Genome variation and function 2

Marie Saitou
Assistant Professor, CIGENE, BIOVIT
17. Oct. 2022

Learning Outcomes
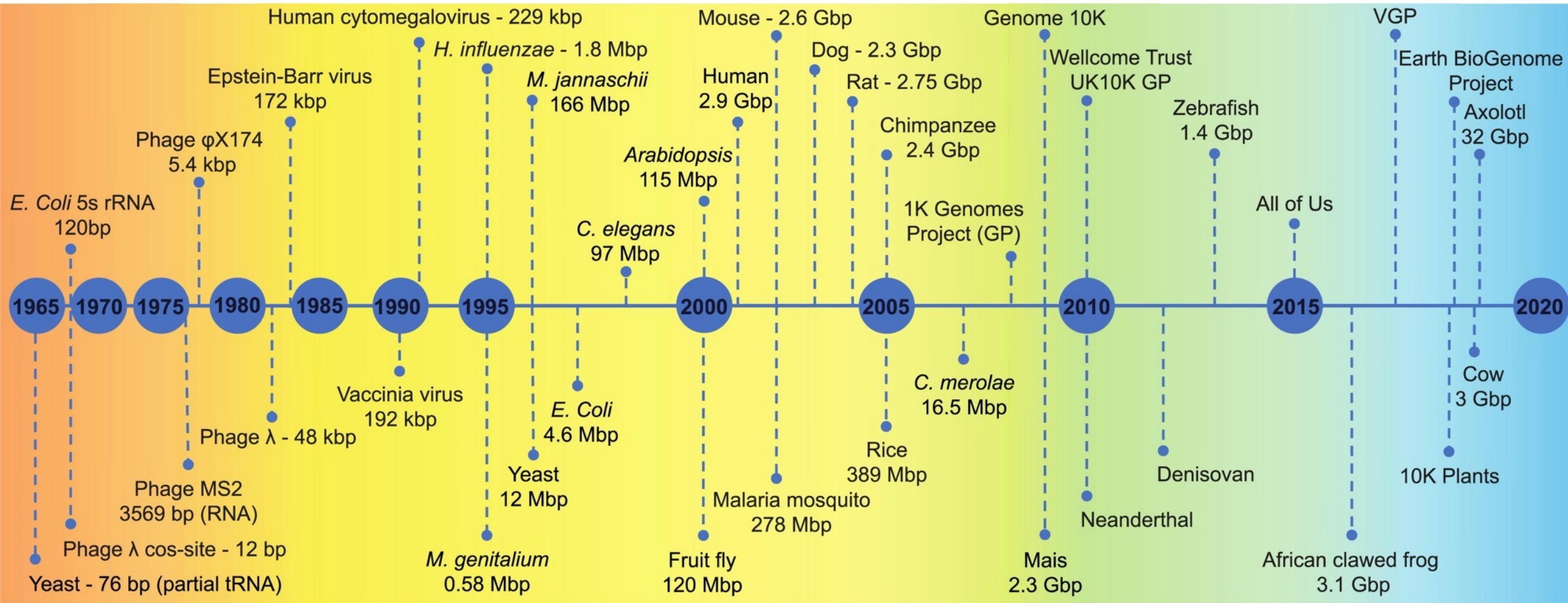
Prologue - Current topics in genomics

- Human genome sequence

- Nobel prize 2022

We will learn about:

- Methods and significance of functional genomics

- Challenges of human genomics

- How genetics is related to societal issues

Brief Review: The timeline of genomics

The complete human genome (2022)

https://www.biorxiv.org/content/10.1101/2021.05.26.445798v1.full

- Adds nearly 200 million base pairs to the 2013 version of the human genome sequence.

- Lasers to scan long stretches of DNA isolated from cells — up to 20,000 base pairs at a time. (PacBio)

- Insight into traits/disease/evolution with further studies

4

Adds nearly 200 million base pairs to the 2013 version of the human genome sequence.

chr21
0 10 20 30 40

chr22
0 10 20 30 40 50

Segmental duplications
- 0 per 10 kbp
- 500 per 10 kbp
- 1000 per 10 kbp

- Centromeric satellites
- Known genes
- Novel genes
- GRCh38 gaps/issues

The latest technology

First Generation

Second Generation
(Next Generation Sequencing)

Third Generation

Sanger Sequencing
Maxam and Gilbert
Sanger Chain-termination

454, Solexa,
Ion Torrent
Illumina

PacBio
Oxford Nanopore

- Infer nucleotide identity using dNTPs then visualize with electrophoresis
- 500-1000 bp fragments

- High throughput from the parallelization of sequencing reactions
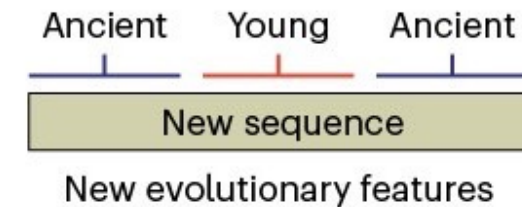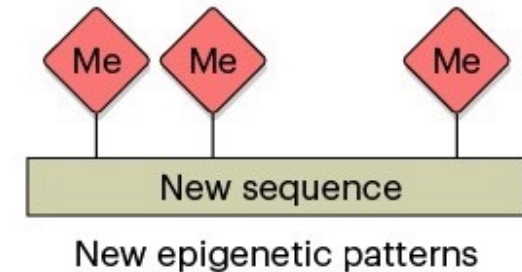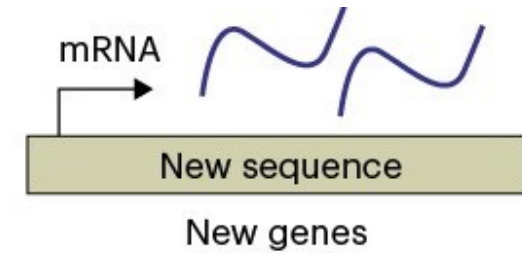- ~50-500 bp fragments

- Sequence native DNA in real time with single-molecule resolution
- Tens of kb fragments, on average

Short-read sequencing

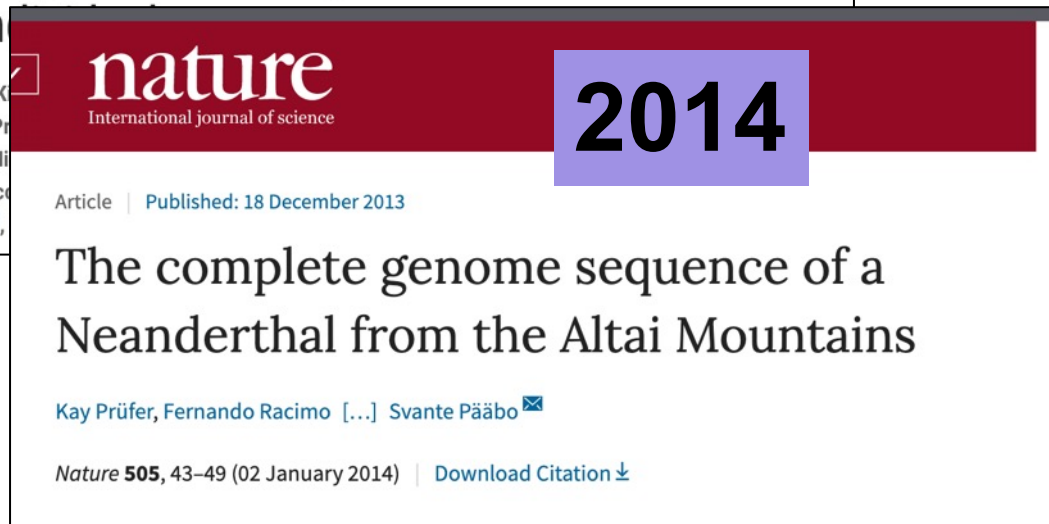**Long-read sequencing**

Evolution/Function/Disease...

- New genes, epigenetic patterns and

  evolutionary features to be identified

  in previously unseen regions

Ancient human genomics

Cell
Volume 90, Issue 1, 11 July 1997, Pages 19-30

CellPress

1997

Article

Neandertal DNA Sequences and the Origin

Science   Contents ▾   ...ers ▾   Journals ▾

2012

..., Svante Pääbo

SHARE   RESEARCH ARTICLE

A High-Coverage Genome Sequence from an Archaic
Denisovan In...

Matthias Meyer[1,*,‡], Martin K...
Schraiber[4], Flora Jay[4], Kay Pr...
Rohland[2,3], Arti Tandon[2,3], Mi...
Dabney[1], Jay Shendure[6], Jacc...
Patterson[2], Aida M. Andrés[1], ...

nature
International journal of science

2014

Article | Published: 18 December 2013

The complete genome sequence of a
Neanderthal from the Altai Mountains

Kay Prüfer, Fernando Racimo [...] Svante Pääbo ✉

Nature **505**, 43–49 (02 January 2014) | Download Citation ⬇

EXPRESS **explained.**
SCI-TECH

Nobel Prize in Medicine, 2022

Archaic Hominins
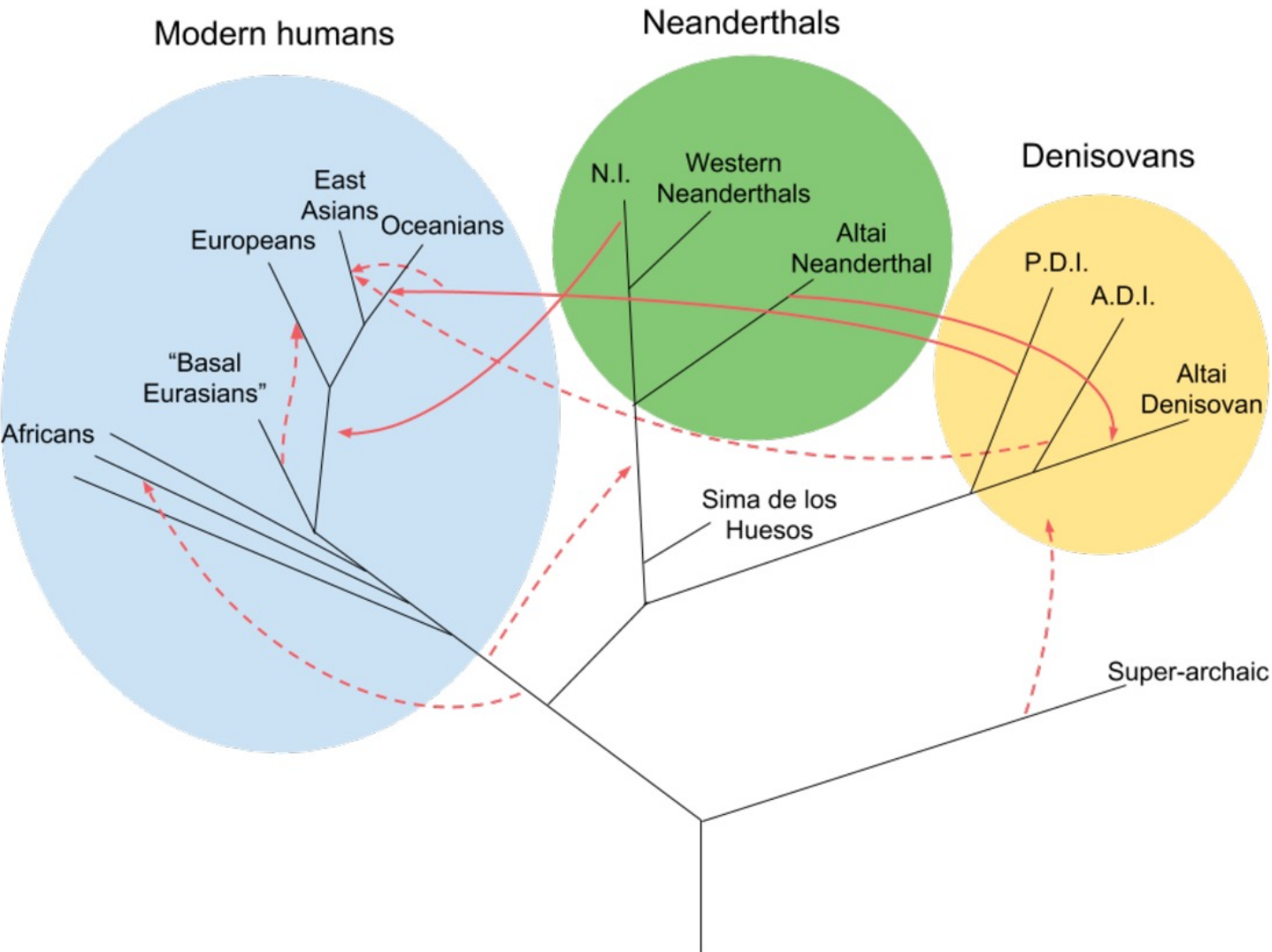
The genome of the offspring of a Neanderthal mother and a Denisovan father

Viviane Slon ✉, Fabrizio Mafessoni, Benjamin Vernot, Cesare de Filippo, Steffi Grote, Bence Viola, Mateja Hajdinjak, Stéphane Peyrégne, Sarah Nagel, Samantha Brown, Katerina Douka, Tom High Maxim B. Kozlikin, Michael V. Shunkov, Anatoly P. Derevianko, Janet Kelso, Matthias Meyer, Kay Prüfer & Svante Pääbo ✉

10 Years of genome-wide association studies

Visscher et al., 2017
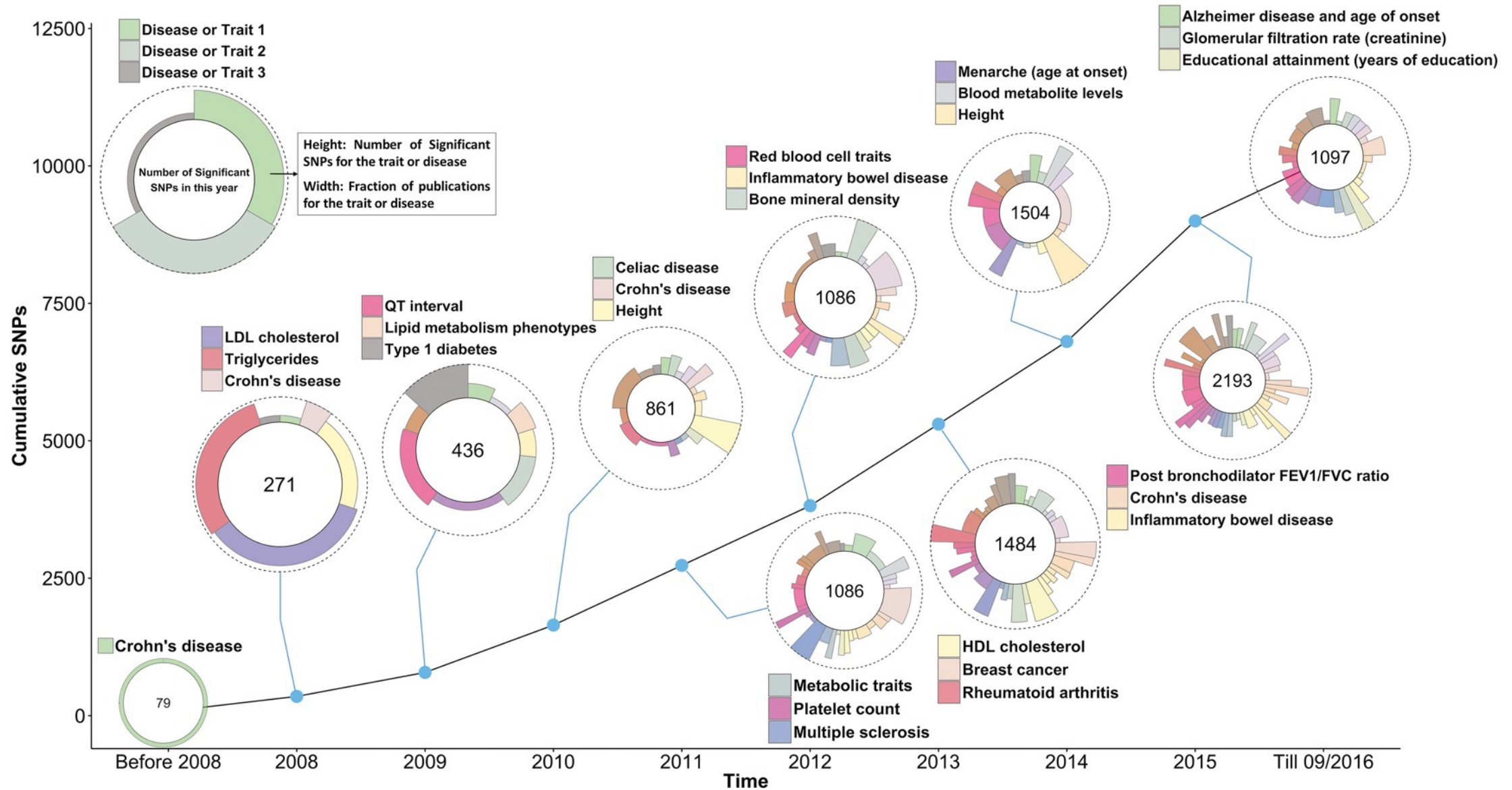
Genome-wide association study (GWAS)
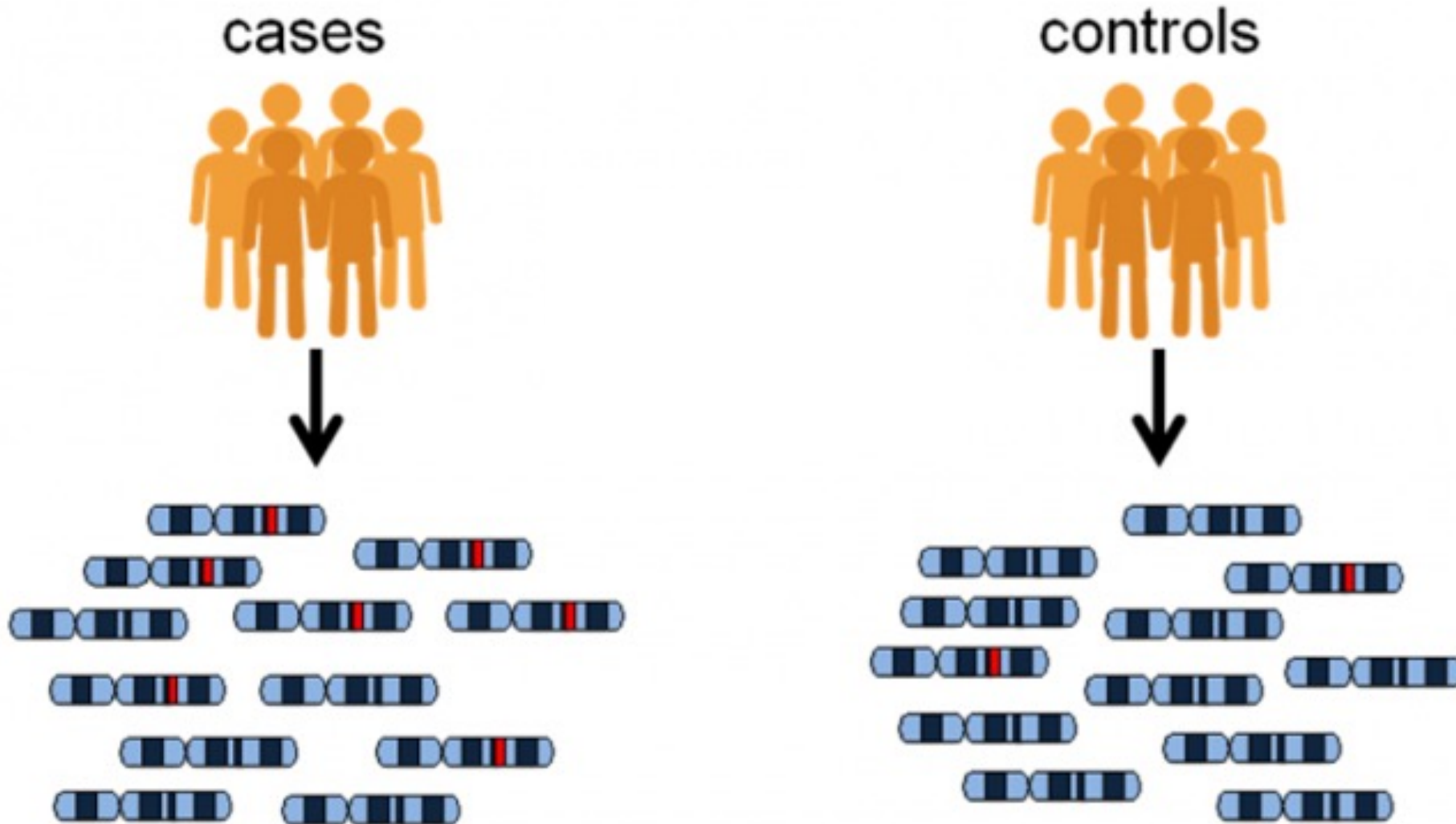
GWAS –
estimate the association between
the phenotype and all observed variants

$$\hat{\beta}_i = cor\,(Genotype_{SNP_i},\ Phenotype)$$

AA   AT   TT
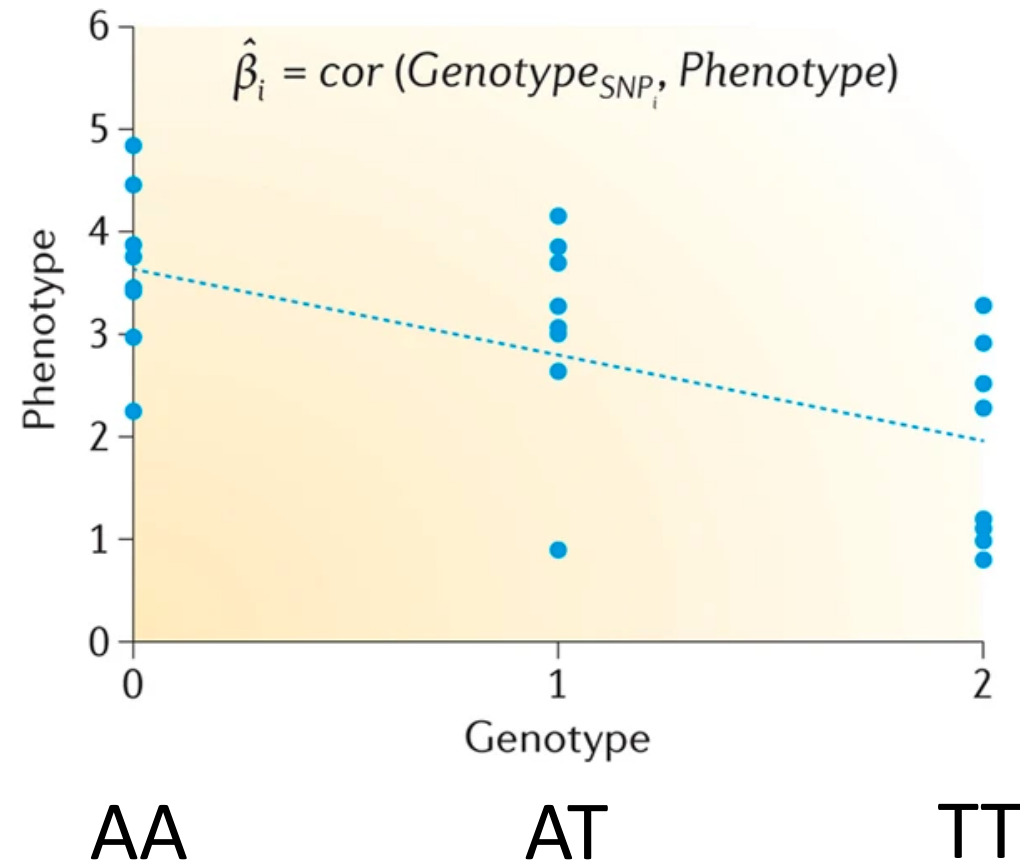
How to read a Manhattan plot – a GWAS result representation

variants

chromosomes
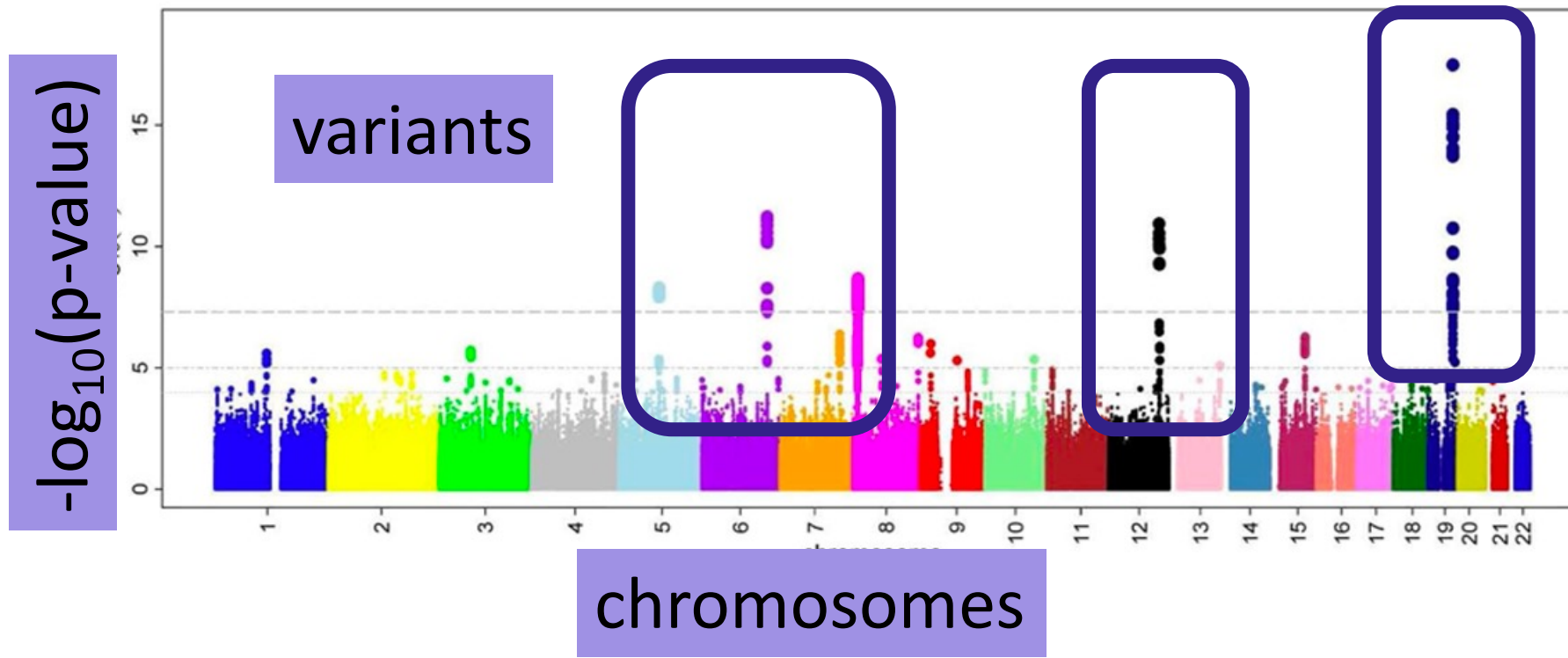
$-\log_{10}(\text{p-value})$

What is the p-value in GWAS?

*the probability of obtaining results as extreme
as the observed results of a statistical hypothesis test,
assuming that the <u>null hypothesis </u>is correct.*

- <u>alternative hypothesis</u>: the variant is associated with the trait
- <u>null hypothesis</u>: the variant is not associated with the trait

Multiple comparisons problem

However, When you conduct the for 100 variants, 5 of the genes can show $p < 0.05$ under the null hypothesis.

-> When you conduct the for 10,000 variants, 500 of the genes can show $p < 0.05$ under the null hypothesis (randomly, by chance).

There are various ways to "correct" this problem, considering how many times one conducts the same comparison.

Example:

Bonferroni correction - the local significance level is divided by the number of tests performed.

Why genome-wide analysis is beneficial

- Most quantitative traits are influenced by many genetic variants each with small effects

- **GWAS** can identify these individual genetic loci with small effects in a large enough sample

- **Polygenic score/breeding value** predicts phenotype based on genomic information aggregating the small effect of genetic variants

Desirable traits: mostly polygenic traits
Polygenic traits vs Mendelian traits

Mendelian traits: shaped by a single gene alone.

example: Wet or dry earwax, Face freckles

Polygenic traits: influenced by multiple (~ thousands) genes.

example: Hair color, height and most of quantitative (continuous) traits

Polygenic trait: example (three loci, A, B and C affect color)
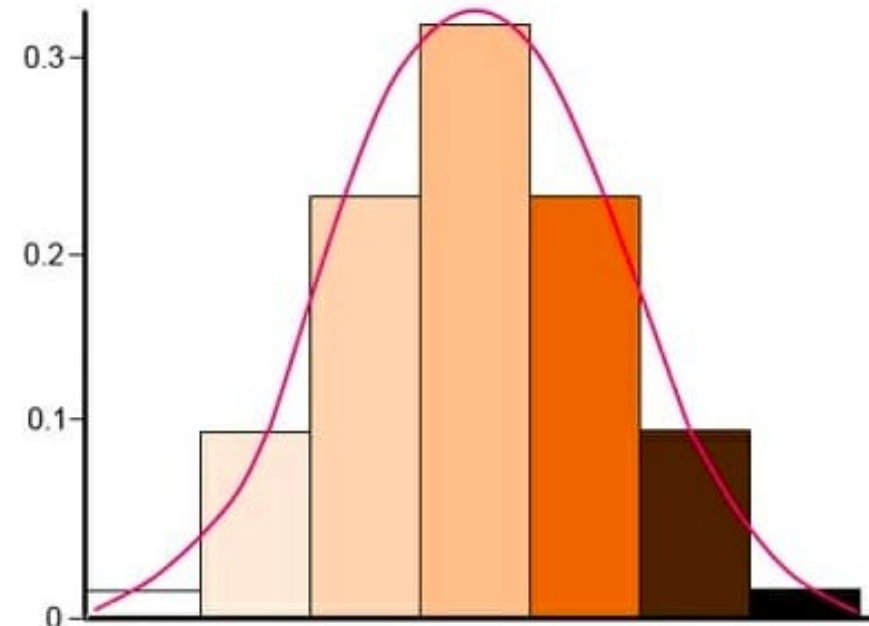
aabbcc (light color) x AABBCC (dark color)

AaBbCc (intermediate)

genotype and color

	ABC	ABc	AbC	aBC	Abc	aBc	abC	abc
ABC	6	5	5	5	4	4	4	3
ABc	5	4	4	4	3	3	3	2
AbC	5	4	4	4	3	3	3	2
aBC	5	4	4	4	3	3	3	2
Abc	4	3	3	3	2	2	2	1
aBc	4	3	3	3	2	2	2	1
abC	4	3	3	3	2	2	2	1
abc	3	2	2	2	1	1	1	0

23andMe® Commercial DNA test
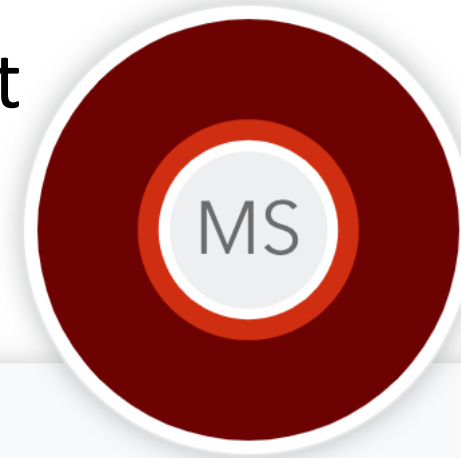
You have <2% **Neanderthal DNA**

MARIE SAITO   **100%**

East Asian & Native American   **100%**

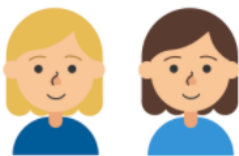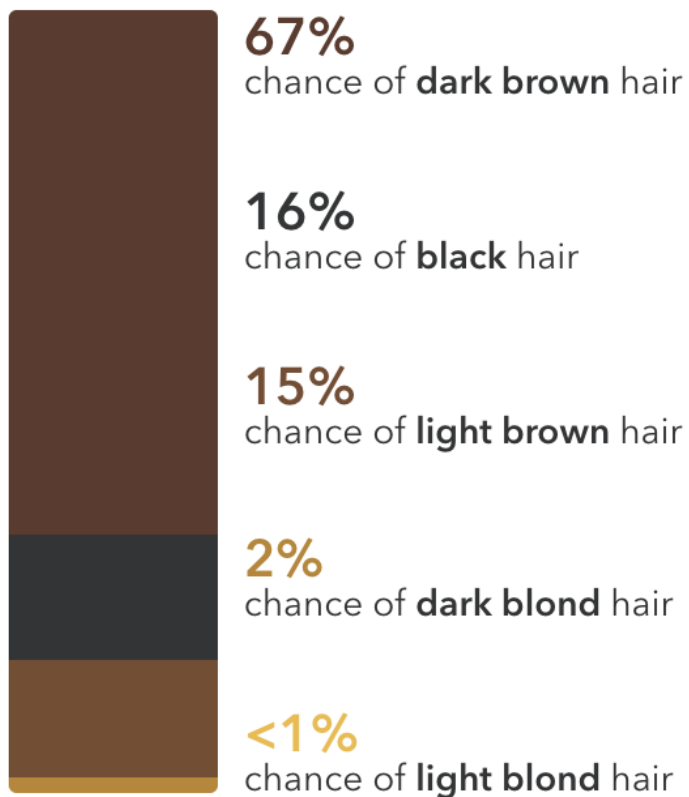● Japanese & Korean   100%

   ● Japanese   100% >

DNA Relatives

Get started with your predicted relationships, then connect and message to learn more.
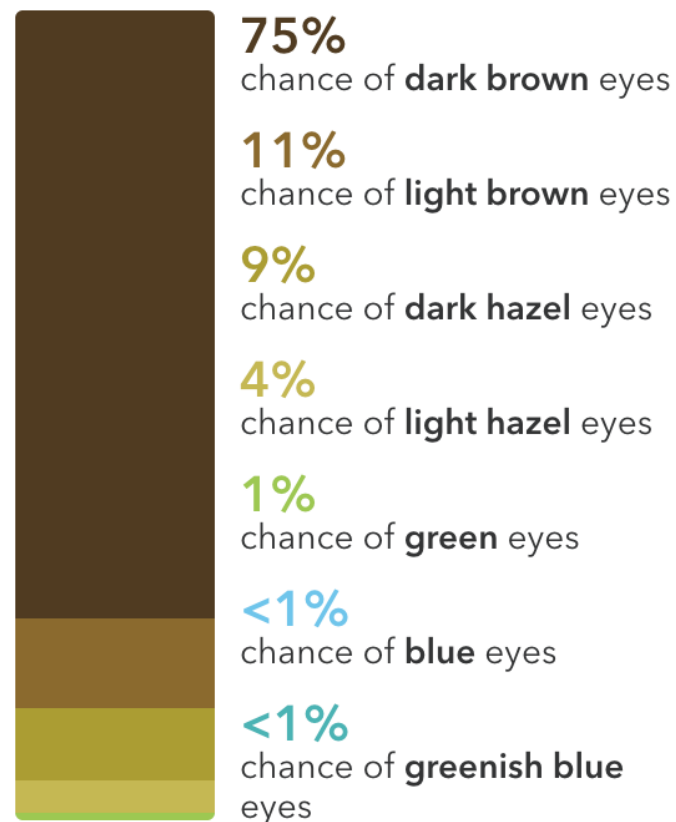
Light or Dark Hair

MARIE, your genetics predict

67%
chance of **dark brown** hair

16%
chance of **black** hair

15%
chance of **light brown** hair

2%
chance of **dark blond** hair

<1%
chance of **light blond** hair

Eye Color

MARIE, your genetics predict

75%
chance of **dark brown** eyes

11%
chance of **light brown** eyes

9%
chance of **dark hazel** eyes

4%
chance of **light hazel** eyes

1%
chance of **green** eyes

<1%
chance of **blue** eyes

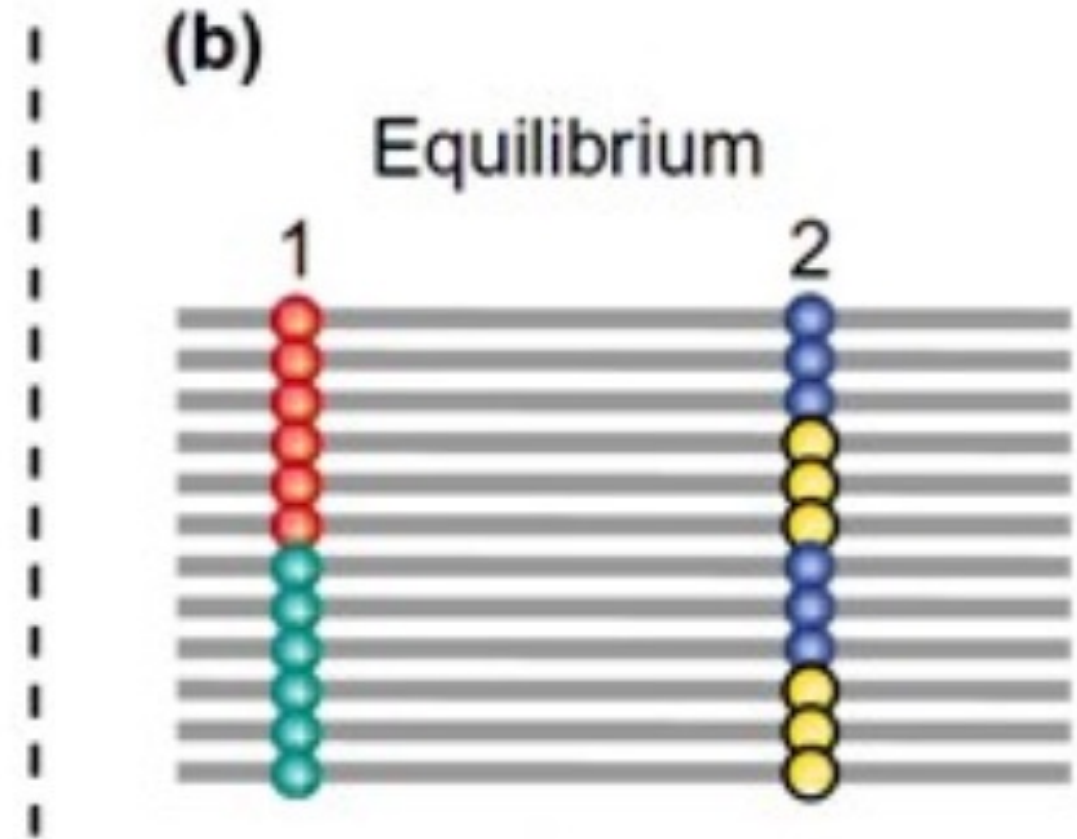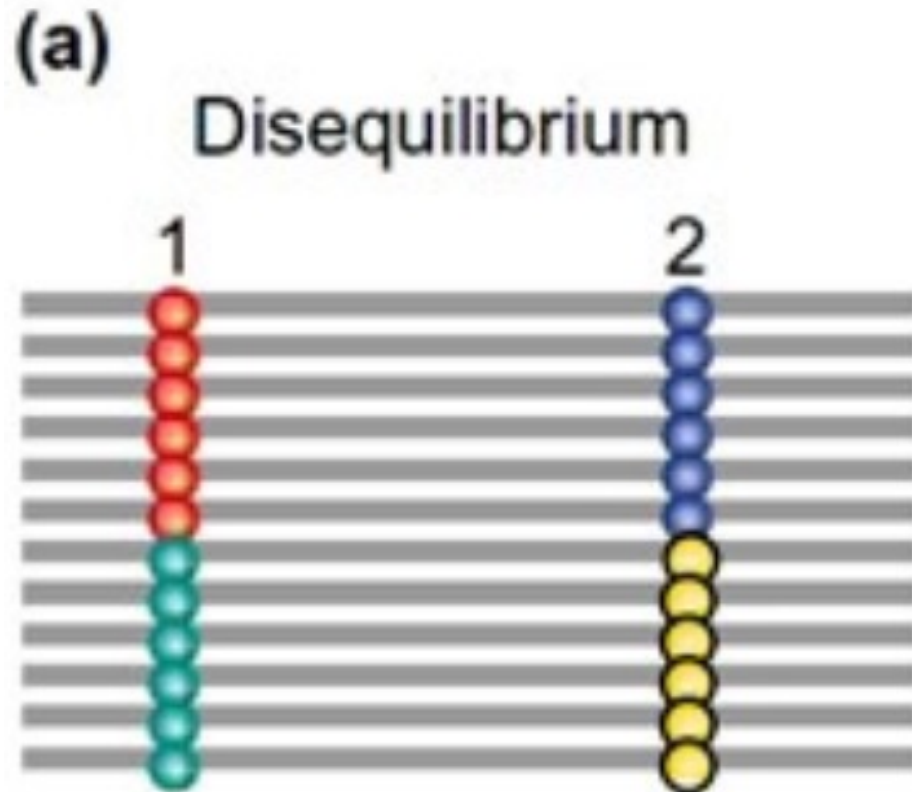**<1%**
chance of **greenish blue** eyes

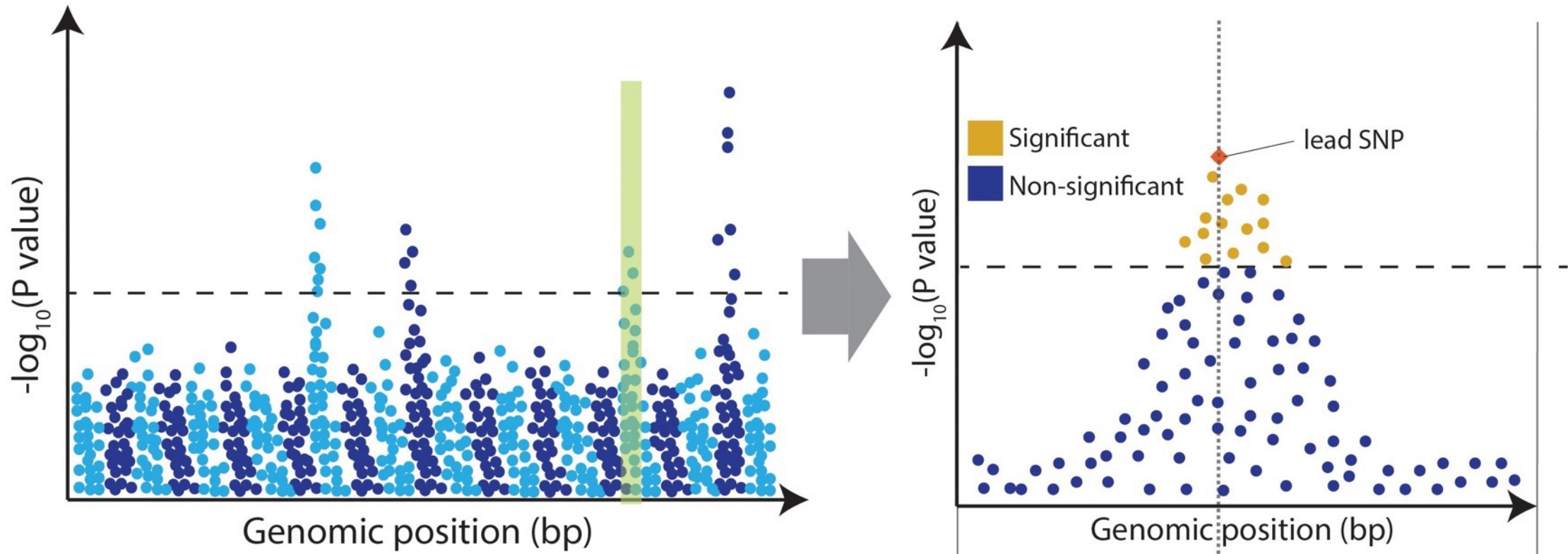# Limitation of genome-wide association studies

- The difficulty in figuring out true causal associations
- They tell us statistical associations, *not biological mechanisms*

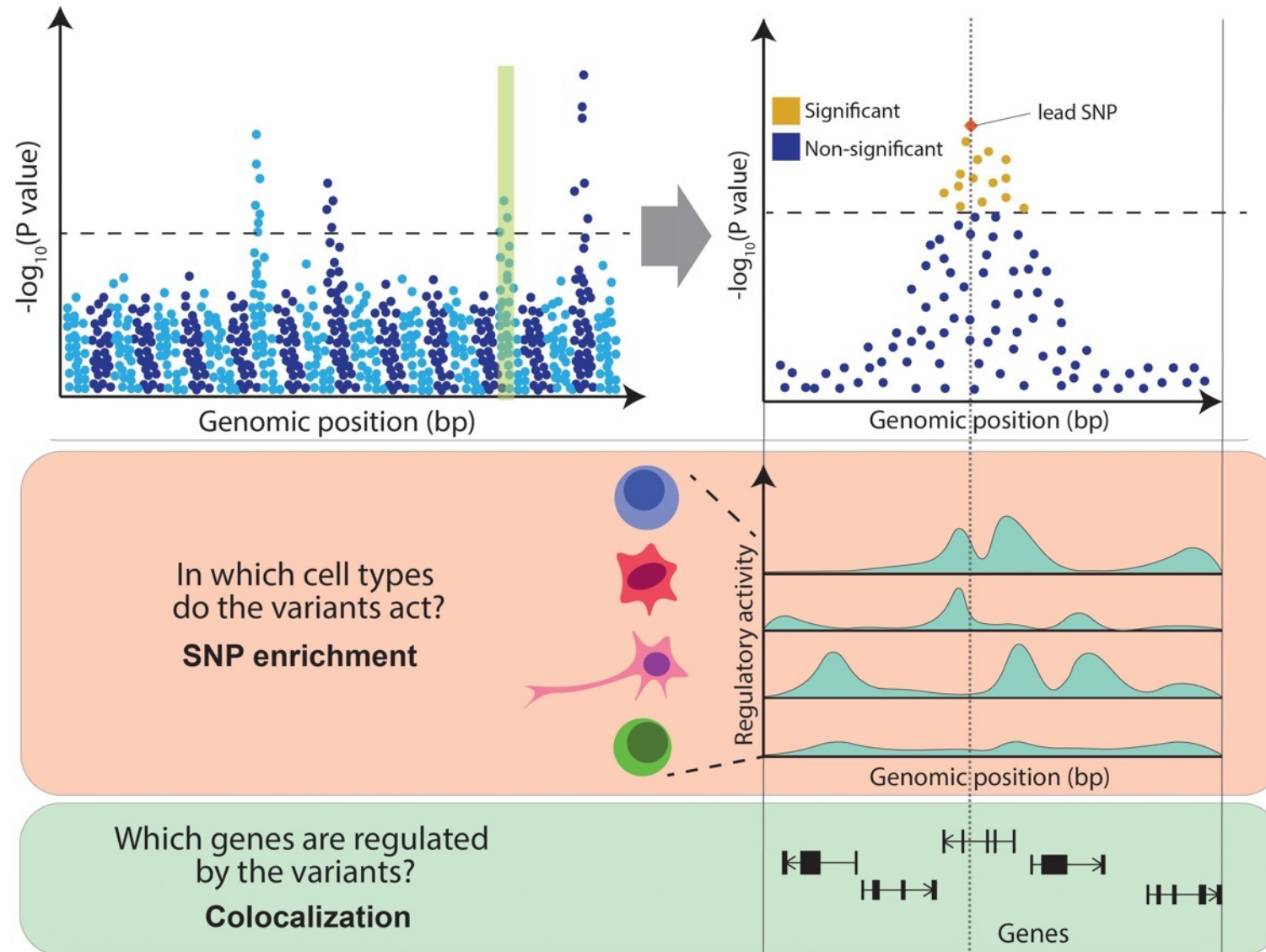# The difficulty in figuring out *true causal associations* - Linkage Disequilibrium
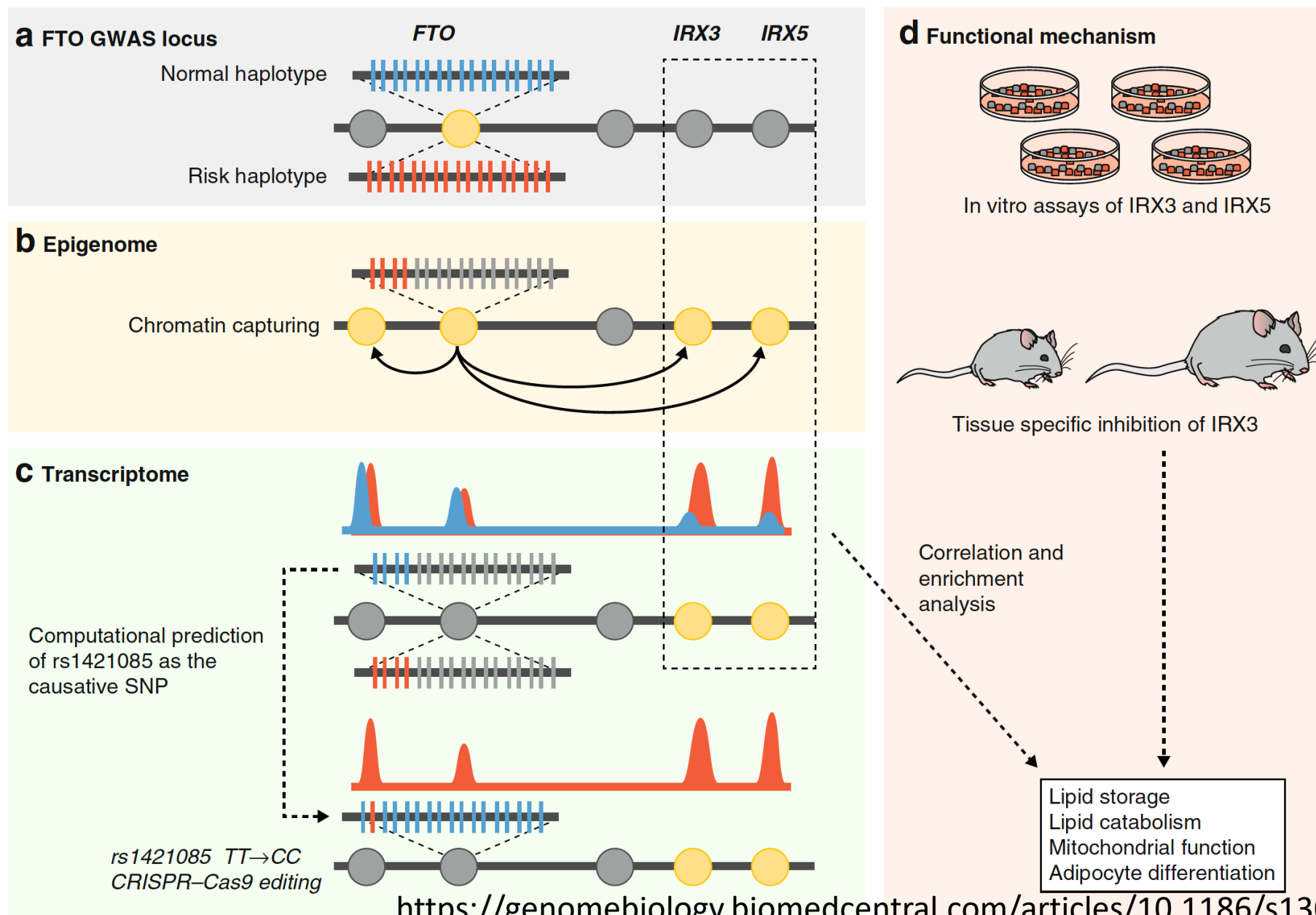

Rafalski. 2002. COPB 5: 94-100

The difficulty in figuring out true causal associations

Multi-omics approach to narrow down causal associations

Multi-omics approach to narrow down causal associations

a FTO GWAS locus

FTO   IRX3   IRX5

Normal haplotype

Risk haplotype

b Epigenome

Chromatin capturing

c Transcriptome

Computational prediction of rs1421085 as the causative SNP

rs1421085 TT→CC
CRISPR–Cas9 editing

d Functional mechanism

In vitro assays of IRX3 and IRX5

Tissue specific inhibition of IRX3

Correlation and enrichment analysis

Lipid storage
Lipid catabolism
Mitochondrial function
Adipocyte differentiation

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1215-1

25

Hands-on exercise