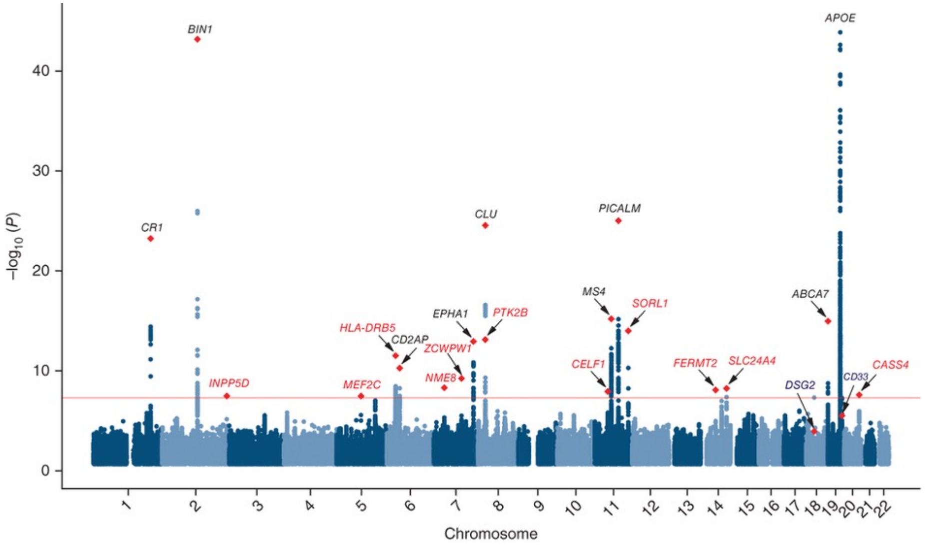
CS 364: Computational Biology

Prof. Sara Mathieson

Fall 2024

Haverford College



High-level Outline

- Genome-Wide Association Studies (GWAS)
- Go over Midterm 2

Notes:

- This week and Tuesday in class: special topics
- Office hours TODAY 2:30-3:30pm (Zubrow)
- In-lab this Thursday and next Thursday: project meetings
- Thursday next week: project presentations

Outline

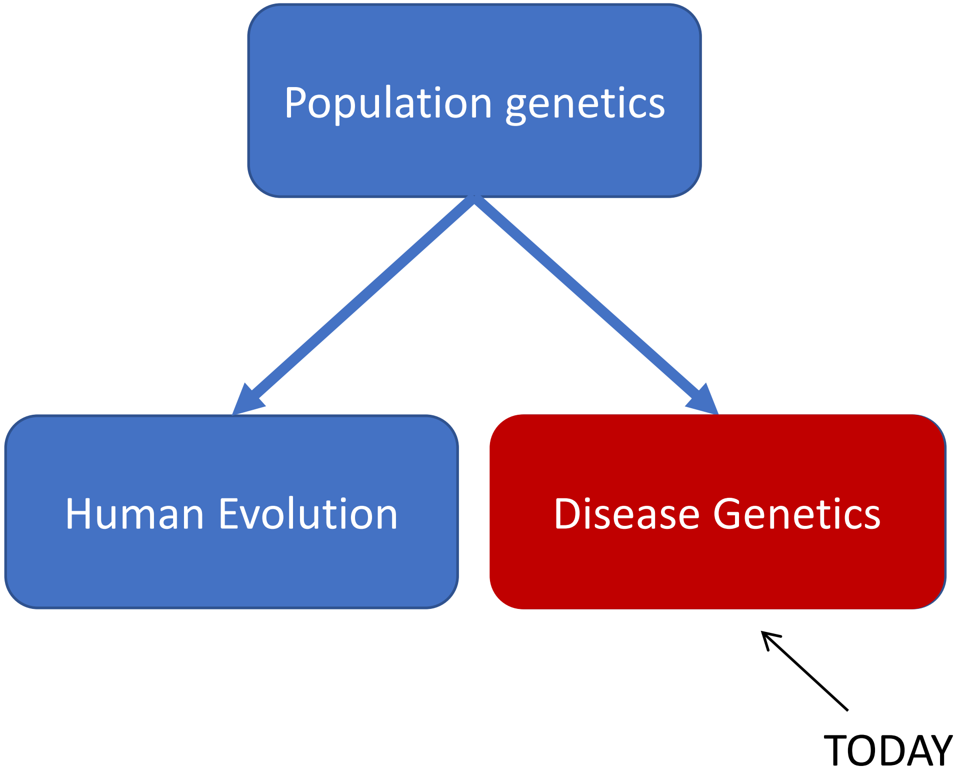
1. **Introduction: what is a GWAS? Why do we do them?**

2. Details and practical applications

3. Drug discovery

4. Trends and active research

Applications of genetic sequencing and method development (in humans)



Population genetics

Human Evolution

Disease Genetics

TODAY

Slide: modified from Iain Mathieson

Human vs nonhuman genetics

Nonhuman	**Human**
Can do experiments	Have to use natural variation
Small sample s	sample sizes (n=1,000,000)
Large effects	Large and small effects
Can easily chose phenotypes	Medical phenotypes usually involve complex biology

"Effect" meaning effect on the phenotype (i.e. the physical manifestation of a trait)

Slide: modified from Iain Mathieson

What is the point?

Two big goals of human genetics:

GWAS

Goal 1: Identify genetic variants (mutations, alleles) that are associated with phenotype, particularly disease

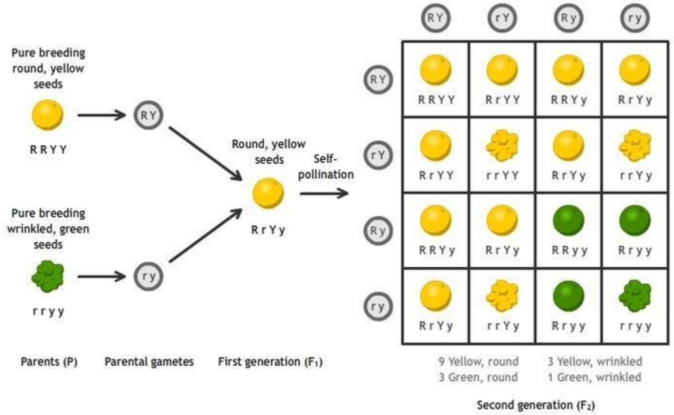
Goal 2: Understand the biological mechanisms through which those variants act.

Hard!

Slide: modified from Iain Mathieson

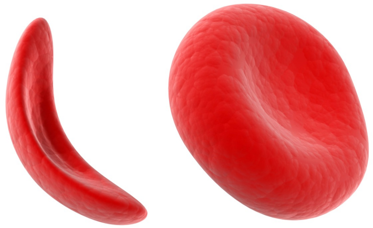
What are we looking for?

Mendelian traits



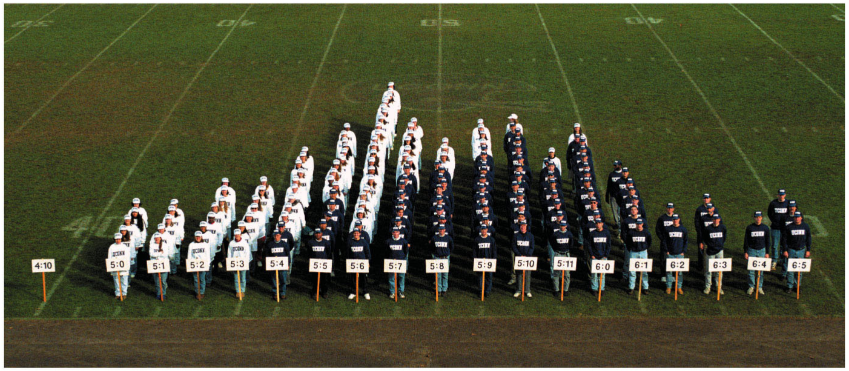
	RY	rY	Ry	ry
RY	RRYY	RrYY	RRYy	RrYy
rY	RrYY	rrYY	RrYy	rrYy
Ry	RRYy	RrYy	RRyy	Rryy
ry	RrYy	rrYy	Rryy	rryy

© 2005-2011 The University of Waikato | www.biotechlearn.org.nz



Thalassemia
Fragile X
Tay-Sachs
Haemophilia

Complex traits



Type II Diabetes

Schizophrenia

Heart disease

Cancer susceptibility

Pigmentation

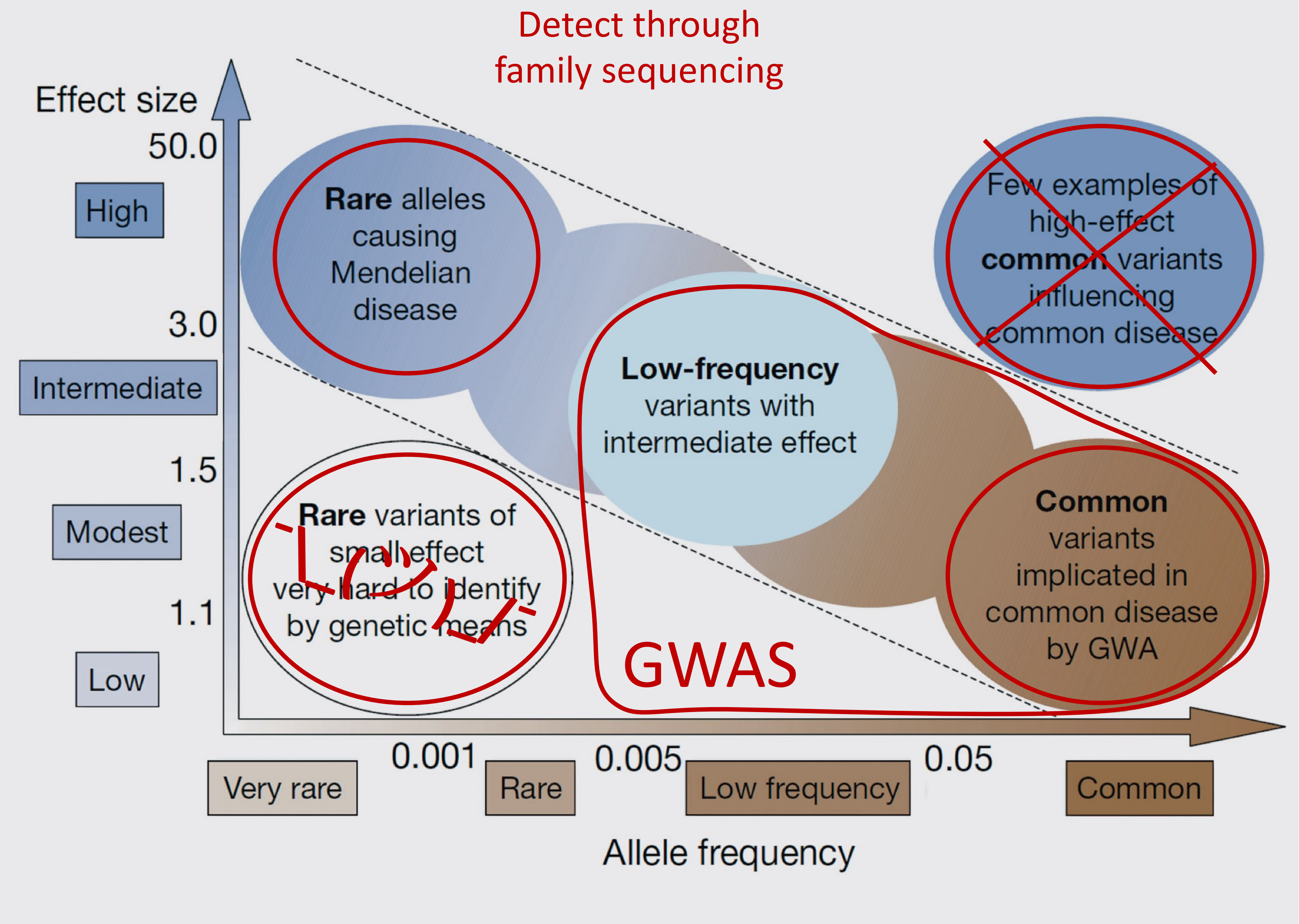
Anxiety

BMI

Cholesterol

Slide: modified from Iain Mathieson

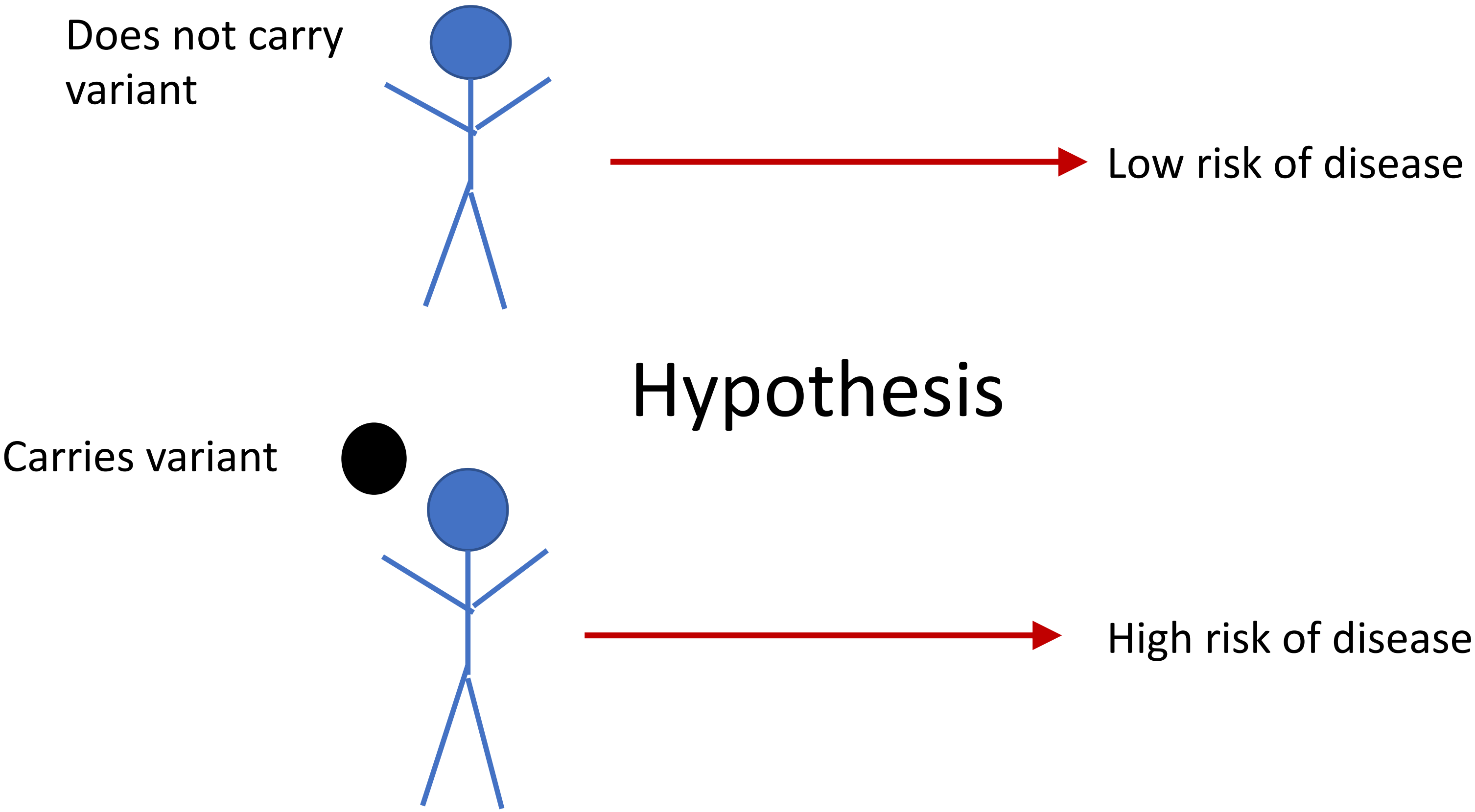
What are we looking for?



Slide: modified from Iain Mathieson

McCarthy et al. *Nat Rev Genetics.* 2008;9:356-369

~~Genome-wide~~ Association Studies



Does not carry variant

Low risk of disease

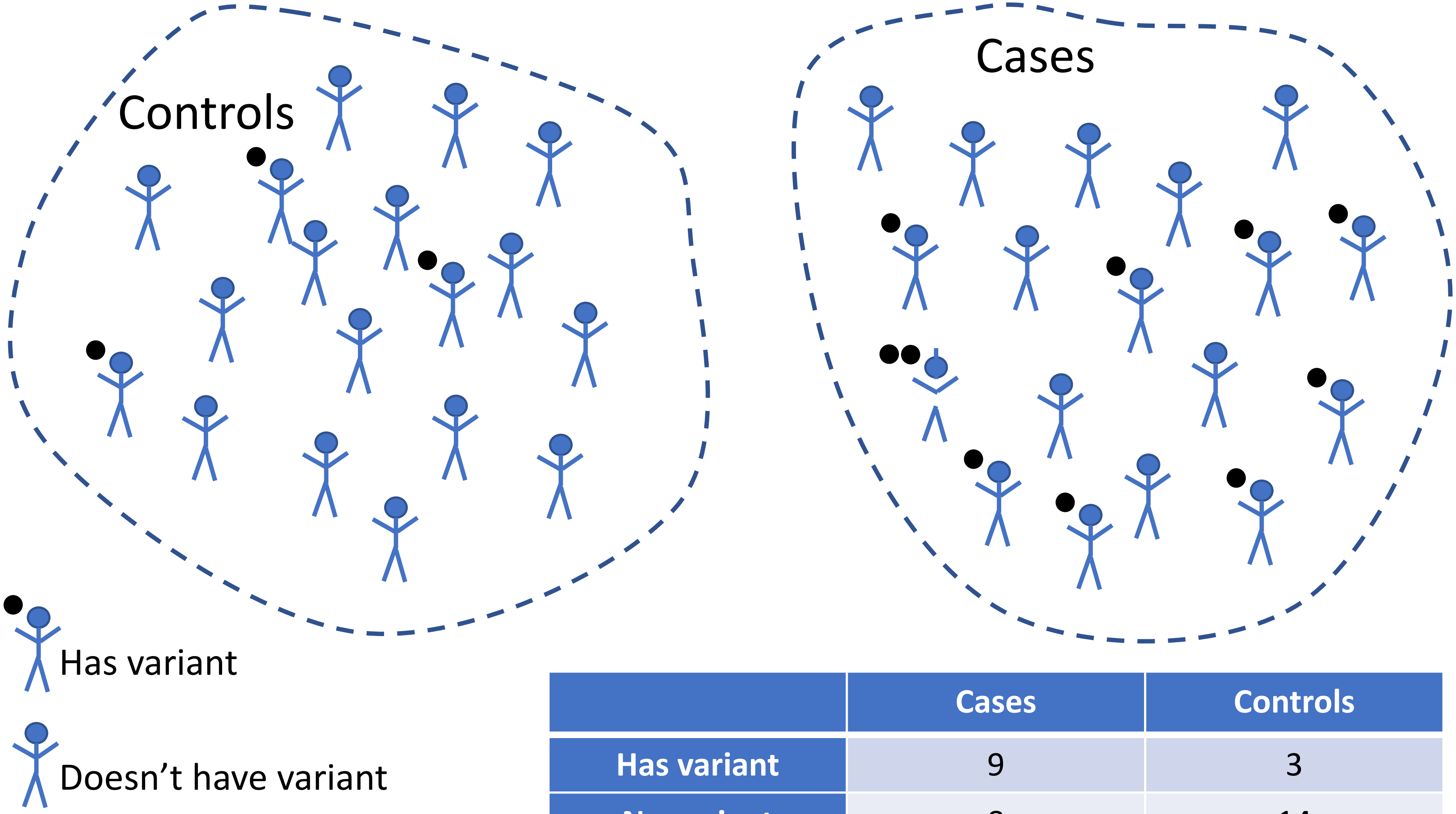
Hypothesis

Carries variant

High risk of disease

Slide: modified from Iain Mathieson

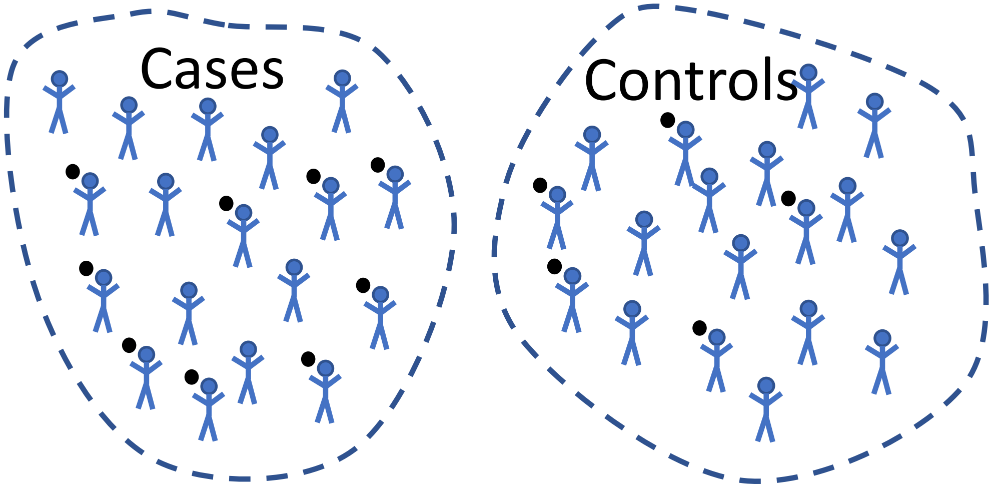
Test hypothesis: Case-control study



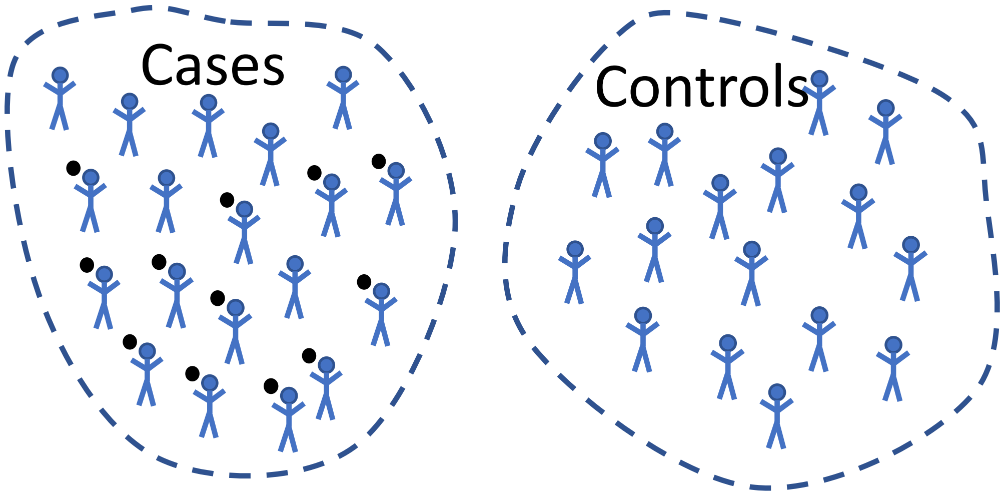
	Cases	Controls
Has variant	9	3
No variant	8	14

Slide: modified from Iain Mathieson

P-value measures non-randomness



P=1
Variant is equally common in cases and controls.



P=0.05
Variant is much more common in one group (here cases).

P=0.05 means that there is a 1 in 20 (5%) chance of seeing a more extreme result, if the variant is not actually associated with the trait.

Slide: modified from Iain Mathieson

P-values: is this result significant?

	Cases	Controls	TOTAL
Has variant	9	3	12
No variant	8	14	22
TOTAL	17	17	34

- Expected number of cases with variant = 17*12/34 = 6
- Expected number of controls with variant = 17*12/34 = 6
- Expected number of cases without variant = 17*22/34 = 11
- Expected number of controls without variant = 17*22/34 = 11

Compute a χ^2 statistic = $\sum \frac{(observed - expected)^2}{expected}$

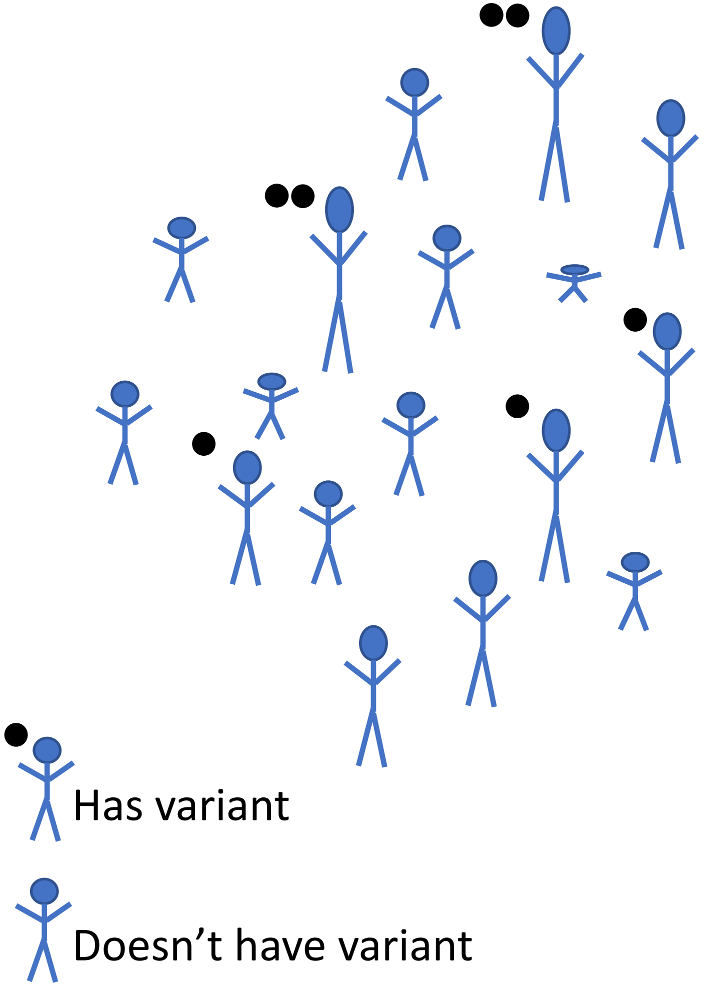
$$= \frac{(9-6)^2}{6} + \frac{(3-6)^2}{6} + \frac{(8-11)^2}{11} + \frac{(14-11)^2}{11}$$

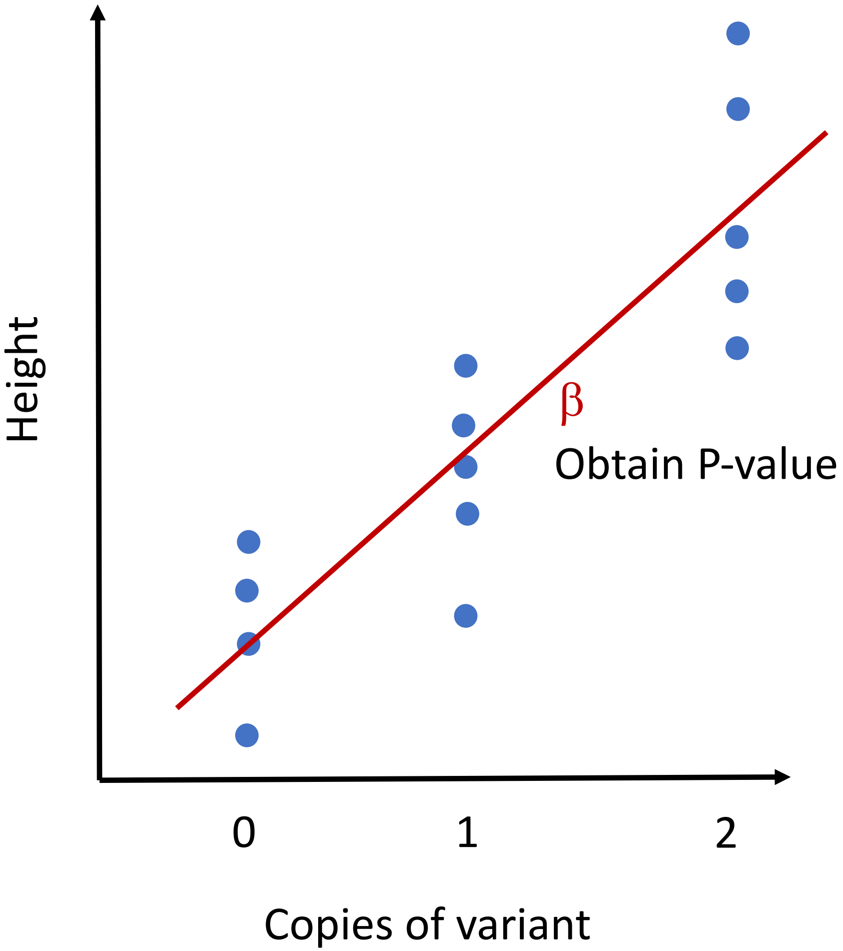
= 4.636

Yes, at a 0.05 significance level

Is this significant? $\boxed{P=0.0313}$ *[R code: 1-pchisq(4.636, df=1)]*

Continuous (“quantitative”) traits





Slide: modified from Iain Mathieson

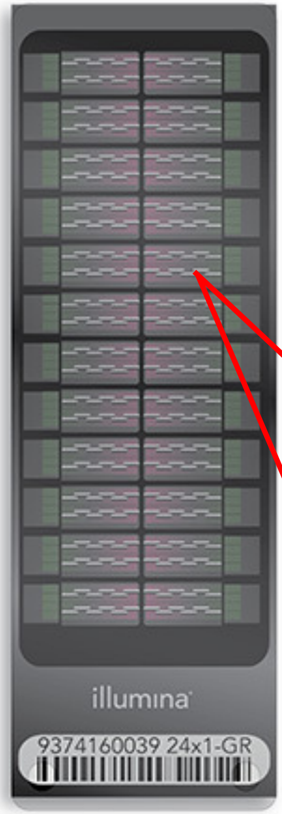
The problem with candidate gene studies

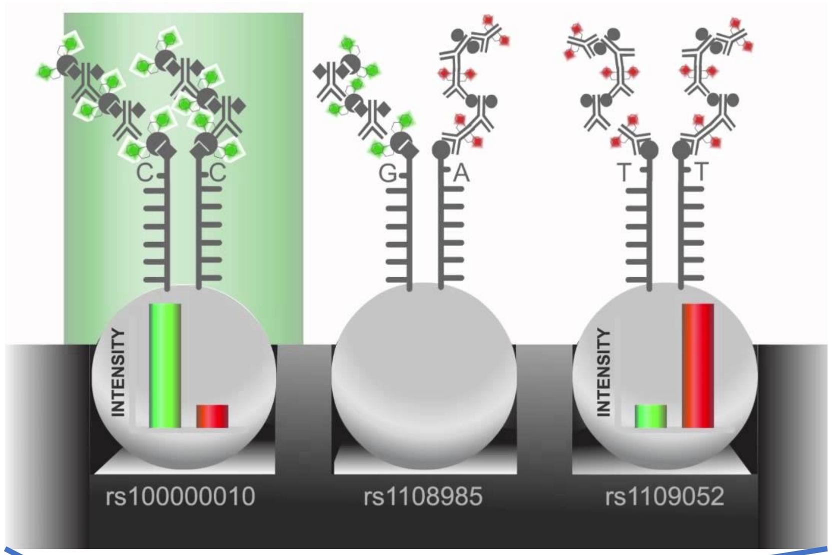
This study design led to a large number of spurious findings for several reasons

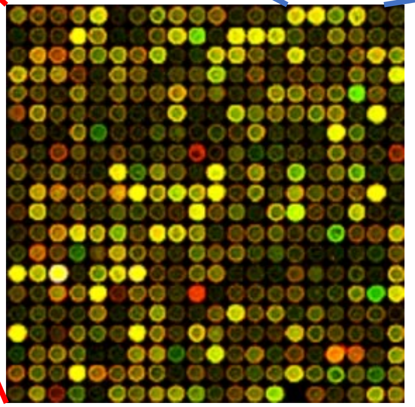
- Guessing the right genes is hard
- Underpowered studies – turns out effects are very small
- Population structure – lots of false positive findings
- Other statistical issues (multiple testing etc…)
- Publication bias

Solution: Test lots of variants in the whole genome ("genome-wide") with very large sample (N = 10,000 – millions).

SNP Genotyping Arrays







Genotype millions of variants
Cost ~$100 per-sample

Genome-wide Association Studies

- Lots of people. Number of people depends on the effect size. Most GWAS today have n=10,000-1,000,000.

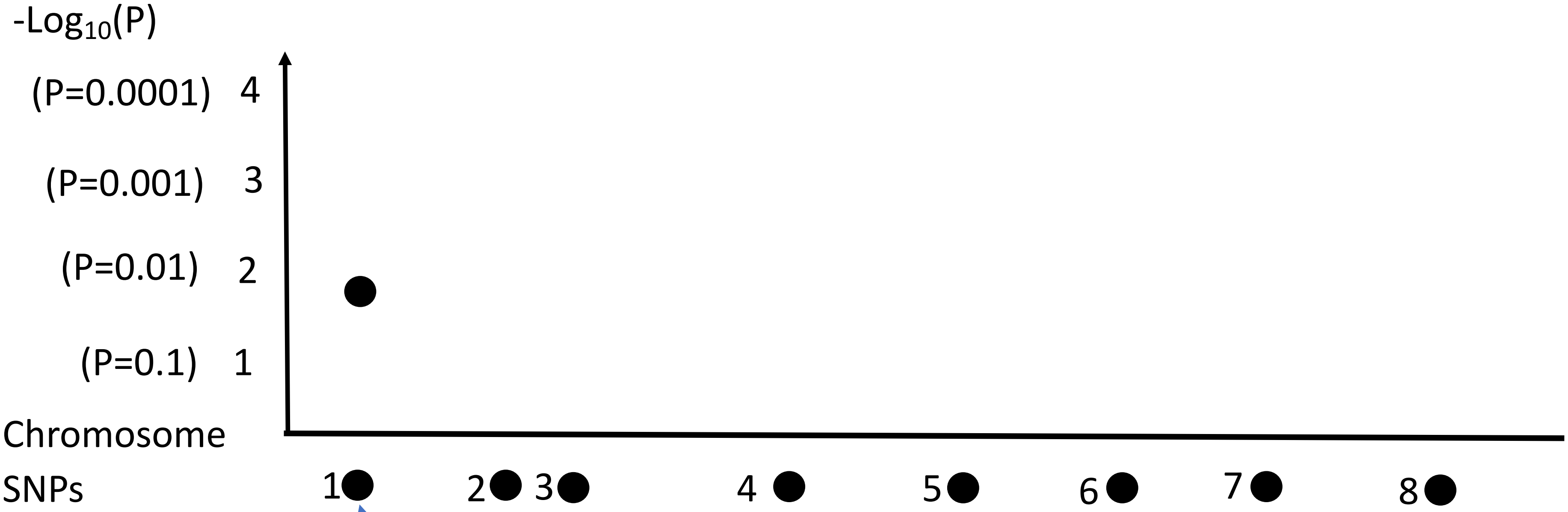
- Genome-wide data. Usually SNP-array data. Typically 100,000-1,000,000 SNPs across the genome

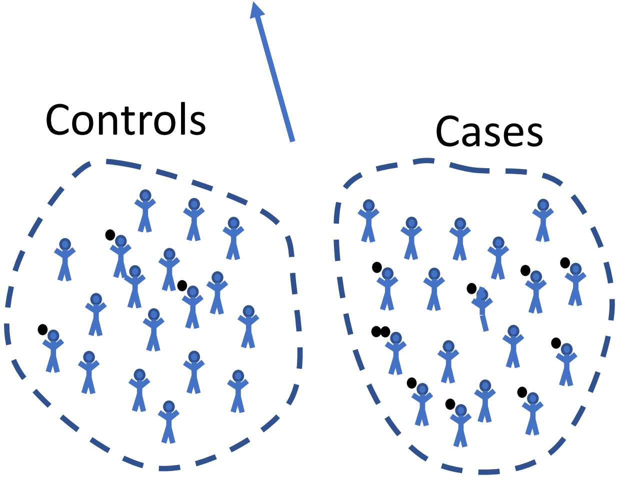
- A phenotype. Anything! GWAS have been carried out for 3,357 traits.

GWAS catalog https://www.ebi.ac.uk/gwas/

Slide: modified from Iain Mathieson

Genome-wide Association Studies



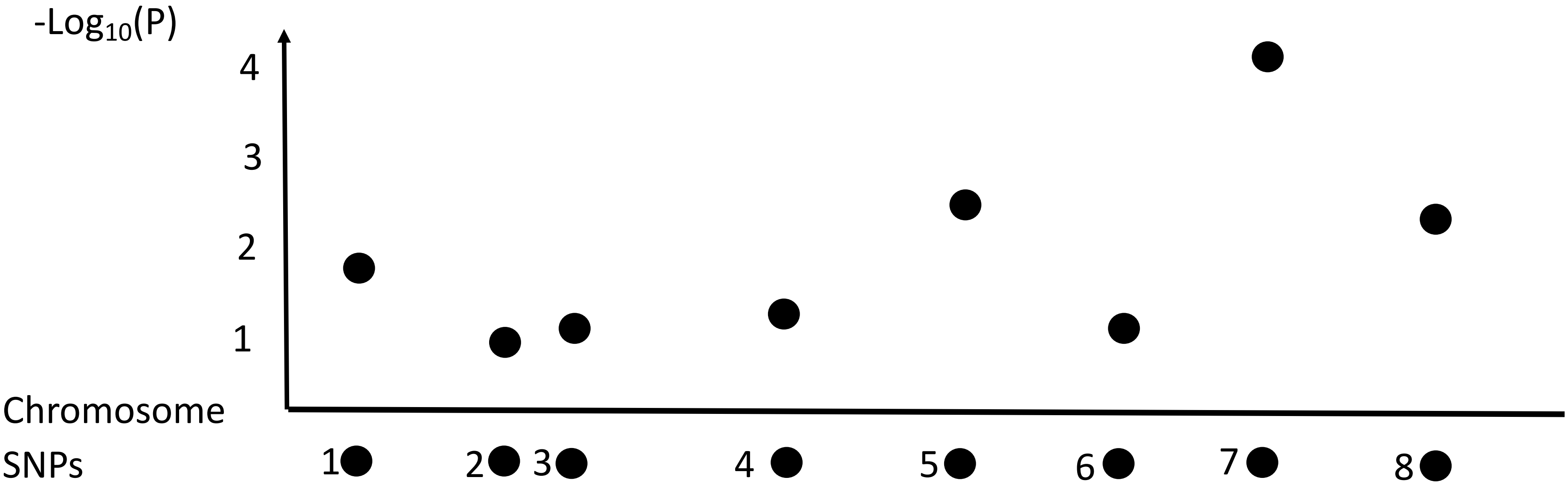


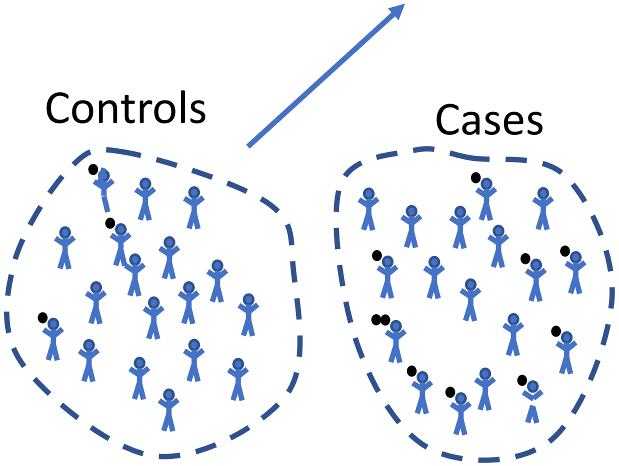
	Cases	Controls
Has variant	9	3
No variant	8	14

P=0.03

Slide: modified from Iain Mathieson

Genome-wide Association Studies

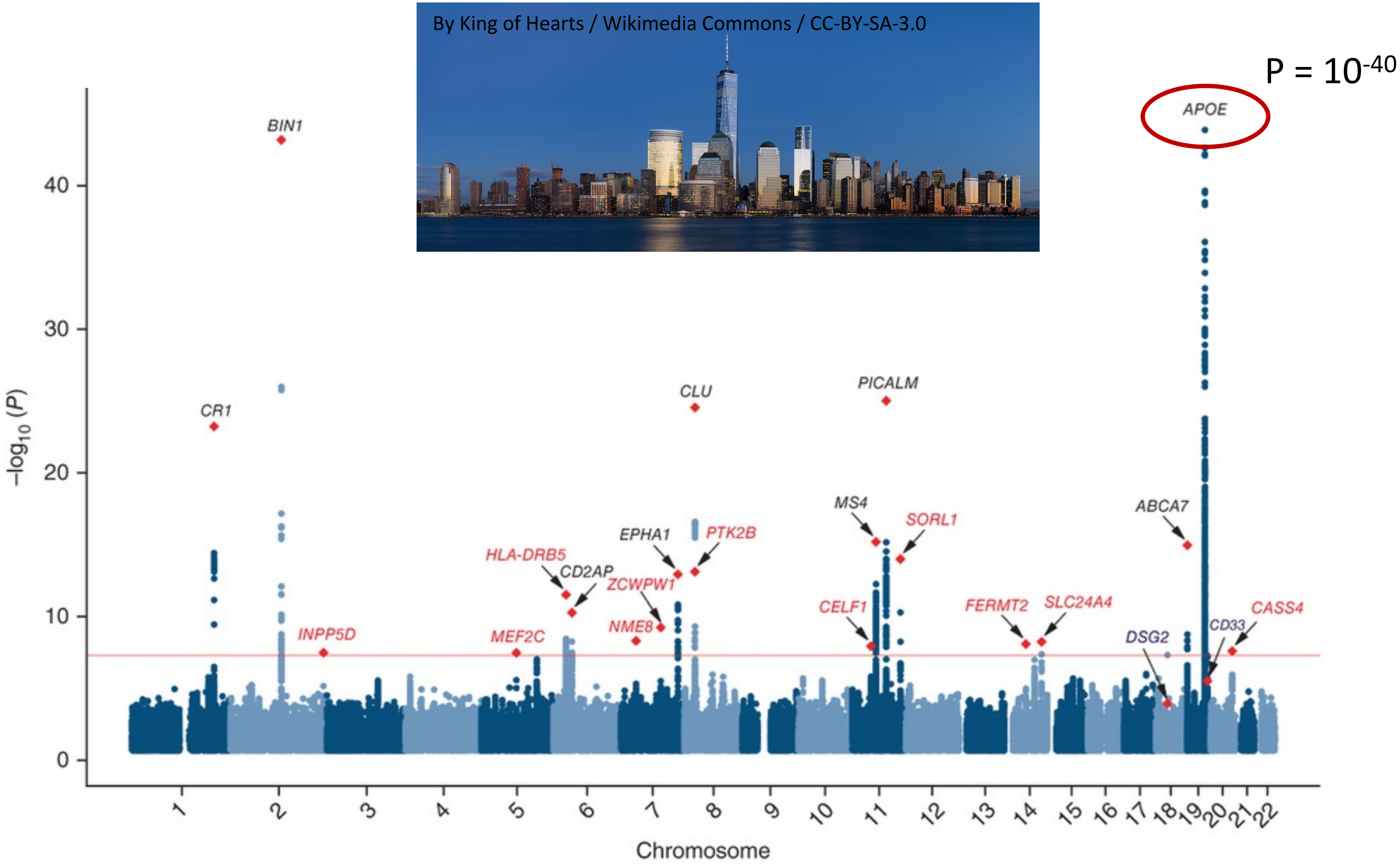




	Cases	Controls
Has variant	6	6
No variant	8	12

Slide: modified from Iain Mathieson

Manhattan plot


By King of Hearts / Wikimedia Commons / CC-BY-SA-3.0

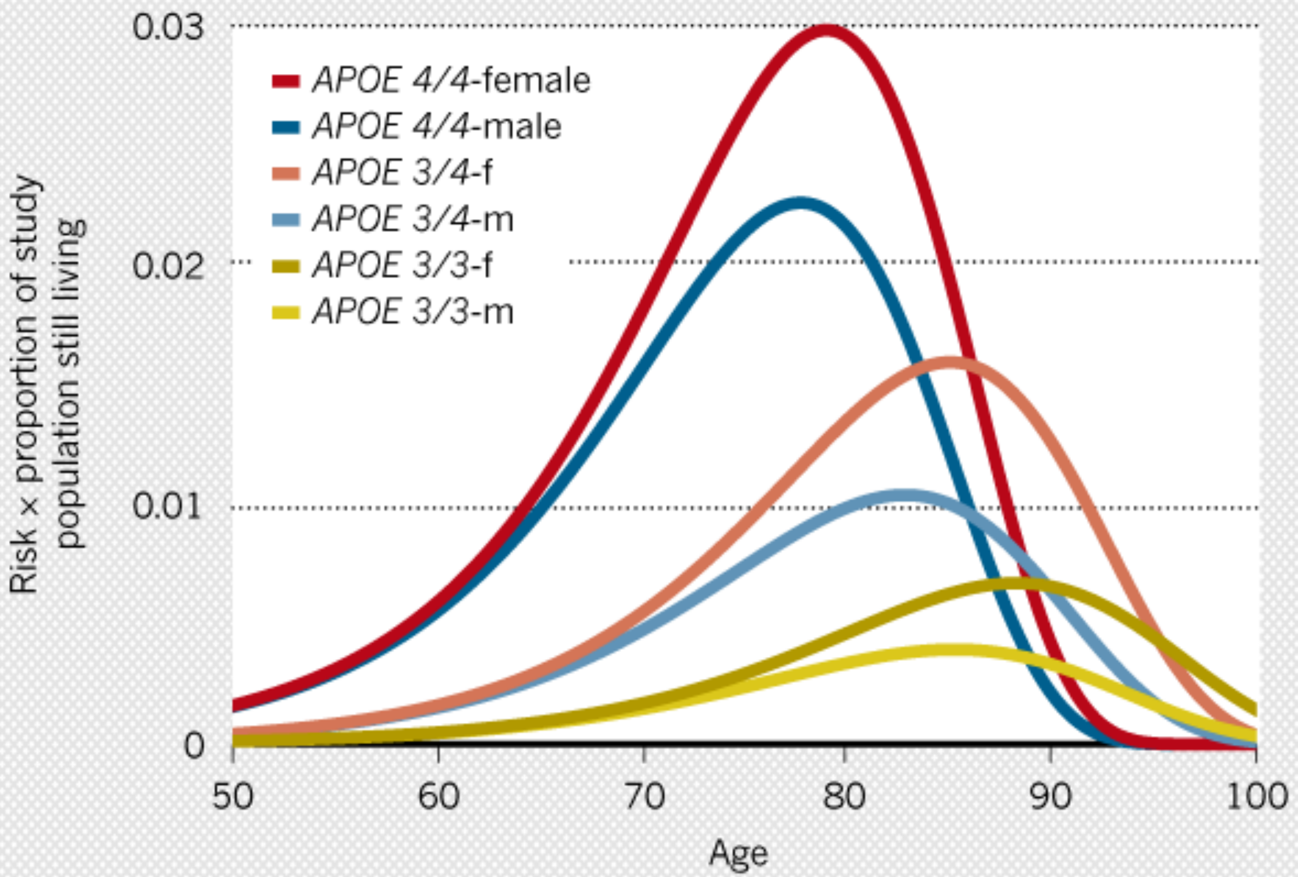
P = 10^{-40}

Slide: modified from Iain Mathieson

“Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease” (2013)

RISKY INHERITANCE

People who carry the gene variant *APOE4* tend to develop Alzheimer's at a younger age than those with two copies of *APOE3*.



Nature 2014

Outline

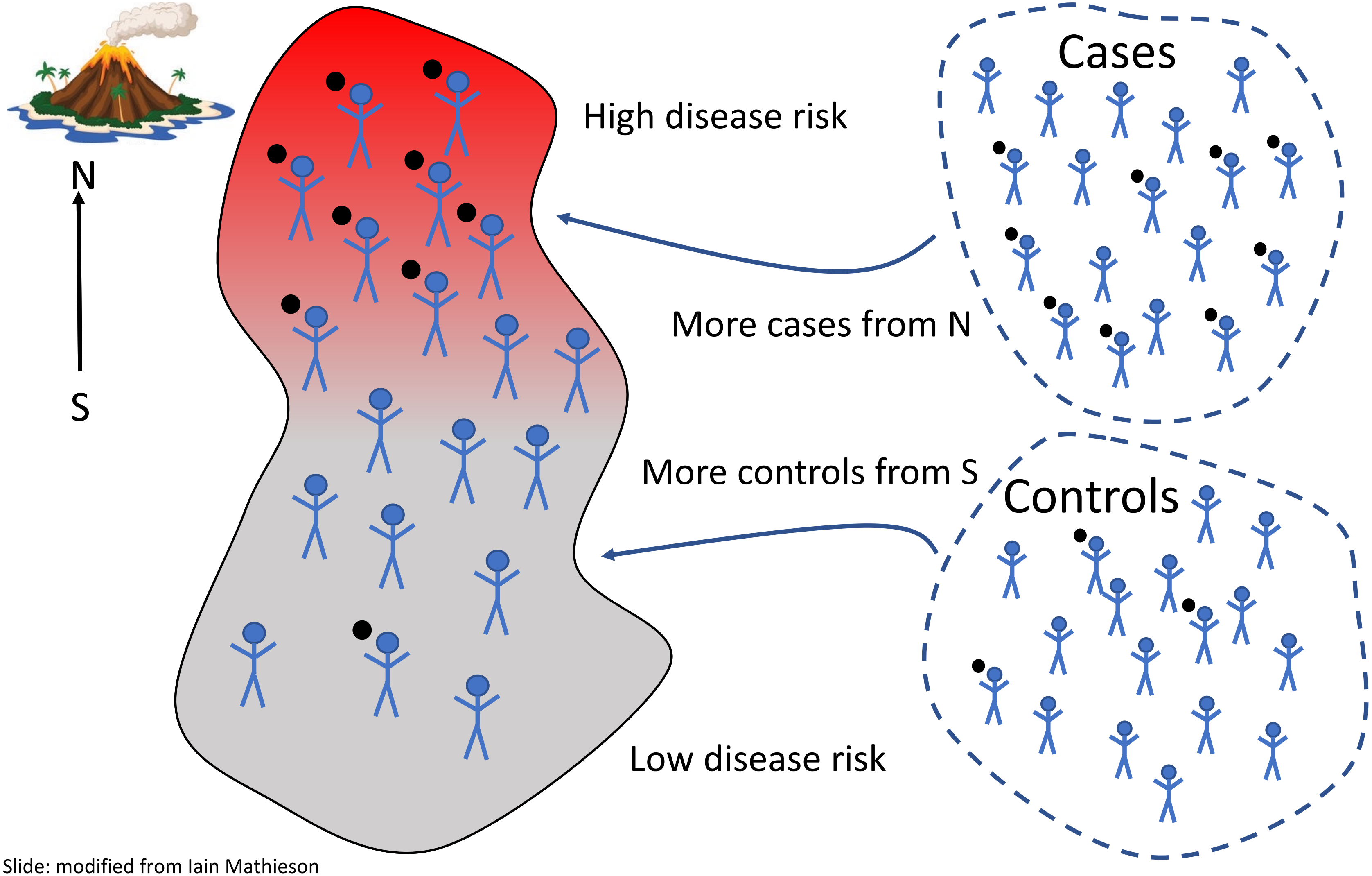
1. Introduction: what is a GWAS? Why do we do them?

2. Details and practical applications

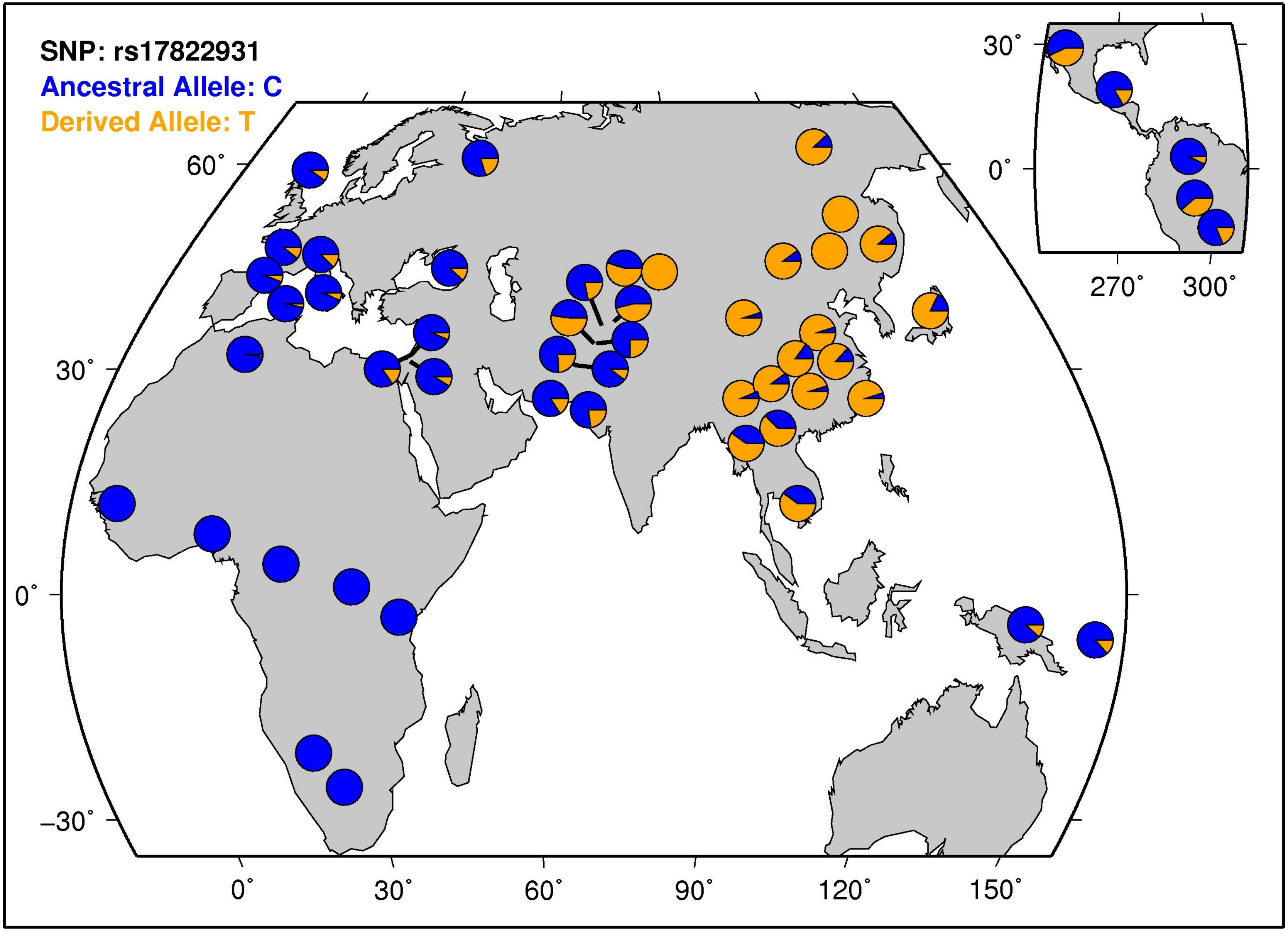
3. Drug discovery

4. Trends and active research

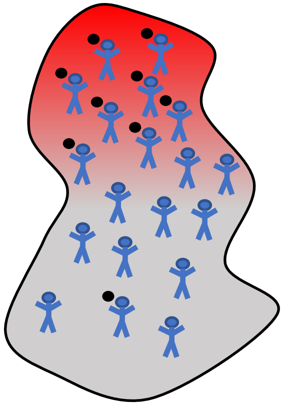
Population structure



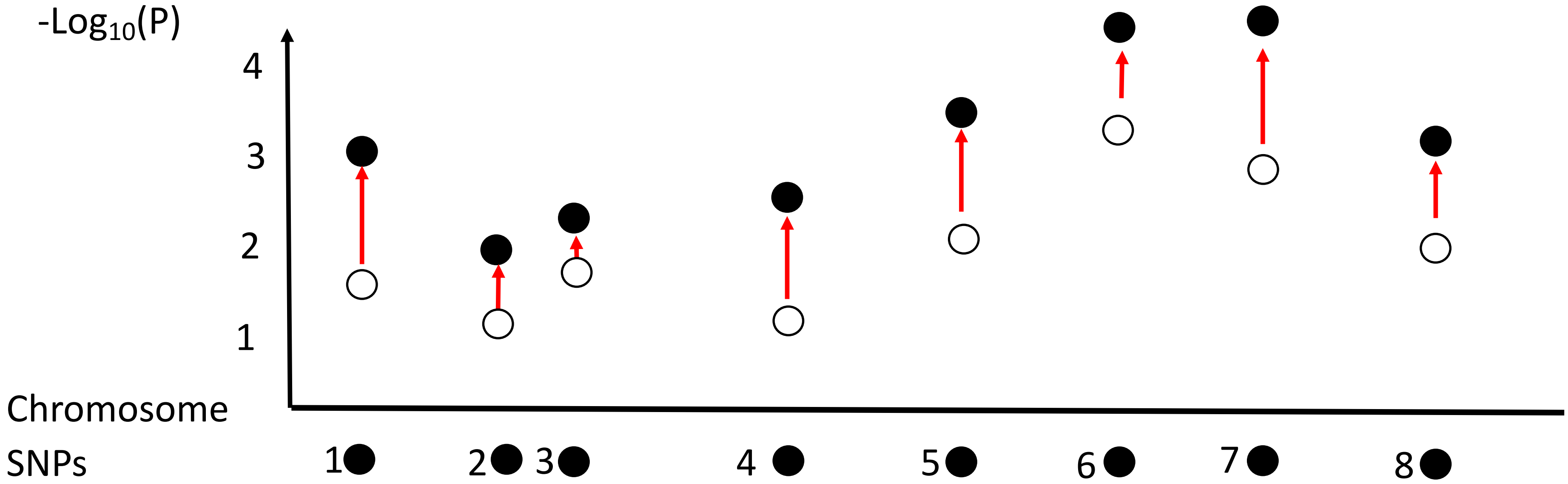
Slide: modified from Iain Mathieson



Slide: modified from Iain Mathieson



Population structure



Slide: modified from Iain Mathieson

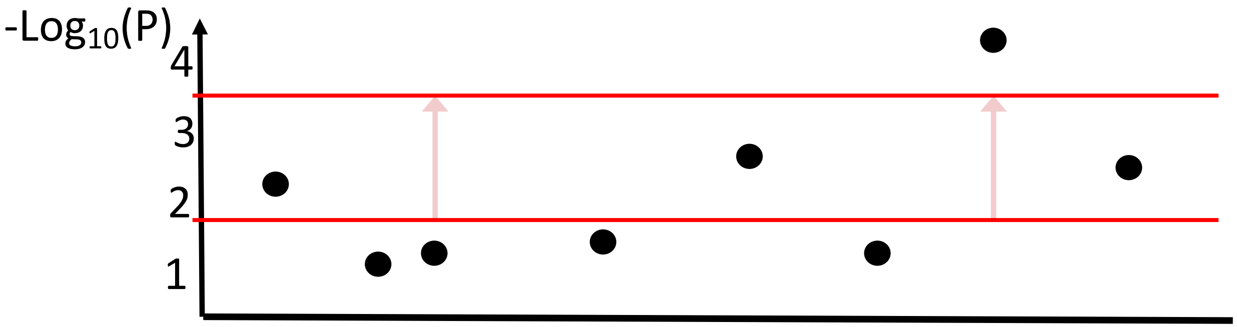
Multiple testing

$P < 0.05$ means that there is less than a 5% chance that the result happens by chance.

	Cases	Controls
Has variant	9	3
No variant	8	14

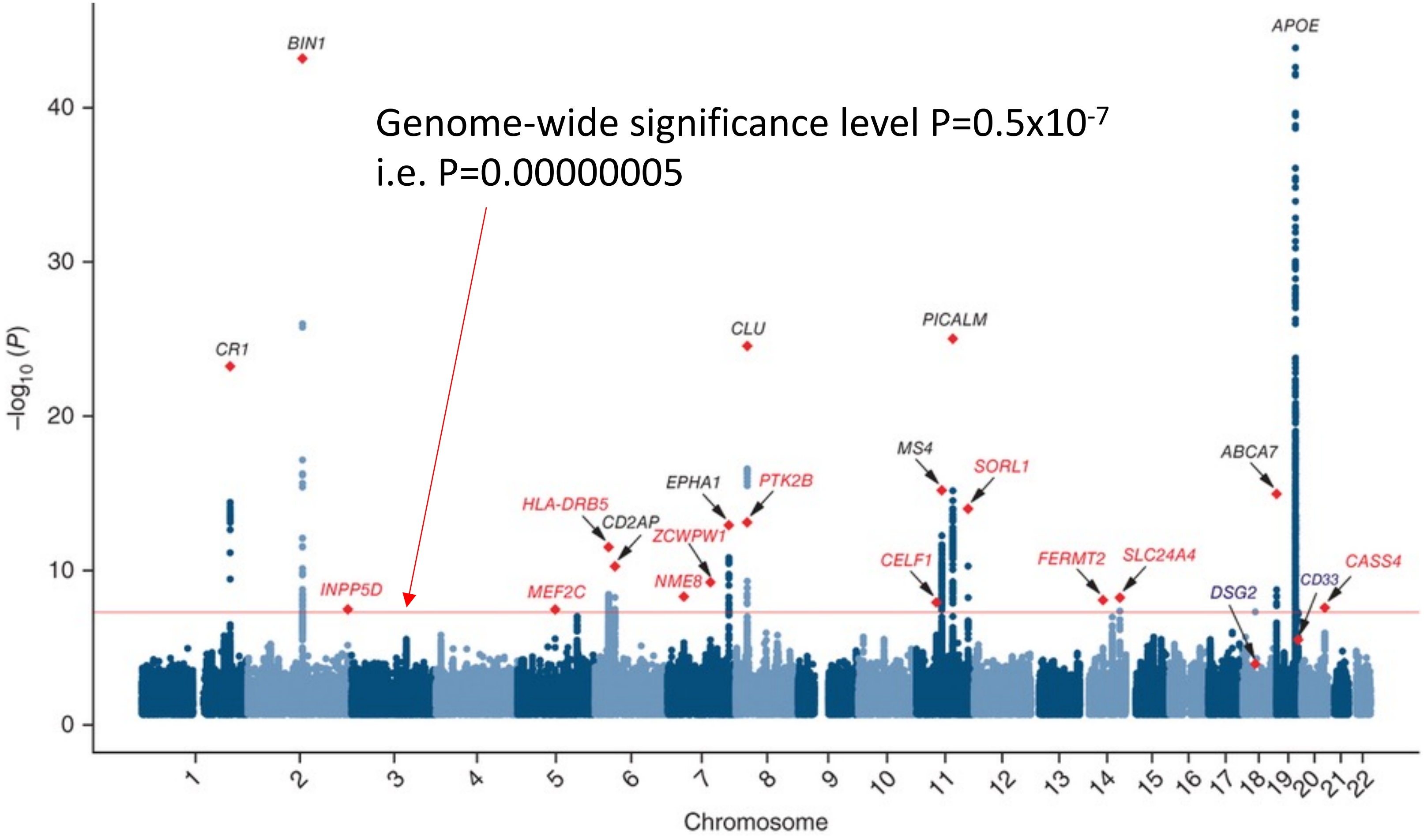
P=0.03

But if you try lots of tests, then the chance that one of them is significant is high
So we need to only look at things that are extremely significant



Slide: modified from Iain Mathieson

Multiple testing



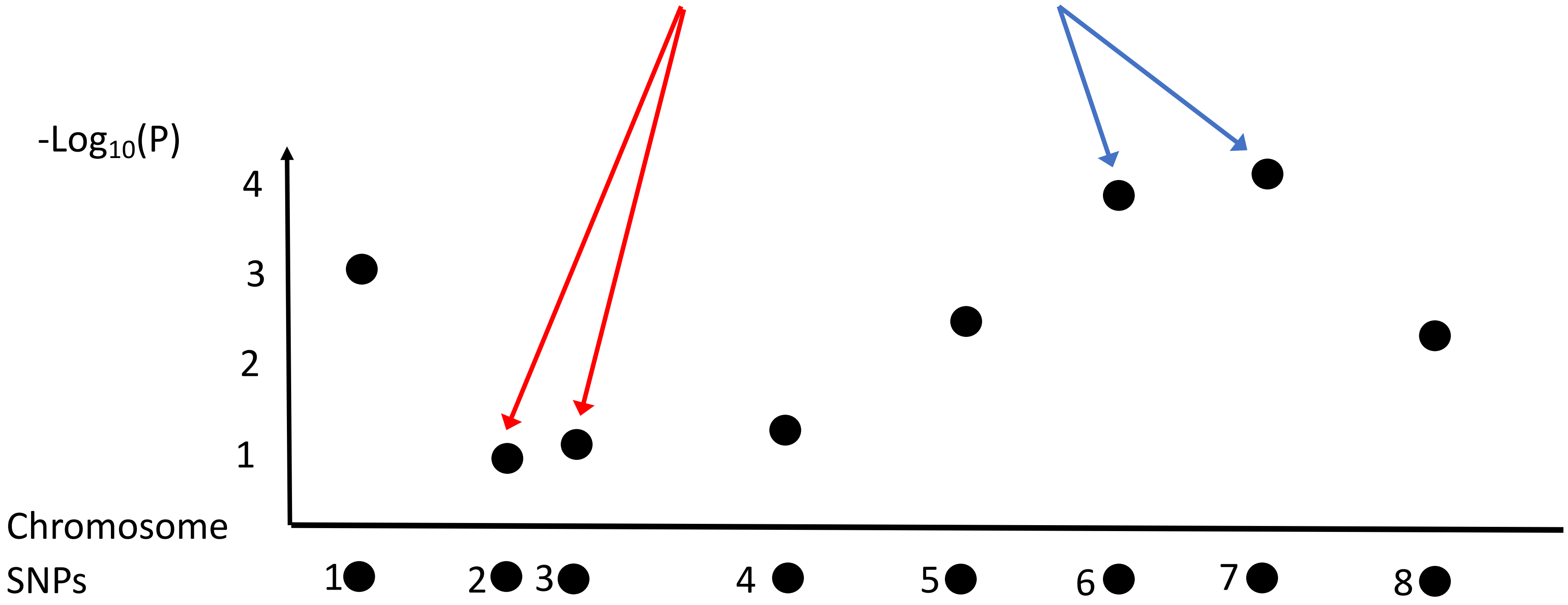
Genome-wide significance level $P=0.5\text{x}10^{-7}$

i.e. P=0.00000005

Slide: modified from Iain Mathieson

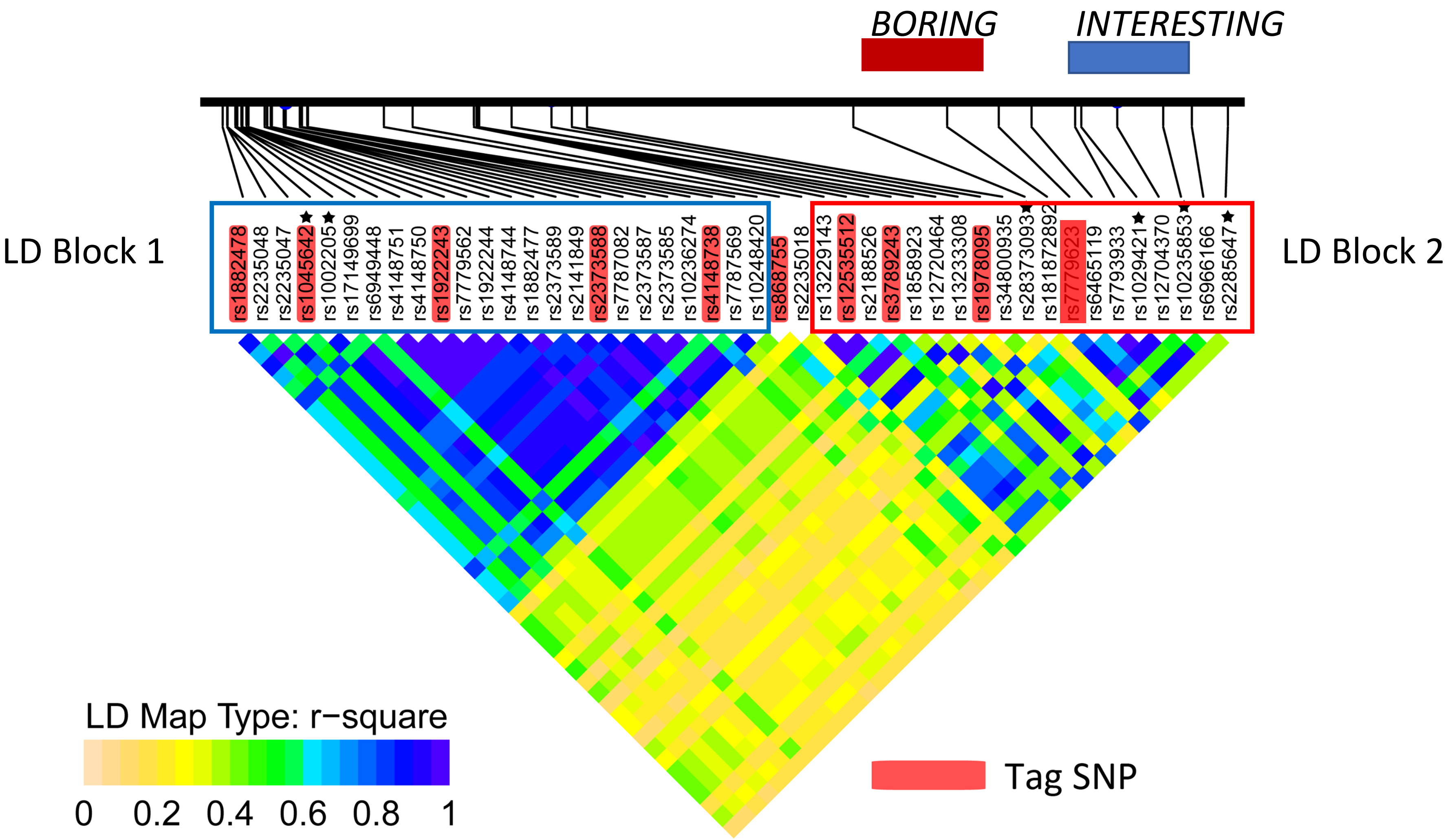
Linkage

SNPs that are close together tend to behave similarly, not broken up by recombination!



Slide: modified from Iain Mathieson

Linkage blocks and tag SNPs



Slide: modified from Iain Mathieson

Shou et al 2011 http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0046295

Fine-Mapping

LD Block

ACGA**T**ACCAG**C**ACGATTCGAT**C**TTT**A**CGCGGGG**C**ACGTAGCTA**G**CTGTGTGT**C**ACGTG**C**TAGCTG

Fine-Mapping

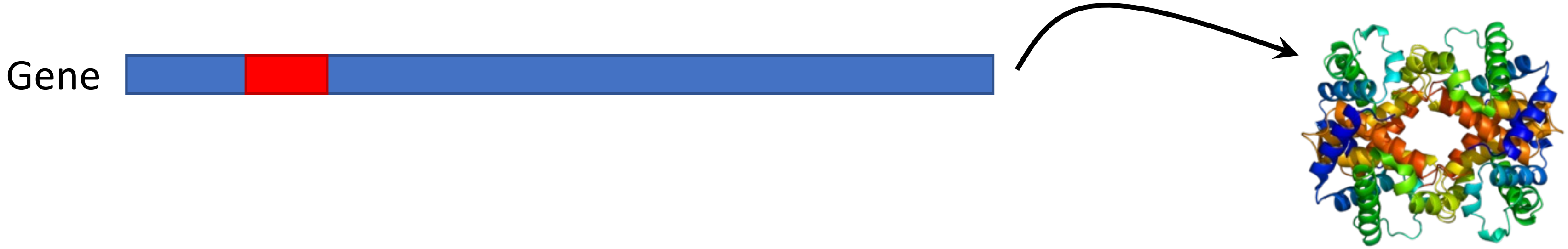
LD Block

ACGA**T**ACCAG**C**ACGATTCGAT**C**TTT**A**CGCGGGG**C**ACGTAGCTA**G**CTGTGTGT**C**ACGTG**C**TAGCTG

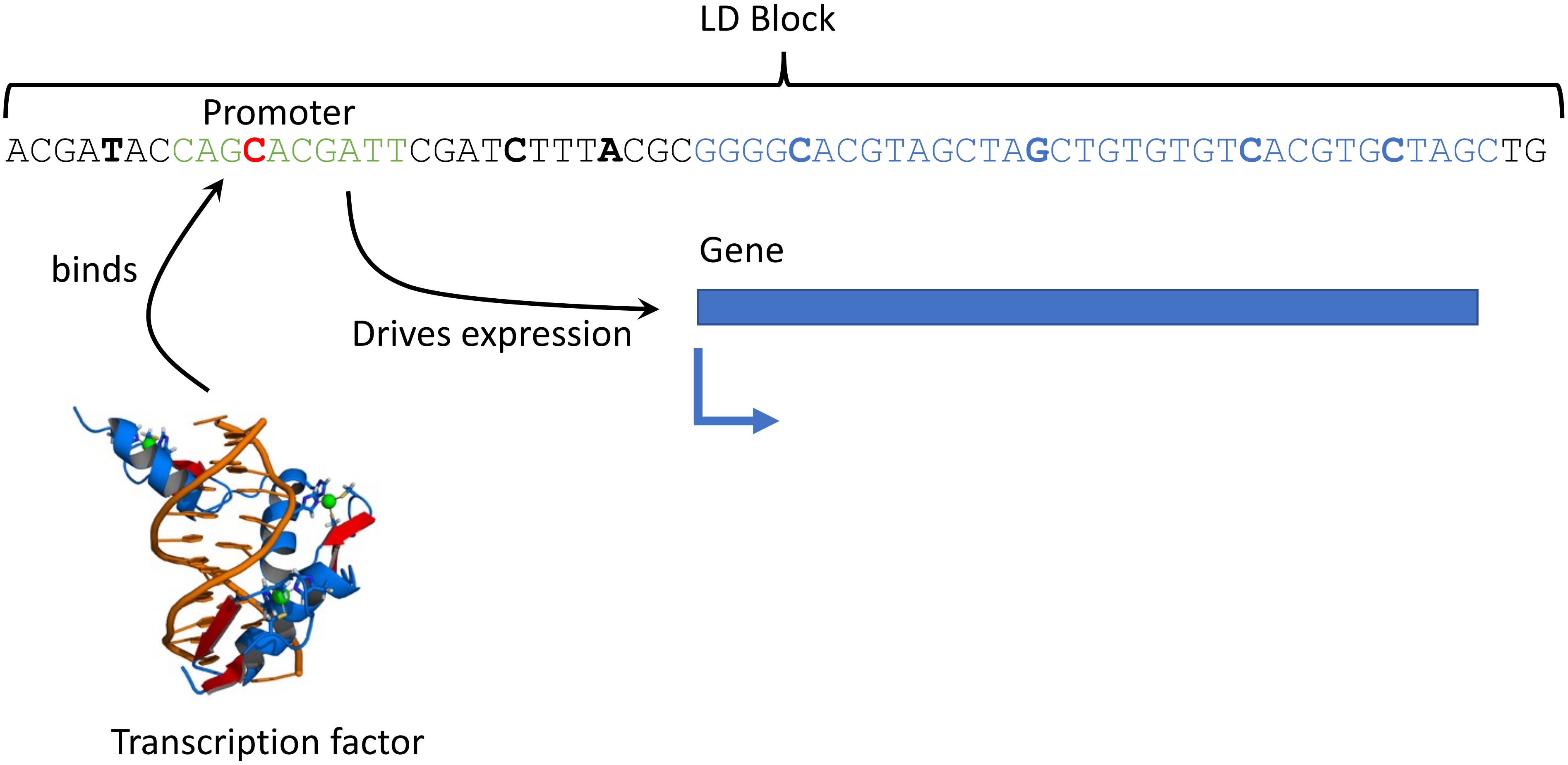
Fine-Mapping

LD Block

ACGA**T**ACCAG**C**ACGATTCGAT**C**TTT**A**CGCGGGG**C**ACGTAGCTA**G**CTGTGTGT**C**ACGTG**C**TAGCTG

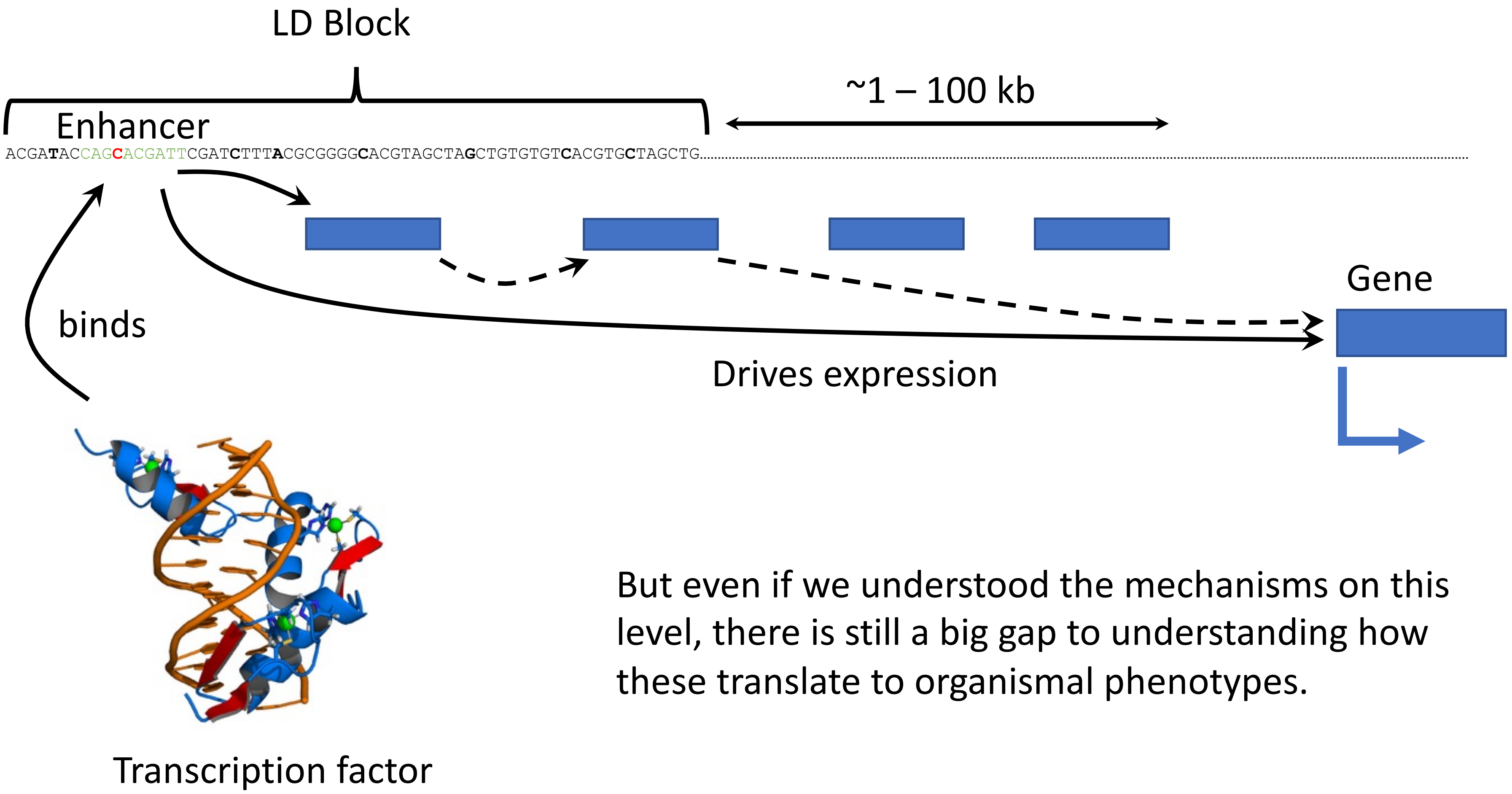


Fine-Mapping



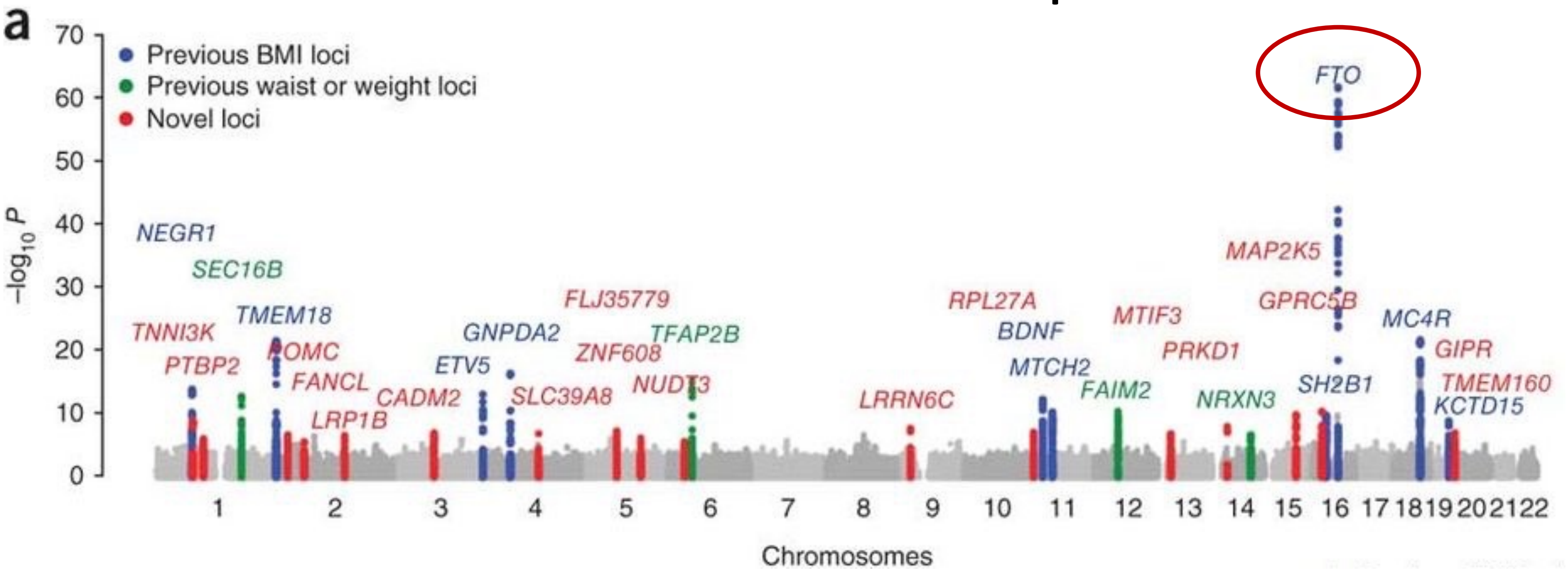
ACGA**T**ACCAG**C**ACGATTCGAT**C**TTT**A**CGCGGGG**C**ACGTAGCTA**G**CTGTGTGT**C**ACGTG**C**TAGCTG

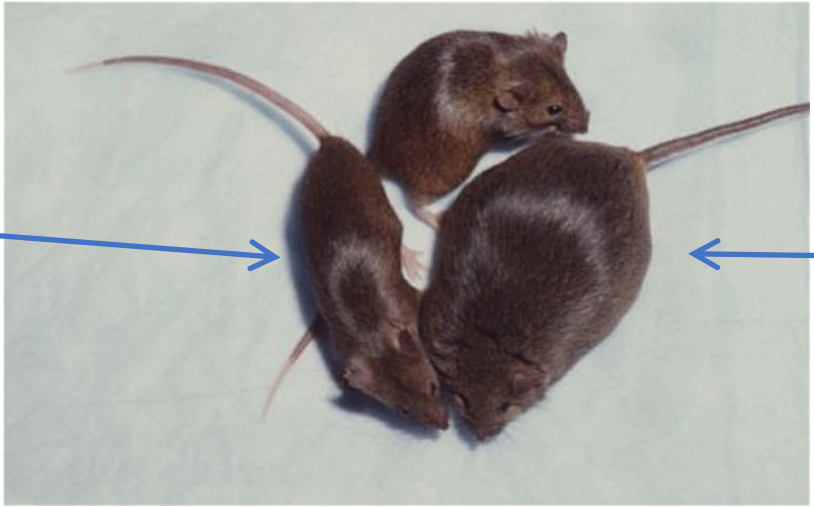
Fine-Mapping



But even if we understood the mechanisms on this level, there is still a big gap to understanding how these translate to organismal phenotypes.

Functional follow-up





Wild-type mouse

Mouse with extra Copies of *fto*

Slide: modified from Iain Mathieson

Speliotes et al. 2010

Fine-Mapping

Once we find an association in a linkage block, how do we identify which specific variant is affecting the trait. “What is the causal variant?”

- Sequence the whole region so that we can find all variants, not just the tag SNPs

- Use functional information – e.g. information about which variants affect gene expression or protein function

- Use prior information about what genes are likely to be associated with a trait (but now we are back to step 1)

Slide: modified from Iain Mathieson

Outline

1. Introduction: what is a GWAS? Why do we do them?

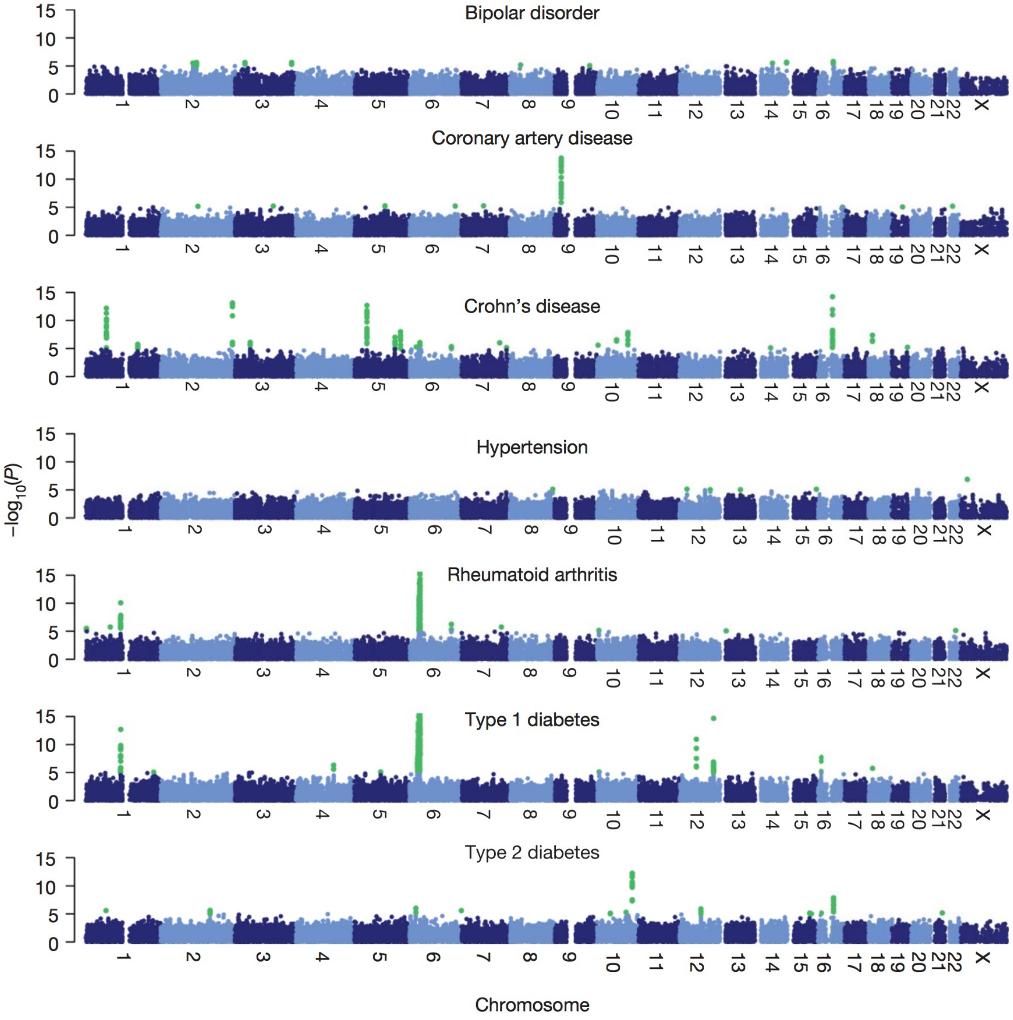
2. Details and practical applications

3. Drug discovery

4. Trends and active research

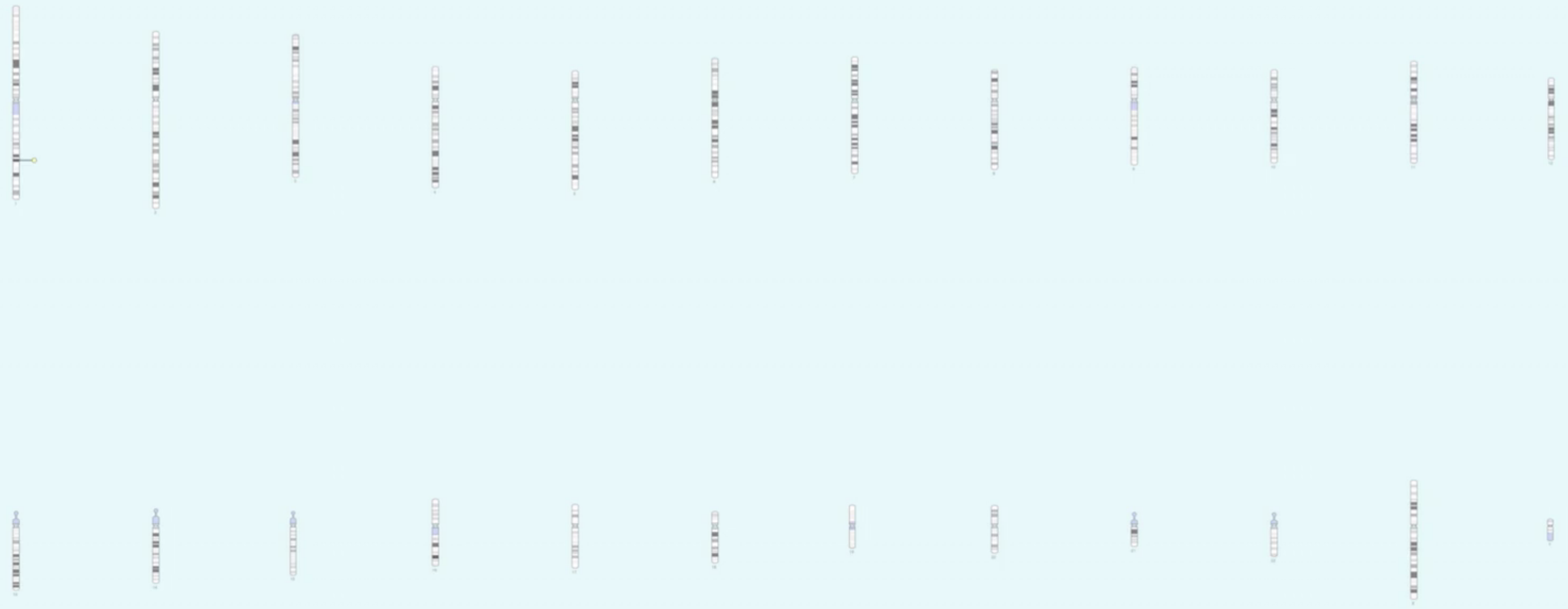
Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*



Wellcome Trust Case Control Consortium, 2007

2006 Jan



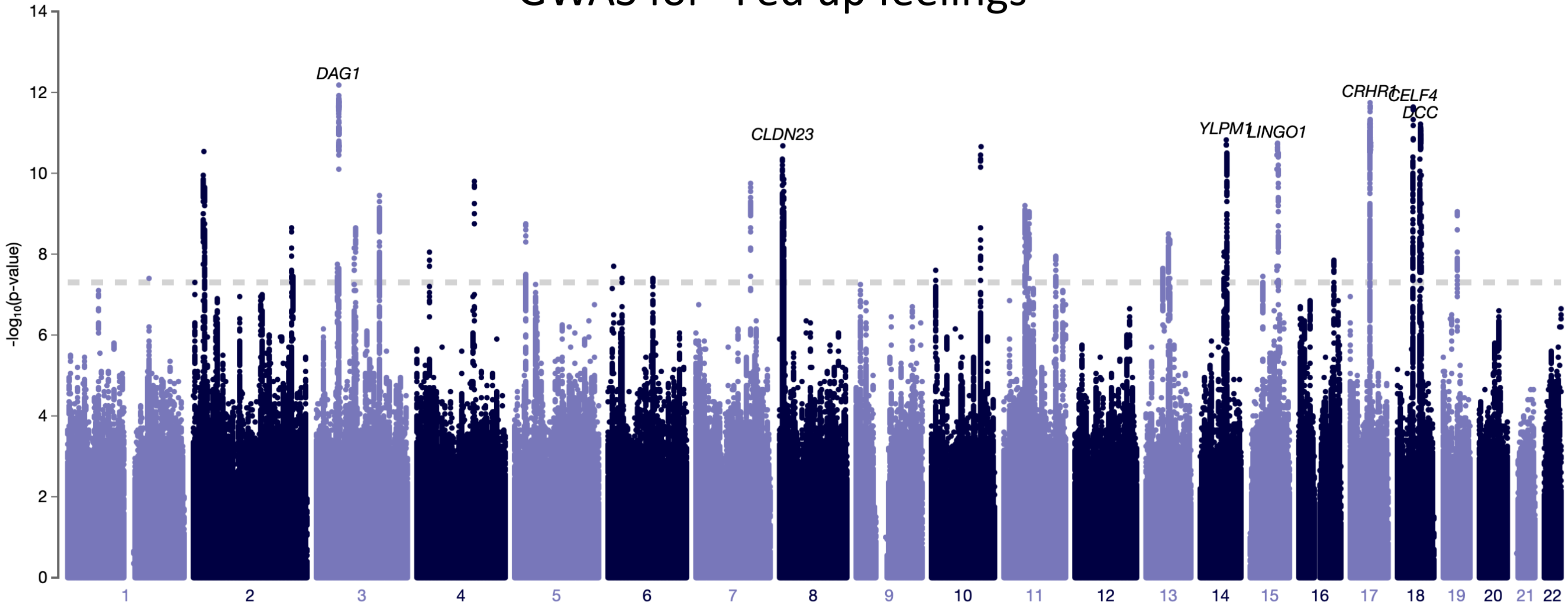
www.ebi.ac.uk/gwas

571,148 Genome-wide significant associations from 6715 publications

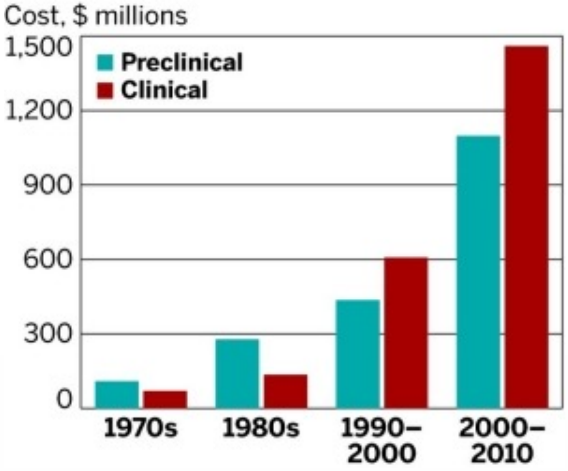
Can GWAS help us to develop new drugs?

GWAS → Find associations → Understand function → Develop drugs

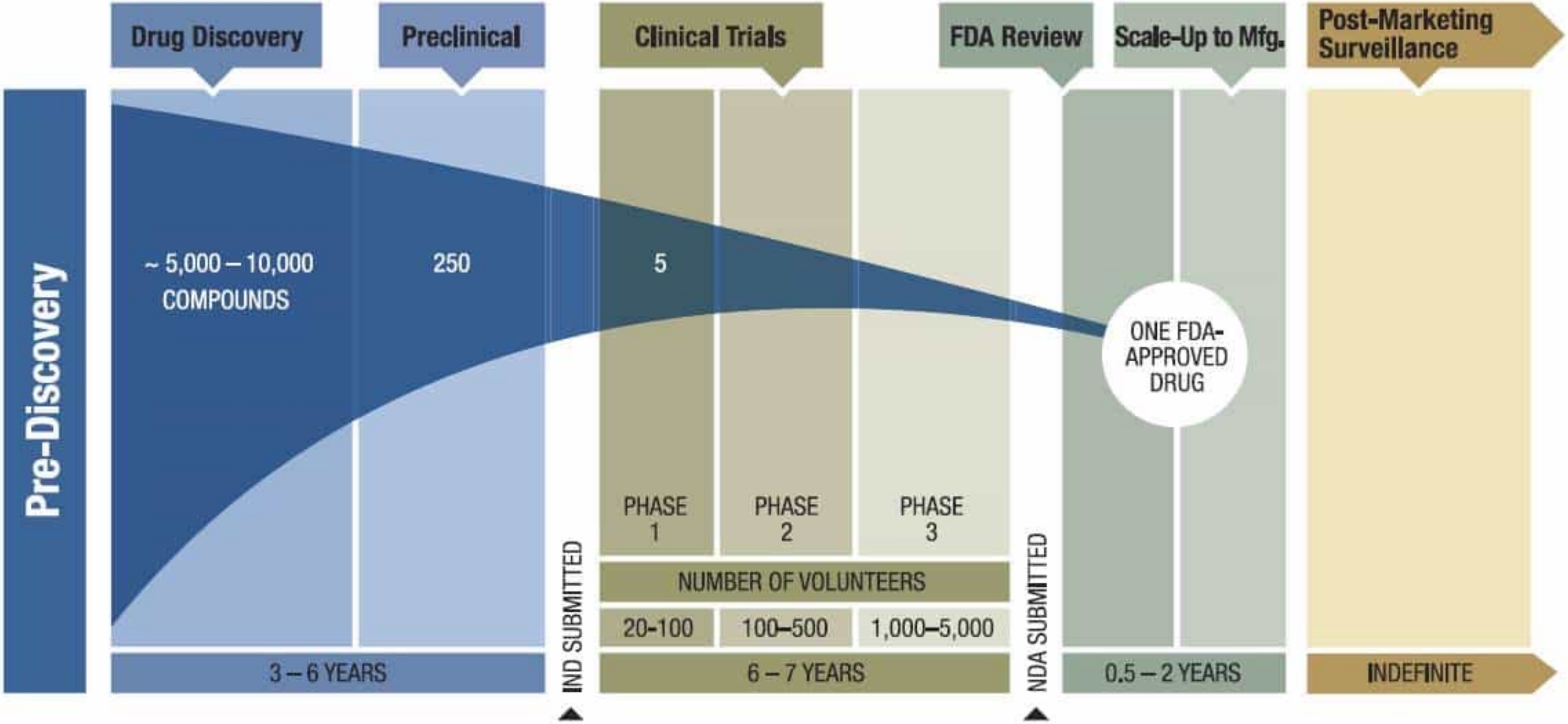
GWAS for “Fed up feelings”



Estimated cost to develop a new drug ~$2-10 Billion



Drug Discovery and Development Timeline



Pre-Discovery

Drug Discovery

Preclinical

Clinical Trials

FDA Review

Scale-Up to Mfg.

Post-Marketing Surveillance

~ 5,000 – 10,000 COMPOUNDS

250

5

ONE FDA-APPROVED DRUG

PHASE 1

PHASE 2

PHASE 3

NUMBER OF VOLUNTEERS

20–100

100–500

1,000–5,000

3 – 6 YEARS

IND SUBMITTED

6 – 7 YEARS

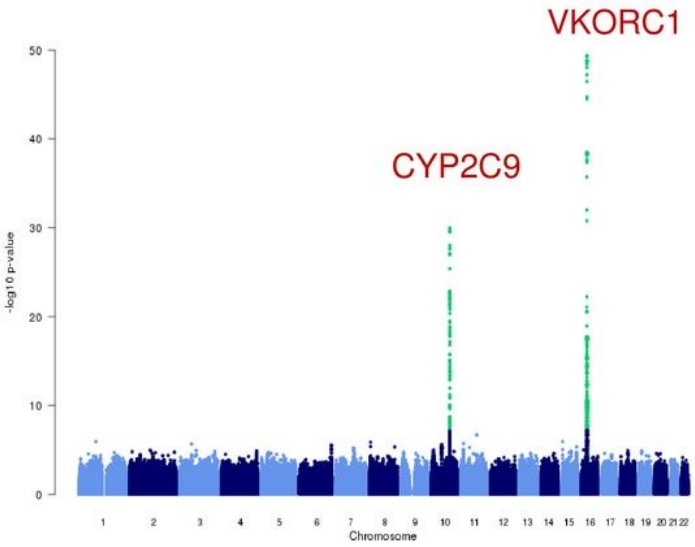
NDA SUBMITTED

0.5 – 2 YEARS

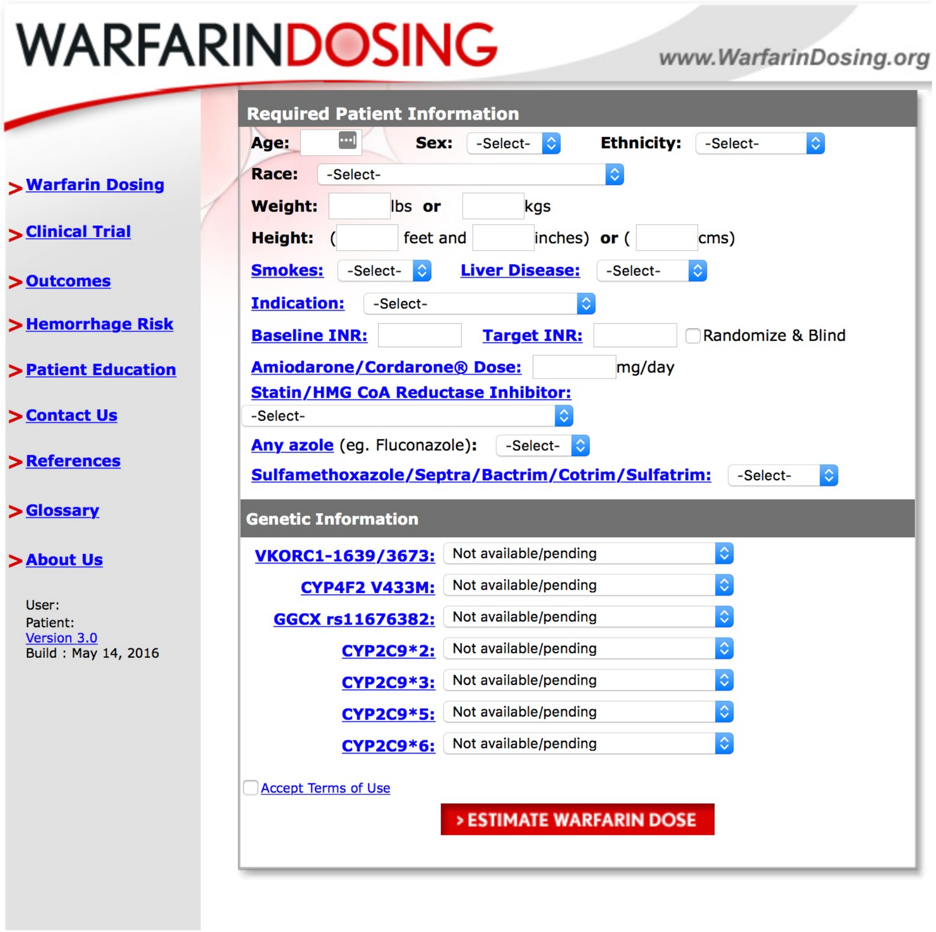
INDEFINITE

Pharmacogenomics

Example: Warfarin is a commonly used anticoagulant, but the appropriate dose is highly patient-specific.



GWAS showed that part of the patient-specific effect is genetic.



User:
Patient:
Version 3.0
Build : May 14, 2016

Required Patient Information

Age: **Sex:** -Select- **Ethnicity:** -Select-

Race: -Select-

Weight: lbs **or** kgs

Height: (feet and inches) **or** (cms)

Smokes: -Select- **Liver Disease:** -Select-

Indication: -Select-

Baseline INR: **Target INR:** □ Randomize & Blind

Amiodarone/Cordarone® Dose: mg/day

Statin/HMG CoA Reductase Inhibitor:
-Select-

Any azole (eg. Fluconazole): -Select-

Sulfamethoxazole/Septra/Bactrim/Cotrim/Sulfatrim: -Select-

Genetic Information

VKORC1-1639/3673:	Not available/pending
CYP4F2 V433M:	Not available/pending
GGCX rs11676382:	Not available/pending
CYP2C9*2:	Not available/pending
CYP2C9*3:	Not available/pending
CYP2C9*5:	Not available/pending
CYP2C9*6:	Not available/pending

□ Accept Terms of Use

Genetic information can be used to guide Rx

- GWAS is most useful as one of many tools in the toolbox for identifying drug targets

- Even fractional increases in the efficiency of discovery can be enormously valuable

- Using GWAS results to effectively manage the drugs we already have might be just as important as developing new drugs.

Outline

1. Introduction: what is a GWAS? Why do we do them?

2. Details and practical applications

3. Drug discovery

4. Trends and active research

Trends and active research

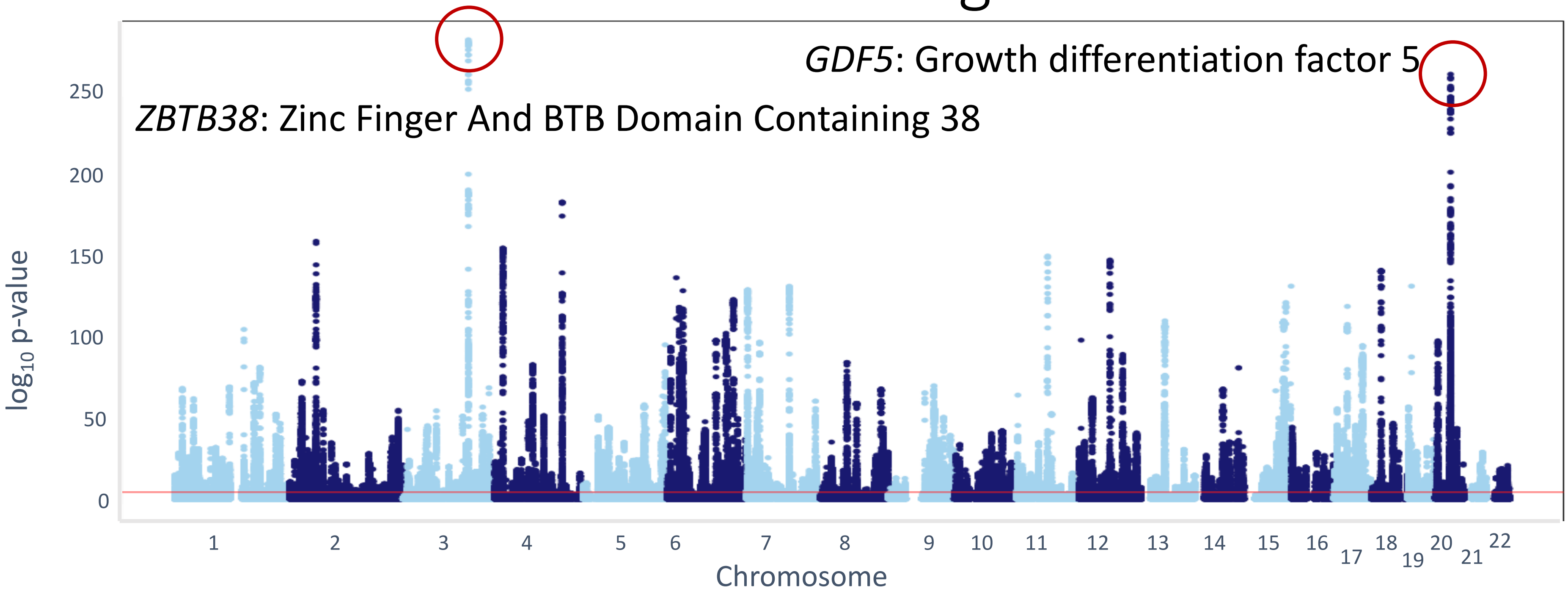
Exome sequencing

Polygenic risk scores

Gene editing

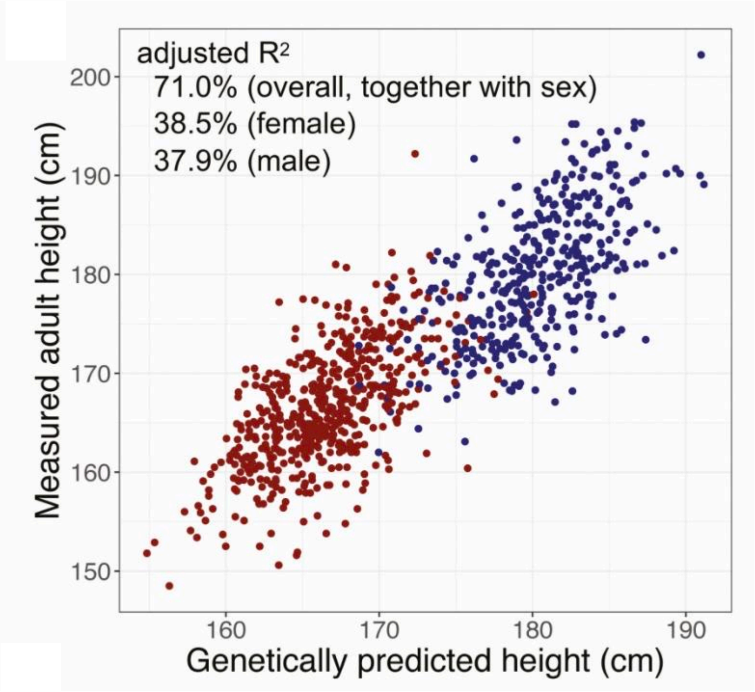
Polygenic risk scores

GWAS for height



Over 3,000 independent loci significantly associated with height in UK Biobank
Together explain about 20-30% of the phenotypic variance

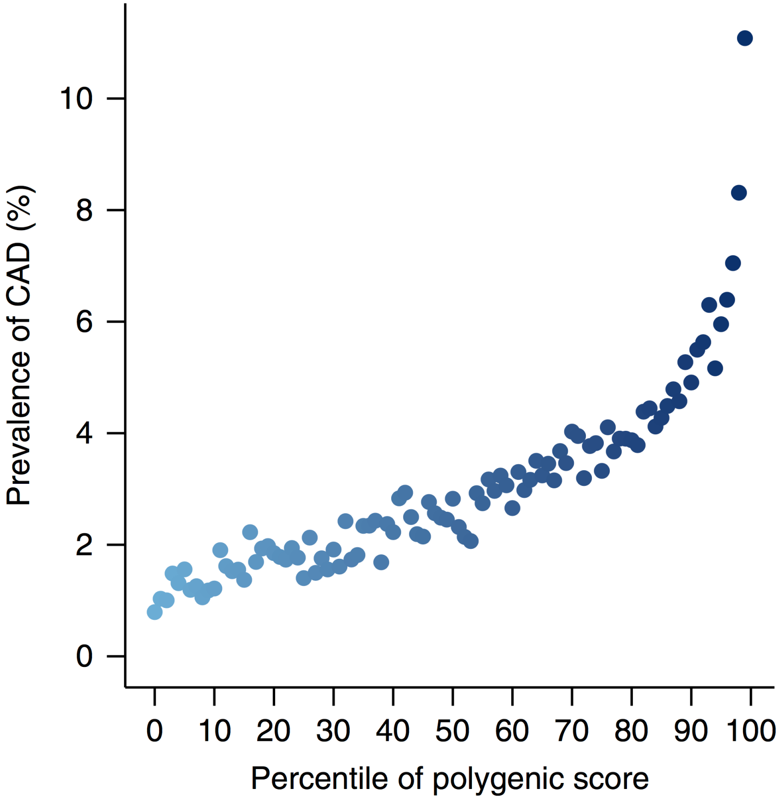
Polygenic risk scores



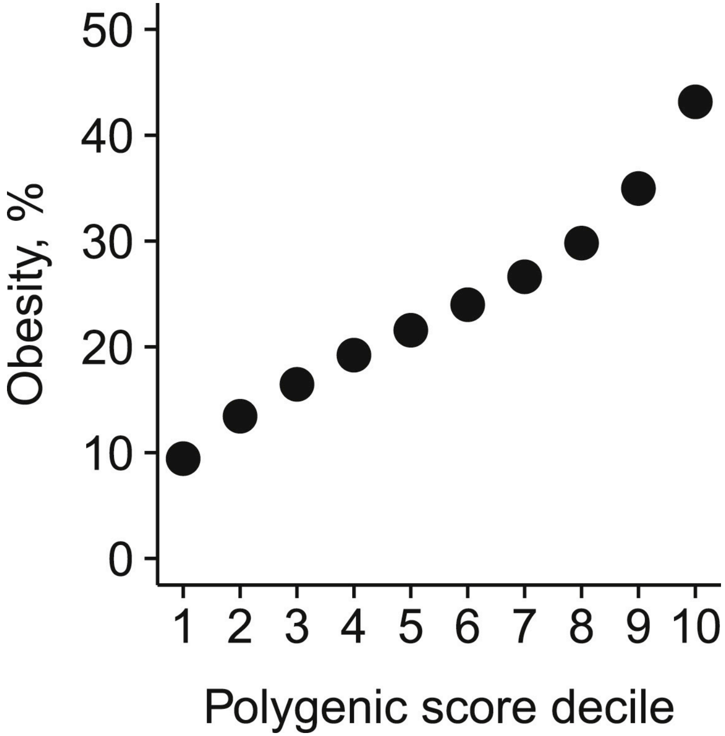
Marouli et al 2017

- Remember, the effect size of any individual GWAS variant is tiny
- So knowing your genotype at any single risk variant doesn’t really help with prediction.
- So let’s just add up the effects over all SNPs!
- This gives us **a polygenic risk score**

Polygenic risk scores can identify people at high risk of disease



Coronary artery disease

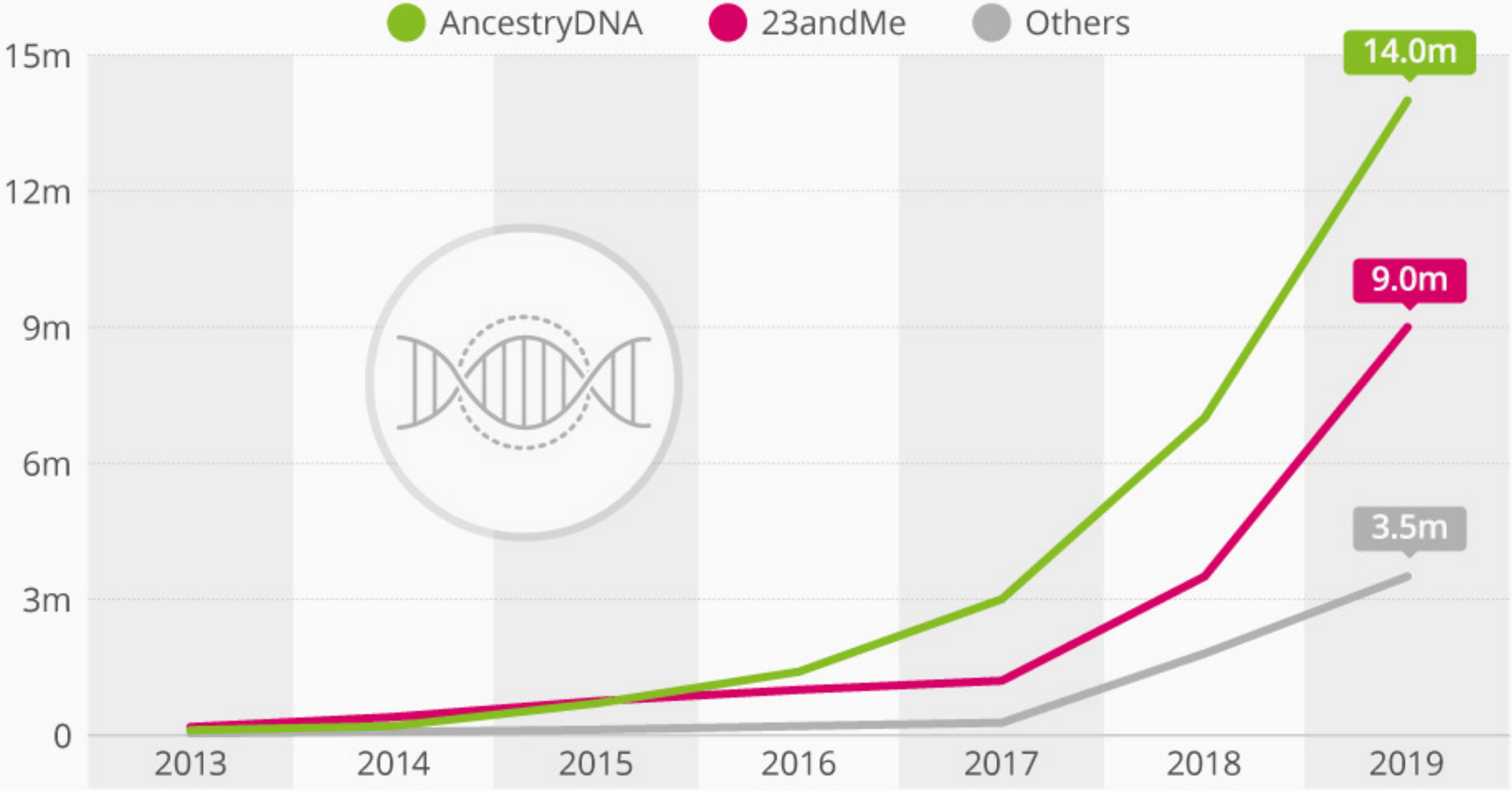


Obesity

Khera et al. 2018, Khera et al 2019

Commercial Genetic Testing Is Gaining Momentum

Estimated total number of people tested by consumer genetic companies*



* Direct-to-consumer genetic testing uses DNA samples, such as saliva, to track a person's ancestry; find family members; disclose a limited array of possible health risks; or brief someone on their personal preferences, like a taste for cilantro or wine.

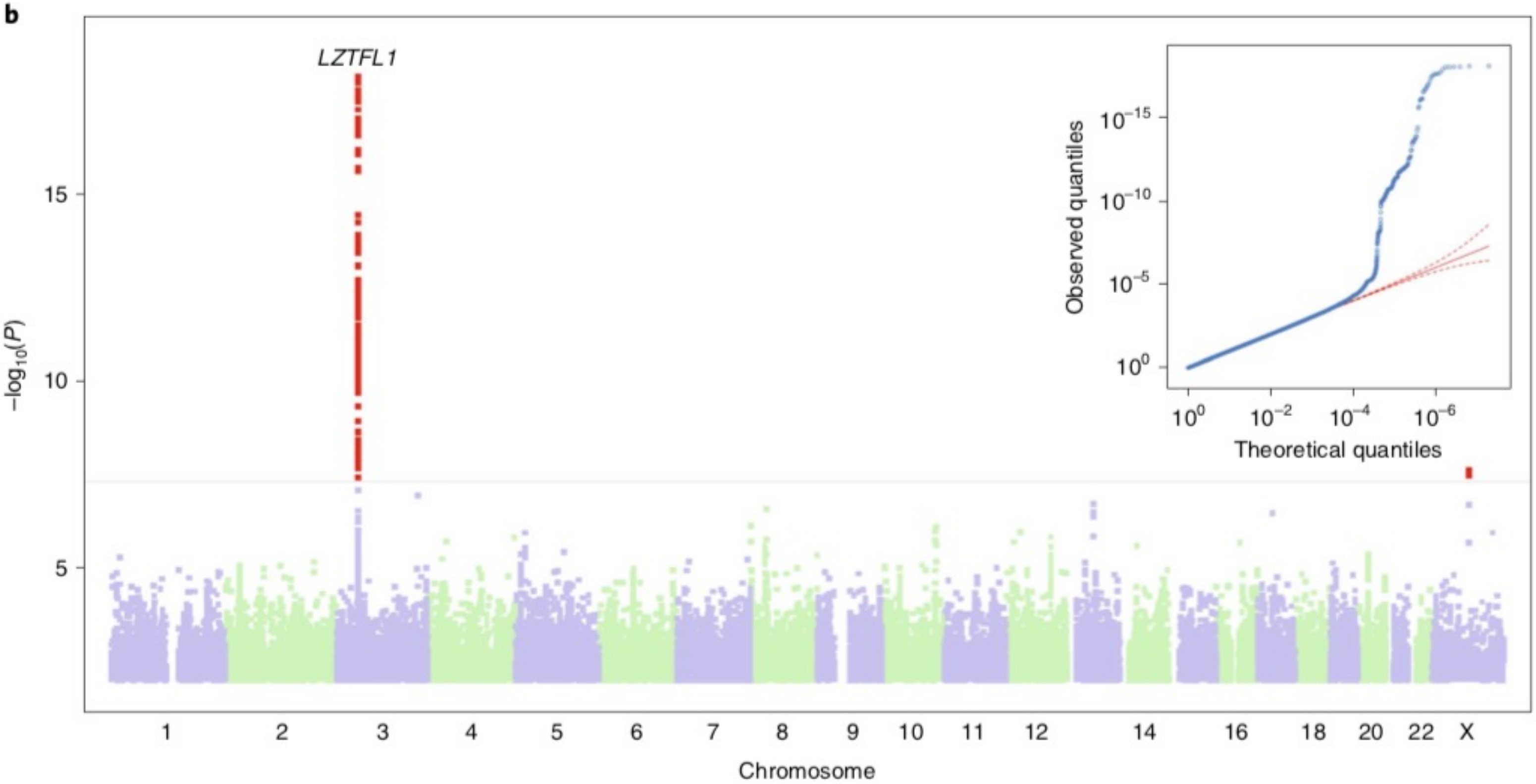
Sources: Company reports, Leah Larkin, ISOGG via MIT Technology Review

statista

Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity

Janie F. Shelton[1,3], Anjali J. Shastri[1,3], Chelsea Ye[1], Catherine H. Weldon[1], Teresa Filshtein-Sonmez[1], Daniella Coker[1], Antony Symons[1], Jorge Esparza-Gordillo[2], The 23andMe COVID-19 Team*, Stella Aslibekyan[1] and Adam Auton[1] ✉

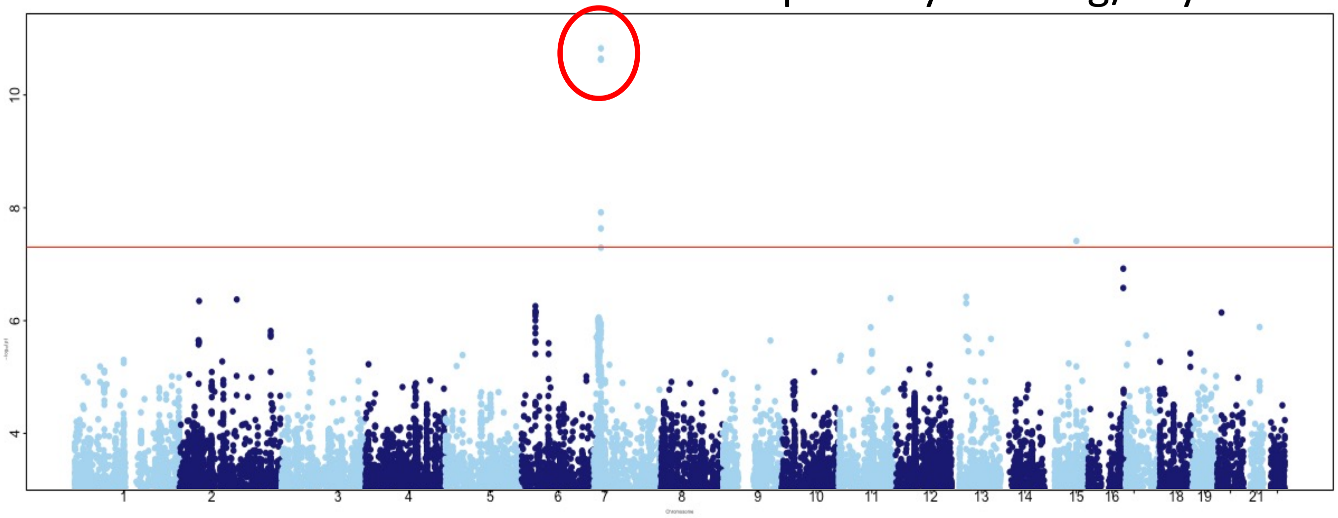
b



PLOS Genetics 2022

Personal genomics

AHR: rs4410790 C/T
Each C allele increases caffeine consumption by 0.15 mg/day



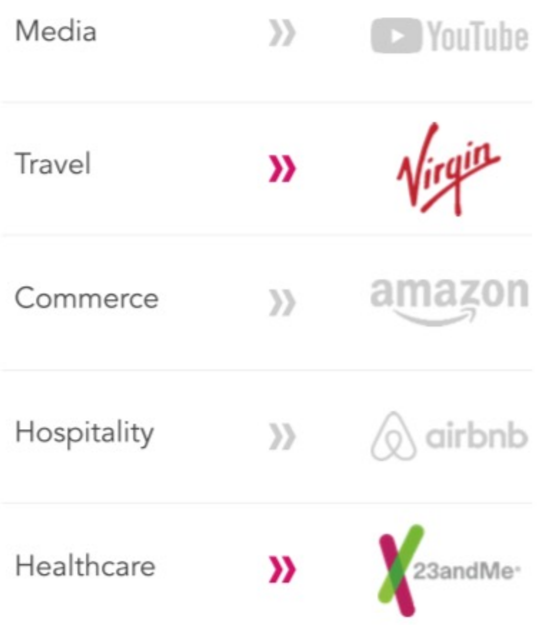
UK Biobank phenotype 100240: ”Did you drink any coffee yesterday?”

Prognosis

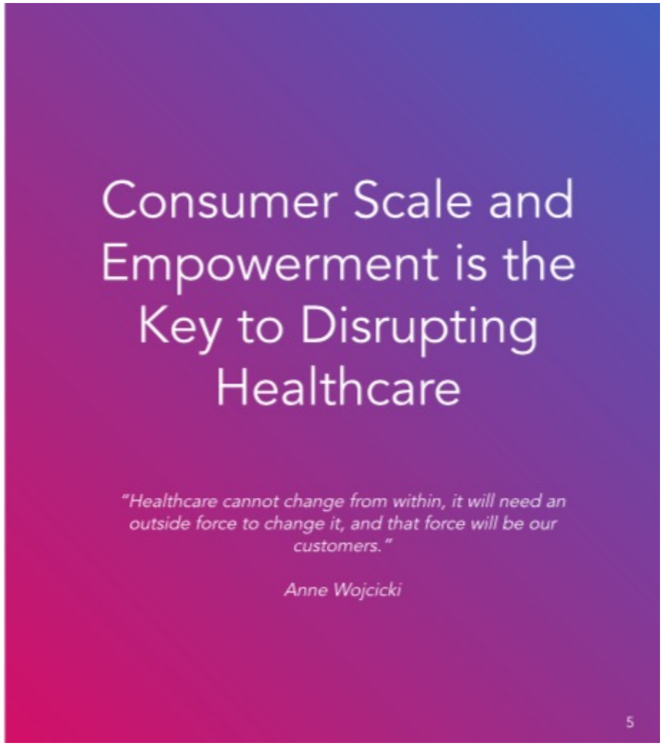
23andMe Goes Public as $3.5 Billion Company With Branson Aid

By Kristen V Brown

February 4, 2021, 7:28 AM EST *Updated on February 4, 2021, 11:39 AM EST*



Media	»	YouTube
Travel	»	Virgin
Commerce	»	amazon
Hospitality	»	airbnb
Healthcare	»	23andMe

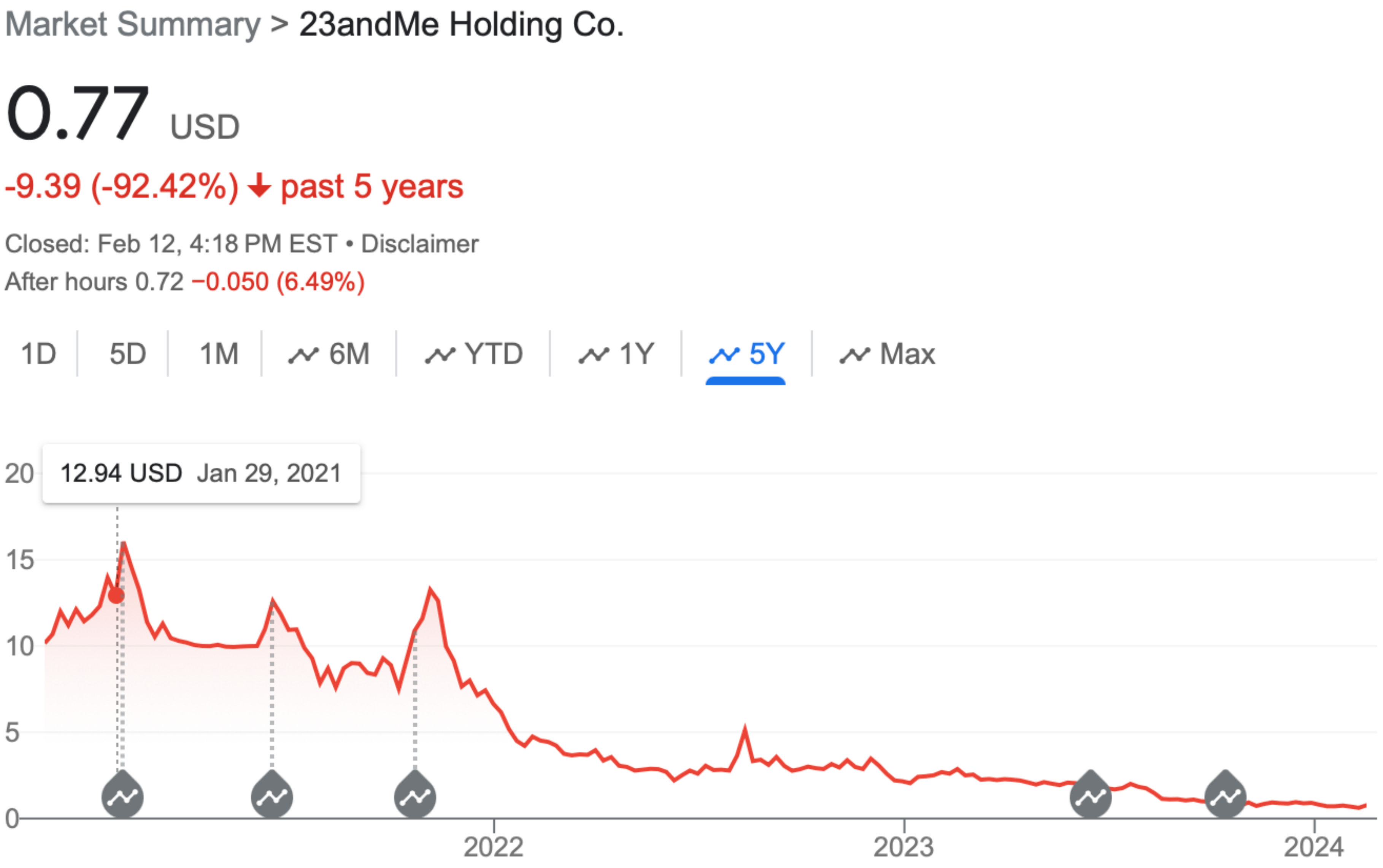


Consumer Scale and Empowerment is the Key to Disrupting Healthcare

"Healthcare cannot change from within, it will need an outside force to change it, and that force will be our customers."

Anne Wojcicki

23andMe investor slide deck: https://mediacenter.23andme.com/company-2/investors/



Market Summary > 23andMe Holding Co.

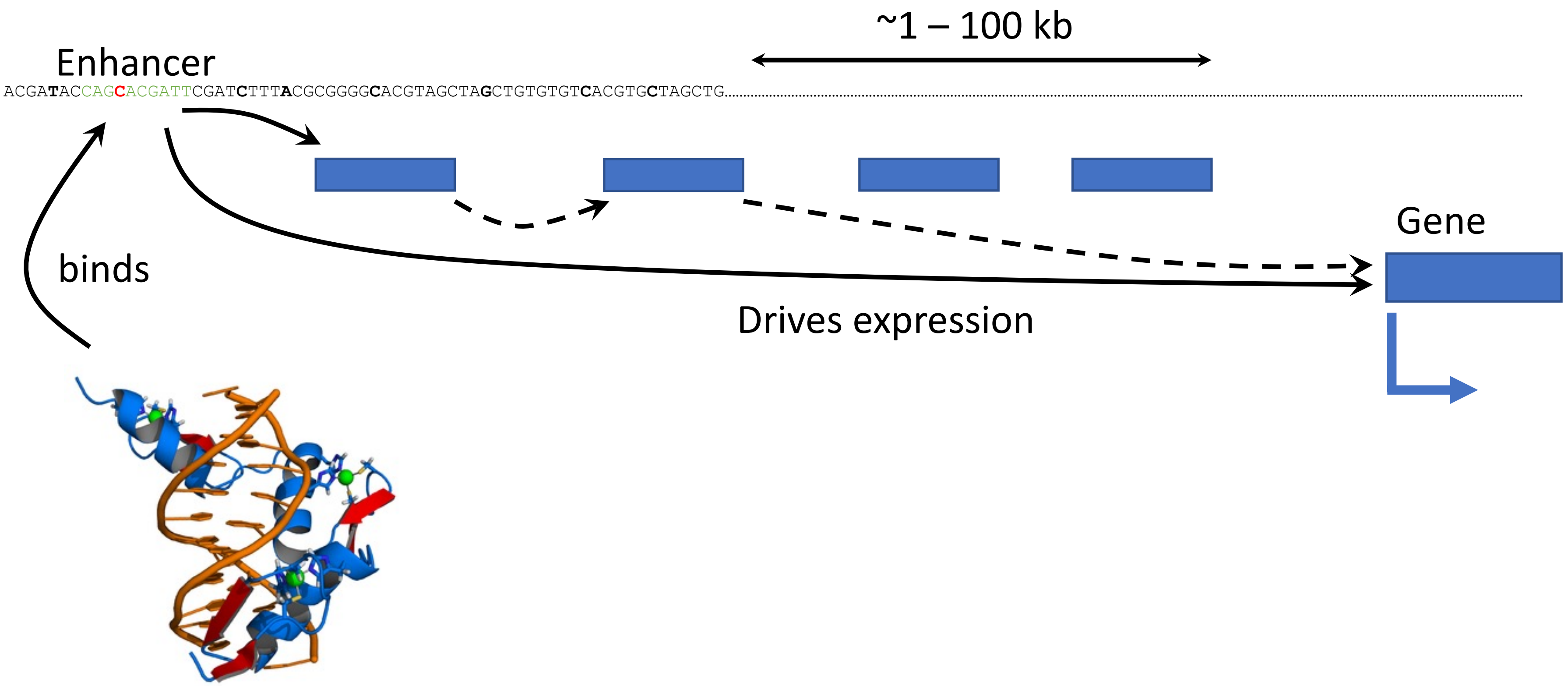
0.77 USD

-9.39 (-92.42%) ↓ past 5 years

Closed: Feb 12, 4:18 PM EST • Disclaimer

After hours 0.72 −0.050 (6.49%)

Exome sequencing



~1 – 100 kb

Enhancer

ACGA**T**ACCAG**C**ACGATTCGAT**C**TTT**A**CGCGGGG**C**ACGTAGCTA**G**CTGTGTGT**C**ACGTG**C**TAGCTG..

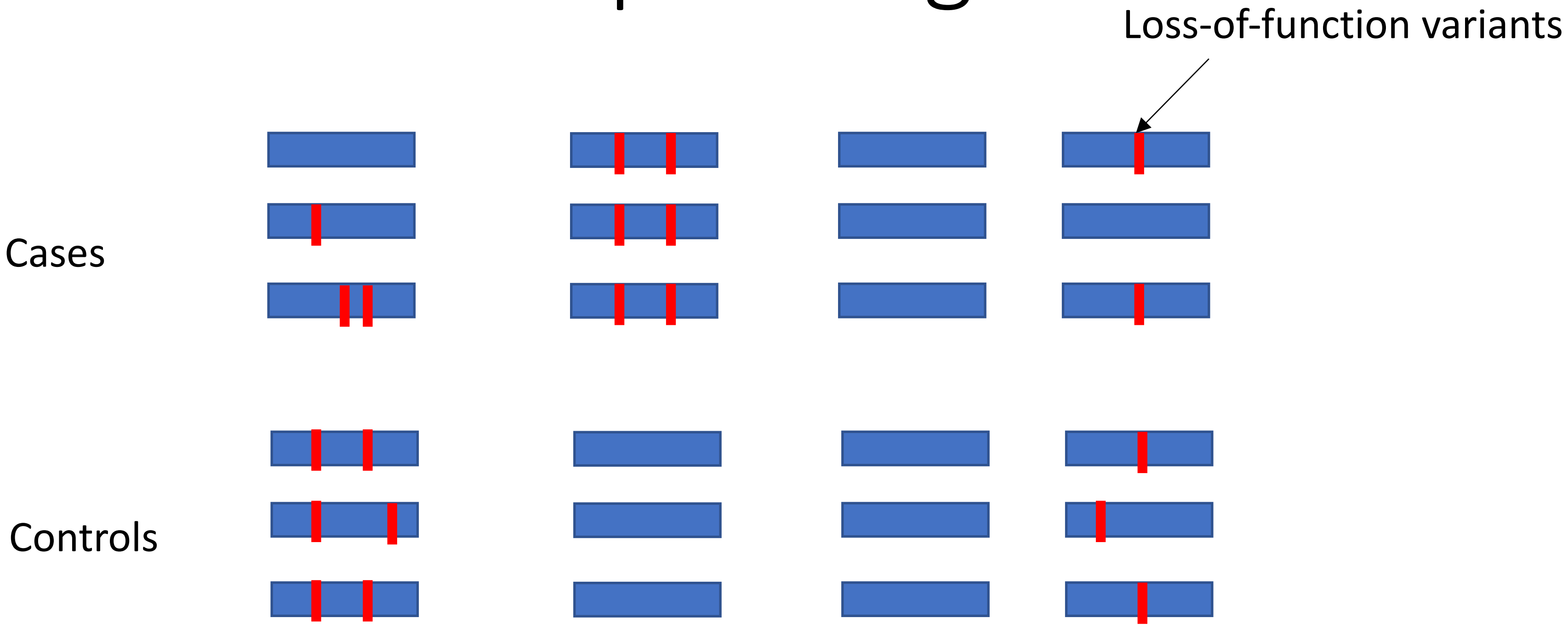
Gene

binds

Drives expression

Transcription factor

Exome sequencing



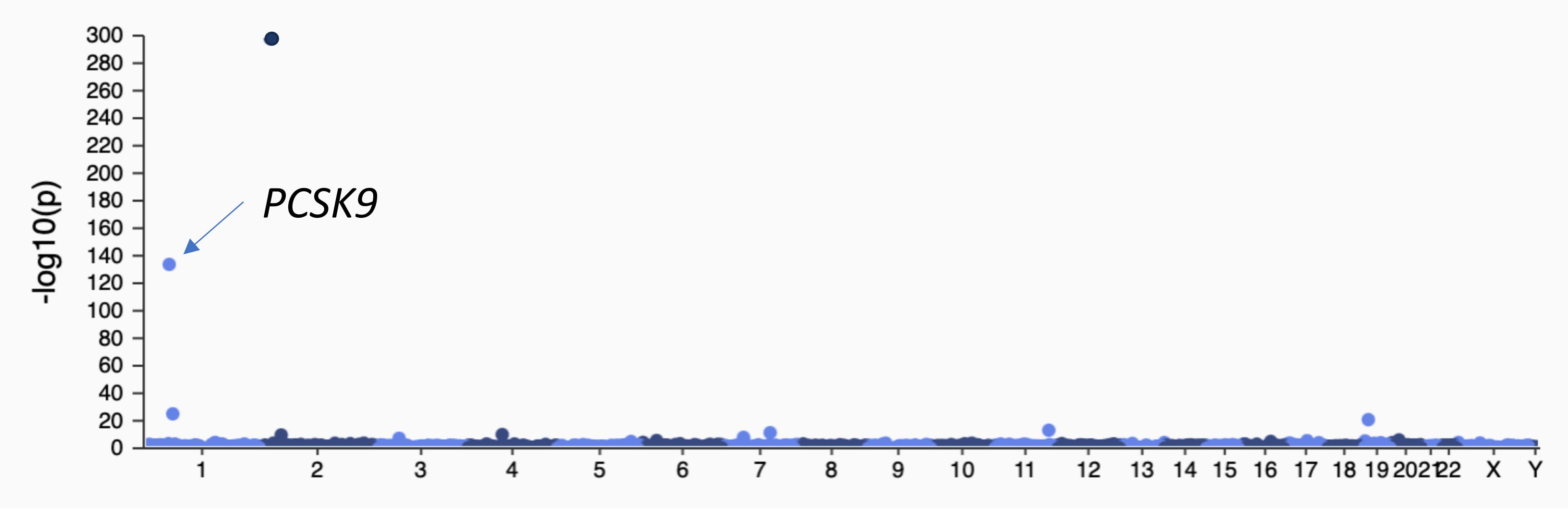
Downside:

- We know that these variants are rare (so need even larger samples)
- We know that we miss most of the genetic effects

Upside:

- We know what the gene is
- We know what the variants are doing
- Cheaper than whole-genome sequencing (more expensive than arrays)

Exome-wide association study for LDL cholesterol

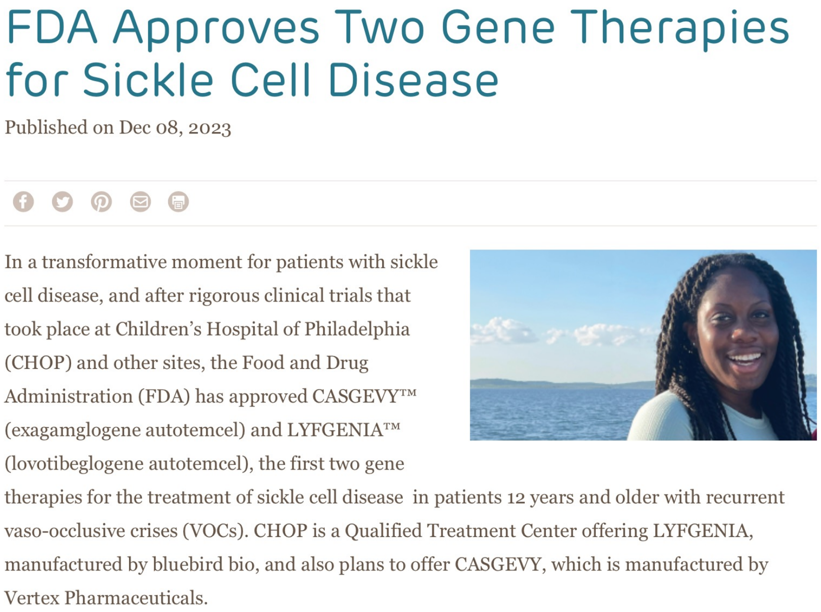


- Loss of function mutations in *PCSK9* cause low LDL cholesterol
- Low LDL cholesterol protects against heart disease
- Leading to drugs (PCSK9 inhibitors) and gene therapy (Vertex pharmaceuticals)

Gene therapy for rare diseases – what about common diseases?

Gene Therapy Allows an 11-Year-Old Boy to Hear for the First Time

The genetic treatment targeted a particular kind of congenital deafness and will soon be tried in children who are younger.



FDA Approves Two Gene Therapies for Sickle Cell Disease

Published on Dec 08, 2023

In a transformative moment for patients with sickle cell disease, and after rigorous clinical trials that took place at Children's Hospital of Philadelphia (CHOP) and other sites, the Food and Drug Administration (FDA) has approved CASGEVY™ (exagamglogene autotemcel) and LYFGENIA™ (lovotibeglogene autotemcel), the first two gene therapies for the treatment of sickle cell disease in patients 12 years and older with recurrent vaso-occlusive crises (VOCs). CHOP is a Qualified Treatment Center offering LYFGENIA, manufactured by bluebird bio, and also plans to offer CASGEVY, which is manufactured by Vertex Pharmaceuticals.

FDA NEWS RELEASE

FDA approves innovative gene therapy to treat pediatric patients with spinal muscular atrophy, a rare disease and leading genetic cause of infant mortality