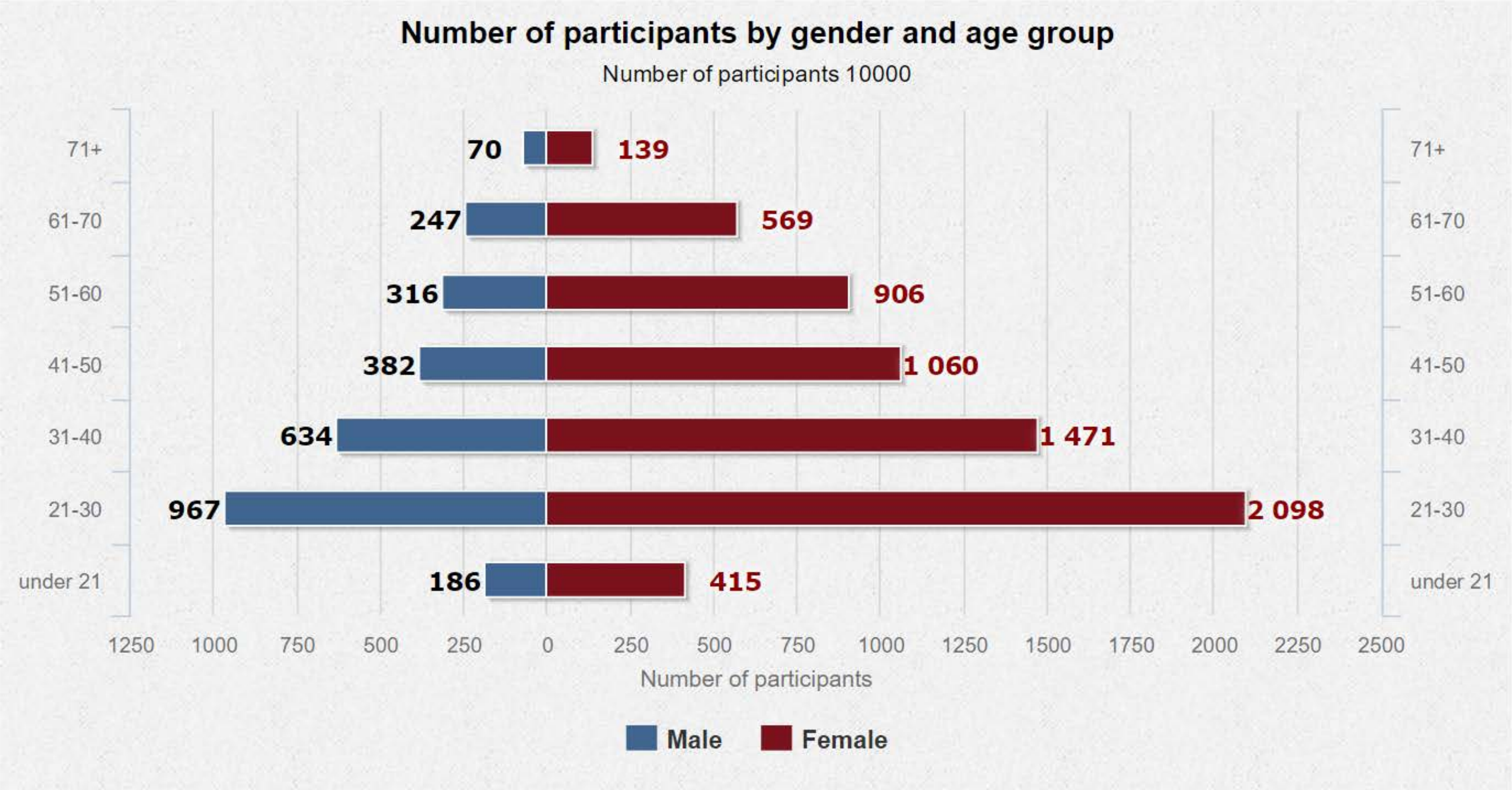


- Exploring new ways to engage populations in research
- Continuous Engagement, Web, Mobile Devices
- Currently, >25,000 participants
- www.genesforgood.org

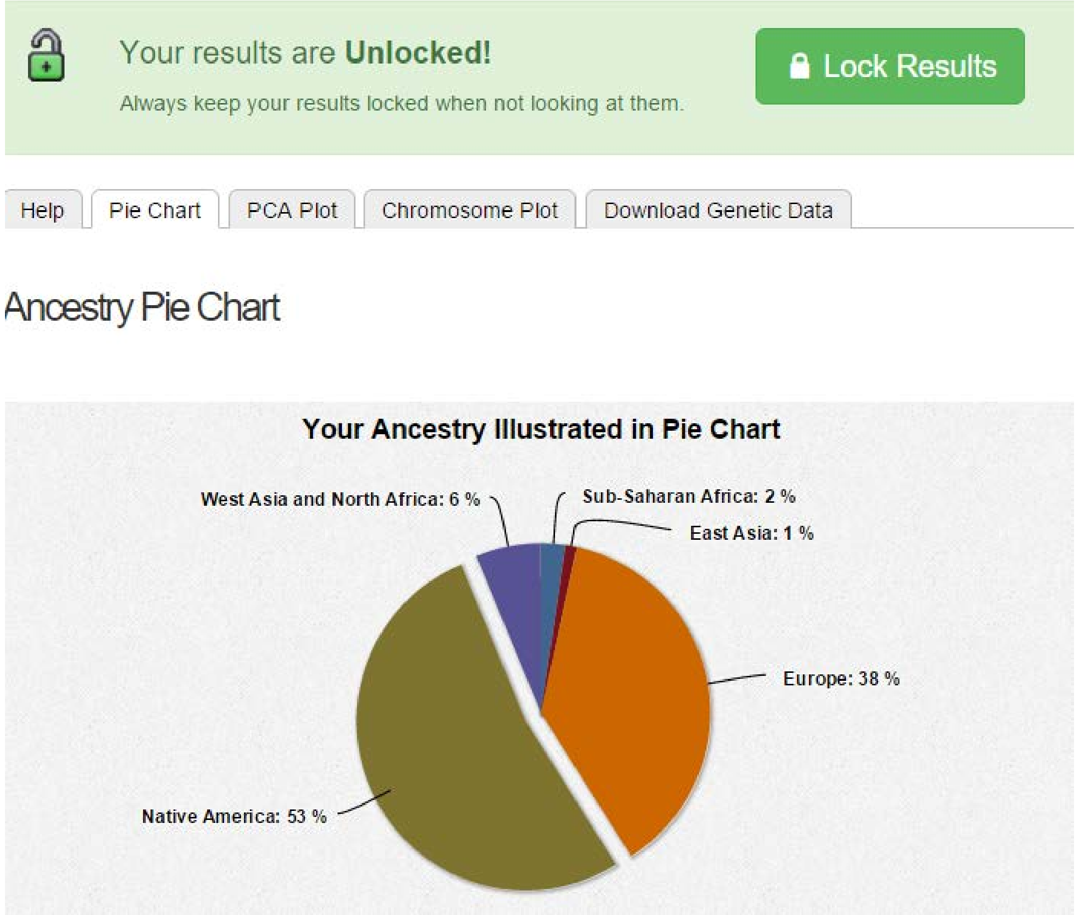
Genes for Good - Geographical Distribution of App Usage

2015-01-01 Participants: 0 Completed Surveys: 0

10,000 Participants…



Return of Results

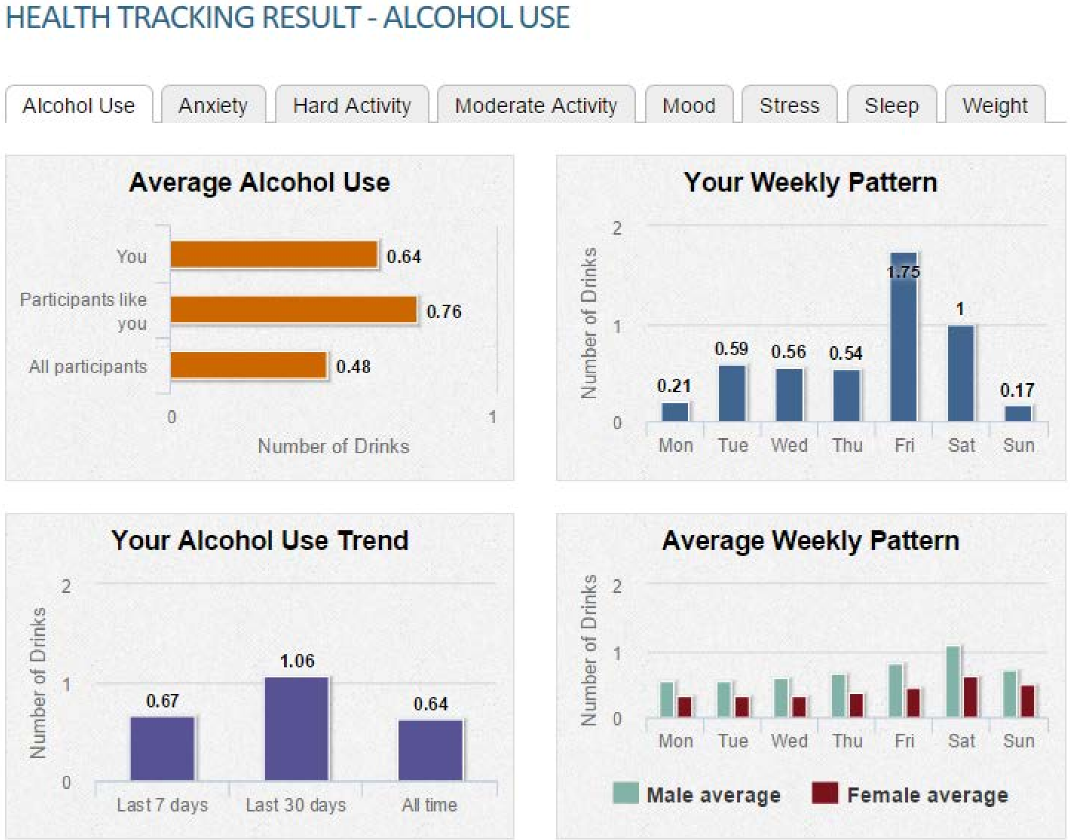


Your results are **Unlocked!**

Always keep your results locked when not looking at them.

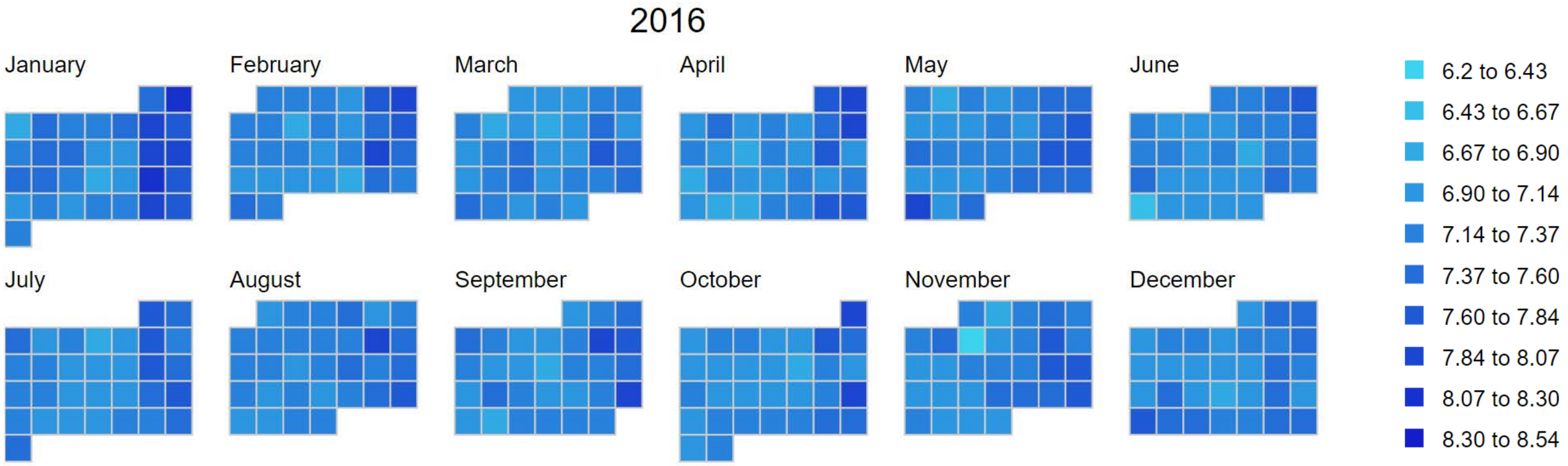
Ancestry Pie Chart

Your Ancestry Illustrated in Pie Chart
West Asia and North Africa: 6 %
Sub-Saharan Africa: 2 %
East Asia: 1 %
Europe: 38 %
Native America: 53 %

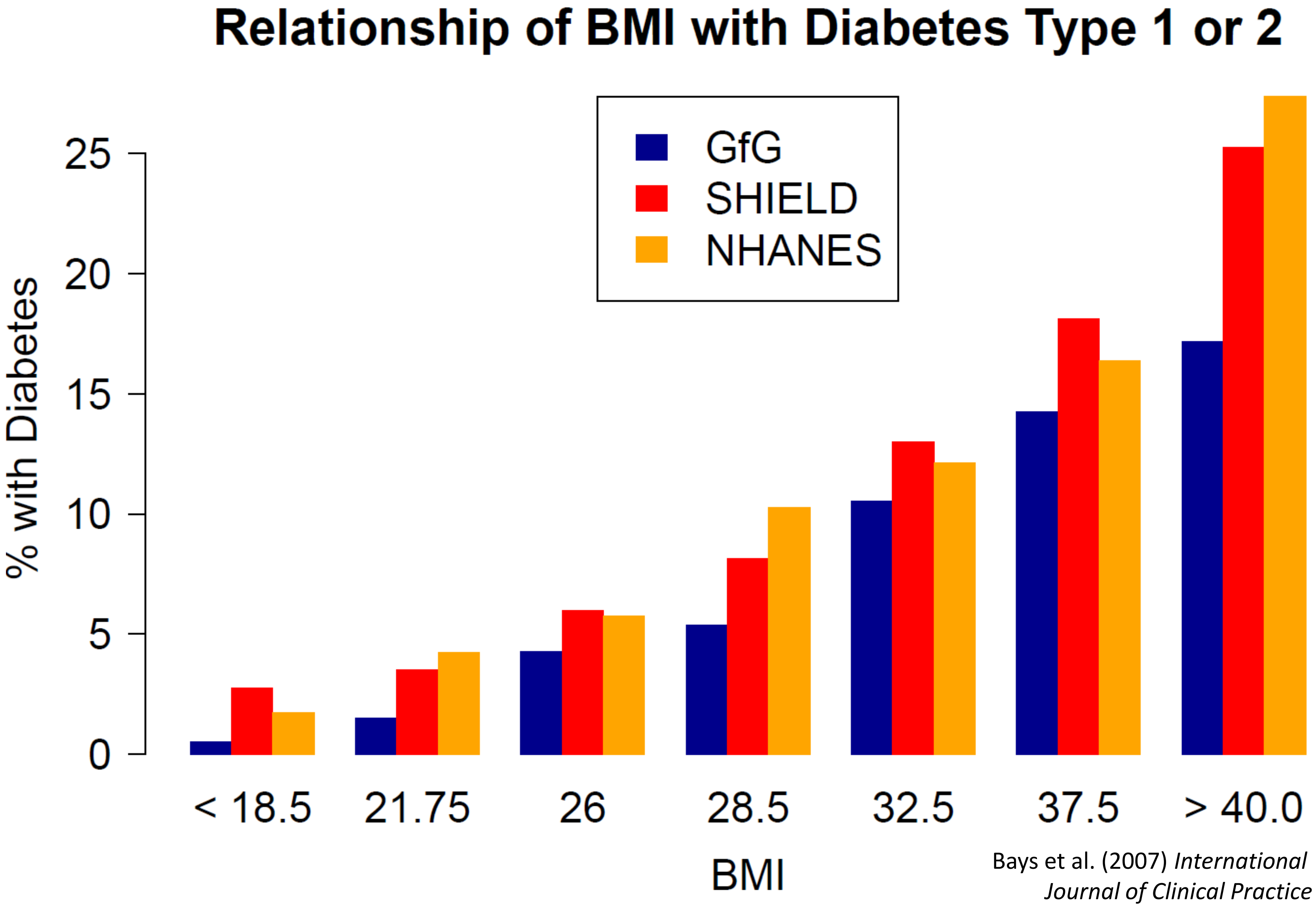


HEALTH TRACKING RESULT - ALCOHOL USE

Average Reported Sleep Hours Over One Year



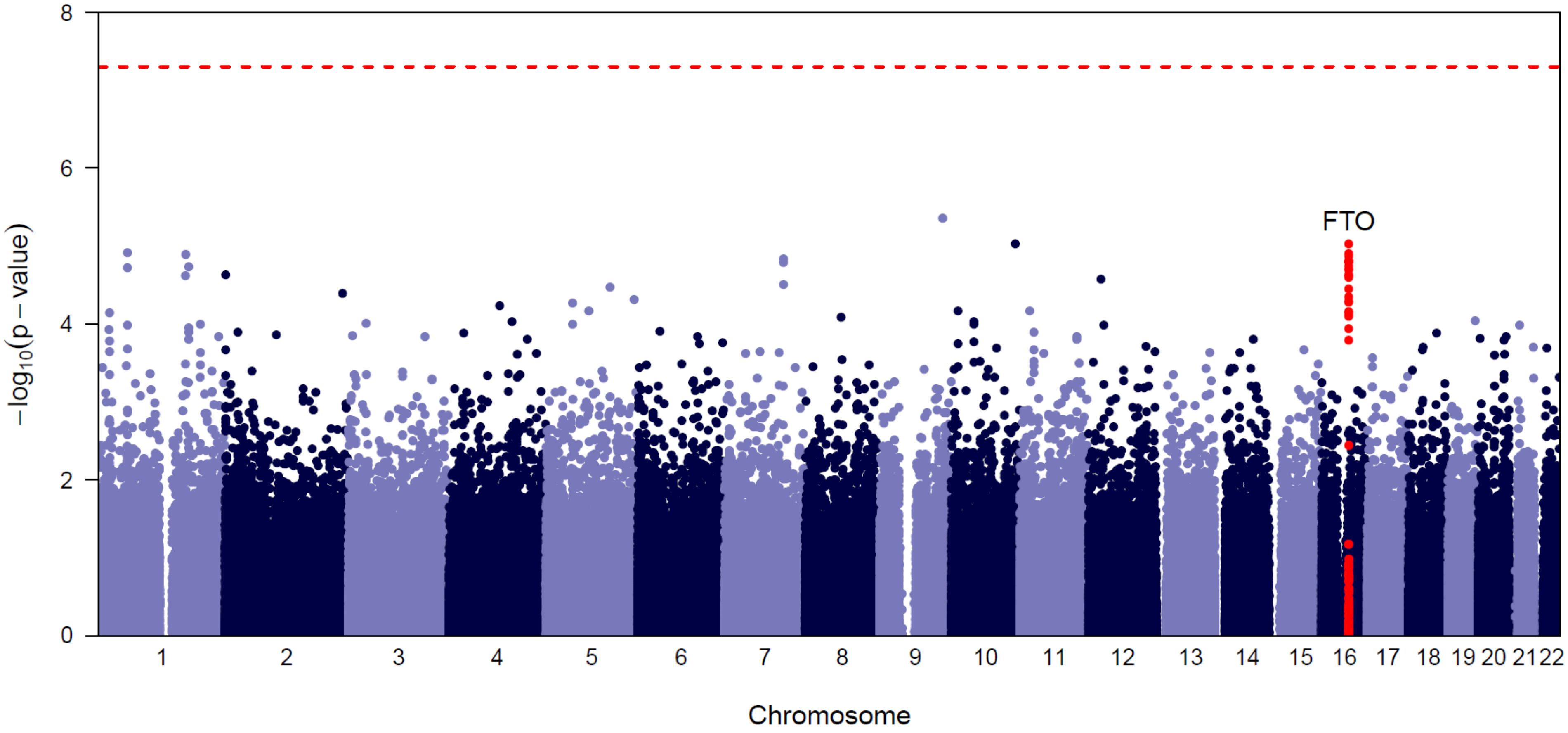
BMI, Age & Diabetes



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

Results: BMI GWAS

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*

Your poll will show here

1

Install the app from
pollev.com/app

2

Make sure you are in
Slide Show mode

Still not working? Get help at pollev.com/app/help

or

Open poll in your web browser

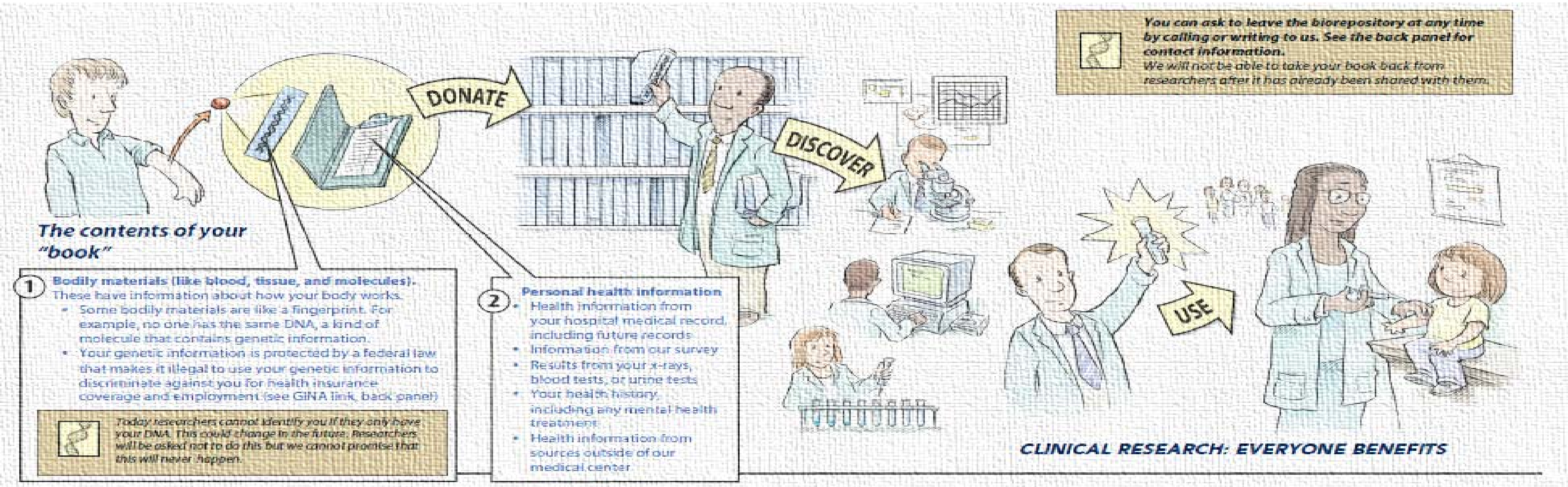
How Can We Engage 10,000s of Research Participants?

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)**



You can ask to leave the biorepository at any time by calling or writing to us. See the back panel for contact information.
We will not be able to take your book back from researchers after it has already been shared with them.

The contents of your "book"

1 **Bodily materials (like blood, tissue, and molecules).** These have information about how your body works.

- Some bodily materials are like a fingerprint. For example, no one has the same DNA, a kind of molecule that contains genetic information.
- Your genetic information is protected by a federal law that makes it illegal to use your genetic information to discriminate against you for health insurance coverage and employment (see GINA link, back panel)

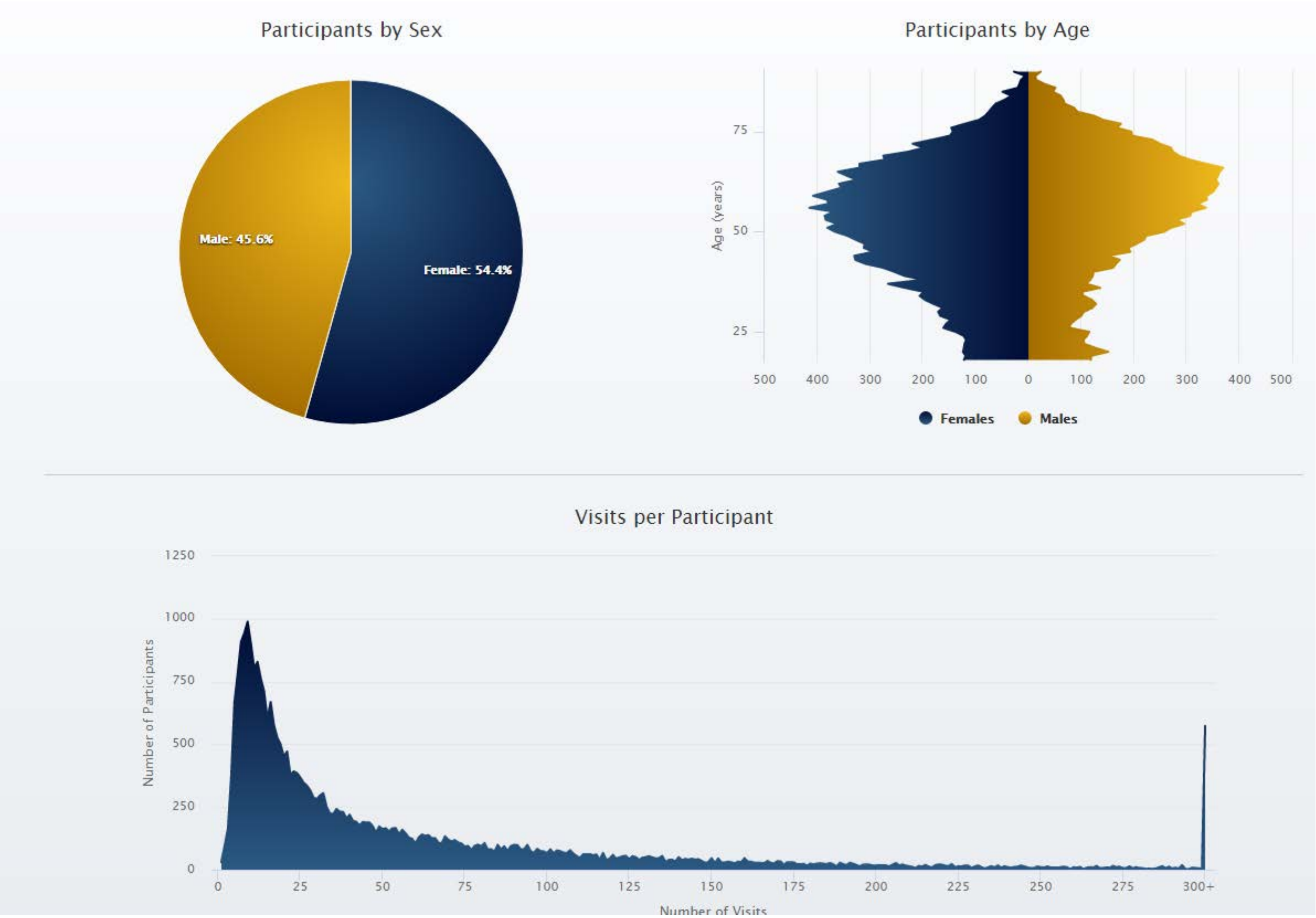
Today researchers cannot identify you if they only have your DNA. This could change in the future. Researchers will be asked not to do this but we cannot promise that this will never happen.

2 **Personal health information**

- Health information from your hospital medical record, including future records
- Information from our survey
- Results from your x-rays, blood tests, or urine tests
- Your health history, including any mental health treatment
- Health information from sources outside of our medical center

MGI demographics

http://www.michigangenomics.org/

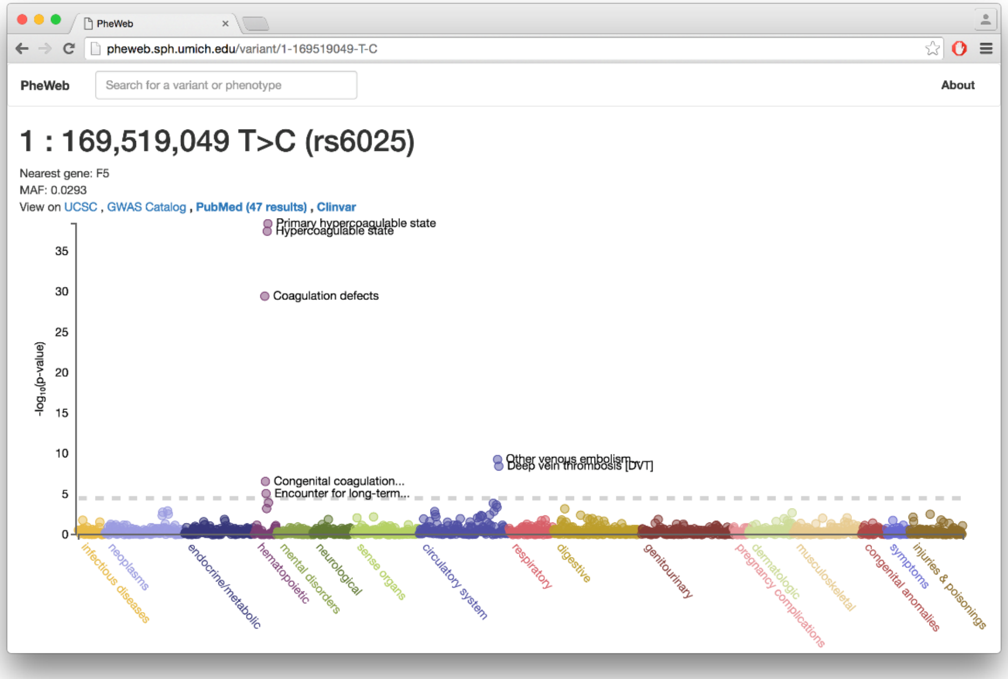


Disease	%
Hypertension	28%
Obesity	24%
Arrhythmias	22%
Sleep Apnea	12%
Skin Cancer	12%
Asthma	8%
Cystic Fibrosis	0.1%

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

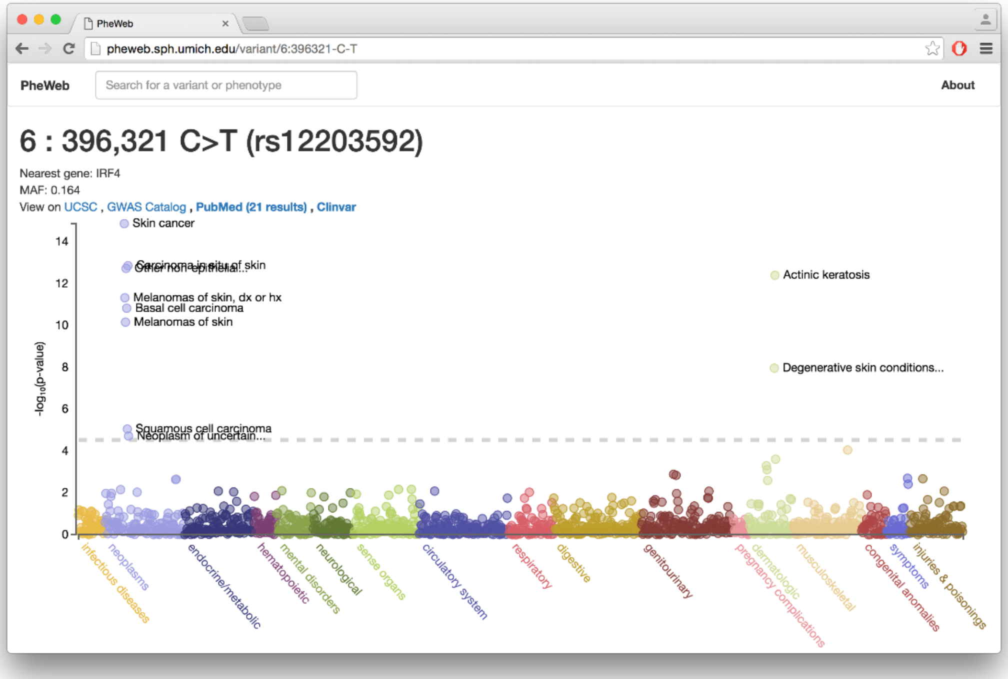
Ellen
Schmidt

Peter
VandeHaar



1 : 169,519,049 T>C (rs6025)

Nearest gene: F5
MAF: 0.0293
View on UCSC , GWAS Catalog , PubMed (47 results) , Clinvar

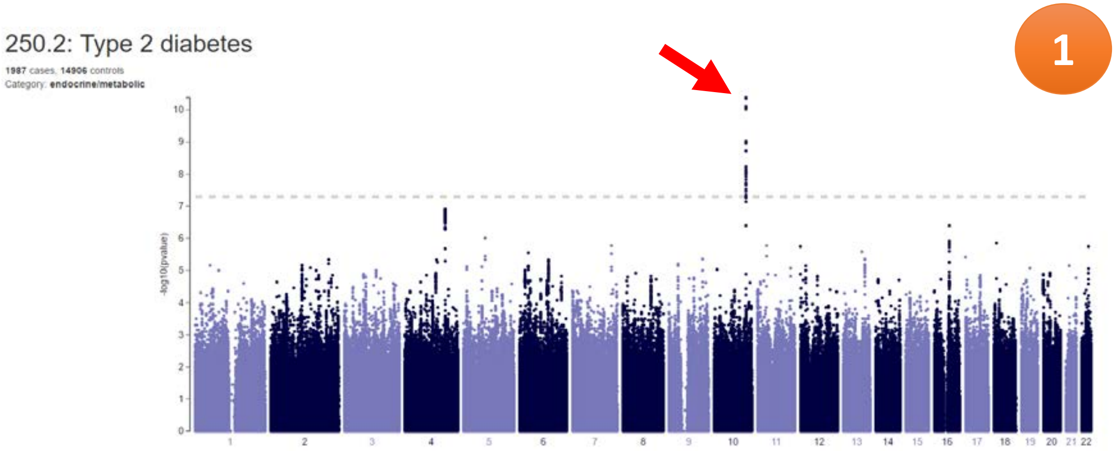


6 : 396,321 C>T (rs12203592)

Nearest gene: IRF4
MAF: 0.164
View on UCSC , GWAS Catalog , PubMed (21 results) , Clinvar

Michigan Genomics Initiative Association Statistics
http://pheweb.sph.umich.edu

1



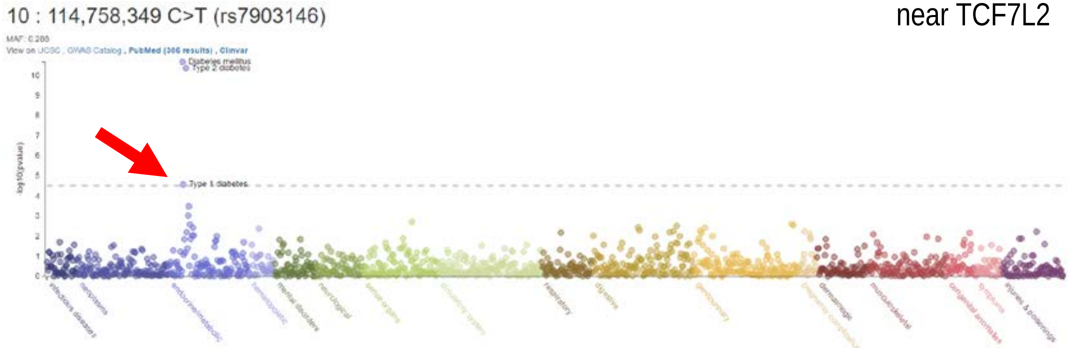
250.2: Type 2 diabetes

1987 cases, 14906 controls

Category: endocrine/metabolic

2

near TCF7L2



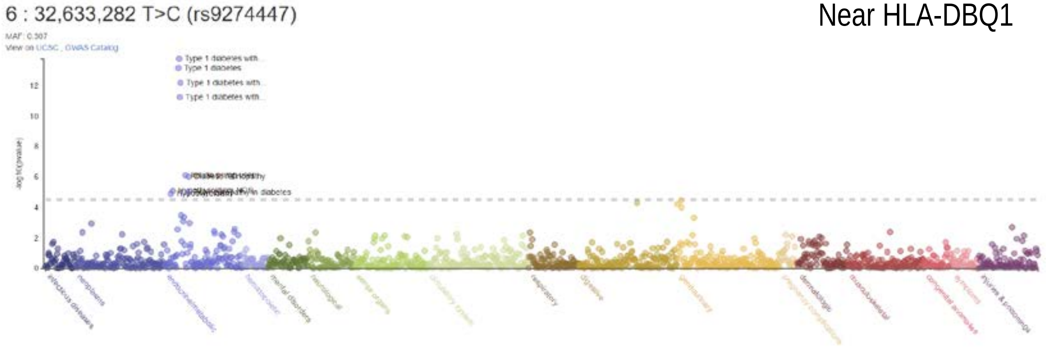
10 : 114,758,349 C>T (rs7903146)

MAF: 0.288

View on UCSC , GWAS Catalog , PubMed (366 results) , Clinvar

4

Near HLA-DBQ1

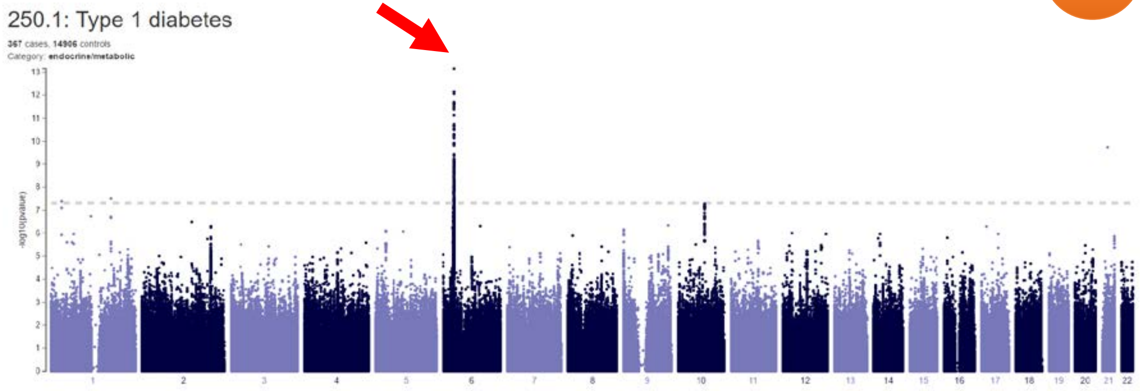


6 : 32,633,282 T>C (rs9274447)

MAF: 0.307

View on UCSC , GWAS Catalog

3

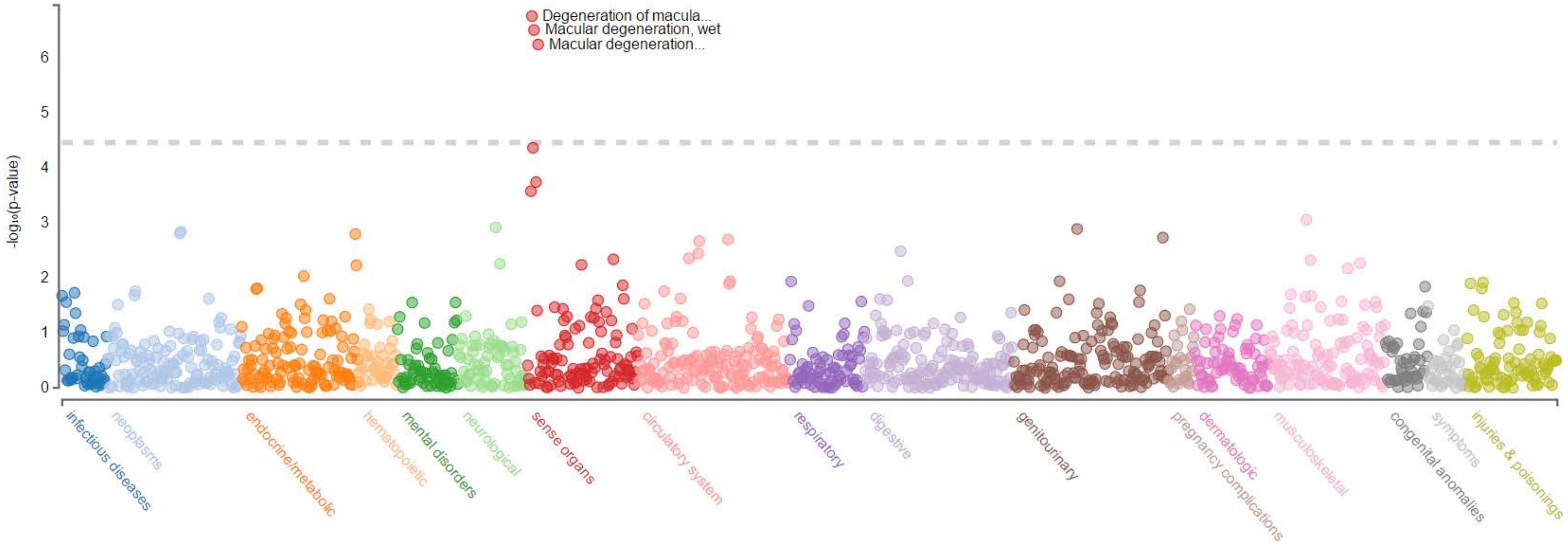


250.1: Type 1 diabetes

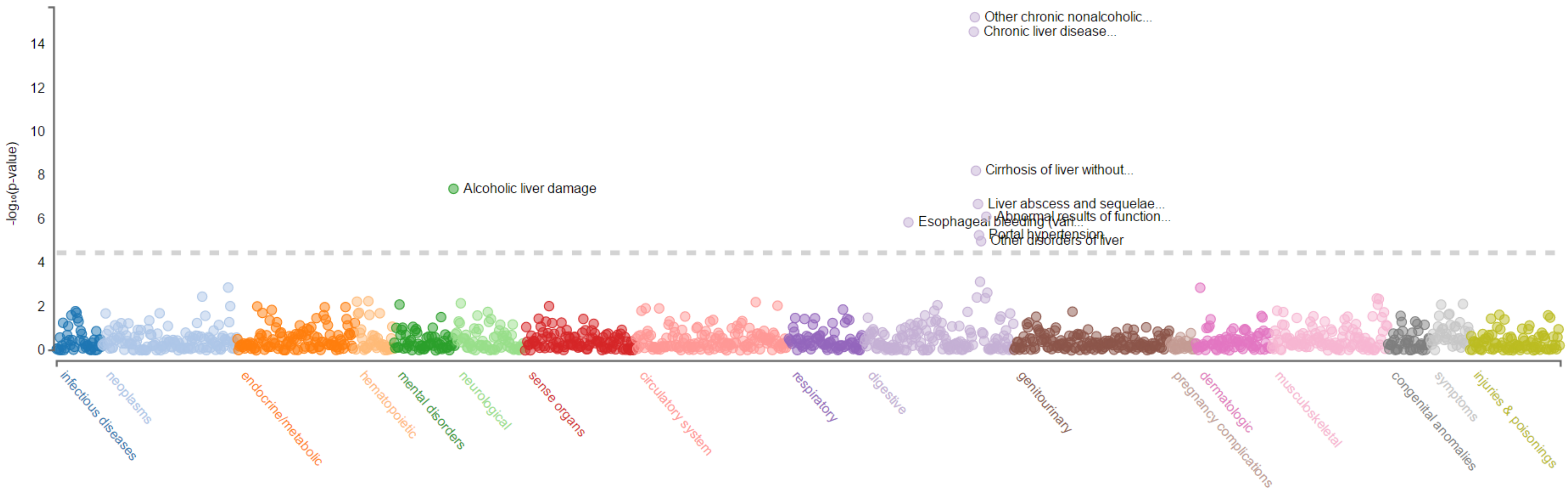
367 cases, 14906 controls

Category: endocrine/metabolic

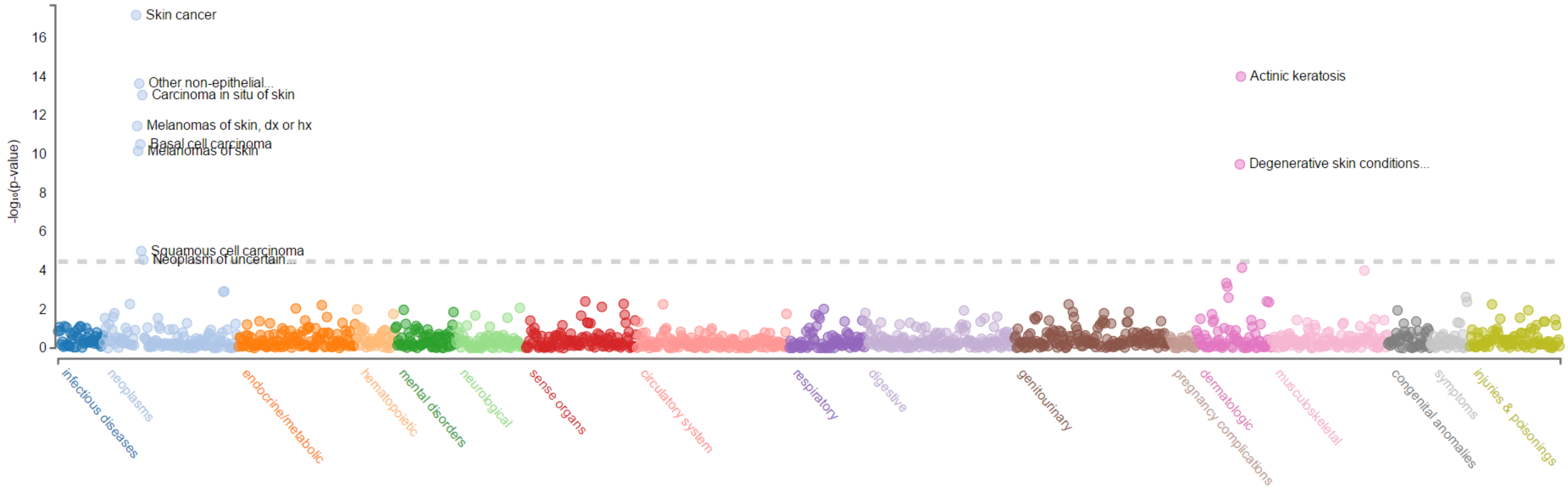
I heard rs10490924 in ARMS2 is associated with macular degeneration …



I heard rs738409 in PNPLA3 is associated with liver disease …



I heard rs12203592 in IRF4 is associated with freckling, skin color …

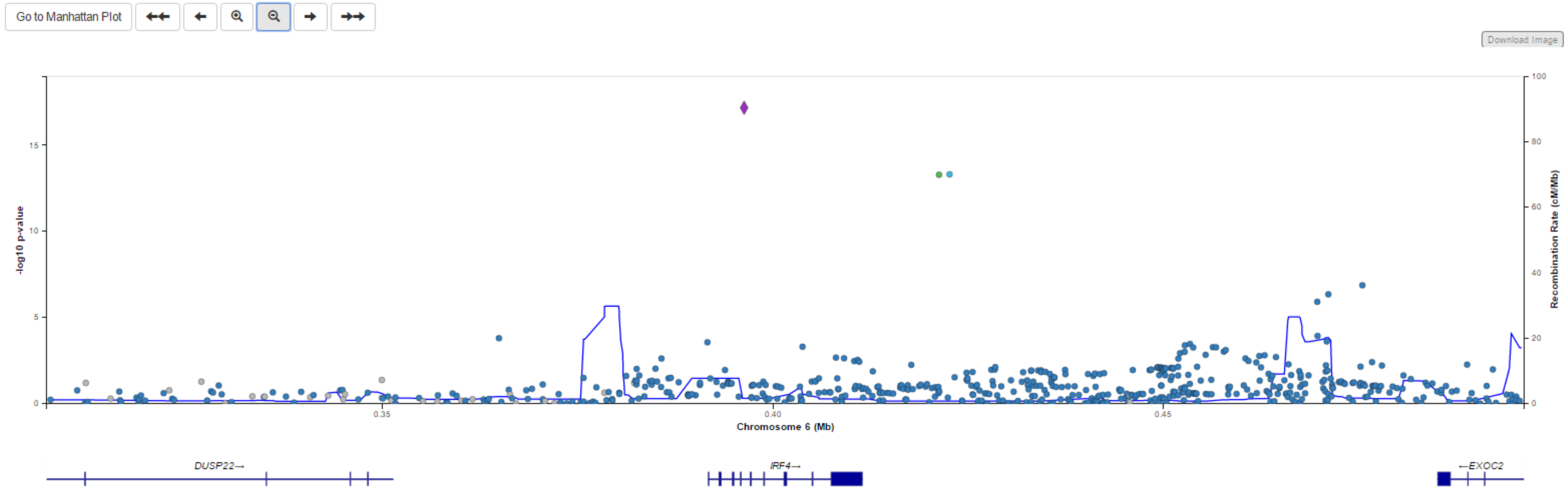


What signals map near IRF4?

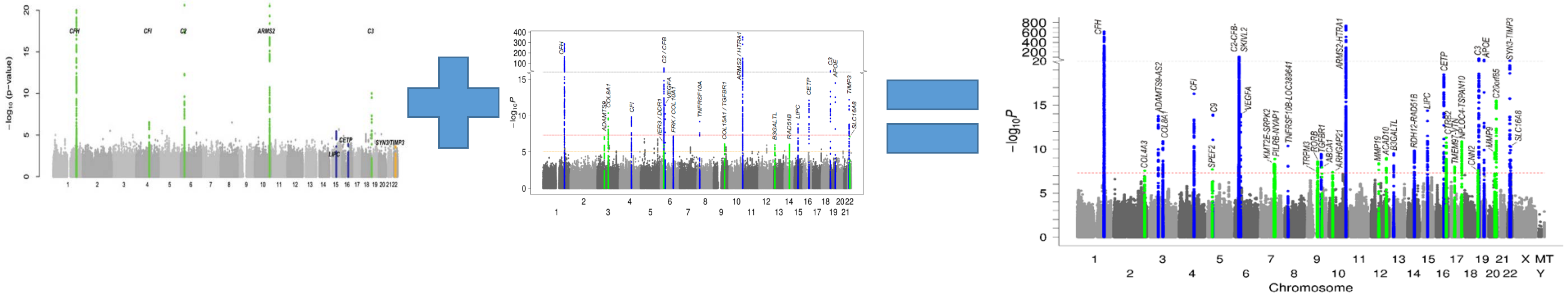
Strongest association is with “Skin Cancer”

Other signals in this locus …

Top p-value	Phenotype	Top Variant
6.7e-18	Skin cancer	rs12203592
9.8e-15	Actinic keratosis	rs12203592
3.1e-11	Basal cell carcinoma	rs12203592
6.7e-11	Melanomas of skin	rs12203592
3.3e-10	Degenerative skin conditions and other dermatoses	rs12203592



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

Your poll will show here

1

Install the app from pollev.com/app

2

Make sure you are in Slide Show mode

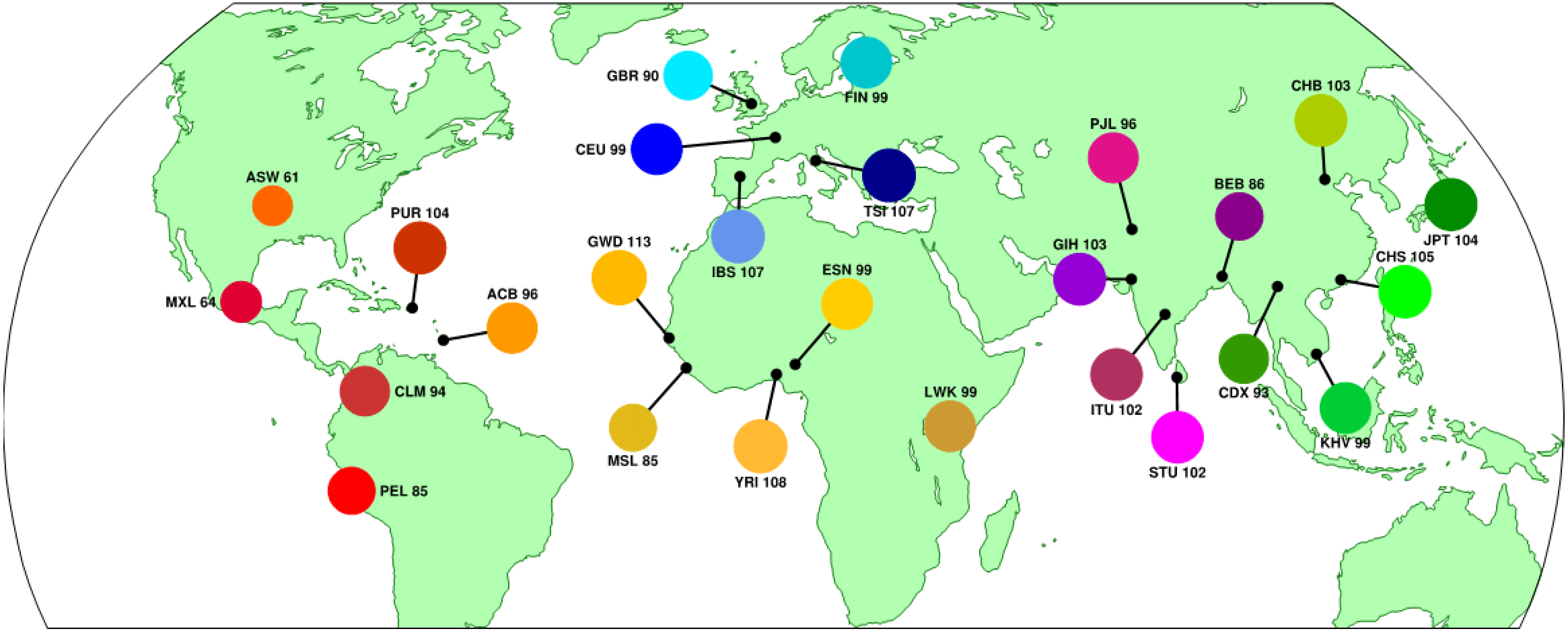
Still not working? Get help at pollev.com/app/help

or

Open poll in your web browser

The Scale of Genetic Data

The 1000 Genomes Project (2008 – 2015)



Shotgun Sequence Data

★

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

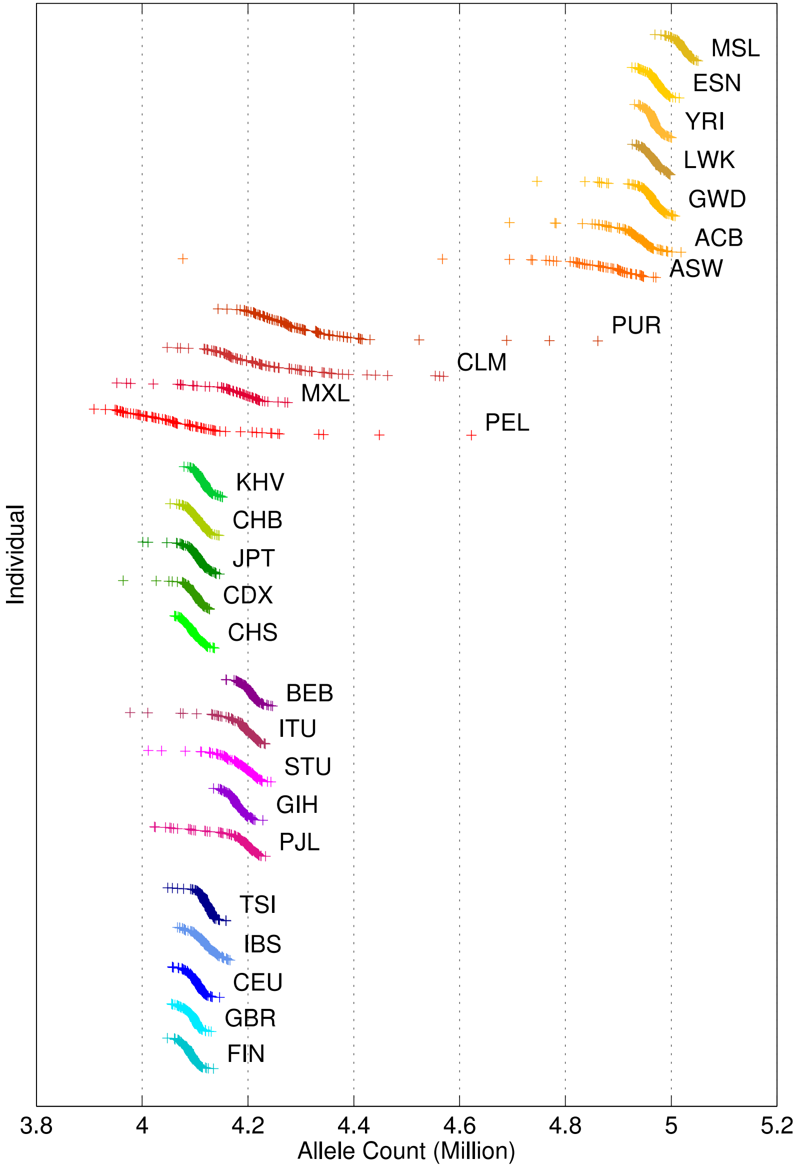
Reference Genome

A/C

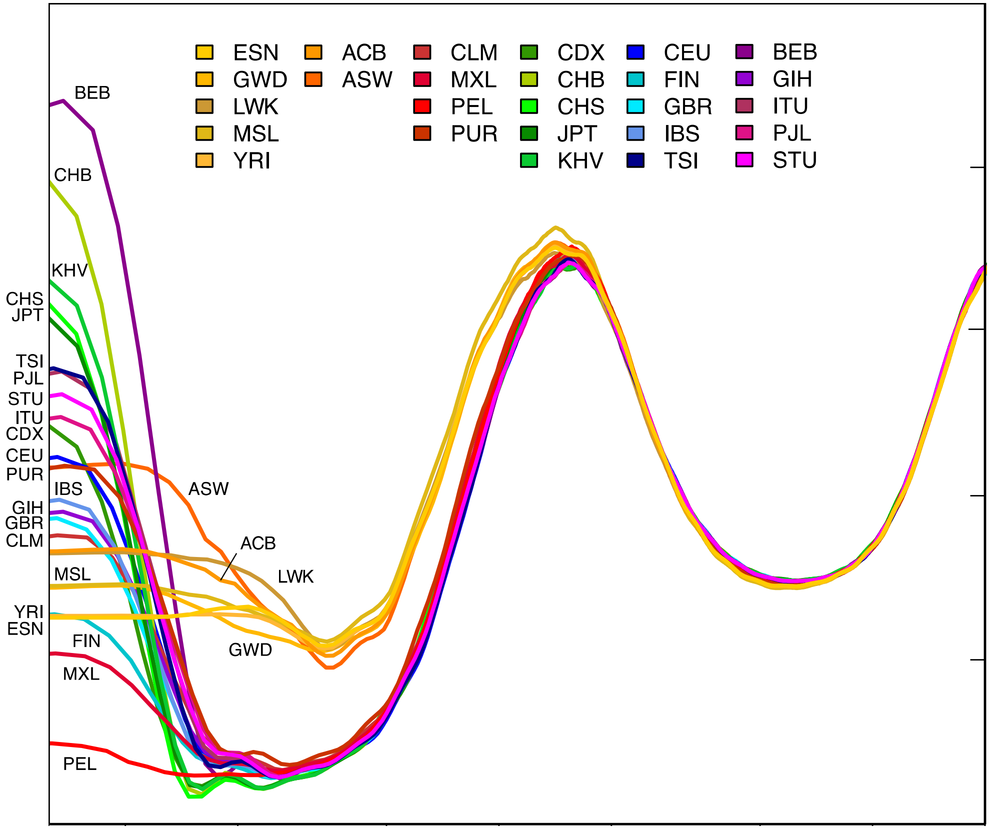
Predicted Genotype

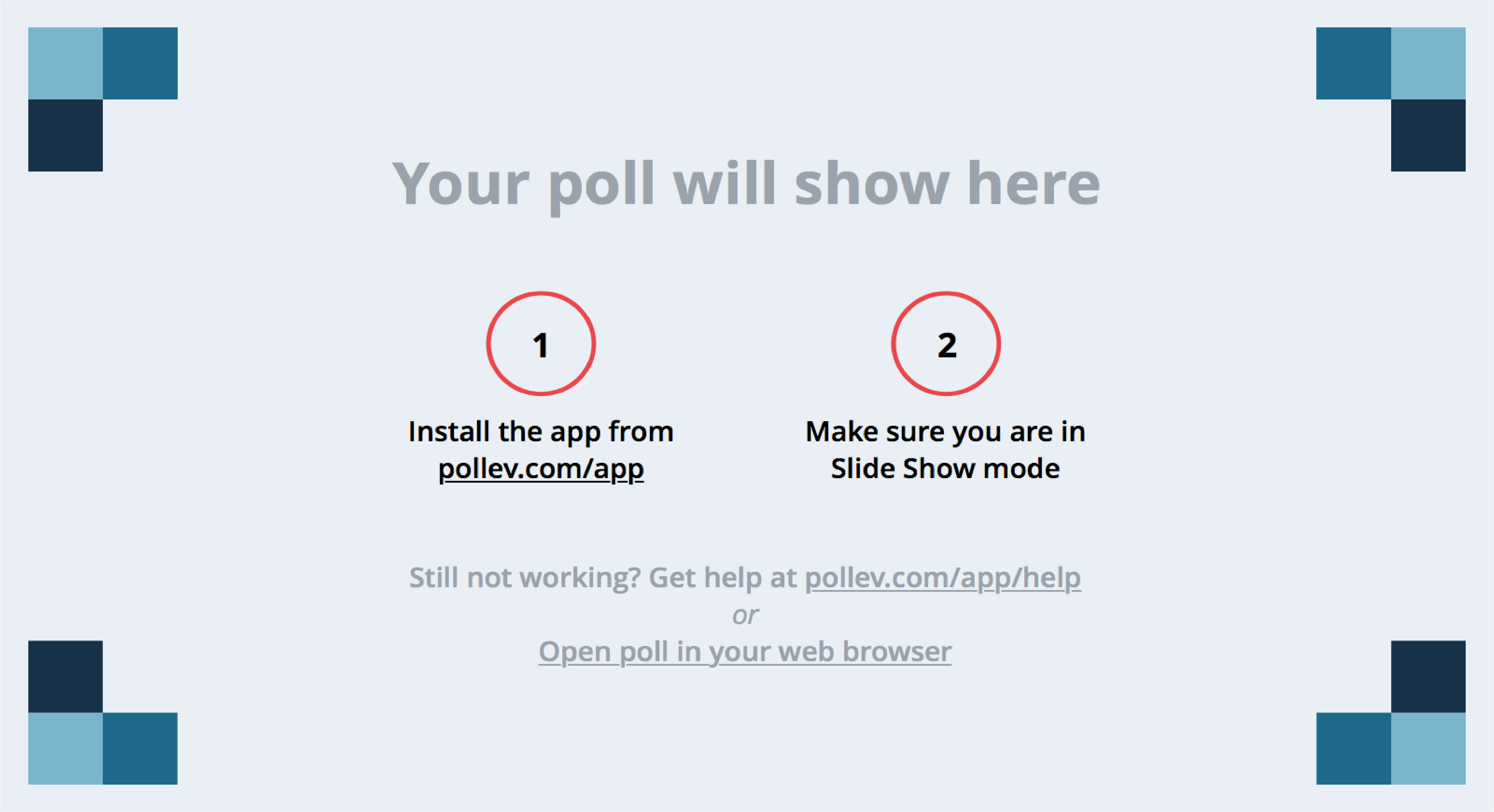
Variants per genome

Type	Variant sites / genome
SNPs	~3,800,000
Indels	~570,000
Mobile Element Insertions	~1000
Large Deletions	~1000
CNVs	~150
Inversions	~11



Population histories





Your poll will show here

1

Install the app from pollev.com/app

2

Make sure you are in Slide Show mode

Still not working? Get help at pollev.com/app/help

or

Open poll in your web browser

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

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 …

```
                              ★
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
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 …

```
                              ★
ACTAGTCGATGCTGGCTGATAGCTAGCTAGATGATGAGCCCGTTCGCTCCTAGCTCGACG
ACTAGTCGATGCTGGCTGATAGCTAGCTAGATGATGAGCCCGTTCGCTCCTAGCTCGACG
ACTAGTCGATGCTGGCTGATAGCTAGCTAGATGATGAGCCCGTTCGCTGCTAGCTCGACG
ACTAGTCGATGCTGGCTGATAGCTAGCTAGATGATGAGCCCGTTCGCTCCTAGCTCGACG
ACTAGTCGATGCTGGCTGATAGCTAGCTAGATGATGAGCCCGTTCGCTCCTAGCTCGACG
ACTAGTCGATGCTGGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTAGTCGATGCTGGCTGATAGCTAGCTAGATGATGAGCCCGTTCGCTCCTAGCTCGACG
ACTAGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGTTCGCTCCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG

5′-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3′
```

Variant Filtering

- Modern callers start with a candidate list of sites and annotate these …
  - Likely good sites: variants in HapMap or Omni 2.5M arrays
  - Likely problematic sites: variants that deviate from HWE or don’t segregate in families

- Then, build a model that separates likely good sites from likely bad ones …
  - SVM, VQSR, self-organizing maps, ….

- Possible features …
  - What is the mapping quality of reads with the variant?
  - How many other differences in reads with the variant?
  - How many individuals are heterozygotes and homozygotes?
  - How many reads with the variant are on the forward and reverse strand?
  - What fraction of reads have the variant in heterozygotes?
  - …

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

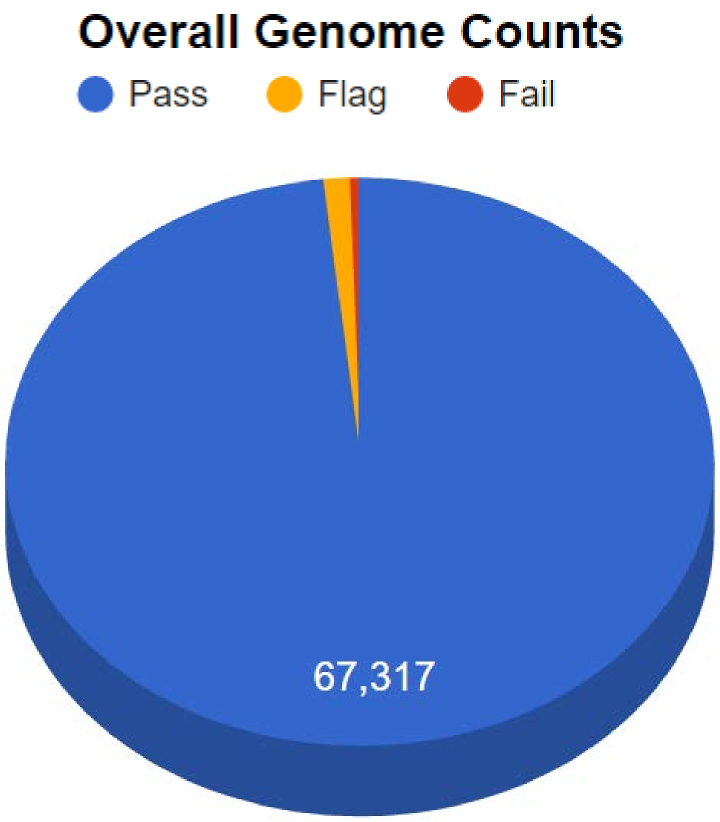
TOPMed Sequencing as of February 15, 2017

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

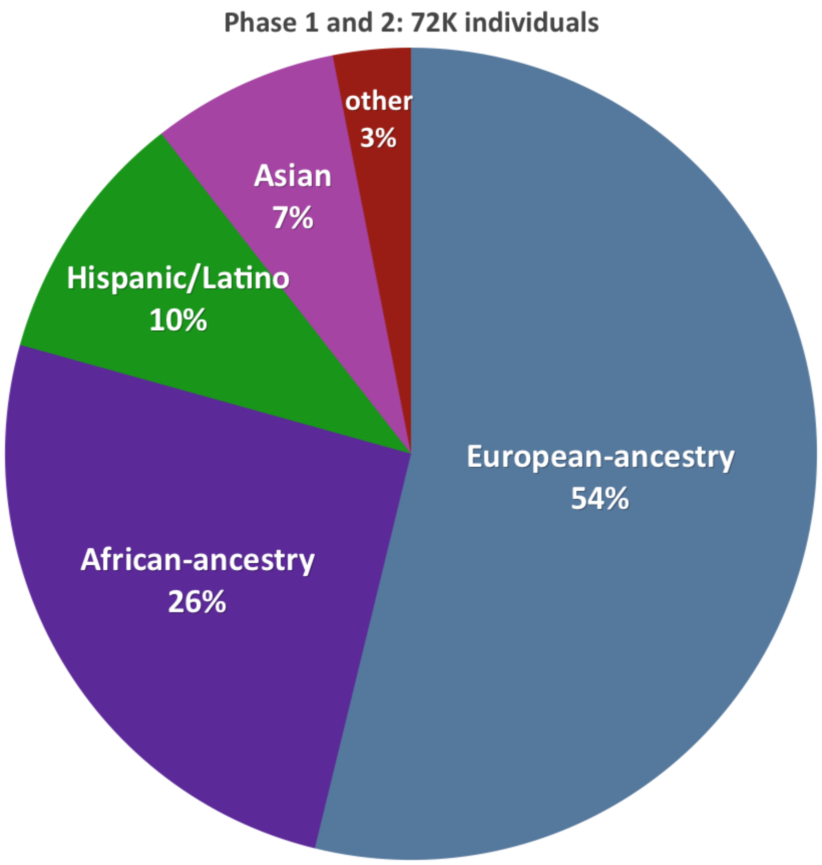
Number of snowflakes covering ~9 square miles in a 10-inch deep snowstorm.
100x more data than the 1000 Genomes Project.

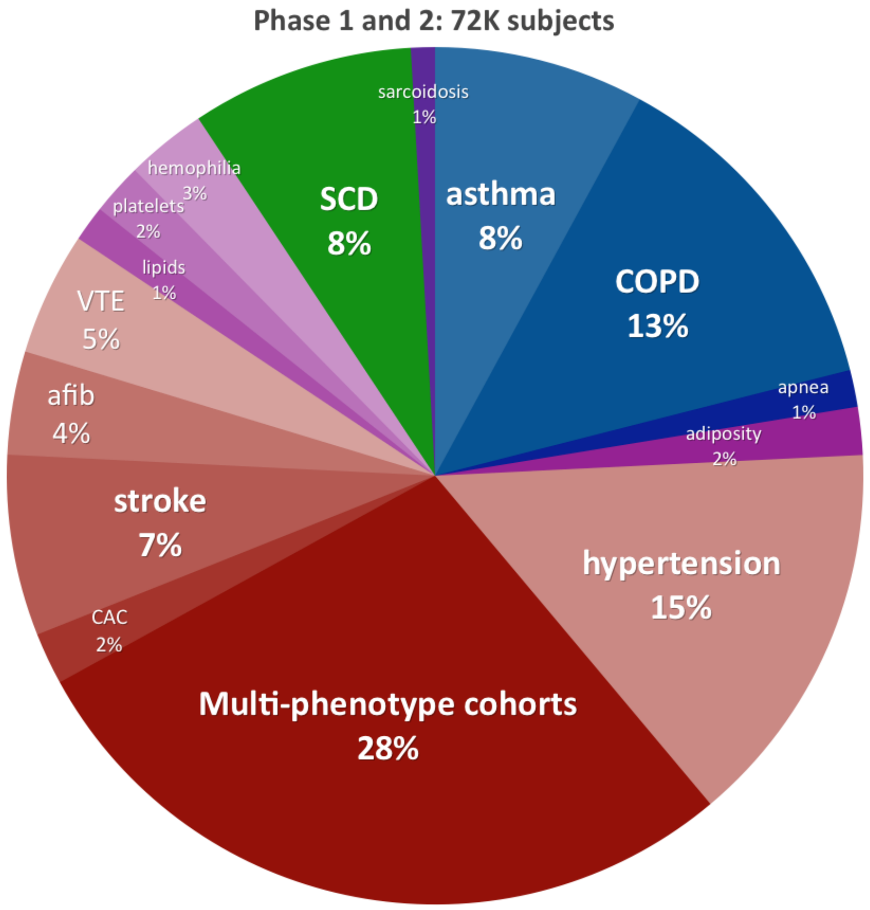
9 x 10^{15} sequenced bases

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

Image: Patrick Porter @ Smug Mug

Ancestry and Focus Phenotypes in TOPMed





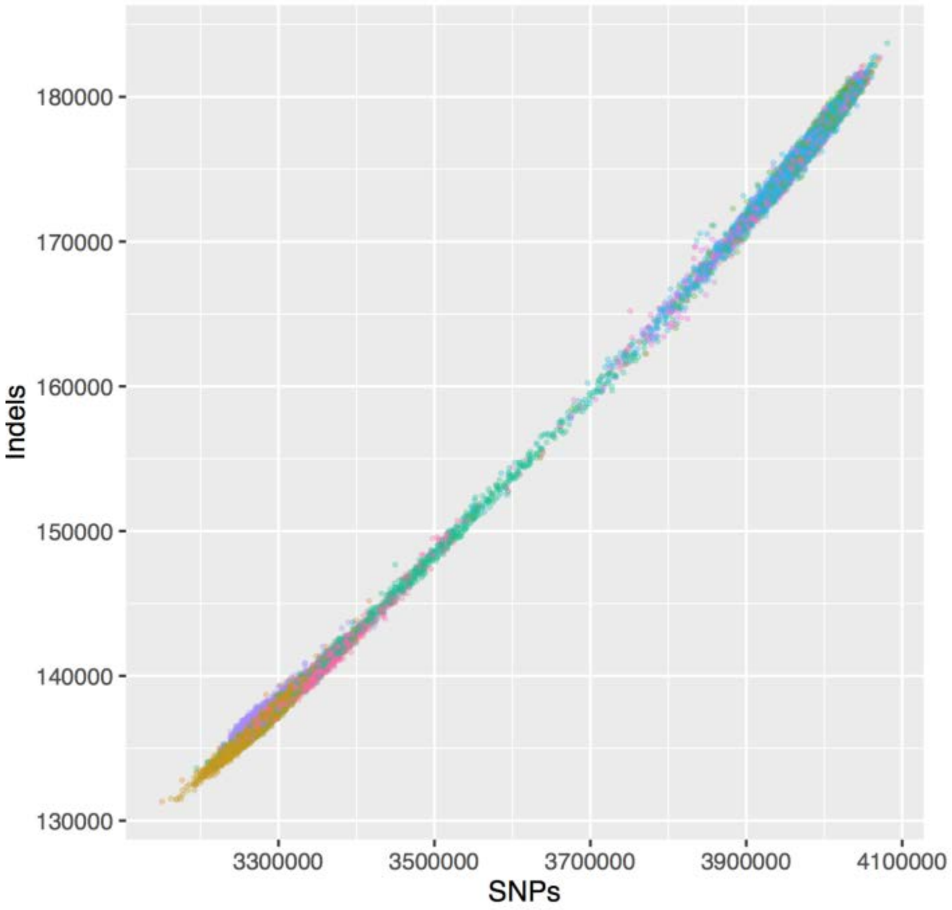
Quenna Wong and Cathy Laurie, University of Washington

Current TOPMed Data Freeze

- 18,877 samples sequenced by early May
  - 4,047 individuals in 1,301 nuclear families

- 191 million SNPs
- 10.1 million indels

- Genotype VCF is very cumbersome
  - Extracting subsets of individuals can take days!
  - Post VCF formats are more compact
  - Post VCF formats not supported by analysis tools



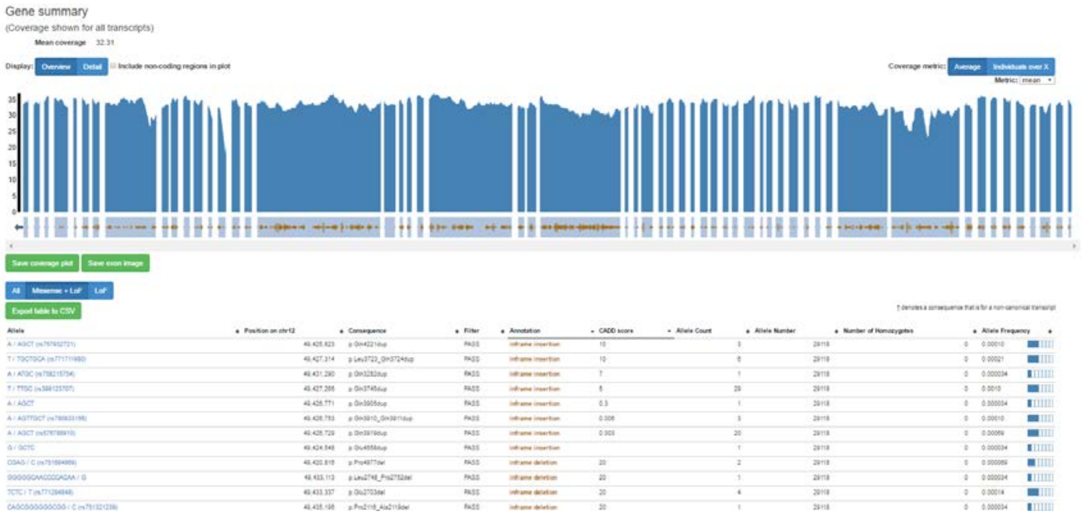
1.6M Coding Variants

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

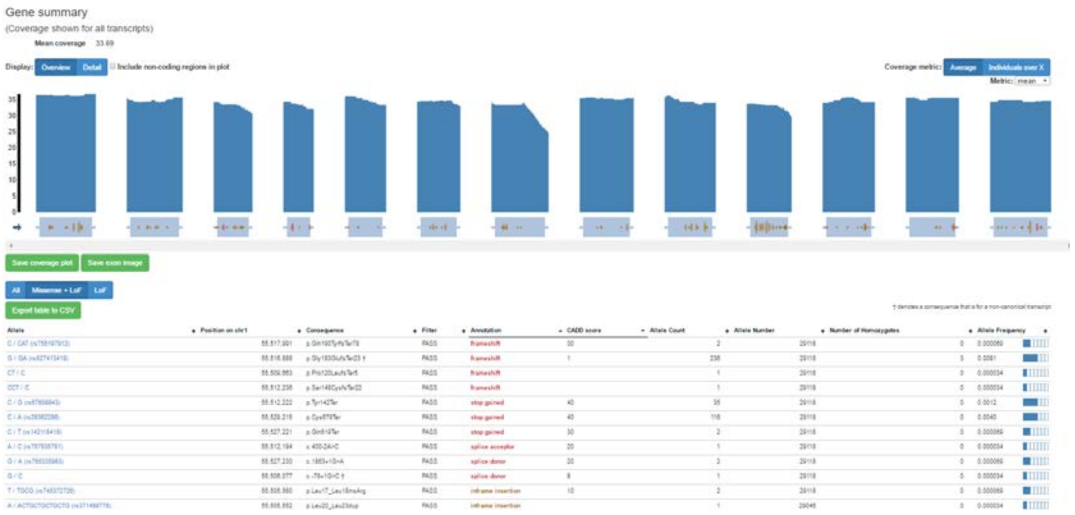
KMT2D


Gene summary
(Coverage shown for all transcripts)
Mean coverage 32.31
Display: Overview Detail □ Include non-coding regions in plot
Coverage metric: Average Individuals over X
Metric: mean
Save coverage plot | Save exon image
All | Missense + LoF | LoF
Export table to CSV
† denotes a consequence that is for a non-canonical transcript

Allele	Position on chr12	Consequence	Filter	Annotation	CADD score	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
A / AGCT (rs757932721)	49,425,823	p.Gln4221dup	PASS	inframe insertion	10	3	29118	0	0.00010
T / TGCTGCA (rs771711880)	49,427,314	p.Leu3723_Gln3724dup	PASS	inframe insertion	10	6	29118	0	0.00021
A / ATGC (rs758215754)	49,431,290	p.Gln3282dup	PASS	inframe insertion	7	1	29118	0	0.000034
T / TTGC (rs388125707)	49,427,266	p.Gln3745dup	PASS	inframe insertion	5	29	29118	0	0.0010
A / AGCT	49,425,771	p.Gln3905dup	PASS	inframe insertion	0.3	1	29118	0	0.000034
A / AGTTGCT (rs780833198)	49,425,783	p.Gln3910_Gln3911dup	PASS	inframe insertion	0.006	3	29118	0	0.00010
A / AGCT (rs578788910)	49,425,729	p.Gln3919dup	PASS	inframe insertion	0.003	20	29118	0	0.00069
G / GCTC	49,424,548	p.Glu4858dup	PASS	inframe insertion		1	29118	0	0.000034
CGAG / C (rs751694969)	49,420,818	p.Pro4977del	PASS	inframe deletion	20	2	29118	0	0.000069
GGGGGGAAGCCCAGAA / G	49,433,113	p.Leu2748_Pro2752del	PASS	inframe deletion	20	1	29118	0	0.000034
TCTC / T (rs771284848)	49,433,337	p.Glu2703del	PASS	inframe deletion	20	4	29118	0	0.00014
CAGCGGGGGCGG / C (rs751321238)	49,435,198	p.Pro2118_Ala2119del	PASS	inframe deletion	20	1	29118	0	0.000034

496 missense, 26 inframe indels, 0 stop or frameshifts

PCSK9

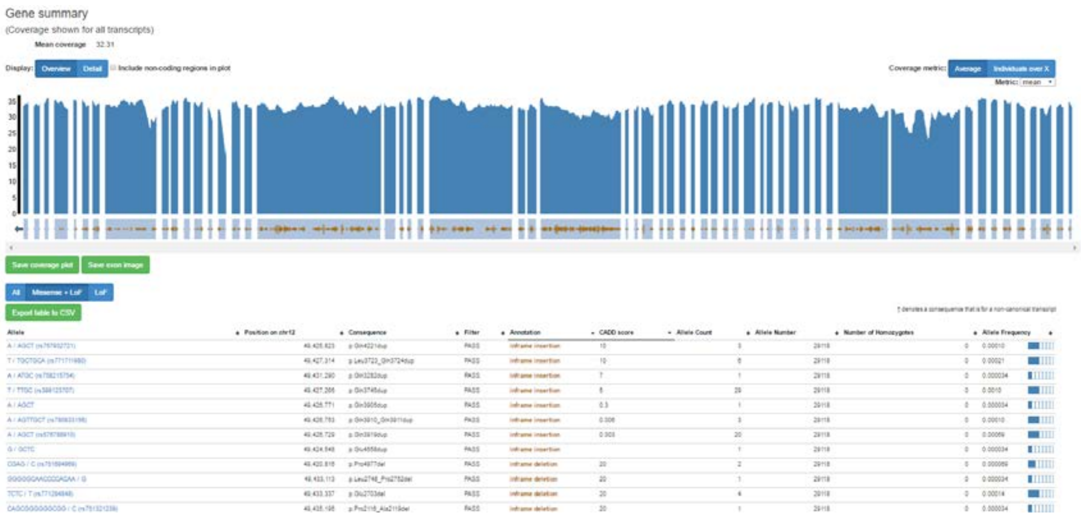

Gene summary
(Coverage shown for all transcripts)
Mean coverage 33.69
Display: Overview Detail □ Include non-coding regions in plot
Coverage metric: Average Individuals over X
Metric: mean
Save coverage plot | Save exon image
All | Missense + LoF | LoF
Export table to CSV
† denotes a consequence that is for a non-canonical transcript

Allele	Position on chr1	Consequence	Filter	Annotation	CADD score	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
C / CAT (rs758197912)	55,517,991	p.Gln190TyrfsTer78	PASS	frameshift	30	2	29118	0	0.000069
G / GA (rs527413419)	55,516,888	p.Gly183GlufsTer23 †	PASS	frameshift	1	236	29118	3	0.0081
CT / C	55,509,663	p.Pro120LeufsTer5	PASS	frameshift		1	29118	0	0.000034
CCT / C	55,512,238	p.Ser148CysfsTer22	PASS	frameshift		1	29118	0	0.000034
C / G (rs67608943)	55,512,222	p.Tyr142Ter	PASS	stop gained	40	35	29118	0	0.0012
C / A (rs28362286)	55,529,215	p.Cys679Ter	PASS	stop gained	40	118	29118	0	0.0040
C / T (rs142116418)	55,527,221	p.Gln619Ter	PASS	stop gained	30	2	29118	0	0.000069
A / C (rs787938781)	55,512,194	c.400-2A>C	PASS	splice acceptor	20	1	29118	0	0.000034
G / A (rs765335983)	55,527,230	c.1863+1G>A	PASS	splice donor	20	2	29118	0	0.000069
G / C	55,508,077	c.-78+1G>C †	PASS	splice donor	8	1	29118	0	0.000034
T / TGCG (rs745372726)	55,505,580	p.Leu17_Leu18insArg	PASS	inframe insertion	10	2	29118	0	0.000069
A / ACTGCTGCTGCTG (rs371488778)	55,505,552	p.Leu20_Leu23dup	PASS	inframe insertion		1	29046	0	0.000034

91 missense, 4 inframe indels, 7 stop or frameshifts

Federate!

KMT2D - BRAVO



Gene summary

(Coverage shown for all transcripts)

Mean coverage 32.31

Display: Overview Detail □ Include non-coding regions in plot

Coverage metric: Average Individuals over X

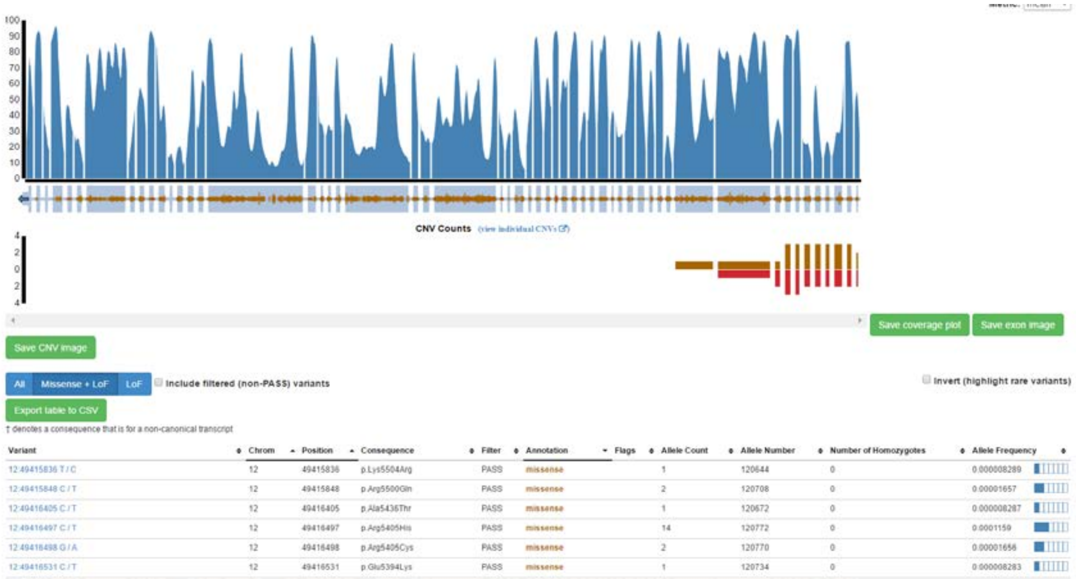
Metric: mean

† denotes a consequence that is for a non-canonical transcript

Allele	Position on chr12	Consequence	Filter	Annotation	CADD score	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
A / AGCT (rs757932721)	49,425,823	p.Gln4221dup	PASS	inframe insertion	10	3	29118	0	0.00010
T / TGCTGCA (rs771711680)	49,427,314	p.Leu3723_Gln3724dup	PASS	inframe insertion	10	6	29118	0	0.00021
A / ATGC (rs758215754)	49,431,290	p.Gln3282dup	PASS	inframe insertion	7	1	29118	0	0.000034
T / TTGC (rs368125707)	49,427,266	p.Gln3745dup	PASS	inframe insertion	5	29	29118	0	0.0010
A / AGCT	49,426,771	p.Gln3905dup	PASS	inframe insertion	0.3	1	29118	0	0.000034
A / AGTTGCT (rs780833198)	49,426,783	p.Gln3910_Gln3911dup	PASS	inframe insertion	0.006	3	29118	0	0.00010
A / AGCT (rs578788910)	49,426,729	p.Gln3919dup	PASS	inframe insertion	0.003	20	29118	0	0.00069
G / GCTC	49,424,548	p.Glu4558dup	PASS	inframe insertion		1	29118	0	0.000034
CGAG / C (rs751594869)	49,420,818	p.Pro4877del	PASS	inframe deletion	20	2	29118	0	0.000069
GGGGGCAACGGCAGAA / G	49,433,113	p.Leu2748_Pro2752del	PASS	inframe deletion	20	1	29118	0	0.000034
TCTC / T (rs771284848)	49,433,337	p.Glu2703del	PASS	inframe deletion	20	4	29118	0	0.00014
CAGCGGGGGCGG / C (rs751321239)	49,435,198	p.Pro2116_Ala2119del	PASS	inframe deletion	20	1	29118	0	0.000034

496 missense, 26 inframe indels, 0 stop or frameshifts

KMT2D - ExAC



□ Include filtered (non-PASS) variants

□ Invert (highlight rare variants)

† denotes a consequence that is for a non-canonical transcript

Variant	Chrom	Position	Consequence	Filter	Annotation	Flags	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
12:49415836 T / C	12	49415836	p.Lys5504Arg	PASS	missense		1	120644	0	0.000008289
12:49415848 C / T	12	49415848	p.Arg5500Gln	PASS	missense		2	120708	0	0.00001657
12:49416405 C / T	12	49416405	p.Ala5436Thr	PASS	missense		1	120672	0	0.000008287
12:49416497 C / T	12	49416497	p.Arg5405His	PASS	missense		14	120772	0	0.0001159
12:49416498 G / A	12	49416498	p.Arg5405Cys	PASS	missense		2	120770	0	0.00001656
12:49416531 C / T	12	49416531	p.Glu5394Lys	PASS	missense		1	120734	0	0.000008283

1842 missense, 23 inframe indels, 11 stop or frameshifts

Functional Variants in Non-coding Regions?

CADD Score	Variants	% Singletons
0 – 9	149,808,517	40%
10 – 14	9,481,664	42%
15 – 19	3,168,122	43%
20 – 24	885,525	45%
>= 25	334,964	50%

- Starting genome-wide explorations for signatures of selection and function

- Strongest outlier regions (>100kb) in singleton proportion are near *HLA* and *ABO.*
- A few large examples of regions (>100kb) with high singleton proportions (e.g. *TP53BP1*)

Functional Variants in Non-coding Regions?

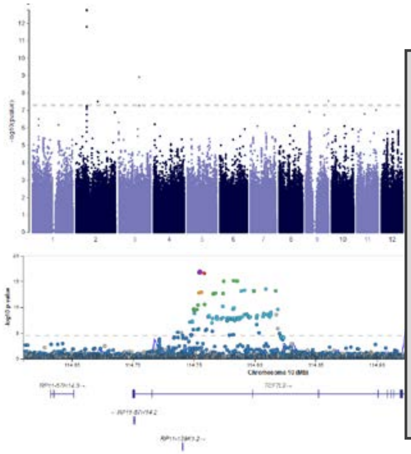
CADD Score	Variants	% Coding	% Singletons
0 – 9	149,808,517	0.5%	40%
10 – 14	9,481,664	4%	42%
15 – 19	3,168,122	12%	43%
20 – 24	885,525	39%	45%
>= 25	334,964	100%	50%

- Starting genome-wide explorations for signatures of selection and function

- Strongest outlier regions (>100kb) in singleton proportion are near *HLA* and *ABO.*
- A few large examples of regions (>100kb) with high singleton proportions (e.g. *TP53BP1*)

How to help TOPMed advance discoveries?

- Genomewide analyses at scale are challenging
- Even simple analysis can require 1,000s of CPU days to complete
- Need to engage diverse teams in analysis and interpretation



```
snp,pvalue
rs1234,0.05
rs4343,0.0002
rs51101,0.61
rs981,0.000018
rs2223,0.72
```

Plasma Lipids and Whole Genome Sequences

- Total Cholesterol, LDL, HDL, Triglycerides
- 8,394 TOPMed Participants
  - Jackson Heart Study
  - Framingham Heart Study
  - Amish Heart Study
- TOPMed Lipids Working Group
  - Leads: S. Kathiresan, C. Willer
  - PIs: A. Correa, A. Cupples, J. O'Connell, J. Wilson

Pradeep Natarajan

May Montasser

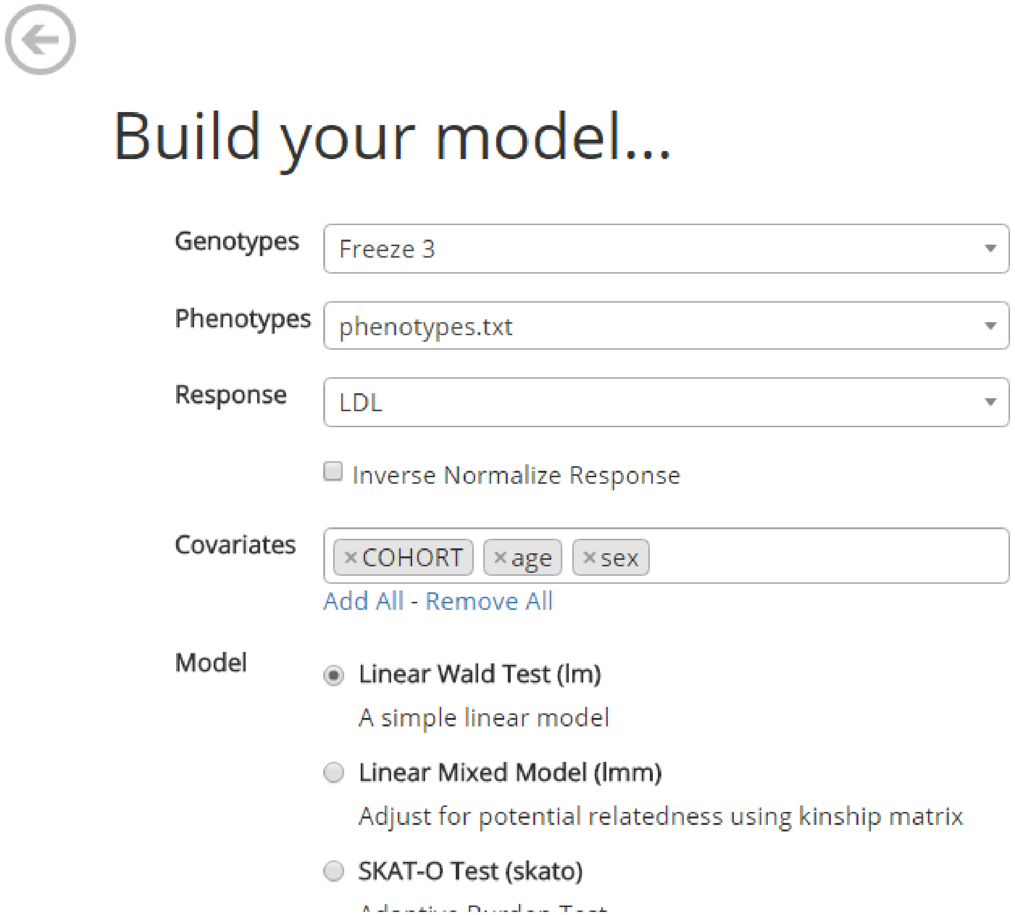
Maryam Zekavat

Gina Peloso

How ENCORE works …

Matthew Flickinger

Jonathon LeFaive



Build your model…

Genotypes: Freeze 3

Phenotypes: phenotypes.txt

Response: LDL

☐ Inverse Normalize Response

Covariates: × COHORT × age × sex

Add All - Remove All

Model

● Linear Wald Test (lm)

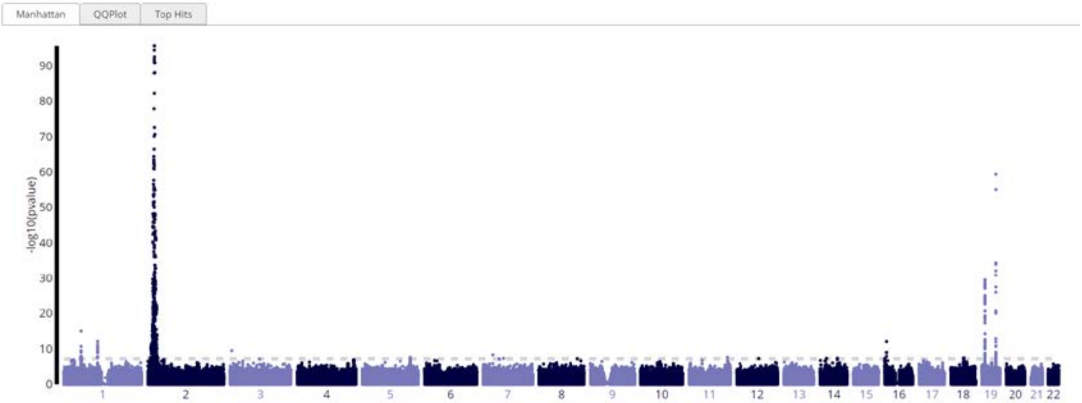
A simple linear model

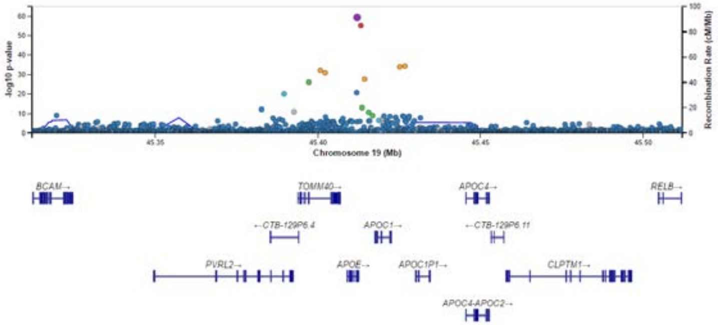
○ Linear Mixed Model (lmm)

Adjust for potential relatedness using kinship matrix

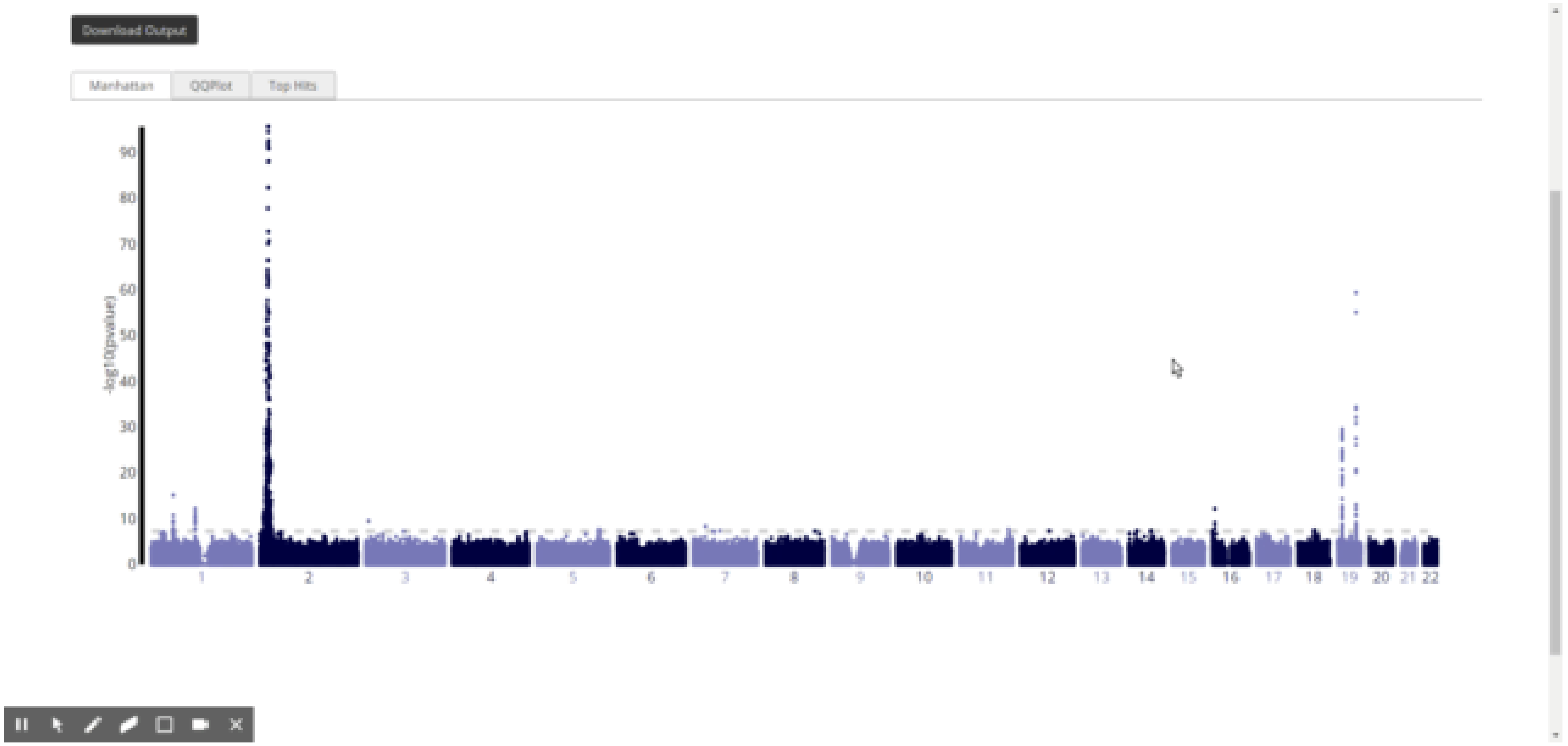
○ SKAT-O Test (skato)

Adaptive Burden Test





LDL Genomewide Analysis in ENCORE



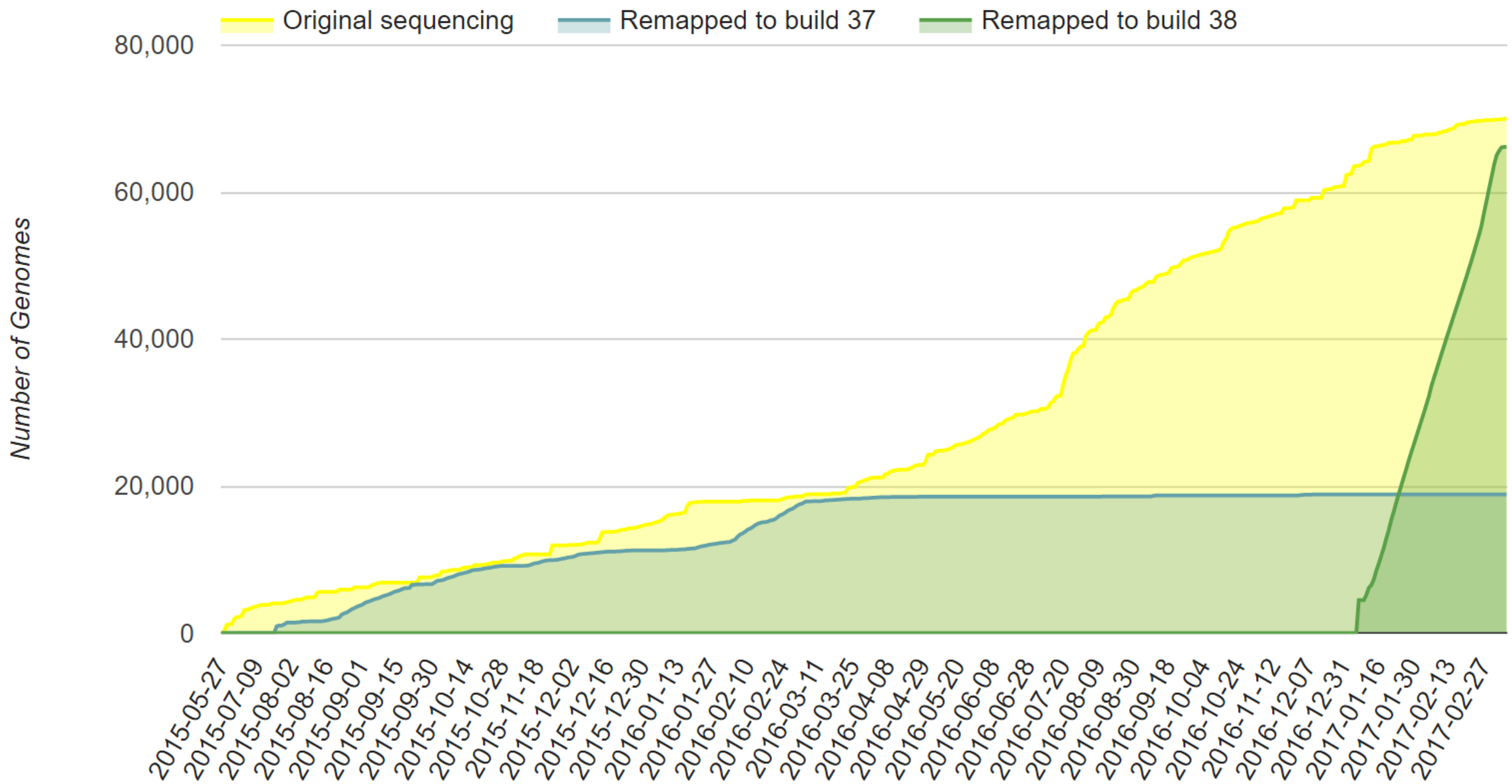
Mendelian Lipid Loci with LoF signals

Phenotype	Gene	Cumulative LOF Frequency	Association P-value
LDL	LDLR	0.00024	0.009
	APOB	0.00061	0.00005
	PCSK9	0.011	9×10^{-28}
HDL	LCAT	0.0004	0.035
	ABCA1	0.0006	0.012
	CETP	0.001	0.001
Triglycerides	APOC3	0.008	2×10^{-19}

Loci examined:

LDL (***LDLR***, ***APOB, PCSK9***, *LDLRAP1, ABCG5, ABCG8*)
HDL (*APOA1*, ***ABCA1, LCAT, CETP***, *LIPC*, *LIPG, SCARB1*)
Triglycerides (*LPL, APOC2*, *APOA5*, ***APOC3***, *GPIHBP1*, *LMF1, ANGPTL3, ANGPTL4*)

TOPMed Production and Processing



Your poll will show here

1

Install the app from pollev.com/app

2

Make sure you are in Slide Show mode

Still not working? Get help at pollev.com/app/help

or

Open poll in your web browser

Sequencing on the Cheap: Imputation

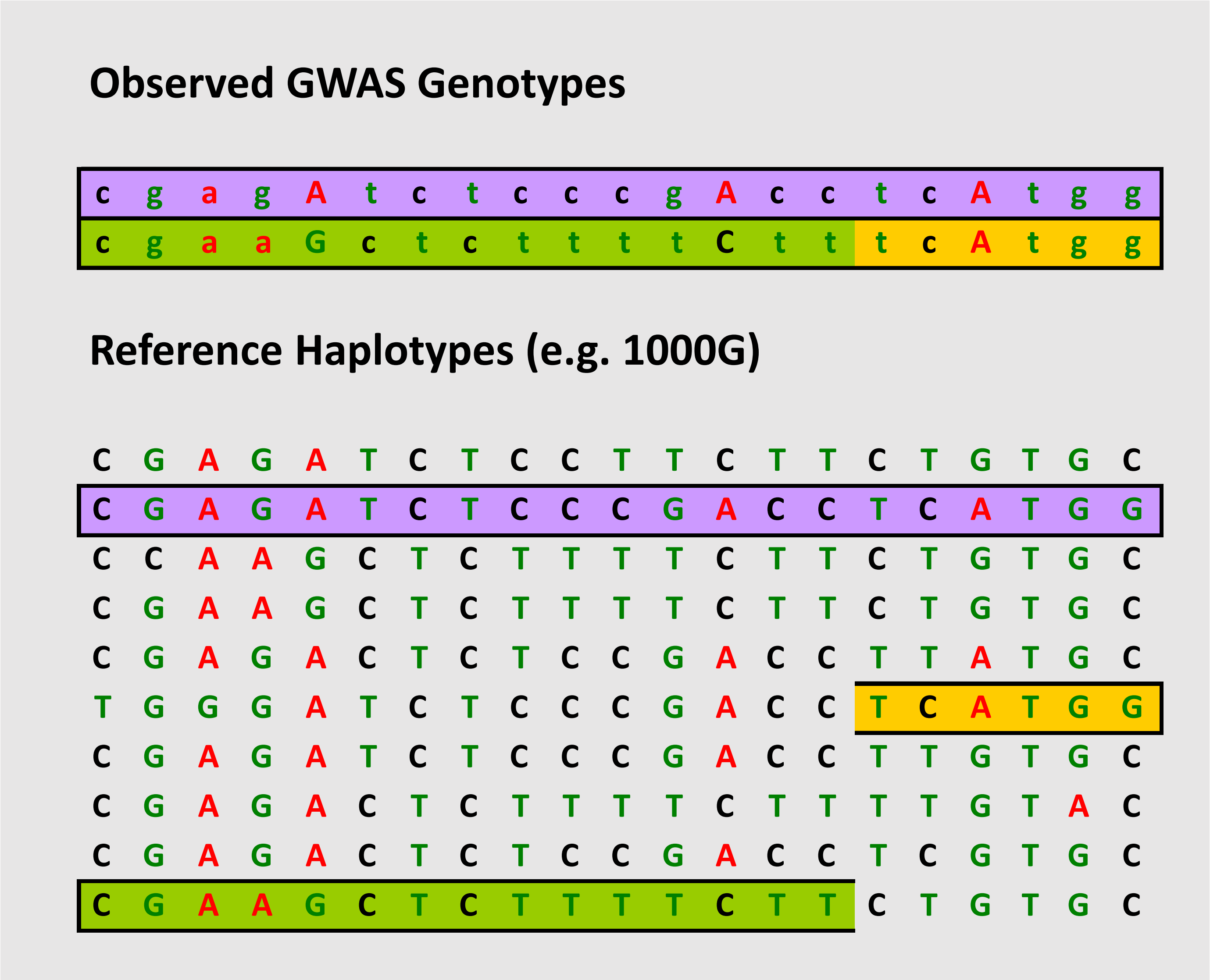
Observed GWAS Genotypes

.	.	.	.	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

Sequencing on the Cheap: Imputation



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

How long does it take to impute one genome?

- Depends on the reference size …

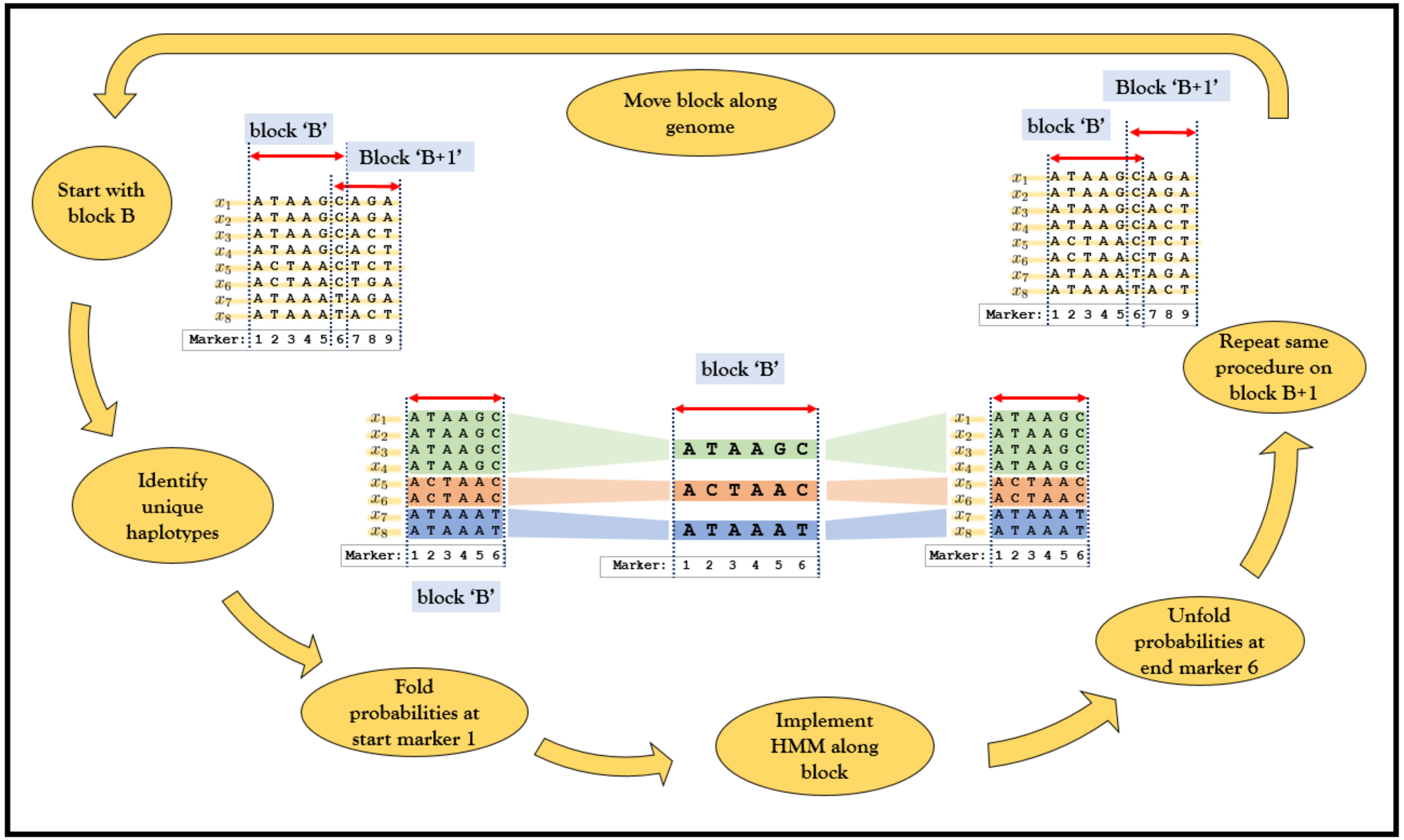
• 2007:	60 samples, 2.5M SNPs	14 min
• 2009:	60 samples, 7.3M SNPs	41 min
• 2011:	283 samples, 11.6M SNPs	1,287 min
• 2012:	381 samples, 37.4M SNPs	7,800 min
• 2015:	33,000 samples, 37M SNPs	63,000,000 min (estimated)

How long does it take to impute one genome?

- Depends on the computational methods …

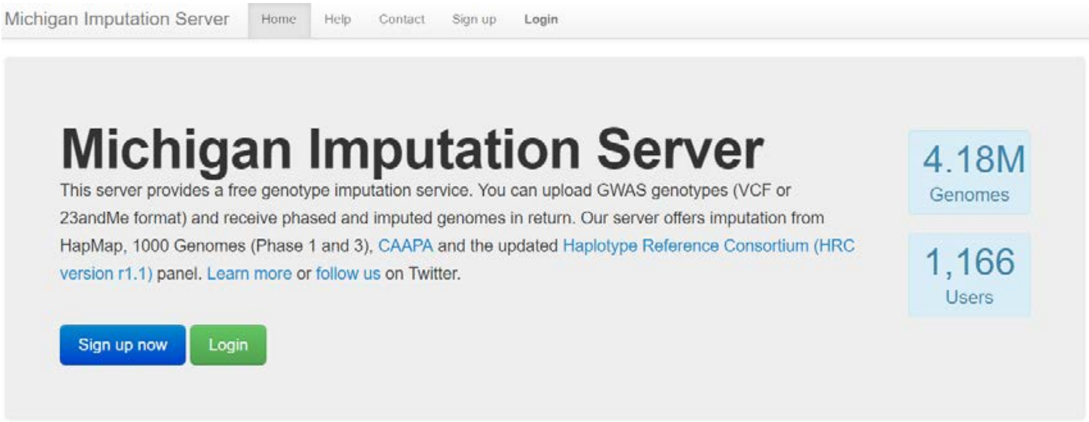
• 2007 software:	381 samples, 37.4M SNPs	7,800 min	$O(MH^2)$
• 2010 software:	381 samples, 37.4M SNPs	512 min	O(MH)
• 2012 software:	381 samples, 37.4M SNPs	24 min	O(MH)
• 2015 software:	381 samples, 37.4M SNPs	1 min	<O(MH)
• 2016 software:	381 samples, 37.4M SNPs	<5 secs	<O(MH)

Most Recent Imputation Improvements
Minimac3



Imputation Servers

https://imputationserver.sph.umich.edu

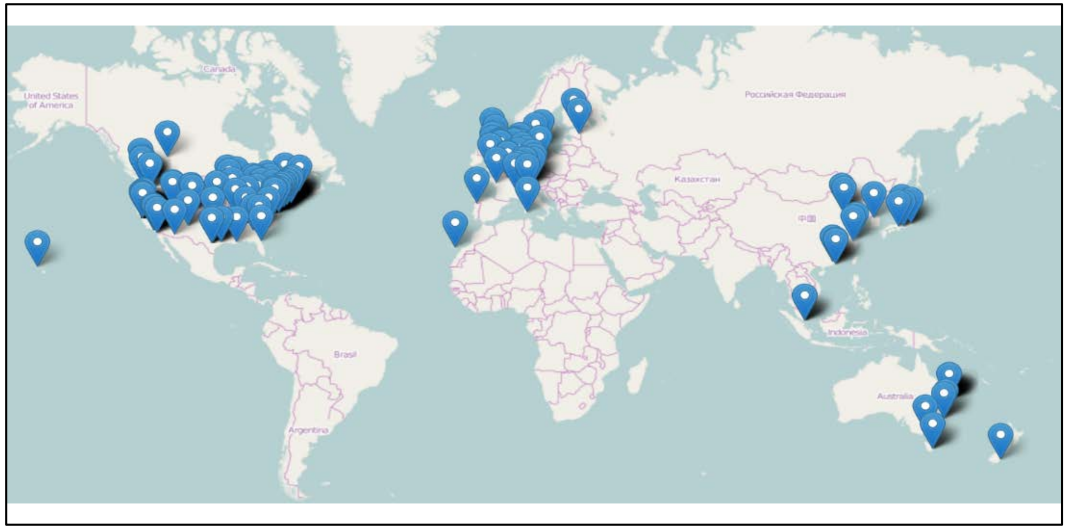


Michigan Imputation Server

This server provides a free genotype imputation service. You can upload GWAS genotypes (VCF or 23andMe format) and receive phased and imputed genomes in return. Our server offers imputation from HapMap, 1000 Genomes (Phase 1 and 3), CAAPA and the updated Haplotype Reference Consortium (HRC version r1.1) panel. Learn more or follow us on Twitter.

4.18M Genomes

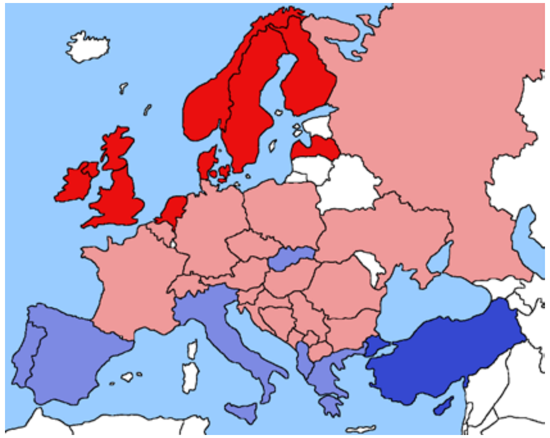
1,166 Users



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.



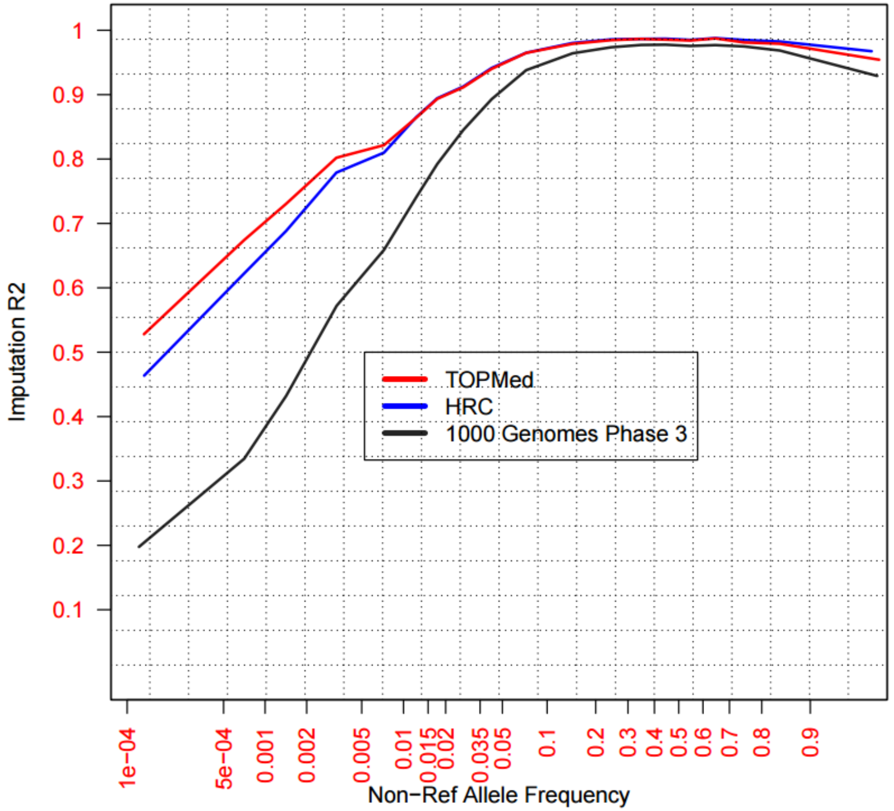
Imputation Accuracy w/TOPMed

Sayantan Das

- Imputation Accuracy in EUR
  - $r^2 > 0.85$ at 1% frequency
  - $r^2 > 0.75$ at 0.2% frequency

- Outperforms alternatives, including
  - 1000 Genomes (2,500 diverse genomes)
  - HRC (33,000 mainly European genomes)

- Working to define set of TOPMed samples that can be included in imputation panel



TOPMed Data Resources

- 1st release of sequenced genomes and phenotypes now in dbGAP/SRA
    - >8,000 genomes now available without embargo

- Browsable Variant Catalog at http://bravo.sph.umich.edu
    - 169,454,024 variants in 14,559 individuals
    - Variant lists will also be deposited in dbSNP

- Track sequence data production at http://nhlbi.sph.umich.edu

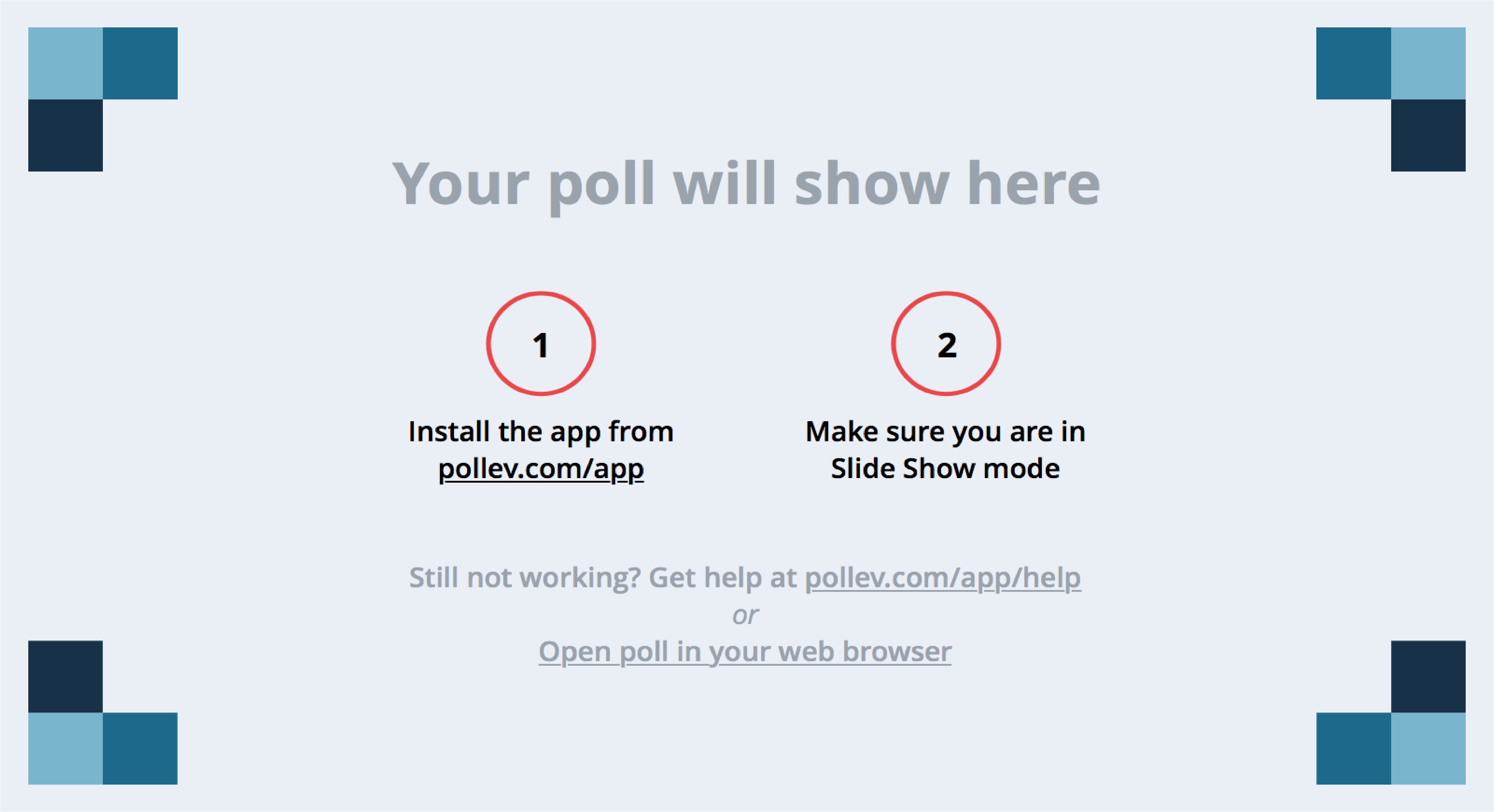
- Imputation resource that will improve ability to reconstruct diverse genomes affordably planned

Many new directions and opportunities …

- New techniques for exploring genomic function at scale
  - Use sequencing to measure enhancer activity after mass mutagenesis
  - Patwardhan et al (*Nature Biotechnology*, 2012)

- New techniques for dissecting biology of single cells
  - Use sequencing to profile expression in individual cells
  - Macosko et al (*Cell*, 2015)

- Open access resources like UK Biobank



Your poll will show here

1

Install the app from pollev.com/app

2

Make sure you are in Slide Show mode

Still not working? Get help at pollev.com/app/help

or

Open poll in your web browser

TOPMed Phase 1 and 2 – Acknowledgments

Asthma in African Descent Populations
Kathleen C. Barnes

Cardiovascular Disease Risk
Joanne E. Curran, David C. Glahn

Race and Ethnic Disparity in Ashtma
Esteban González Burchard

Early Atrial Fibrillation
Patrick T. Ellinor

Genes Influencing LDL Cholesterol
Braxton D. Mitchell

Framingham Heart Study
Vasan S. Ramachandran

Genetic Epidemiology of COPD
Edwin K. Silverman

Asthma in Costa Rica
Scott T. Weiss

Jackson Heart Study
Adolfo Correa, James Wilson

Samoan Family Obesity Study
Stephen T. McGarvey

Cleveland Family Study
Susan Redline

San Antonio Family Heart Study
John Blangero

Genotyping for Hemophilia
Barbara Konkle

Severe Asthma Research Program
Deborah A. Meyers

Women's Health Initiative
Charles Kooperberg

MESA Atherosclerosis Study
Jerome I. Rotter, Stephen S. Rich

AA Coronary Artery Calcification
Kent Taylor

HyperGEN and GENOA
Donna K. Arnett

Atherosclerosis Risk Study
Rasika Mathias

Atherosclerosis Risk Study and VTE
Eric Boerwinkle

Taiwan Study of Hypertension
D. C. Rao

GenNet Study of Salt Sensitivity
Jiang He

Sarcoidosis in African Americans
Courtney Montgomery

Walk-PHaSST
Mark Gladwin

REDS-III Brazil
Brian Custer, Shannon Kelly

Sickle Cell Disease (OMG)
Allison Ashley-Koch

Sickle Cell Disease (HW)
Sergei Nekhai

Sickle Cell Disease (Boston-Brazil)
Vijay G. Sankaran

PharmHU
Eric Boerwinkle

GOLDN
Donna K. Arnett

Severe COPD Gene
Edwin K. Silverman

UW Northwest Genomics Center
Debbie Nickerson

Baylor HGSC
Richard Gibbs

Broad Institute
Stacey Gabriel

NY Genome Center
Soren Germer

DCC
Bruce Weir

IRC
Gonçalo Abecasis

Acknowledgements

- National Institutes of Health
  - **NHLBI**
  - NHGRI
  - NCBI

- TOPMed Study Investigators
- TOPMed Data Coordinating Center
- TOPMed Sequencing Centers

- TOPMed Lipids Working Group

Thank you!
Michigan Team