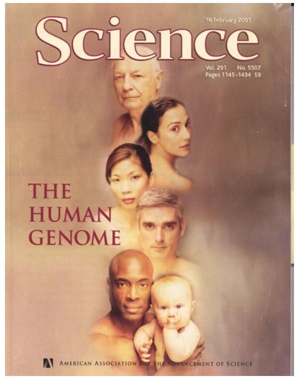
Genome-wide identification of genetic variants and their effects towards gene regulation and disease

Beatrice Borsari
Gerstein Lab – Molecular Biophysics & Biochemistry Dept – Yale University

10.27.2022

Milestones in genetics & genomics



1944:
DNA is the
genetic material

1953:
DNA
double-helix

1966:
Completion of the
genetic code

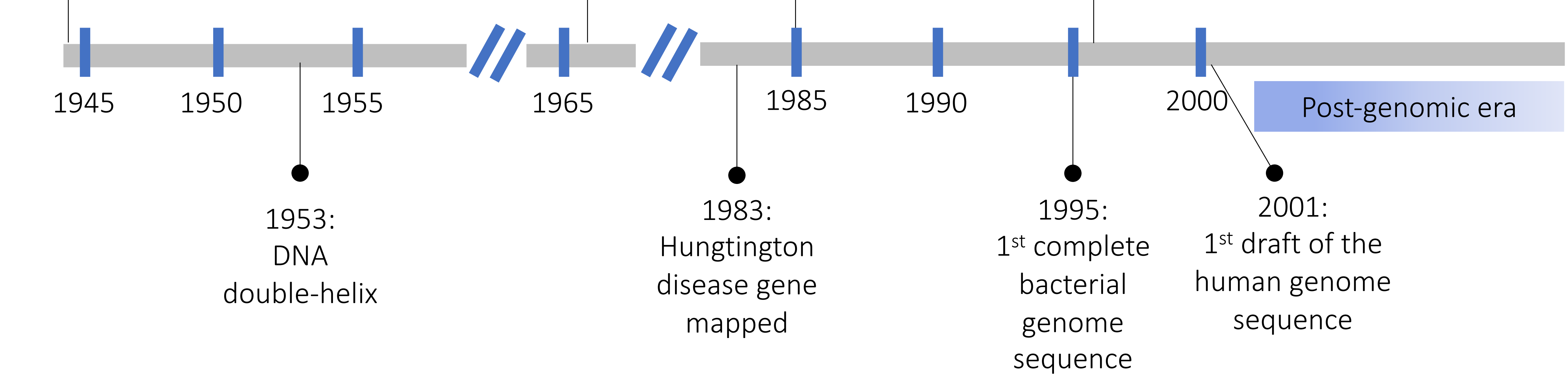
1983:
Hungtington
disease gene
mapped

1985:
PCR

1995:
1st complete
bacterial
genome
sequence

1996:
Complete yeast
genome
sequence

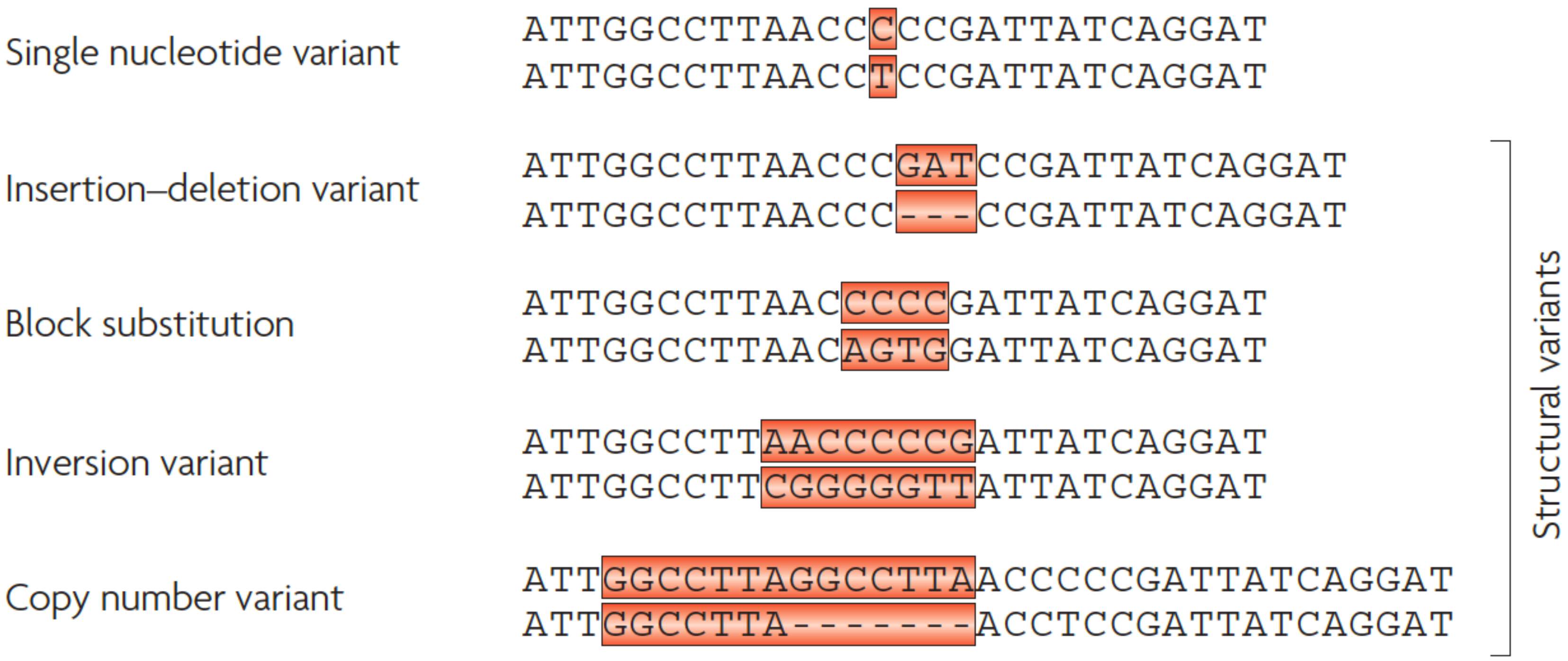
2001:
1st draft of the
human genome
sequence



Post-genomic era

Some key definitions and numbers

- Human genome length: $3\cdot10^9$ bp
- Human-to-human variation: ~0.1% (1:1000 bp); Human-to-chimp variation: ~1-2%
    - $3\cdot10^9$ bp · 0.1% = ~$3\cdot10^6$ DNA variants in an individual
- Different types of variants exist



Single nucleotide variant

ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACCTCCGATTATCAGGAT

Insertion–deletion variant

ATTGGCCTTAACCCGATCCGATTATCAGGAT
ATTGGCCTTAACCC---CCGATTATCAGGAT

Block substitution

ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACAGTGGATTATCAGGAT

Inversion variant

ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTCGGGGGTTATTATCAGGAT

Copy number variant

ATTGGCCTTAGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTA--------ACCTCCGATTATCAGGAT

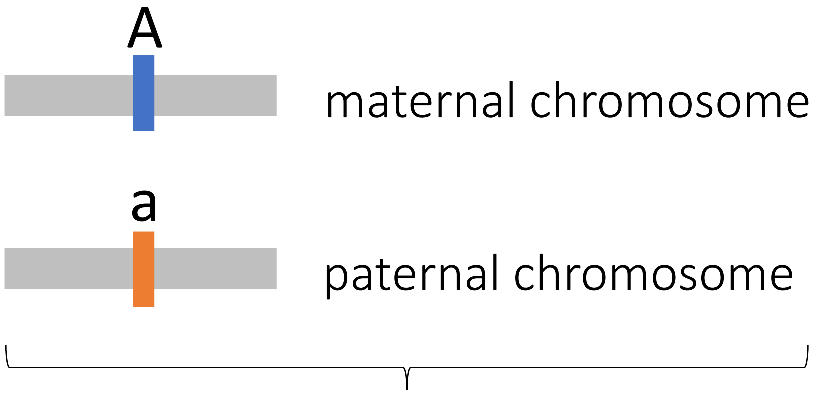
Structural variants

Frazer et al, Nat Rev Genet 2009

Some key definitions and numbers

Alleles and genotypes:

- Allele is the nucleotide present at a given locus (position) in the DNA sequence
- The pair of maternal and paternal alleles at a given locus is the genotype



Maternal allele:	*A*
Paternal allele:	*a*
Genotype:	*Aa*

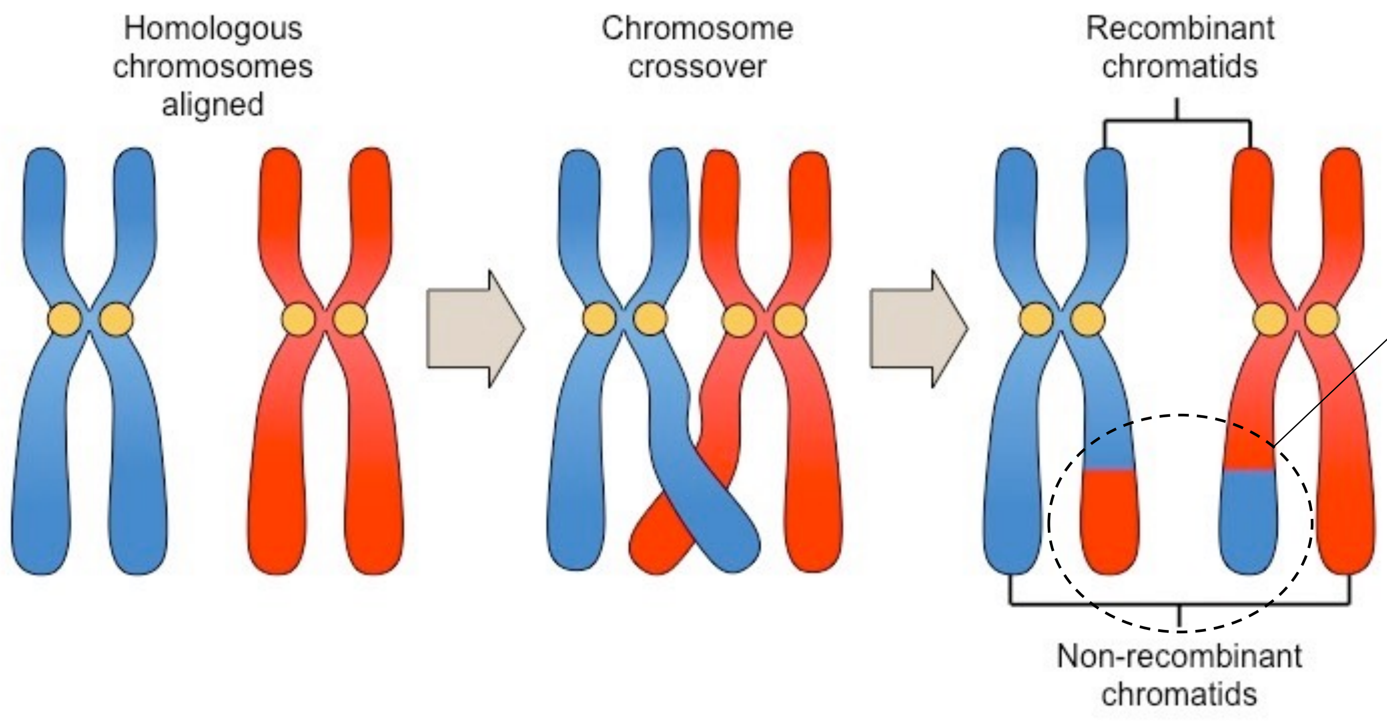
Frequency of variants in the population

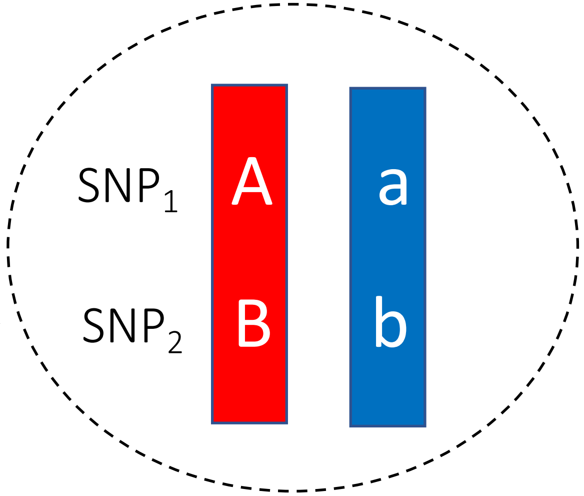
- At a locus, there is usually an allele more frequently observed in the population (*major* allele, e.g. *A*), and one less frequently observed (*minor* allele, e.g. *a*)
- The frequency of the minor allele (MAF) is an important metric to distinguish between common and rare variants
  - MAF is usually retrieved from pilot cohort studies (1000 genomes project)
  - Rare variants usually defined for MAF < 1%

Linkage Disequilibrium (LD)

LD is the non-random association of alleles at nearby loci in the genome

- Alleles of SNPs that reside near one another on a chromosome tend to occur in non-random combinations: their frequency of co-occurrence is higher than one would expect if the loci were independent
- Several factors contribute to LD, but the most important one is chromosomal recombination during gametogenesis (**crossing over**)



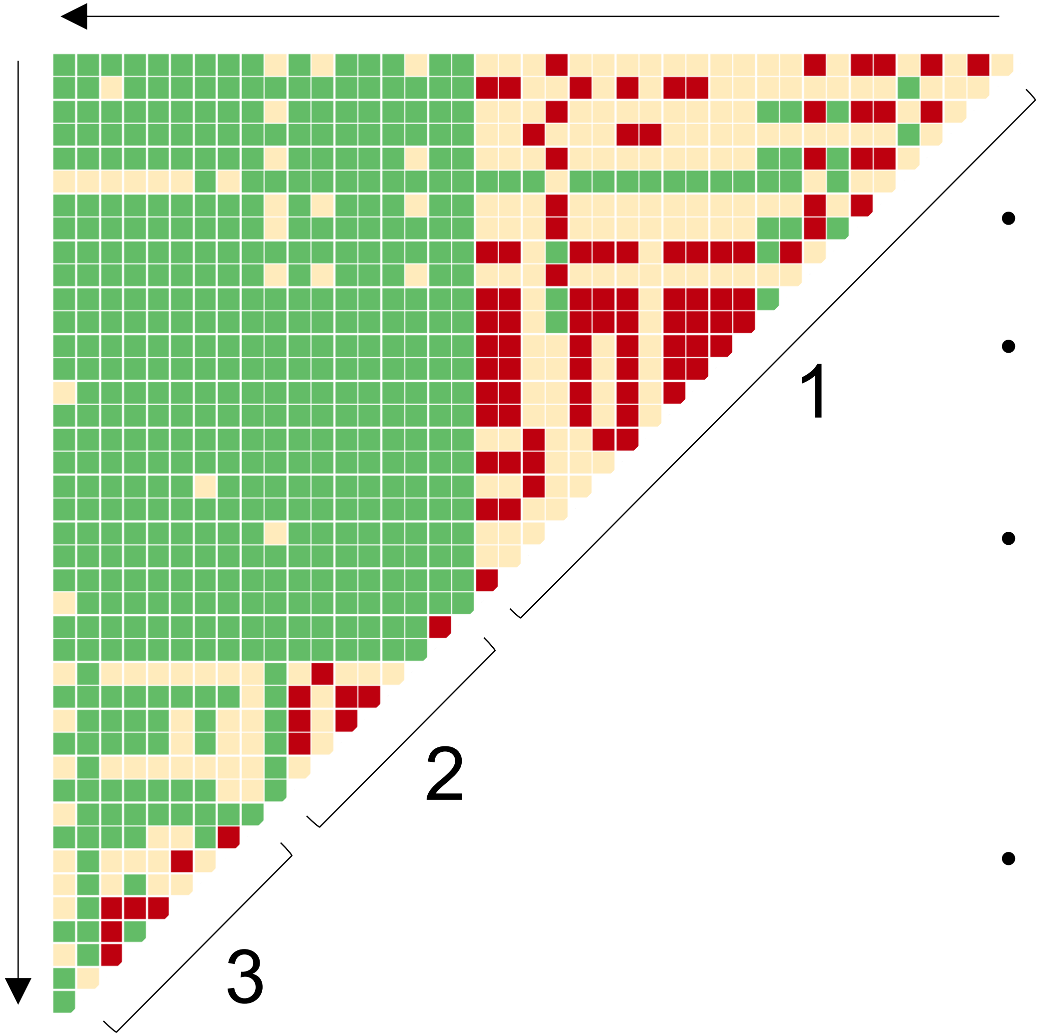


$$p_{AB} > p_A \cdot p_B$$
$$p_{ab} > p_a \cdot p_b$$
$$p_{Ab} < p_A \cdot p_b$$
$$p_{aB} < p_a \cdot p_B$$

Minor alleles *a* and *b* at nearby SNPs tend to be co-inherited on the same chromatid after crossing over

Linkage Disequilibrium (LD)

~300 Kb region in an European population



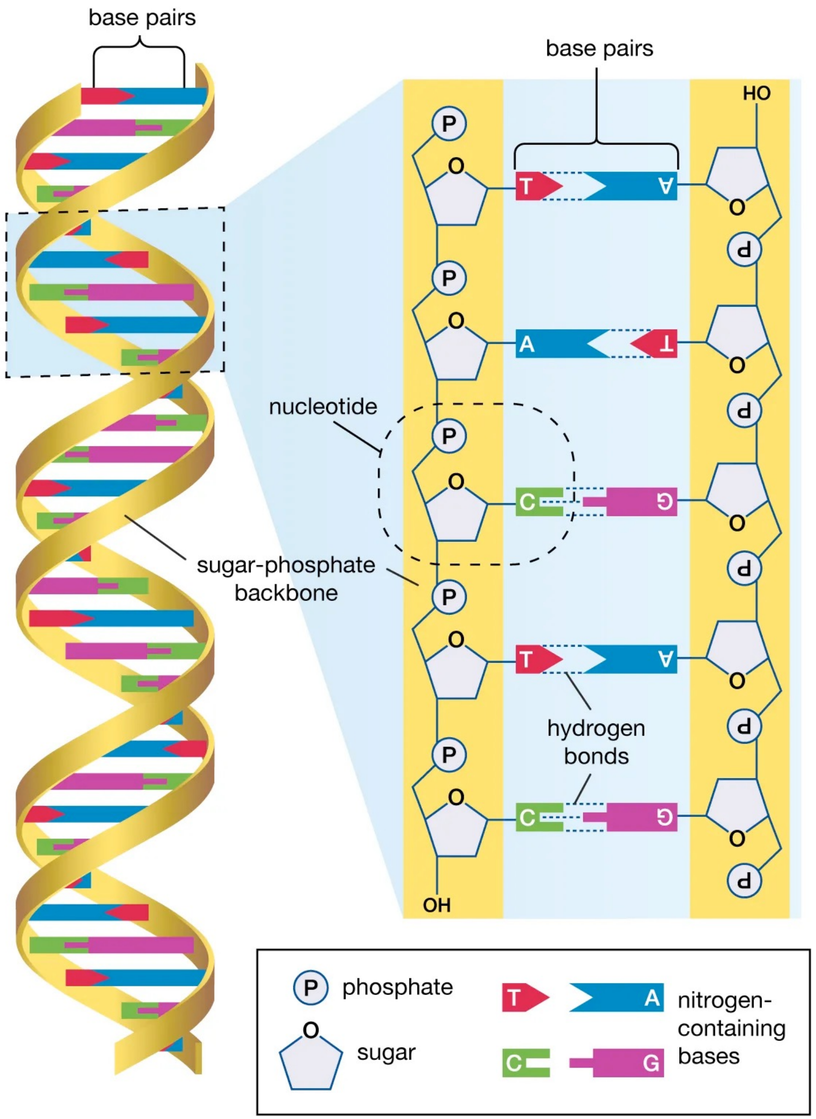
Wall and Pritchard, Nat Rev Genet 2003

LD between alleles at two loci is usually measured as:

$$D_{AB} = p_{AB} - p_A \cdot p_B$$

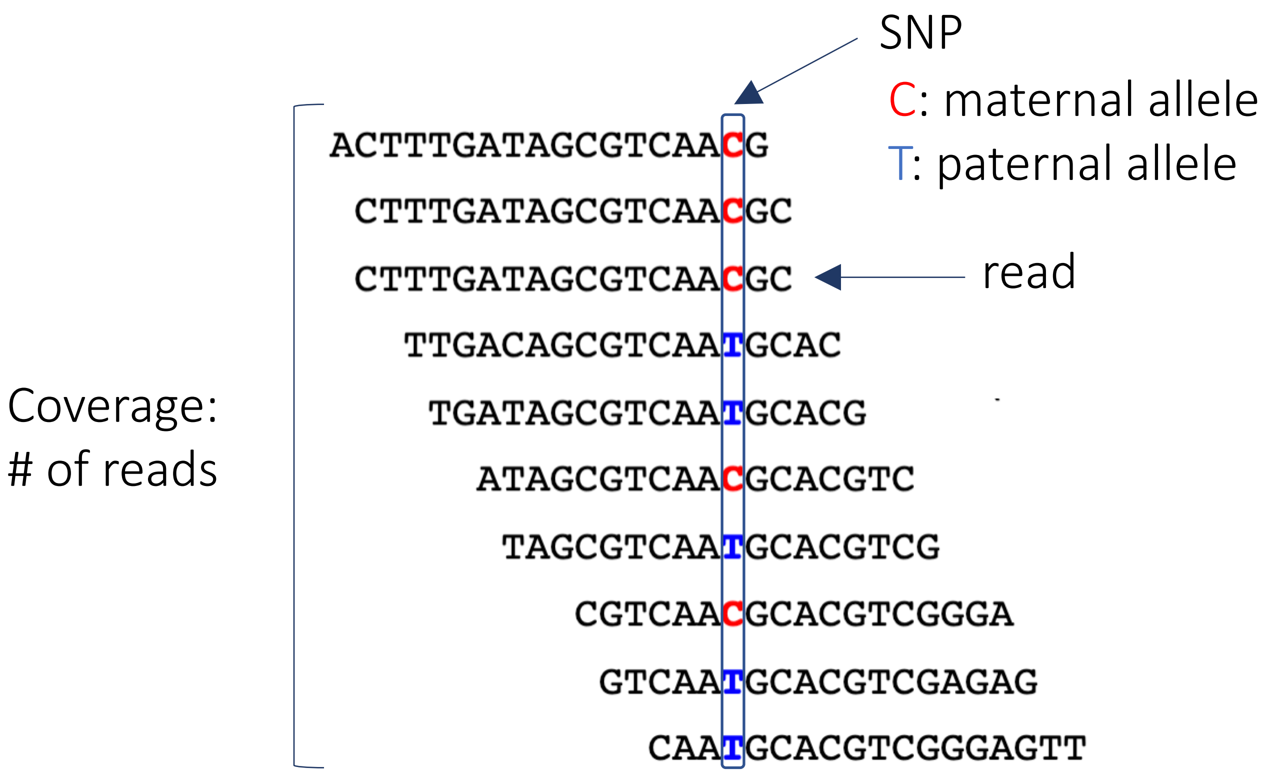
- Three macro LD blocks in this region
- Alleles at SNPs within the same block on one chromosome form a **haplotype**
- The boundaries of these blocks depend on several factors besides recombination rate: natural selection, genetic drift, population bottleneck, inbreeding → LD blocks largely vary across ethnicities
- By tagging a few SNPs within a block, we can infer the allele (major or minor) of most other SNPs in the same block

Genome sequencing: reading a string of letters



© Encyclopædia Britannica, Inc.

- Sequencing consists in deciphering the string of letters of a nucleic acid (either DNA or RNA)
- The main outcome of sequencing are reads
- The coverage is the number of times I’m ”reading” a particular position in the genome



SNP

C: maternal allele

T: paternal allele

read

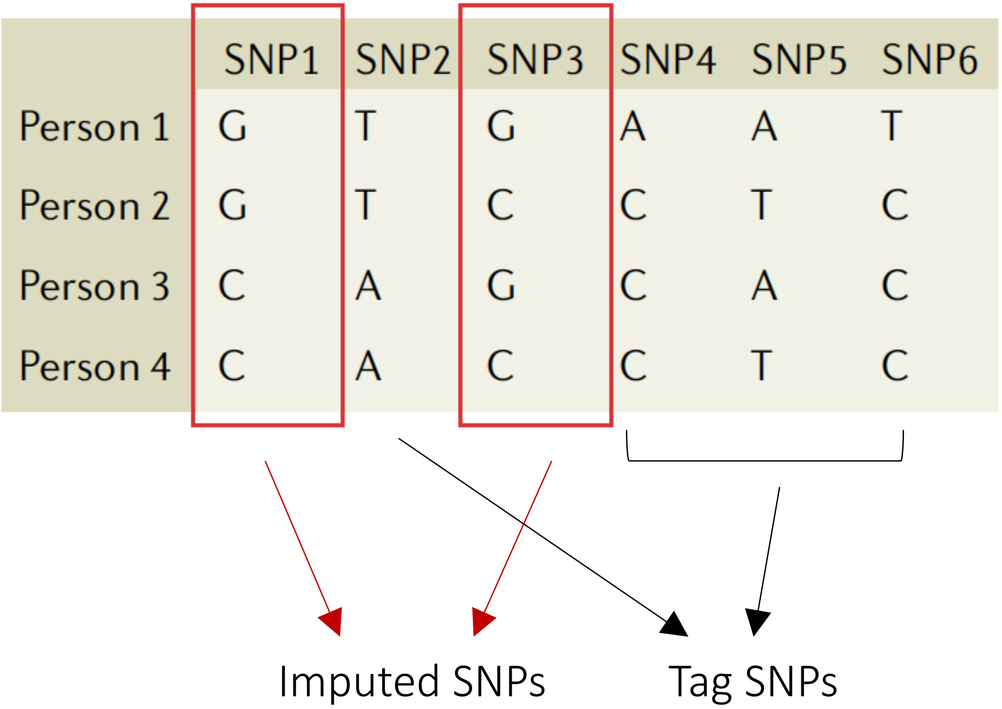
Coverage:
# of reads

```
ACTTTGATAGCGTCAACG
 CTTTGATAGCGTCAACGC
 CTTTGATAGCGTCAACGC
   TTGACAGCGTCAATGCAC
    TGATAGCGTCAATGCACG
      ATAGCGTCAACGCACGTC
       TAGCGTCAATGCACGTCG
          CGTCAACGCACGTCGGGA
           GTCAATGCACGTCGAGAG
             CAATGCACGTCGGGAGTT
```

Identifying variants at a genome-wide scale

STRATEGY #1: Genotyping Arrays

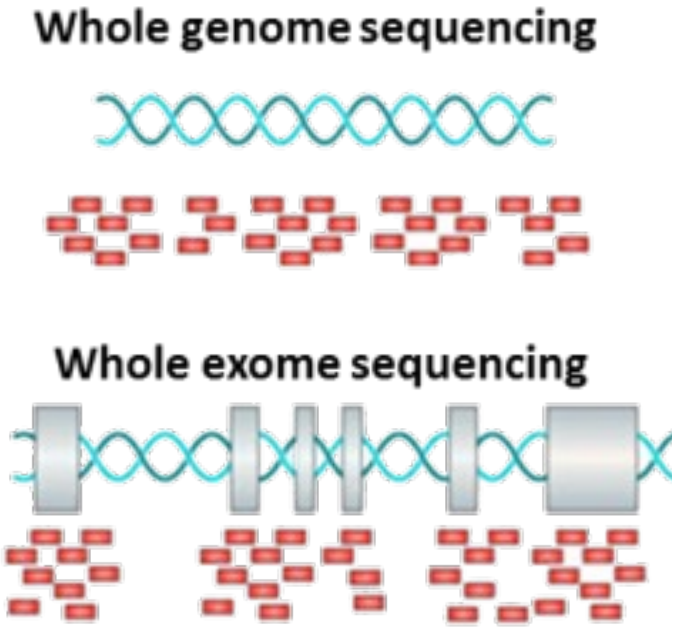
- Platforms that allow to identify the alleles of a pre-determined set of SNPs (tag SNPs)
- Because of LD, we can impute the alleles at all other loci in the same block



	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Person 1	G	T	G	A	A	T
Person 2	G	T	C	C	T	C
Person 3	C	A	G	C	A	C
Person 4	C	A	C	C	T	C

STRATEGY #2: Whole Genome Sequencing

- We obtain allelic information at every of the $3 \cdot 10^9$ bp in the genome
- Highly expensive
- In some cases, whole exome sequencing is preferred as a less expensive strategy



Uffelmann et al., Nature Rev Methods Primers, 2021

Surveying human genome variation

International projects

- 1000 Genomes Project: first catalog of common human genetic variants (~2K healthy individuals, mostly European ancestry)
- International HapMap Project:
    - catalog of LD haplotype blocks across ethnicities
    - ~ 1 M independent common genetic variants (“tag” SNPs)
- Genome Aggregation Database (gnomAD): catalog of allele frequencies (~100K genomes, different ancestries)
- Trans-Omics for Precision Medicine (TOPMed): catalog of genetic variants from disease-specific cohorts (~180K genomes, blood/heart/lung diseases)

gnomAD

National initiatives of precision medicine

- All of Us (USA, 1M participants)
- UK Biobank (UK, 500K participants)
- 2025 France Genomic Medicine Initiative
- Initiative on Rare and Undiagnosed Disease in Japan

All of Us℠ THE FUTURE OF HEALTH BEGINS WITH YOU | The Precision Medicine Initiative®

FRANCE MÉDECINE GÉNOMIQUE 2025 aviesan

biobank uk
Enabling scientific discoveries that improve human health

The post-genomic era: how to fill the gap?

The genome encodes the instructions that determine the biological traits of organisms

But **where** in the genome these instructions are encoded, and **how** they translate into the biological traits of organisms is still mostly **unknown**

Genome-wide association studies (GWAS) can help bridge this gap

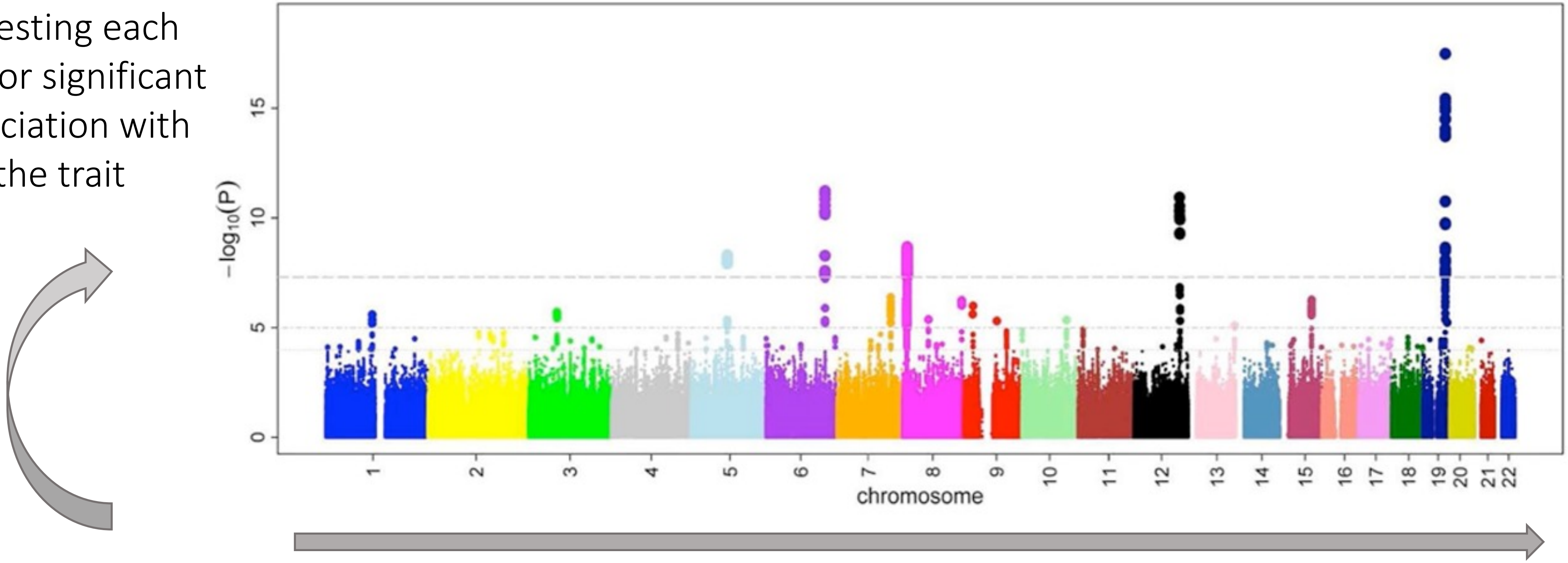
Genome sequence

Organismal traits / diseases

Genome-Wide Association Studies (GWAS)

The basic idea behind a GWAS is to find significant associations between genetic markers and phenotypes (disease / traits) → exploratory "genome-wide" research, non-hypothesis based

Manhattan plot



2. Testing each SNP for significant association with the trait

1. Scanning SNPs across the genome

GWAS: a (multiple) linear regression problem

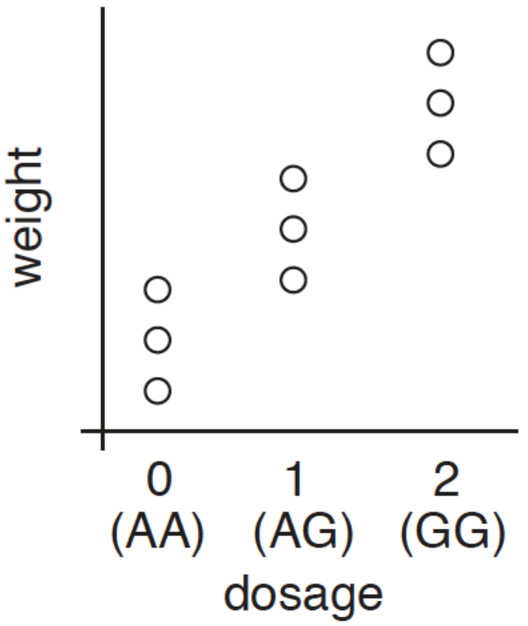
Consider a quantitative trait (eg: weight)

- Consider a SNP *S* with allele$_1$ = A, allele$_2$ = G
- Define three groups of individuals with genotype AA, AG, GG
- The question we try to answer when conducting a GWAS: do we see a significant difference in the weight between these three groups of individuals that correlates with the dosage of allele$_2$?

We can treat this as a linear regression problem:

$$\begin{array}{ccccccccc} y_i & = & \beta_0 & + & \beta_1 & \cdot & x_{1i} & + & \varepsilon_i \\ \swarrow & & \swarrow & & \swarrow & & \searrow & & \searrow \\ \text{weight}_i & = & b_0 & + & b_1 & \cdot & (\text{dosage}_i \text{ of allele}_2) & + & \text{error}_i \end{array}$$

- weight$_i$ = weight of individual *i* = dependent variable
- b$_0$ = intercept
- dosage$_i$ of allele$_2$ = dosage of allele$_2$ in individual *i* = explanatory or independent variable
- b$_1$ = effect of allele$_2$ on the weight of the individual



Theoretical model: assumptions

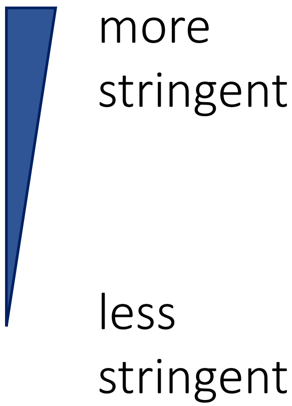
$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \varepsilon_i$$

weight$_i$ = b$_0$ + b$_1$ · (dosage$_i$ of allele$_2$) + error$_i$

error$_i$ is also more commonly called **residual**

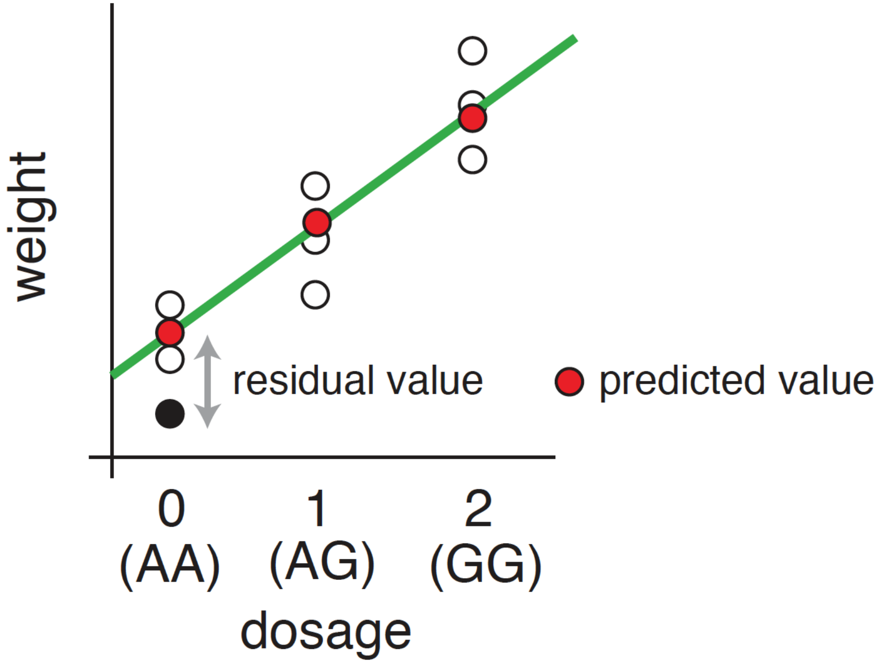
Assumptions

- Linear relationship between y and x
- Homoscedastic residuals (= constant variance)
- Normally-distributed residuals
  - ε_i = ~ Normal(0, σ^2)
- Independent observations



more stringent

less stringent

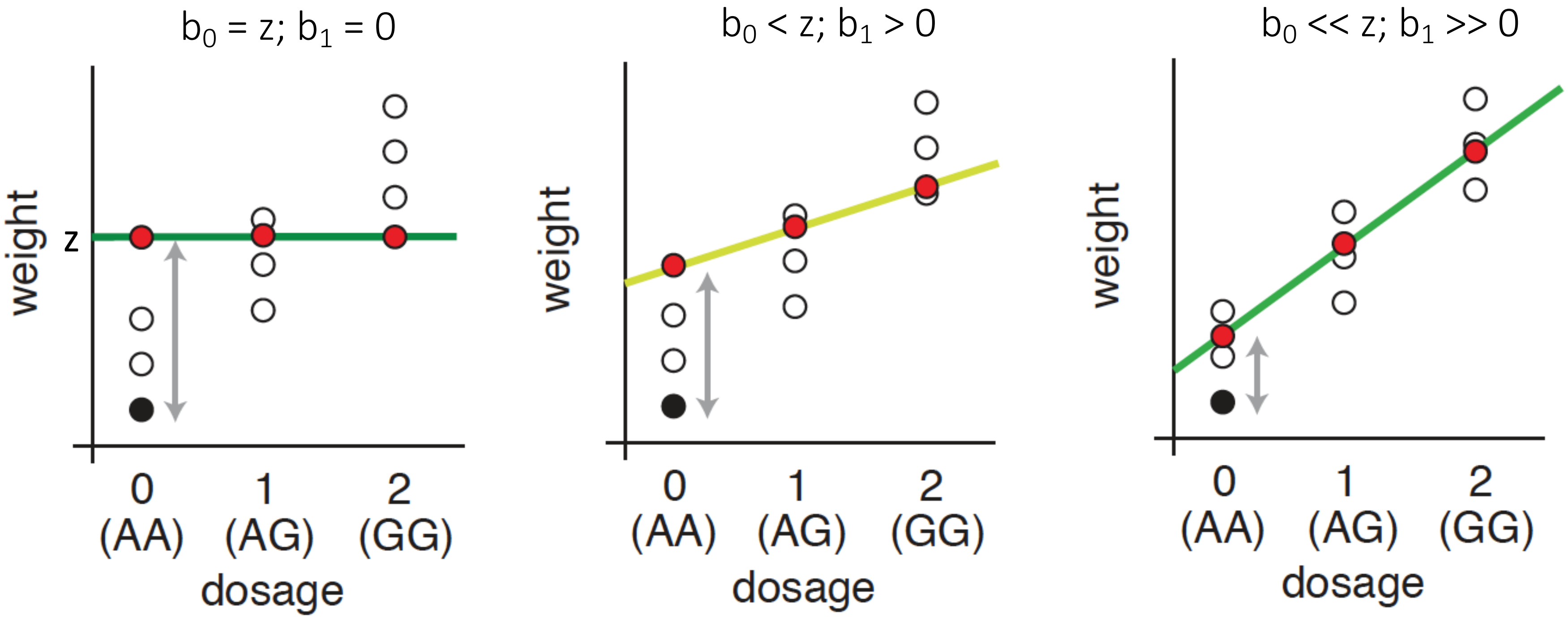


Estimation by Least Squares

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \varepsilon_i$$

weight$_i$ = b$_0$ + b$_1$ · (dosage$_i$ of allele$_2$) + error$_i$

To solve this equation, we apply the **Ordinary Least Squares** criterion: $Q(b_0, b_1) = \boxed{\sum_{i=1}^{n} e_i^2} = \sum_{i=1}^{n} (Y_i - b_0 - b_1 \cdot X_i)^2$



$b_0 = z$; $b_1 = 0$

$b_0 < z$; $b_1 > 0$

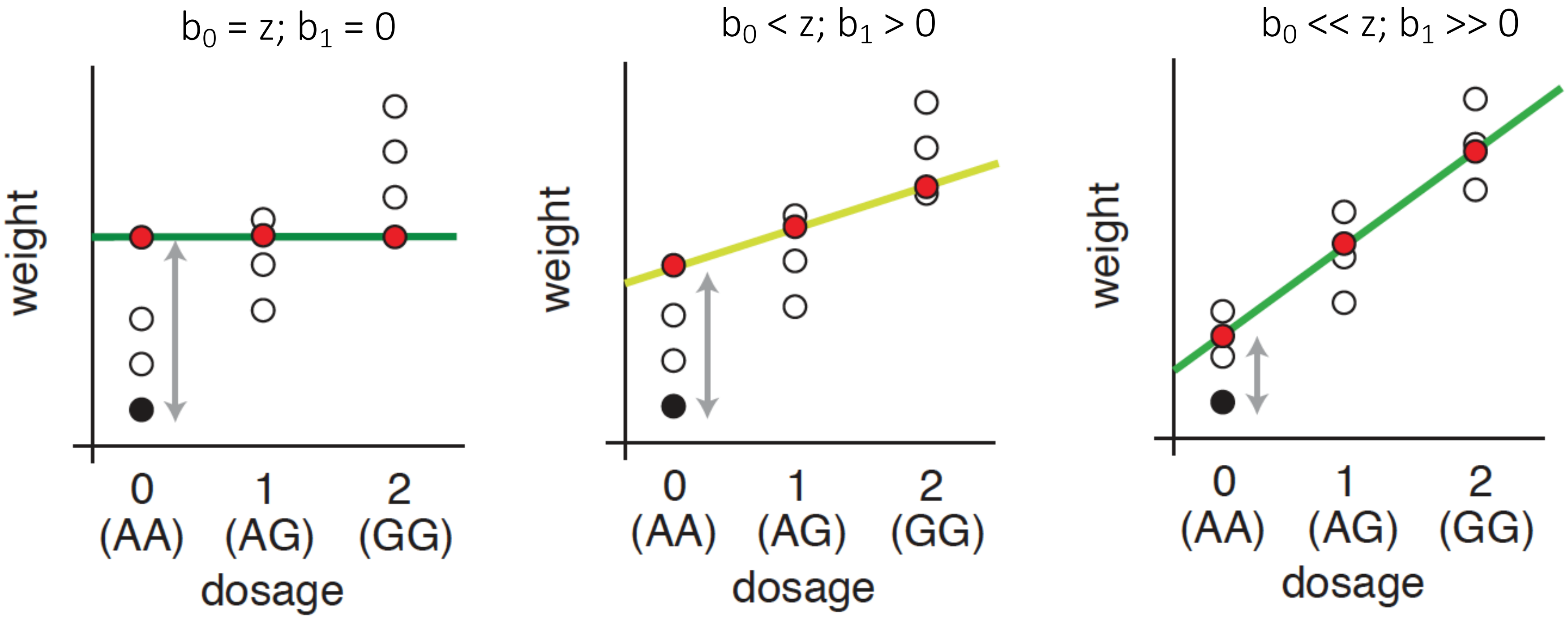
$b_0 \ll z$; $b_1 \gg 0$

In other words, we need to find the combination of b$_0$ and b$_1$ that minimizes the sum of squared residuals across all individuals

Estimation by Least Squares

$$\left.\begin{array}{ll} y_i = \beta_0 + \beta_1 \cdot x_{1i} & \text{statistics} \\ y = mx + b & \text{algebra} \end{array}\right\} \begin{array}{l} b = \beta_0 \\ m = \beta_1 \end{array}$$

Sum of squared residuals across n individuals = $((mx_1 + b) - y_1)^2 + ((mx_2 + b) - y_2)^2 + \ldots + ((mx_n + b) - y_n)^2$



$$m = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$b = \frac{\sum y - m \sum x}{n}$$

GWAS: a (multiple) linear regression problem

Consider a quantitative trait (eg: weight)

- Consider a SNP *S* with allele_1 = A, allele_2 = G
- Define three groups of individuals with genotype AA, AG, GG
- The question we try to answer when conducting a GWAS: do we see a significant difference in the weight between these three individuals that correlates with the dosage of allele_2?

However, things are a little bit more complicated…

Caveat: a phenotype is given by the contribution of both genetic and non-genetic effects

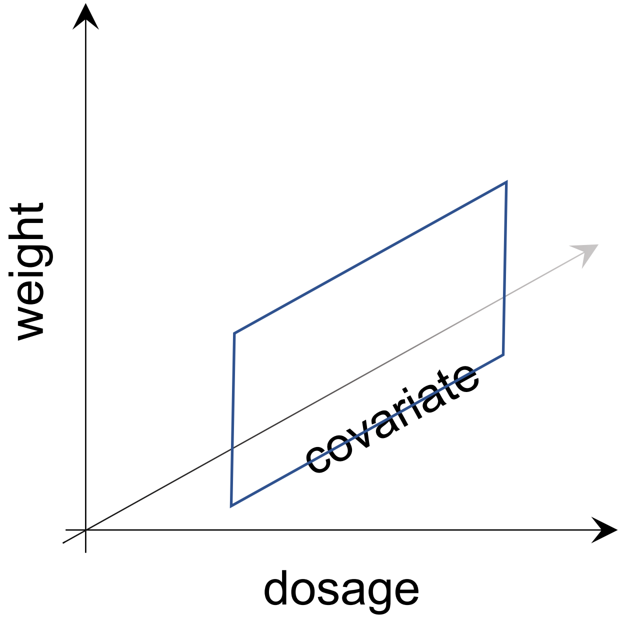
- it might be that, by coincidence, there are more males than females in the GG group, thus we can’t know a priori if the difference in weight is purely given by the effect of the SNP
- it might be that, by coincidence, the diet fatty-acid content varies between the three groups

GWAS: a (multiple) linear regression problem

A multiple regression problem:

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \ldots + \beta_{(p-1)} \cdot x_{(p-1)i} + \varepsilon_i$$

- i = 1 … n observations (individuals / samples)
- y_i = weight of individual i
- x_{1i} = dosage of allele_2 of SNP *S* in individual i (0/1/2)
- $x_{2i} + \ldots + x_{(p-1)i}$ = covariates (age, gender, diet) in individual i
- ε_i = error or residual of the estimated weight for individual i



Goals when performing multiple linear regression:

- Obtain the equation that models the relationship between y and the predictors x
- Test if a specific explanatory variable x has a significant effect in predicting y
  - We are interested in evaluating the effect of SNP *S* on weight

Determining the effect of a SNP on the trait

Question: Does the genotype of SNP *S* (x_1) have a significant effect on the weight of an individual?

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \ldots + \beta_{(p-1)} \cdot x_{(p-1)i} + \varepsilon_i$$

The estimated effect of SNP *S* on weight is b_1 (or $\widehat{\beta_1}$)

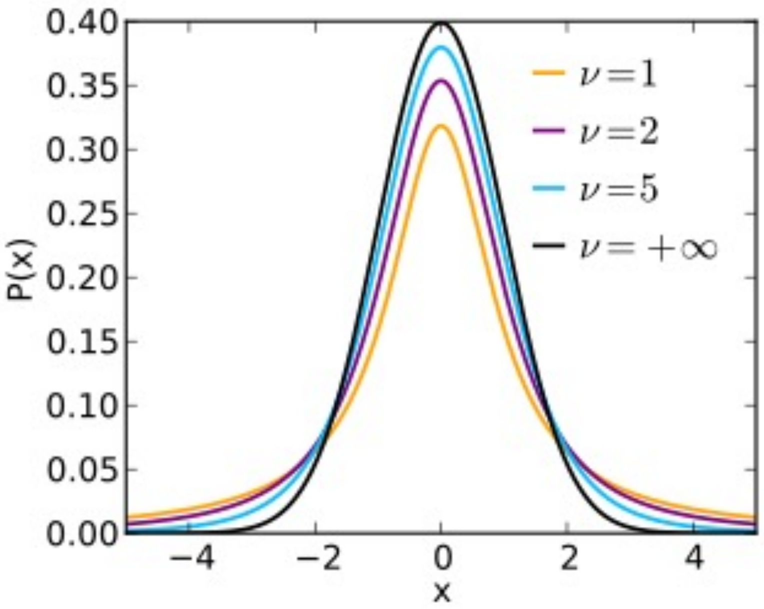
- Under the null hypothesis (no effect of SNP *S* on weight), $\beta_1 = 0$
- We can use the *t*-statistic to compute whether b_1 is significantly different from β_1 (0)

$$t = \frac{b_1 - \beta_1}{SE_{b1}} \rightarrow$$

$$t \sim t_{\mathrm{STUDENT}}$$

$$\upsilon = n - 2$$

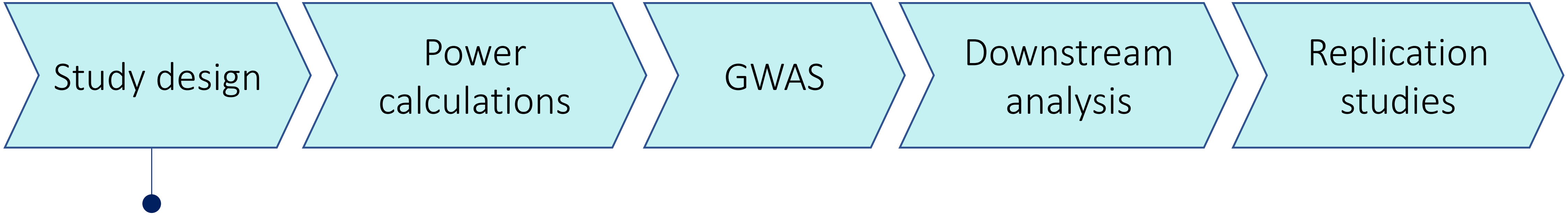
$$n = \text{n of indivs.}$$



→

- p-value < α: reject the null hypothesis, the SNP has a significant effect on weight
- p-value ≥ α : accept the null hypothesis, the SNP does not have a significant effect on weight
- α can be 0.05, 0.01, 0.001

GWAS workflow



Study design

Power calculations

GWAS

Downstream analysis

Replication studies

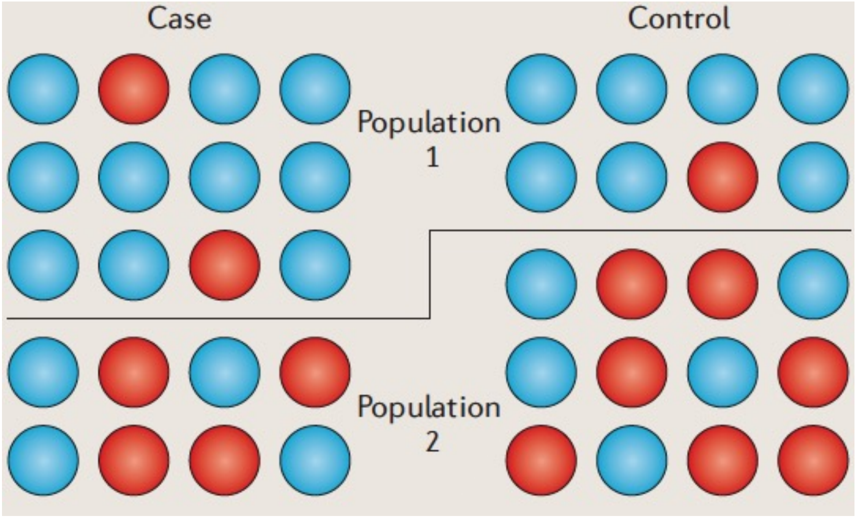
Type of study

- Quantitative trait
- Case-Control study (example: disease vs. healthy)

Choice of relevant covariates

Population stratification

- Some SNPs might have different allele frequencies in different subpopulations (eg. Asian vs. European)

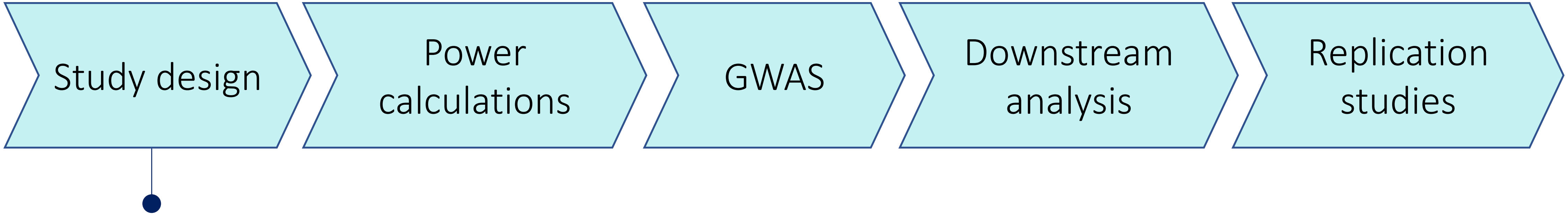


Allele$_2$ = blue

- Enriched in cases
- BUT cases are enriched in population 1, where allele$_2$ is more frequent

Balding, Nat Rev Genet, 2006

GWAS workflow



Study design

Power calculations

GWAS

Downstream analysis

Replication studies

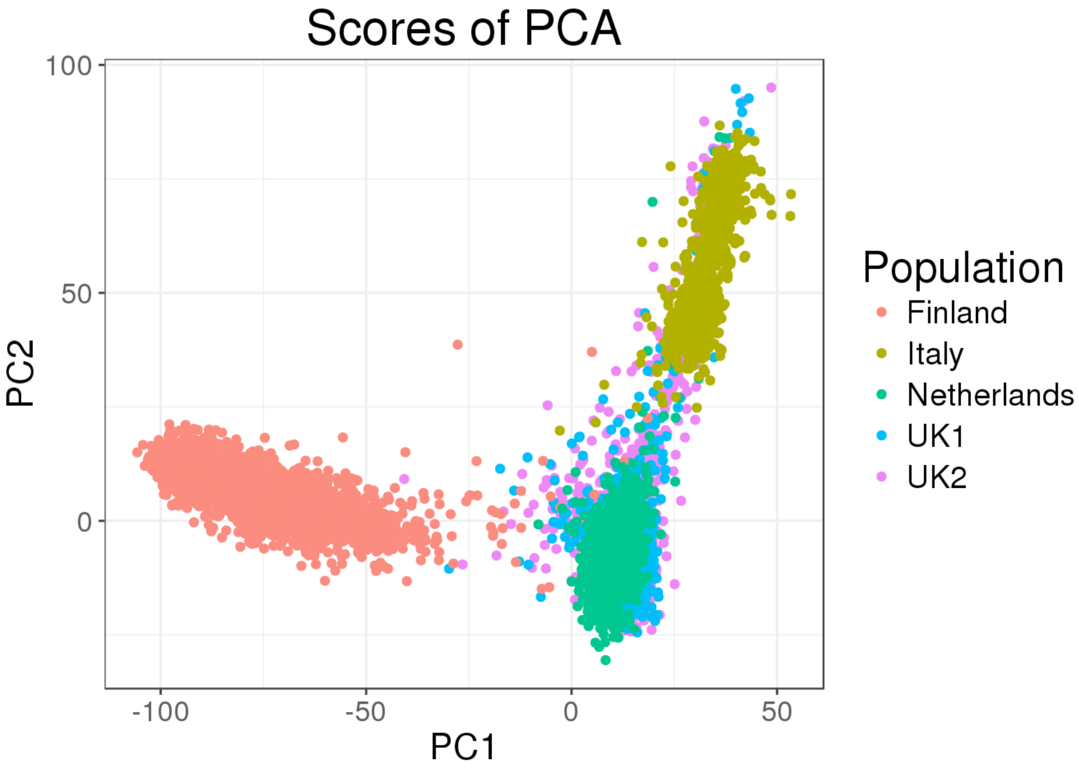
Type of study

- Quantitative trait
- Case-Control study (example: disease vs. healthy)

Choice of relevant covariates

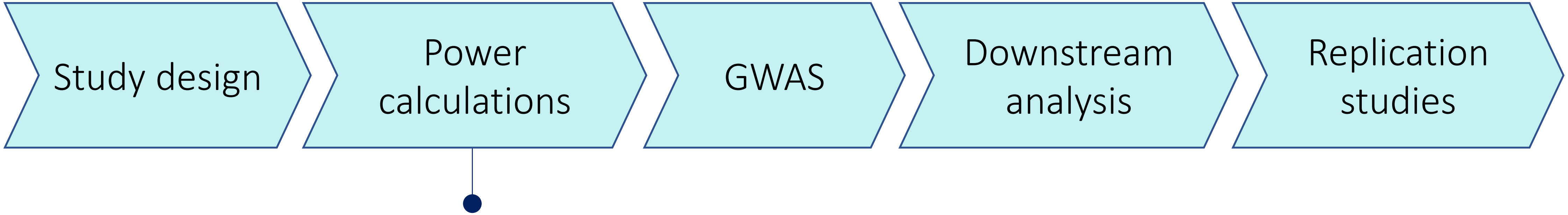
Population stratification

- Some SNPs might have different allele frequencies in different subpopulations (eg. Asian vs. European)
- First 5 or 6 Principal Components based on ancestry are usually included as model covariates



https://privefl.github.io/bigsnpr/articles/how-to-PCA.html

GWAS workflow



Study design

Power calculations

GWAS

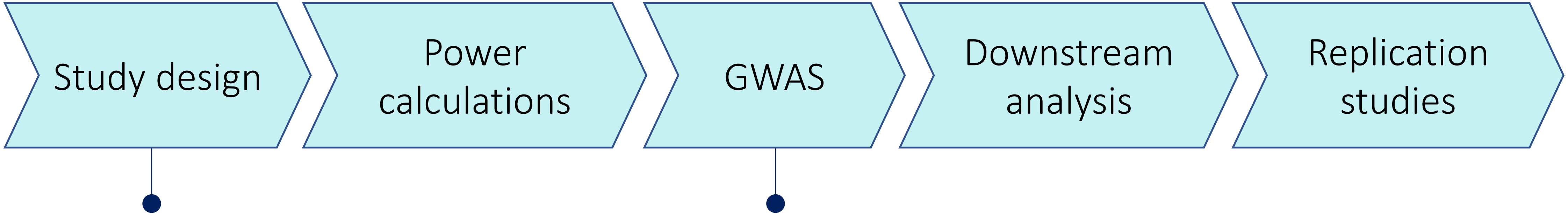
Downstream analysis

Replication studies

- Power is the probability that a SNP is truly associated with a trait
- It depends on sample size, allele frequency and effect size

    - Larger sample size *n* and MAF *f* result in a more accurate estimate of the SNP effect β
    - Larger absolute values of β increase the difference from the null model (e.g. same mean value of the trait across genotype groups)

GWAS workflow



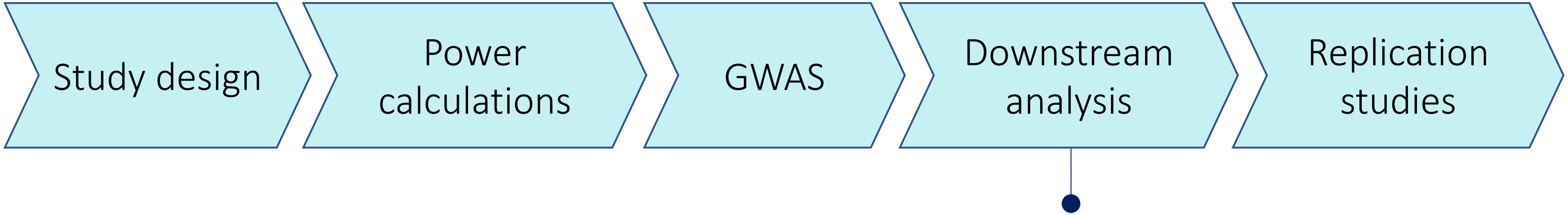
Type of study

- Quantitative trait
- Case-Control study

Statistical model

- Linear regression (beta values)
- Logistic regression (OR)

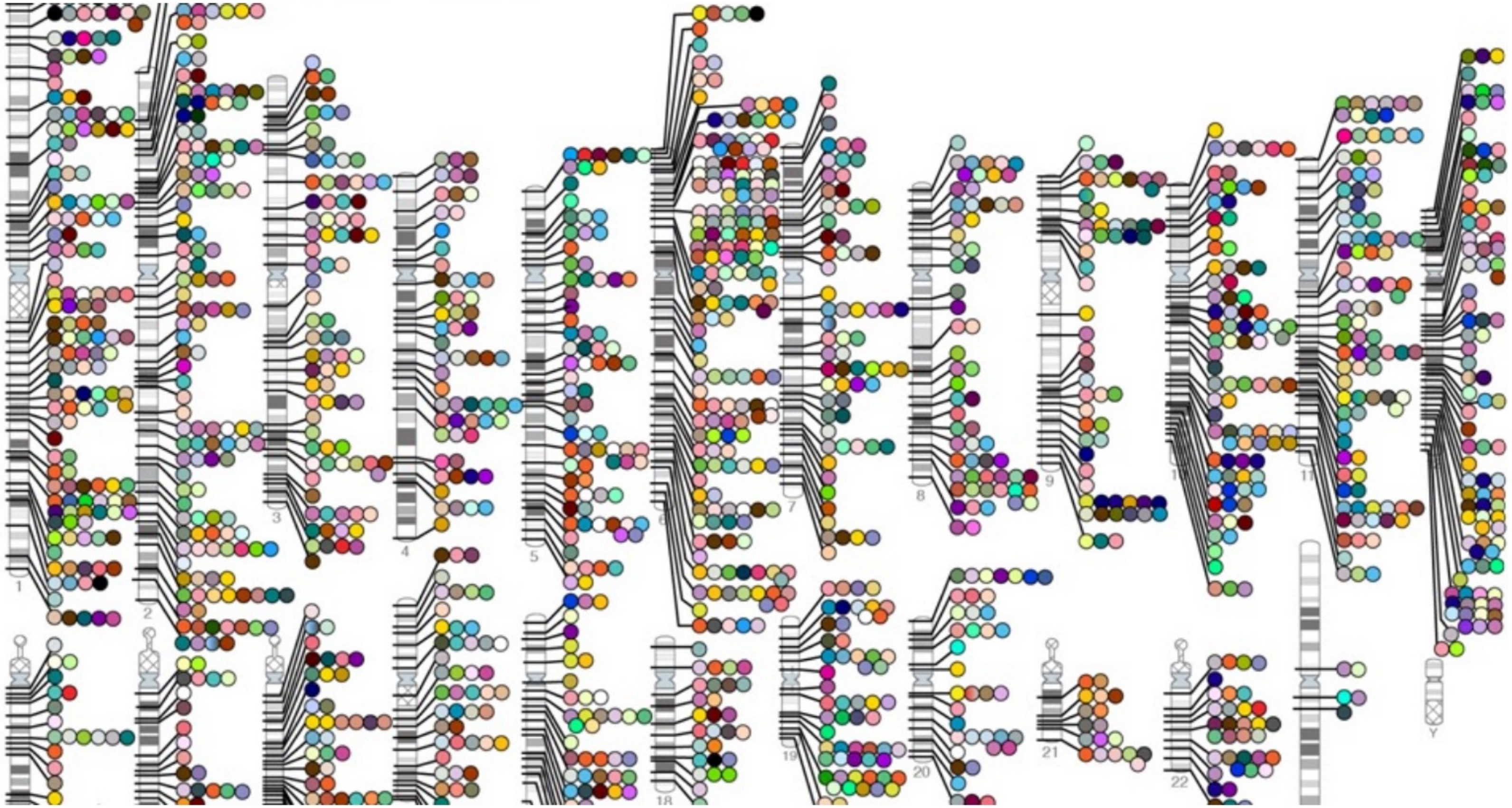
GWAS workflow



- Because of LD, many significant SNPs are indeed the result of indirect associations
- Multiple testing Bonferroni correction:
  - FWER $= \frac{\alpha}{m}$, m = # of independent hypotheses
  - # of independent common variants = 10^6
  - FWER = $0.05/10^6 = 5 \cdot 10^{-8}$

The NHGRI-EBI GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies: https://www.ebi.ac.uk/gwas/



As of 2022-10-08, the GWAS Catalog contains 6041 publications and 427870 associations.
GWAS Catalog data is currently mapped to Genome Assembly GRCh38.p13 and dbSNP Build 154.

The post-genomic era: how to fill the gap?

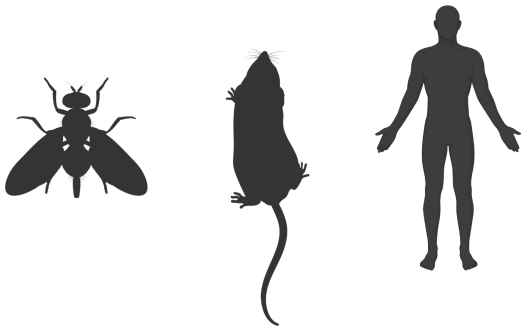
Genome-wide association studies (GWAS) can help bridge this gap

… but most of the times we don’t know what are the molecular mechanisms explaining the effect of a specific variant

Genome sequence → **Molecular phenotypes** → **Organismal traits / diseases**

Molecular phenotypes

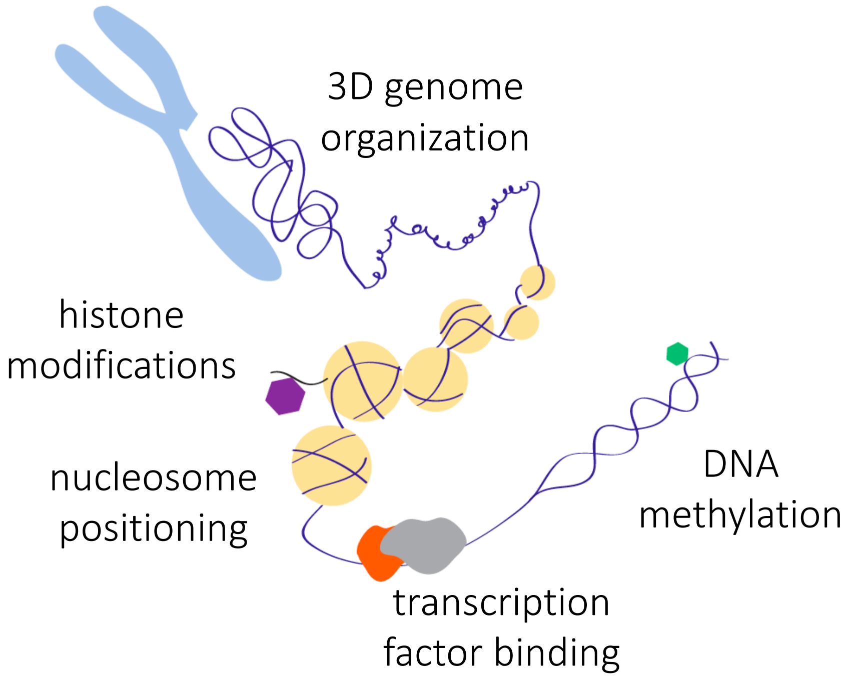
Epigenetic marking
RNA expression
RNA translation
Protein levels

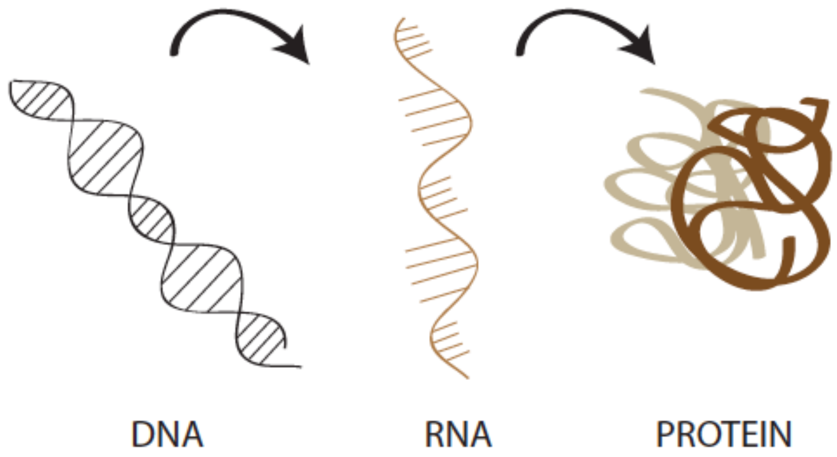


Functional Genomics

Central dogma of molecular biology

Epigenetic regulation (DNA)
ChIP-seq, ATAC-seq, DNase-seq, FAIRE-seq, Hi-C, WGBS, …



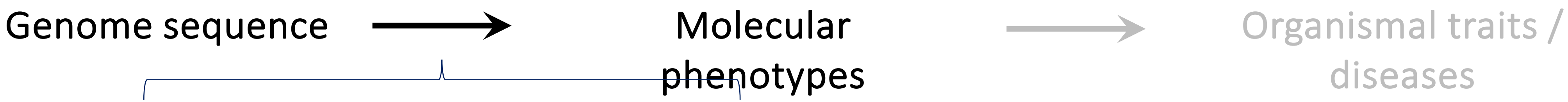


Translation (RNA)
Ribo-seq

Transcription (RNA)
RNA-seq, BRU-seq, …

All these *-*seq* experiments rely on the same principle as genome sequencing: a molecular event is measured in terms of number of reads sequenced at a particular position in the genome

The post-genomic era: how to fill the gap?



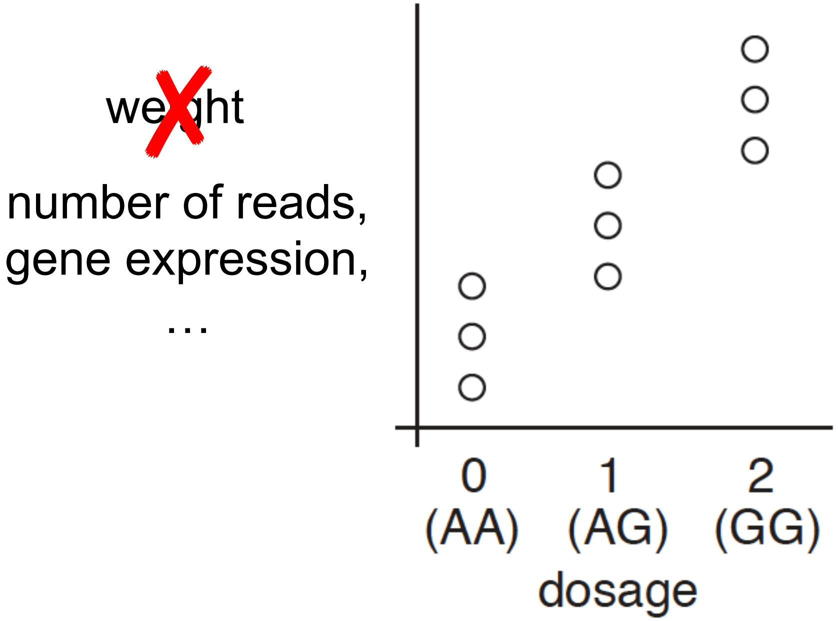
Genome sequence → Molecular phenotypes → Organismal traits / diseases

- molecular quantitative trait loci (QTLs)
- allele-specific (AS) events

Molecular quantitative trait loci

Genome sequence → Molecular phenotypes → Organismal traits / diseases

- molecular quantitative trait loci (QTLs)
- allele-specific (AS) events



- Population-scale analysis
- Same concept as GWAS for quantitative traits (linear models, effects modeled as beta coefficients)

Allele-Specific events

Genome sequence → Molecular phenotypes → Organismal traits / diseases

- molecular quantitative trait loci (QTLs)
- allele-specific (AS) events

e.g. a SNV @ chr 7 position 4345325 ☒

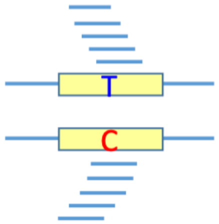
RNA-/ChIP-Seq Reads

```
ACTTTGATAGCGTCAACG
  CTTTGATAGCGTCAACGC
  CTTTGATAGCGTCAACGC
      TTGACAGCGTCAATGCAC
        TGATAGCGTCAATGCACG
          ATAGCGTCAACGCACGTC
            TAGCGTCAATGCACGTCG
               CGTCAACGCACGTCGGGA
                 GTCAATGCACGTCGAGAG
                   CAATGCACGTCGGGAGTT
```

5 x T (ref)

5 x C

Allelic ratio = 0.5
(i.e. 'null' expectation)



e.g. a SNV @ chr 5 position 12455 ☑

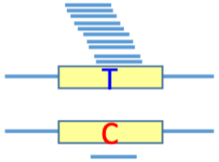
RNA-/ChIP-Seq Reads

```
ACTTTGATAGCGTCAATG
  CTTTGATAGCGTCAATGC
  CTTTGATAGCGTCAATGC
      TTGACAGCGTCAATGCAC
        TGATAGCGTCAATGCACG
          ATAGCGTCAATGCACGTC
            TAGCGTCAATGCACGTCG
               CGTCAACGCACGTCGGGA
                 GTCAATGCACGTCGAGAG
                   CAATGCACGTCGGGAGTT
```

9 x T (ref)

1 x C

Allelic ratio = 0.9



- Not a population-scale analysis
- Can be performed for all heterozygous SNPs within a single genome
- Allelic ratio is modelled with a binomial distribution
  - n = total # of reads at the SNP
  - k = # of ref allele
  - p = 0.5

Chen et al, Nat Comms 2016