Genome-wide Association Studies (GWAS)

Vinod Kumar, PhD
Dept. of Genetics, UMCG

Big subject

Lots of methods, tools and software packages

Out of scope for today:

- Structural variants and copy number variations (CNVs)
- Meta-analysis
- Imputation

Why is genetics useful in Medicine?

Identifying the genetic cause(s) of diseases will help to

clarify underlying biological mechanisms, which are needed to develop drugs

identify individuals at high risk for specific care and follow up

provide personalized treatments based on genetic defects

GWAS steps

The following steps are taken to perform a GWAS:

1. Quality control per SNP

1.1 Remove SNPs with minor allele frequency (MAF) < 0.01

1.2 Remove SNPs that deviate from Hardy Weinberg equilibrium (HWE) using a P threshold < 1e-6

1.3 Remove SNPs with missing genotype rate < 0.99

2. Quality control per individual

2.1 Identify and remove related individuals

2.2 Perform principal component analysis (PCA)and plot the first two components to identify and remove population outliers

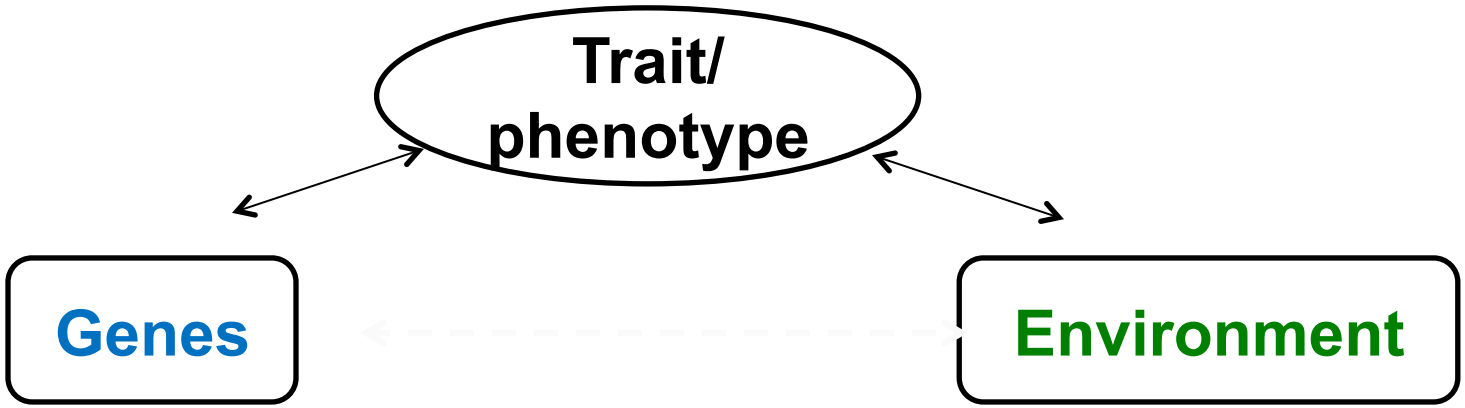
2.3 Perform genome-wide association testing

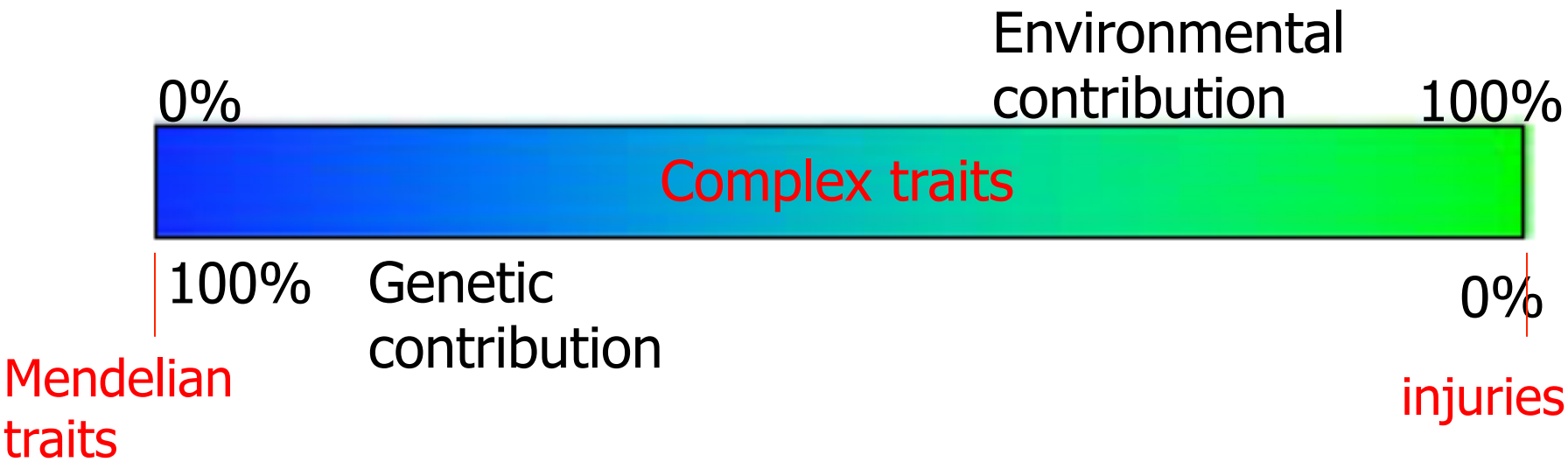
2.4 do initial association analysis, linear

2.5 do association analysis, logistic (Bonus)

2.6 prepare a manhattan plot

Genetic contribution can vary





Multifactorial phenotypes - GWAS

- Height
- Weight
- Behavior
- Chronic diseases
- Susceptibility to infectious diseases
- Response to drugs
- …many other phenotypes

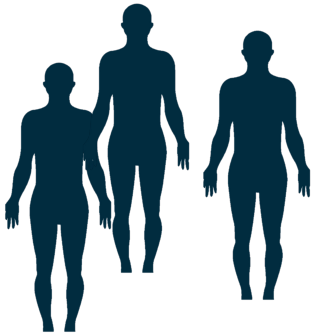
How do we study the role of genetics in a disease?

Read the genome of affected individuals and look for similarities among them, and differences compared to healthy samples

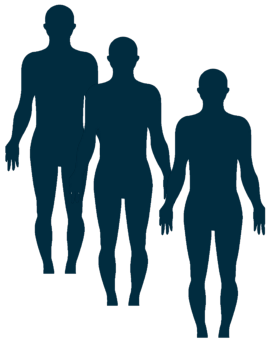
Candidate-gene approach 2000-2005

Old approach:

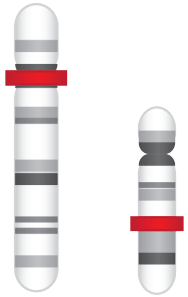
Search for “association” of genetic variants with a disease using independent cases

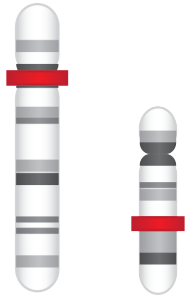


patients (cases)

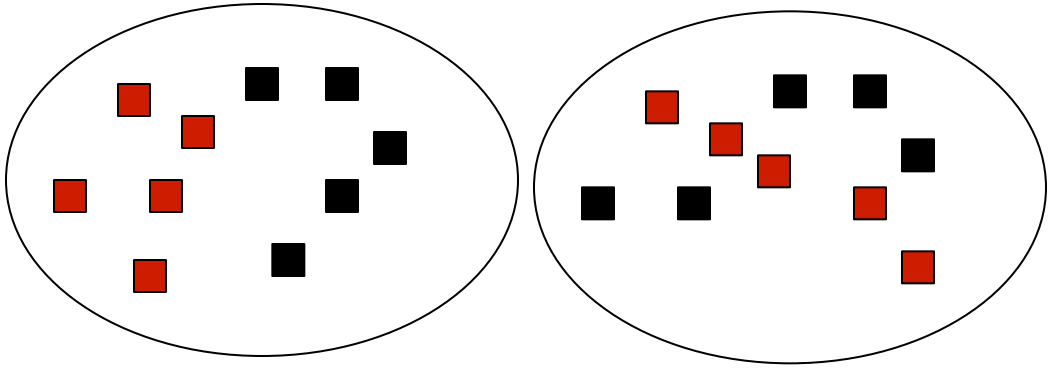


healthy samples (controls)





Variant #1

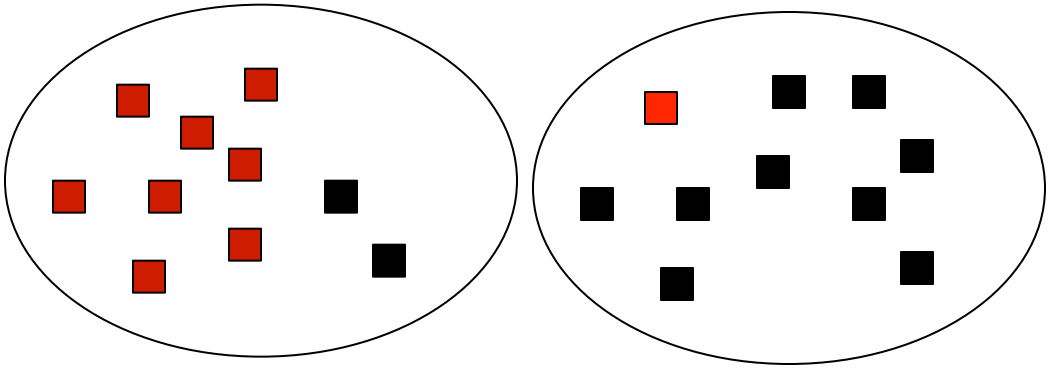


patients (cases)

healthy samples (controls)

No association

Variant #2

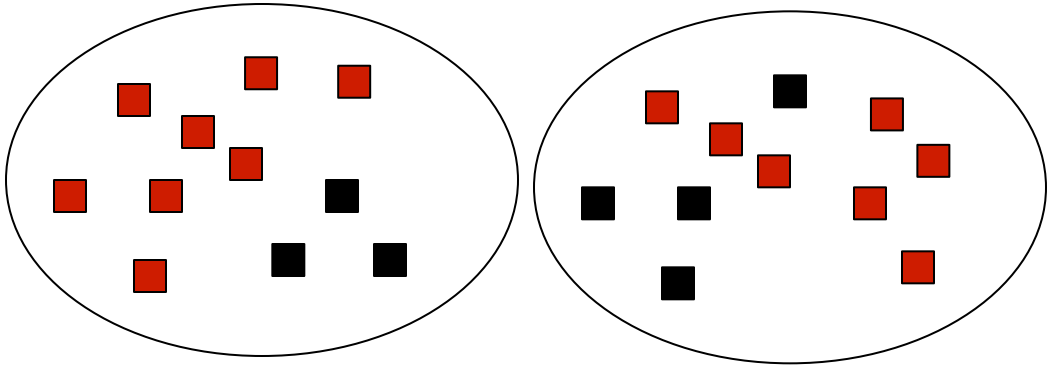


patients (cases)

healthy samples (controls)

Association of red type with disease

Variant #3

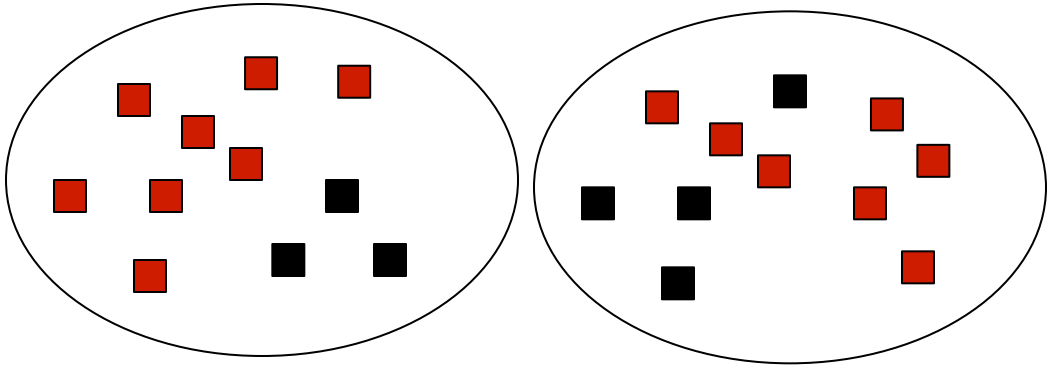


patients (cases)

healthy samples (controls)

Can we claim association?

Variant #3



patients (cases)

healthy samples (controls)

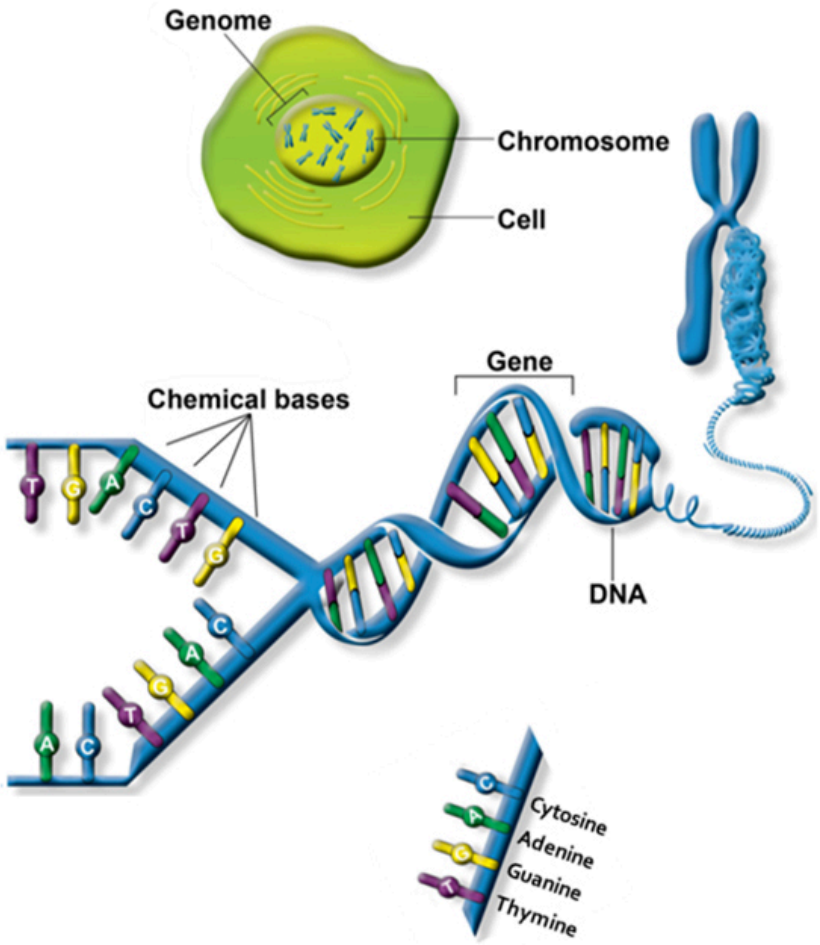
Can we claim association?

Need to calculate probability (pvalue) with statistical test:

how likely is that the red variant is NOT associated to the disease?

which genetic variants?

- SNP – single nucleotide polymorphisms
- Microsatellites
- Copy number variants

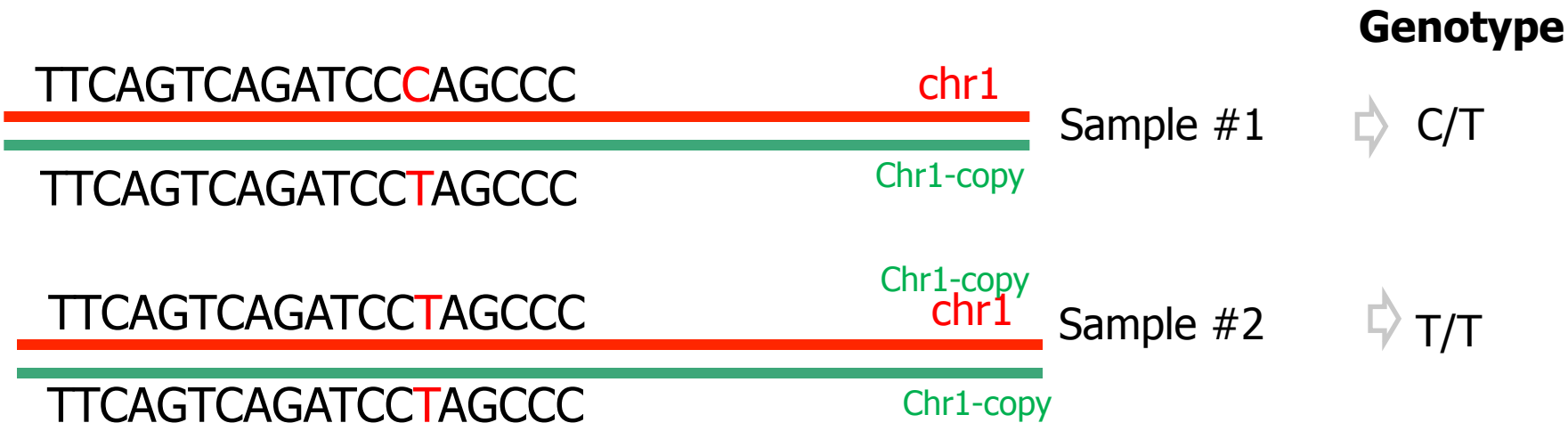


which genetic variants?

SNP:

Single Nucleotyde Polymorphism C -> T (alleles)

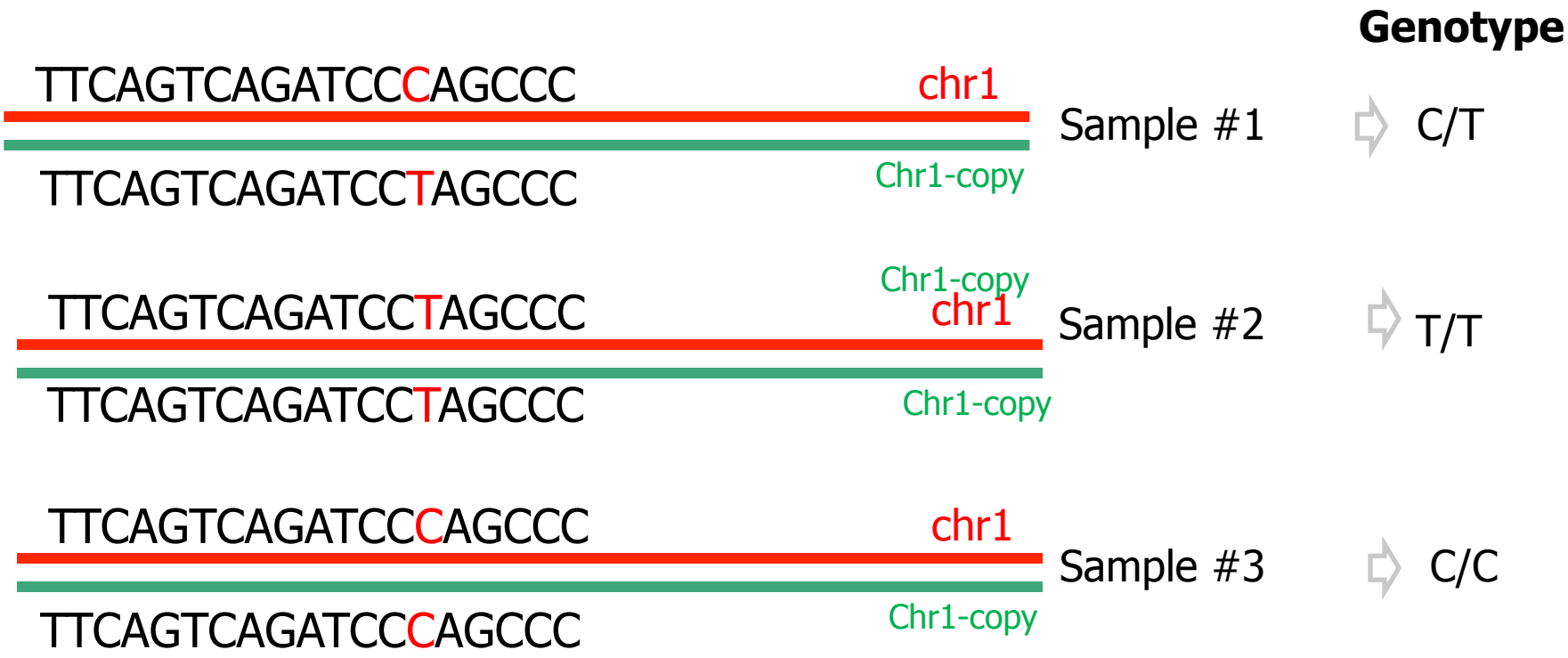
which genetic variants?



SNP:

Single Nucleotyde Polymorphism C -> T (alleles)

which genetic variants?



SNP:

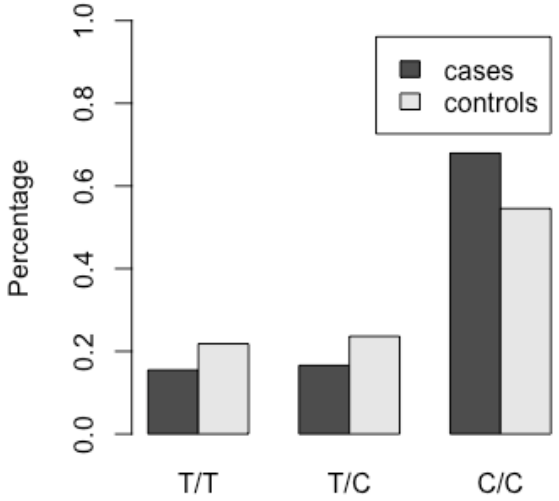
Single Nucleotyde Polymorphism C -> T (alleles)

Statistical test for association

- Test for differences in allele is the most used and simplest.

is there a particular allele more (less) present in the patient group?

Genotype	Cases	Controls
T/T	98	120
T/C	105	130
C/C	430	300

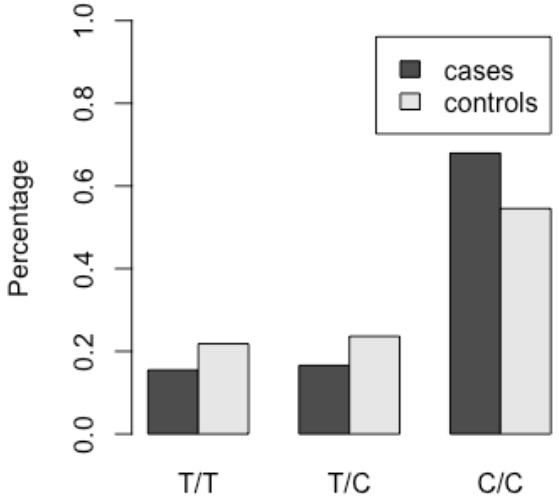


Statistical test for association

- Test for differences in allele is the most used and simplest.

is there a particular allele more (less) present in the patient group?

Genotype	Cases	Controls
T/T	98	120
T/C	105	130
C/C	430	300



T = 2*T/T + T/C

C = 2*C/C + T/C

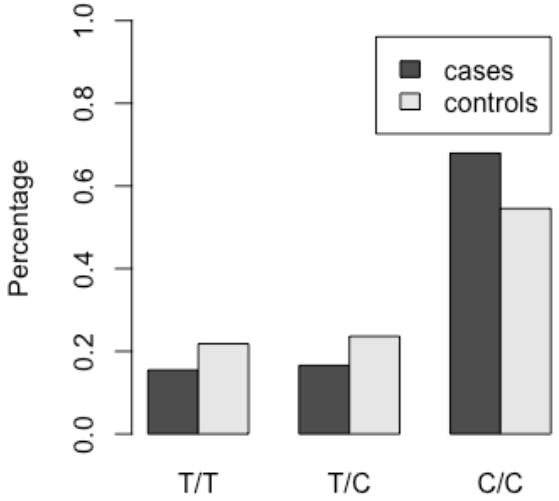
Allele	Cases	Controls
T	98*2+105	120*2+130
C	105+430*2	130+300*2

Statistical test for association

- Test for differences in allele is the most used and simplest.

is there a particular allele more (less) present in the patient group?

Genotype	Cases	Controls
T/T	98	120
T/C	105	130
C/C	430	300



T = 2*T/T + T/C

C = 2*C/C + T/C

Allele	Cases	Controlls
T	301	370
C	965	730

Pvalue and OR

- *Pvalue from a 2x2 Chisquare:* probability that the SNP is NOT related to disease

P value is 0.0001

- *Odds ratio:* quantifies the impact of the SNP on the disease

$$\text{OR (T)} = \frac{\text{Ratio (odd) between patients and controls with allele T}}{\text{Ratio (odd) between patients and controls with allele C}}$$

if **OR > 1, T increase risk** of disease
if **OR < 1, T decrease risk** of disease

OR (T)= 0.61

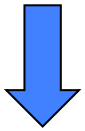
Candidate gene approach

- Investigates only genetic variants in genes that are potential candidates, for example:
  - for Diabetes, only in genes involved in glucose/insulin metabolism
  - for Multiple Sclerosis, only in genes expressed in brain and CNS

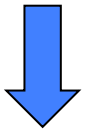
Sometimes refers to mouse homologous to gather information on genes

Workflow of candidate gene based association studies

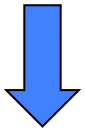
10 genes of interest



1,000 SNPs to test for association in the 10 genes



Perform statistical test for all 1,000 SNPs



Get the SNP(s) with the pvalue below your desired threshold (0.05, or 0.0005)

easy to apply, but with many limitations

- Observations in mice not always reflect human mechanisms
- Based on previous knowledge on genes and genetic variants
- Genes not previously studied or genetic variants not yet discovered, will never be assessed

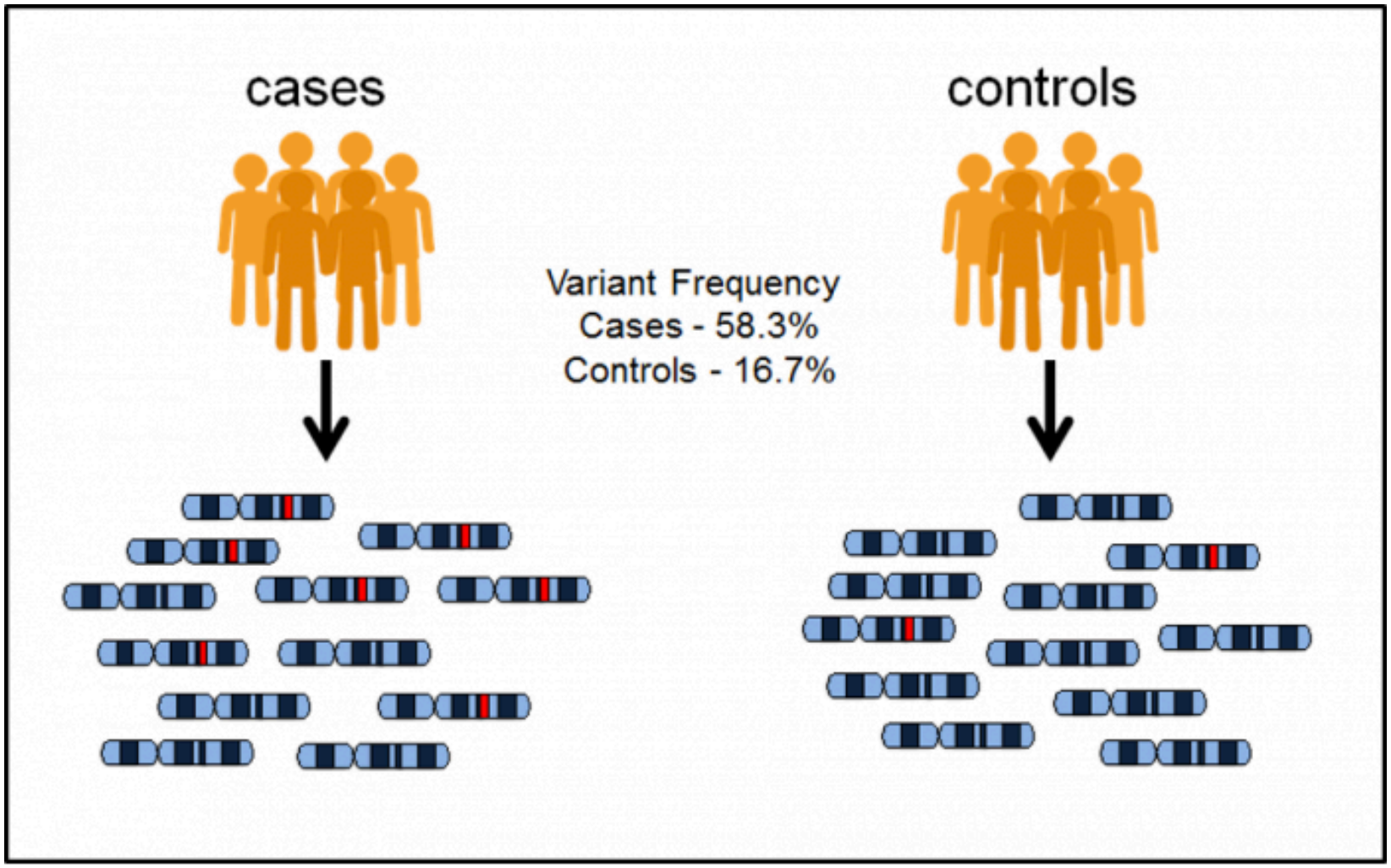
Genome-wide association studies – GWAS

Study all genome, as possible, without prior knowledge.

This includes studying genetic variation also outside genes

500,000 – 3 million SNPs

Genome wide association studies (GWAS) are hypothesis free methods to identify associations between genetic regions (loci) and traits (including diseases)



Linkage disequilibrium (LD)

“non-random association of alleles at different loci”

Low LD:

A	C
A	T
A	C
A	T
G	C
G	T
G	C
G	T

High LD:

A	C
A	C
A	C
A	C
G	T
G	T
G	T
G	T

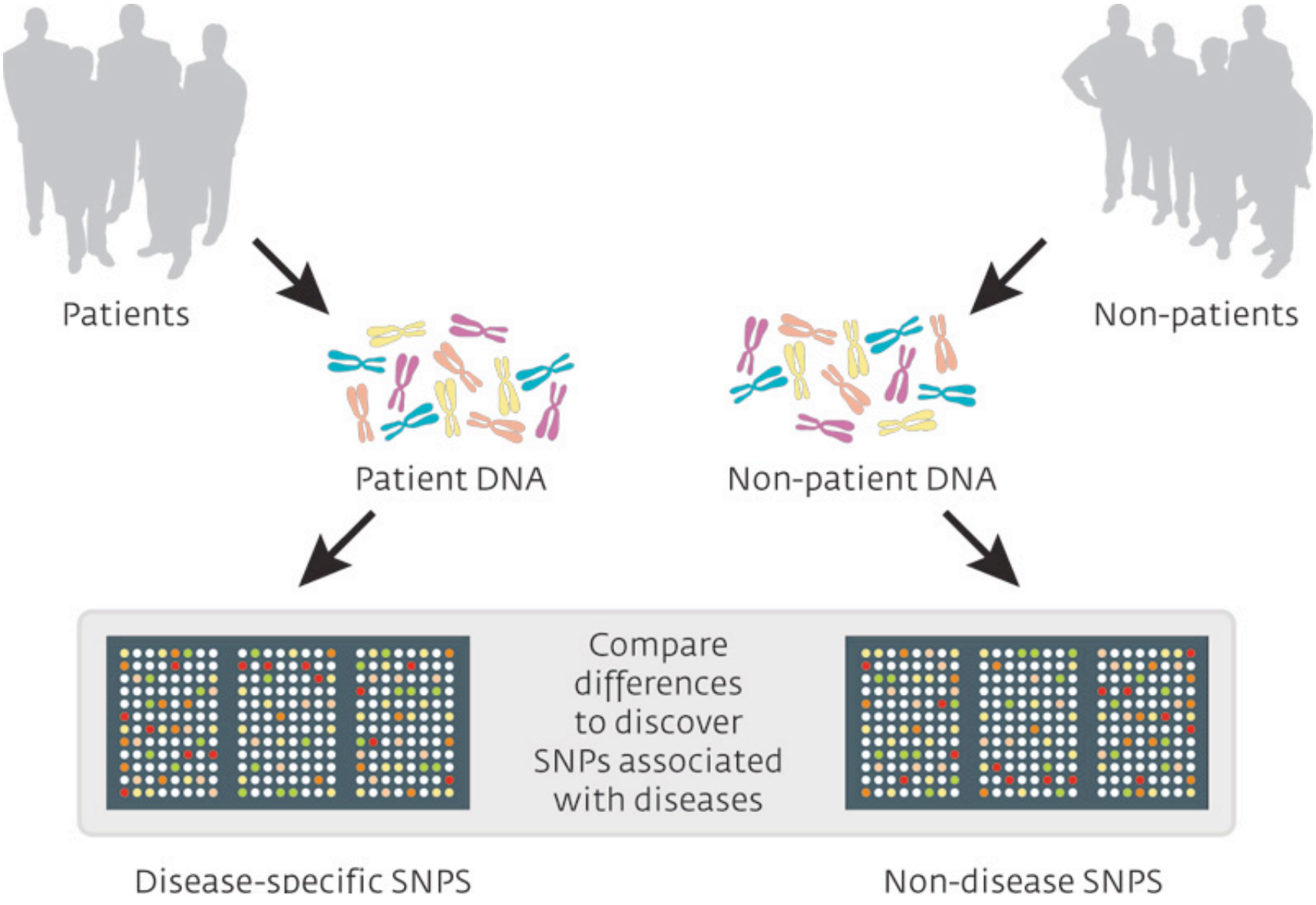
Why do we want to know about LD?

LD and association mapping

For association studies, the appropriate marker density depends on the background levels of LD

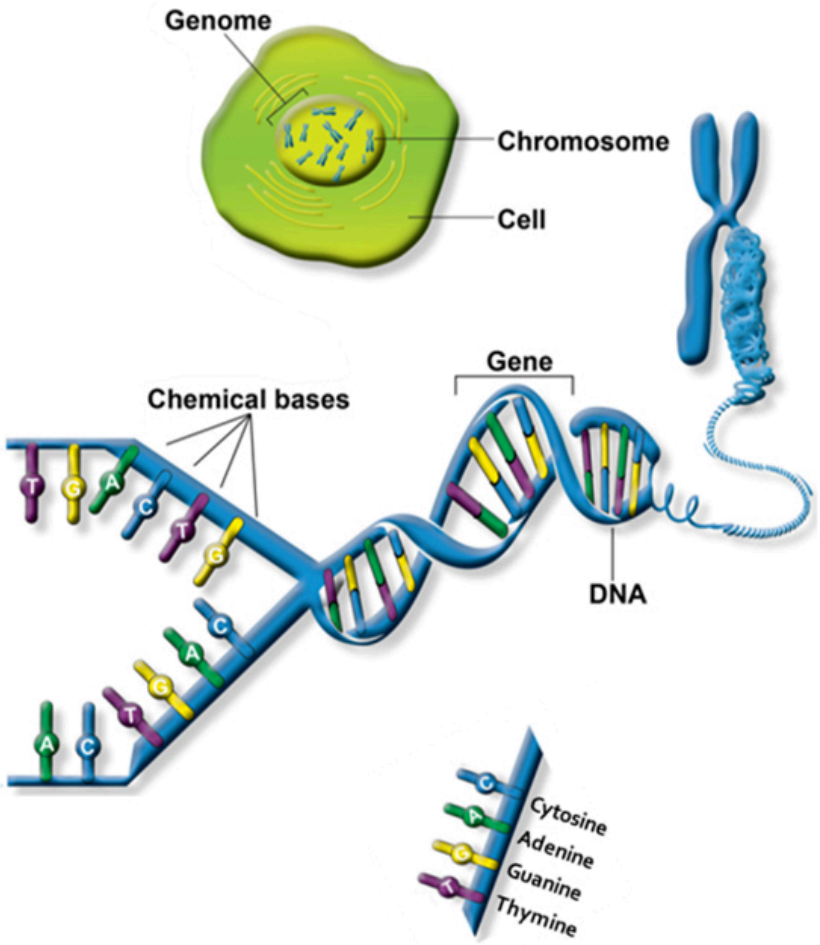
- Low LD populations need higher marker densities
- High LD populations require lower marker densities

Genome Wide Association Study (GWAS)

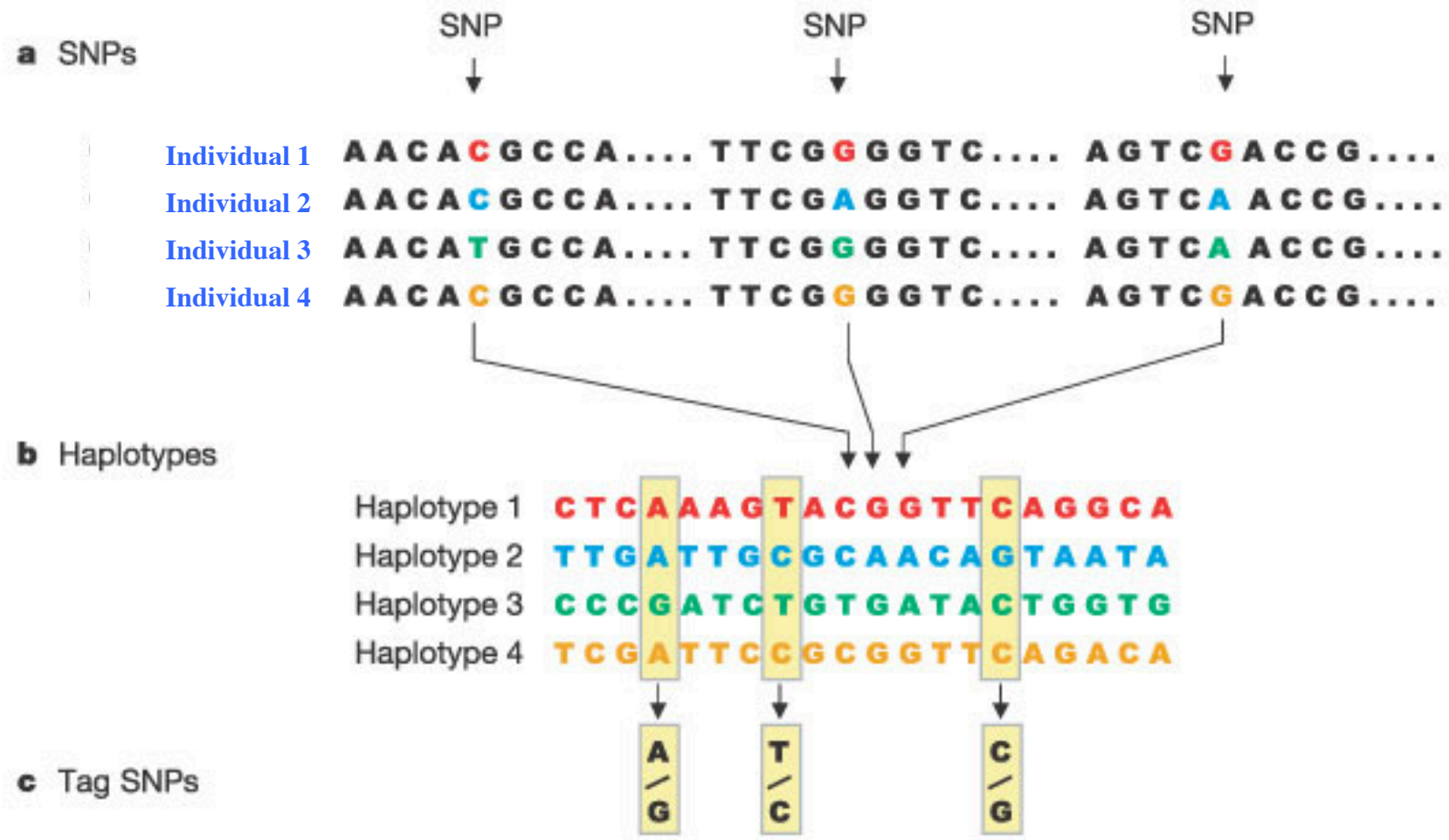


Genetic variants: SNPs have higher resolution to map genes for diseases

- SNP – single nucleotide polymorphisms
- Microsatellites
- Copy number variants



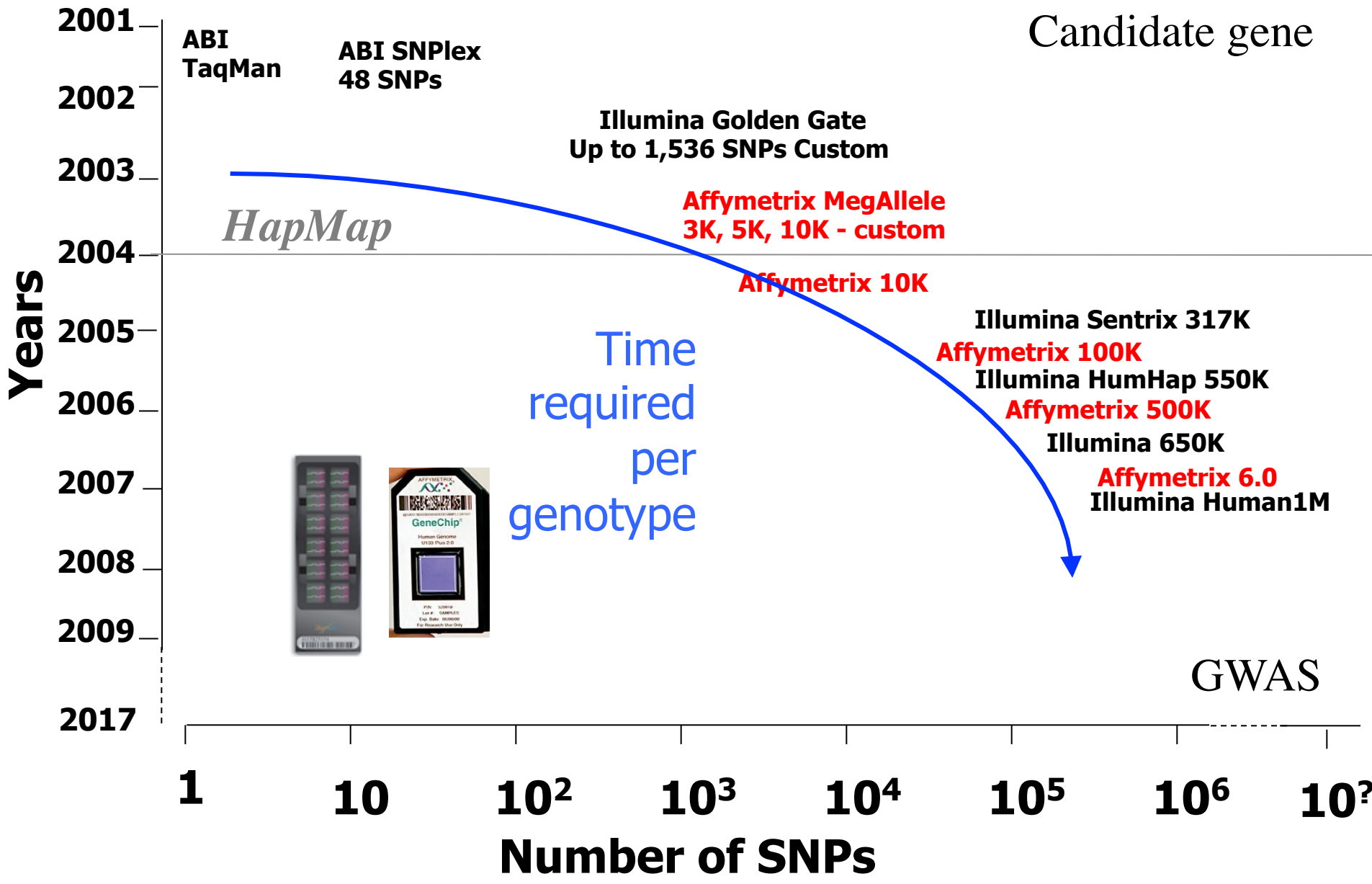
Concept of tag SNPs to minimize the number of SNPs to test



tag SNPs are reference SNPs that can serve as proxies of unknown and/or un-typed SNPs.

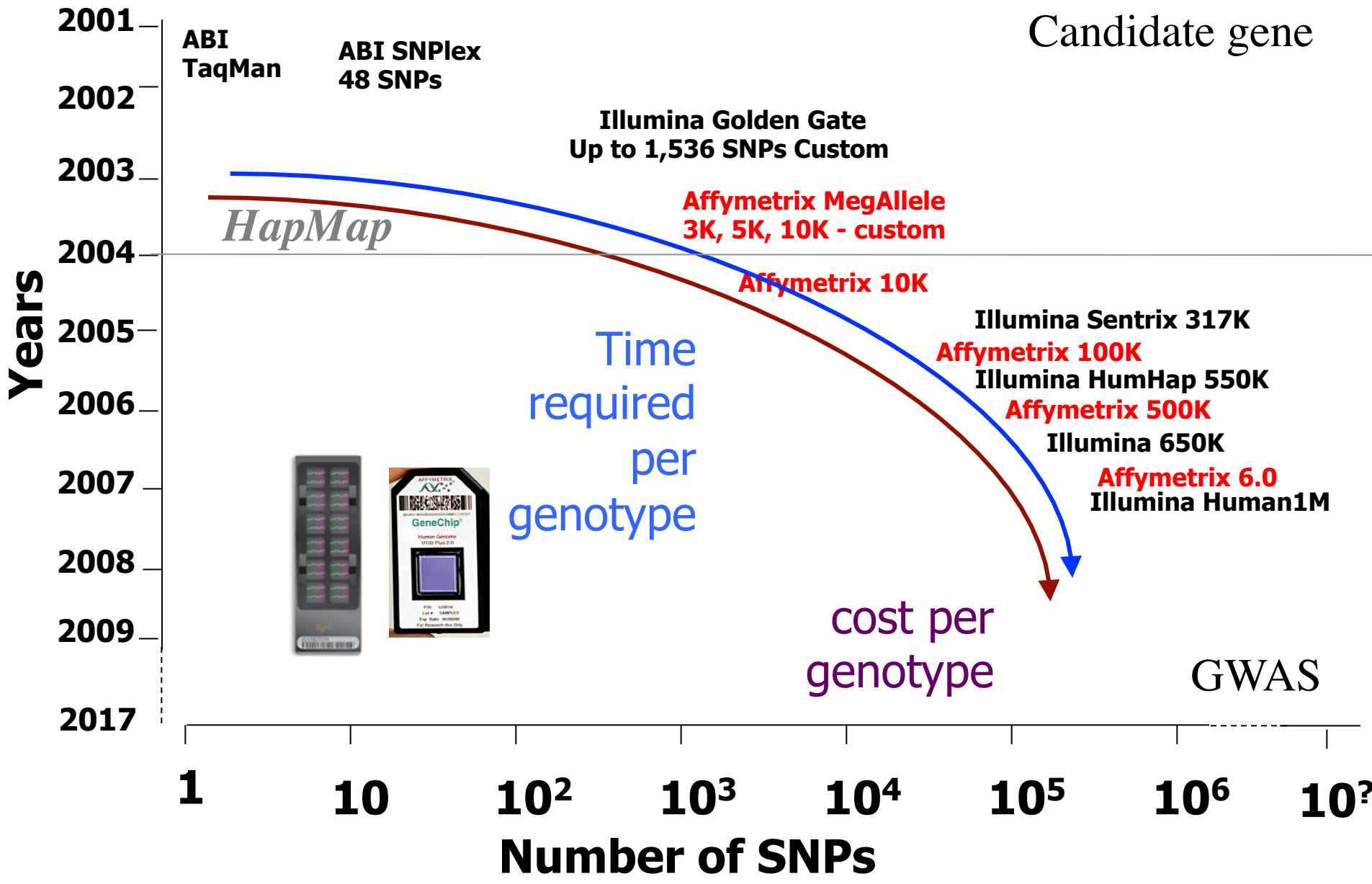
Ignoring redundant SNPs: selection of a subset of tag SNPs (from more than 3 million SNPs) that capture most of the un-typed common variation

Progress in Genotyping Technology



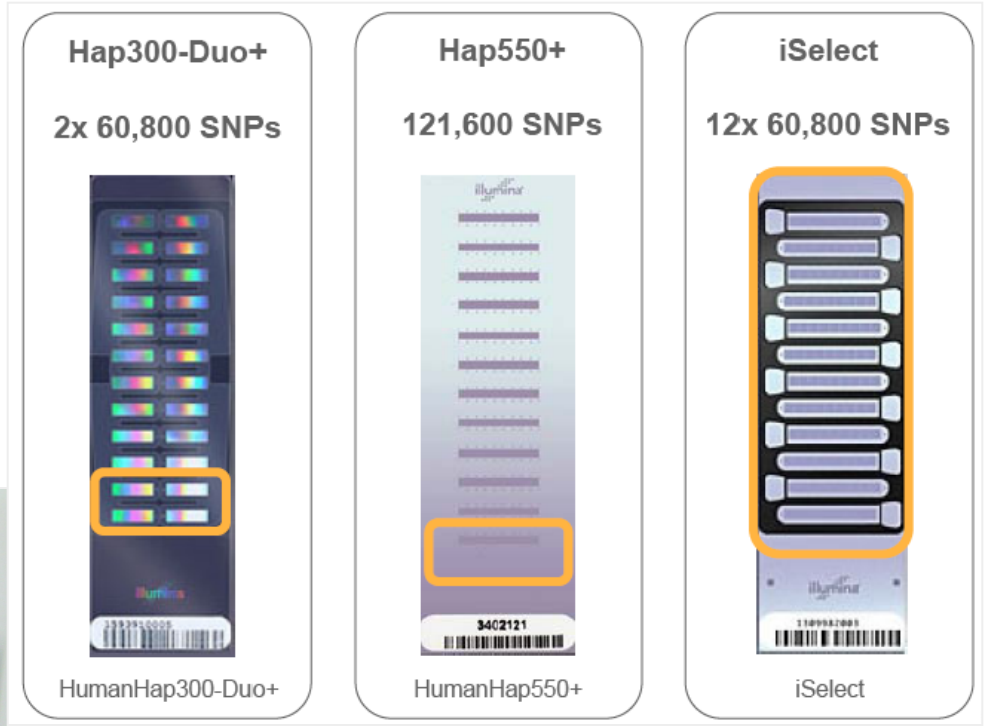
*data are not exhaustive

Progress in Genotyping Technology



*data are not exhaustive

High throughput technology



Genome-wide SNP arrays (e.g. affymetrix, Illumina) up till 1000K

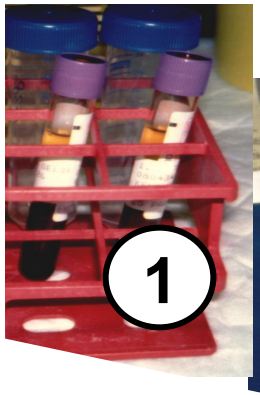
Illumina Human Quad670

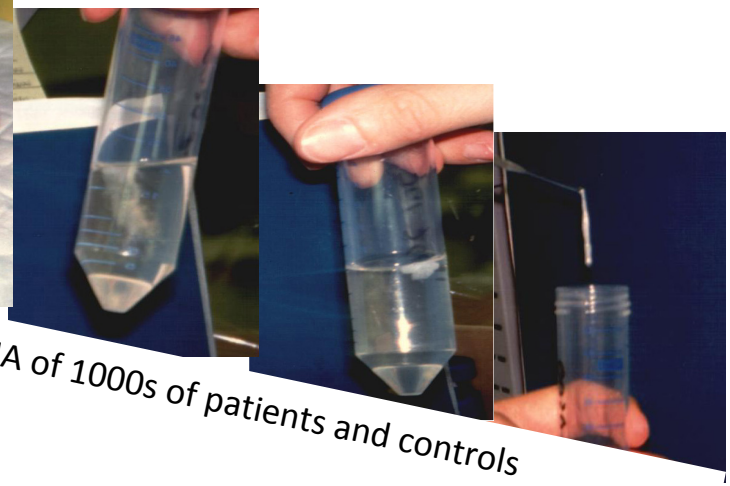
Each slide can be hybridized to DNA of 4 individuals
Each slide contains ~670,000 markers
Each slide costs ~ $ 740,- ($ 200, - p.p)

The price is reducing → $ 30, - p.p

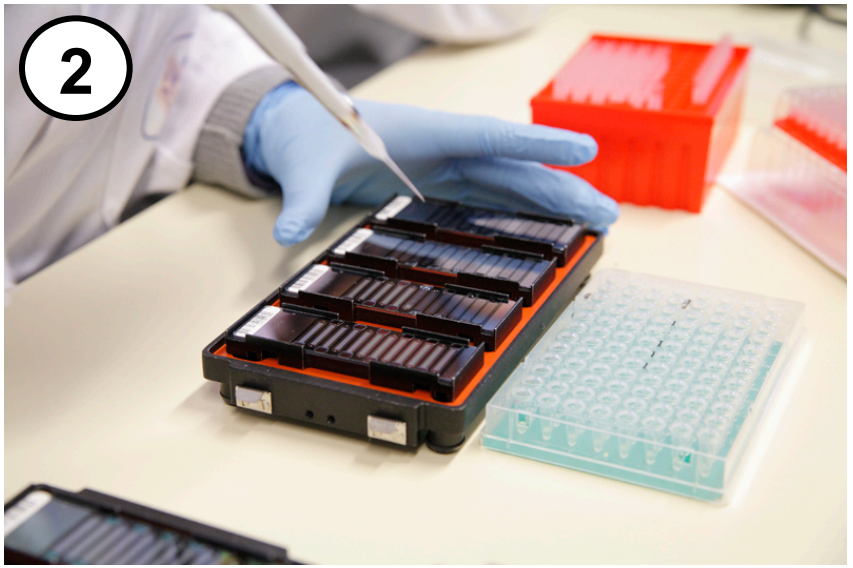
How does a GWAS work in practice?

How does a GWAS work in practice?





DNA of 1000s of patients and controls



Let's see real examples!

Aim of these examples

- Note the sample size used in each study
- Number of genes found
- Amount of variance explained

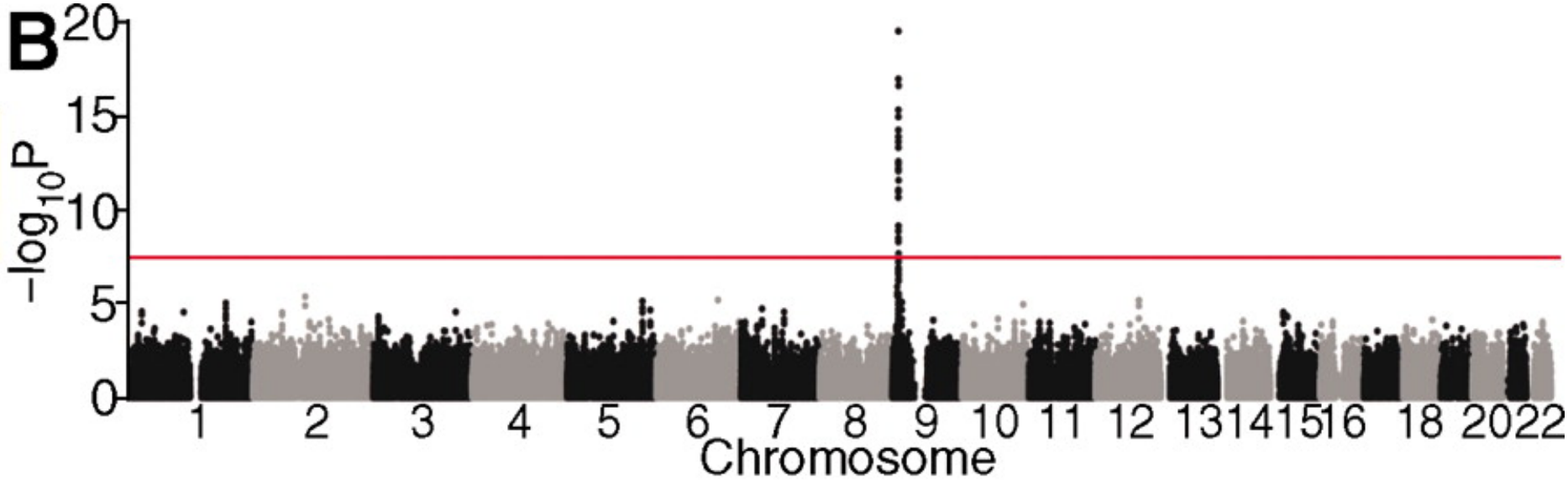
Gene for naturally blond hair: A GWA study using Solomon Islanders

Although individuals from the Solomon Islands and Equatorial Oceania have the darkest skin pigmentation outside of Africa, they also have the highest prevalence of **blond hair** (5 to 10%) outside of Europe

Source: Kenny et al. **Science,** 336:554(2012)

Gene for Naturally blond hair: A GWA study using Solomon Islanders

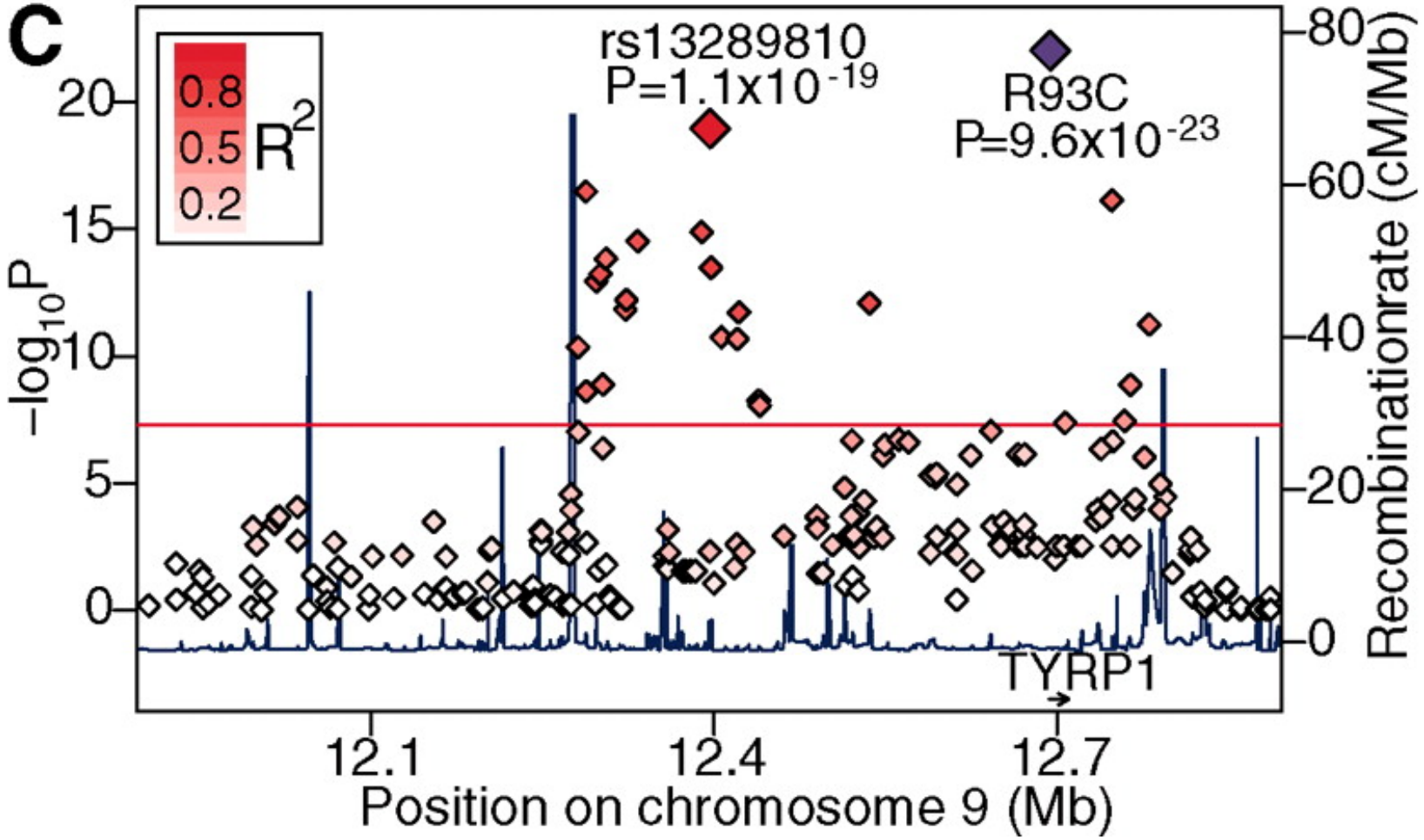
Case-control genome-wide association (GWA) study on
43 blond- and **42 dark-haired** Solomon Islanders



Source: Kenny et al. **Science,** 336:554(2012)

Gene for Naturally blond hair: A GWA study using Solomon Islanders

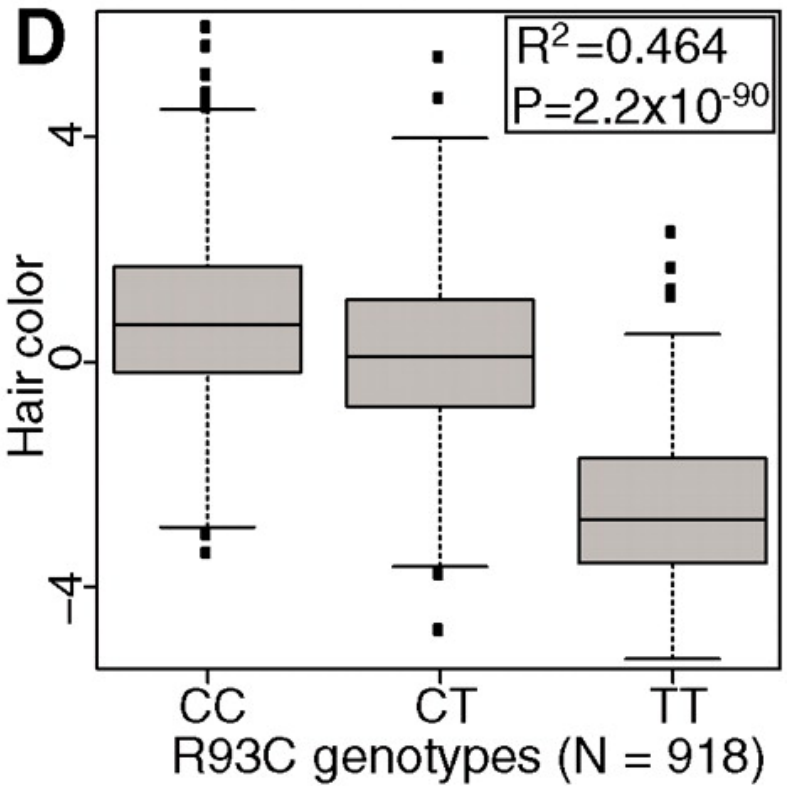
The mapping interval contained one known gene, tyrosinase-related protein 1 (*TYRP1*), which encodes a melanosomal enzyme involved in mammalian pigmentation



Source: Kenny et al. **Science,** 336:554(2012)

Gene for Naturally blond hair: A GWA study using Solomon Islanders

R93C was more strongly associated with blond hair ($P = 9.60 \times 10^{-23}$) than **the top GWA SNP and accounted for 46.4% of the variance** in hair color (linear regression; $P = 2.19 \times 10^{-90}$)



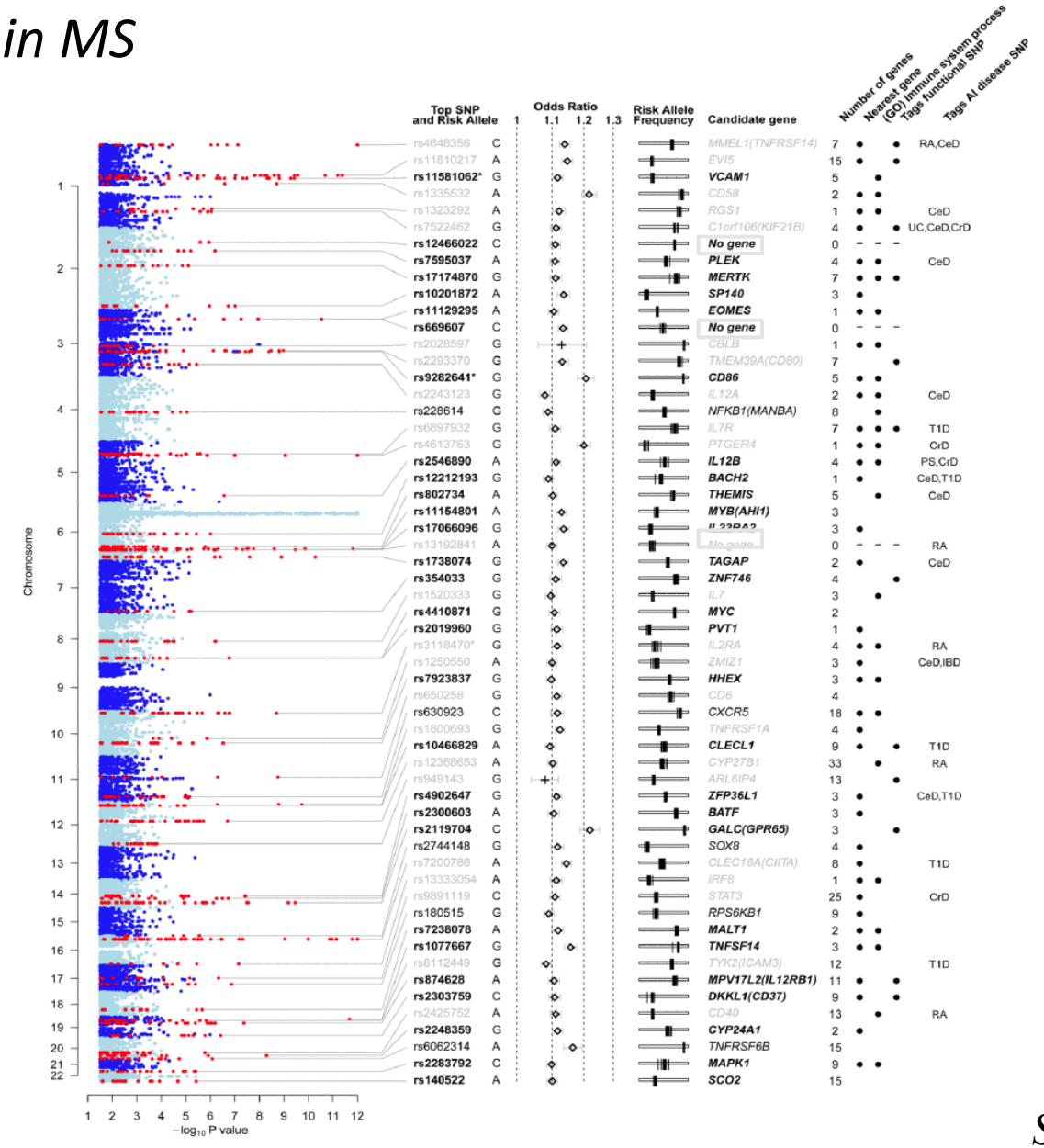
Source: Kenny et al. **Melanesian blond hair is caused by an amino acid change in TYRP1. Science,** 336:554(2012)

Let’s see a second example of a GWAS

GWAS of multiple sclerosis (MS)

- Disease of the central nervous system
- Gets its name from the buildup of scar tissue (sclerosis) in the brain and/or spinal cord
- The scar tissue forms when the insulating myelin covering the nerves is destroyed (demyelination)
- Different forms of MS

*Genetic risk and **a primary role for cell-mediated immune mechanisms** in MS*



10,000 patients
17,000 controls

52 loci (odd's ratio <1.5) involved, that together explain ~20% of the disease

Sawcer et al. Nature (2011) 476: 214

Aim of these examples

- Note the sample size used in each study
- Number of genes found
- Amount of variance explained

GWAS steps

The following steps are taken to perform a GWAS:

1. Quality control per SNP

1.1 Remove SNPs with minor allele frequency (MAF) < 0.01

1.2 Remove SNPs that deviate from Hardy Weinberg equilibrium (HWE) using a P threshold < 1e-6

1.3 Remove SNPs with missing genotype rate < 0.99

2. Quality control per individual

2.1 Identify and remove related individuals

2.2 Perform principal component analysis (PCA)and plot the first two components to identify and remove population outliers

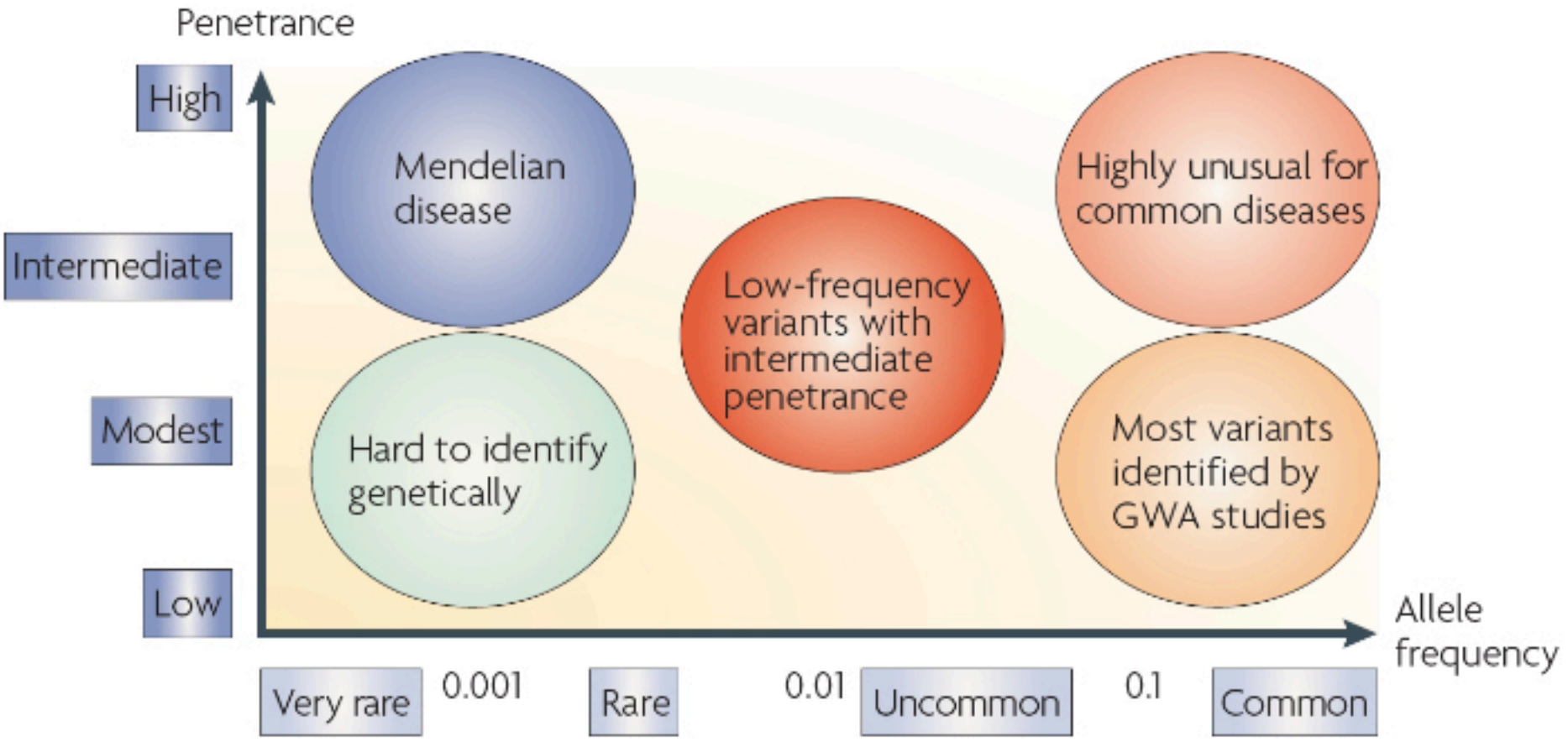
2.3 Perform genome-wide association testing

2.4 do initial association analysis, linear

2.5 do association analysis, logistic (Bonus)

2.6 prepare a manhattan plot

1.1 Remove SNPs with minor allele frequency (MAF) < 0.01



1.2 Remove SNPs that deviate from Hardy Weinberg equilibrium (HWE)

The Hardy-Weinberg equilibrium is a principle stating that the genetic variation in a population will remain constant from one generation to the next in the absence of disturbing factors.

The Hardy-Weinberg equation is expressed as:

$p2 + 2pq + q2 = 1$

p = frequency of the “A” allele
q= frequency of the “a” allele

In population genetics studies, the Hardy-Weinberg equation can be used to measure whether the observed genotype frequencies in a population differ from the frequencies predicted by the equation.

1.3 Remove SNPs with missing genotype rate < 0.99

To include only SNPs with a 99% genotyping rate (1% missing)

-

2. Quality control per individual

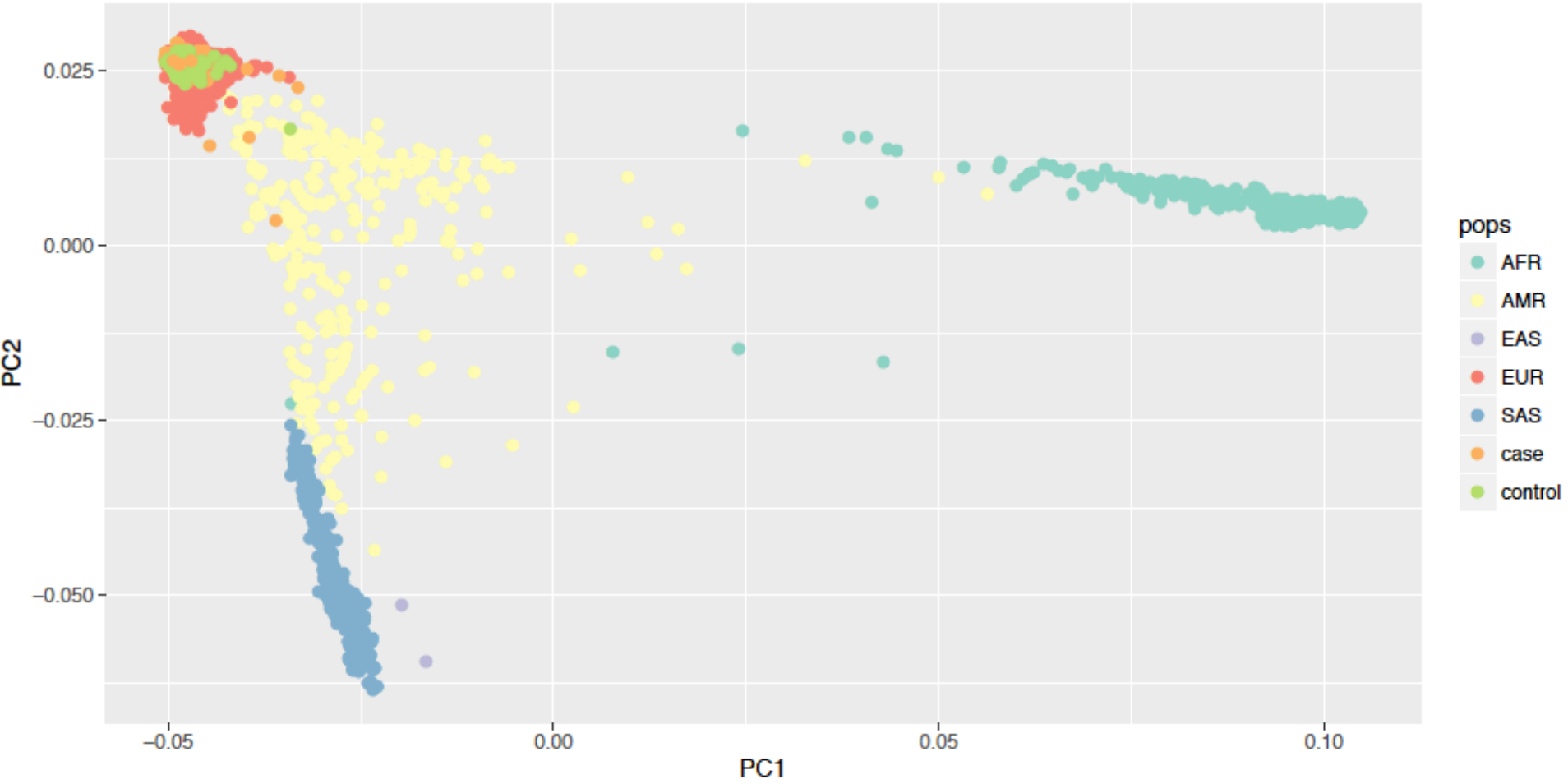
2.1 Identify and remove related individuals

To identify related individuals, we make a genetic relationship matrix. This matrix identifies how much related the individuals are by comparing all genotypes of the individuals. The more similar are two individuals on the DNA level, the more likely it is to be related.

2.2 Perform principal component analysis (PCA)and plot the first two components to identify and remove population outliers

There are often known differences in phenotype prevalence due to ethnicity, and allele frequencies are highly variable across human subpopulations, meaning that in a sample with multiple ethnicities, ethnic specific SNPs will likely be associated to the trait due to population stratification.

2.2 Perform principal component analysis (PCA)and plot the first two components to identify and remove population outliers

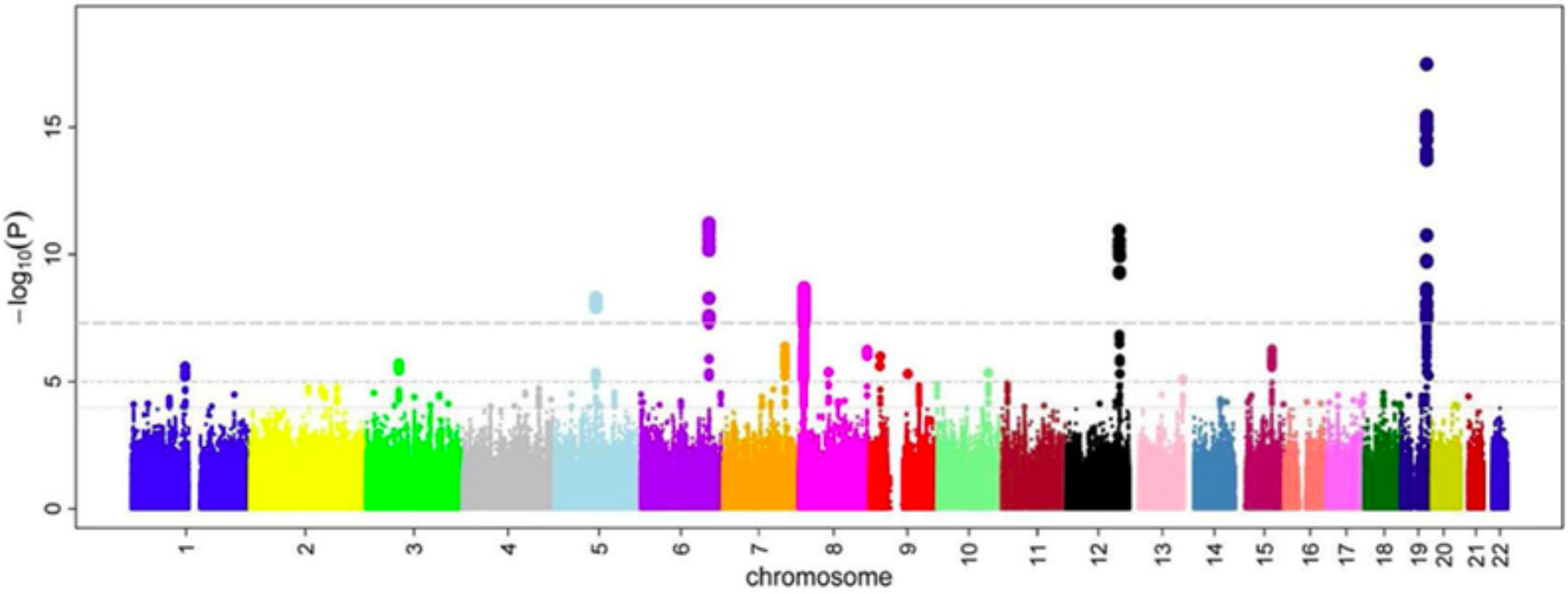


How to asses significance of genetic association statistically?

- The frequency of a SNP marker is collected in a group of cases and a group of controls
- A 2 x 2 Contingency Table is constructed and the Chi Square statistics is calculated to test for deviations of the observed frequencies from expected frequencies
- The chi square value corresponds to a p-value;
  - $p < 0.05$ is regarded significant (but needs correction for multiple testing)
  - For GWA a p value $< 5x10^{-8}$ is considered significant

Manhattan plots?

GWAS result: plot for all ~500,000 SNPs their P values across the genome = Manhattan plot



Genome-wide significant p value: 5×10^{-8}

P value corrects for the multiple testing of ~ 500,000 SNPs

Systems genetics

Genetic variation

Transcriptome:
All messenger RNA molecules ('transcripts')

Proteome:
All proteins in cell or organism

Metabolome:
all metabolites in a biological organism

Microbiome:
human gut, oral, skin microbiota