Can we predict the phenotype of an individual from DNA?

Edith Le Floch
DRF/ Institut de biologie François Jacob/
Centre National de Recherche en Génomique Humaine

How to predict Alzheimer's Disease from DNA?

Alzheimer's Disease:

- A neurodegenerative disease associated with cognitive disorders and memory loss
- Prevalence: almost 20% in people over 80

Some genetic origins:

- Common form caused at ~75% by genetic factors
- But the known causal genes account only for 8% (main gene APOE accounts for 6%)

How to predict Alzheimer's Disease from DNA?

Alzheimer's Disease:

- A neurodegenerative disease associated with cognitive disorders and memory loss
- Prevalence: almost 20% in people over 80

Some genetic origins:

- Common form caused at ~75% by genetic factors
- But the known causal genes account only for 8% (main gene APOE accounts for 6%)

Alzheimer's Disease Neuroimaging Initiative (ADNI)

- **Clinical information on 809 individuals:**
  - 188 patients with Alzheimer's Disease (AD)
  - 393 patients with Mild Cognitive Impairment (MCI)
  - 228 controls

- **Genetic Data** available

- **Brain imaging data** (MRI) also available

How to predict the patient/control status from DNA?

Outline

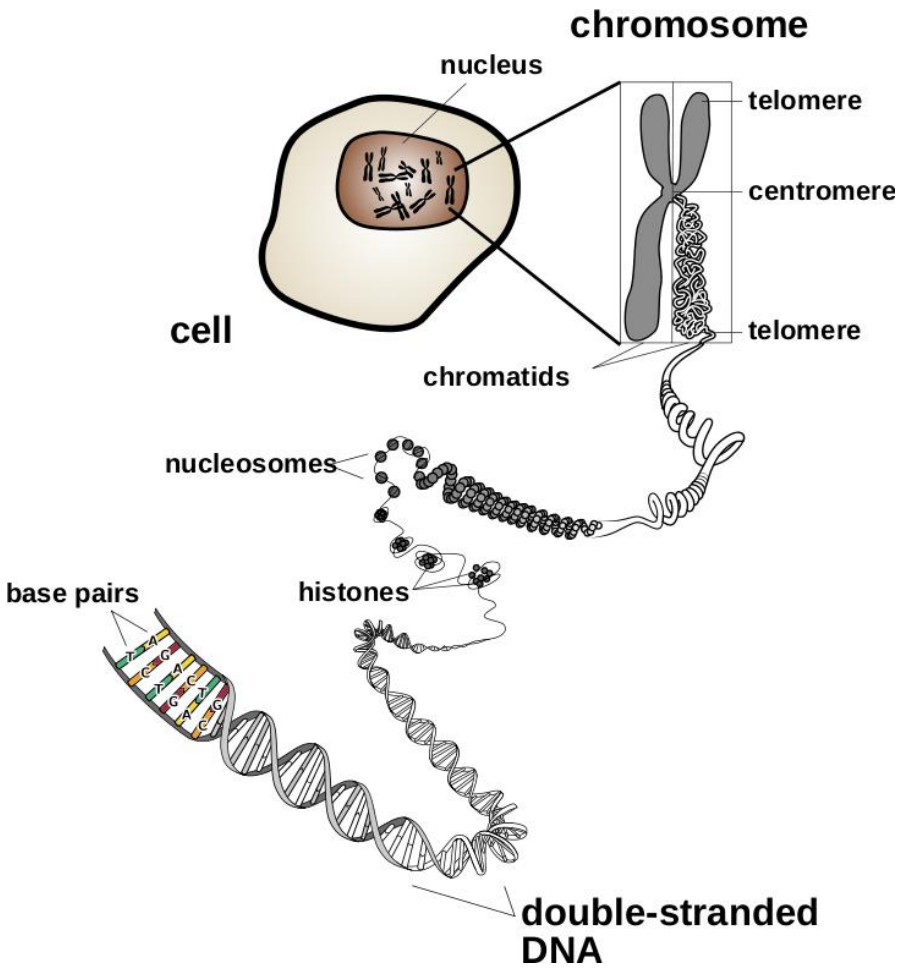
Part 1 : A few notions in Genetics

Part 2: The univariate approach
2.1 Genotyping data
2.2 Sequencing data

Part 3: The multivariate approach (machine learning)
3.1 Sequencing data
3.2 Genotyping data

Part 1:
A few notions in Genetics

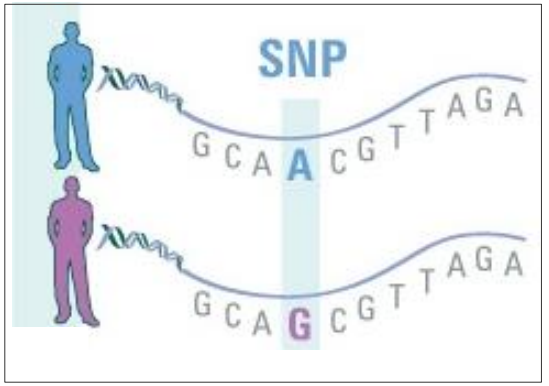
Human genome



- **22 pairs** of homologous chromosomes + X Y
- **2 identical chromatids** per chromosome
- **2 complementary strands** per chromatid
- Each strand: **sequence of nuceotides** (Adenine, Thymine, Cytosine and Guanine)
- **3 billion** base pairs
- About 2% of DNA coding for proteins: **25 000 genes**

Single Nucleotide Polymorphisms (SNPs) 1/2

- **SNP**: position on the genome where a single nucleotide varies in the population (>1% of individuals)
→ due to an ancestral mutation



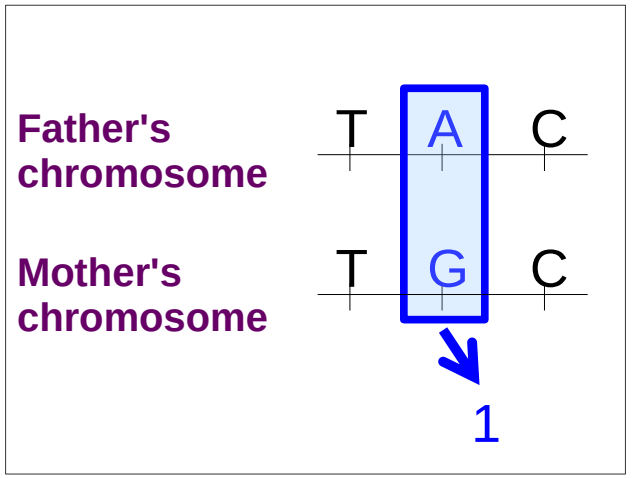
- **Main form of DNA variability** in the population (about **30 million** SNPs)

- About **3-4 million single nucleotide differences** between 2 individuals

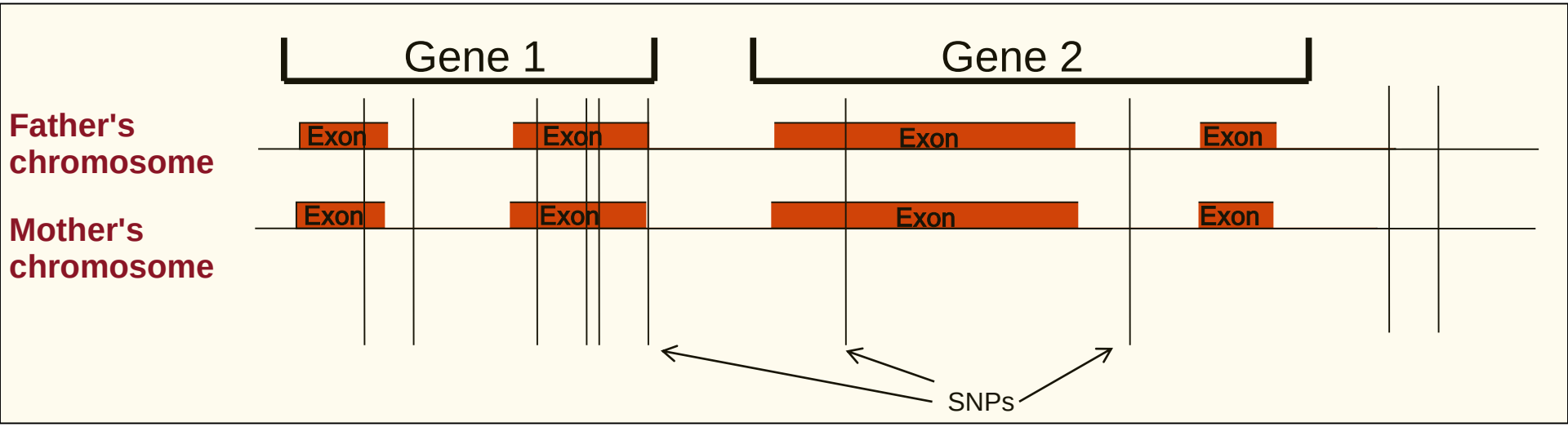
- Usually only **two possible alleles for a SNP:** one major (e.g. A) and one minor (e.g. G)

- The **genotype** of an individual is defined by considering the pair of homologous chromosomes: 3 possibilities (e.g. AA, AG or GG)

→ often coded as the **number of minor alleles**: 0 (AA), 1 (AG) or 2 (GG)



Single Nucleotide Polymorphisms (SNPs) 2/2



SNPs may be located :

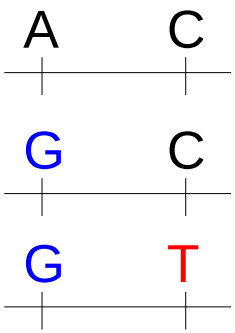
- **inside a gene**:
  - in an **exon** (coding for the protein) : synonymous or not
  - in an **intron** (non-coding)

- **outside a gene**

CNRGH
CENTRE NATIONAL DE RECHERCHE EN GÉNOMIQUE HUMAINE

Linkage Disequilibrium (LD) 1/3

. **LD: non-random association of alleles** between two SNPs
→ often due to physical **linkage** (*ie* SNPs on the same chromosome)



→ The 2 SNPs transmitted together trough generations

• **Recombination** between homologous chromosomes during meiosis

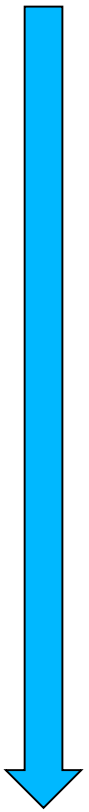
→ Probability of recombination increases (and thus LD decreases) with the distance between the 2 SNPs

Linkage Disequilibrium (LD) 2/3

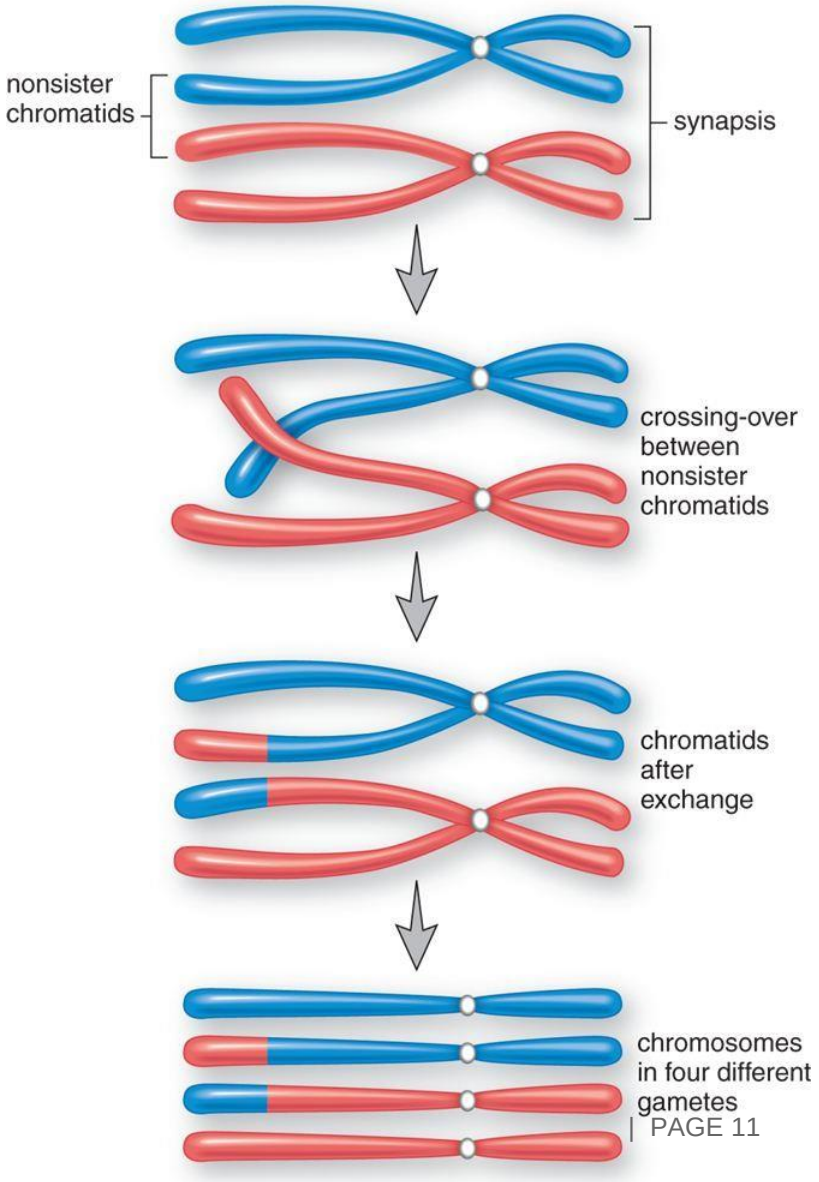
Meiosis : formation of

reproductive cells (gametes)

1 cell



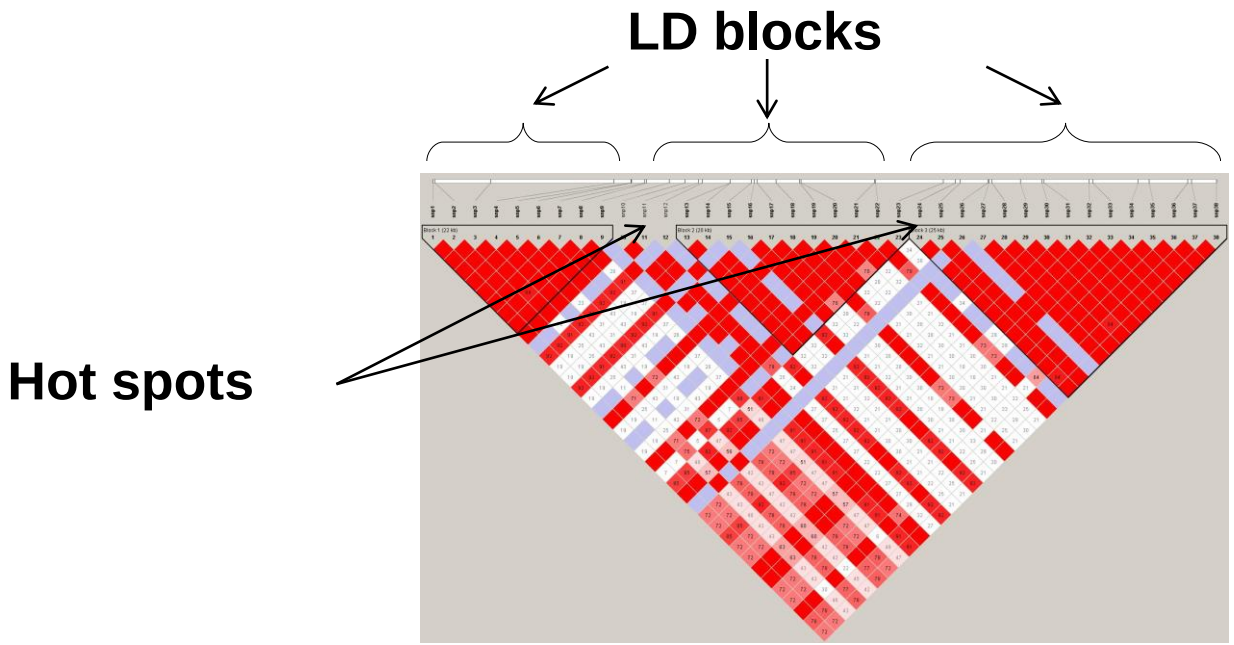
4 cells



Linkage Disequilibrium (LD) 3/3

- **Non-homogeneous recombination** between homologous chromosomes during meiosis: **hot spots** of recombination

→ **LD blocks**



Genetic diseases

Monogenic (Mendelian) diseases:

- Caused by one single gene
- High effect (often lethal)
- Rare mutations (due to genetic selection)

Polygenic (complex) disease:

- Caused by several genes (not the same in every patient)
- Moderate effect
- Common polymorphisms

→ The common type of Alzheimer's disease in a complex disease

Genotyping data

- **Not the whole genome sequence** is observed
- **Only some of the known SNPs** all over the genome
- **1 million** SNPs on common chips today
- **Mainly common SNPs** (>5% in the population)
- Enough to capture **most of the common genetic variability** and to guess all other SNPs by knowing LD

Whole Genome Sequencing data

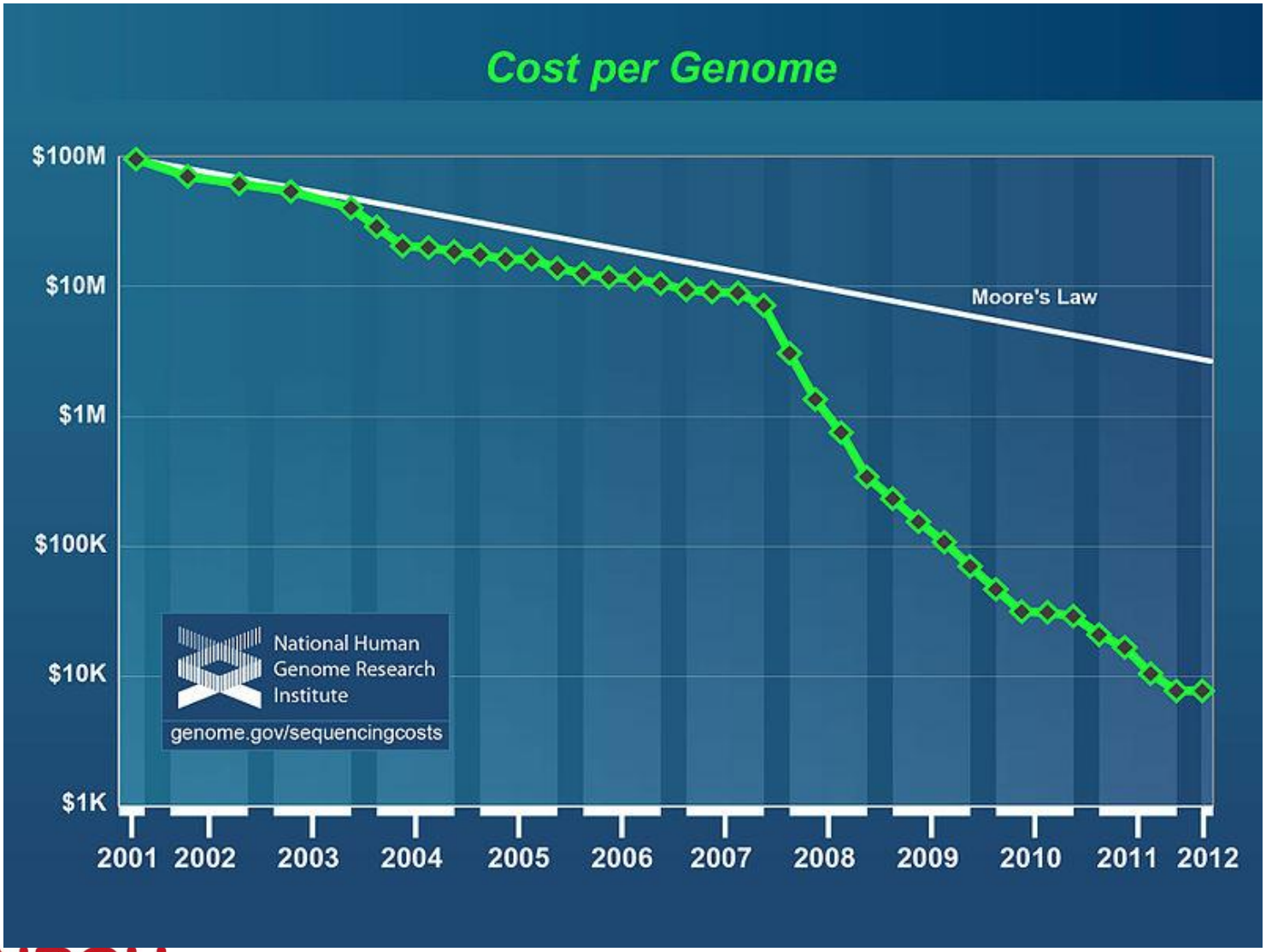
- **The whole genome is sequenced** (3 billion bases) for each individual

- Chromosomes **not sequenced in one piece:**
  Short "reads" of about 100 nucleotides are sequenced

- **Bioinformatics tools needed** to reconstruct the whole sequence:
  - Each read is aligned on a reference sequence
  - Variations from the reference are identified (e.g. SNPs, SNVs)
  - Only variations from the reference are stored in the final file (3-4 million SNPs/SNVs per individual)

- Each nucleotide is **sequenced about 30 times** to avoid errors

Cost of Whole Genome Sequencing



Part 2:
The univariate approach

Part 2.1:
The univariate approach on ADNI genotyping data

Alzheimer's Disease Neuroimaging Initiative Genotyping data

- 809 individuals:
  - 188 Alzheimer's Disease (AD)
  - 393 Mild Cognitive Impairment (MCI)
  - 228 controls

- The 809 individuals were genotyped with SNP array with 2.5 million SNPs

Classical statistical analysis of genotyping data

Univariate approach:

- Test for the association of each SNP (with the phenotype) **independently**
- If the phenotype is **disease (case) / not disease (control):**
  Is the distribution of the 3 genotypes the same for cases and controls?

For example for a SNP with two possible alleles A and T:

	Cases	Controls
AA	20	10
AT	40	66
TT	75	163

→ p-value of a **Chi-square test** = 0.007 but many tests (= nb SNPs) !!

→ Need to correct for **multiple comparisons**

Different univariate tests for case/control study

Genotypic test:

- The most general
- Not very powerful on average to detect moderate associations (higher nb of degrees of freedom)

	Cases	Controls
AA	20	10
AT	40	66
TT	75	163

Allelic test:

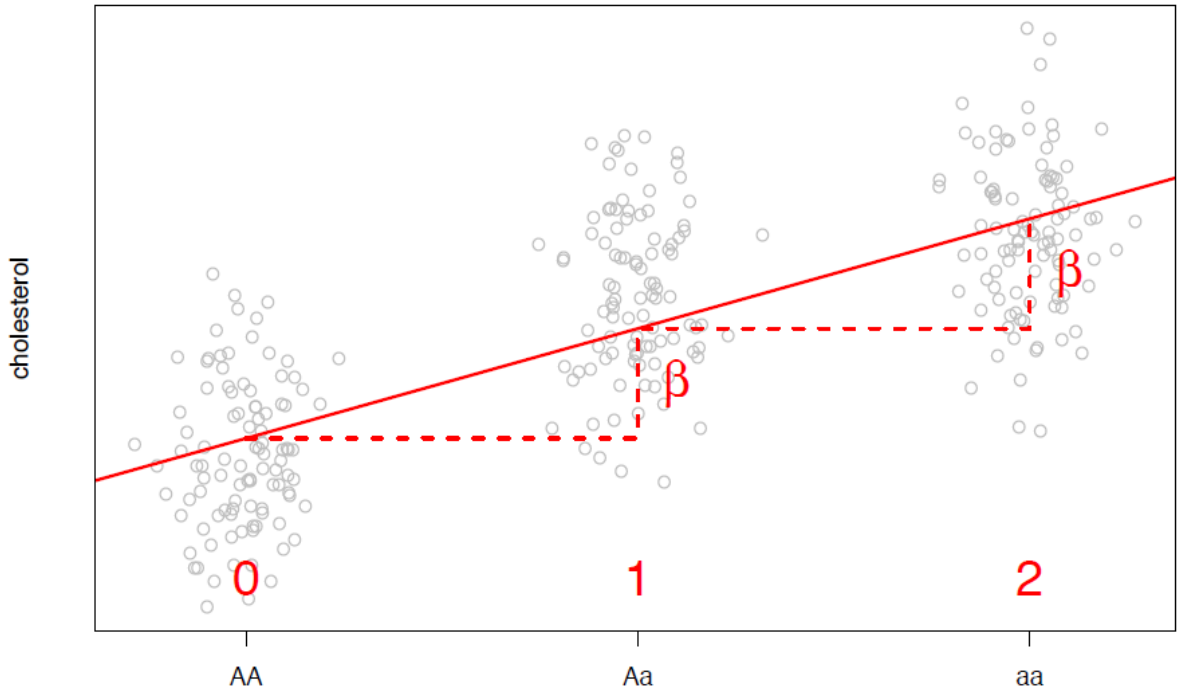
- Assumes the 2 alleles of an individual are independent (Hardy-Weinberg)
- Assumes additive effects of alleles
- Powerful in most cases

	Cases	Controls
Minor allele A	20*2+40=80	10*2+66=86
Major allele T	75*2*40=190	163*2+66=392

Univariate tests for a quantitative phenotype

- If the phenotype (y) is quantitative → **simple linear regression**
- Like for case/control studies, an **additive model** is usually used:

$$y = \beta_0 + \beta \times \#\text{minor alleles}$$

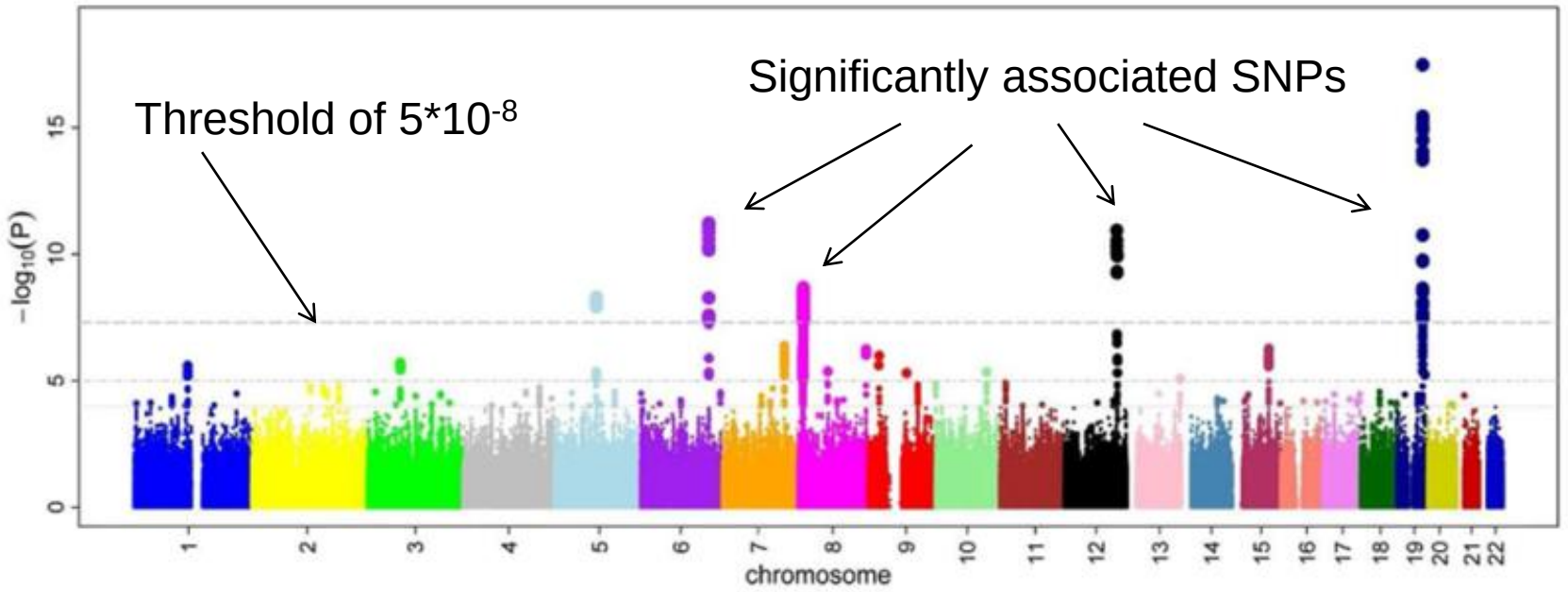


β ≠ 0 ?
(T-test)

From lectures by Timothy Thornton

P-value correction in Genome-Wide Association Studies (GWAS)

"Manhattan plot" of the p-values of the SNPs along the genome:



- **Bonferroni correction** commonly used to correct for multiple tests:

1 million SNPs on common chips
→ genome-wide significance threshold: $5*10^{-2} / 10^{6} = 5*10^{-8}$

CNRGH
CENTRE NATIONAL DE RECHERCHE EN GÉNOMIQUE HUMAINE

Power of GWAS

- The test will be more powerful to detect an association:

  - with **high sample size** (often 10s of 1000s of individuals)

  - with **frequent polymorphisms**

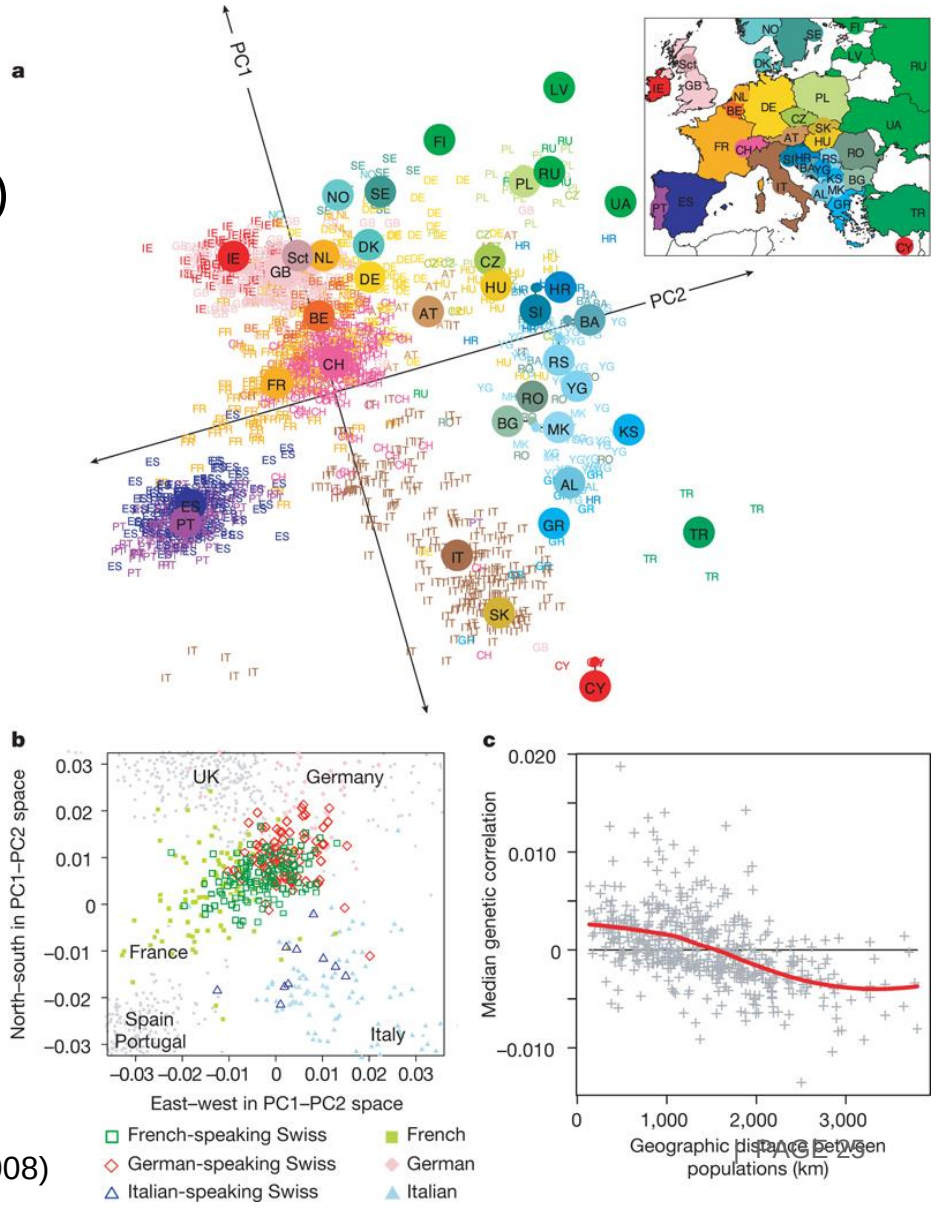
  - with **strong effects**

Population structure within Europe

Principal component analysis on 197,146 SNPs (coded 0, 1 or 2) in 1387 individuals

→ when plotting the two first principal components, **the map of Europe appears!**

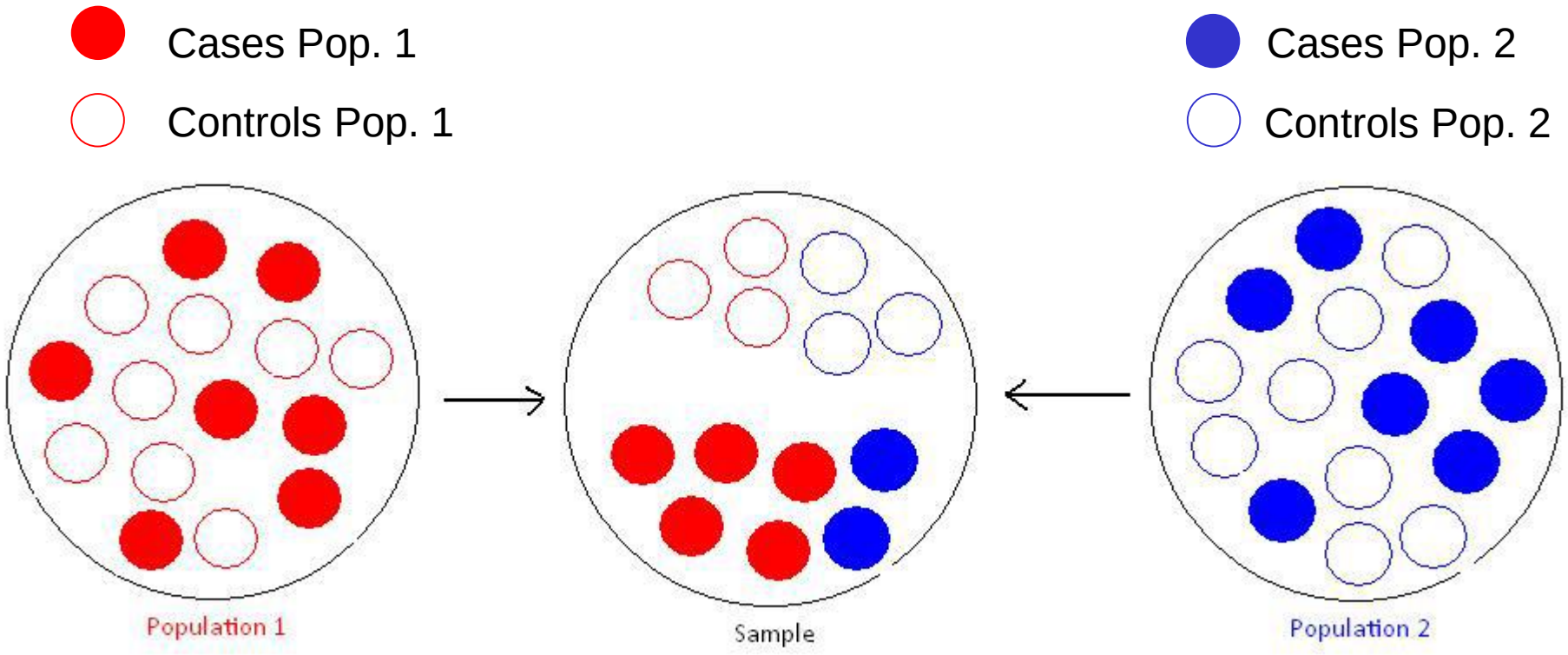
→ even possible to distinguish between :
- French-speaking Swiss
- German-speaking Swiss
- Italian-speaking Swiss



J Novembre *et al. Nature*, 1-4 (2008)

Problems with population structure in GWAS 1/2

- **Sampling bias for a case/control study:**
  If the % of each population different in cases and controls

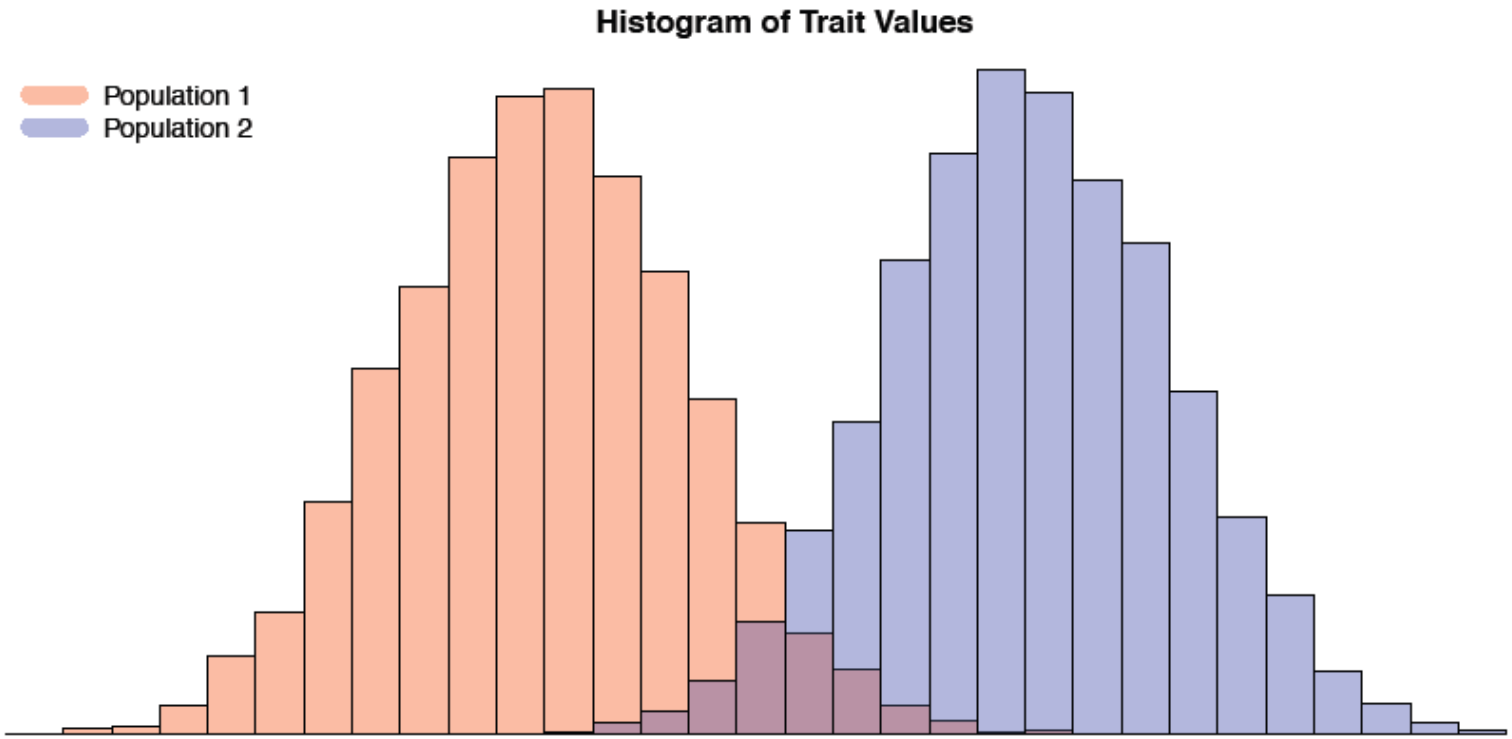


→ Alleles specific to Pop. 1 **artificially associated** with the disease!

From lectures by Timothy Thornton

Problems with population structure in GWAS 2/2

- **For a quantitative trait:**
  If the 2 populations have different means (due to sampling bias, to different lifestyles)



→ Alleles specific to 1 population **artificially associated** with the trait!

From lectures by Timothy Thornton

GWAS results of ADNI genotyping data

- Comparison of Alzheimer patients (188) versus others (621) on the 2.5 million SNPs

- **3 variants with a significant p-value** after Bonferroni correction ($p<2*10^{-8}$) with χ^2 test or Fisher exact test
  - 1 in APOE intron and in regulatory region ($p=1.4*10^{-12}$)
  - 2 in intergenic regions near APOE ($p=3*10^{-13}$ and $p=6*10^{-14}$)

- **Associated variants are frequent** (Minor Allele Freq. 20-40%)

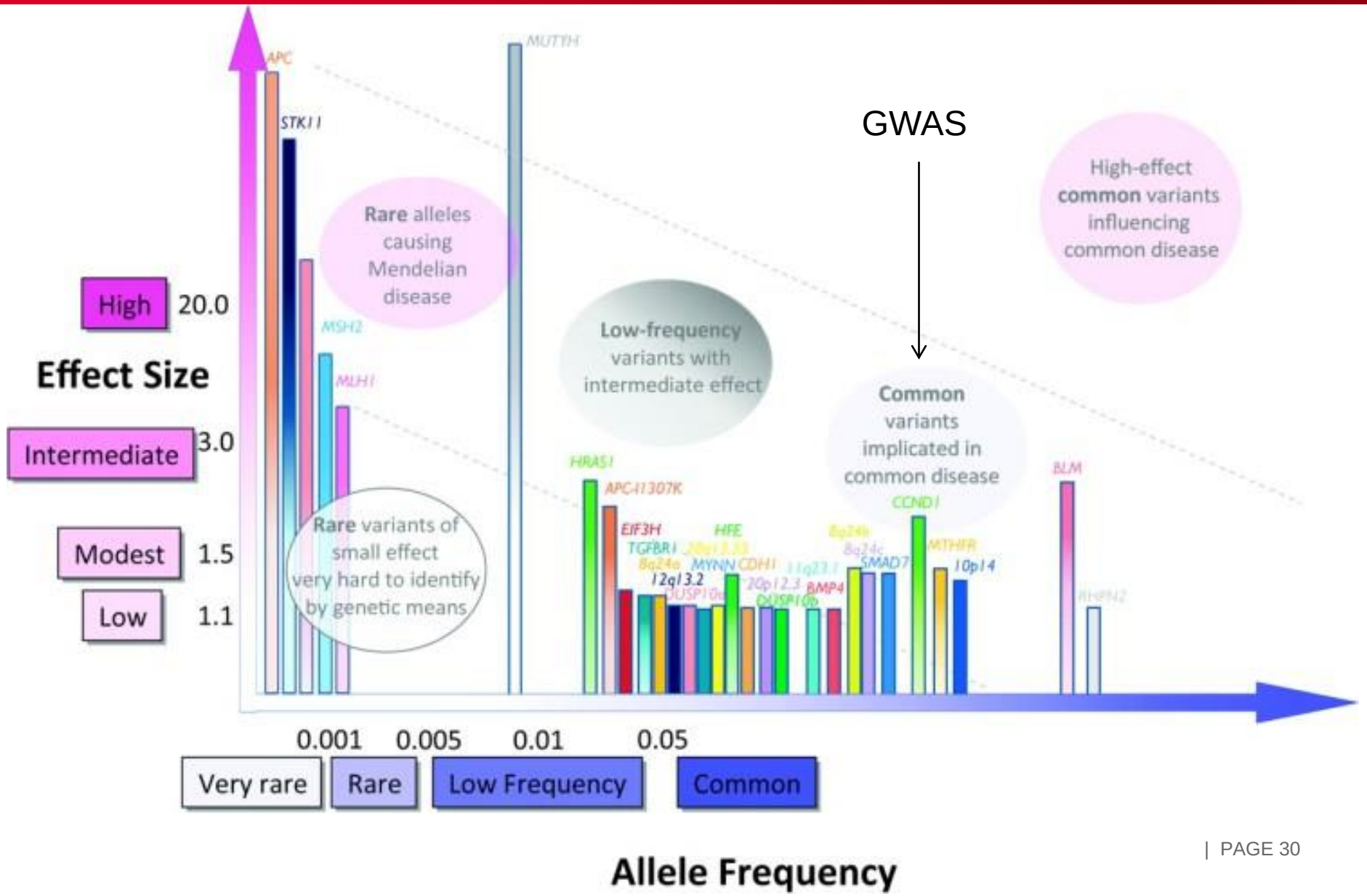
Global results of GWAS

- GWAS on genotyping data **have identified many SNPs** (14000) significantly associated with more than 1500 phenotypes

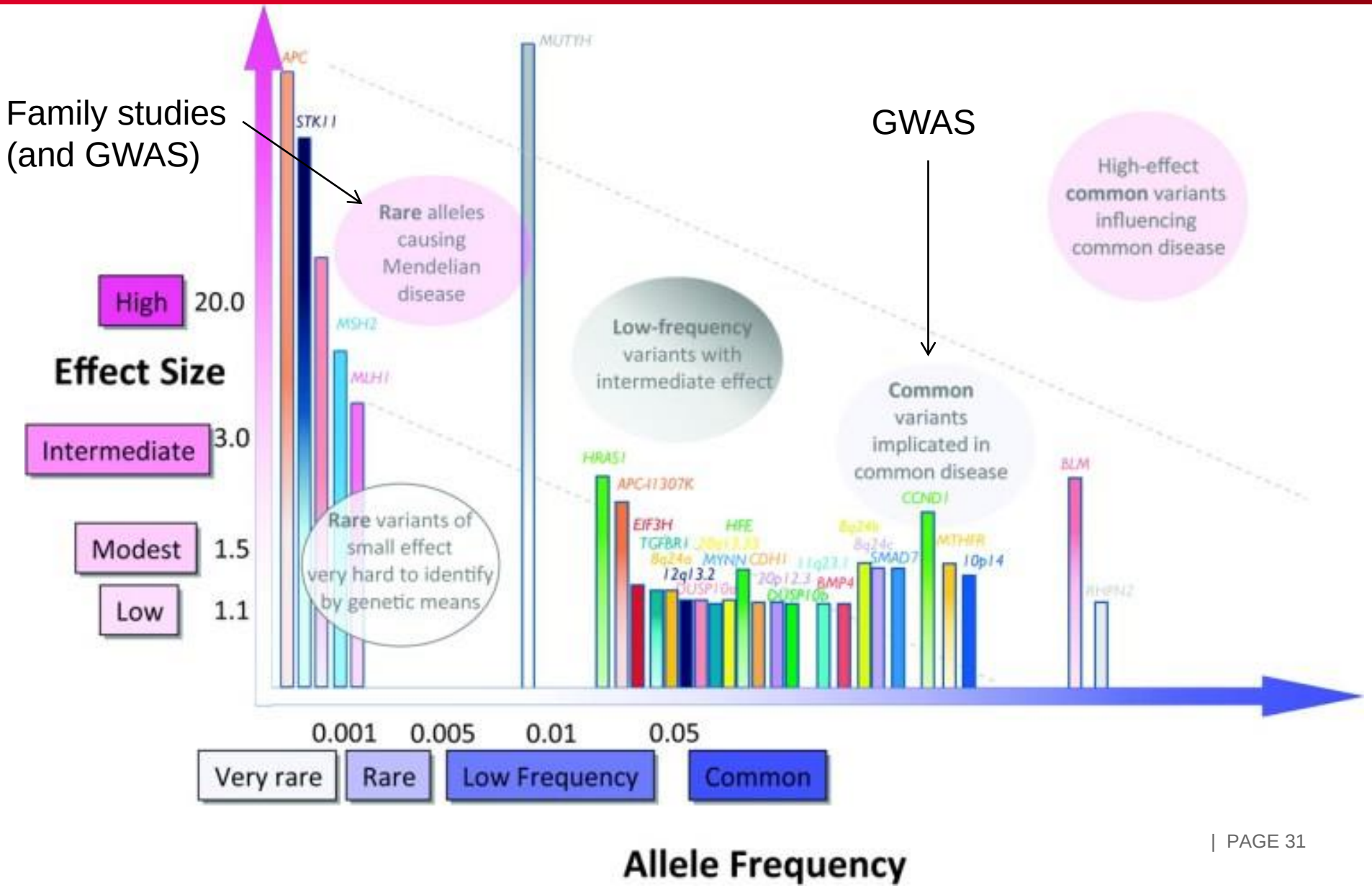
- But they only explain **a small portion of the phenotypic variance** (8 % for Alzheimer's disease instead of 75%!)
  → **Missing heritability**

- **Many possible reasons** for missing heritability:
  - rare variants
  - interaction effects between variants
  - many small effects that cannot be detected with current sample sizes

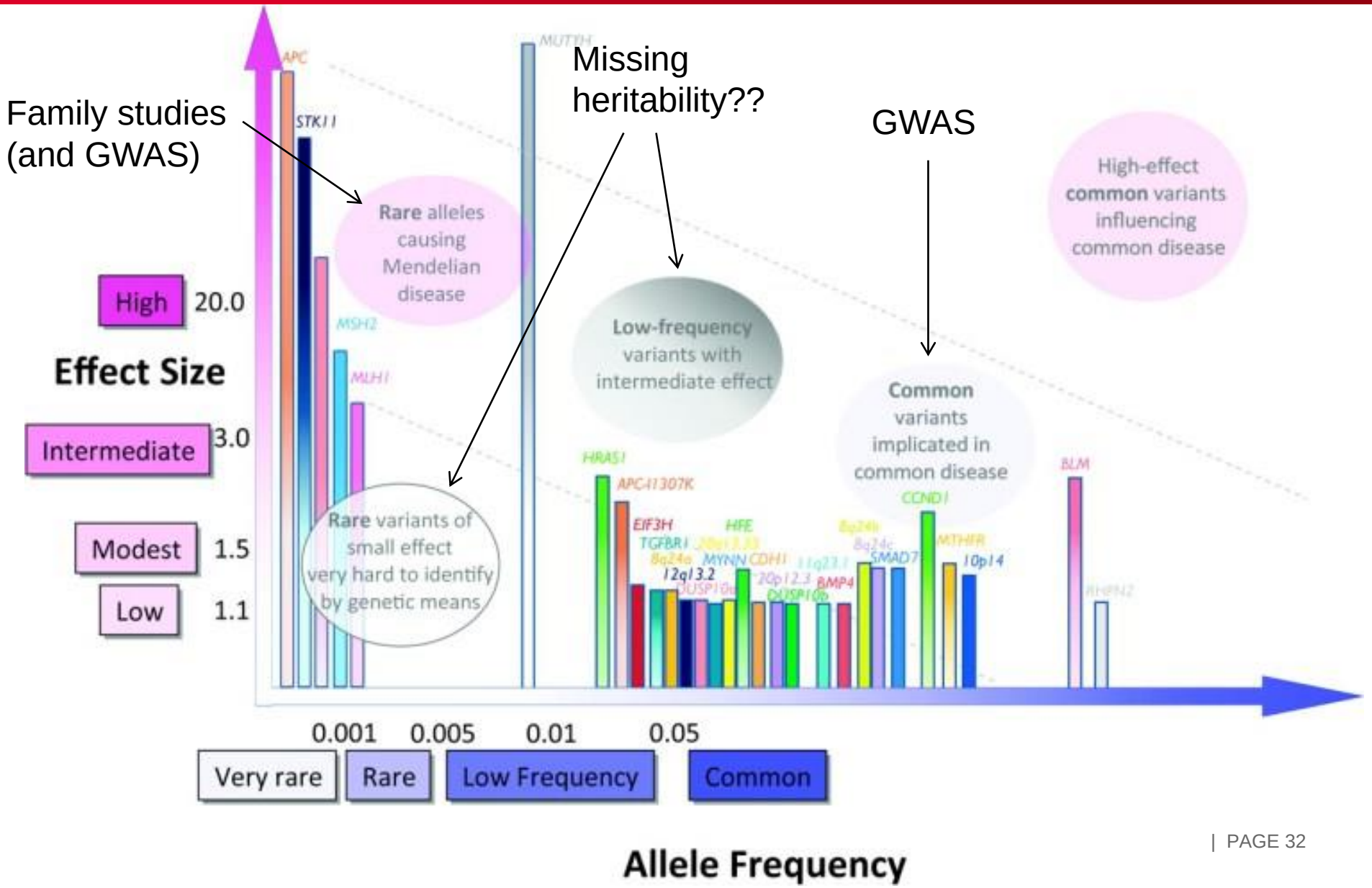
Global results of GWAS on diseases



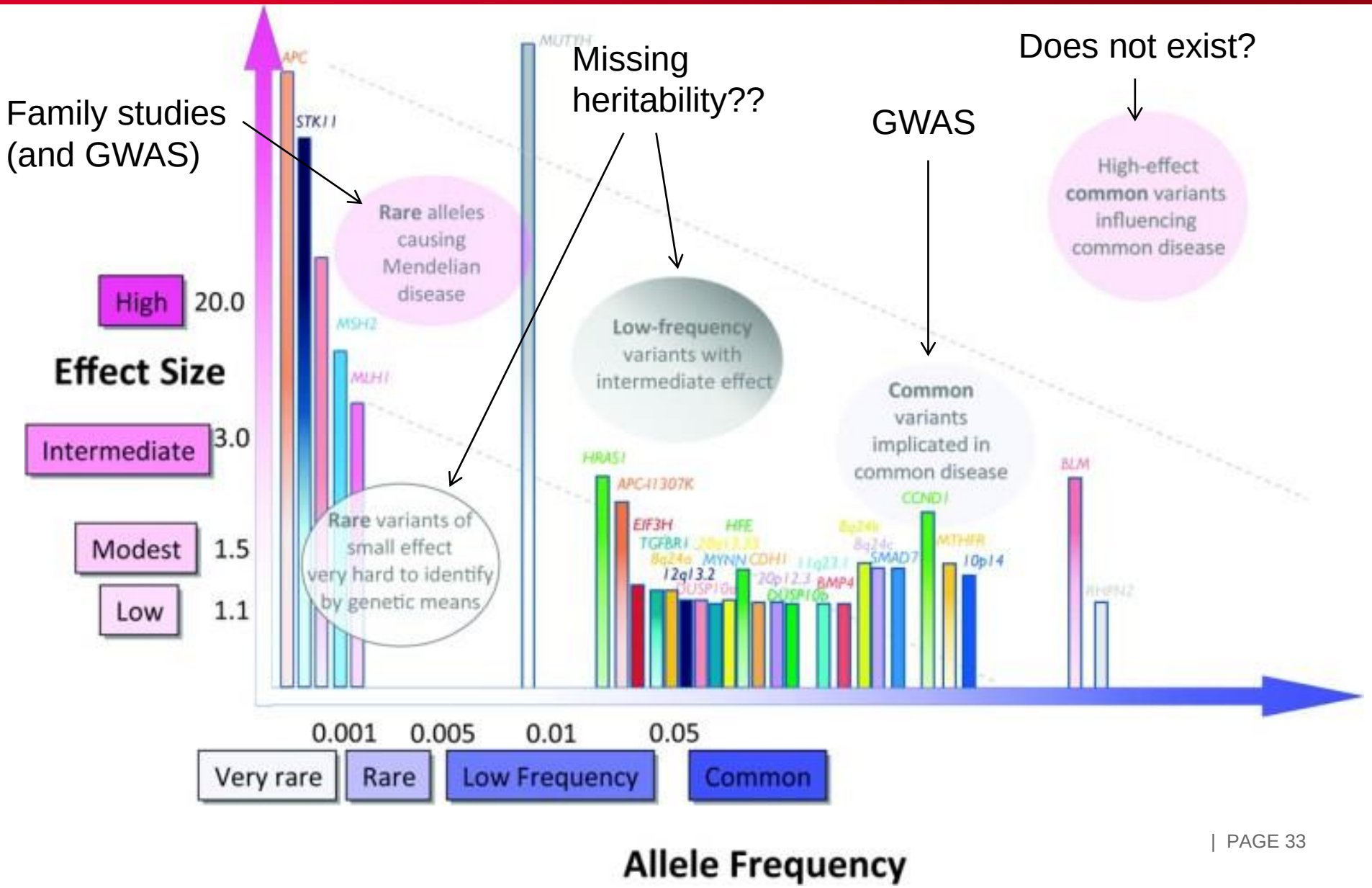
Global results of GWAS on diseases



Global results of GWAS on diseases



Global results of GWAS on diseases



Part 2.2:
The univariate approach on ADNI sequencing data

Alzheimer's Disease Neuroimaging Initiative Sequencing data

- **809 individuals** with Whole Genome Sequencing data
  - 188 Alzheimer's Disease (AD)
  - 393 Mild Cognitive Impairment (MCI)
  - 228 controls

- ~4 million variants per individual
  → ~**60 million variants** for all individuals

- **63% of the SNVs with good quality**
  → ~40 million variants of good quality

Many rare variants and a few annotations in ADNI

Among the 40 million variants of good quality :

- **46% of variants are specific to only 1 individual**:
  18 400 000 variants

⇒ <1% of the variants of an individual are specific to this individual : 20000-25000 variants

- **2% of the variants located in genes** (coding for proteins)

- **15% of the variants are annotated** by epigenetic markers seen in brain cells (DNA regions not necessarily coding for proteins but influencing the transcription into RNA)

Statistical analysis of WGS ADNI data

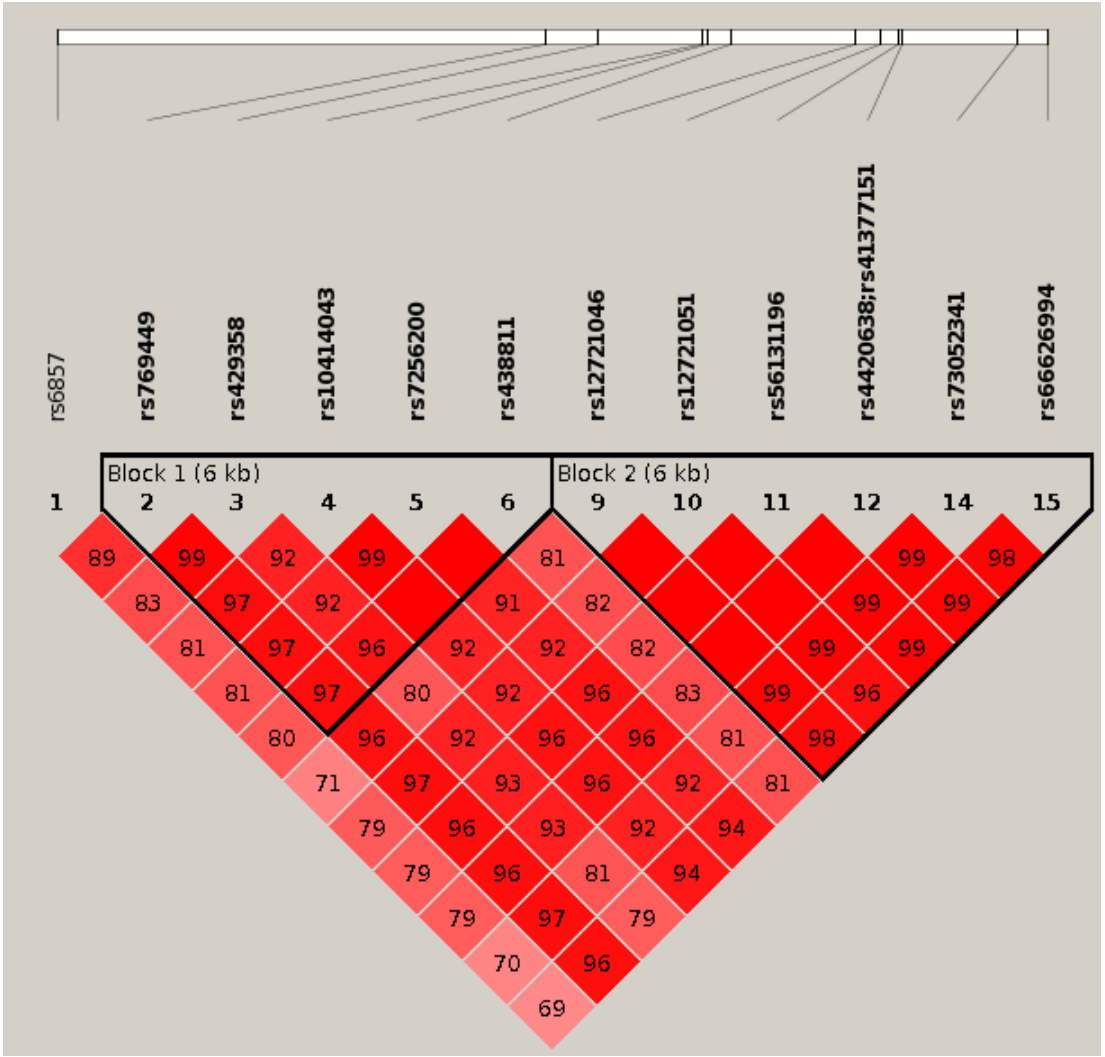
- Comparison of AD patients (188) versus others (621)

- **16 SNPs with a significant** p-value after Bonferroni correction (χ^2 test/Fisher's exact test) : $10^{-18} < p < 10^{-9}$
  in the APOE region (36kb)

- **Top associated SNP:**
  rs429358 (non-synonymous) : $p=5*10^{-18}$
  One of the 2 SNPs of **APOε4 allele**

- **Associated variants are frequent** (MAF 20-40%)

LD structure between the significant SNPs



CNRGH
CENTRE NATIONAL DE RECHERCHE EN GÉNOMIQUE HUMAINE

Among top associations (35 SNPs with $p<10^{7}$)

APOE region on chromosome 19

rs429358 (missense) : $p=5*10^{-18}$

One of the 2 SNPs of APOε4 allele

PCDH11X on chromosome X

rs2750788 (intron) : $p=4*10^{-8}$

Already associated with Alzheimer's Disease

LINGO2 on chromosome 9

rs2578253 (intron) : $p=5*10^{-8}$

Already associated with Parkinson's Disease

ATP11C on chromosome X

rs2485724 (intron) : $p=2.5*10^{-8}$

Limitations of univariate analysis of Whole Genome Sequencing data

- **Univariate GWAS methods** (linear regression, chi-square) may be applied BUT:

  - much more multiple comparisons (tens of millions)

  - very low power to detect association for rare variants

- **More individuals needed** but expensive (>1000$ per individual)

- **Statistical methods need to be adapted** by collapsing nearby variants: **Region-based analysis (multivariate approach)**

  → stronger signals and fewer tests (20000-25000 genes)

Part 2:
The multivariate approach

Part 3.1:
The multivariate approach on sequencing data

Region-based analysis for sequencing data

Multiple regression model :

- p variants in **a certain region** (e.g. a gene)
- **Genotypes** of individual i : $\boldsymbol{X}_i$ $(1 \times p)$, coded 0, 1 or 2
- **Covariates** of individual i : $\boldsymbol{Z}_i$ $(1 \times k)$ such as age, sex, pop. structure

- For a **case/control** (1/0) phenotype Y_i :

$$logit(p_i) = ln\left(\frac{p_i}{1-pi}\right) = \alpha_0 + \boldsymbol{Z}_i\boldsymbol{\alpha} + \boldsymbol{X}_i\boldsymbol{\beta} \qquad \text{with } p_i = P(Y_i = 1|\boldsymbol{X}_i, \boldsymbol{Z}_i)$$

- **Test of no region effect**: H_0: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T = \mathbf{0}$

Sequence Kernel Association Test (SKAT)

- **Recall of the multiple regression model**:

$$E(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i)/logit(p_i) = \alpha_0 + \boldsymbol{Z}_i\boldsymbol{\alpha} + \boldsymbol{X}_i\boldsymbol{\beta}$$

- **Assume random effects:**

$$\beta_j \sim distribution\ (0, w_j{}^2\tau)$$

where w_j^2 is an optional weight for variant *j* (higher for rare variants)

- **Test of no region effect**:

$$\mathrm{H}_0 : \beta_1 = \cdots = \beta_p = 0 \ \leftrightarrow\ \tau = 0$$

SKAT results on ADNI sequencing data

SKAT tests the association of a group of variants (a gene) with the phenotype, assuming additive effects of variants:

- SKAT **on each full gene: no significant results** after correction even on candidate genes

- SKAT **on each gene with exons only: 2 significant genes** ($p=10^{-6}$) after correction **APOE** and **SORBS3** (already associated with Alzheimer's disease)

Part 3.2:
The multivariate approach on genotyping data

Heritabilty on genotyping data to predict Alzheimer's disease

Heritability on SNPs:
Same model as SKAT (logistic regression with additive random effects) but on genome-wide common SNPs

Results obtained on Alzheimer's disease genotyping data:

- with 809 individuals (188 AD/621 controls) and 2.5M SNPs from ADNI : **heritability of 10% but high variance**

- with 9900 individuals (2400 AD/7500 controls) and 500K SNPs from CNRGH : **heritability of 75%!**

Results of multivariate methods on genotyping data to predict Alzheimer's disease

Results obtained on Alzheimer's disease genotyping data using AdaBoost (trees) or Random forests:

- with 809 individuals (188 AD/621 controls) and 2.5M SNPs from ADNI

- with 9900 individuals (2400 AD/7500 controls) and 500K SNPs from CNRGH

Data set	Accuracy AD	Accuracy controls	Global accuracy
ADNI genotyping data	6%	100%	63%
CNRGH genotyping data	46%	93%	82%

Summary of multivariate analysis of ADNI data

- At the gene level with sequencing data, **significant association for APOE and SORBS3 only** and driven by common SNPs

- At the whole genome level with genotyping data, **classification algorithms failed on ADNI data**

- **Improvement when much more samples** and fewer SNPs

Conclusion on the prediction of a phenotype from genome-wide data

- At the gene/SNP level, **a few significant associations and mainly on common SNPs** (no great improvement with sequencing data and rare variants yet)

- At the whole genome level **multivariate algorithms are promising on genotyping with common SNPs data** suggesting cummulative effects of many SNPs

- But we **need many samples!** A lot are coming…. (even companies like Google)

- We need to **integrate other sources of omics data** (RNA, proteins, DNA methylation, DNA 3D structure, …) and biological knowledge (such as gene networks)