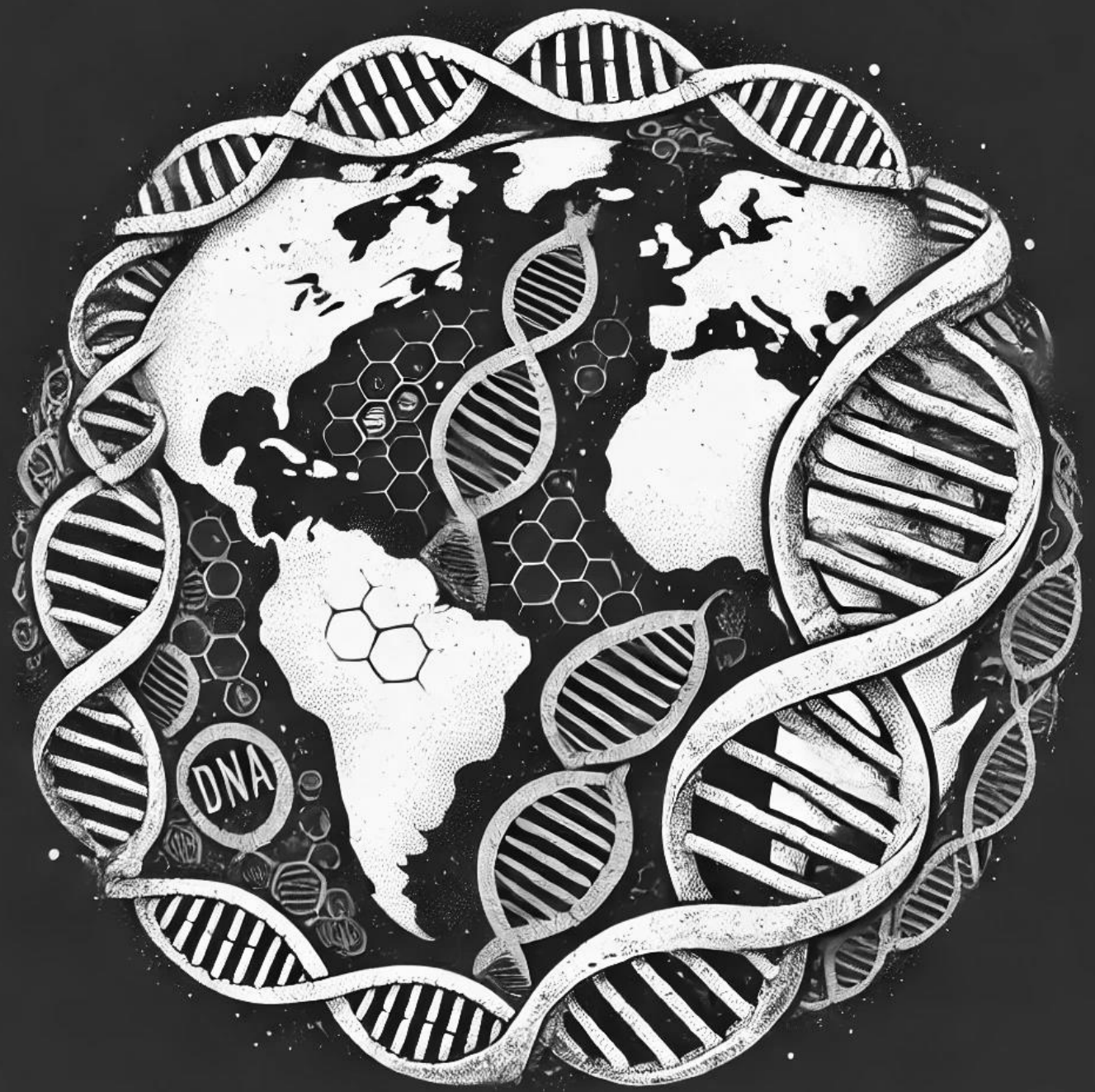


MASTER COURSE IN PERSONALISED MEDICINE

POLYGENIC SCORES

Palle Duun Rohde

[palledr@hst.aau.dk | p.rohde@rn.dk]

https://pdrohde.github.io/

Associate Professor & Research Group Leader

Department of Health Science and Technology

Aalborg University

Clinical Academic

Department of Clinical Genetics

Aalborg University Hospital

THE PURPOSE OF TODAY

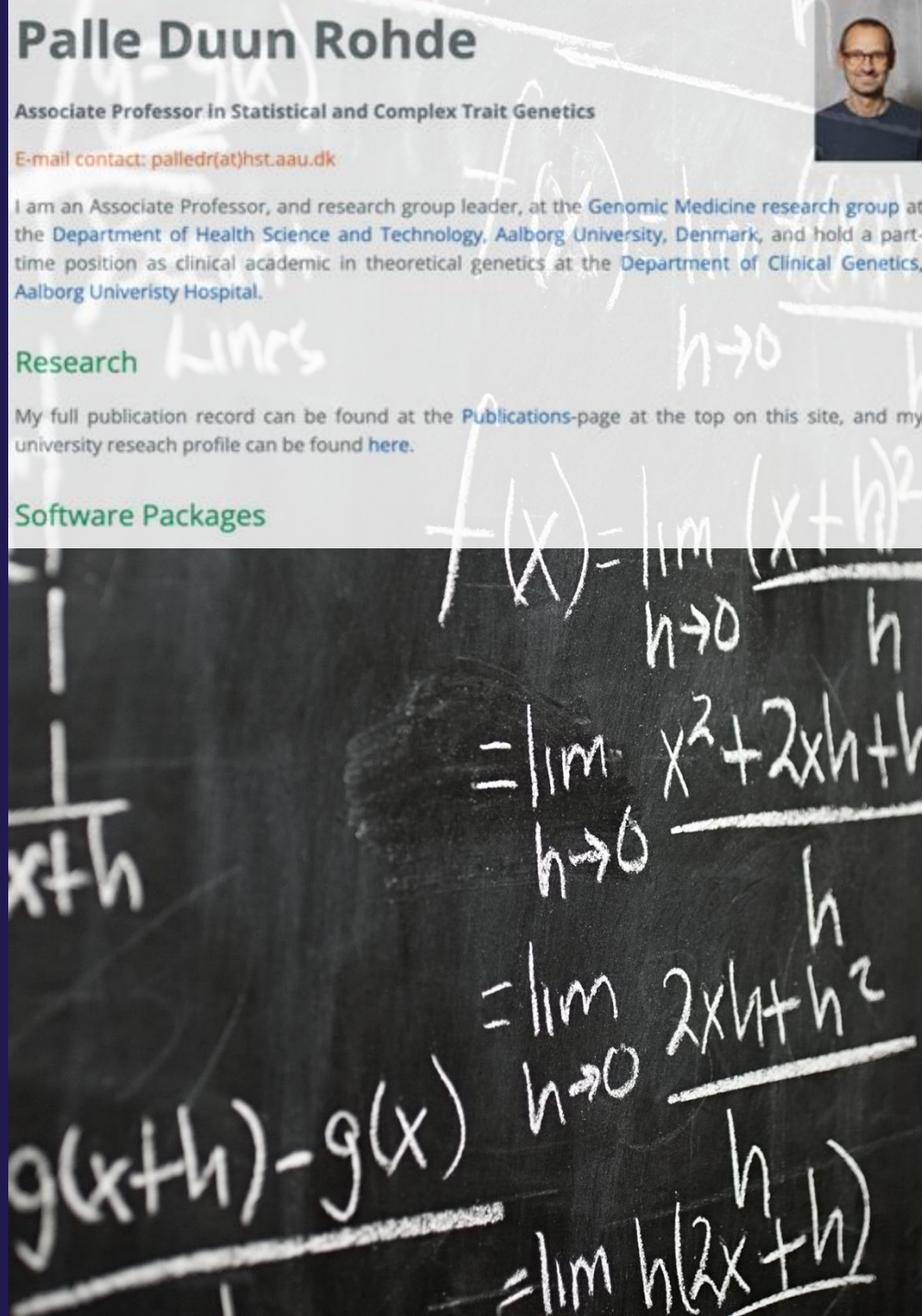
- **Give an introduction to polygenic scores (PGS)**
- **How genetic variation shapes phenotypic diversity**
- **Provide an introduction to complex trait genetics**
  - Monogenic vs multifactorial aetiology
- **How we can utilize genomic data to elucidate molecular genetic aetiology underlying complex traits**
  - Genome-wide association studies (GWAS)

Slides → https://pdrohde.github.io/teachingmaterial/

WHO AM I?

- BCs in biology from AU
- MSc in genetics from AU
- PhD in statistical and quantitative genetics from AU

- Research group leader aiming to understanding the genetic architecture of common complex diseases and translating genomic discoveries into improved prevention, diagnosis, and treatment
- A central research priority for the group is advancing statistical genomics through AI-driven methods.



AGENDA

09:00 – 09:15	Welcome
09:15 – 09:45	Session 1: Genetic variation and heredity
09:45 – 10:15	Group work 1: Genetic variation and heredity
10:15 – 10:30	Break
10:30 – 11:00	Session 2: From SNP to GWAS to disease risk
11:00 – 11:40	Group work 2: Construct your own PGS
11:40 – 12:00	Common discussion and evaluation

Join at menti.com | use code 7798 1843

AALBORG UNIVERSITET

What comes to mind when you hear *personalised medicine*?

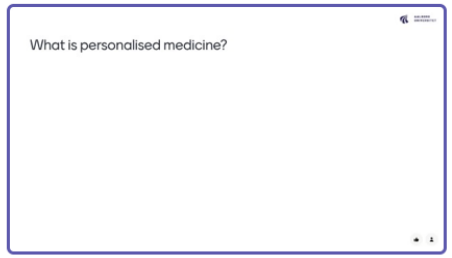
9 / 10

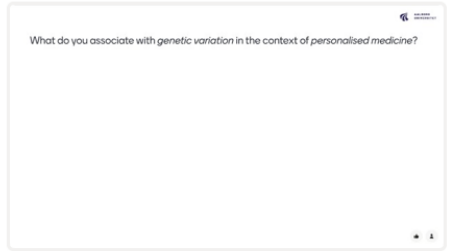
PD

Menti

PM-genetik2025

Choose a slide to present



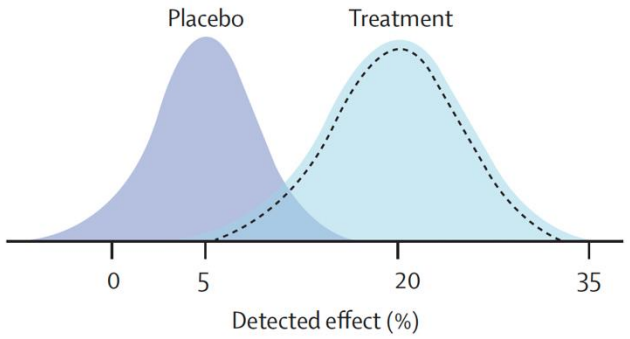


IMPLEMENTATION OF PRECISION MEDICINE

EPPOS [evidence-based precision personalised objective subjective]

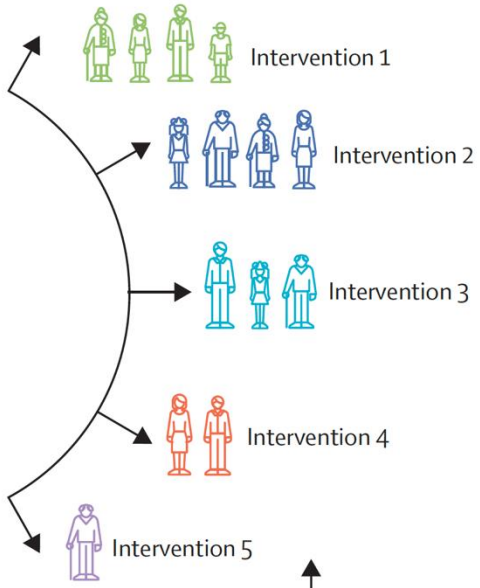
Evidence-based Medicine

(1) Contemporary evidence-based medicine
Estimate average risk or response using epidemiological and clinical trial cohorts



Precision Medicine

(2) Probability scoring and stratification
Maximise response and minimise risk using subclassification

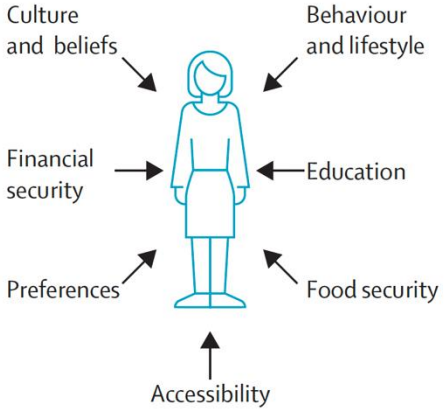


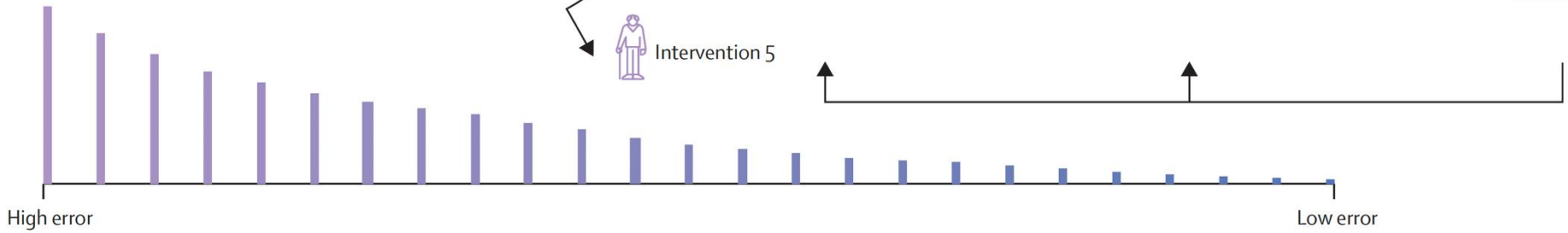
Personalised Medicine

(3) Personalisation (objective)
Monitor response to optimise dose, timing, and delivery

Individualised Medicine

(4) Personalisation (subjective)
Adapt intervention to fit the person's needs, capabilities, and preferences





Franks, P. W. *et al.*. *Lancet Diabetes Endocrinol.* **11**, 822–835 (2023).

AGENDA

09:00 – 09:15	Welcome
09:15 – 09:45	Session 1: Genetic variation and heredity
09:45 – 10:15	Group work 1: Genetic variation and heredity
10:15 – 10:30	Break
10:30 – 11:00	Session 2: From SNP to GWAS to disease risk
11:00 – 11:40	Group work 2: Construct your own PGS
11:40 – 12:00	Common discussion and evaluation

Join at menti.com | use code 7798 1843

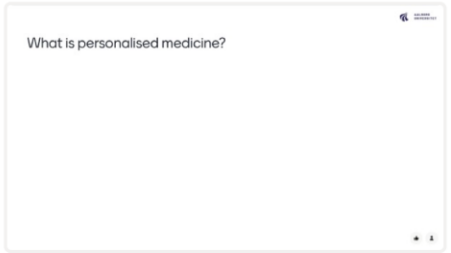
AALBORG UNIVERSITET

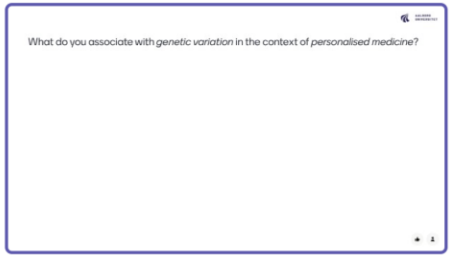
PD

Menti

PM-genetik2025

Choose a slide to present





9 / 10

SESSION 1

- Complex traits
- Genetic variation
- Linkage Disequilibrium (LD)
- Genetic parameters

AALBORG UNIVERSITY

DIFFERENT MODE OF INHERITANCES

- Monogenic (single gene variant)
- Polygenic (many gene variants)
- Multifactorial (many gene variants plus environment exposures)

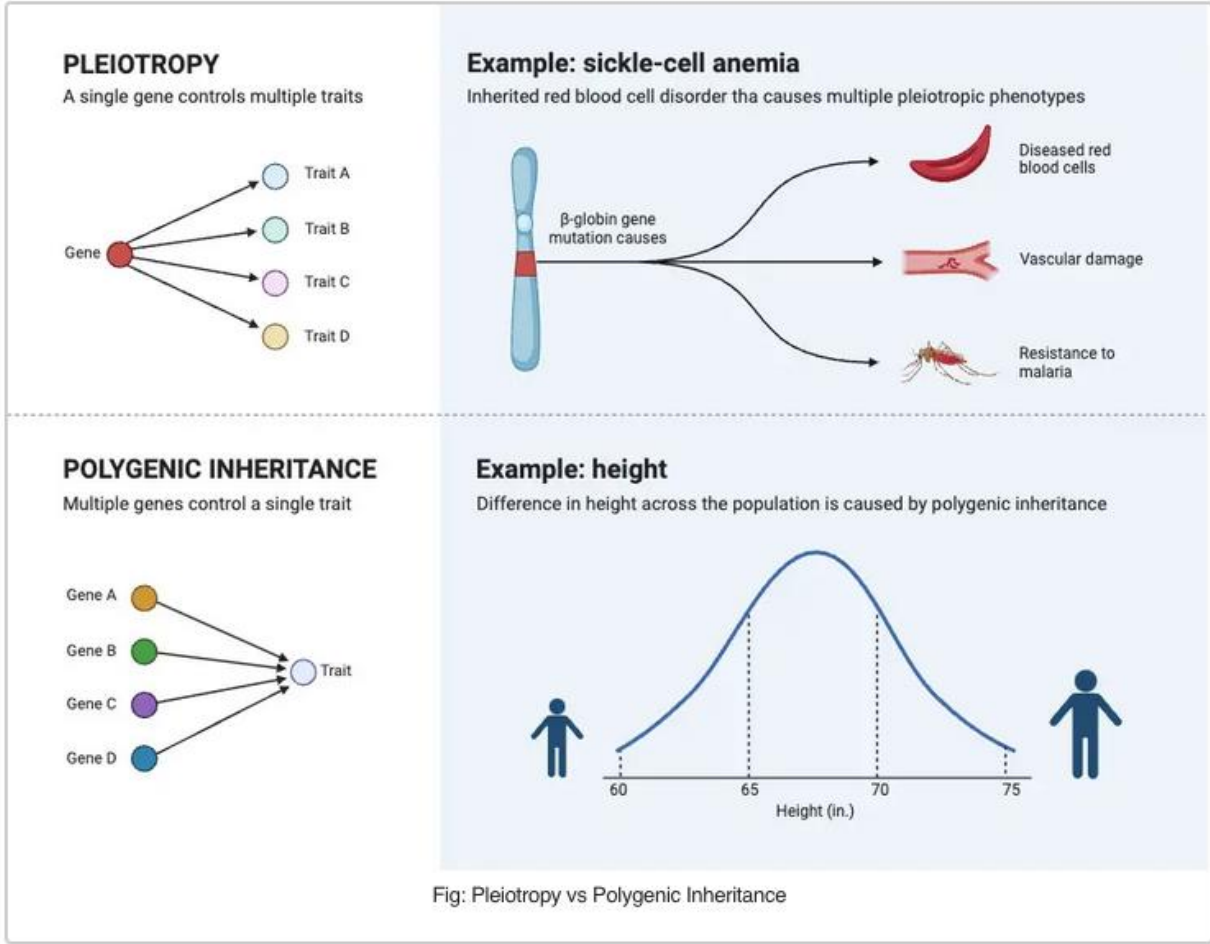
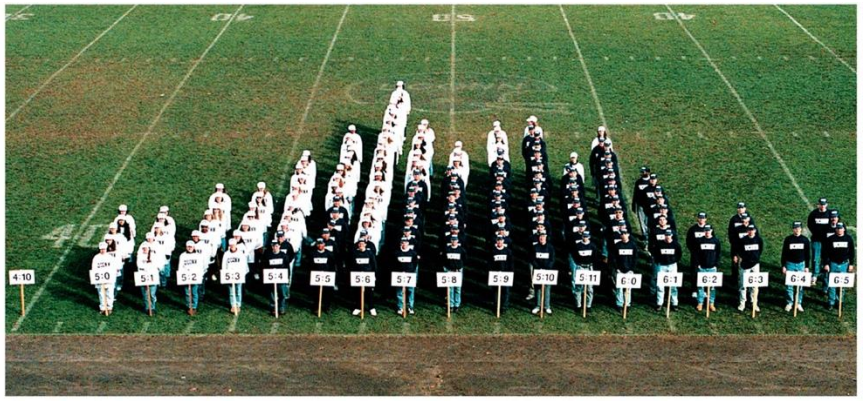


Fig: Pleiotropy vs Polygenic Inheritance

QUANTITATIVE TRAITS

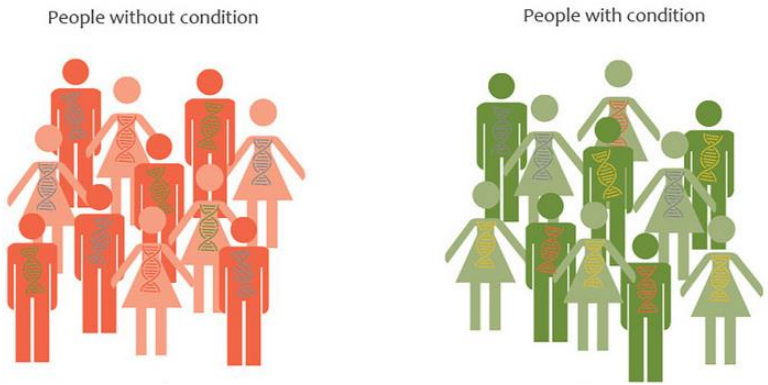
IN DIFFERENT SHAPES

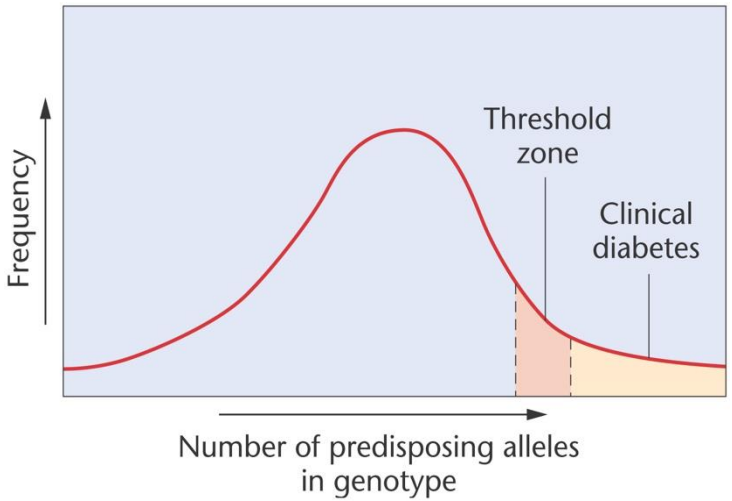
Continous variation



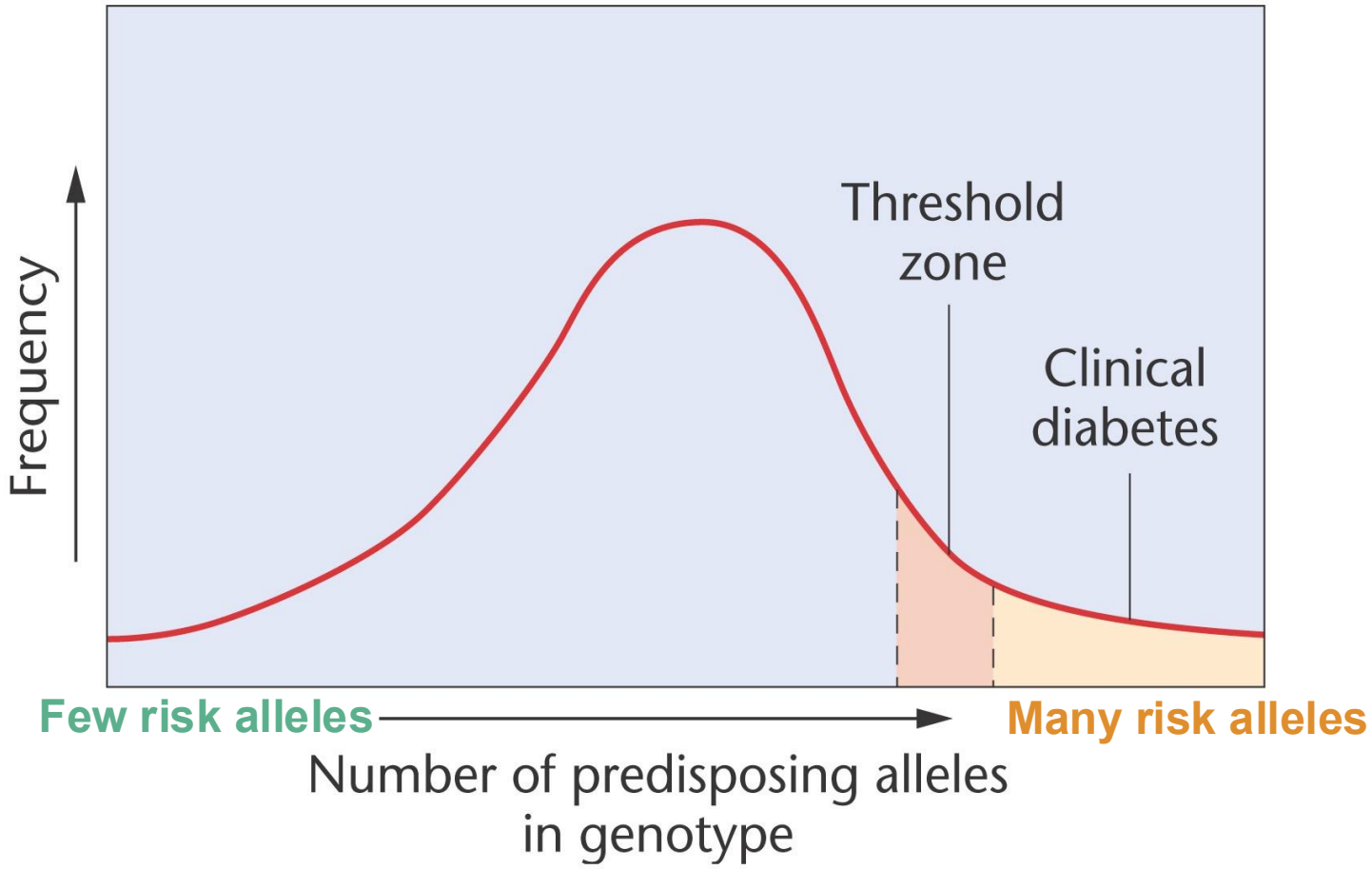
Categorical variation

Threshold traits





LIABILITY (THRESHOLD) MODEL



Liability model

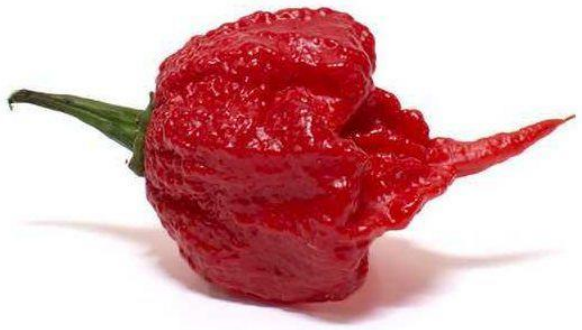
Only individuals with a liability over a certain threshold will become affected

The **sum** of many genetic variants with **small effect/risk**.

Each locus follow Mendelian inheritance pattern, although the trait does not

GENETIC VARIATION IS LIKE CHILI

Carolina reaper



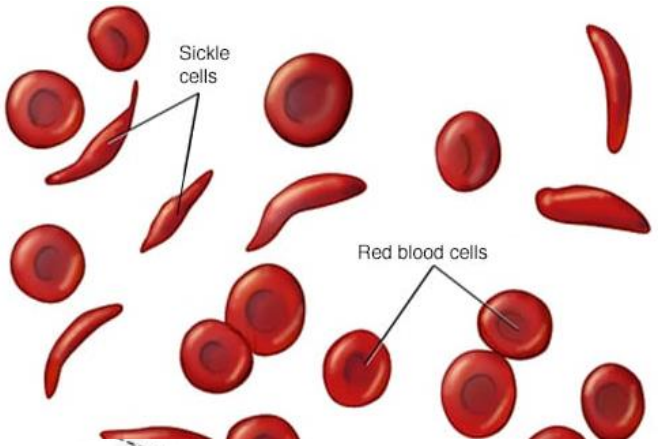
Strong effect on phenotype

Habanero Lemon

Moderate effect on phenotype

Bell pepper

Weak effect on phenotype
(or none at all)



Effect size

50.0

High

3.0

Intermediate

1.5

Modest

1.1

Low

Rare alleles causing Mendelian disease

Low-frequency variants with intermediate effect

Common variants implicated in common disease by GWA

Few examples of high-effect **common** variants influencing common disease

Rare variants of small effect very hard to identify by genetic means

0.001

0.005

0.05

Very rare

Rare

Low frequency

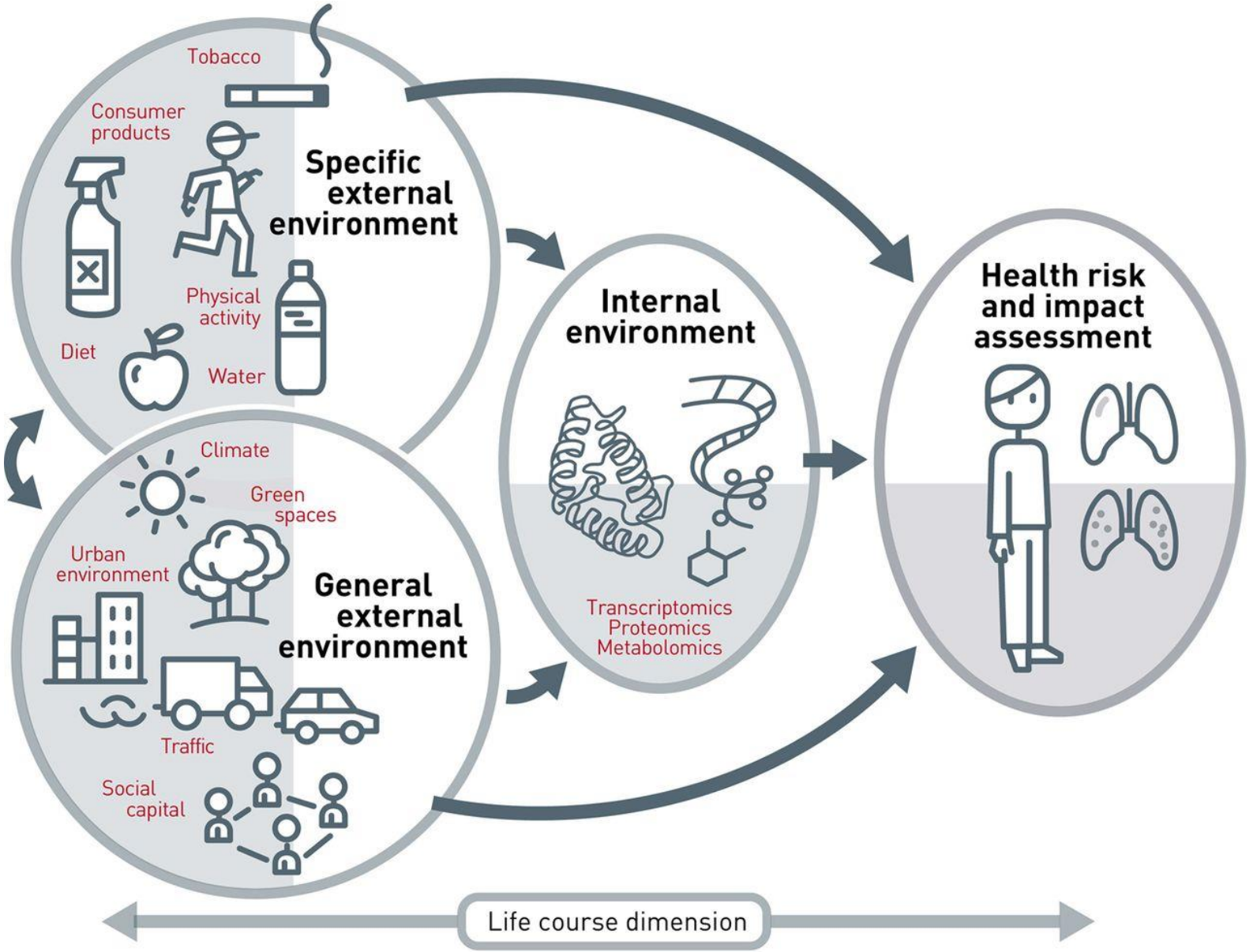
Common

Allele frequency

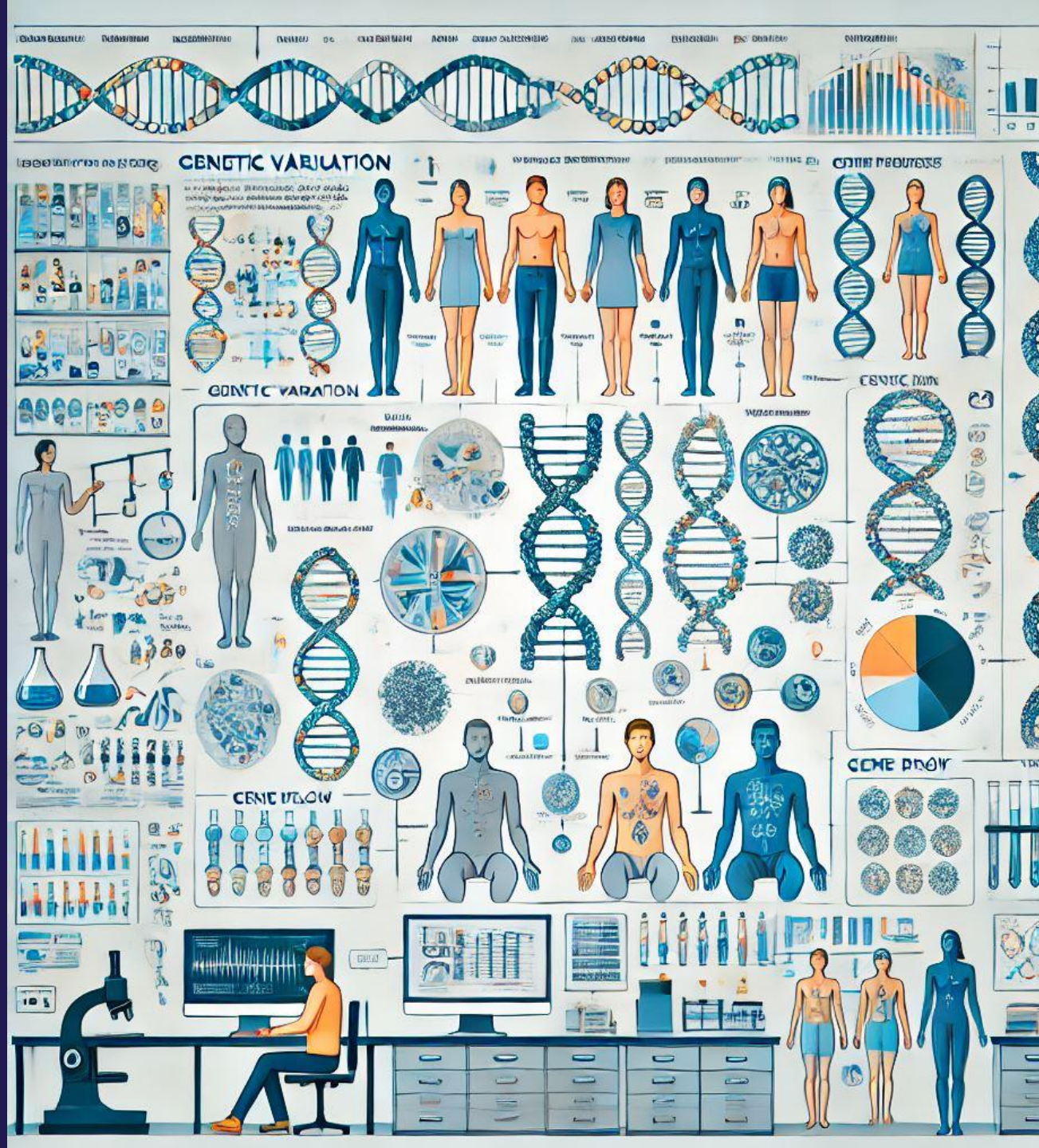
COMPLEX TRAITS are complex...

Complex traits have a **genetic component** and an **environmental component**

The relative genetic contribution is called ***heritability***.



GENETIC VARIATION



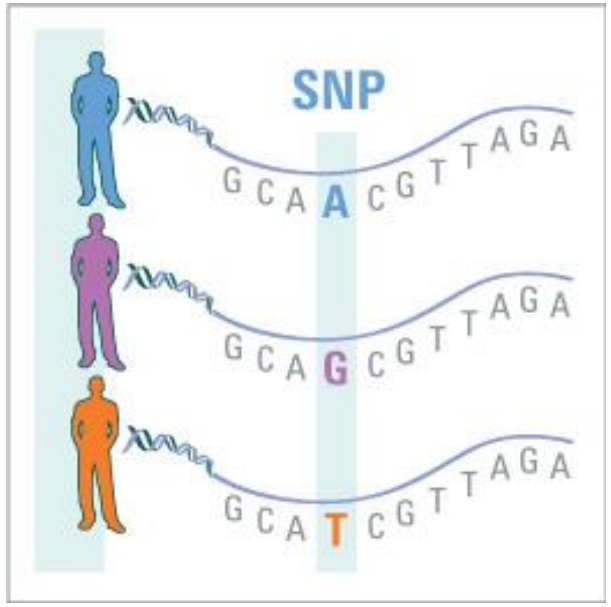
AALBORG UNIVERSITY

GENETIC VARIATION

SINGLE NUCLEOTIDE POLYMORPHISMS (SNPs)

A common change in a single base pair; ~1/1000 bp

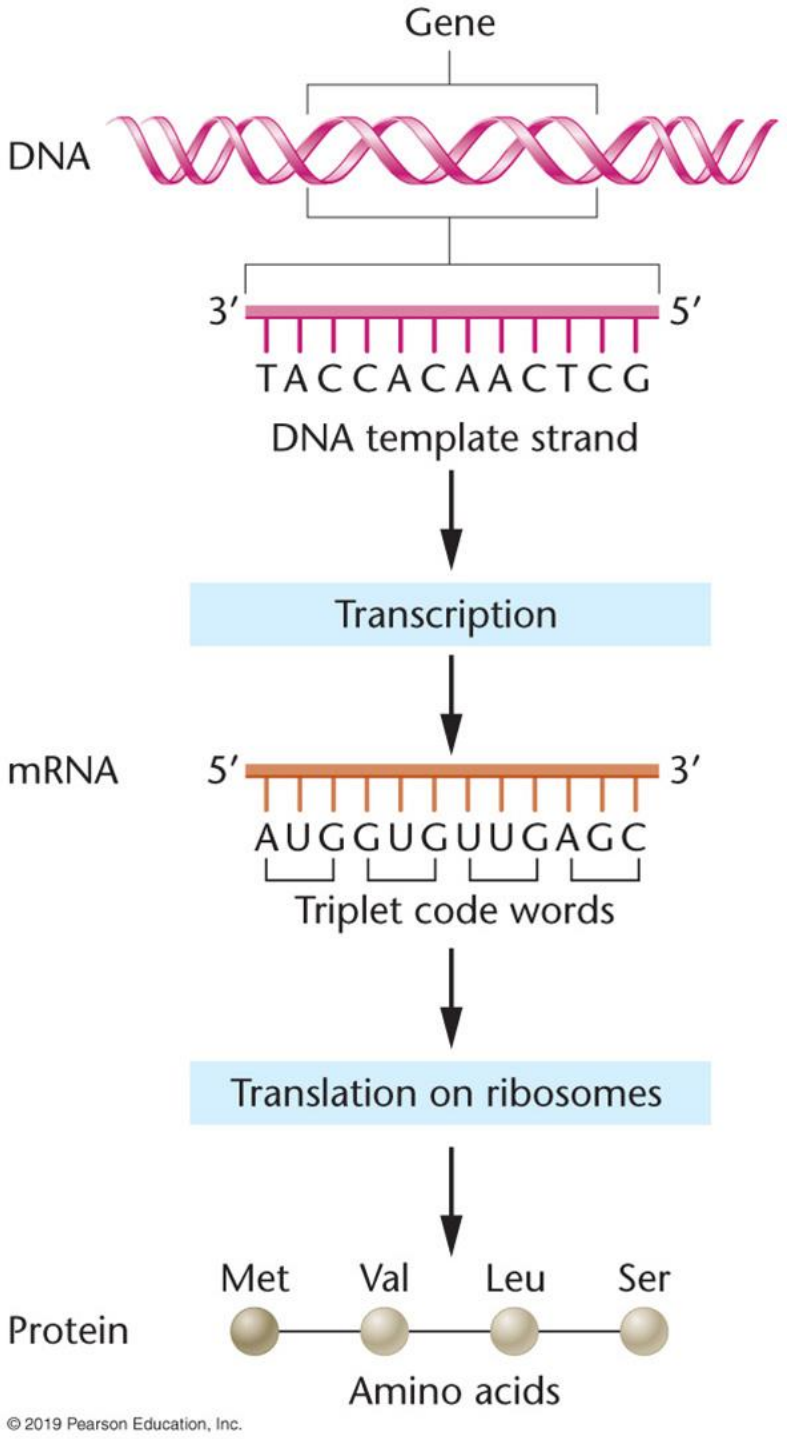
Accounts for ~90% of all variation in the human genome



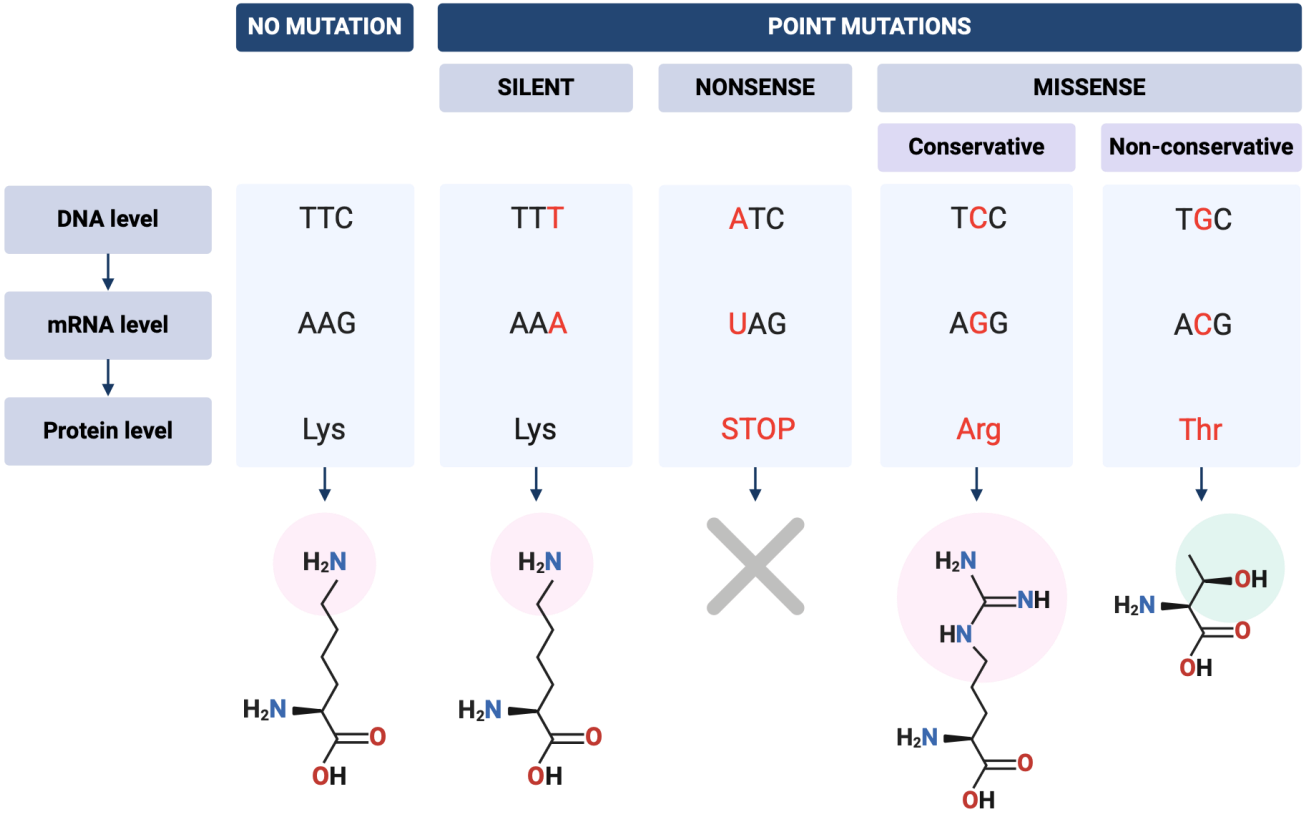
All (known) SNPs has a unique identifier

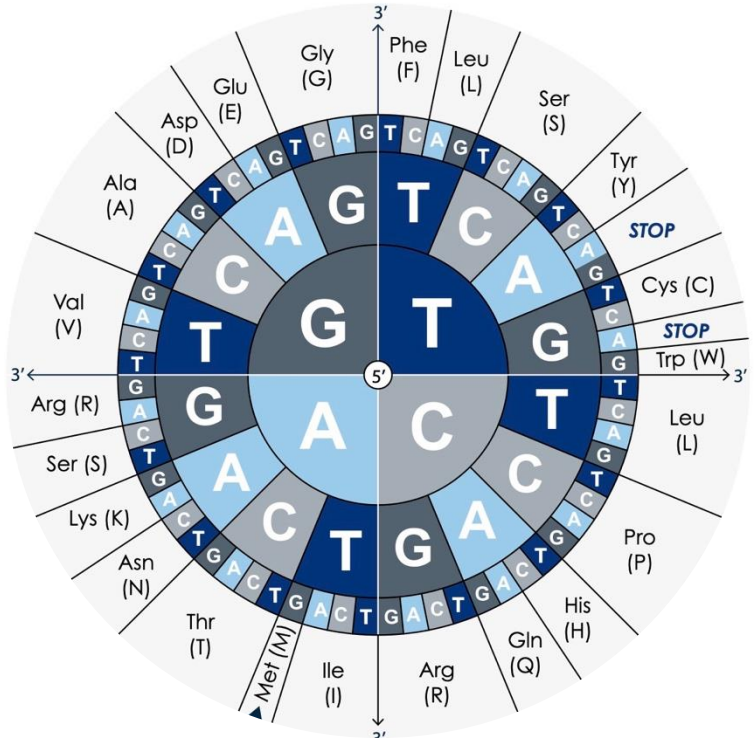
(independent of alleles)

rsXXX – Ref-SNP cluster ID number

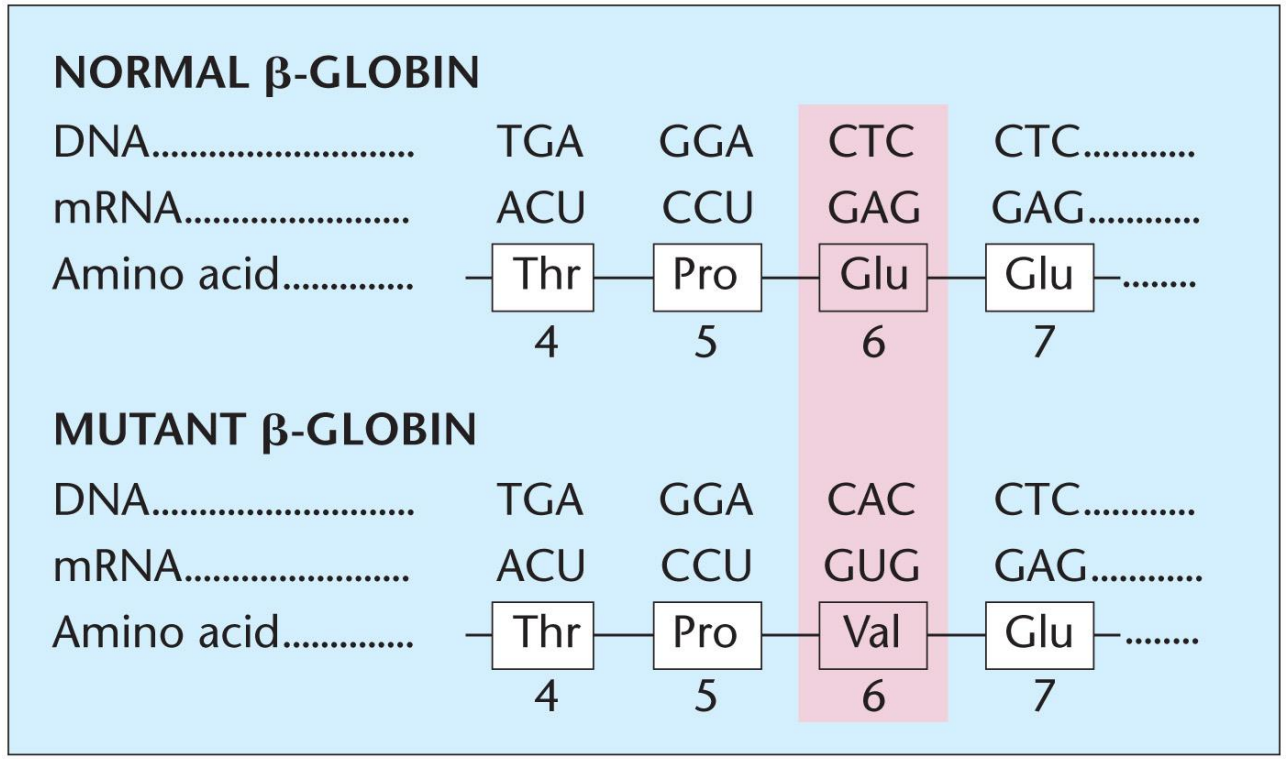


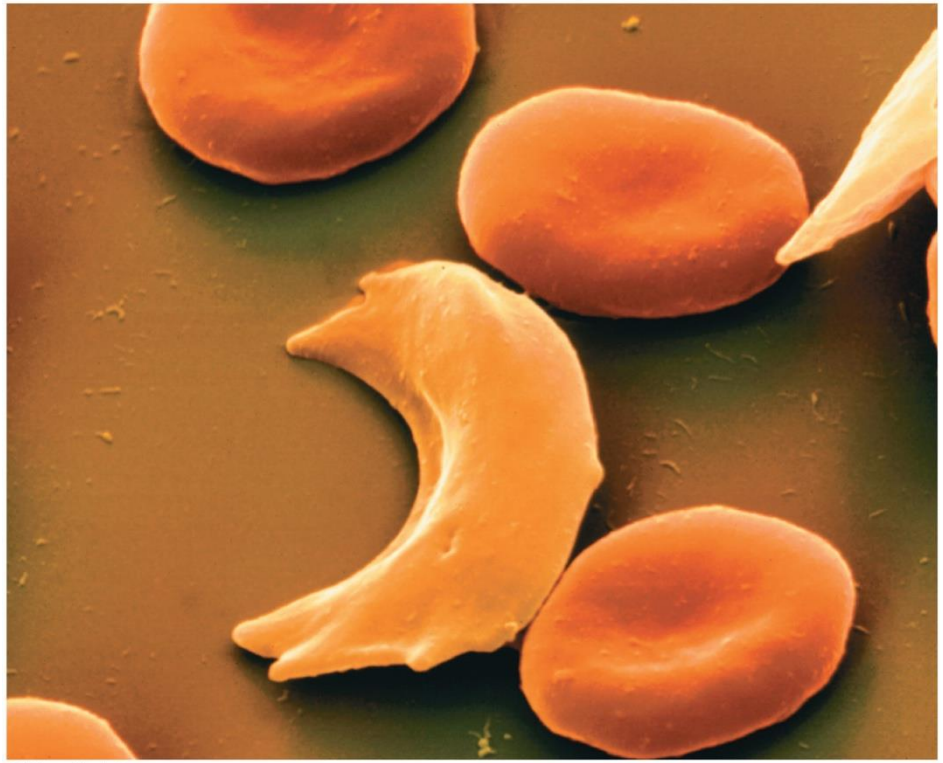
MUTATIONS GENERATE GENETIC VARIATION

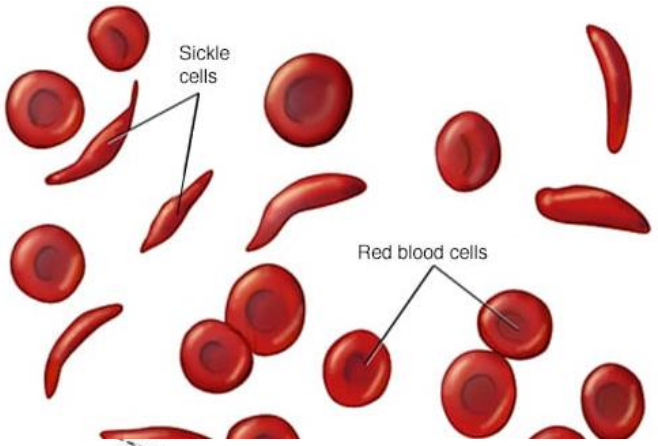




GENOTYPE TO PHENOTYPE







Effect size

50.0

High

3.0

Intermediate

1.5

Modest

1.1

Low

Rare alleles causing Mendelian disease

Low-frequency variants with intermediate effect

Common variants implicated in common disease by GWA

Few examples of high-effect **common** variants influencing common disease

Rare variants of small effect very hard to identify by genetic means

0.001

0.005

0.05

Very rare

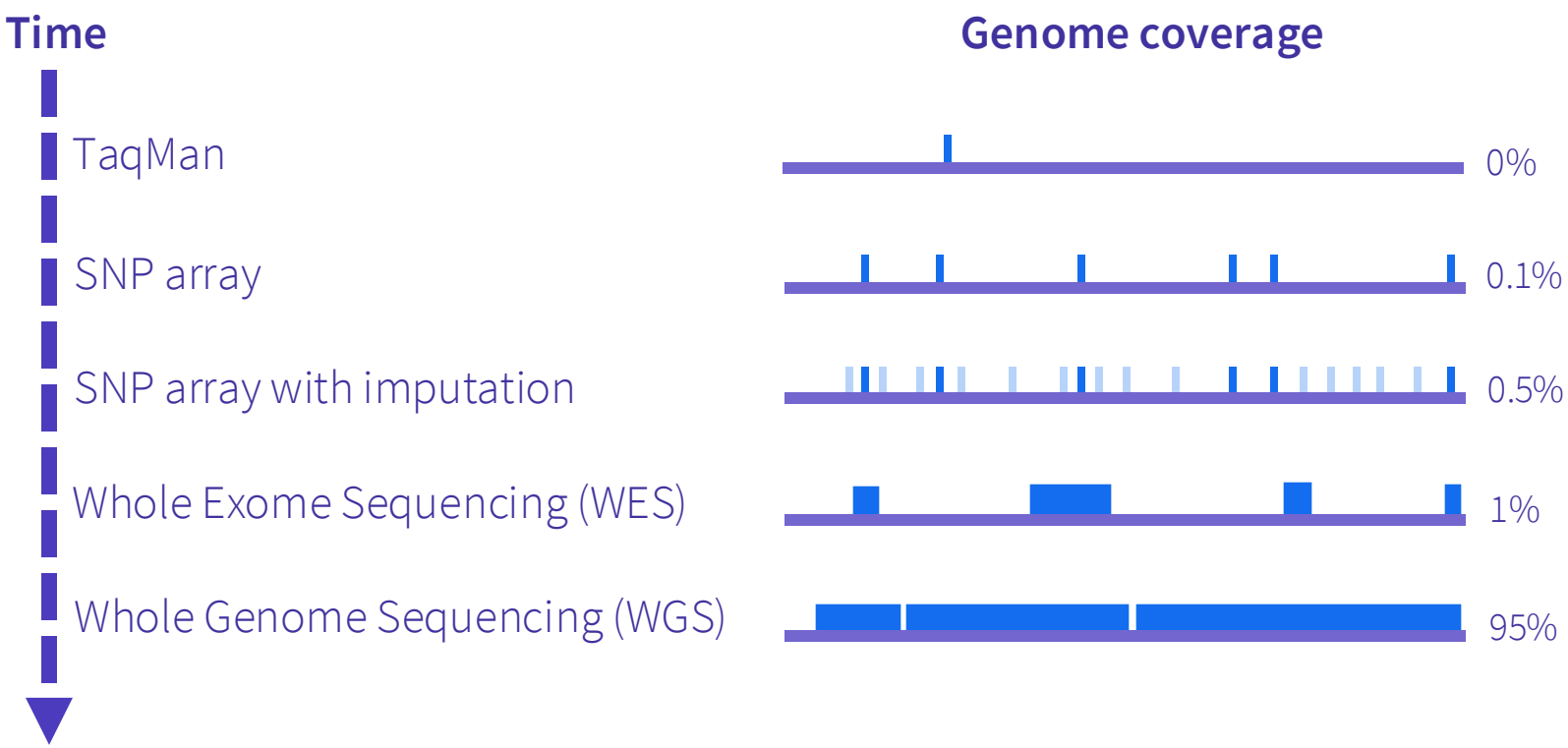
Rare

Low frequency

Common

Allele frequency

GENOMIC COVERAGE

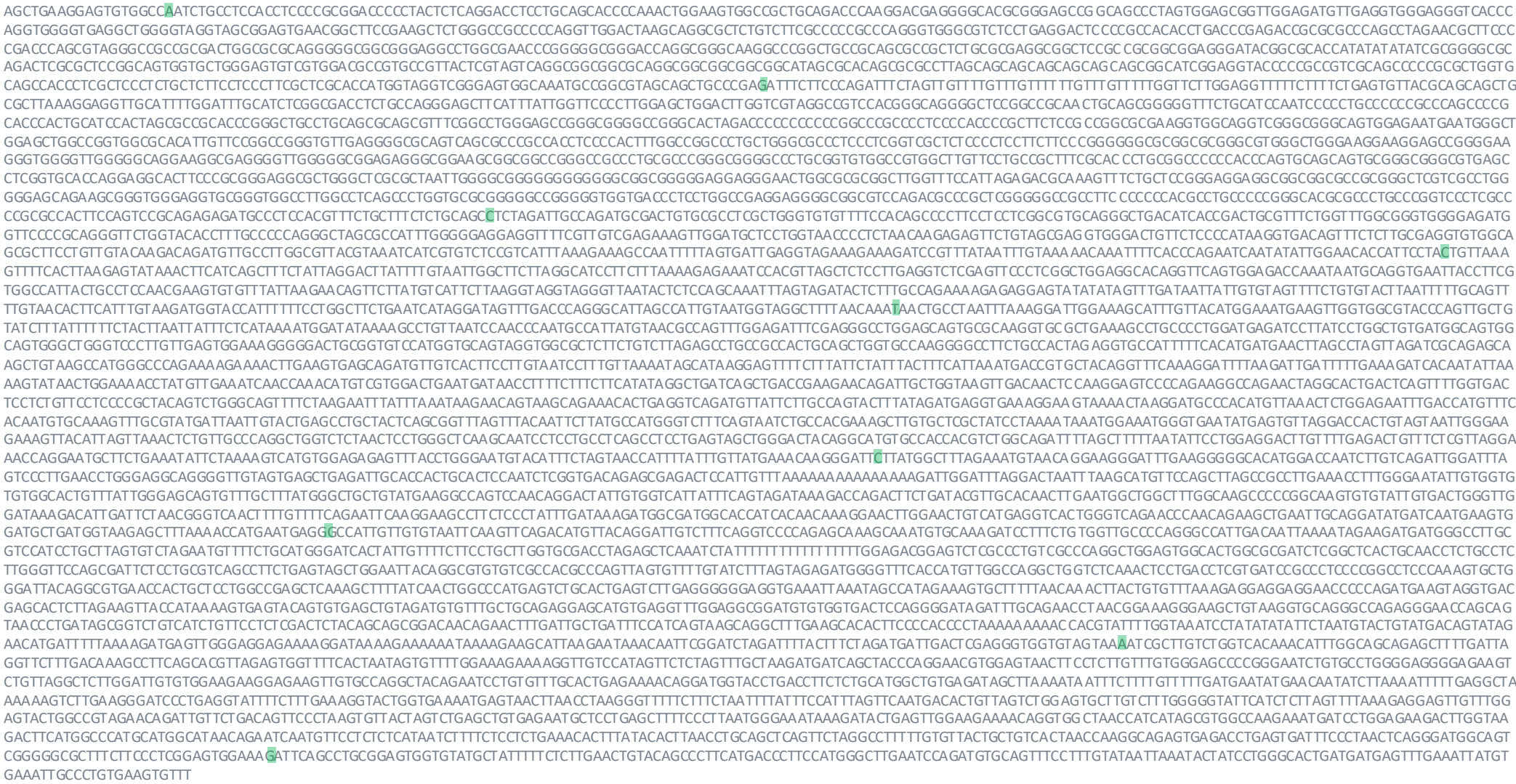


Adapted from Uitterlinden A. (2016) An Introduction to Genome-Wide Association Studies: GWAS for Dummies. Seminars in Reproductive Medicine, 34(4): 196-204.

GENOTYPING VS SEQUENCING

GENOTYPING VS SEQUENCING

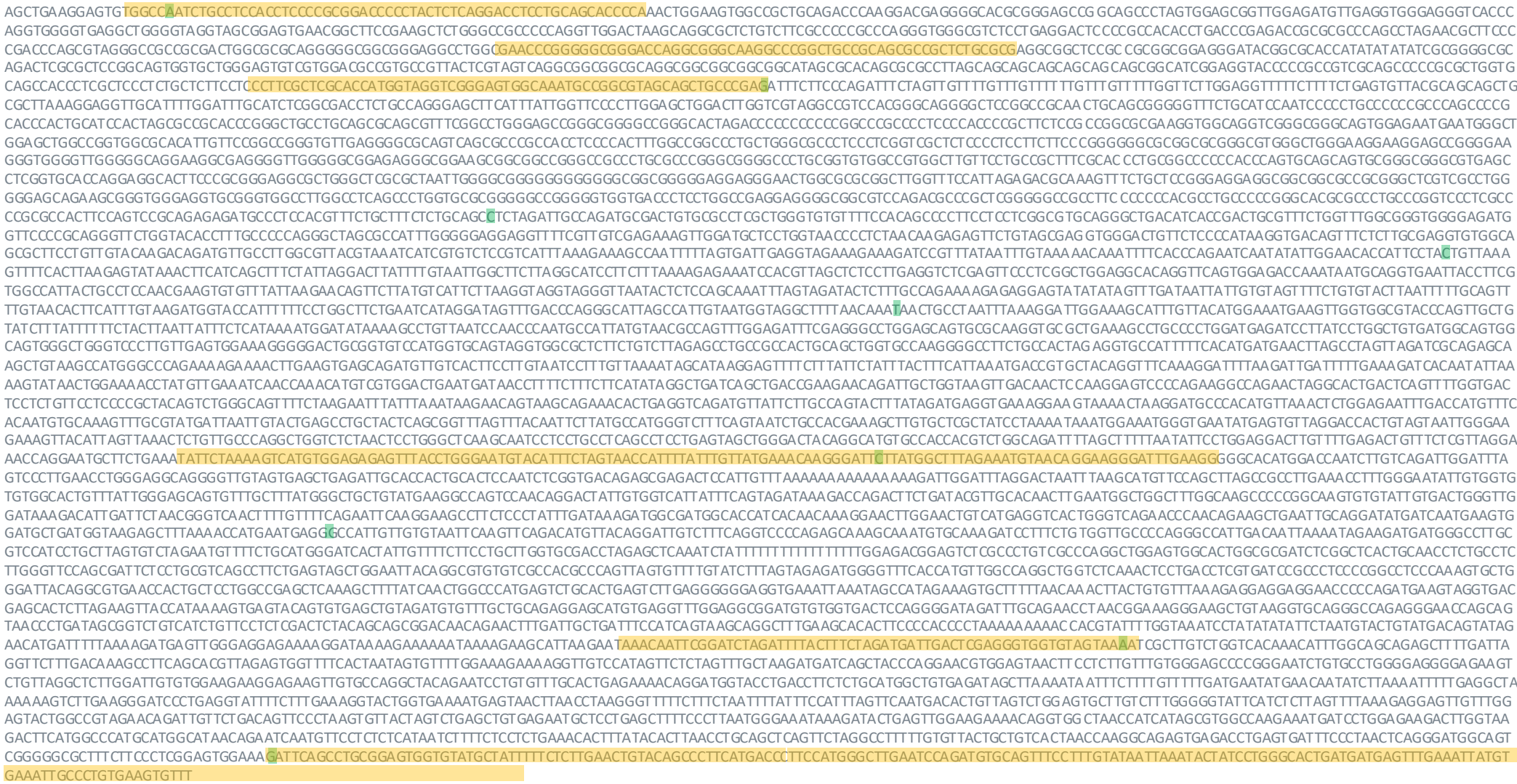
Genotyping



GENOTYPING VS SEQUENCING

Genotyping

WES

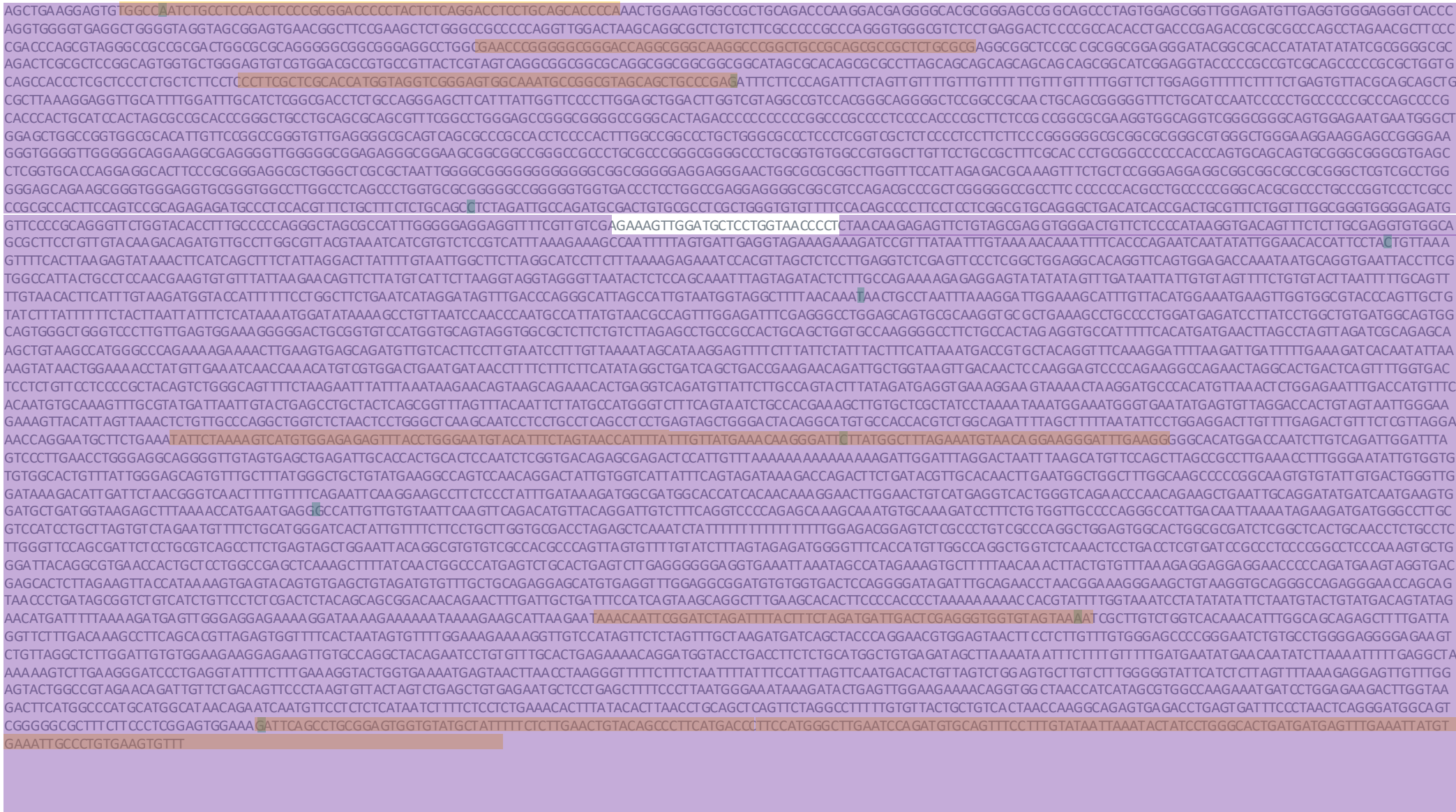


GENOTYPING VS SEQUENCING

Genotyping

WES

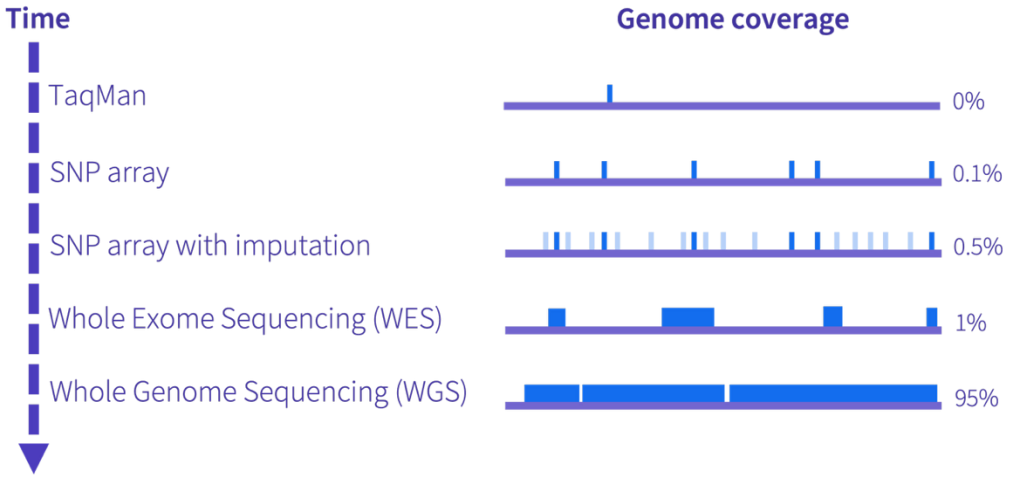
WGS



WHICH TECHNOLOGY?

The choice of technology fir detecting single nucleotide polymorphisms (SNPs) depends upon the application.

For GWAS / PGS we use '*SNP array with imputation*'

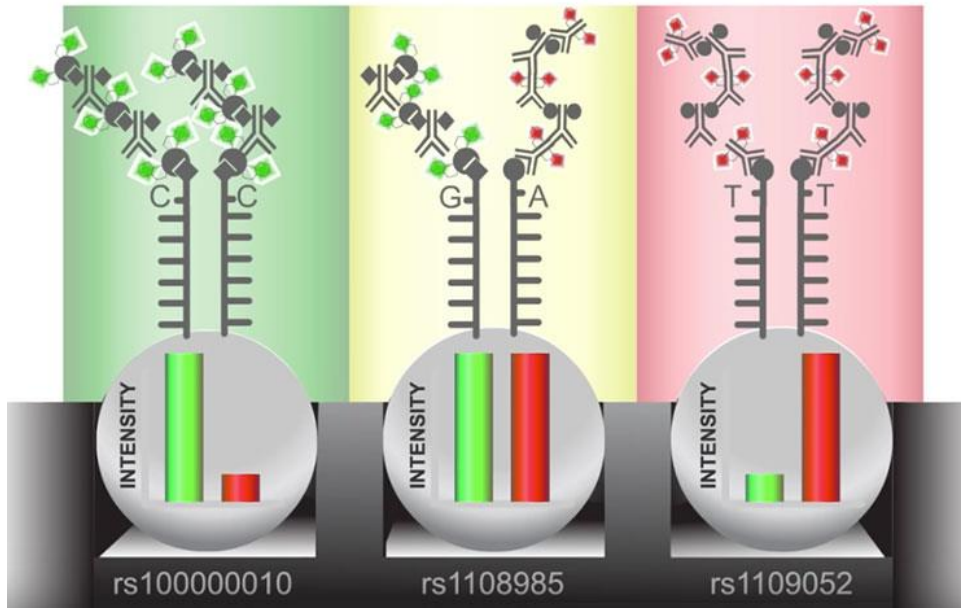


GENOTYPING

For GWAS/PGS we don’t need to analyse all 3 billion positions in the human genome.

Instead, we can measure about 500,000 to 1 million genetic variants that act as markers across the genome.

Because nearby variants are often inherited together, we can use these markers to predict (“impute”) millions of additional common variants — up to about 10 million in total.

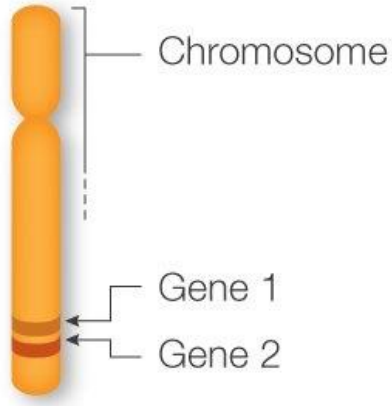


WHY DON'T WE NEED TO TEST EVERY VARIANT?

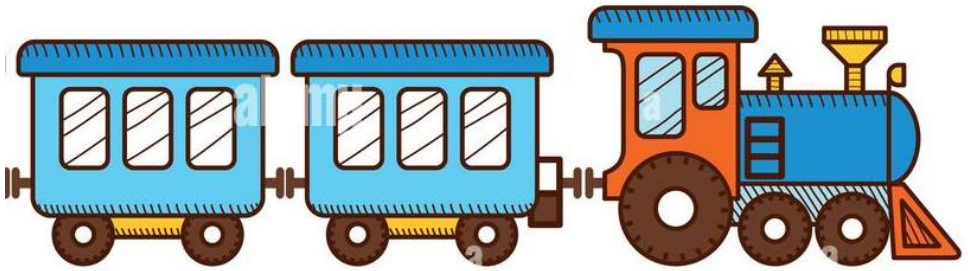
The human genome has about 3 billion positions.

But many variants are inherited together.

Variants that tend to be passed on together are said to be in linkage disequilibrium (LD).

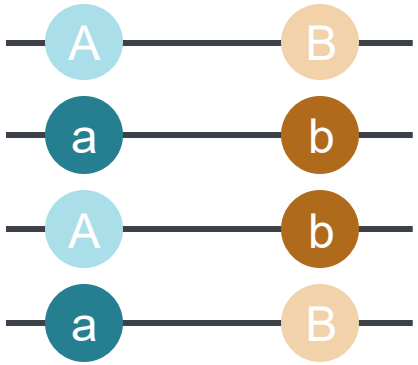


Linked loci
(Physical proximity)



Linked train wagons

WHAT IS LD?



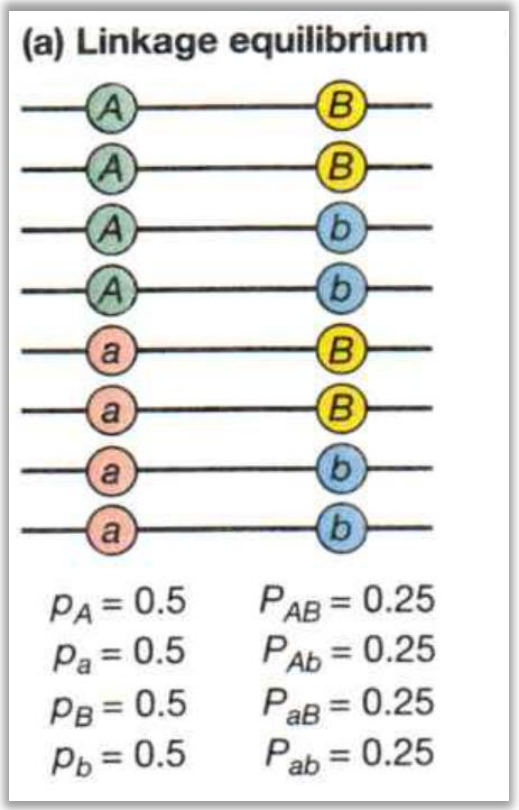
If there is random relationship among alleles at the two loci then the frequency of the haplotypes will be the product of the frequencies of the two alleles:

$P(AB)=P(A)xP(B)$

$P(Ab)=P(A)xP(b)$

$P(aB)=P(a)xP(B)$

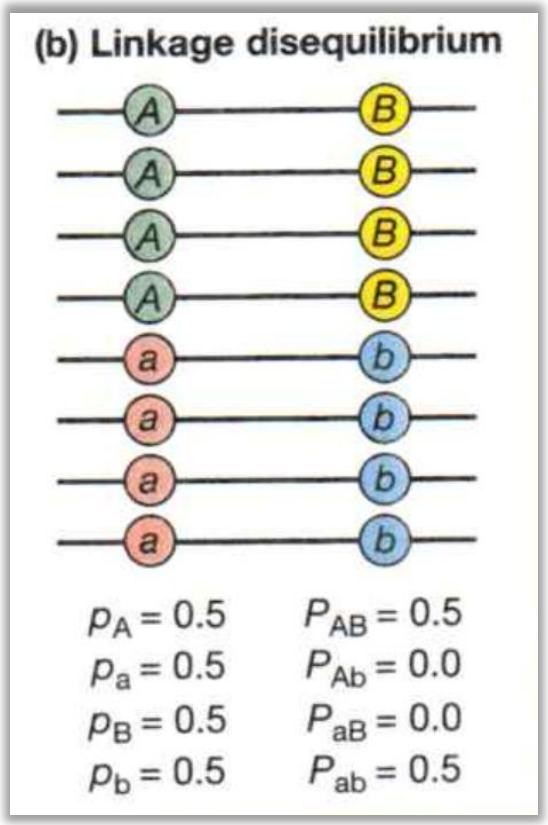
$P(ab)=P(a)xP(b)$



WHAT IS LD?

LD is just a way of describing how strongly two variants are correlated. In practice, we use r^2 to say how well one variant predicts another

$r^2 = 1$ → perfect correlation (completely linked)
$r^2 = 0$ → no correlation (completely independent)

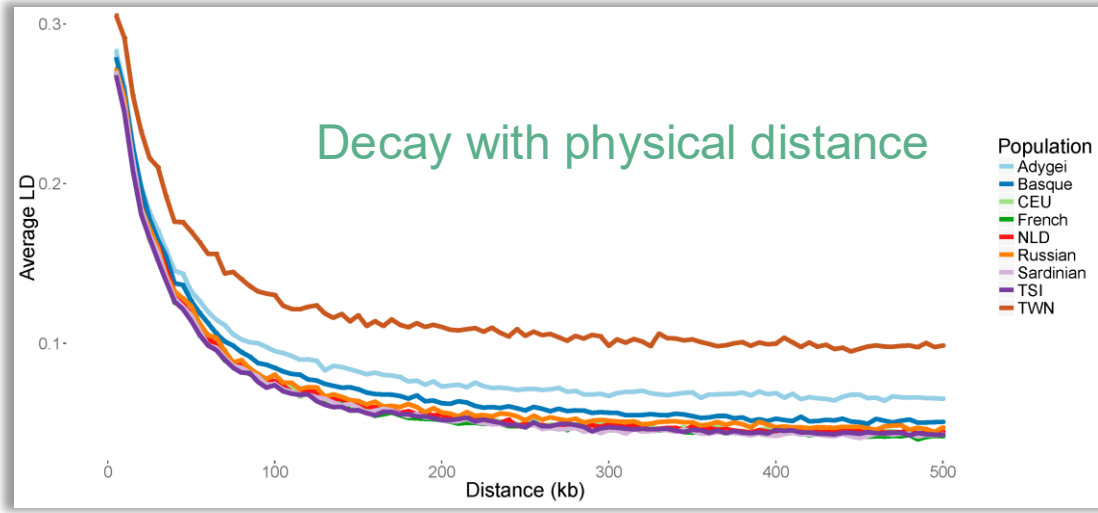


When the association between alleles at two loci is **non-random** they are said to be in **linkage disequilibrium**

The degree of LD can be measure in several ways – the simples one is:

$$D = P_{AB} - P_A P_B$$

If D=0, no LD, if D>0 LD



IMPUTATION
USING HAPLOTYPES

The true haplotypes

A	T	C
G	C	A

This individual has inherited a chromosome with alleles A-T-C from one parent, and G-C-A from the other parent

We observe only the genotypes

A/G T/C C/A

Genotype data does not carry information about the haplotypes.

We do not know whether A at SNP1 is coming from the same parent as T or C at SNP2

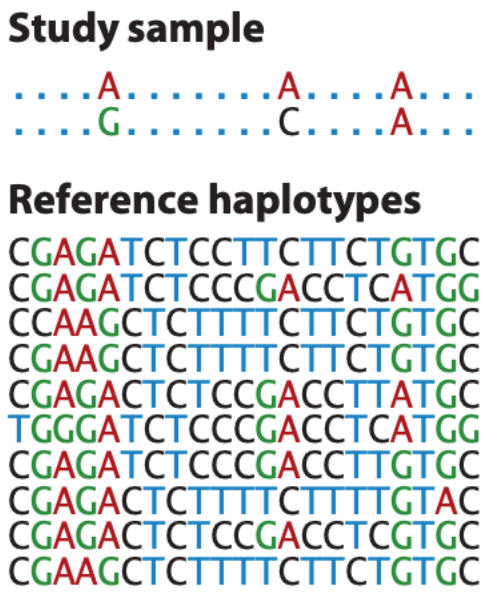
Different haplotypes

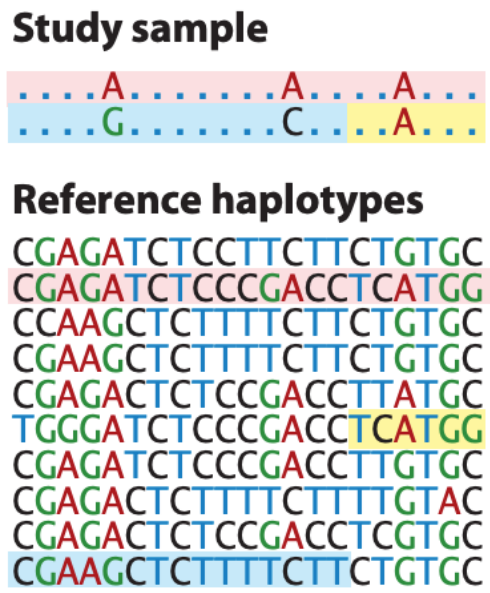
A	C	A
A	C	C
A	T	A
A	C	A
G	C	A
G	C	C
G	T	A
G	T	C

Phasing = estimate the most likely haplotypes

IMPUTATION

Genotypes





From a sequencing study

GENETIC PARAMETERS

- Separate V_G and V_E
- Heritability
- Genetic correlation

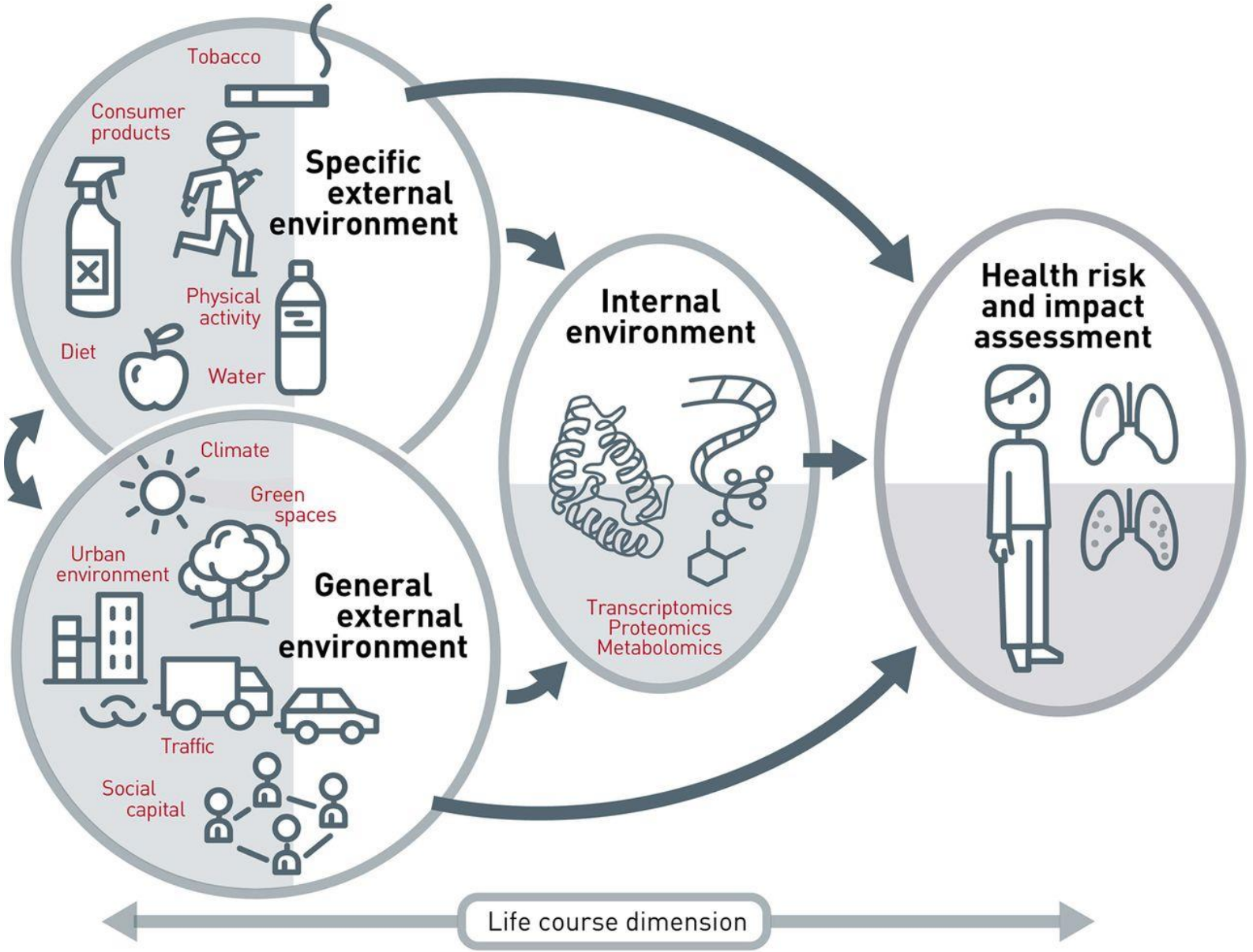
COMPLEX TRAITS are complex...

Complex traits have a **genetic component** and an **environmental component**

The relative genetic contribution is called ***heritability***.

$$V_G / (V_G+V_E)$$

Thus, set a limits for the genetic predictor



MULTIFACTORIAL TRAITS

Multifactorial traits = polygenic effect + environmental effect

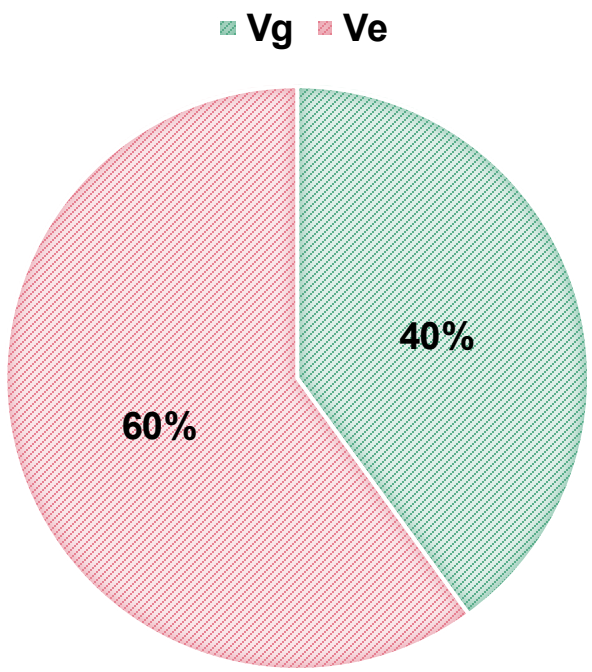
many genes/alleles

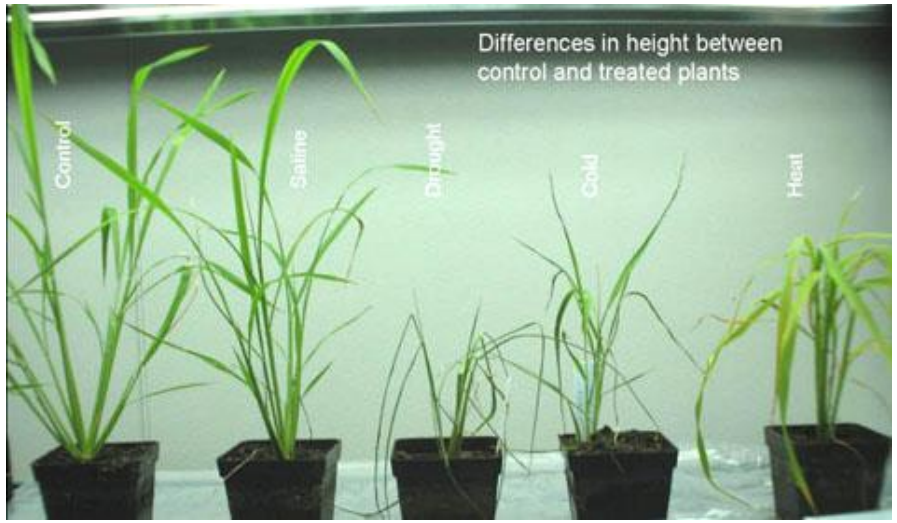
External effects that modulates the phenotypic value.

… or things that we cannot explain

$V_P = V_G + V_E$

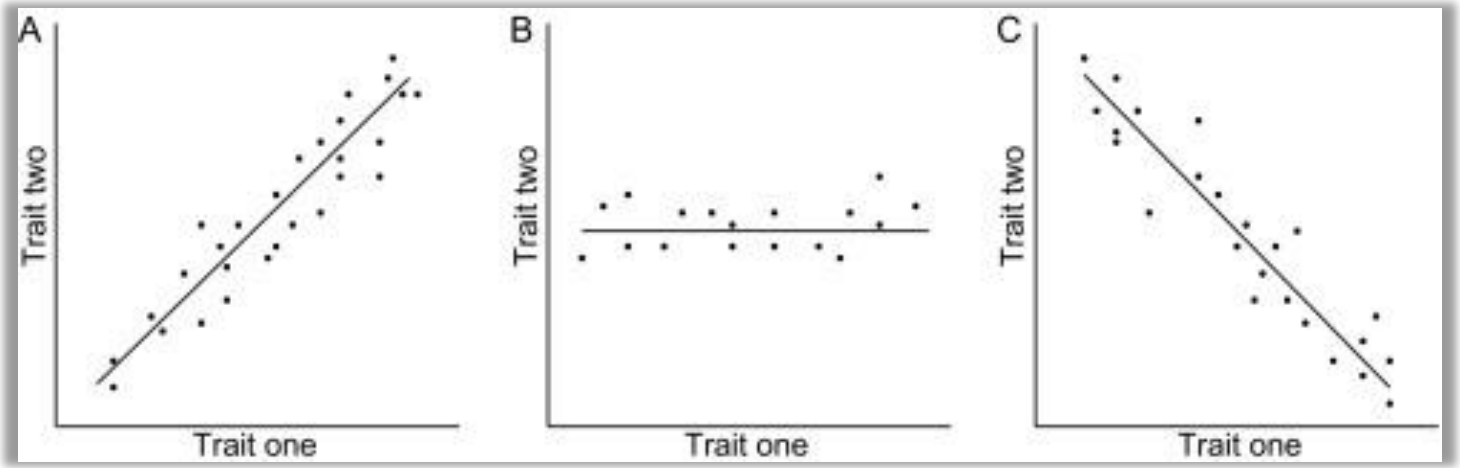
$H^2 = V_G / (V_G + V_E)$





GENETIC CORRELATION

- The **genetic correlation** refers to the genetic link between two traits, which can help elucidate the shared biological pathways and/or causal relationships between them.



Genetic correlation quantifies how much two traits share in their genetic basis — not because of environment or chance, but because *the same DNA variants affect both.*

SESSION 1

- **Complex traits**
  - *Most diseases are influenced by many genetic and environmental factors — understanding their complexity helps us predict risk more accurately*
- **Genetic variation**
  - *Individual differences in DNA underlie why people respond differently to diseases and treatments.*
- **Haplotypes and LD**
  - *Patterns of correlated variants (LD) allow us to identify genetic regions associated with disease risk even when the causal variant is unknown.*
- **Genetic parameters**
  - *Measures like heritability and genetic correlation help quantify how much of disease risk is due to genetics and how traits are biologically connected.*

GROUP WORK

PART 1 (15 min)
Make 3 groups

- ❑ Each group work with one exercise
- ❑ Explain the concepts & discuss how these concepts relates to identification of genetic risk variants

PART 2 (15 min)
We make 3 new groups with 1 person from each of the previous groups

- ❑ Present key messenges from each exercise


BREAK

AGENDA

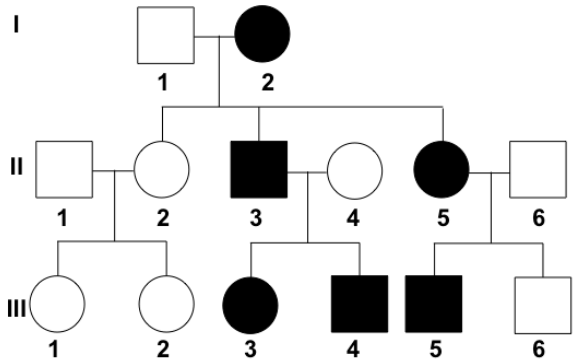
09:00 – 09:15	Welcome
09:15 – 09:45	Session 1: Genetic variation and heredity
09:45 – 10:15	Group work 1: Genetic variation and heredity
10:15 – 10:30	Break
10:30 – 11:00	Session 2: From SNP to GWAS to disease risk
11:00 – 11:40	Group work 2: Construct your own PGS
11:40 – 12:00	Common discussion and evaluation

SESSION 2

- Genetic associations
- Polygenic score

AALBORG UNIVERSITY

Monogenic disorders



Mutation

Each genetic variant is both necessary and sufficient

Linkage analysis
(pedigree)

Common complex disorders

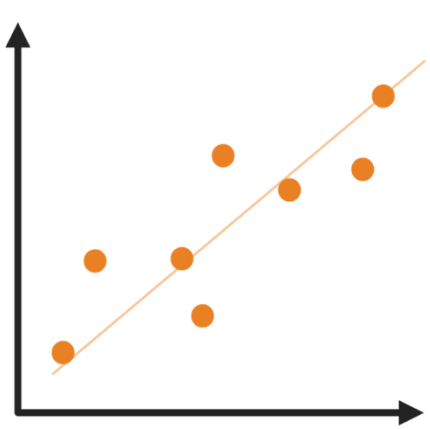
Each genetic variant is neither necessary nor sufficient

Association study
(unrelated population)

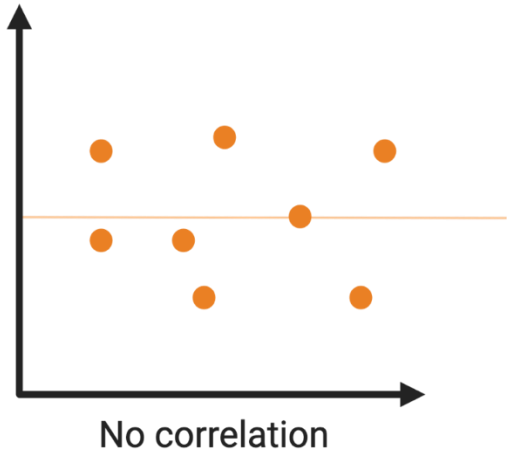
? What does association mean

ASSOCIATION

An association defines a relationship between two entity objects based on common attributes.



High positive correlation



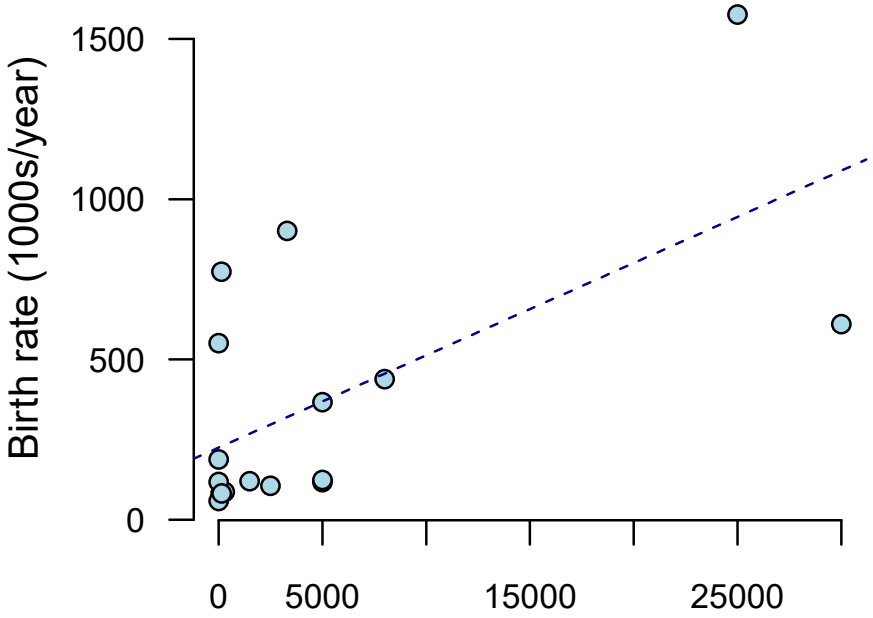
No correlation

Is there an association between exposure and outcome?

ASSOCIATION *NOT CAUSATION*

Country	Area (km^2)	Storks (pairs)	Humans (10^6)	Birth rate (10^3/yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	38	610
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

Table 1. Geographic, human and stork data for 17 European countries

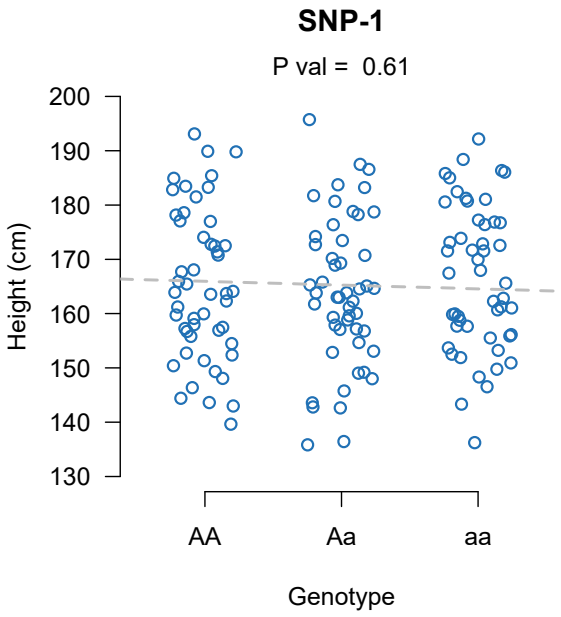



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.250e+02  9.356e+01   2.405   0.0295 *
Storks      2.879e-02  9.402e-03   3.063   0.0079 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

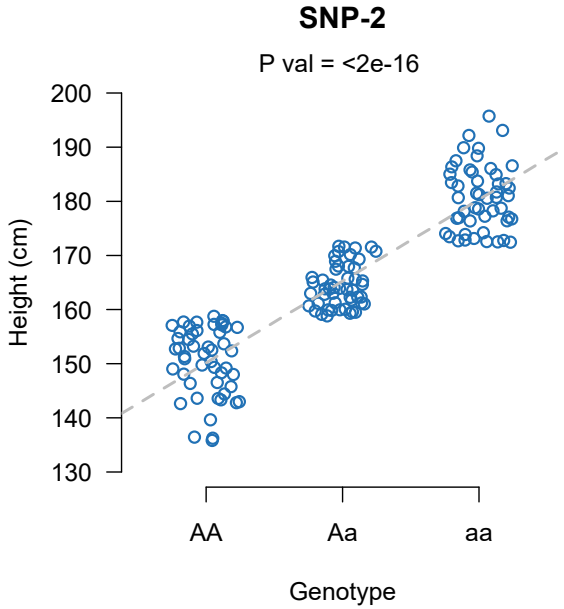

Matthews, R. (2000), Storks Deliver Babies (p= 0.008). Teaching Statistics, 22: 36-38.

IDENTIFY RISK VARIANTS

GENOME-WIDE ASSOCIATION STUDY (GWAS)

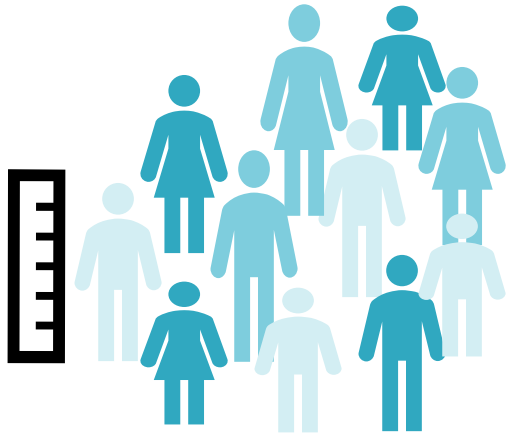


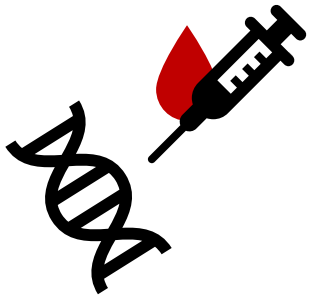
No association

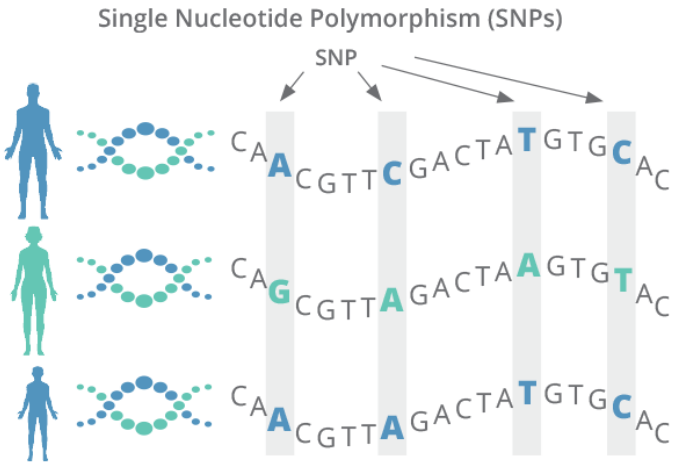


Association

Systematic hypothesis-free scanning of all
common genetic variants



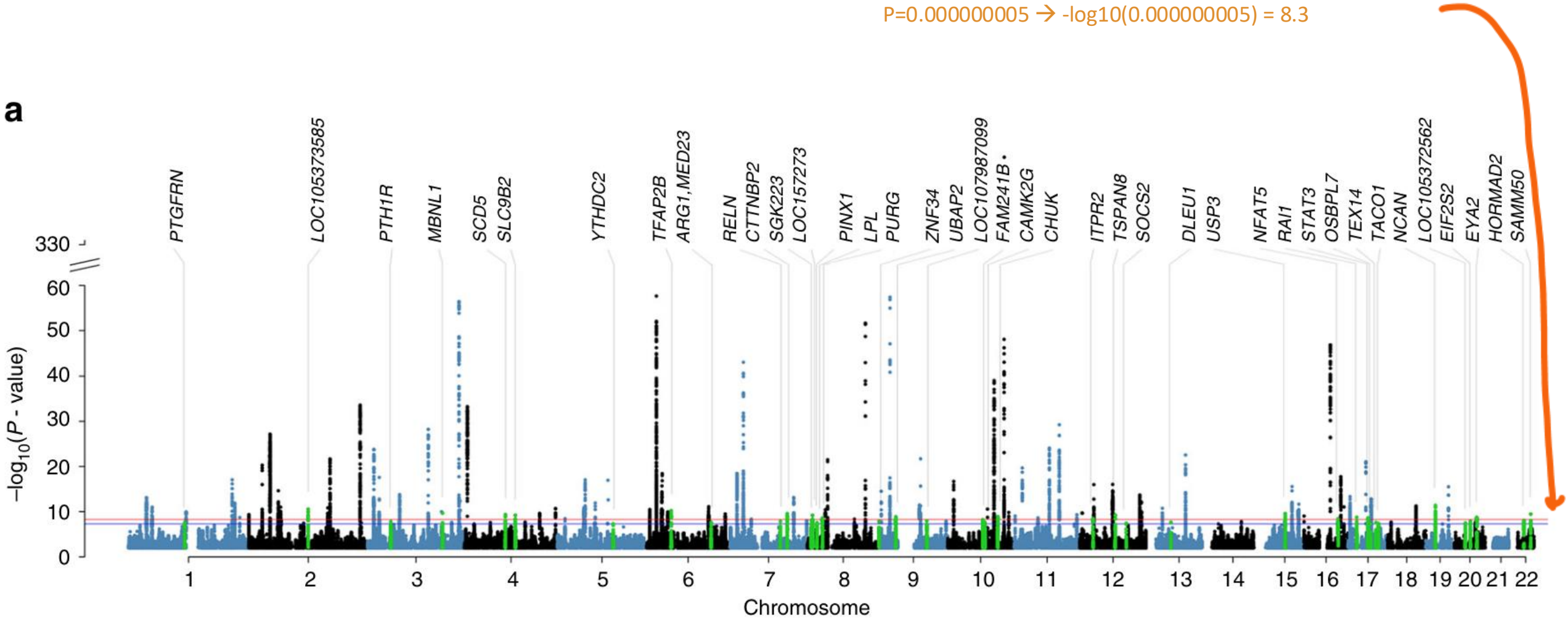




MANHATTAN PLOT

For each variant, plot the –log10(P-value) as function of chromosomal position.

P=0.05 → -log10(0.05) = 1.3
P=0.001 → -log10(0.001) = 3
P=0.000000005 → -log10(0.000000005) = 8.3



HYPOTHESIS TESTING

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis (H_0)	Type I error α **False positive**	Correct outcome **True positive**
Accecpt null hypothesis (H_0)	Correct outcome **True negative**	Type II error β **False negative**

The probability (*P*) of making a type I error is denoted by α; we reject the null hypothesis if the inferred *P* value is less than the significance level (α=0.05). I.e., the probability of rejecting the null hypothesis when should be accepted.

Why multiple testing correction? If we test 500,000 SNPs, then by chance we expect 25.000 SNPs to be significant (if α=0.05) → i.e., **25.000 false-positive associations**.

One solution is to correct for number of tests performed; Bonferroni correction;

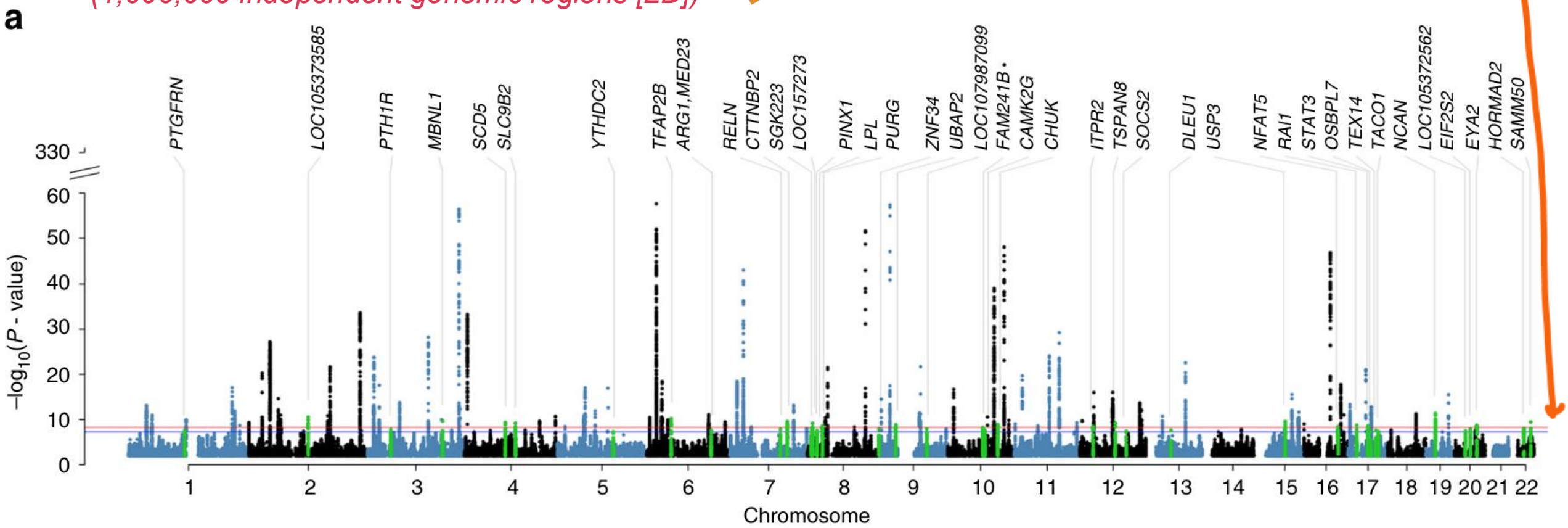
Corrected *P*-value = $P \times n_{tests} \leq 0.05$ OR $\frac{\alpha}{n_{tests}}$ =**new significance threshold**

MANHATTAN PLOT

For each variant, plot the –log10(P-value) as function of chromosomal position.

P=0.05 → -log10(0.05) = 1.3
P=0.001 → -log10(0.001) = 3
P=0.000000005 → -log10(0.000000005) = 8.3

Genome-wide significance level?
-adjust for no. of independent statistical tests (1,000,000 independent genomic regions [LD])

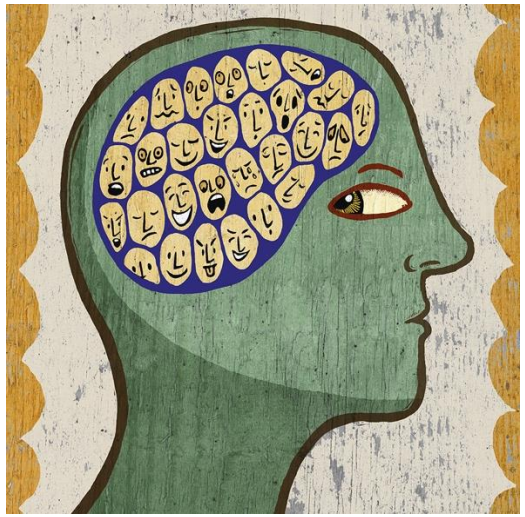


POWER IS EVERYTHING IN GWAS

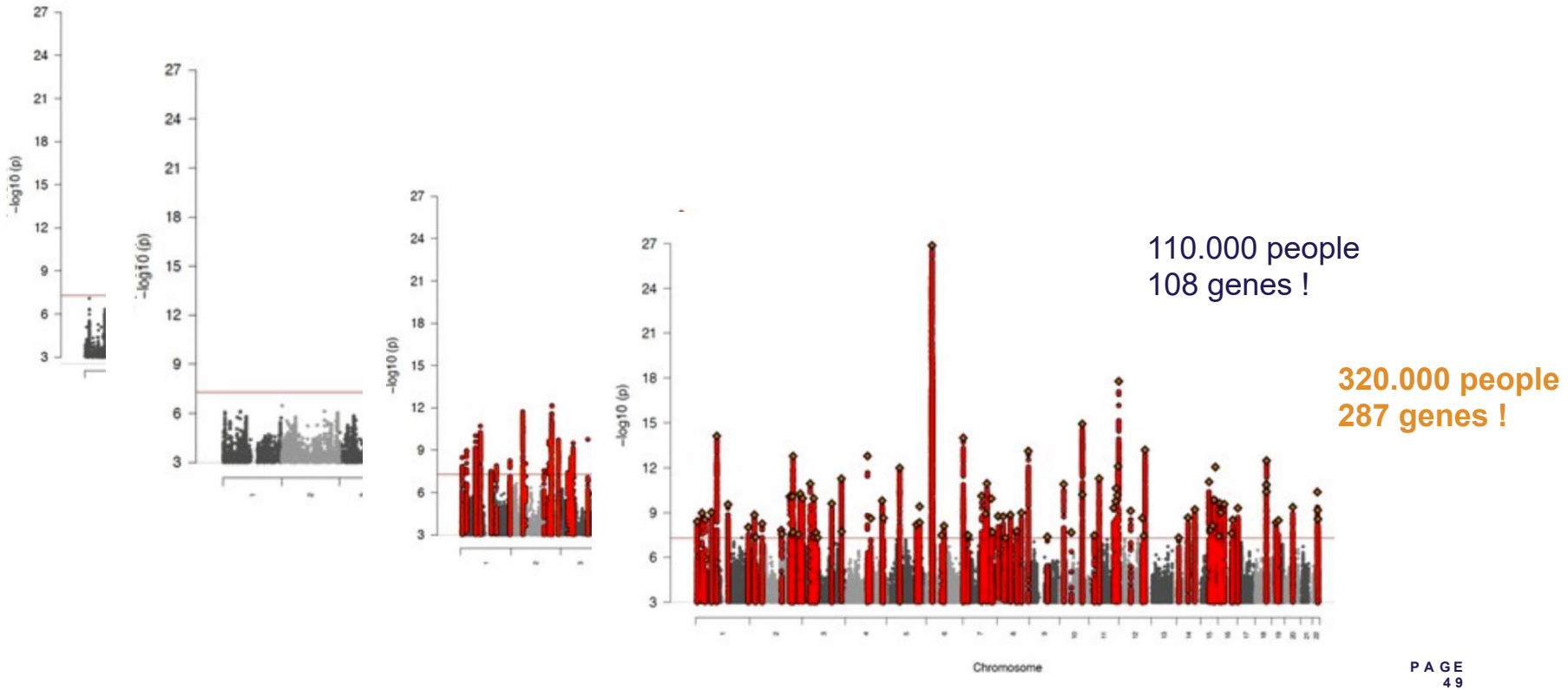
the probability of detecting an effect, if there is a true effect present to detect

depends among otherthings on sample size

Schizophrenia



Has high heritabilitet (h^2= 0.85)
The population prevalence is 1%
Emerge in late teens
Molecular aetiology is unknown

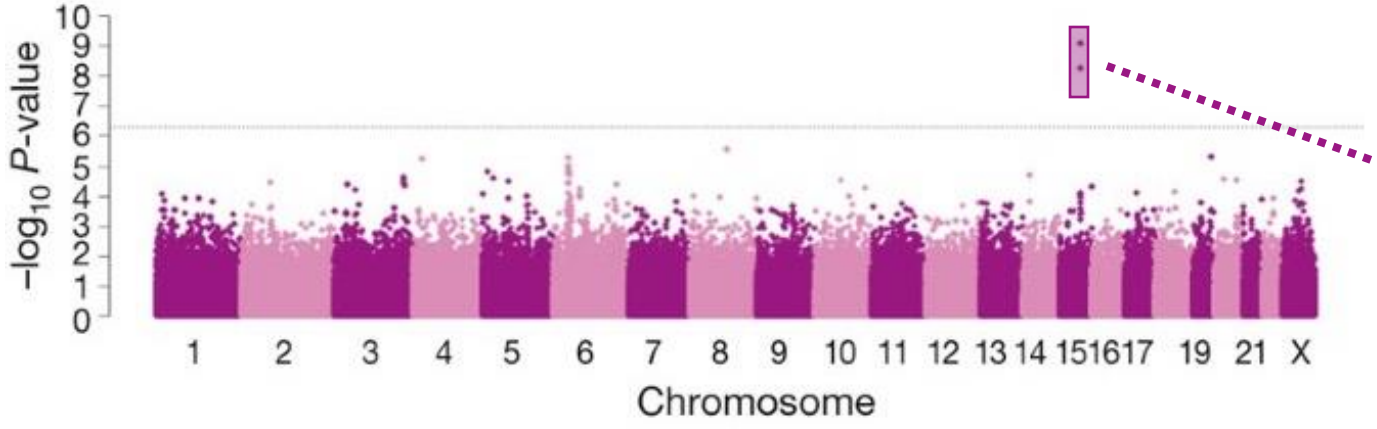


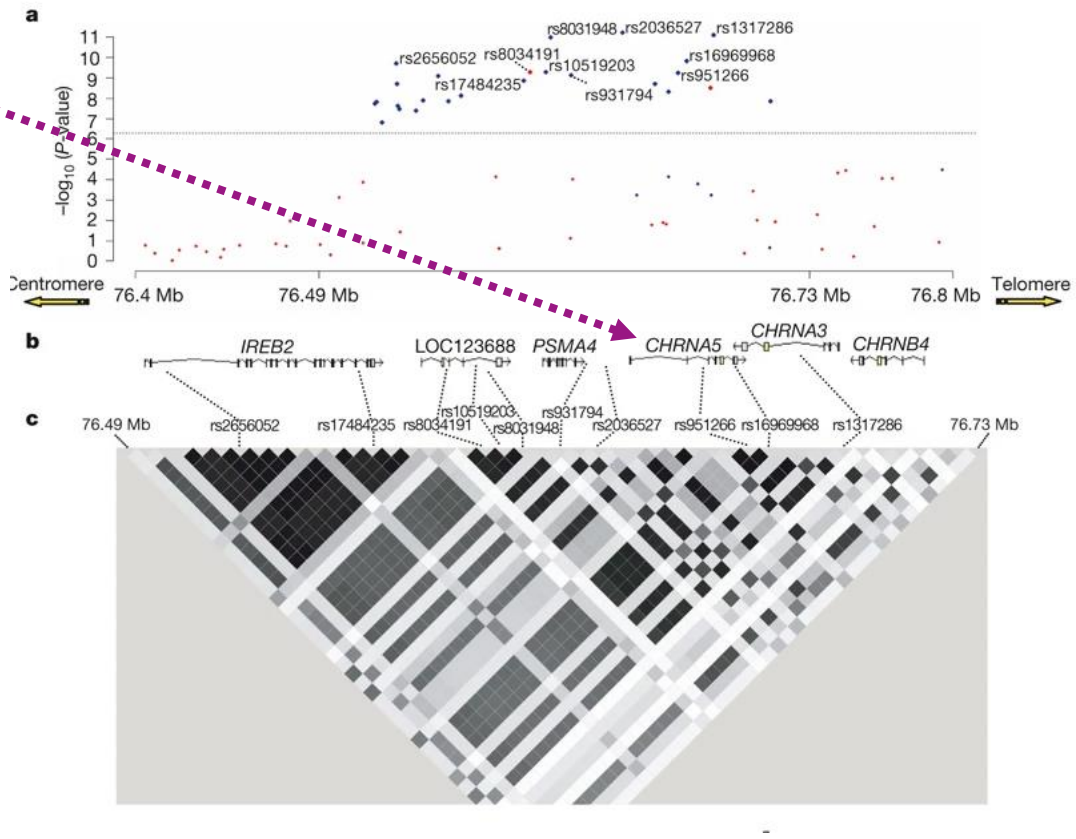
GENETIC RISK FOR LUNG CANCER?

Vol 452|3 April 2008|doi:10.1038/nature06885 nature

LETTERS

A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25





A genetic studie of lung cancer (LC, 1989 cases and 2625 controls) found a nicotine receptor (*CHRNA3/5*) to be associated with the risk of developing lung cancer.

Does that then mean that *CHRNA3/5* is a risk loci for LC?

Risk loci for nicotine addition → addicted to smoking → increase LC risk

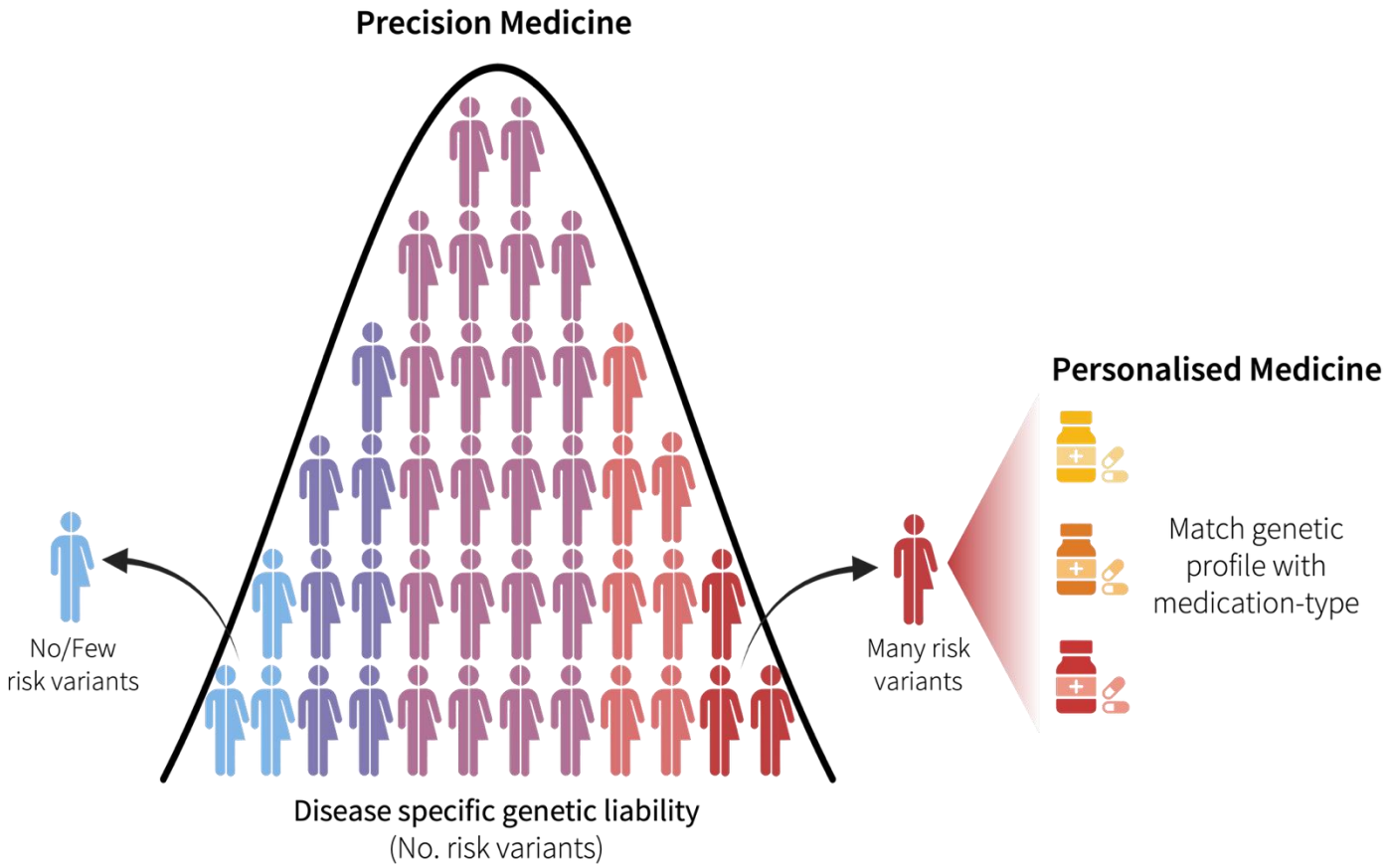
PREDICTING DISEASE RISK FROM GENETIC DATA

A “polygenic score” is one way by which people can learn about their risk of developing a disease, based on the total number of changes (i.e., SNPs) related to the disease (NHI)

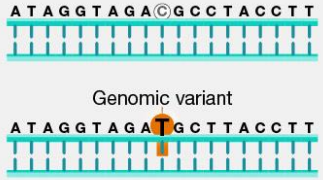
DIFFERENT NAMES

BUT THE SAME

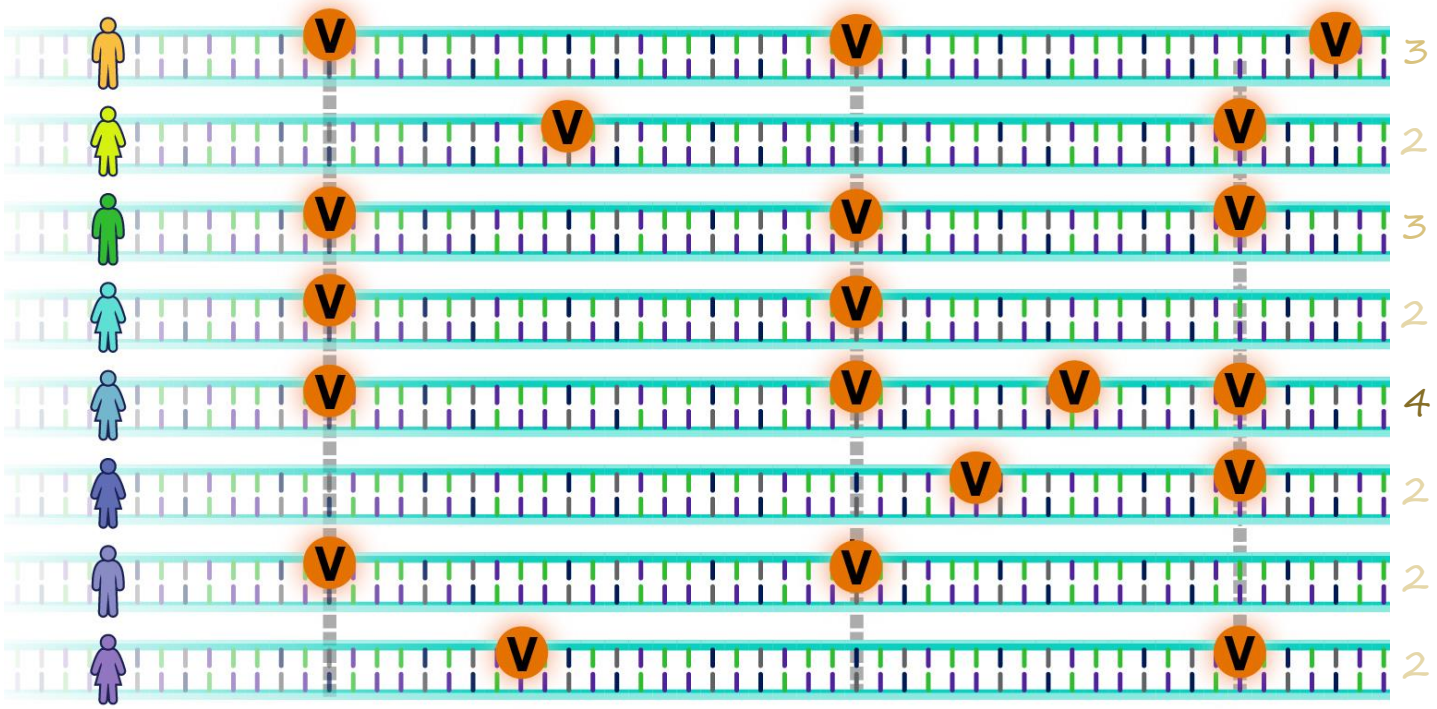
- Polygenic risk score (PRS)
- Polygenic score (PGS)
- Genetic score (GS)
- Genetic risk score (GRS)
- Genetic value
- Genetic liability
- Breeding value
- …



THE INHERENT DISEASE RISK



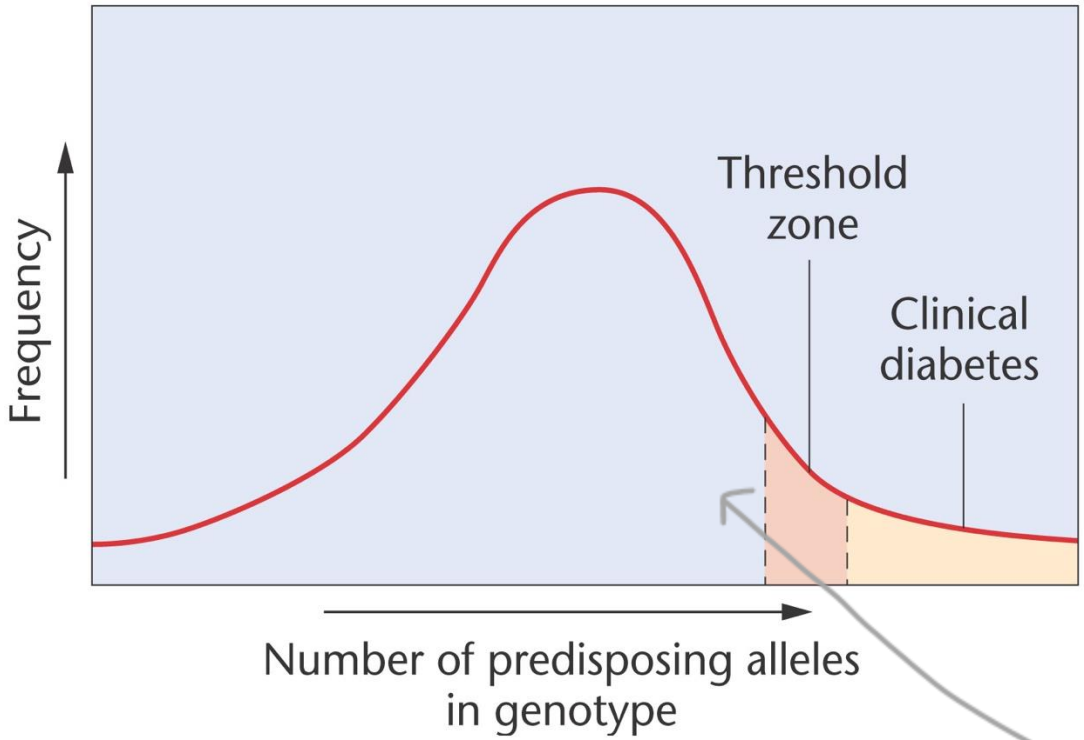
THE INHERENT DISEASE RISK



Variation in the number of risk variants
→
Polygenic score (PGS)

WHAT IS A PGS?

A RELATIVE RISK



SNP1 A G

SNP2 T T

SNP3 A C

SNP4 T A

SNP5 C C

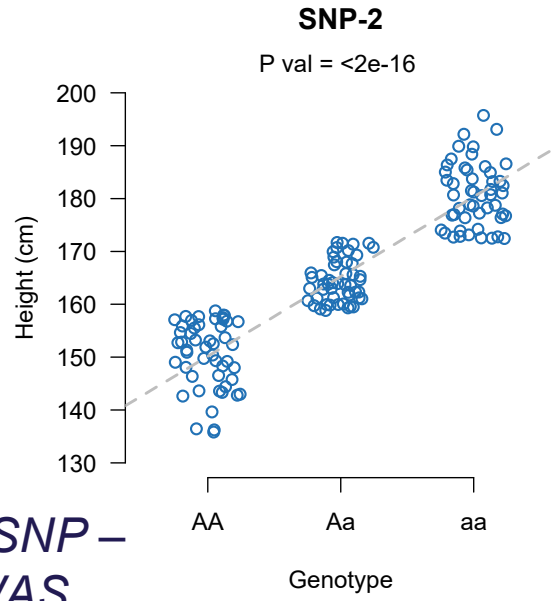
PGS =7

How do we find the 'orange' alleles?

WHAT IS A PGS?

"A PGS combines information from large numbers of markers across the genome (hundreds to millions) to give a single numerical score for an individual's risk for developing a specific disease on the basis of the DNA variants they have inherited."

b is the slope (effect size) from reggresion



The effect size of the SNP – obtain from the GWAS

$$PGS = \sum X_i b_i$$

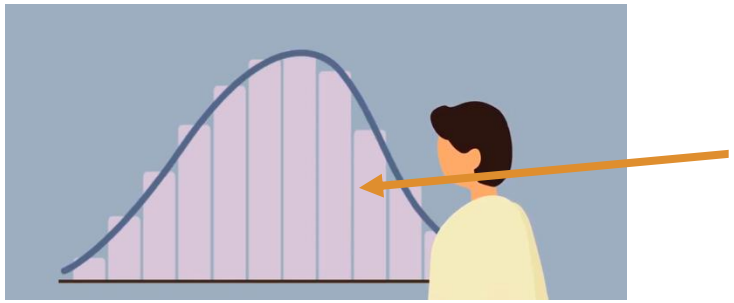
The genotype of the individual for SNP i (0, 1, 2 – counting the number of the alternative allele)

AA = 0
Aa = 1
aa = 2

$$PGS = \sum X_i b_i$$

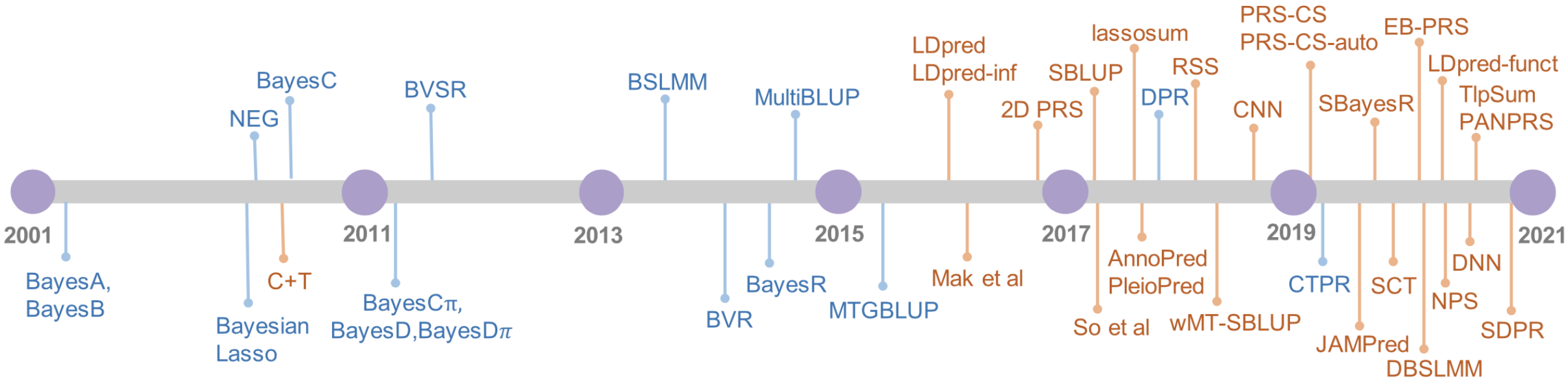
HOW TO COMPUTE A (simple) PGS?

SNPs	Adams Genotypes	Ref allele	Alt allele
SNP-1	TC	T	C
SNP-2	GG	G	T
SNP-3	CC	A	C
SNP-4	TG	T	G
SNP-5	AA	A	G



A LARGE PALETTE OF PGS METHODS

$$PGS = \sum X_i b_i$$



Ma, Y., & Zhou, X. (2021). Trends in Genetics, 37(11), 995–1011.

WHY DIFFERENT PGS SCORING METHODS?

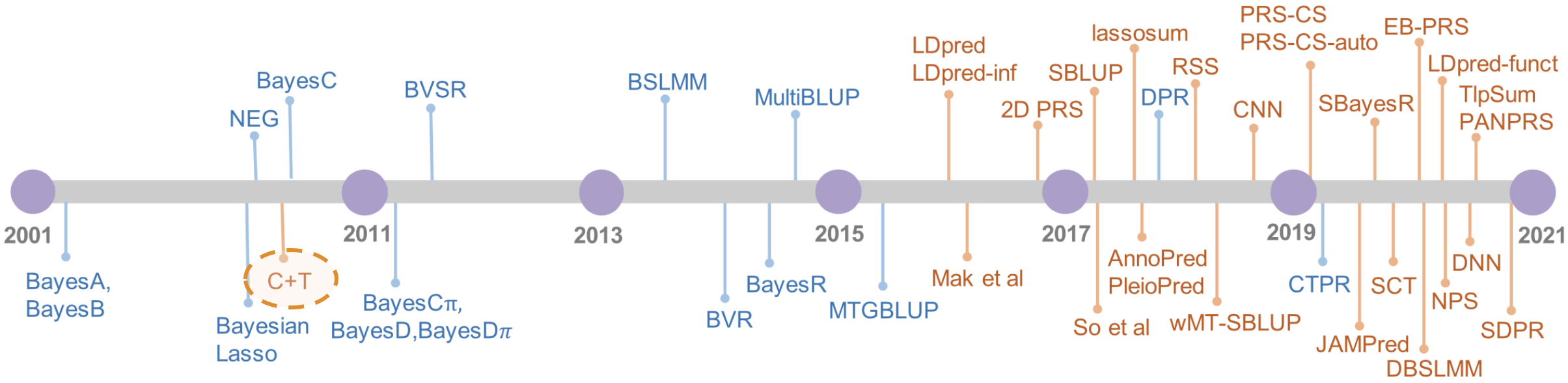
Complex traits have different underlying genetic architectures

- some are influenced by <100 genetic loci
- some are influenced by >1000 genetic loci
- some loci have very small effects
- some loci have moderate effects
- correlation structure among loci (linkage disequilibrium)

The different scoring algorithms attempts to account for this.

A LARGE PALETTE OF PGS METHODS

$$PGS = \sum X_i b_i$$



Ma, Y., & Zhou, X. (2021). Trends in Genetics, 37(11), 995–1011.

CLUMPING AND THRESHOLDING (C+T)

0: Set LD (=0.8) and *P* values (0.01)

SNP	b	p
1	0.21	0.005
2	0.22	0.0048
3	0.25	0.0003
4	0.1	0.04
5	0.05	0.15
6	0.02	0.49
7	0.03	0.87
8	0.12	0.003
9	0.14	0.0034
10	0.18	0.0004
11	0.21	0.00003
12	0.12	0.15
13	0.14	0.12
14	0.03	0.84
15	0.02	0.32

1: Sort by *P*-value

SNP	b	p
11	0.21	0.00003
3	0.25	0.0003
10	0.18	0.0004
8	0.12	0.003
9	0.14	0.0034
2	0.22	0.0048
1	0.21	0.005
4	0.1	0.04
13	0.14	0.12
5	0.05	0.15
12	0.12	0.15
15	0.02	0.32
6	0.02	0.49
14	0.03	0.84
7	0.03	0.87

2: Compute LD and select variants based of thresholds

SNP	b	p	r^2
11	0.21	0.00003	
3	0.25	0.0003	0.96
10	0.18	0.0004	0.93
8	0.12	0.003	0.88
9	0.14	0.0034	0.74
2	0.22	0.0048	0.4
1	0.21	0.005	0.03
4	0.1	0.04	0.04
13	0.14	0.12	0.05
5	0.05	0.15	0.03
12	0.12	0.15	0.04
15	0.02	0.32	0.01
6	0.02	0.49	0.01
14	0.03	0.84	0.01
7	0.03	0.87	0.01

1st variant in LD-pair

Have LD>r^2 – ignore those

CLUMPING AND THRESHOLDING (C+T)

0: Set LD (=0.8) and *P* values (0.01)

SNP	b	p
1	0.21	0.005
2	0.22	0.0048
3	0.25	0.0003
4	0.1	0.04
5	0.05	0.15
6	0.02	0.49
7	0.03	0.87
8	0.12	0.003
9	0.14	0.0034
10	0.18	0.0004
11	0.21	0.00003
12	0.12	0.15
13	0.14	0.12
14	0.03	0.84
15	0.02	0.32

1: Sort by P-value

SNP	b	p
11	0.21	0.00003
3	0.25	0.0003
10	0.18	0.0004
8	0.12	0.003
9	0.14	0.0034
2	0.22	0.0048
1	0.21	0.005
4	0.1	0.04
13	0.14	0.12
5	0.05	0.15
12	0.12	0.15
15	0.02	0.32
6	0.02	0.49
14	0.03	0.84
7	0.03	0.87

2: Compute LD and select variants based of thresholds

SNP	b	p	r^2
11	0.21	0.00003	
3	0.25	0.0003	
10	0.18	0.0004	
8	0.12	0.003	
9	0.14	0.0034	
2	0.22	0.0048	0.98
1	0.21	0.005	0.96
4	0.1	0.04	0.96
13	0.14	0.12	0.52
5	0.05	0.15	0.34
12	0.12	0.15	0.10
15	0.02	0.32	0.04
6	0.02	0.49	0.01
14	0.03	0.84	0.01
7	0.03	0.87	0.01

1st variant in LD-pair

Have LD>r^2 – ignore those

CLUMPING AND THRESHOLDING (C+T)

0: Set LD (=0.8) and *P* values (0.01)

SNP	b	p
1	0.21	0.005
2	0.22	0.0048
3	0.25	0.0003
4	0.1	0.04
5	0.05	0.15
6	0.02	0.49
7	0.03	0.87
8	0.12	0.003
9	0.14	0.0034
10	0.18	0.0004
11	0.21	0.00003
12	0.12	0.15
13	0.14	0.12
14	0.03	0.84
15	0.02	0.32

1: Sort by P-value

SNP	b	p
11	0.21	0.00003
3	0.25	0.0003
10	0.18	0.0004
8	0.12	0.003
9	0.14	0.0034
2	0.22	0.0048
1	0.21	0.005
4	0.1	0.04
13	0.14	0.12
5	0.05	0.15
12	0.12	0.15
15	0.02	0.32
6	0.02	0.49
14	0.03	0.84
7	0.03	0.87

2: Compute LD and select variants based of thresholds

SNP	b	p	r^2
11	0.21	0.00003	
3	0.25	0.0003	
10	0.18	0.0004	
8	0.12	0.003	
9	0.14	0.0034	
2	0.22	0.0048	
1	0.21	0.005	
4	0.1	0.04	
13	0.14	0.12	
5	0.05	0.15	0.86
12	0.12	0.15	0.82
15	0.02	0.32	0.81
6	0.02	0.49	0.85
14	0.03	0.84	0.85
7	0.03	0.87	0.81

1st variant in LD-pair

Have LD>r^2 – ignore those

CLUMPING AND THRESHOLDING (C+T)

0: Set LD (=0.8) and *P* values (0.01)

SNP	b	p
1	0.21	0.005
2	0.22	0.0048
3	0.25	0.0003
4	0.1	0.04
5	0.05	0.15
6	0.02	0.49
7	0.03	0.87
8	0.12	0.003
9	0.14	0.0034
10	0.18	0.0004
11	0.21	0.00003
12	0.12	0.15
13	0.14	0.12
14	0.03	0.84
15	0.02	0.32

1: Sort by P-value

SNP	b	p
11	0.21	0.00003
3	0.25	0.0003
10	0.18	0.0004
8	0.12	0.003
9	0.14	0.0034
2	0.22	0.0048
1	0.21	0.005
4	0.1	0.04
13	0.14	0.12
5	0.05	0.15
12	0.12	0.15
15	0.02	0.32
6	0.02	0.49
14	0.03	0.84
7	0.03	0.87

2: Compute LD and select variants based on LD

SNP	b	p	r^2
11	0.21	0.00003	
3	0.25	0.0003	
10	0.18	0.0004	
8	0.12	0.003	
9	0.14	0.0034	
2	0.22	0.0048	
1	0.21	0.005	
4	0.1	0.04	
13	0.14	0.12	
5	0.05	0.15	
12	0.12	0.15	
15	0.02	0.32	
6	0.02	0.49	
14	0.03	0.84	
7	0.03	0.87	

3: Compute PGS based on effect sizes (b) and *P*-values

$$PGS = \sum X_i b_i$$

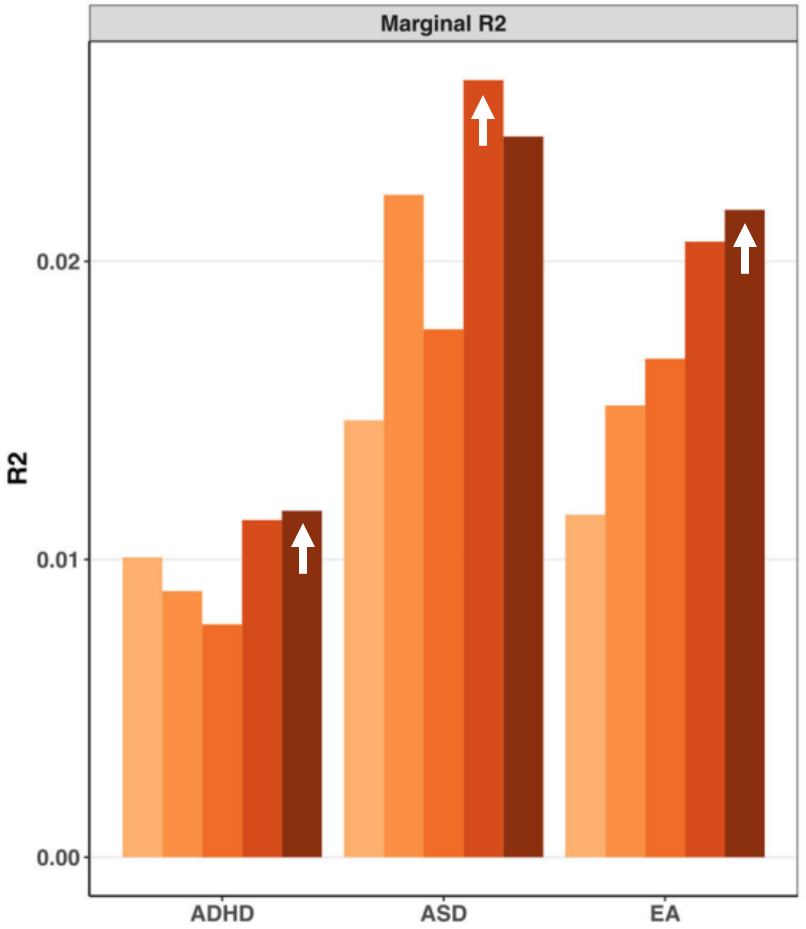
$$= \boldsymbol{X}_{11} \times 0.21 + \boldsymbol{X}_{9} \times 0.14$$

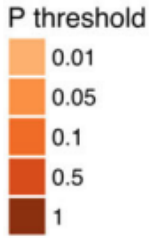
CLUMPING AND THRESHOLDING (C+T)

Repeat for other *P*-value cutoffs (and LD values)

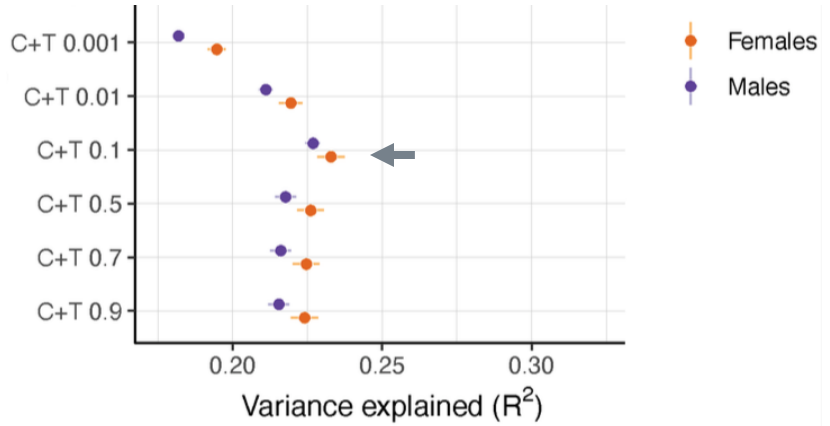
How does the PGS associate with the disease

$$y_{trait} = PGS + \varepsilon$$





Finding optimal r^2 and *P*-cutoff and apply in second cohort



CLUMPING AND THRESHOLDING (C+T)

Clumping and Thresholding (C+T) and Stacked C+T (SCT)

We compute C+T scores for each chromosome separately and for several parameters:

- Threshold on imputation INFO score INFO_T within {0.3, 0.6, 0.9, 0.95}.
- Squared correlation threshold of clumping r_c^2 within {0.01, 0.05, 0.1, 0.2, 0.5, 0.8, 0.95}.
- Base size of clumping window within {50, 100, 200, 500}. The window size w_c is then computed as the base size divided by r_c^2. For example, for $r_c^2 = 0.2$, we test values of w_c within {250, 500, 1000, 2500} (in kb). This is motivated by the fact that linkage disequilibrium is inversely proportional to genetic distance between variants.[11]
- A sequence of 50 thresholds on p values between the least and the most significant p values, equally spaced on a log-log scale.

Thus, for individual i, chromosome k, and the four hyper-parameters INFO_T, r_c^2, w_c, and p_T, we compute C+T predictions

$$V_i^{(k)}(\text{INFO}_T, r_c^2, w_c, p_T) = \sum_{\substack{j \in S_{\text{clumping}}(k, \text{INFO}_T, r_c^2, w_c) \\ p_j < p_T}} \widehat{\beta}_j \cdot G_{i,j} \ ,$$

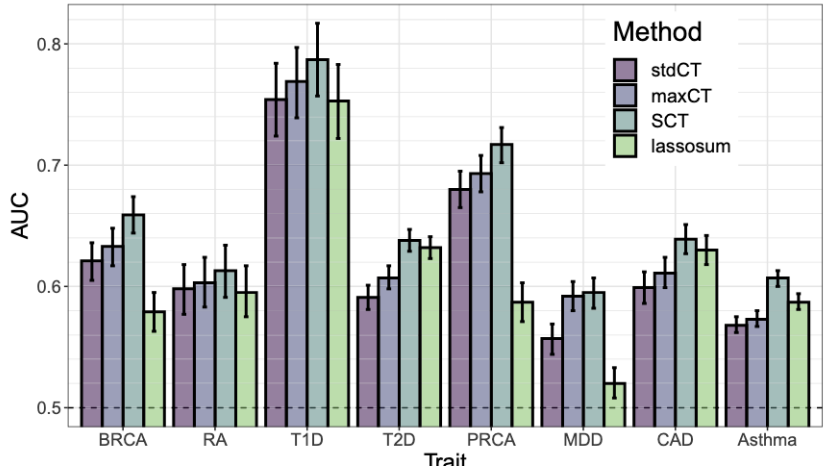


Table 2. Optimal Choices of C+T Parameters

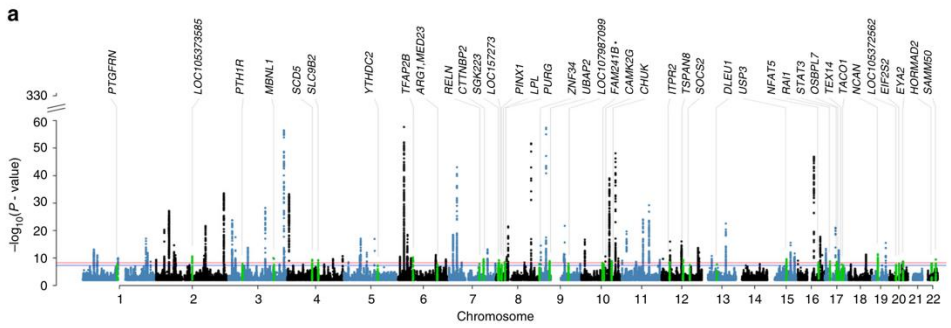
Trait	w_c	r_c^2	INFO_T	p_T
Breast cancer (BRCA)	2,500	0.2	0.95	2.2×10^{-4}
Rheumatoid arthritis (RA)	200	0.5	0.95	7.5×10^{-2}
Type 1 diabetes (T1D)	10K–50K	0.01	0.90	2.6×10^{-5}
Type 2 diabetes (T2D)	625	0.8	0.95	1.1×10^{-2}
Prostate cancer (PRCA)	10K–50K	0.01	0.90	4.2×10^{-6}
Depression (MDD)	625	0.8	0.95	1.0×10^{-1}
Coronary artery disease (CAD)	526	0.95	0.95	3.5×10^{-2}
Asthma	2,500	0.2	0.90	2.2×10^{-4}

Choice of C+T parameters is based on the maximum AUC in the training set. Hyper-parameters of C+T are the squared correlation threshold r_c^2 and the window size w_c of clumping, the p value threshold p_T and the threshold on the quality of imputation INFO_T. Choosing optimal hyper-parameters for C+T use 63%–90% of the individuals reported in Table 1. Resulting predictions of maxCT in the test set are reported in Figure 2.

Prive et al (2019) Am J Hum Gen, 105:1213–1221

WHAT DO YOU NEED?

1. A large well-powered GWAS for your trait of interest



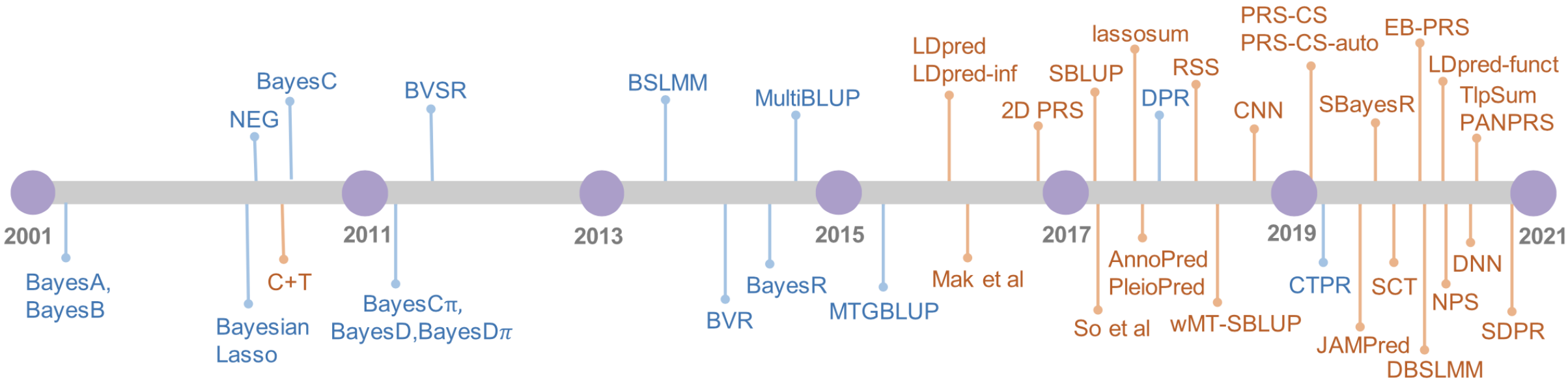
2. An independent cohort that has been genotyped

(3. That some individuals in the cohort has the phenotype)

A LARGE PALETTE OF PGS METHODS

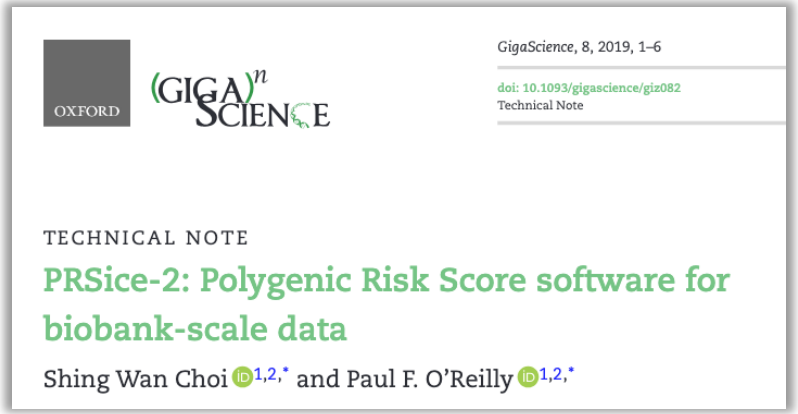
-- Require individual level data --
(can be tricky because of GDPR)

-- Require summary data --



Ma, Y., & Zhou, X. (2021). Trends in Genetics, 37(11), 995–1011.

COMMONLY USED METHODS



OXFORD (GIGA)n SCIENCE

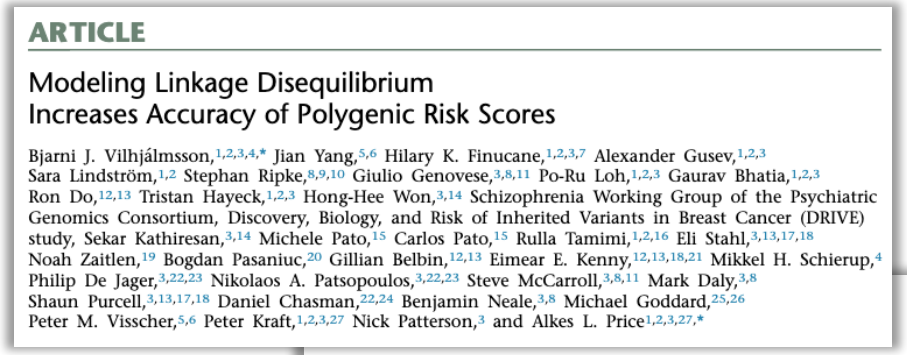
GigaScience, 8, 2019, 1–6

doi: 10.1093/gigascience/giz082
Technical Note

TECHNICAL NOTE

PRSice-2: Polygenic Risk Score software for biobank-scale data

Shing Wan Choi [1,2,*] and Paul F. O'Reilly [1,2,*]



ARTICLE

Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores

Bjarni J. Vilhjálmsson,[1,2,3,4,*] Jian Yang,[5,6] Hilary K. Finucane,[1,2,3,7] Alexander Gusev,[1,2,3] Sara Lindström,[1,2] Stephan Ripke,[8,9,10] Giulio Genovese,[3,8,11] Po-Ru Loh,[1,2,3] Gaurav Bhatia,[1,2,3] Ron Do,[12,13] Tristan Hayeck,[1,2,3] Hong-Hee Won,[3,14] Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study, Sekar Kathiresan,[3,14] Michele Pato,[15] Carlos Pato,[15] Rulla Tamimi,[1,2,16] Eli Stahl,[3,13,17,18] Noah Zaitlen,[19] Bogdan Pasaniuc,[20] Gillian Belbin,[12,13] Eimear E. Kenny,[12,13,18,21] Mikkel H. Schierup,[4] Philip De Jager,[3,22,23] Nikolaos A. Patsopoulos,[3,22,23] Steve McCarroll,[3,8,11] Mark Daly,[3,8] Shaun Purcell,[3,13,17,18] Daniel Chasman,[22,24] Benjamin Neale,[3,8] Michael Goddard,[25,26] Peter M. Visscher,[5,6] Peter Kraft,[1,2,3,27] Nick Patterson,[3] and Alkes L. Price[1,2,3,27,*]

Bioinformatics, 36(22-23), 2020, 5424–5431
doi: 10.1093/bioinformatics/btaa1029
Advance Access Publication Date: 16 December 2020
Original Paper

OXFORD

Genetics and population analysis

LDpred2: better, faster, stronger

Florian Privé[1,*], Julyan Arbel[2] and Bjarni J. Vilhjálmsson[1,3,*]

[1]National Centre for Register-Based Research, Aarhus University, Aarhus 8210, Denmark, [2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble 38000, France and [3]Bioinformatics Research Centre, Aarhus University, Aarhus 8000, Denmark

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-09718-5 OPEN

Polygenic prediction via Bayesian regression and continuous shrinkage priors

Tian Ge[1,2,3], Chia-Yen Chen[1,2,3,4], Yang Ni[5], Yen-Chen Anne Feng[1,2,3,4] & Jordan W. Smoller[1,2,3]

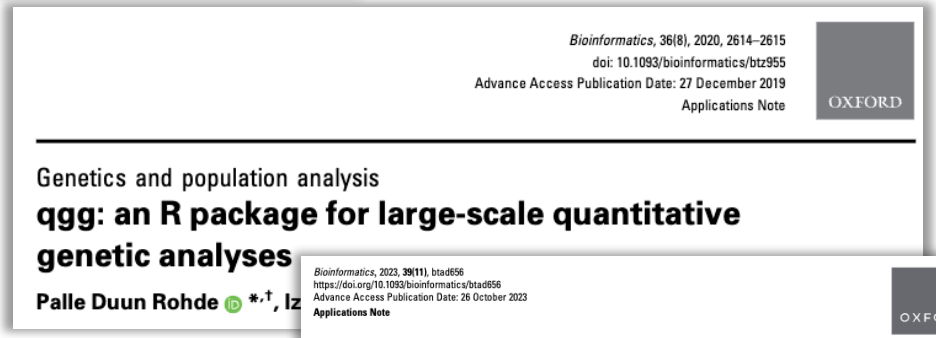
nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-12653-0 OPEN

Improved polygenic prediction by Bayesian multiple regression on summary statistics

Luke R. Lloyd-Jones[1,9*], Jian Zeng[1,9*], Julia Sidorenko[1,2], Loïc Yengo[1], Gerhard Moser[3,4], Kathryn E. Kemper[1], Huanwei Wang[1], Zhili Zheng[1], Reedik Magi[2], Tõnu Esko[2], Andres Metspalu[2,5], Naomi R. Wray[1,6], Michael E. Goddard[7], Jian Yang[1,8*] & Peter M. Visscher[1*]



Bioinformatics, 36(8), 2020, 2614–2615
doi: 10.1093/bioinformatics/btz955
Advance Access Publication Date: 27 December 2019
Applications Note

OXFORD

Genetics and population analysis

qgg: an R package for large-scale quantitative genetic analyses

Palle Duun Rohde [*,†], Iz

Bioinformatics, 2023, **39(11)**, btad656
https://doi.org/10.1093/bioinformatics/btad656
Advance Access Publication Date: 26 October 2023
Applications Note

OXFORD

Genetics and population analysis

Expanded utility of the R package, qgg, with applications within genomic medicine

Palle Duun Rohde [1,*], Izel Fourie Sørensen[2], Peter Sørensen[2,*]

[1]Genomic Medicine, Department of Health Science and Technology, Aalborg University, 9260 Gistrup, Denmark
[2]Center for Quantitative Genetics and Genomics, Aarhus University, 8000 Aarhus, Denmark
*Corresponding authors. Genomic Medicine, Department of Health Science and Technology, Aalborg University, Selma Langerløfs Vej 249, 9260 Gistrup, Denmark. E-mail: palledr@hst.aau.dk (P.D.R.); Center for Quantitative Genetics and Genomics, Aarhus University, C. F. Mallers Allé 3, 8000 Aarhus, Denmark. E-mail: pso@qgg.au.dk (P.S.)

Associate Editor: Christina Kendziorski

AALBORG UNIVERSITY

SESSION 2

- **Genetic associations**
  - *By linking specific genetic variants to traits or diseases, we can identify biological pathways and potential drug targets.*
- **Polygenic score**
  - *Combining information from thousands of variants allows us to estimate an individual's genetic risk — a key step toward personalised prevention and treatment.*

GROUP WORK

1

2

3

AGENDA

09:00 – 09:15	Welcome
09:15 – 09:45	Session 1: Genetic variation and heredity
09:45 – 10:15	Group work 1: Genetic variation and heredity
10:15 – 10:30	Break
10:30 – 11:00	Session 2: From SNP to GWAS to disease risk
11:00 – 11:40	Group work 2: Construct your own PGS
11:40 – 12:00	Common discussion and evaluation

Join at menti.com | use code 7798 1843

Hvilken indsigt tager du med dig fra i dag?

No questions from the audience!

Incoming questions will show up here so that you can answer them one by one.

10

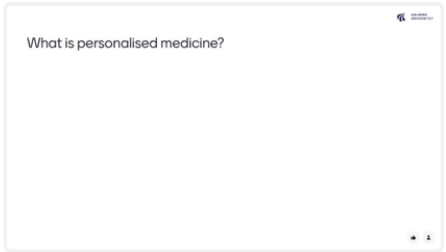
PD

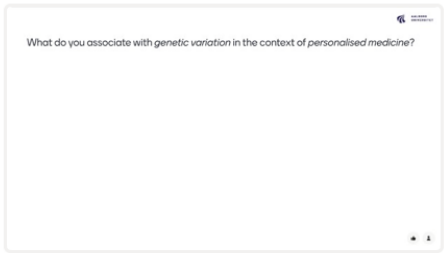
AALBORG UNIVERSITET

Menti

PM-genetik2025

Choose a slide to present





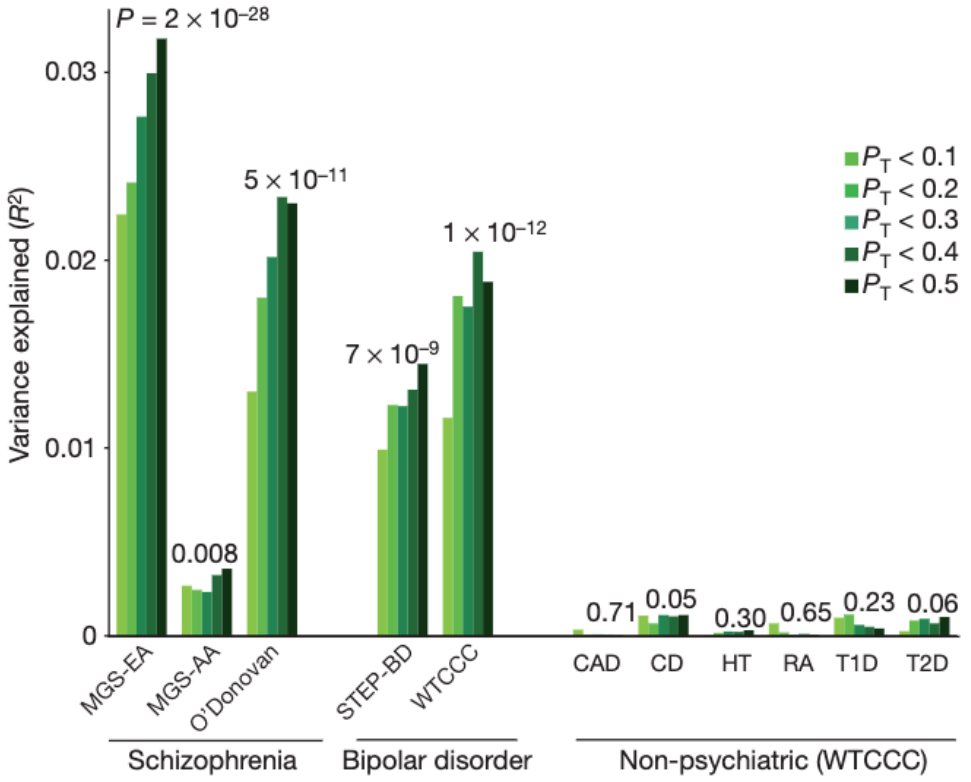
EVALUATE POLYGENIC PROFILES

$$y = Xb + Zc + e$$

$$y = \text{phenotype}; X = \text{PGS}; Z = \text{covariates}$$

Compare variance explained from the full model (with X+covariates) compared to a reduced model (covariates only)

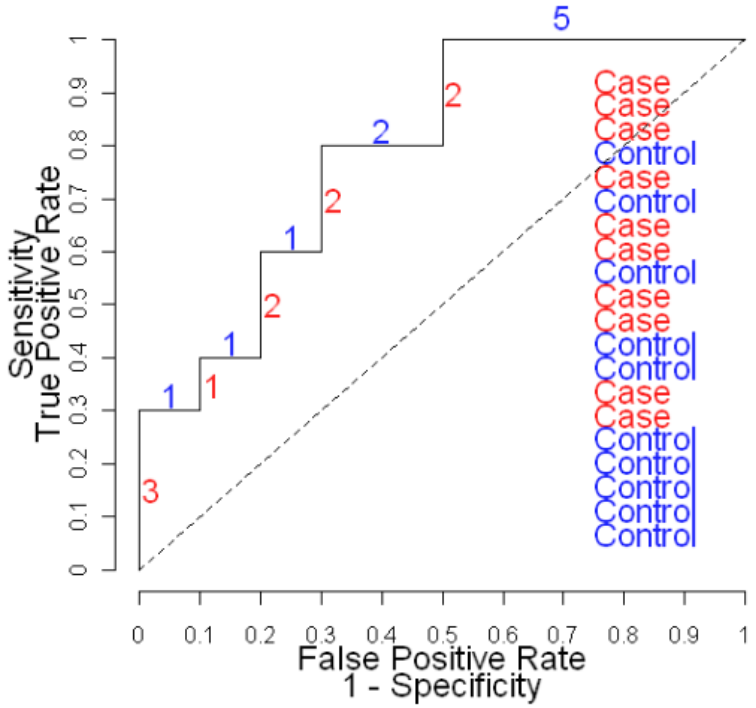
Variance explained (R^2) for quantitative traits, and Nagelkerke's R^2 for binary traits (**however, NagR2 is biased with disease prevalence!**)



The International Schizophrenia Consortium (2009) Nature, 460

EVALUATE POLYGENIC PROFILES

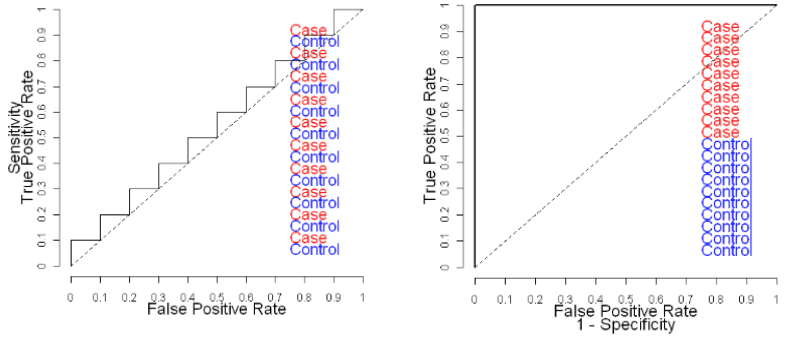
Rank individuals on score from highest ranked to lowest



Area Under Receiver Operator Characteristic Curve (AUC)

Well established measure of validity of tests for classifier diseased vs non-diseased individuals

- Nice property – independent to proportion of cases and controls in sample
- Range 0.5 to 1
- 0.5 the score has no predictive value
- **Probability that a randomly selected case has a score higher than a randomly selected control**



From Wray N

EVALUATE POLYGENIC PROFILES

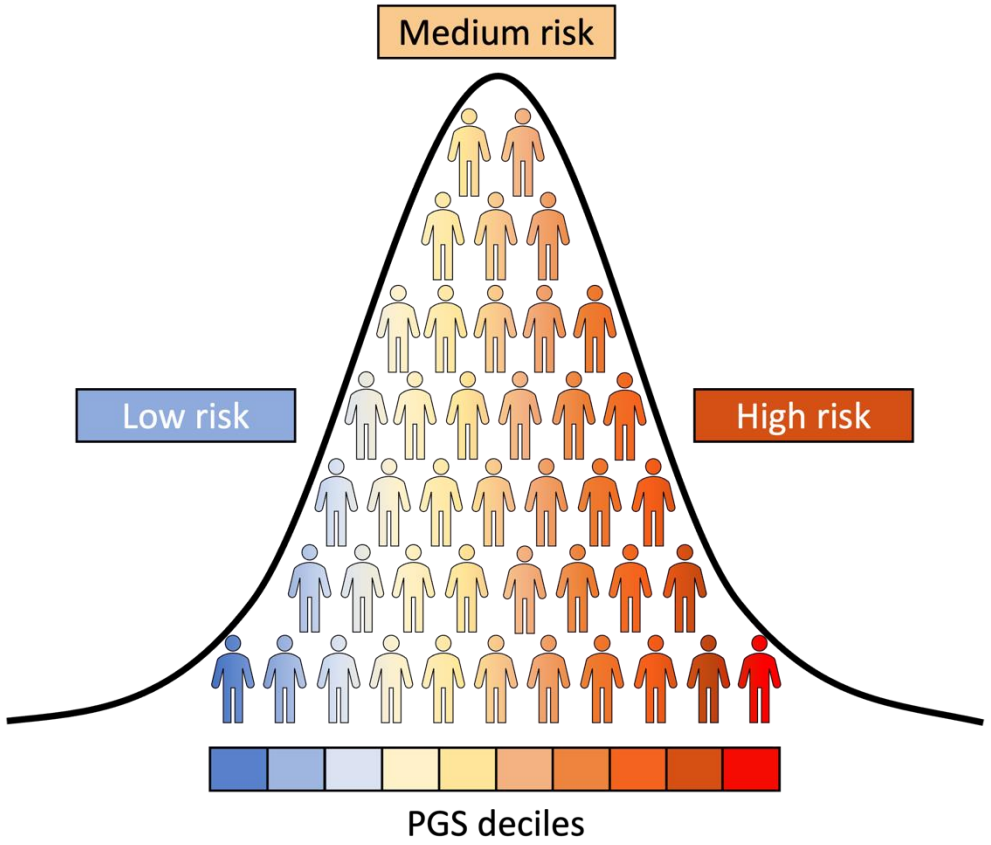
Odds ratio (OR)

Cut the distribution into deciles

Each decile will include both cases and controls

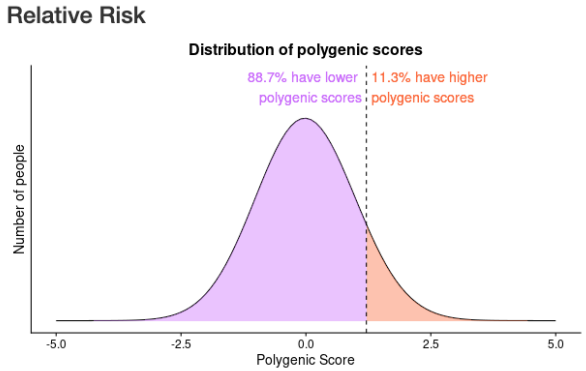
Odds of being a case in each decile

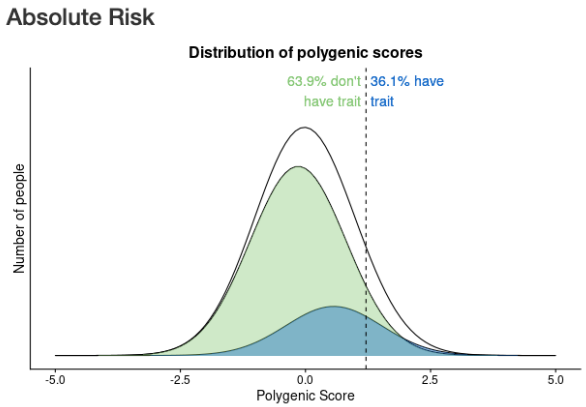
Odds ratio for each decile compared to the first decile



INTERPRETABILITY & RISK COMMUNICATION

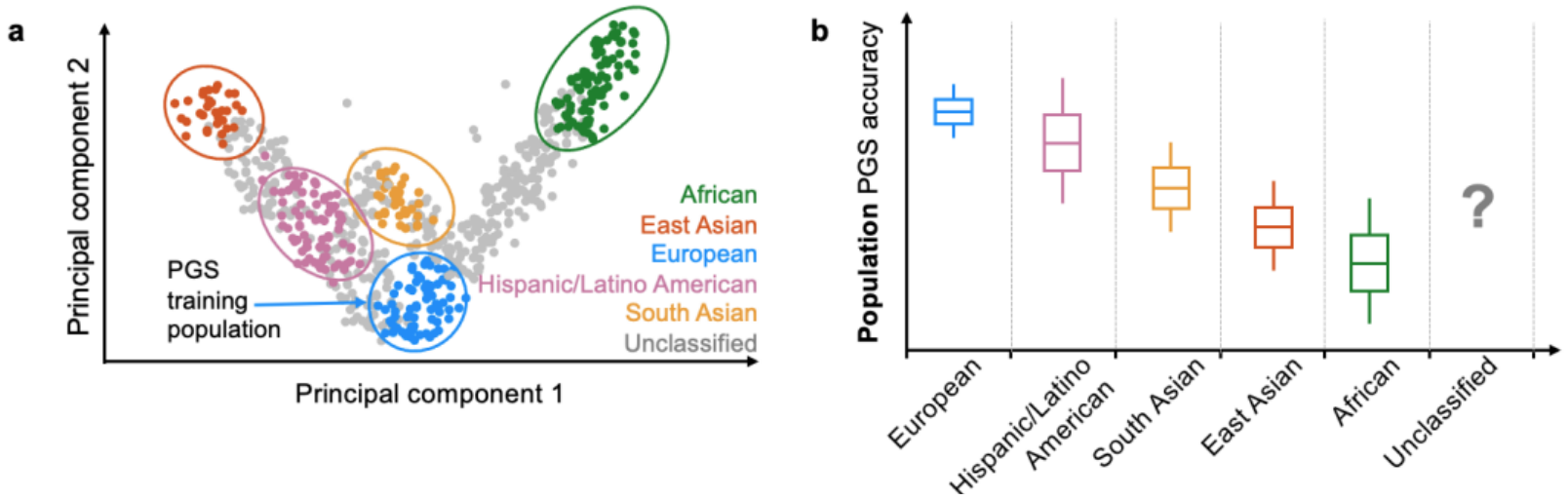
- The risk associated with the PGS is a relative risk
  - People have suggested methods to convert relative PGS risk to absolute risks (https://opain.github.io/GenoPred/PRS_to_Abs_tool.html)
- The relative risk may sound high, but the absolute risk is low
- Hard to use meaningfully in clinical decisions without baseline risk
- Effective for population stratification, less so for individual prediction
- Lack of population transferability

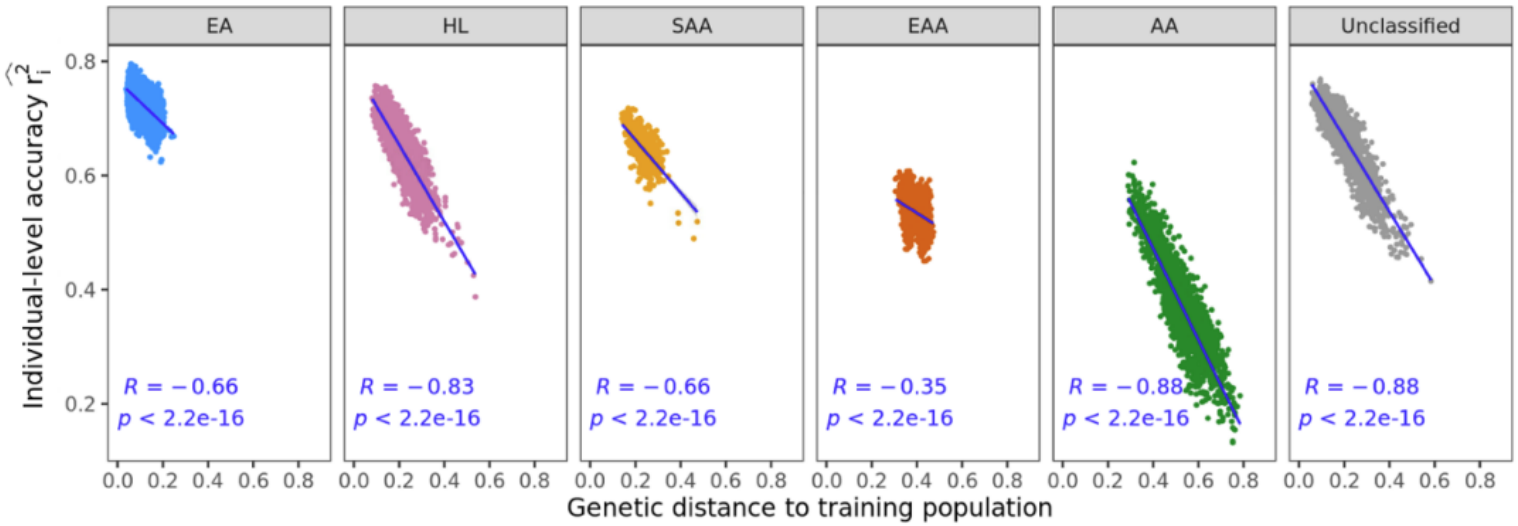


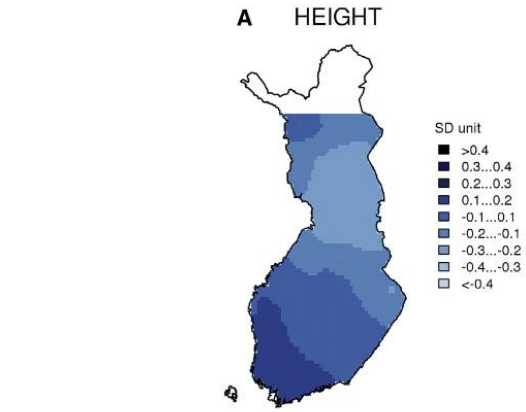


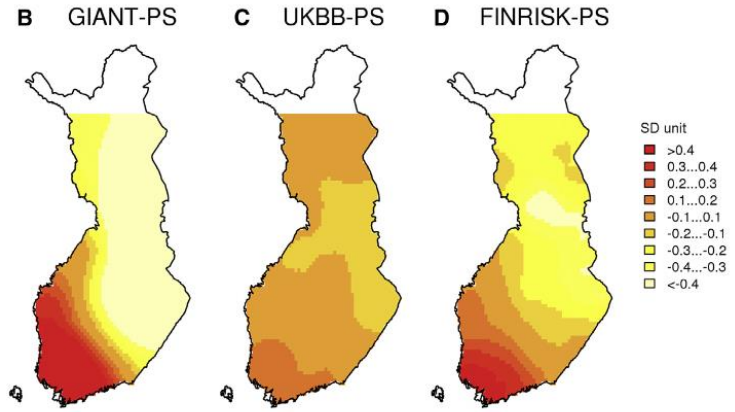
Pain et al 2022, EJHG, 3:339-348

LACK OF TRANSFERABILITY









Ding et al 2022, Nature, 618