http://tinyurl.com/666-week0

Welcome to Biostatistics 666!

Please fill in the survey while we wait to start.

Course Overview and Welcome!

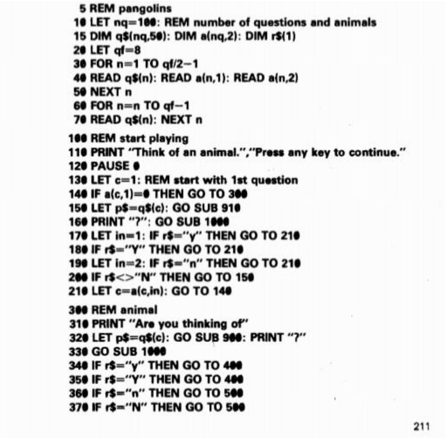
Biostatistics 666

Goncalo Abecasis

My Day Job…

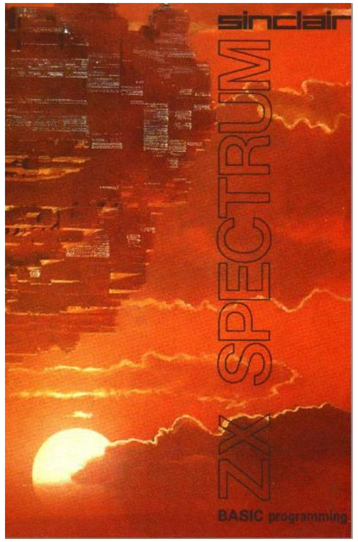
Use Genetic Variation, Math and Computation to Understand Human Disease

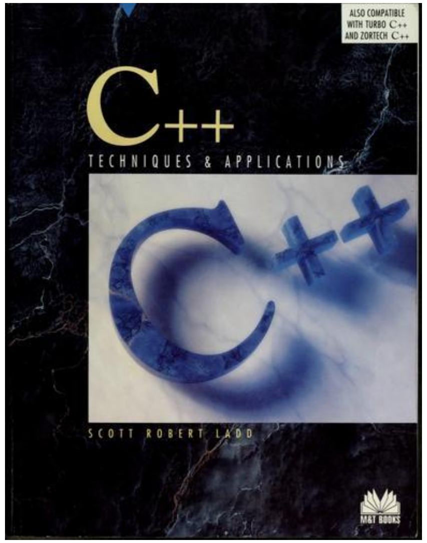
Why?

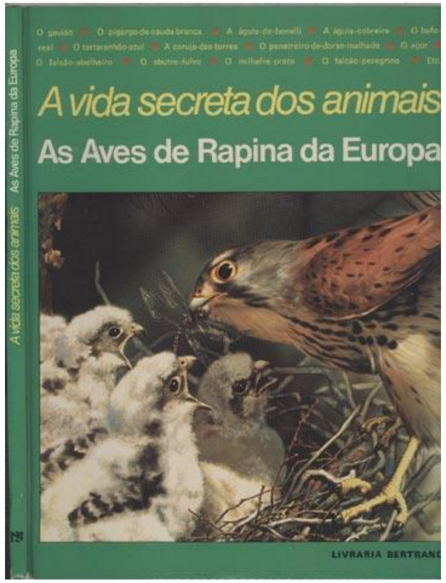


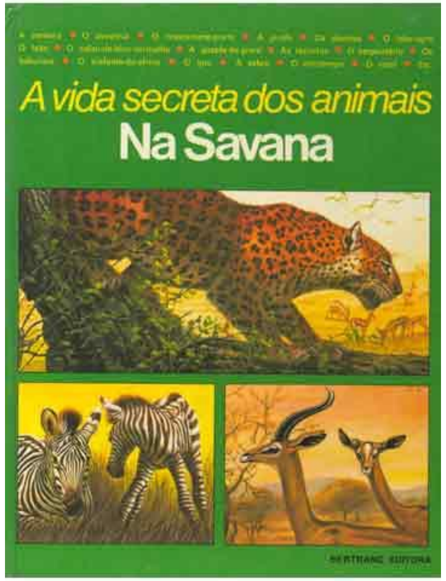
```
  5 REM pangolins
 10 LET nq=100: REM number of questions and animals
 15 DIM q$(nq,50): DIM a(nq,2): DIM r$(1)
 20 LET qf=8
 30 FOR n=1 TO qf/2-1
 40 READ q$(n): READ a(n,1): READ a(n,2)
 50 NEXT n
 60 FOR n=n TO qf-1
 70 READ q$(n): NEXT n

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




ALSO COMPATIBLE WITH TURBO C++ AND ZORTECH C++
C++
TECHNIQUES & APPLICATIONS
SCOTT ROBERT LADD
M&T BOOKS


O gavião • O pigargo-de-cauda-branca • A águia-de-bonelli • A águia-cobreira • O bufo-real • O tartaranhão-azul • A coruja-das-torres • O peneireiro-de-dorso-malhado • O açor • O falcão-abelheiro • O abutre-fulvo • O milhafre-preto • O falcão-peregrino • Etc.
A vida secreta dos animais
As Aves de Rapina da Europa
A vida secreta dos animais   As Aves de Rapina da Europa
LIVRARIA BERTRAND



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)

Course Logistics

Scheduling

Office Hours

Class Notes

Grading

Course Objective

- Introduce statistical models used in gene mapping studies
- Survey common algorithms used for handling genetic data
- Provide foundation for using gene mapping methods
- Provide foundation for refining and developing gene mapping methods

Course Notes

- We will not be using a textbook
  - Extremely important to attend class, and ask questions as needed!
- Copies of slides and additional content available online at
  - http://genome.sph.umich.edu/wiki/666

Assessment

- Grading will be a combination of:
    - Home work assignments (40%, approximately weekly)
    - In-class written assessments (60%, two of these)

- Greg Zajac (Biostatistics PhD student) will be helping me with grading.

Academic Integrity

- All assignments you submit for evaluation must represent your own work.
- If you copy or paraphrase others, you must clearly mark these sections and indicate sources.
- Any involvement in cheating constitutes academic misconduct and is a serious offense.
- See also the School policy on academic conduct.

Academic Integrity

- All assignments you submit for evaluation must represent your own work.
- If you copy or paraphrase others, you must clearly mark these sections and indicate sources.
- Any involvement in cheating constitutes academic misconduct and is a serious offense.
- See also the School policy on academic conduct.

If a set of assignments or exams is broadly identical, each will be scored as zero and referred to Department or School.

Scheduling

- We will try to start classes at 8:30 sharp.
    - Due to prior commitments, I may have to miss several lectures and starting sharply on time should allow us to make up for any lost time.

Office Hours

- We will try to find a time …
- To provide input, please fill in times when **you are available** at:

  http://tinyurl.com/666-office-hours

- **Thanks!**

Goals for Today …

- Overview of the evolution of complex disease studies
    - Current state of the art, challenges, opportunities
- Heritability
    - Estimating the total (additive) contribution of genetic variation to a trait
- Genomewide Association Studies
    - Linkage Disequilibrium, Genotype imputation
- Sequence-based Association Studies

If you are new to genetics …

- 23andMe provides a nice intro to the basic principles of genetics
- http://www.23andme.com/gen101/
- The material is all quite good and the videos are easy to watch.
- If genetics are new to you, I recommend you browse the first 4 sections.

Modern Gene Mapping Studies

A quick overview!

How human genetic studies work …

- DNA is our instruction manual

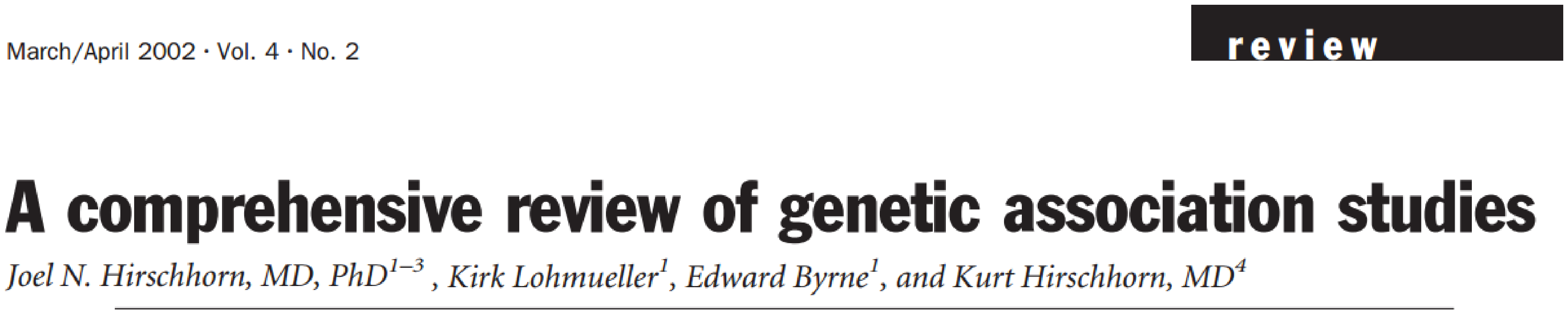
- We are all built mostly to the same plan…
  - Any two human DNA molecules are ~99.9% the same

- We each have our manual, with small variations from the typical plan
  - Most of these variations, as far as know, have no health consequences
  - Some modify key cell processes, a lot (these are very rare) or a little (more commonly)

- Search for variants that modify interesting health outcomes…
- … identify the cellular processes they modify …
- … improve understanding of biology, identify therapeutic targets …

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

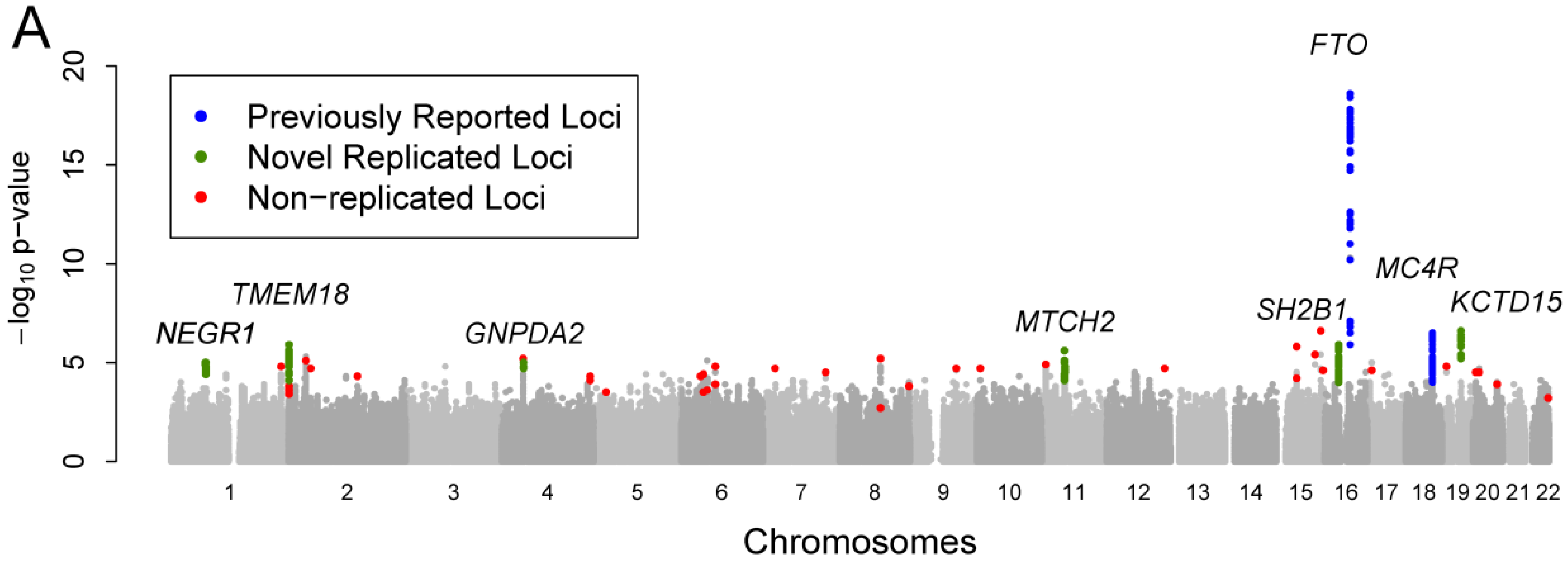
A comprehensive review of genetic association studies

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)

A Genomewide Study of Obesity



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

Willer et al, *Nature Genetics,* 2009

Current State of Genetic Association Studies

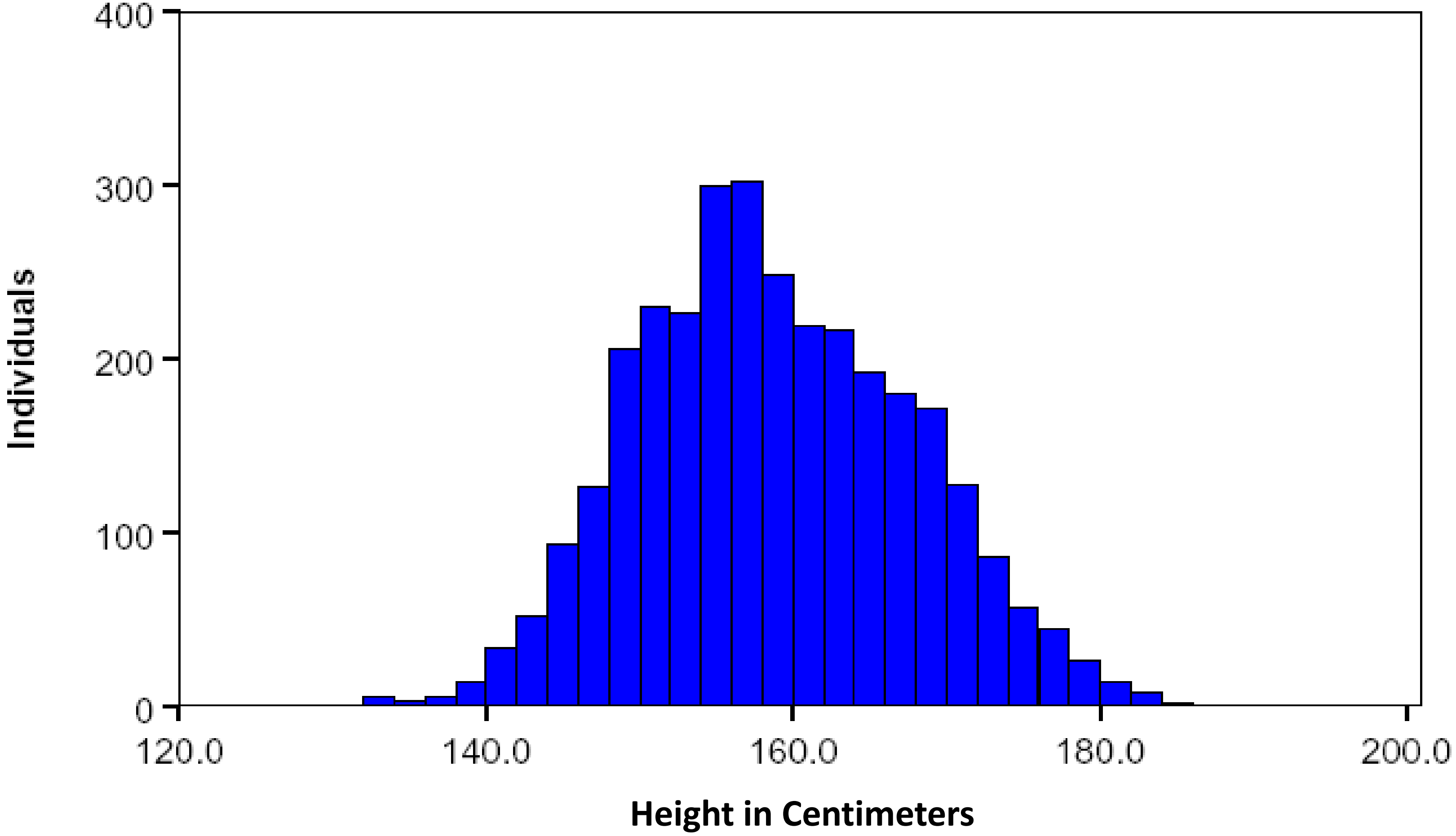
- Surveying common variation across 10,000s - 100,000s of individuals is now routine, using genotyping arrays
- Many common alleles have been associated with a variety of human complex traits
- The functional consequences of these alleles are often subtle, and translating the results into mechanistic insights remains challenging
- Sequencing studies are starting to allow studies to extend to rare variants, which can lead to easier to understand biology

Heritability

How Much of Phenotypic Variation Can Genetic Variation Explain?

Does Genetic Similarity Predict Phenotypic Similarity?

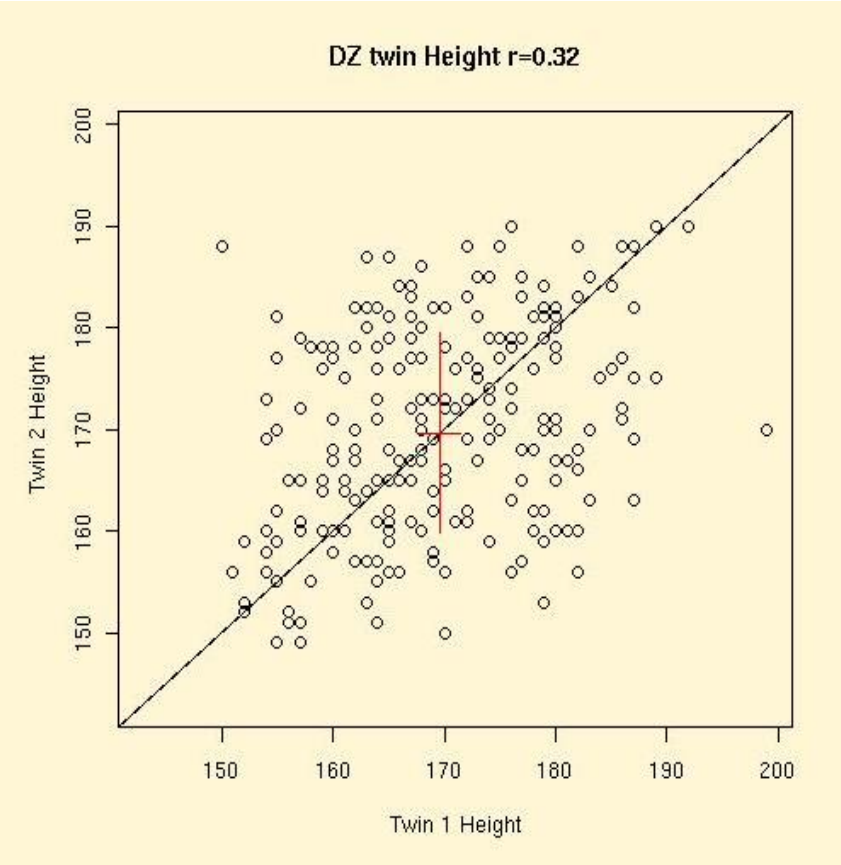
Variability in Height



We might often summarize this distribution with a mean and variance.

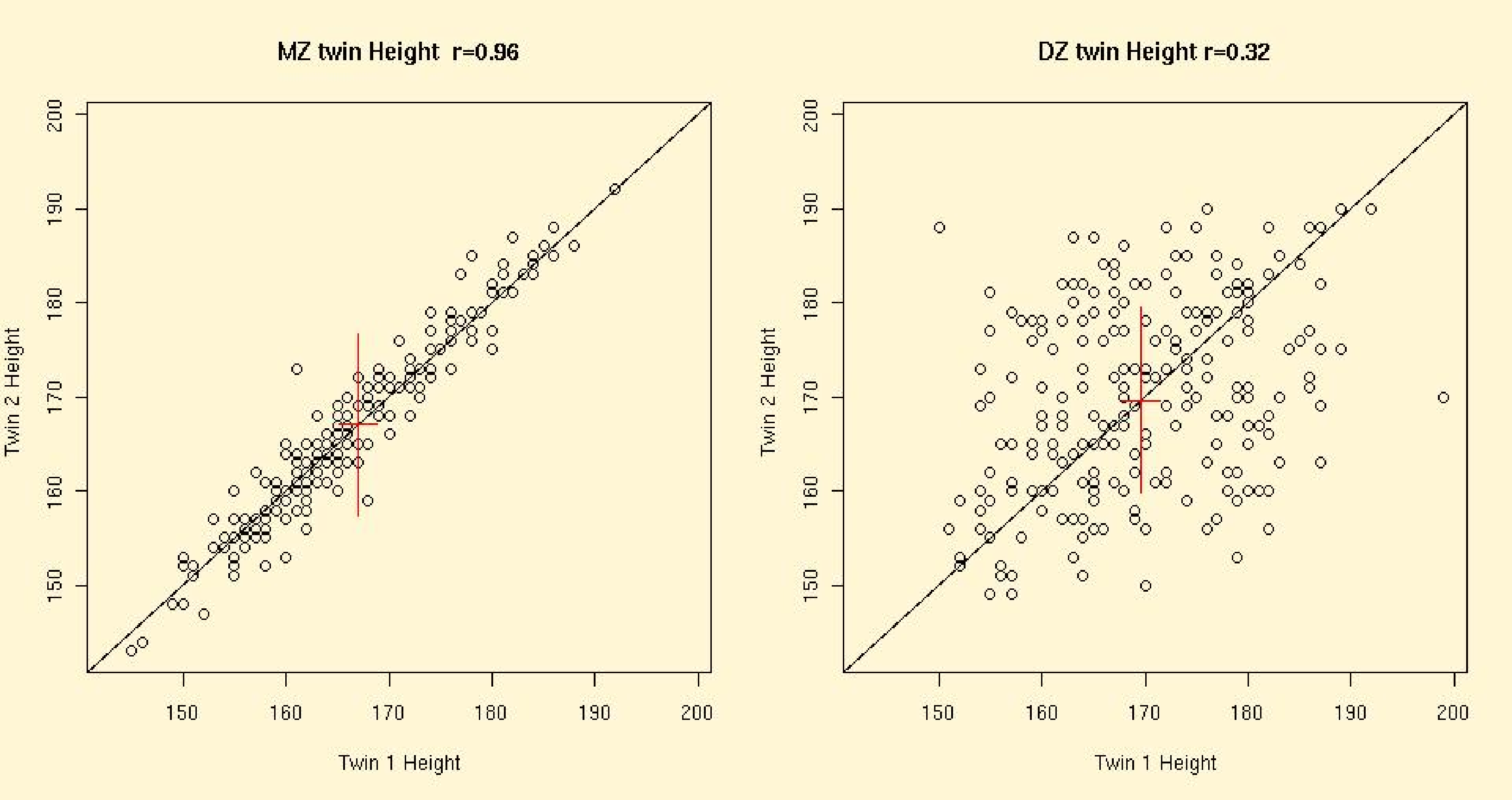
Variability in Height, Pairs of Observations

If sampling pairs of individuals, we might also record covariance between pairs of observations …



(Data from David Duffy)

Height in DZ and MZ twins



(How would you interpret these data from David Duffy?)

Variance-Covariance Matrix

$$\Omega = \begin{bmatrix} V(y_1) & Cov(y_1, y_2) \\ Cov(y_1, y_2) & V(y_2) \end{bmatrix}$$

Model describes not only variance of each observation but also covariance for pairs of observations

A Simple Model for the Variance-Covariance Matrix

$$\Omega = \begin{bmatrix} \sigma_g^2 + \sigma_e^2 & 2\varphi\sigma_g^2 \\ 2\varphi\sigma_g^2 & \sigma_g^2 + \sigma_e^2 \end{bmatrix}$$

Where,

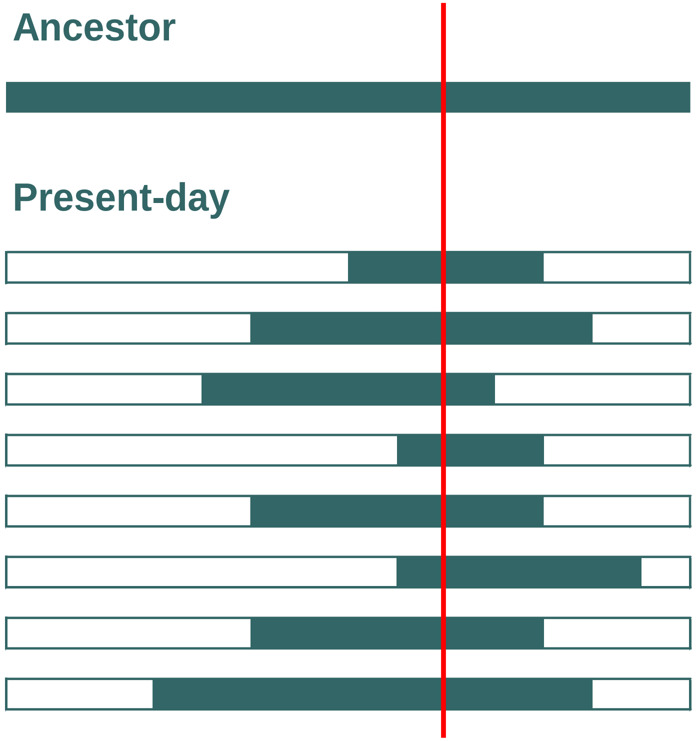
φ is the kinship coefficient for the two individuals

Linkage Disequilibrium and Genetic Association Studies

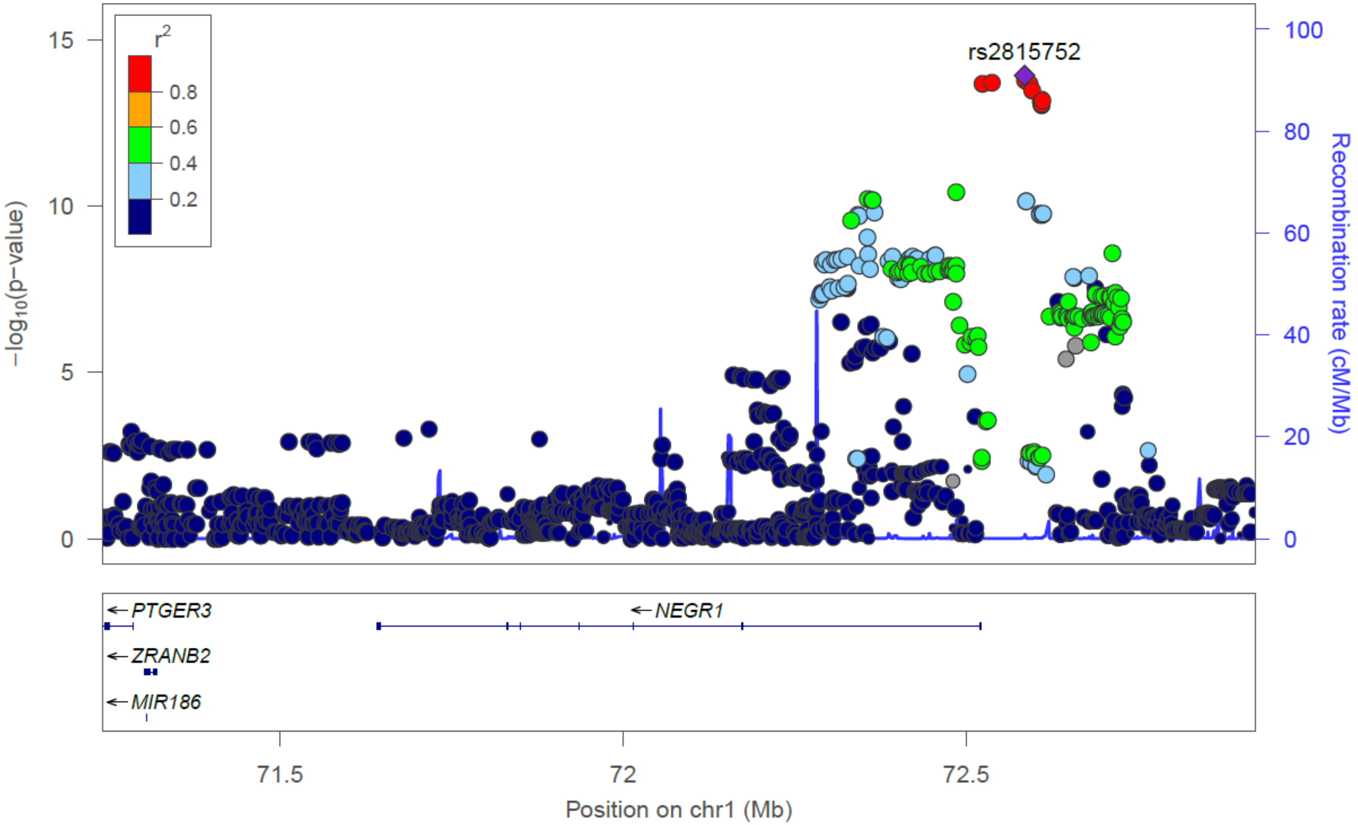
Genetic Association Signals at Nearby Variants …

Linkage Disequilibrium

- Chromosomes are mosaics
- Extent and conservation of mosaic pieces depends on
  - Recombination rate
  - Mutation rate
  - Population size
  - Natural selection
- Combinations of alleles at very close markers reflect ancestral haplotypes



Obesity and the *NEGR1* locus



Multiple nearby SNPs show evidence for association with obesity.
The associated alleles usually appear together, in a haplotype.

Willer et al, *Nature Genetics, 2009*

Observed Genotypes

Observed Genotypes

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

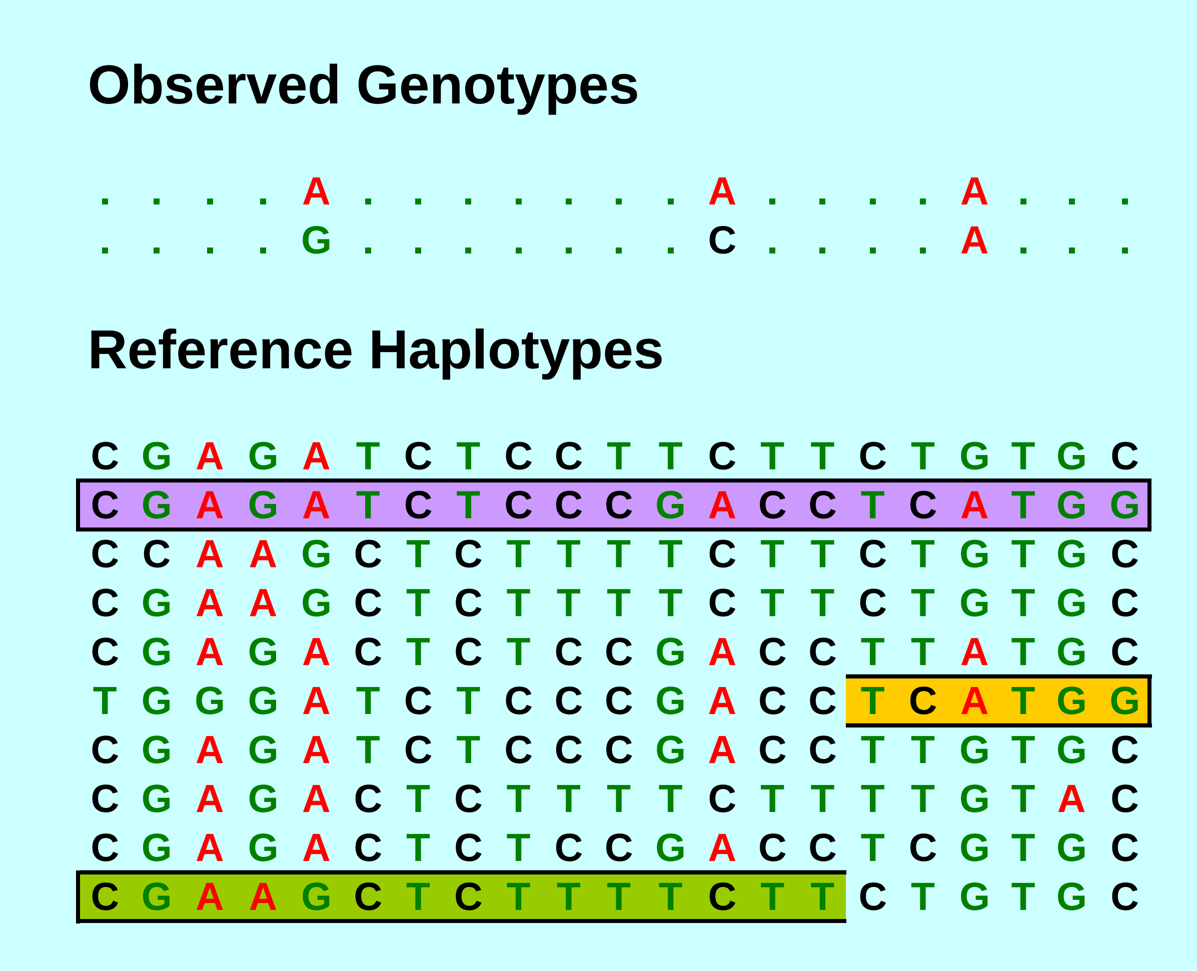
Reference Haplotypes

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

Study Sample
Inexpensive measurements
at 100,000s of markers

Reference Sample
Detailed measurements
of 1,000,000s of markers

Identify Match Among Reference



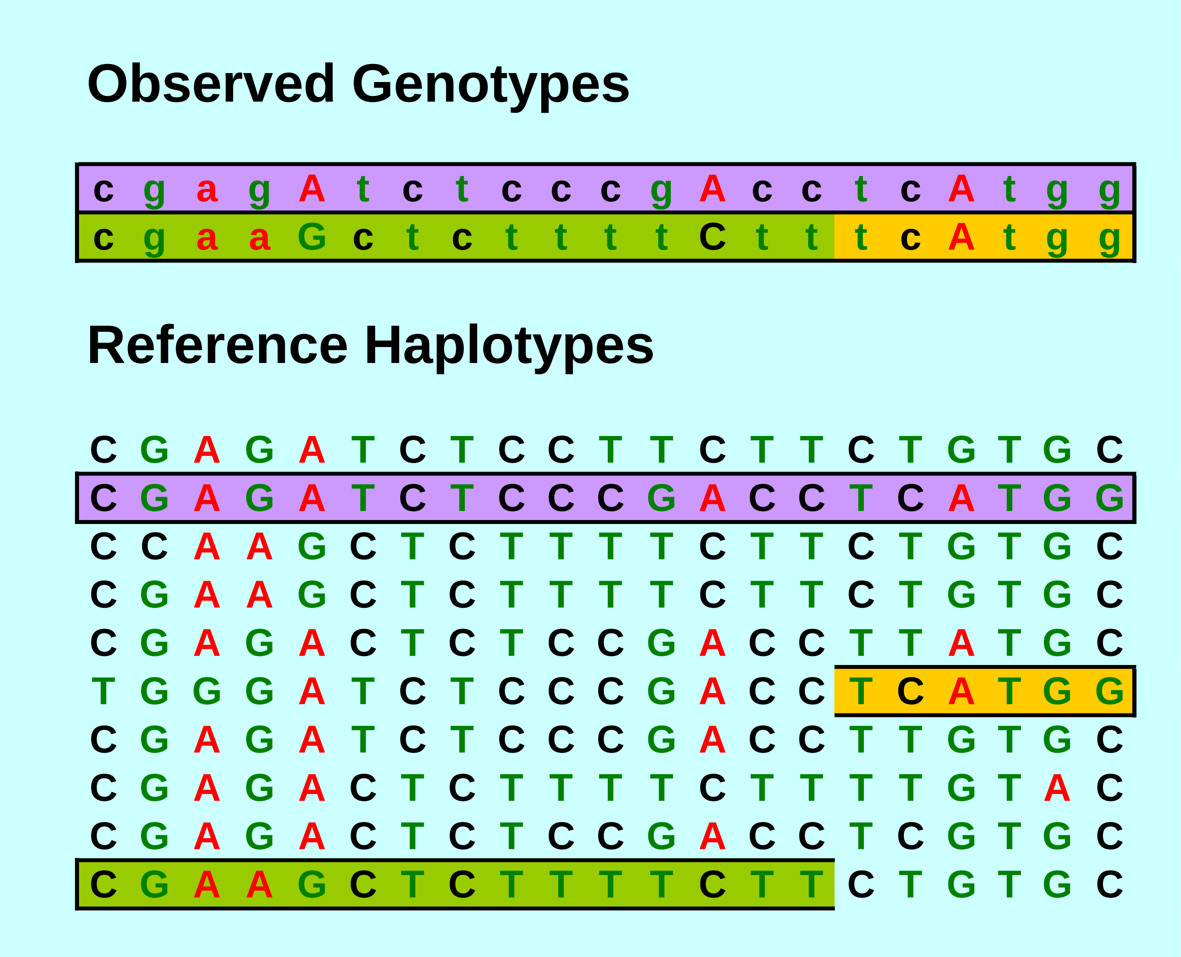
Observed Genotypes

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

Reference Haplotypes

```
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
```

Fill-in Missing Genotypes



Observed 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

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

The Role of Sequencing in Genetic Association Studies

Shotgun Sequence Data

★

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

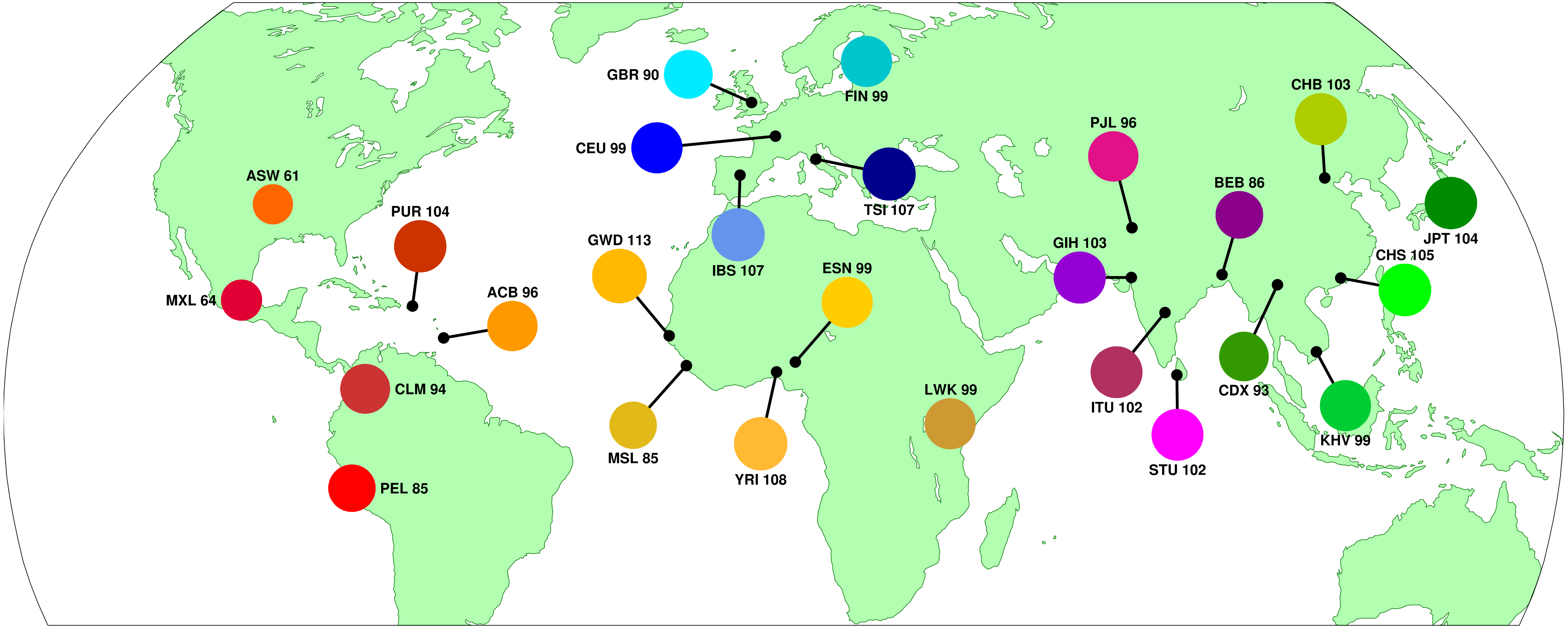
5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

A/C

Predicted Genotype

The 1000 Genomes Project (2008 – 2015)



Optimal Model for Analyzing 1000 Genomes?

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

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

Optimal Model for Analyzing 1000 Genomes?

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

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

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

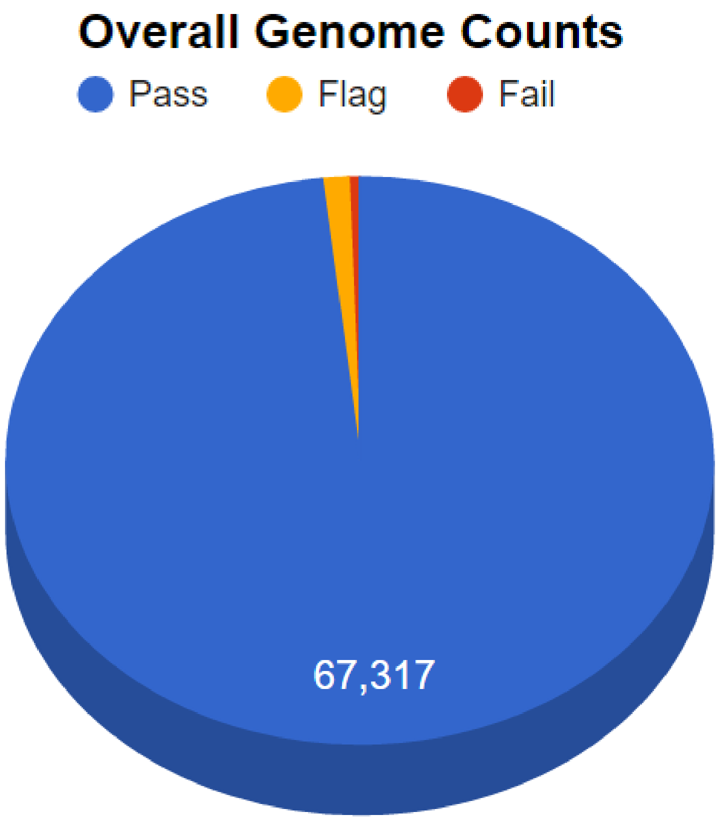
TOPMed Sequencing as of May 25, 2017

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

- 76,436 genomes
  - 74,890 pass quality checks (98.0%)
  - 946 flagged for low coverage (1.2%)
  - 606 fail quality checks (0.8%)

- Mean depth: 38.3x
- Genome covered: 98.6%
- Contamination: 0.28%

- 10^{16} sequenced bases



10^{16} sequenced bases

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

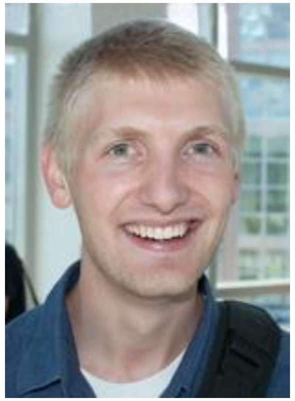
10^{16} sequenced bases

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

Image: Patrick Porter @ Smug Mug

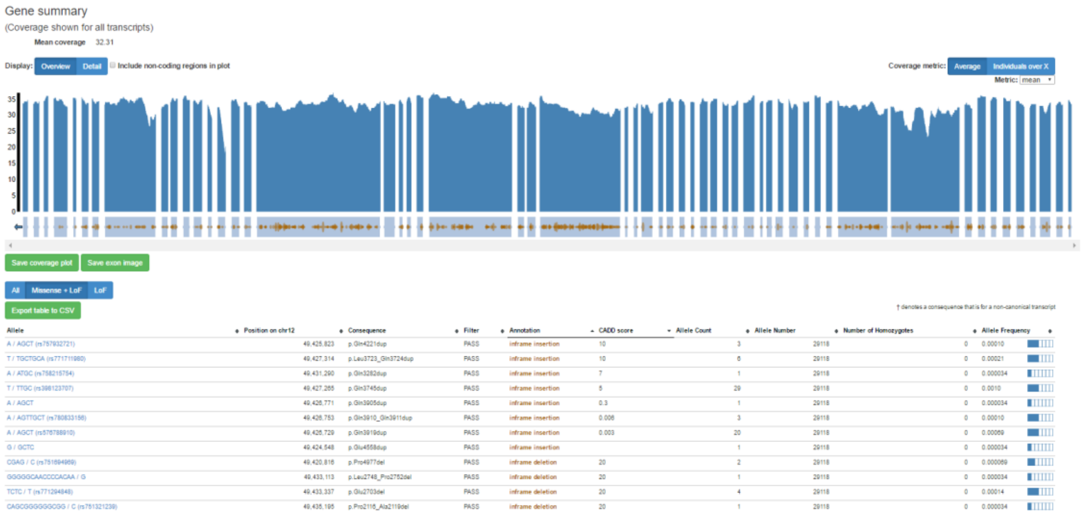
Browse All Variations Online

http://bravo.sph.umich.edu



Peter VandeHaar

KMT2D



Gene summary

(Coverage shown for all transcripts)

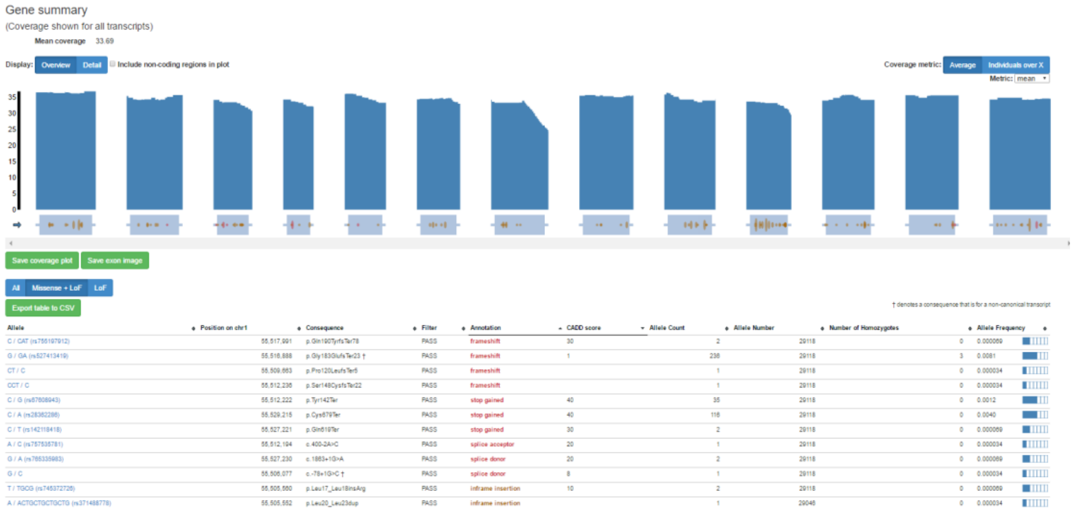
Mean coverage 32.31

† 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 (rs771711980)	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 (rs398123707)	49,427,265	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 (rs780833156)	49,426,753	p.Gln3910_Gln3911dup	PASS	inframe insertion	0.006	3	29118	0	0.00010
A / AGCT (rs576788910)	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 (rs751694969)	49,420,816	p.Pro4977del	PASS	inframe deletion	20	2	29118	0	0.000069
GGGGGCAACCCCACAA / G	49,433,113	p.Leu2748_Pro2752del	PASS	inframe deletion	20	1	29118	0	0.000034
TCTC / T (rs771294848)	49,433,337	p.Glu2703del	PASS	inframe deletion	20	4	29118	0	0.00014
CAGCGGGGGGCGG / C (rs751321239)	49,435,195	p.Pro2116_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

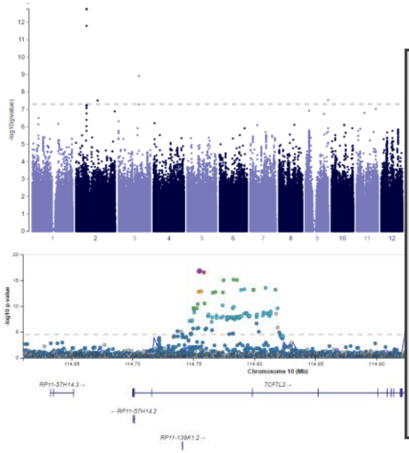
† denotes a consequence that is for a non-canonical transcript

Allele	Position on chr1	Consequence	Filter	Annotation	CADD score	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
C / CAT (rs755197912)	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	55,512,236	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	116	29118	0	0.0040
C / T (rs142118418)	55,527,221	p.Gln619Ter	PASS	stop gained	30	2	29118	0	0.000069
A / C (rs757535781)	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	55,506,077	c.-78+1G>C †	PASS	splice donor	8	1	29118	0	0.000034
T / TGCG (rs745372726)	55,505,560	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

How to help TOPMed advance discoveries?

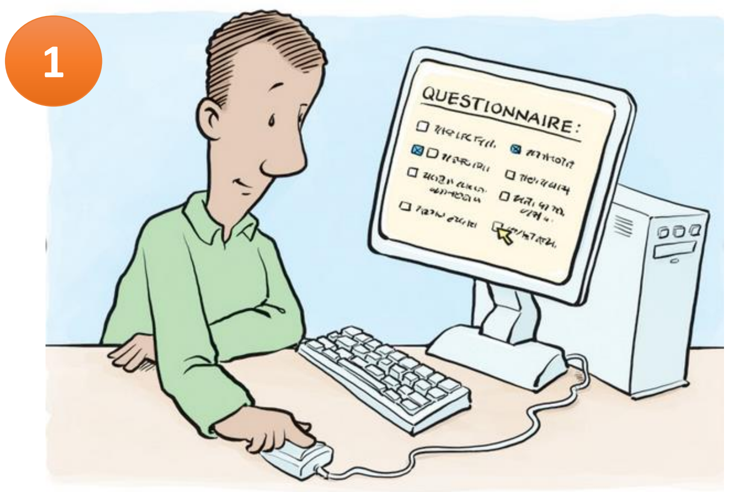
- Genomewide analyses at scale are challenging
- Even simple analysis can require 1,000s of CPU days to complete
- Need to engage diverse teams in analysis and interpretation



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

GENES for GOOD

1



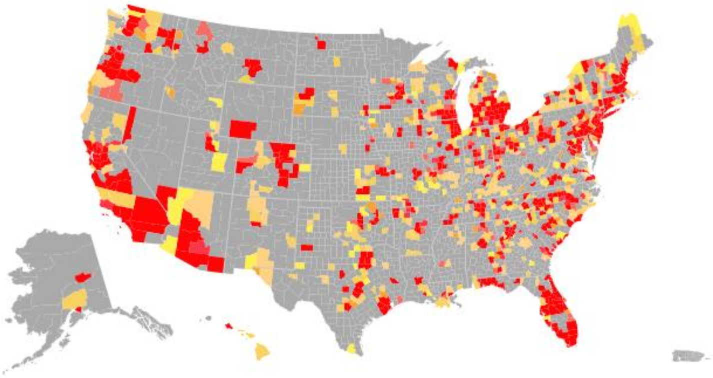
2

3

PROTECTIONS
DNA SAMPLE #273610

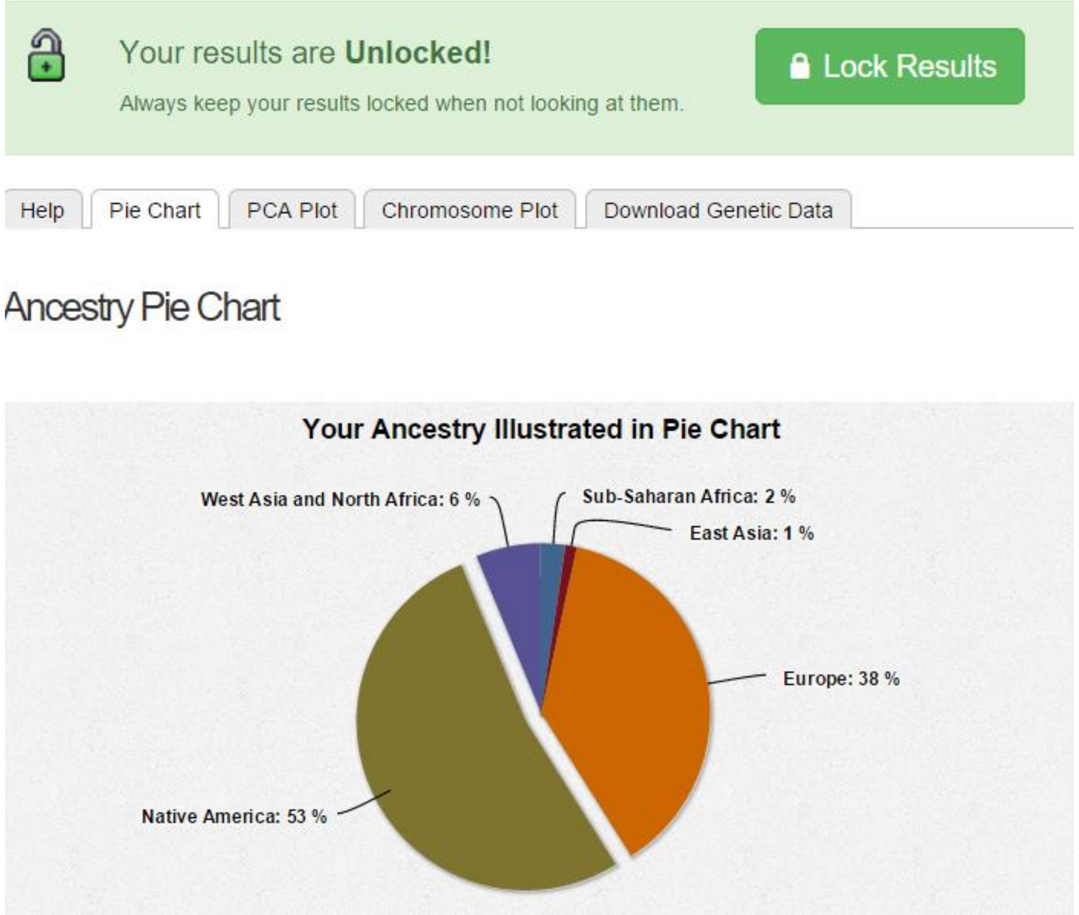
4

DAILY NEWS
ONLINE STUDY LEADS TO BIG BREAKTHROUGH MEDICAL FINDING



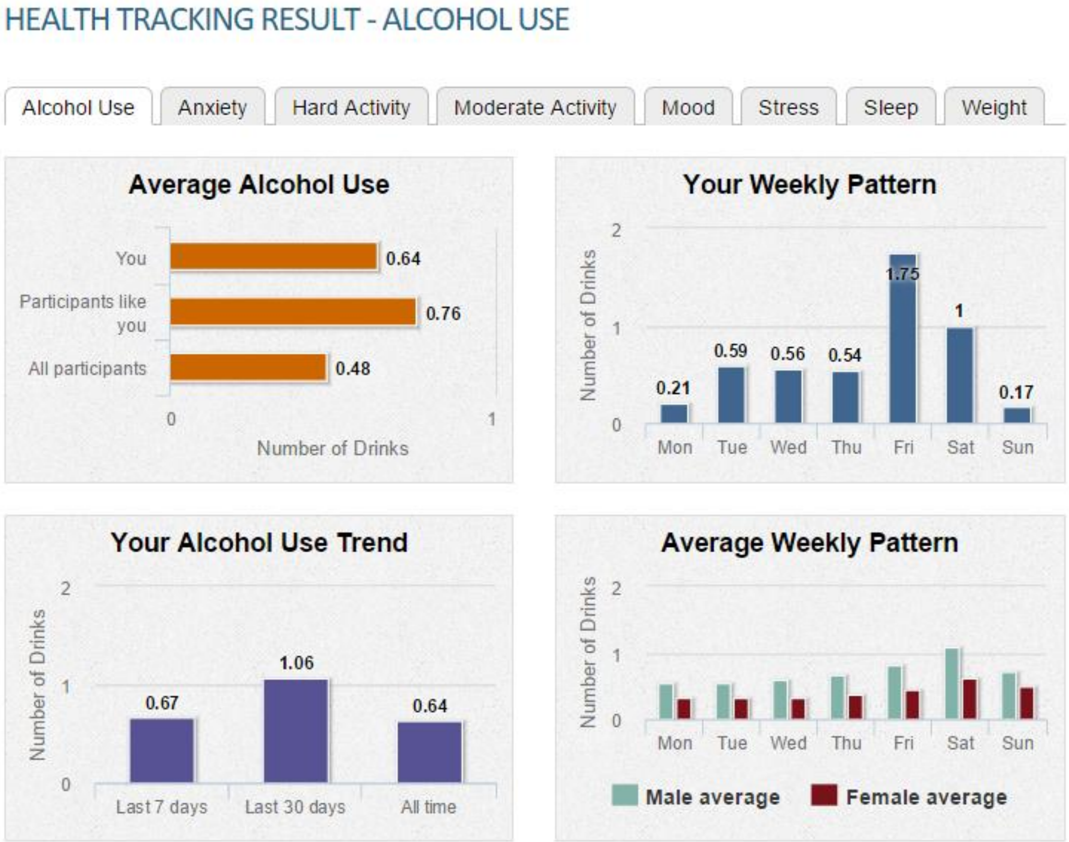
- Exploring new ways to engage populations in research
- Continuous Engagement, Web, Mobile Devices
- Currently, >50,000 participants
- www.genesforgood.org

Return of Results

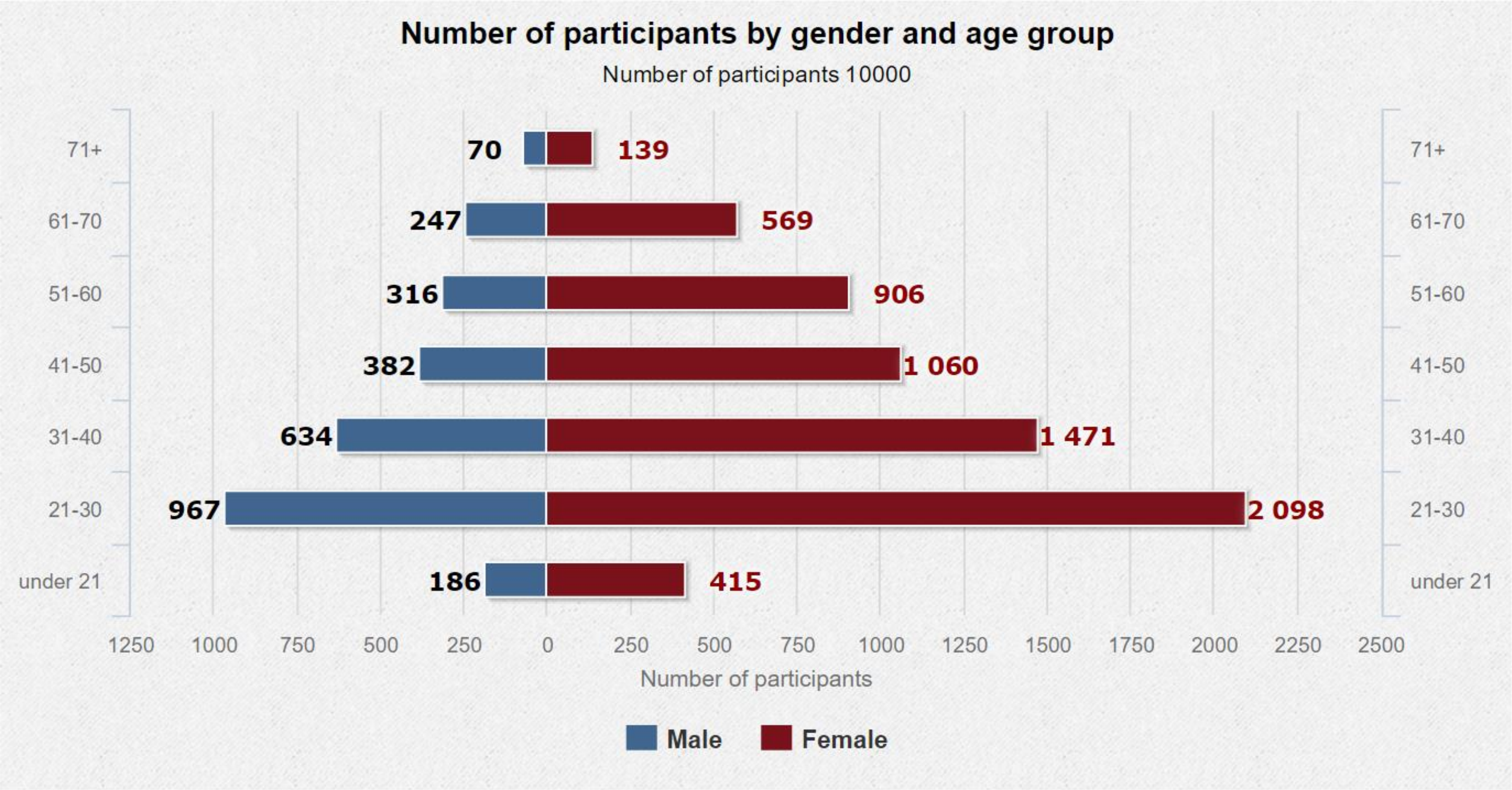


Your results are **Unlocked!**

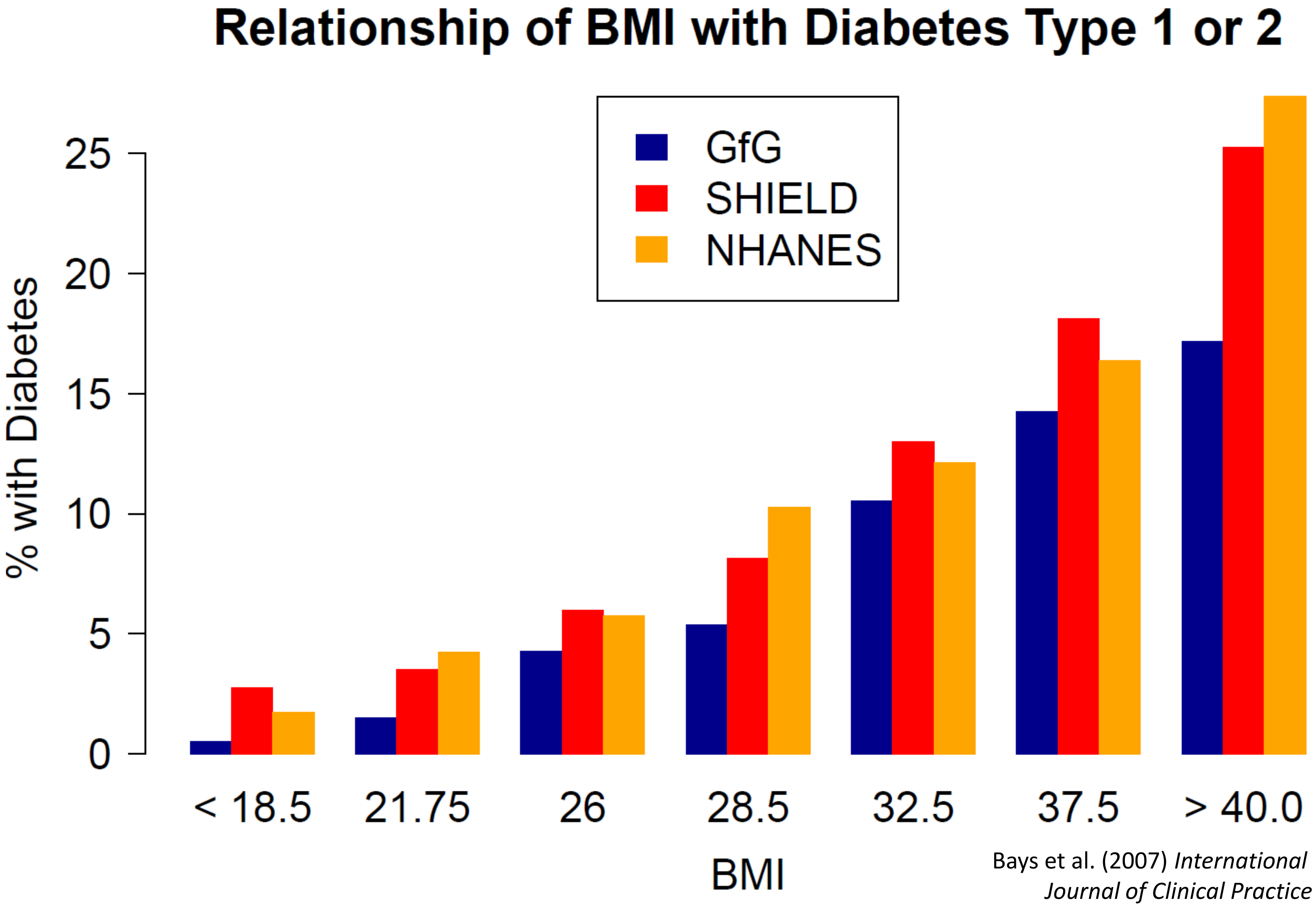
Always keep your results locked when not looking at them.



10,000 Participants…



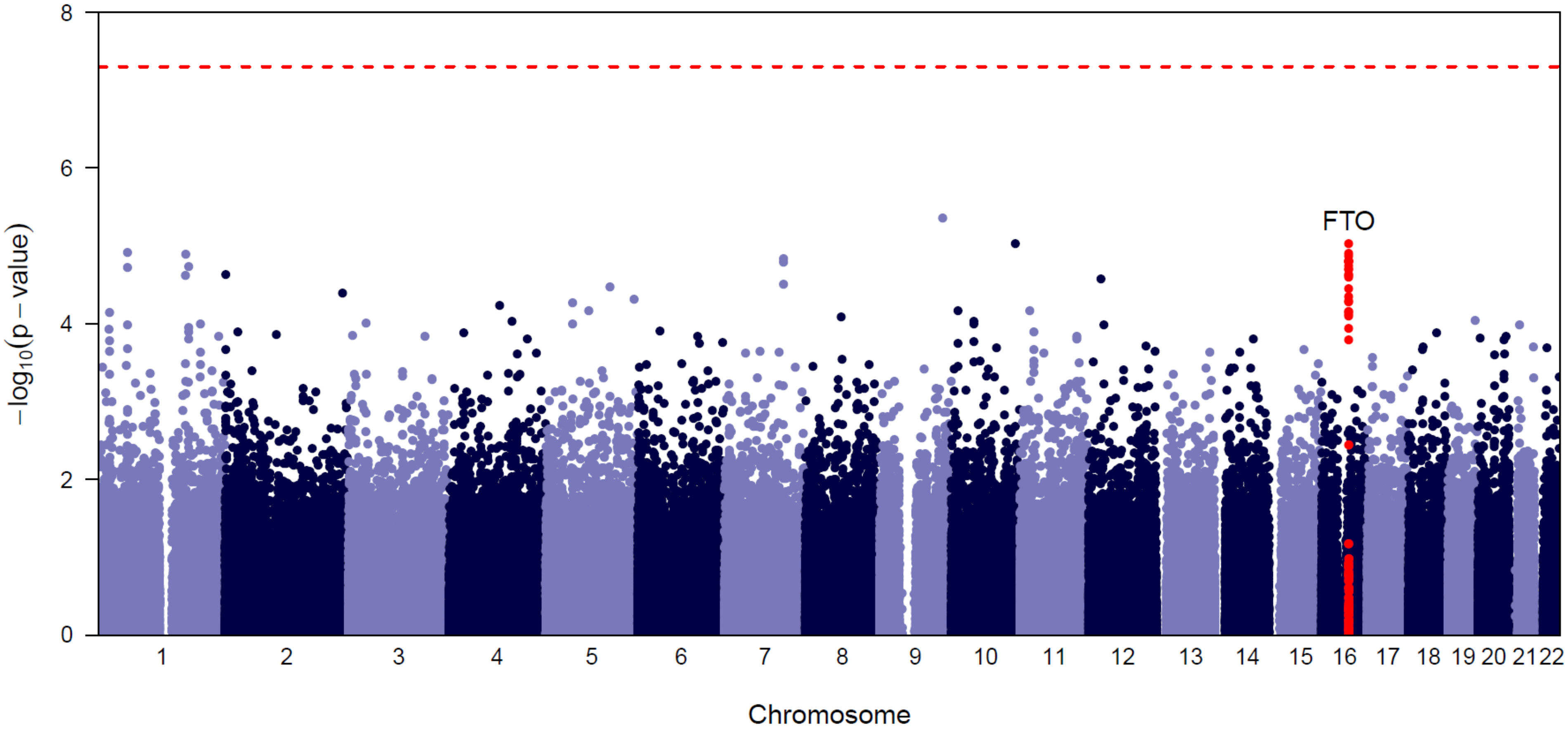
BMI, Age & Diabetes



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

Results: BMI GWAS

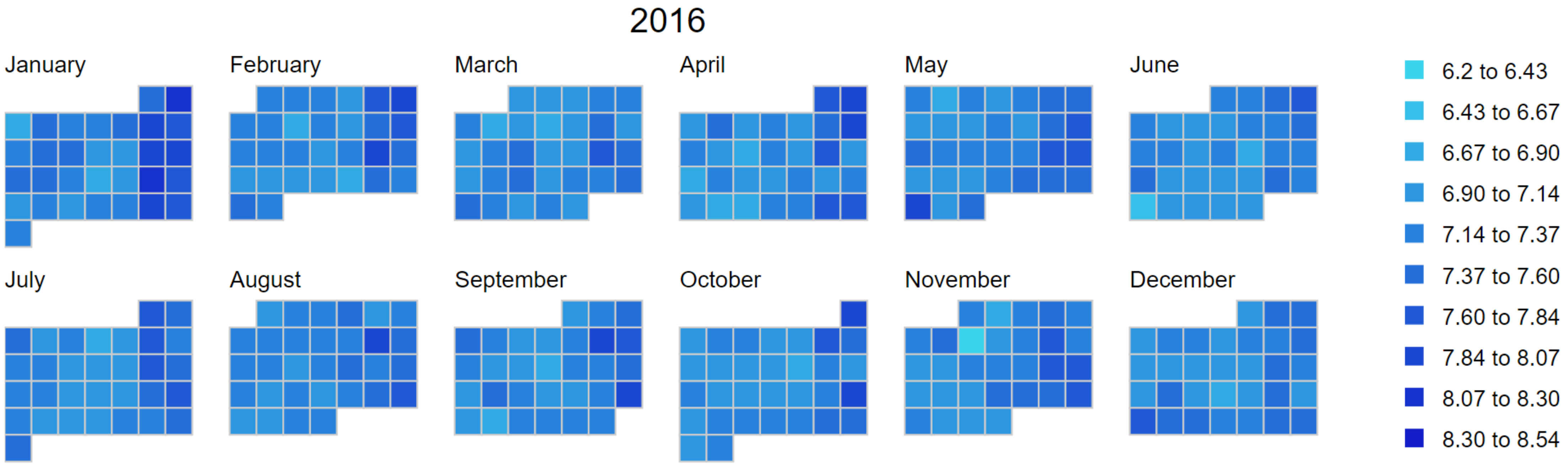
Pheno	n	Chr:Pos	SNP	Gene	Our P	Other P*
BMI	2,851	16:53803574	rs1558902	*FTO*	5×10^{-5}	5×10^{-120}



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

Average Reported Sleep Hours Over a Year
(data from Genes for Good participants)

Anita Pandit



Michigan Genomics Initiative

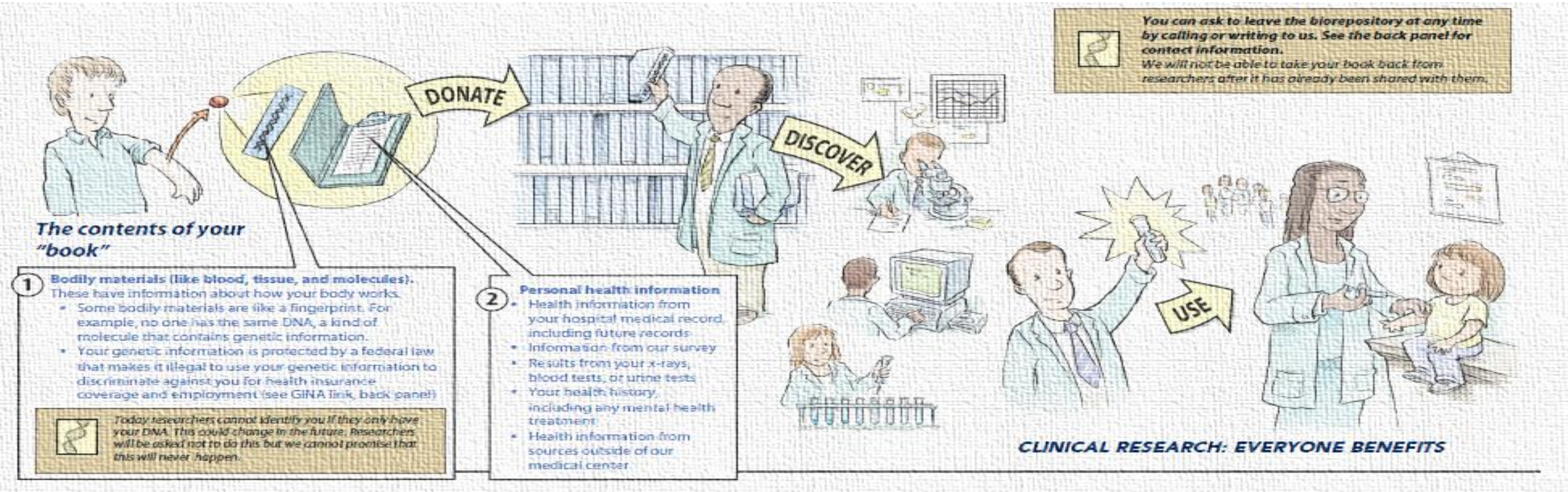
- Combine genetic and electronic health information on 40,000+ patients
- Use genetic information study many traits and diseases
- Build catalog of naturally occurring human knockouts
- Clear, easy to understand consent – full participant buy-in.
- **Team effort: Abecasis (Genetics), Ketherpal (Electronic Health Records), Brummett (Recruitment)**

50 new participants per day

Diverse traits – 40% w/cancer

Speed and improve translation

Key for long term success



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"

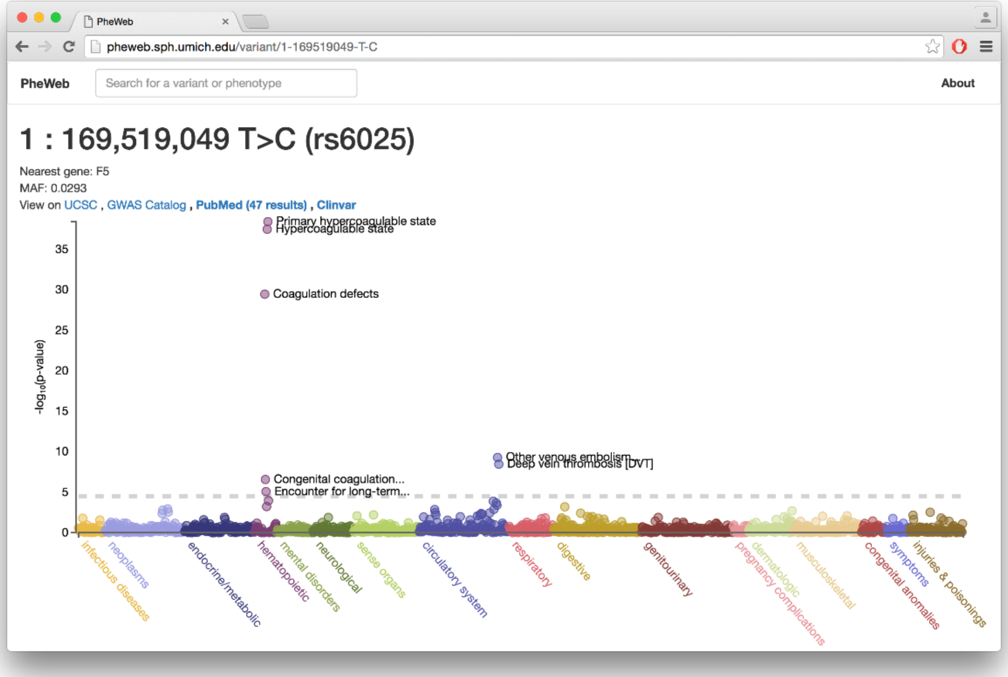
1 **Bodily materials (like blood, tissue, and molecules).** These have information about how your body works.
- Some bodily materials are like a fingerprint. For example, no one has the same DNA, a kind of molecule that contains genetic information.
- Your genetic information is protected by a federal 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

CLINICAL RESEARCH: EVERYONE BENEFITS

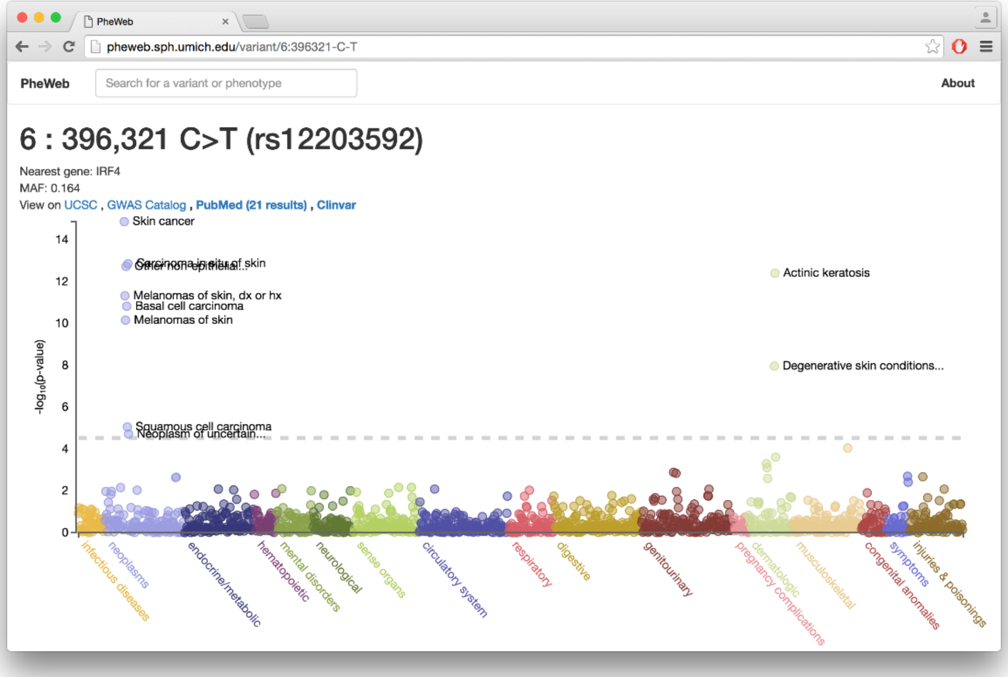
Michigan Genomics Initiative (Freeze 1)
20,000 individuals
7.5 million variants x 1,500 phenotypes



pheweb.sph.umich.edu/variant/1-169519049-T-C

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

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



pheweb.sph.umich.edu/variant/6:396321-C-T

6 : 396,321 C>T (rs12203592)

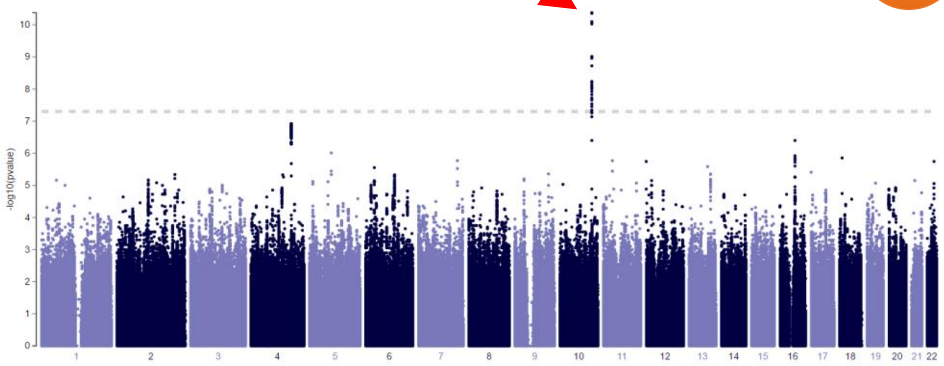
Nearest gene: IRF4
MAF: 0.164
View on UCSC , GWAS Catalog , **PubMed (21 results)** , **Clinvar**

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

1

250.2: Type 2 diabetes

1987 cases, **14906** controls
Category: **endocrine/metabolic**

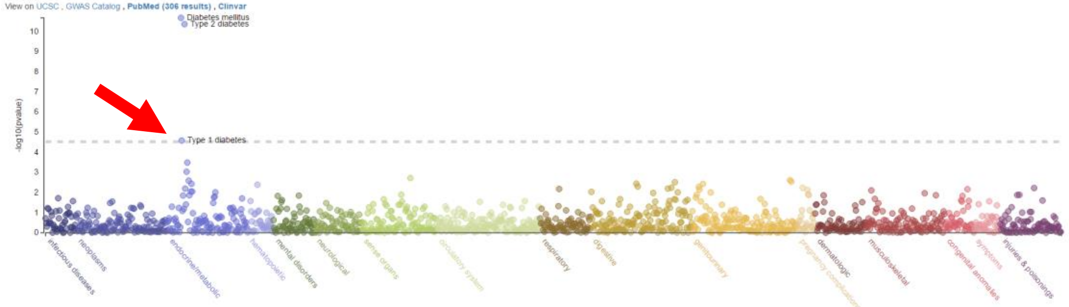


2

near TCF7L2

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

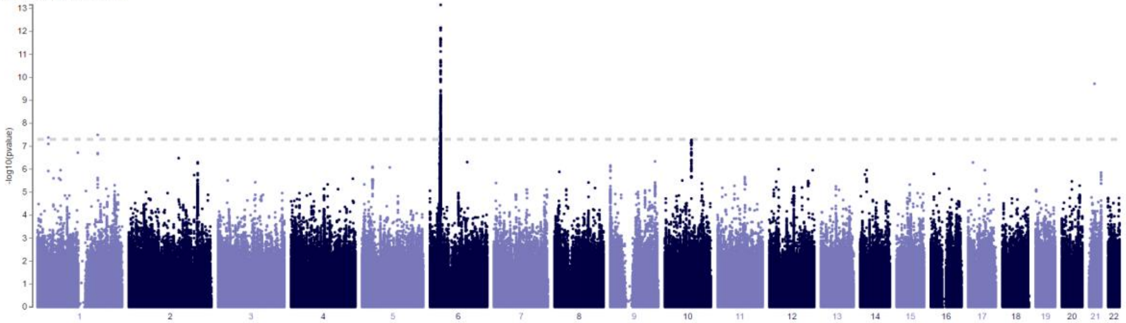
MAF: 0.288
View on UCSC , GWAS Catalog , **PubMed (306 results)** , **Clinvar**



3

250.1: Type 1 diabetes

367 cases, **14906** controls
Category: **endocrine/metabolic**

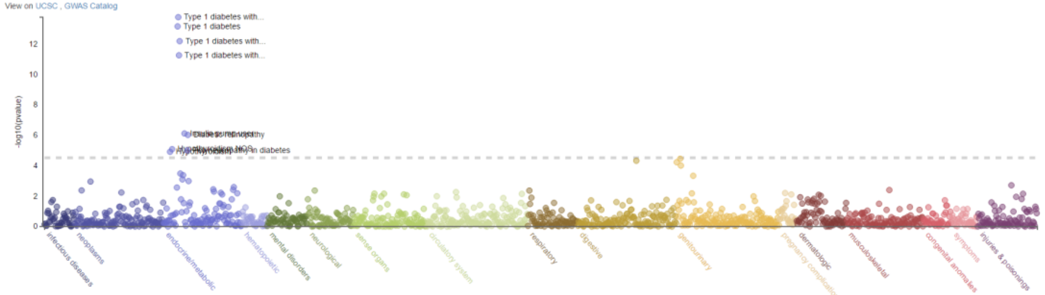


4

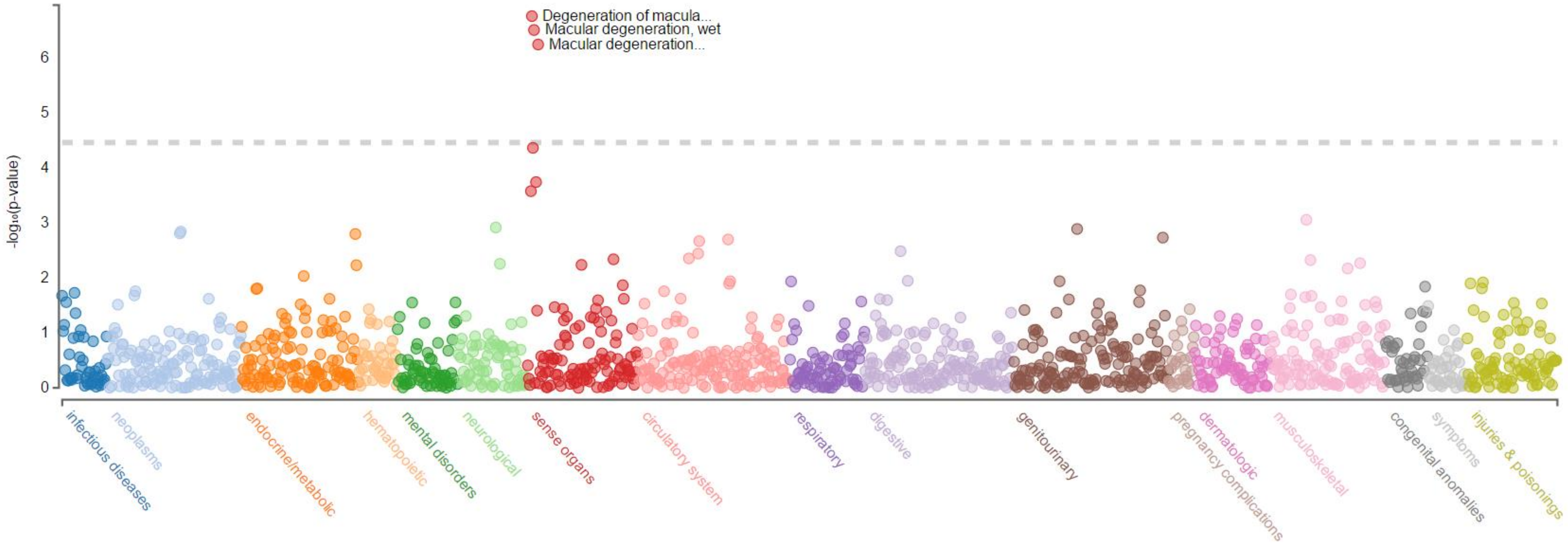
Near HLA-DBQ1

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

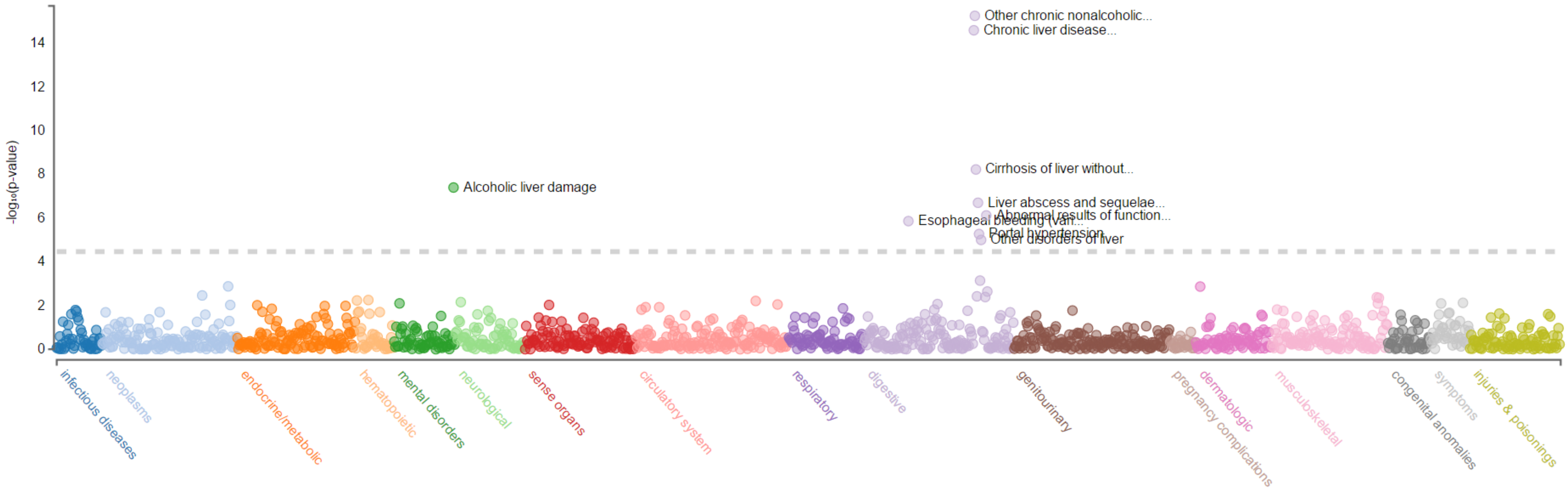
MAF: 0.307
View on UCSC , GWAS Catalog



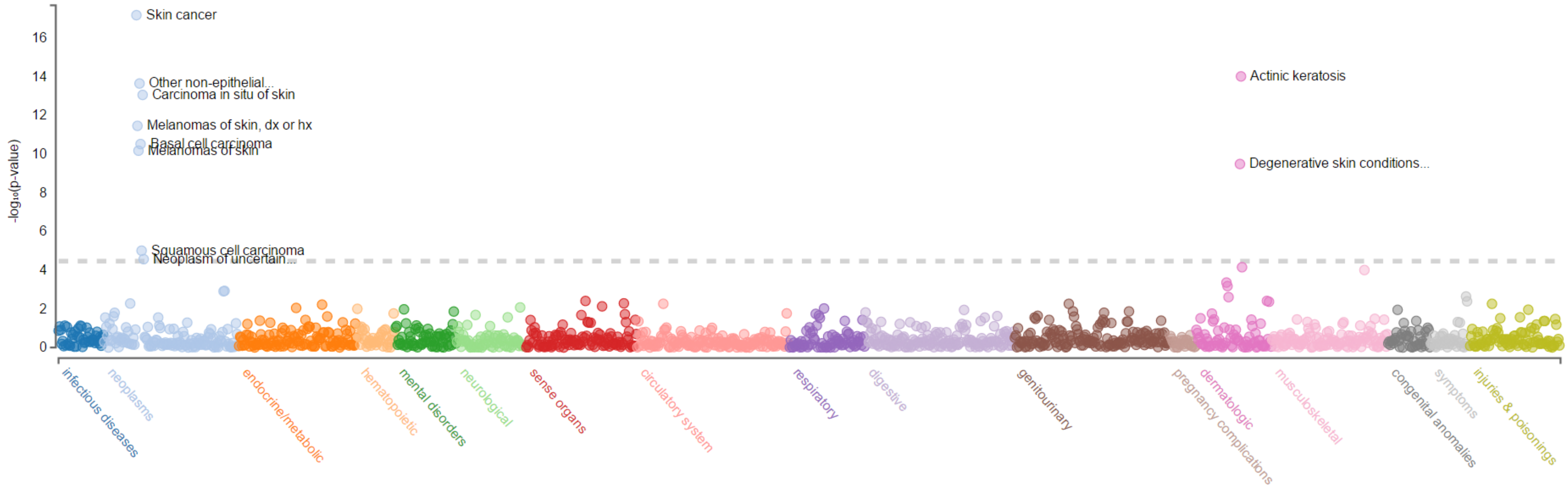
I heard rs10490924 in ARMS2 is associated with macular degeneration …



I heard rs738409 in PNPLA3 is associated with liver disease …



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



The secret of success …

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.