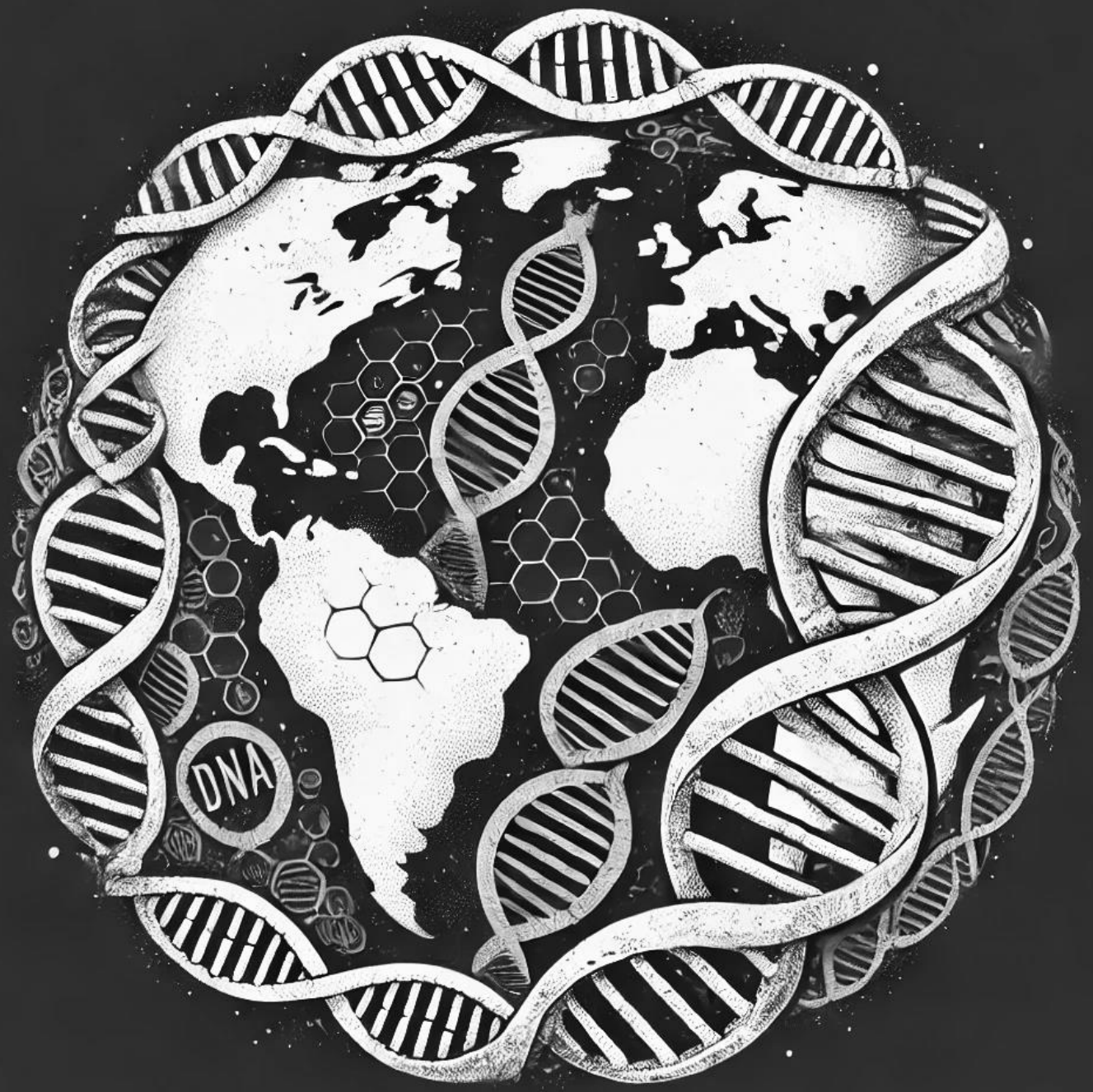


WORKSHOP
POLYGENIC SCORES

Palle Duun Rohde [palledr@hst.aau.dk]

Associate Professor & Research Group Leader

Genomic Medicine

Department of Health Science and Technology

Aalborg University

https://pdrohde.github.io/

THE PURPOSE OF TODAY

- **Give an introduction to polygenic scores (PGS)**
- **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)
- **Stratify a population/cohort by their inherent genetic load towards common complex diseases**
  - Polygenic scores (PGS)
- **Identify future projects of common interests**

AGENDA

08:00 – 08:30	Welcome and common introductions
08:30 – 09:10	Session 1: Introduction to Polygenic Scores (PGS)
09:10 – 09:20	Break
09:20 – 10:00	Session 2: Data Sources and Computational Methods
10:00 – 10:10	Break
10:10 – 10:40	Session 3: Evaluating and Interpreting Polygenic Scores
10:40 – 11:00	Break
11:00 – 11:45	Session 4: Advanced Applications and Future Directions
11:45 – 12:30	Lunch and short walk
12:30 – 15:30	Identification of 2-3 projects of common interest
15:30 – 16:00	Next steps and thank you for today

AGENDA

08:00 – 08:30	Welcome and common introductions
08:30 – 09:10	Session 1: Introduction to Polygenic Scores (PGS)
09:10 – 09:20	Break
09:20 – 10:00	Session 2: Data Sources and Computational Methods
10:00 – 10:10	Break
10:10 – 10:40	Session 3: Evaluating and Interpreting Polygenic Scores
10:40 – 11:00	Break
11:00 – 11:45	Session 4: Advanced Applications and Future Directions
11:45 – 12:30	Lunch and short walk
12:30 – 15:30	Identification of 2-3 projects of common interest
15:30 – 16:00	Next steps and thank you for today

AGENDA

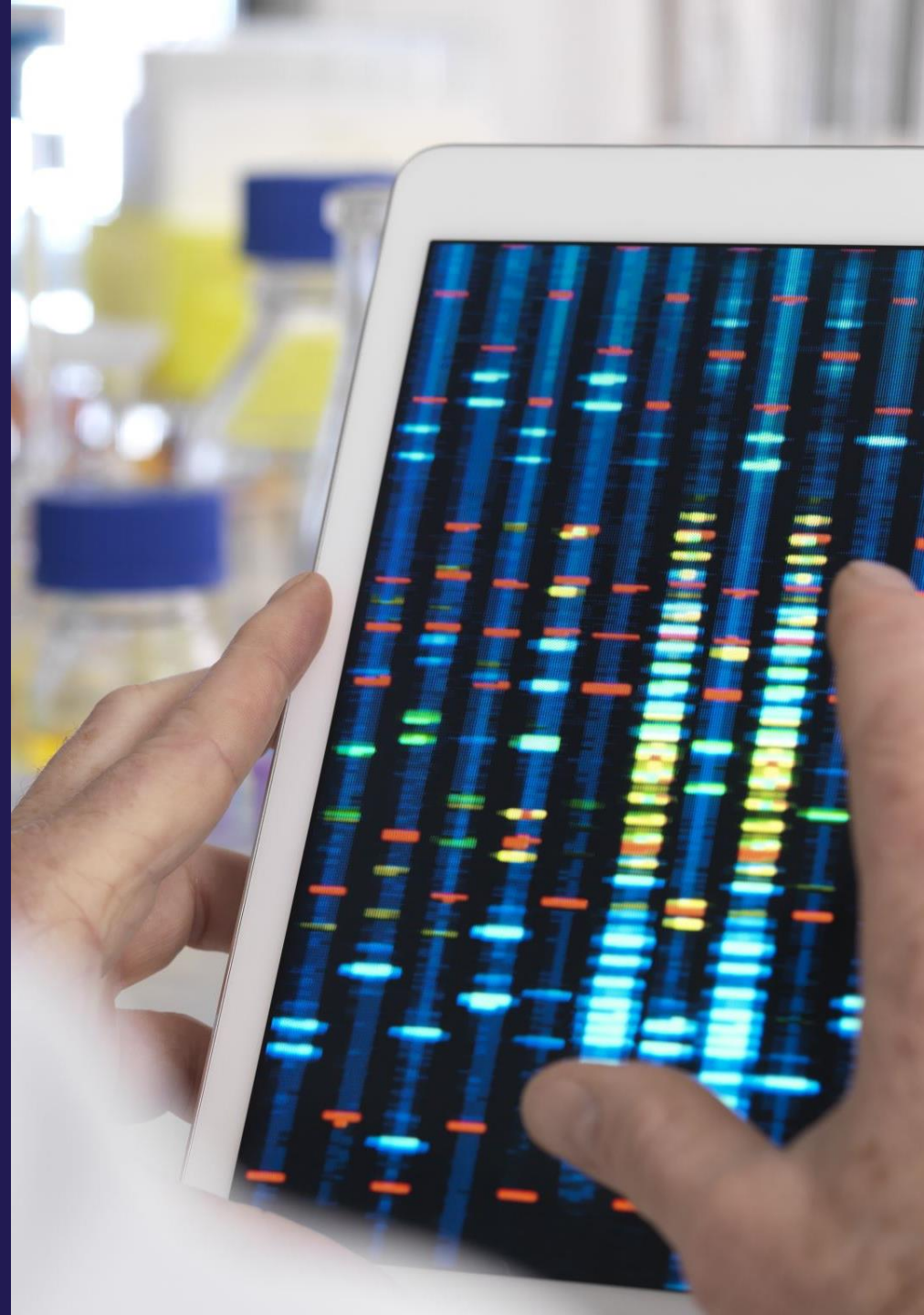
08:00 – 08:30	Welcome and common introductions
08:30 – 09:10	Session 1: Introduction to Polygenic Scores (PGS)
09:10 – 09:20	Break
09:20 – 10:00	Session 2: Data Sources and Computational Methods
10:00 – 10:10	Break
10:10 – 10:40	Session 3: Evaluating and Interpreting Polygenic Scores
10:40 – 11:00	Break
11:00 – 11:45	Session 4: Advanced Applications and Future Directions
11:45 – 12:30	Lunch and short walk
12:30 – 15:30	Identification of 2-3 projects of common interest
15:30 – 16:00	Next steps and thank you for today

SESSION 1

- Precision Medicine?
- Complex traits?
- Genetic variants as chilies
- What is a polygenic score?
- What is needed to compute a polygenic score?
- Why so many different methods?

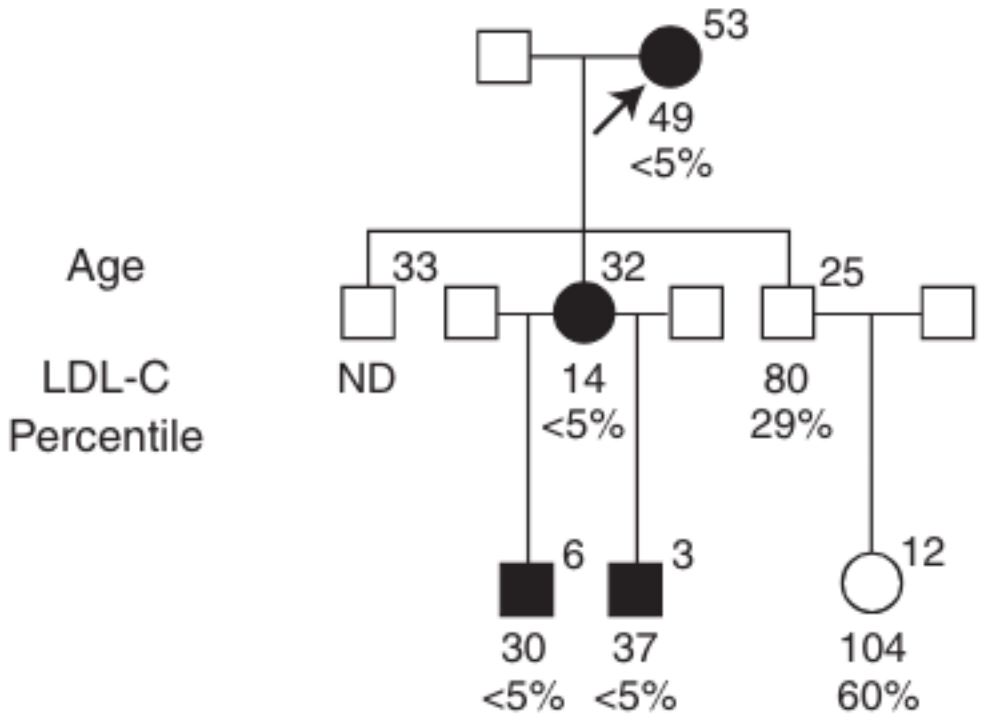
WHY STUDY THE HUMAN GENOME

- A story about the gene *PCSK9.*



A GENETIC VARIANT IN *PCSK9* IS ASSOCIATED WITH CHOLESTROL LEVELS

- LDL is a strong risk factor for heart disease.
- Carriers of nonsense mutation (filled symbols, heterozygous) within the *PCSK9* gene display markedly decreased LDL-C levels.
- Carriers are apparently healthy.
- **Why is this a very interesting observation?**
  - *PCSK9* could be a new drug target for lowering LDL cholesterol.



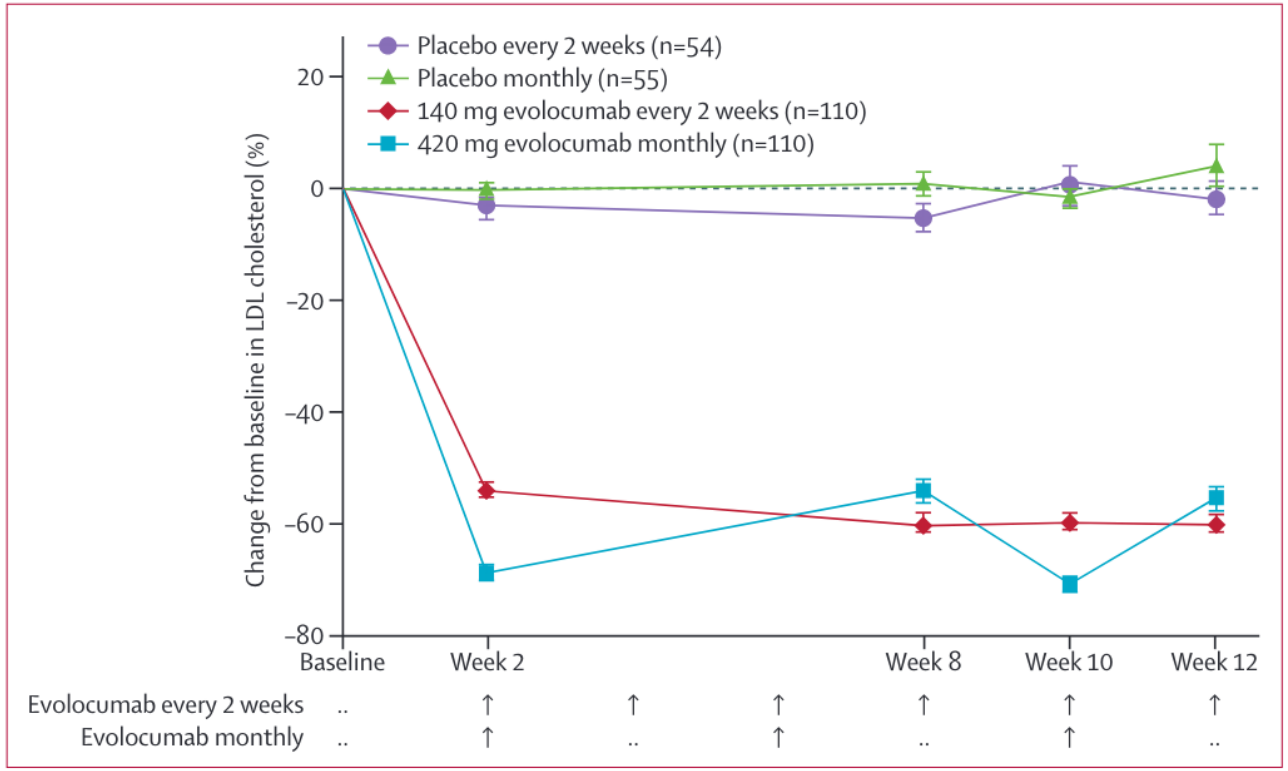
Cohen et al 2003, Nat Gen

PCSK9 INHIBITION WITH EVOLOCUMAB ON LDL CHOLESTEROL

Familial hypercholesterolaemia (FH) is a monogenic disease affecting 1 in 250 people and increases the likelihood of having coronary heart disease at a younger age.

FH increase LDL-C in blood → increase heart disease risk.

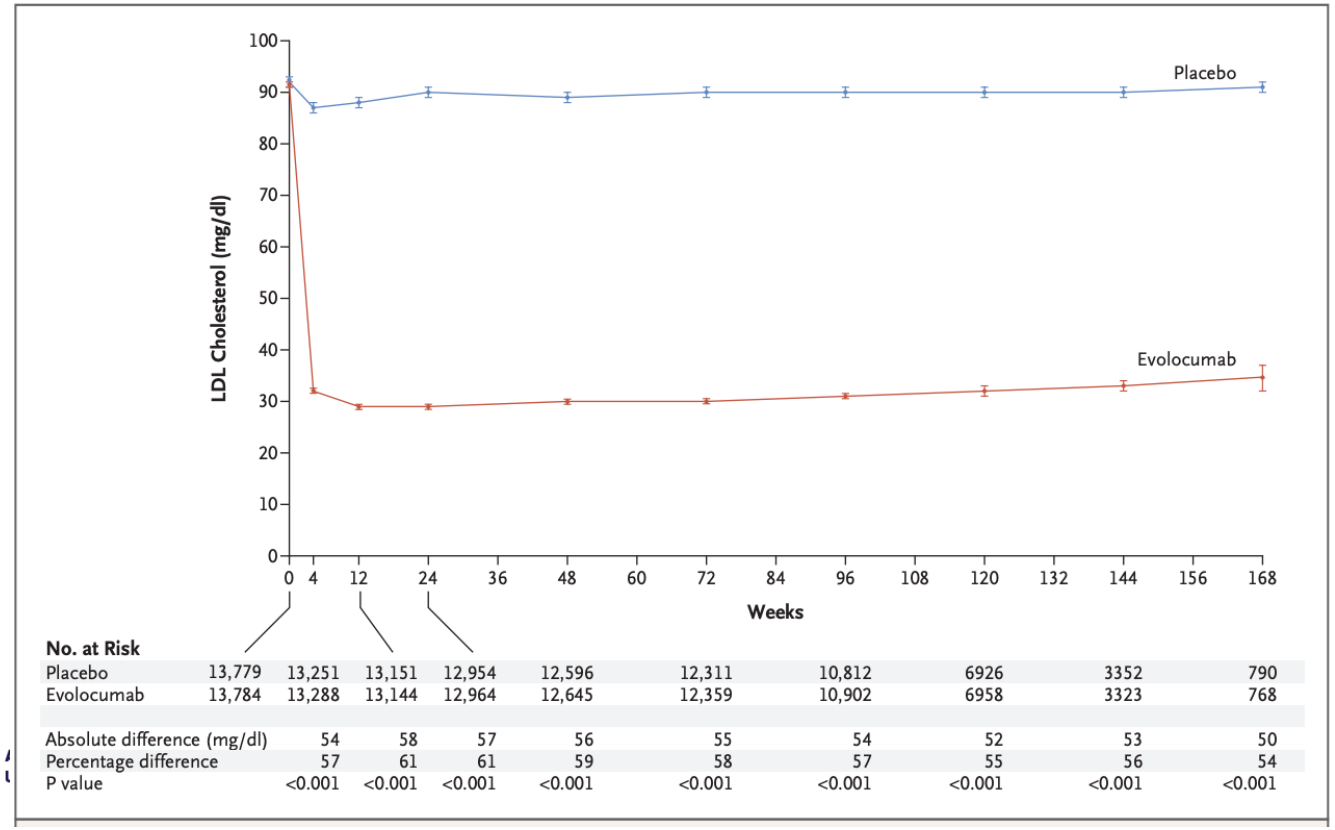
In patients that are heterozygous for FH gene mutation evolocumab reduced LDL-C.



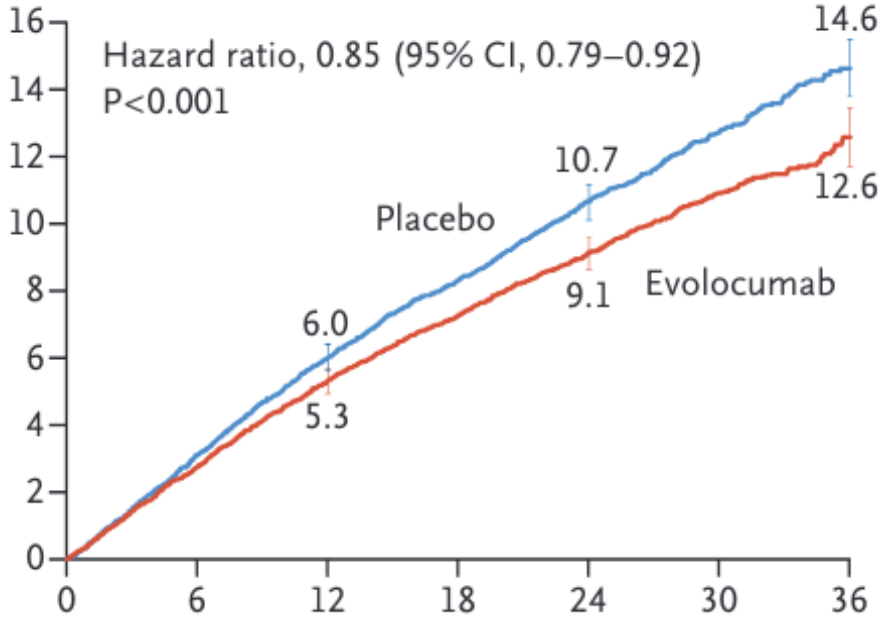
Raal et al 2015, Lancet

PCSK9 INHIBITION WITH EVOLOCUMAB ON LDL CHOLESTEROL

In patients with cardiovascular disease.
At 48 weeks reduction of 59% LDL-C with evolocumab.



No. at Risk										
Placebo	13,779	13,251	13,151	12,954	12,596	12,311	10,812	6926	3352	790
Evolocumab	13,784	13,288	13,144	12,964	12,645	12,359	10,902	6958	3323	768
Absolute difference (mg/dl)		54	58	57	56	55	54	52	53	50
Percentage difference		57	61	61	59	58	57	55	56	54
P value		<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001



Reduction in primary end points in evolocumab group compared to placebo.

Sabatine et al 2017, NEJM

WE LOOK AT VARIATION NATURE CREATED

Nature as a big genetic laboratory

GENOMIC-INFORMED MEDICINE

Personalised medicine

vs

Precision medicine

vs

Individualised medicine

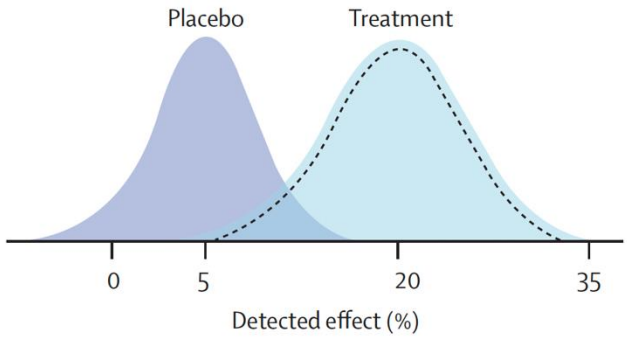
Uhh?!

IMPLEMENTATION OF PRECISION MEDICINE

EPPOS [evidence-based precision personalised objective subjective]

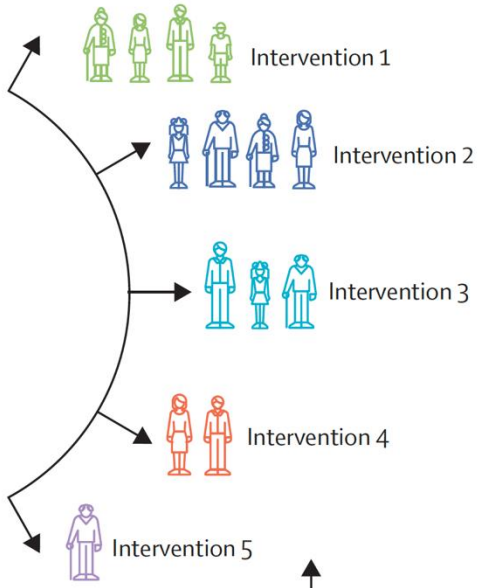
Evidence-based Medicine

(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

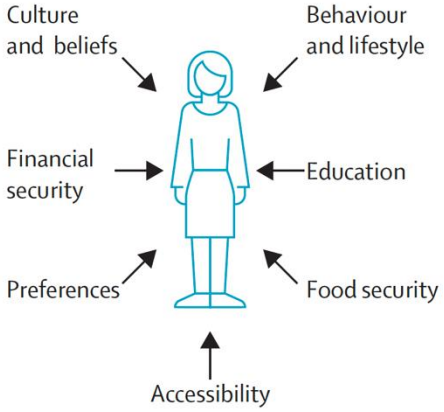


Personalised Medicine

(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



AALBORG UNIVERSITY

The Novo Nordisk Foundation, Precision medicine for cardiometabolic disease 2022

Franks et al, 2023, Lancet Diabetes Endocrinol; 11: 822–35

DIFFERENT MODE OF INHERITANCES

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

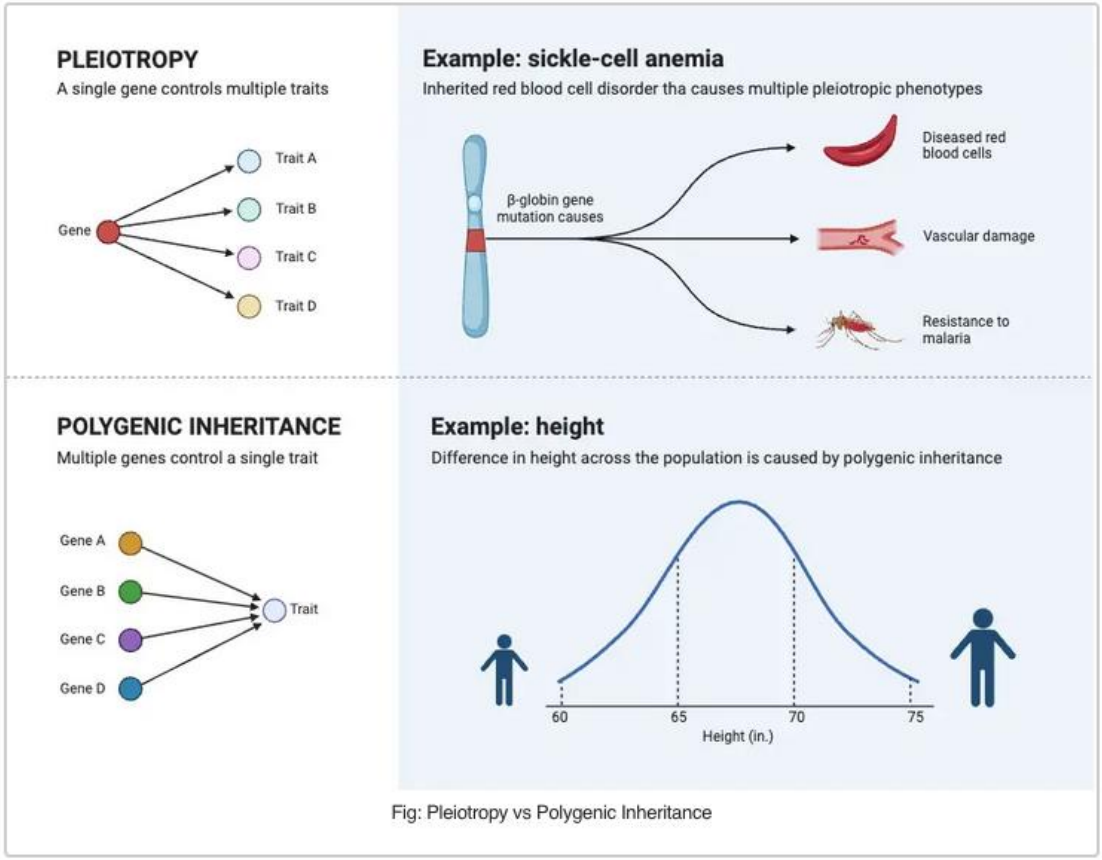
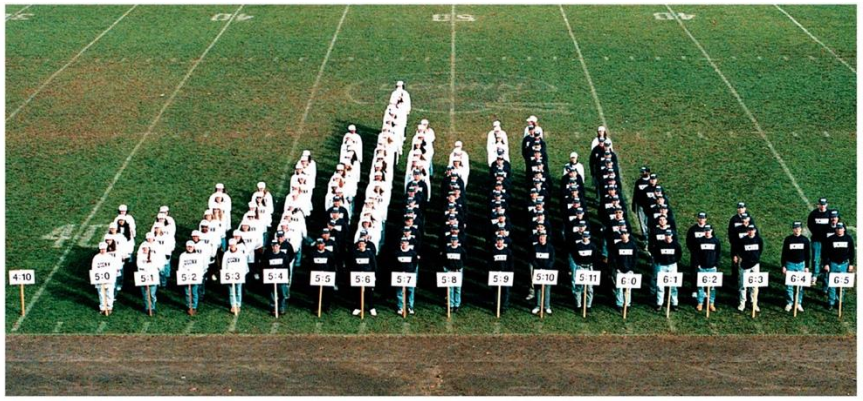


Fig: Pleiotropy vs Polygenic Inheritance

QUANTITATIVE TRAITS

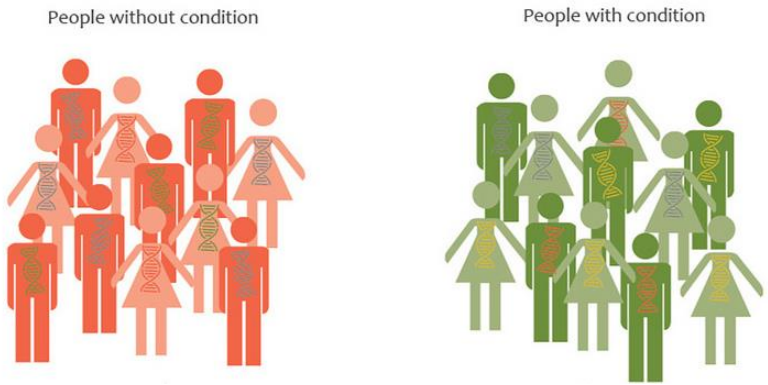
IN DIFFERENT SHAPES

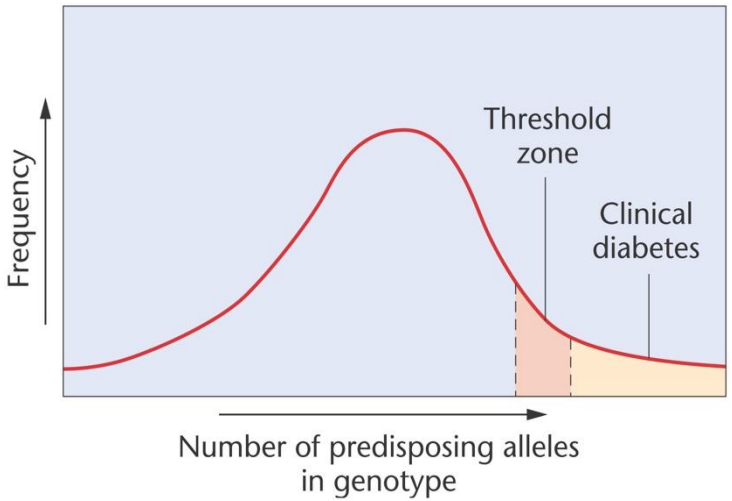
Continous variation



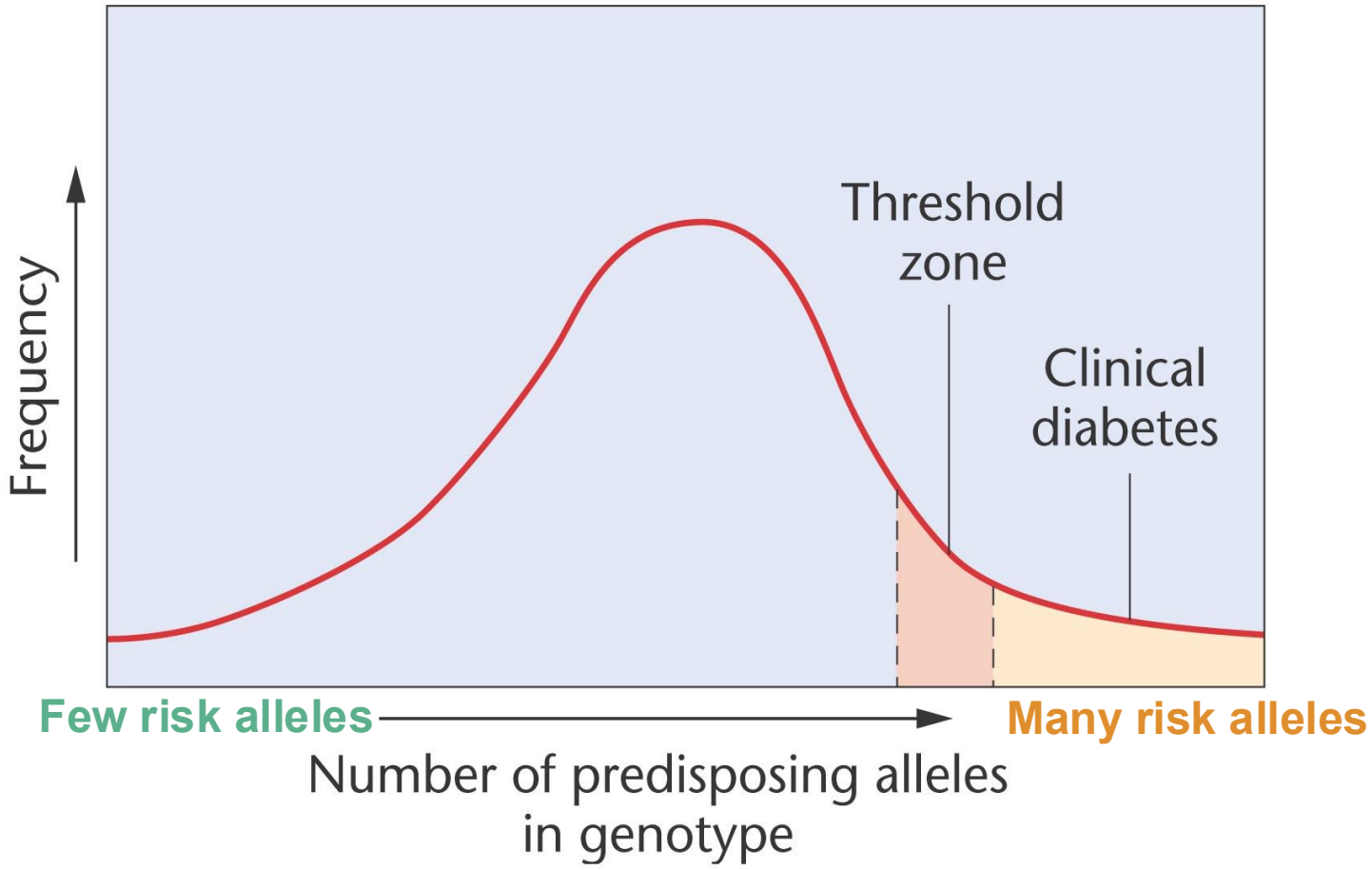
Categorical variation

Threshold traits





LIABILITY (THRESHOLD) MODEL



Liability model

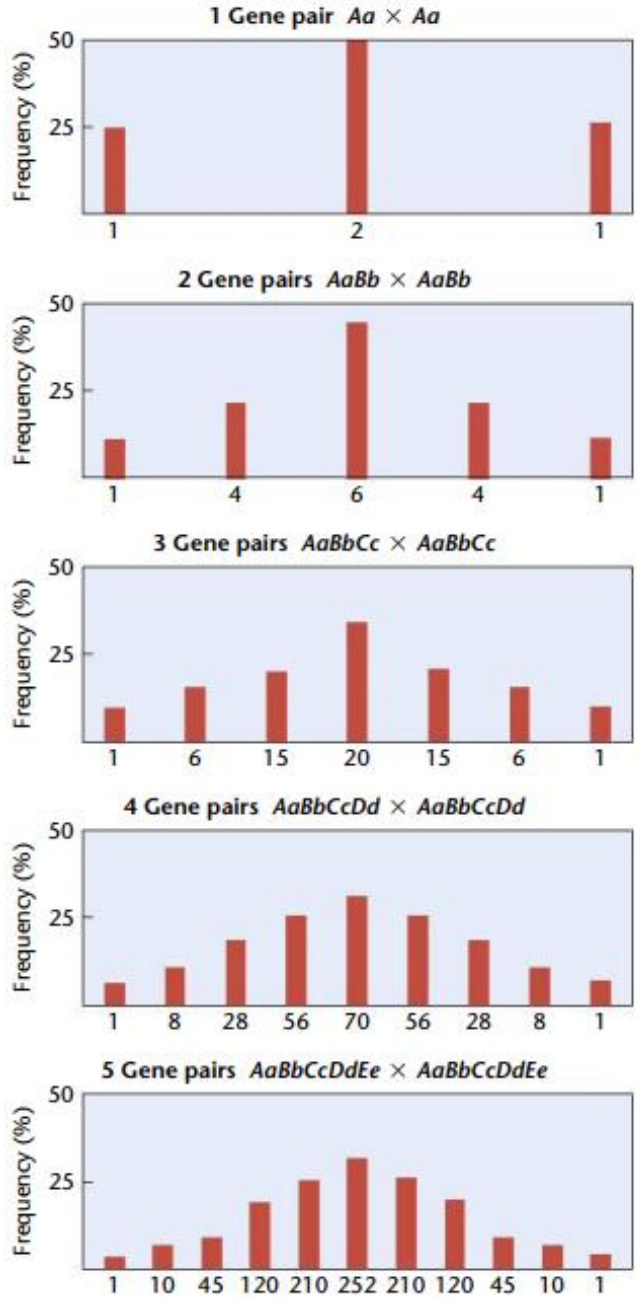
Only individuals with a liability over a certain threshold will become affected

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

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

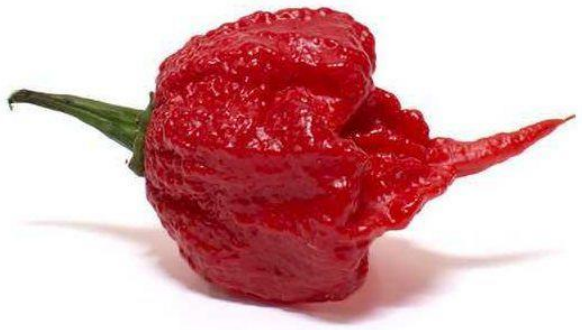
POLYGENIC TRAITS

A simple ‘Mendelian’ explanation for why polygenic traits follows a normal distribution



GENETIC VARIATION IS LIKE CHILI

Carolina reaper



Strong effect on phenotype

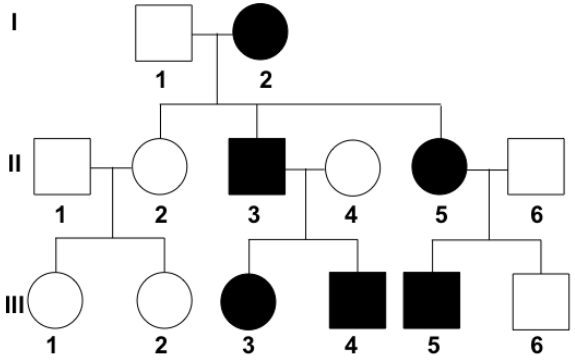
Habanero Lemon

Moderate effect on phenotype

Bell pepper

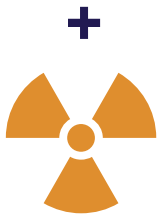
Weak effect on phenotype
(or none at all)

GENETIC VARIATION IS LIKE CHILI

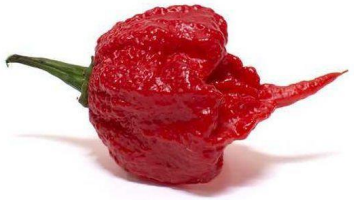


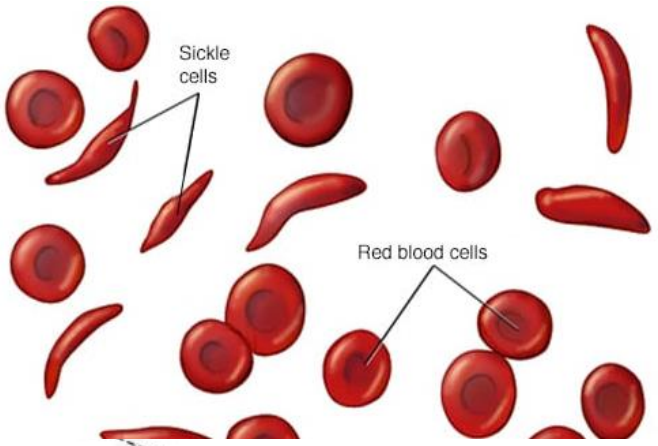
Mutation

Each genetic variant is both necessary and sufficient



Each genetic variant is neither necessary nor sufficient





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

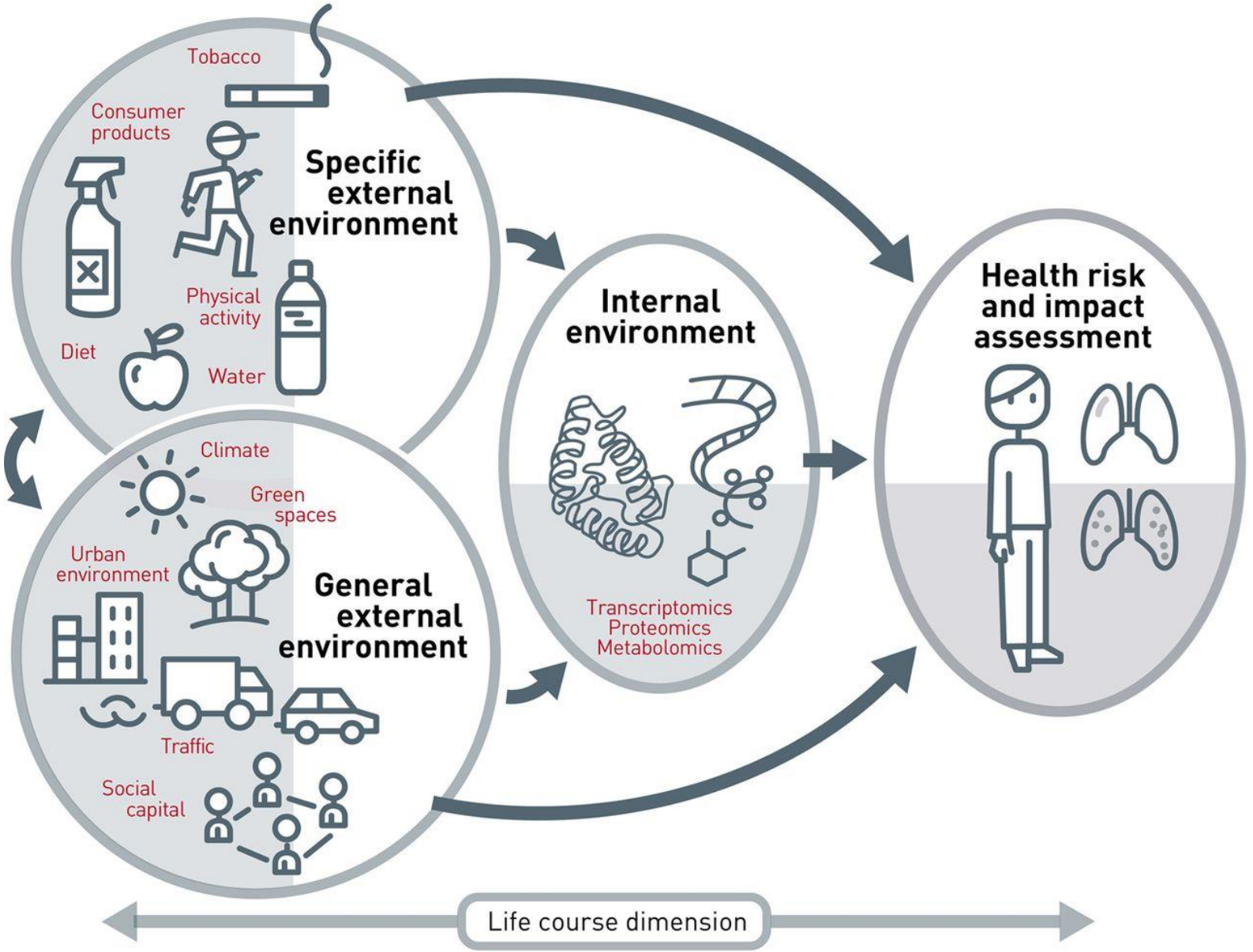
COMPLEX TRAITS are complex...

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

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

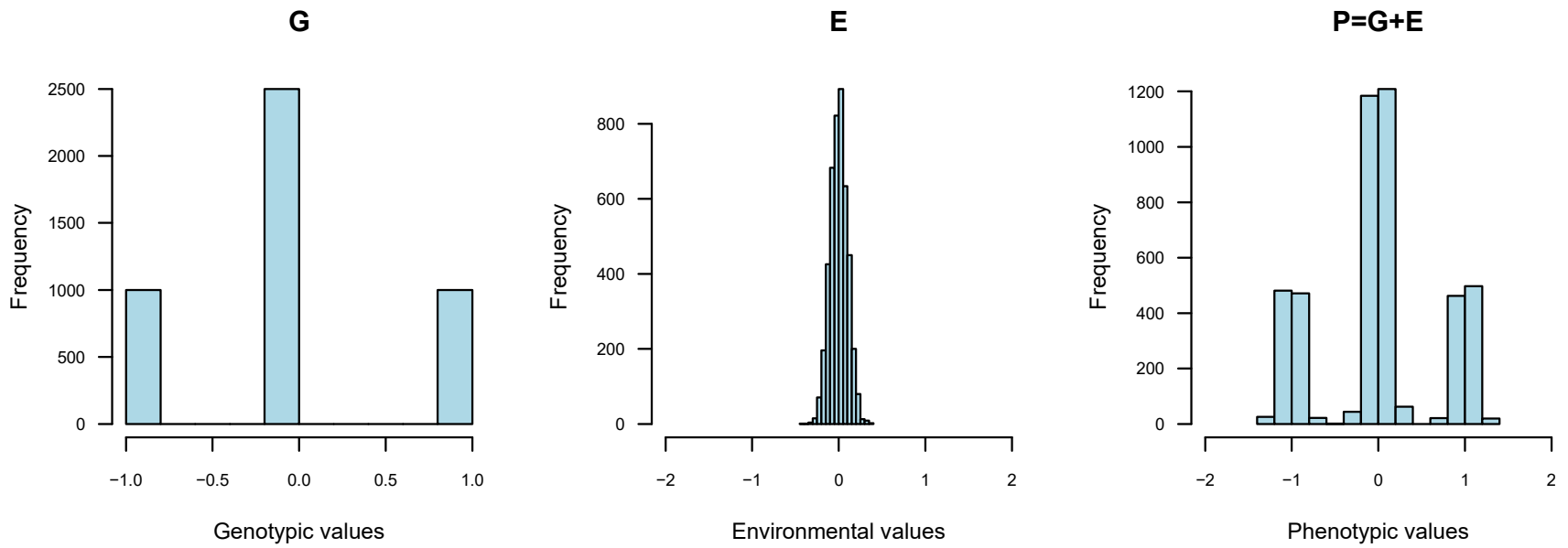
$$V_G / (V_G+V_E)$$

Thus, set a limits for the genetic predictor



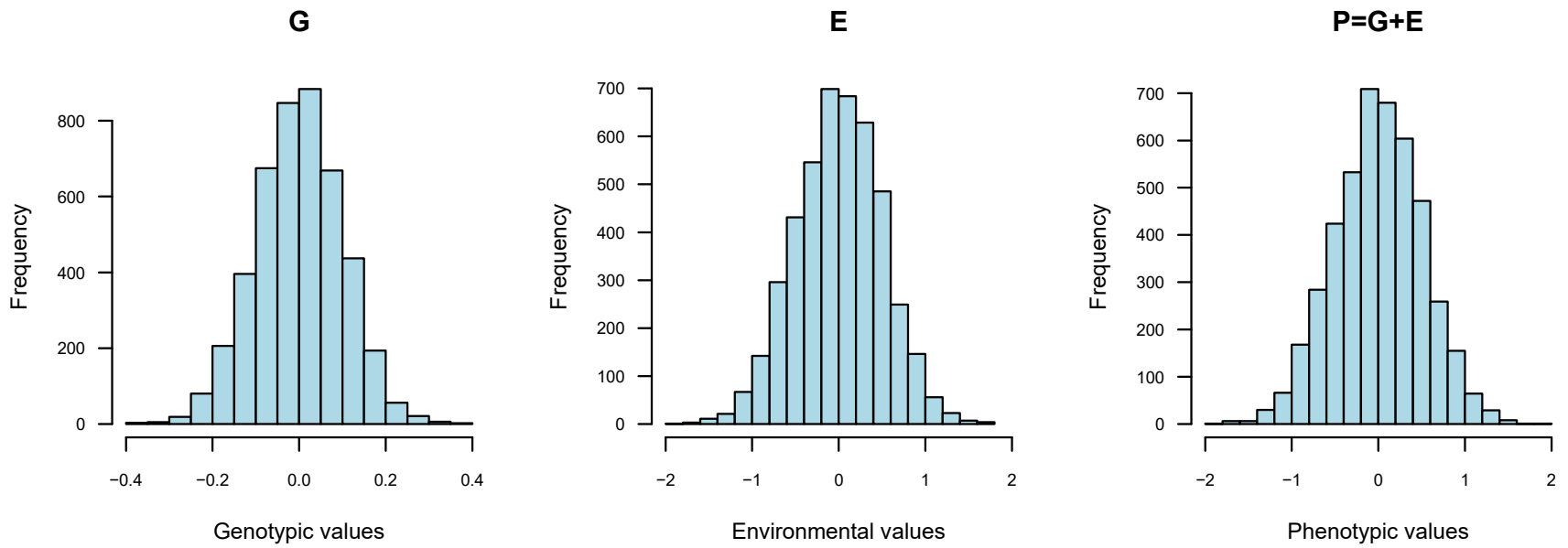
MULTIFACTORIAL TRAITS

Environmental variance (non-genetic factors) blurs phenotypic classes



MULTIFACTORIAL TRAITS

Genotypic and environmental variance creates infinite many phenotypic classes



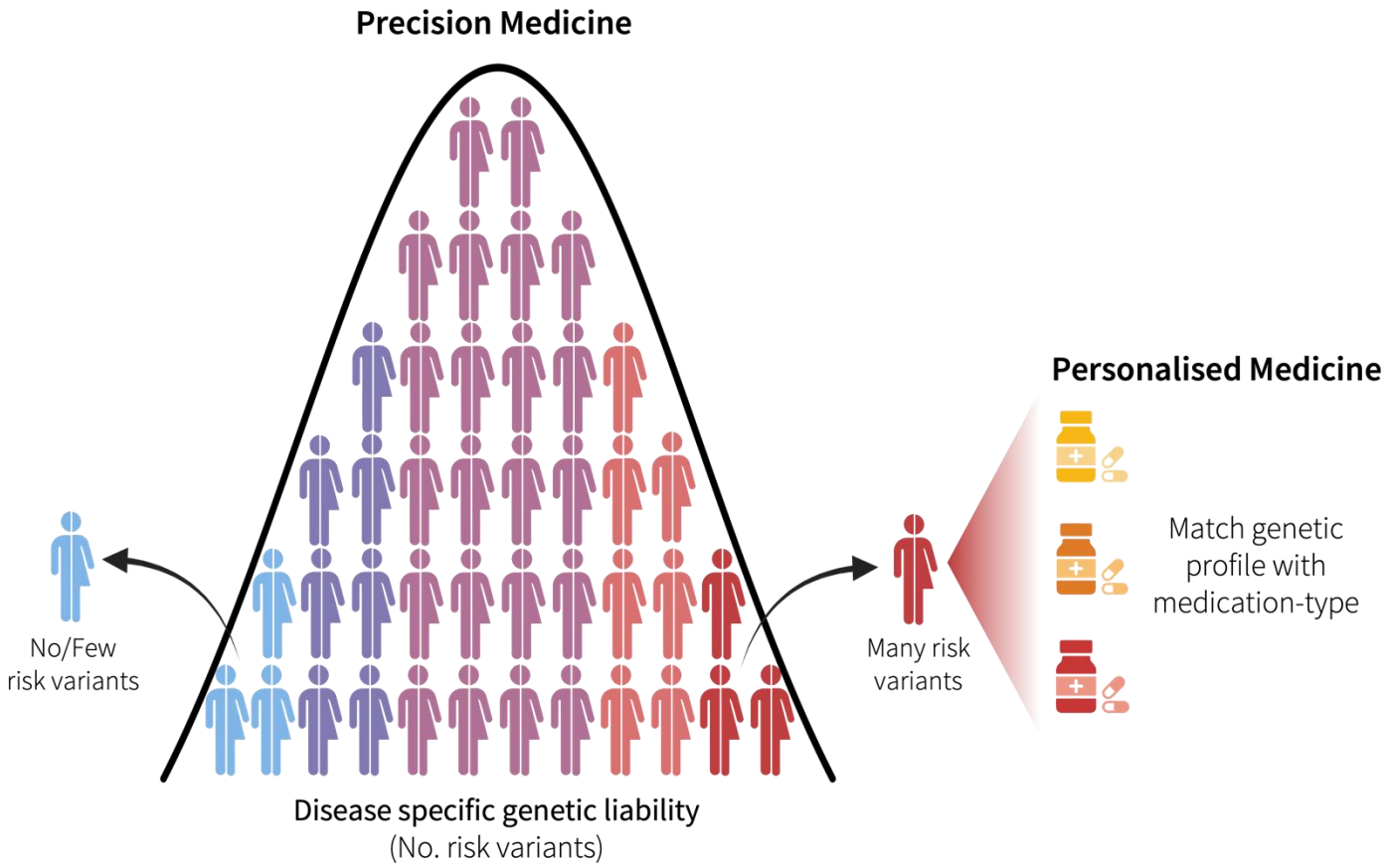
PREDICTING DISEASE RISK FROM GENETIC DATA

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)

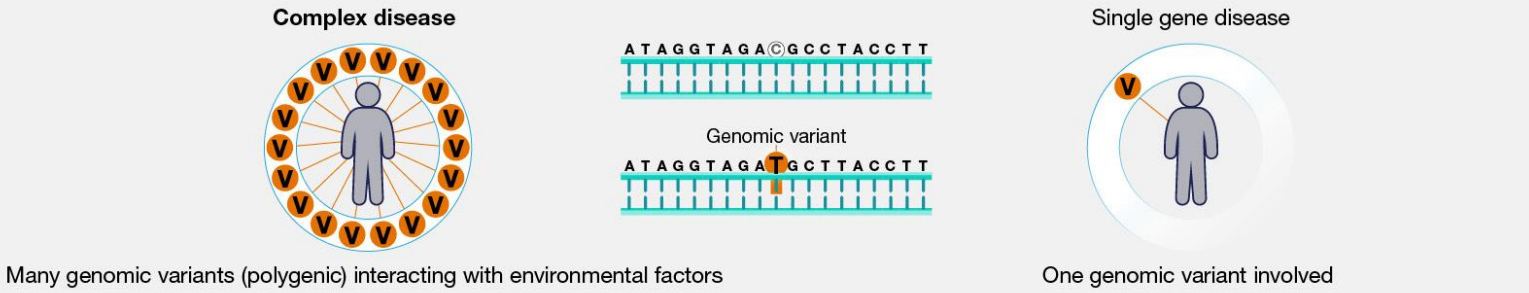
DIFFERENT NAMES

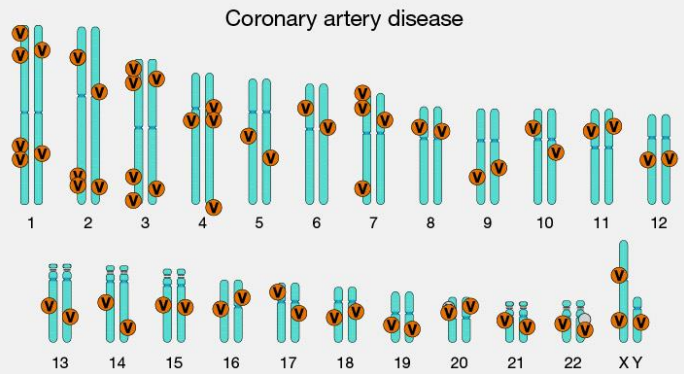
BUT THE SAME

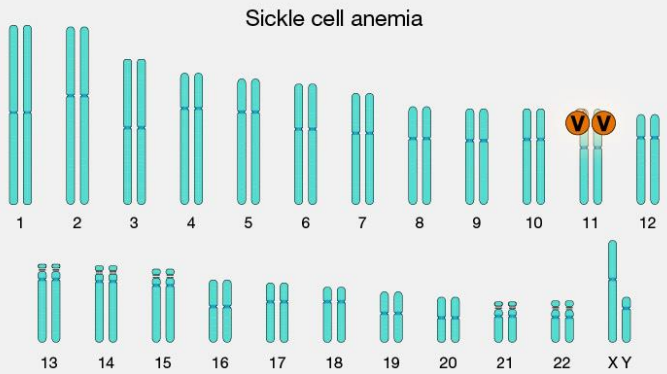
- Polygenic risk score (PRS)
- Polygenic score (PGS)
- Genetic score (GS)
- Genetic risk score (GRS)
- Genetic value
- Genetic liability
- Breeding value
- …



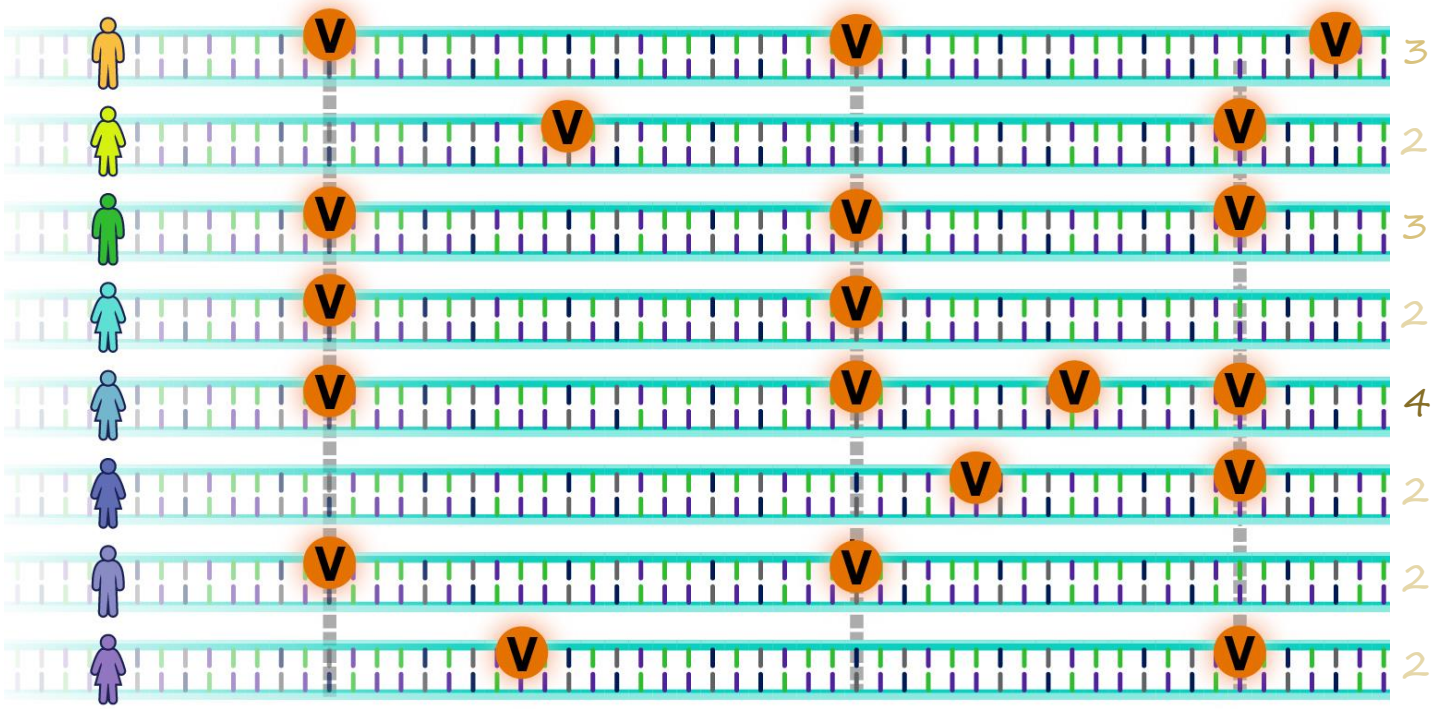
THE INHERENT DISEASE RISK







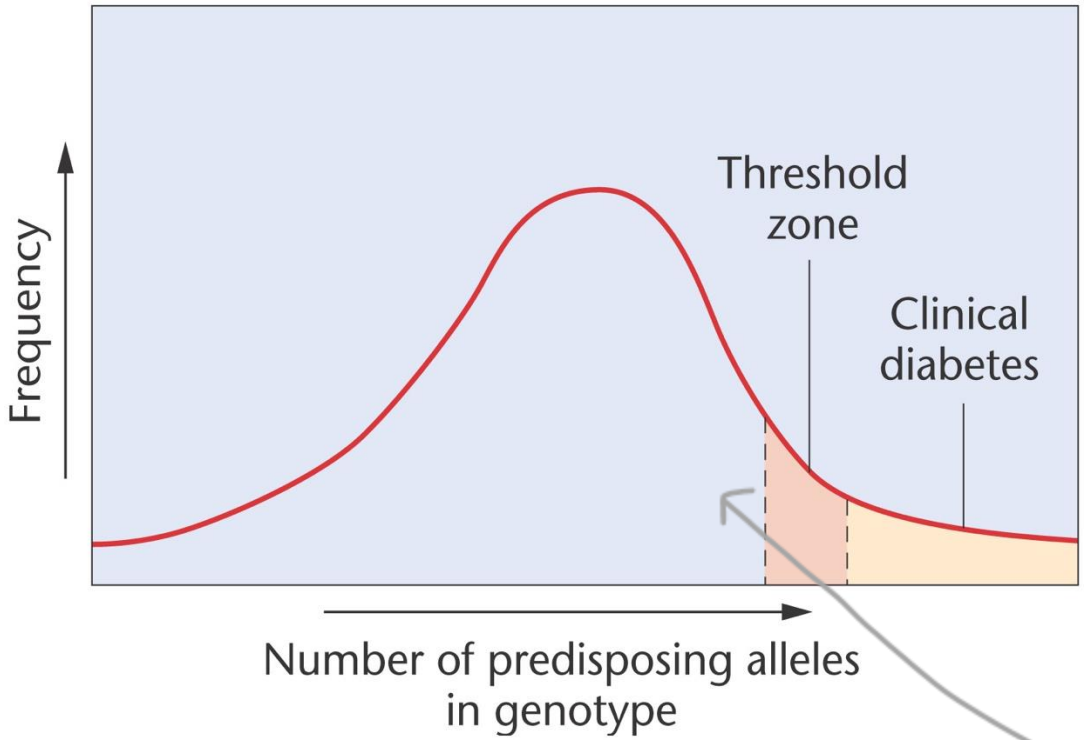
THE INHERENT DISEASE RISK



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

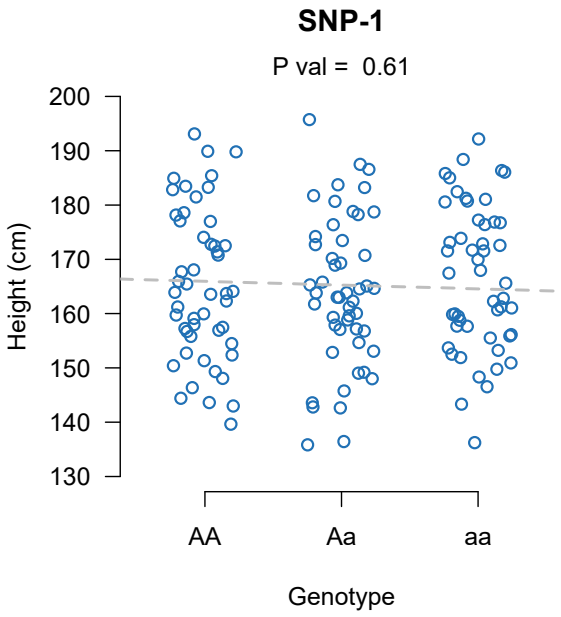
SNP5 C C

PGS =7

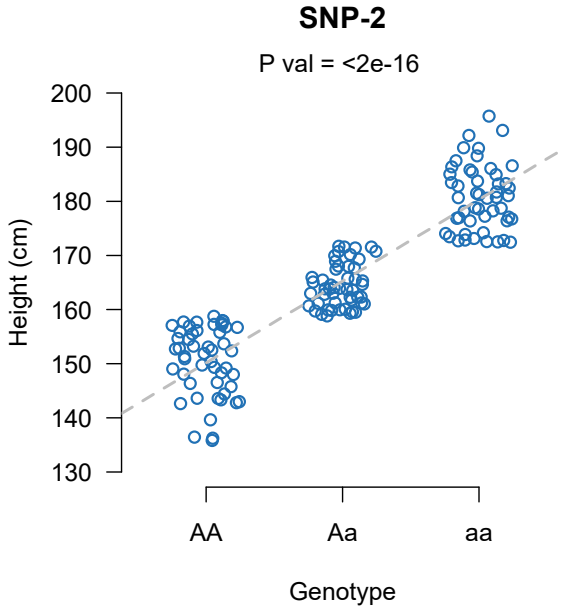
How do we find the 'orange' alleles?

IDENTIFY RISK VARIANTS

GENOME-WIDE ASSOCIATION STUDY (GWAS)

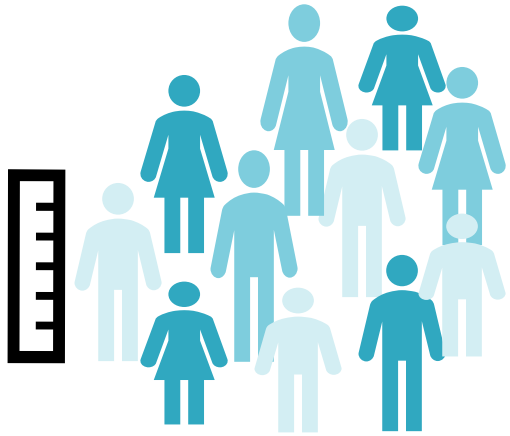


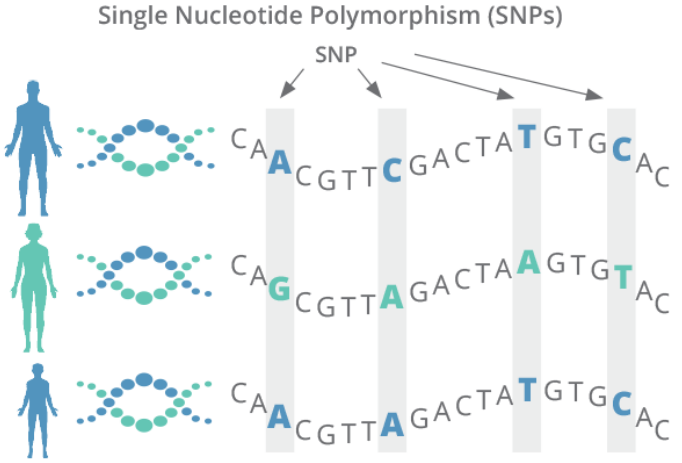
No association



Association

Systematic hypothesis-free scanning of all
common genetic variants

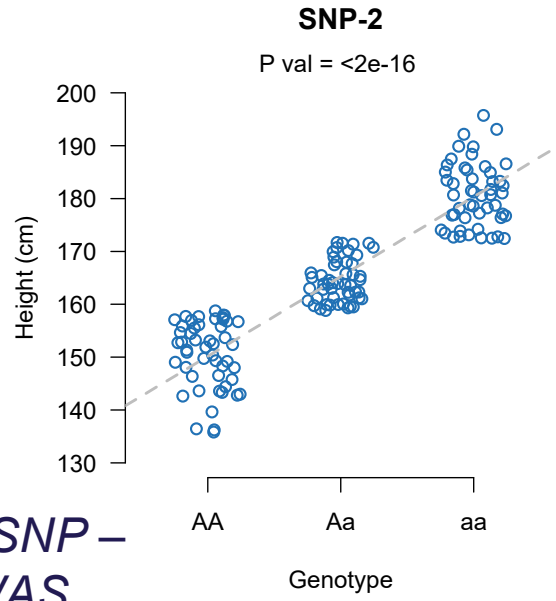




WHAT IS A PGS?

"*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

$$PGS = \sum X_i b_i$$

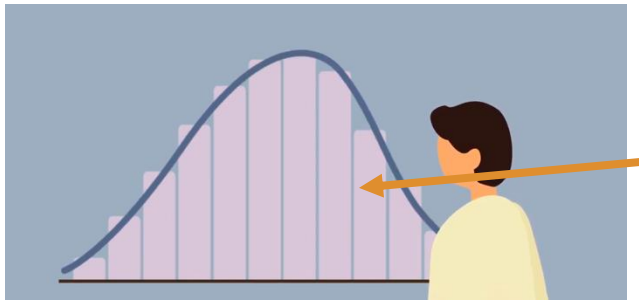
The genotype of the individual for SNP i (0, 1, 2 – counting the number of the alternative allele)

AA = 0
Aa = 1
aa = 2

$$PGS = \sum X_i b_i$$

HOW TO COMPUTE A (simple) PGS?

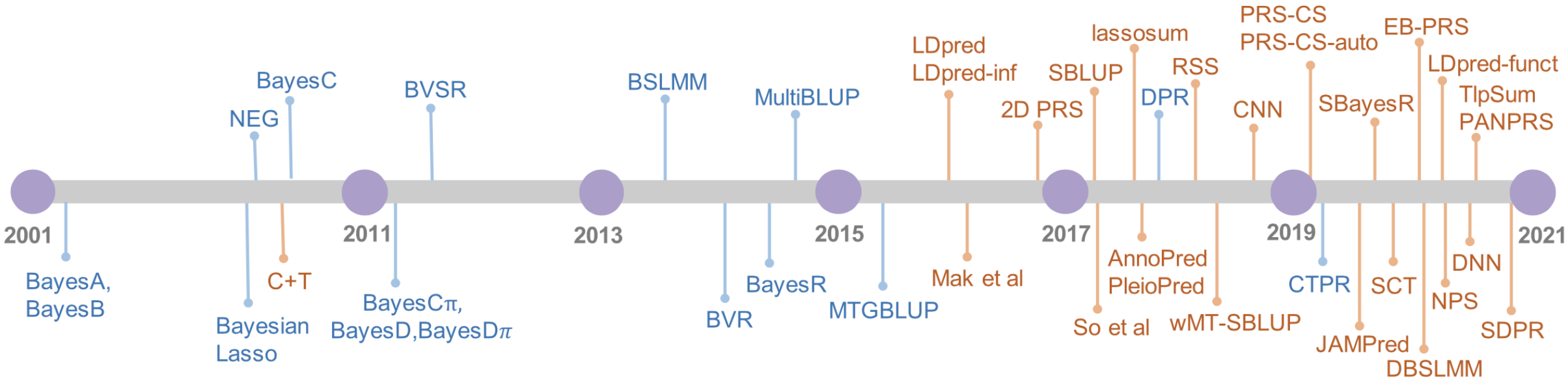
SNPs	Adams Genotypes	Ref allele	Alt allele	X	b	Xb
SNP-1	TC	T	C	1	0.04	0.04
SNP-2	GG	G	T	0	0.02	0.00
SNP-3	CC	A	C	2	0.05	0.10
SNP-4	TG	T	G	1	0.02	0.02
SNP-5	AA	A	G	0	0.06	0.00



PGS = 0.16

A LARGE PALETTE OF PGS METHODS

$$PGS = \sum X_i b_i$$



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

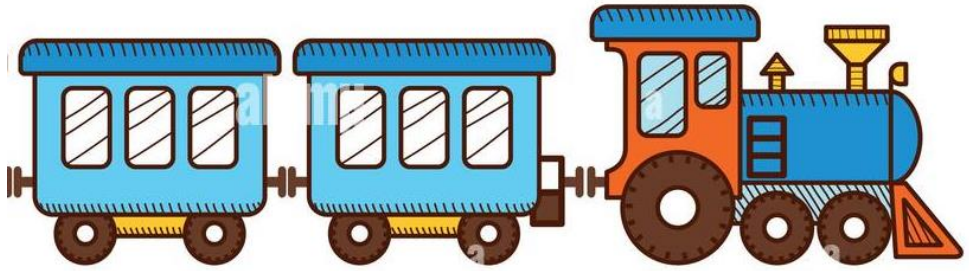
WHY DIFFERENT PGS SCORING METHODS?

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)

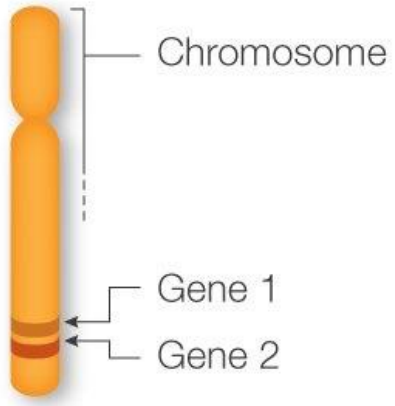
The different scoring algorithms attempts to account for this.

GENETIC LINKAGE



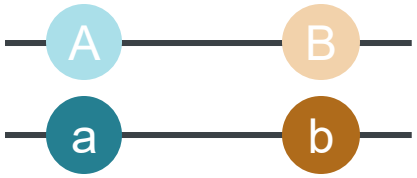
Linked train wagons

Linked prisoners

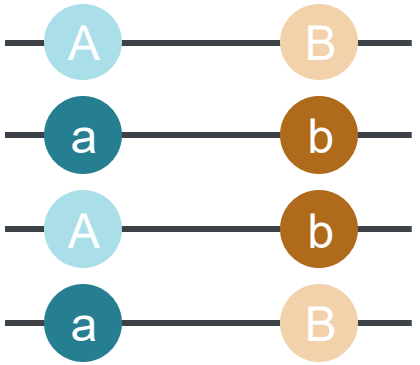


Linked loci
(Physical proximity)

WHAT IS LD?



Which gamtes can be produced?



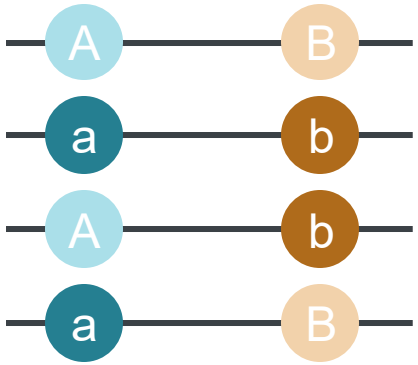
What are the frequencies of the alleles?

P(A), P(a)
P(B), P(b)

What are the frequencies of the haplotypes?

P(AB), P(Ab)
P(aB), P(ab)

WHAT IS LD?



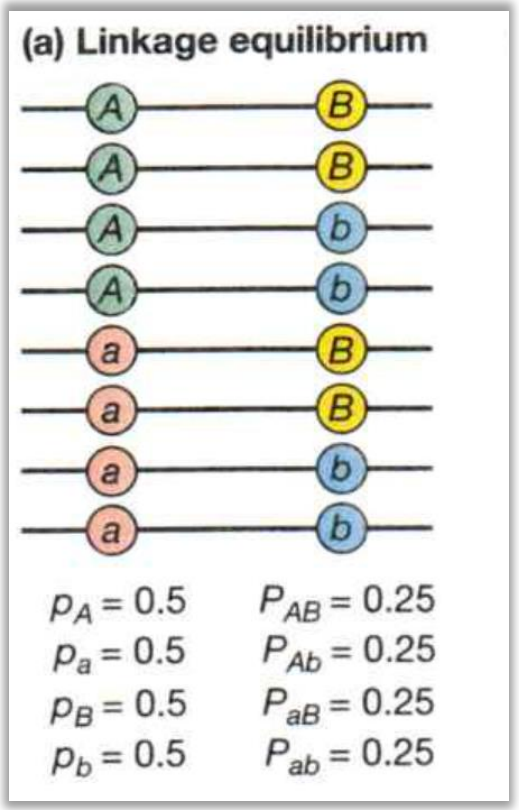
If there is random relationship among alleles at the two loci then the frequency of the haplotypes will be the product of the frequencies of the two alleles:

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

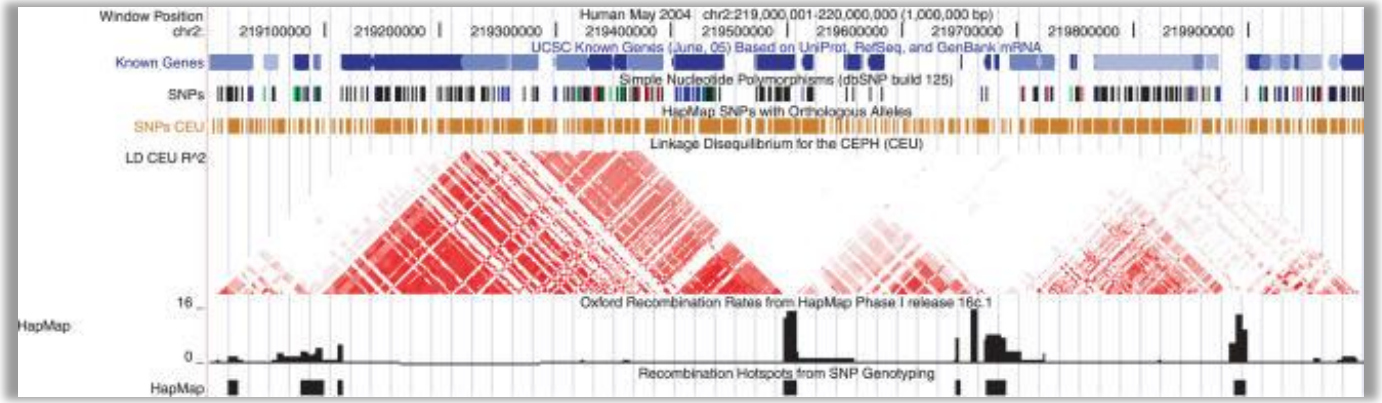
$P(Ab)=P(A)xP(b)$

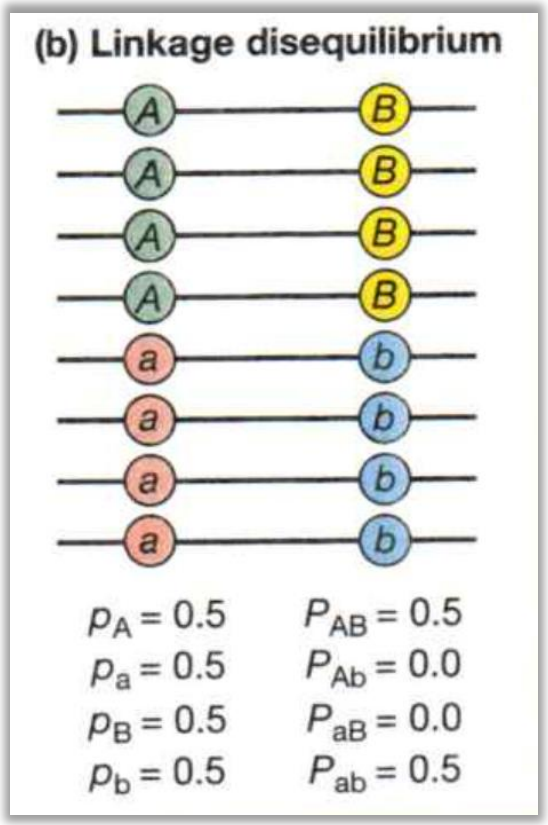
$P(aB)=P(a)xP(B)$

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



WHAT IS LD?



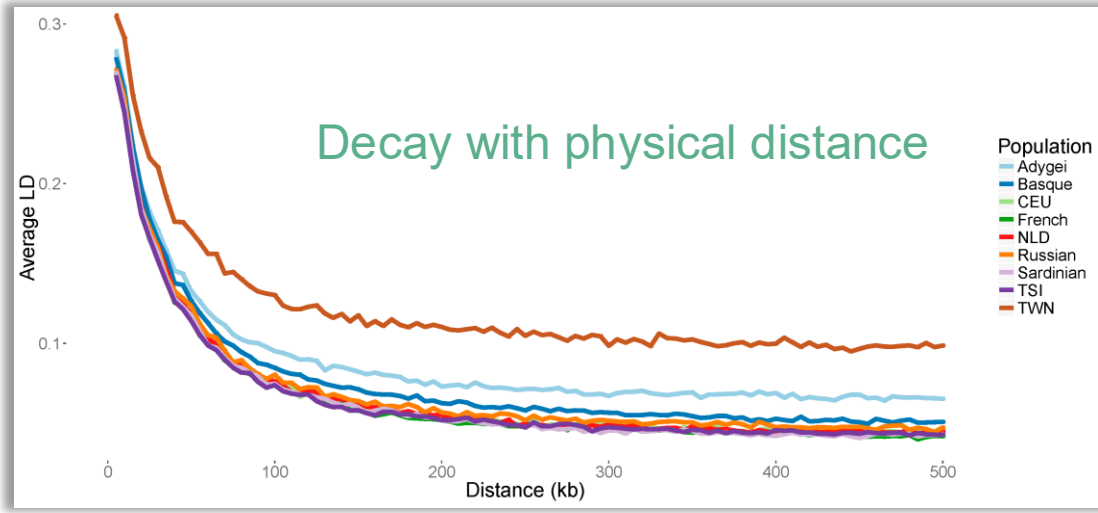


When the association between alleles at two loci is **non-random** they are said to be in **linkage disequilibrium**

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

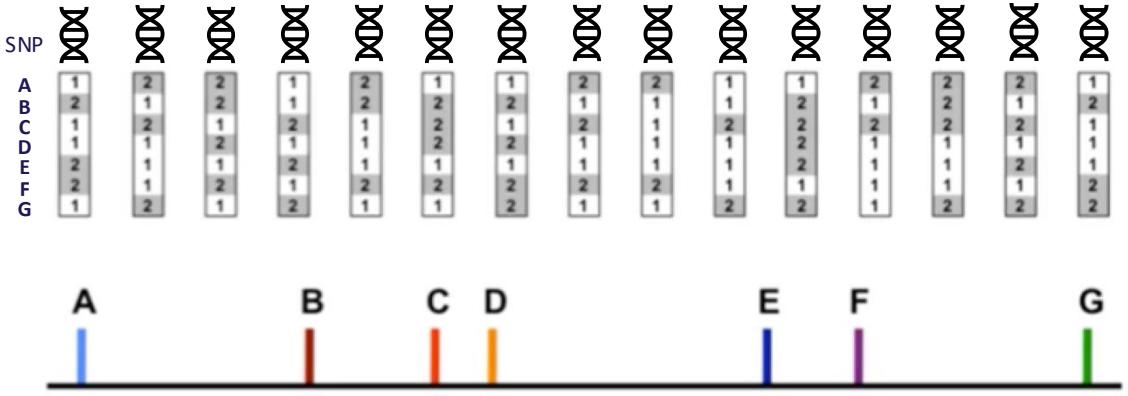
$$D = P_{AB} - P_A P_B$$

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

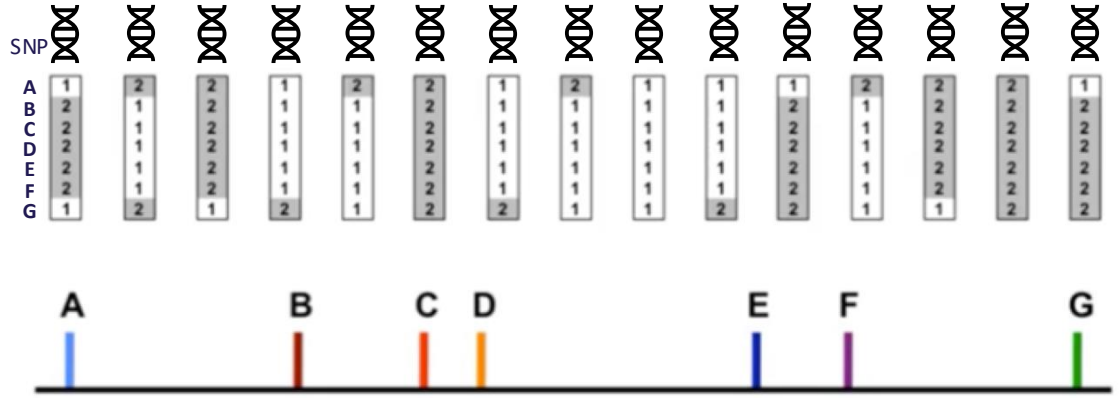


LD AND GENE MAPPING

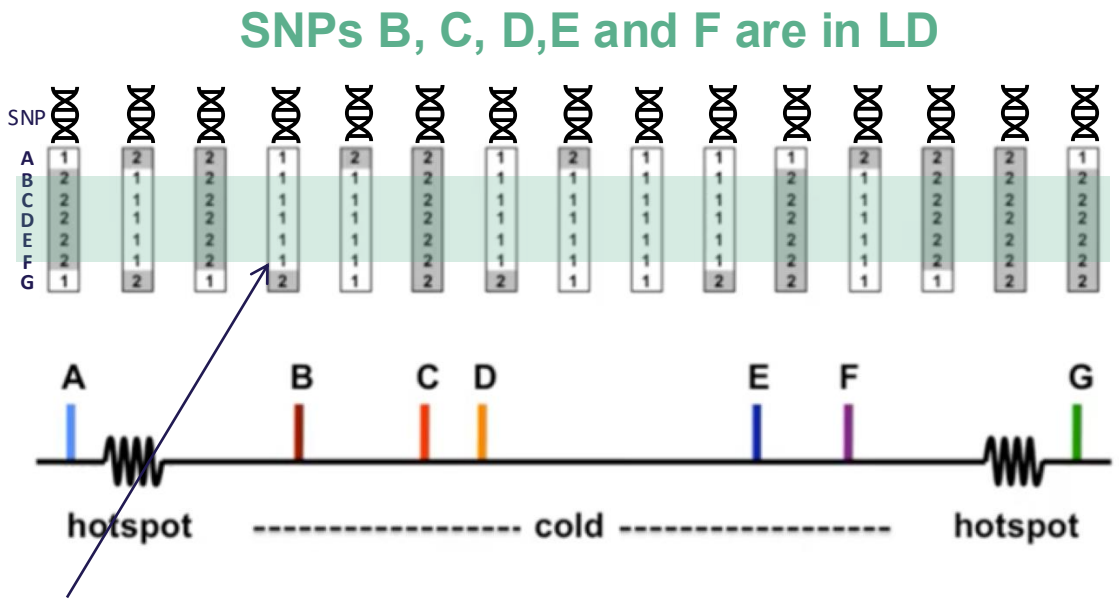
Linkage equilibrium – *random association*



Linkage disequilibrium – *non-random association*

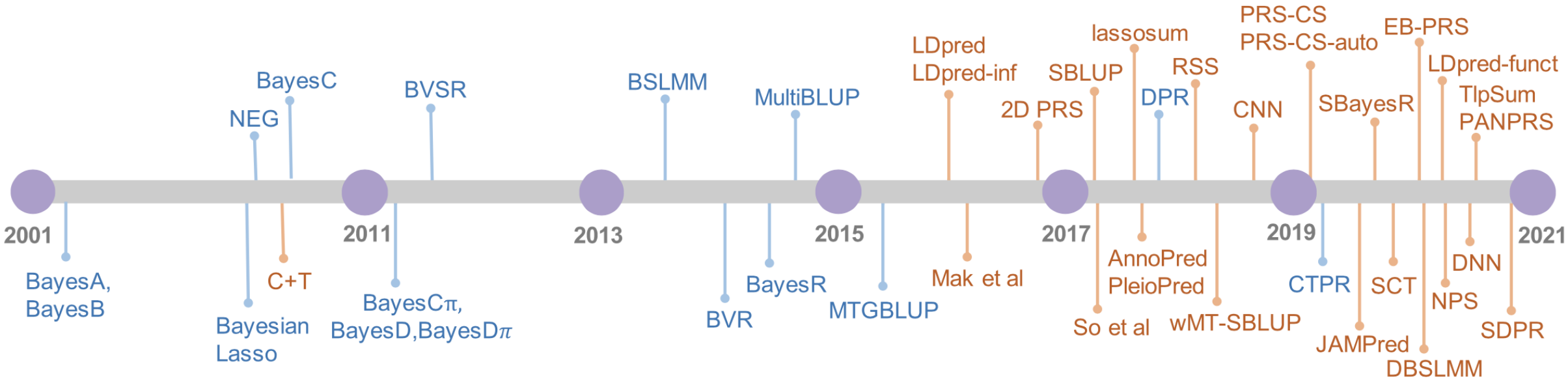


LD AND GENE MAPPING



If you have allele 1 here, I know what you are at the remaing sites in this haploblok

A LARGE PALETTE OF PGS METHODS



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

SESSION 1

- Precision Medicine?
- Complex traits?
- Genetic variants as chilies
- What is a polygenic score?
- What is needed to compute a polygenic score?
- Why so many different methods?


BREAK

AGENDA

08:00 – 08:30	Welcome and common introductions
08:30 – 09:10	Session 1: Introduction to Polygenic Scores (PGS)
09:10 – 09:20	Break
09:20 – 10:00	Session 2: Data Sources and Computational Methods
10:00 – 10:10	Break
10:10 – 10:40	Session 3: Evaluating and Interpreting Polygenic Scores
10:40 – 11:00	Break
11:00 – 11:45	Session 4: Advanced Applications and Future Directions
11:45 – 12:30	Lunch and short walk
12:30 – 15:30	Identification of 2-3 projects of common interest
15:30 – 16:00	Next steps and thank you for today

SESSION 2

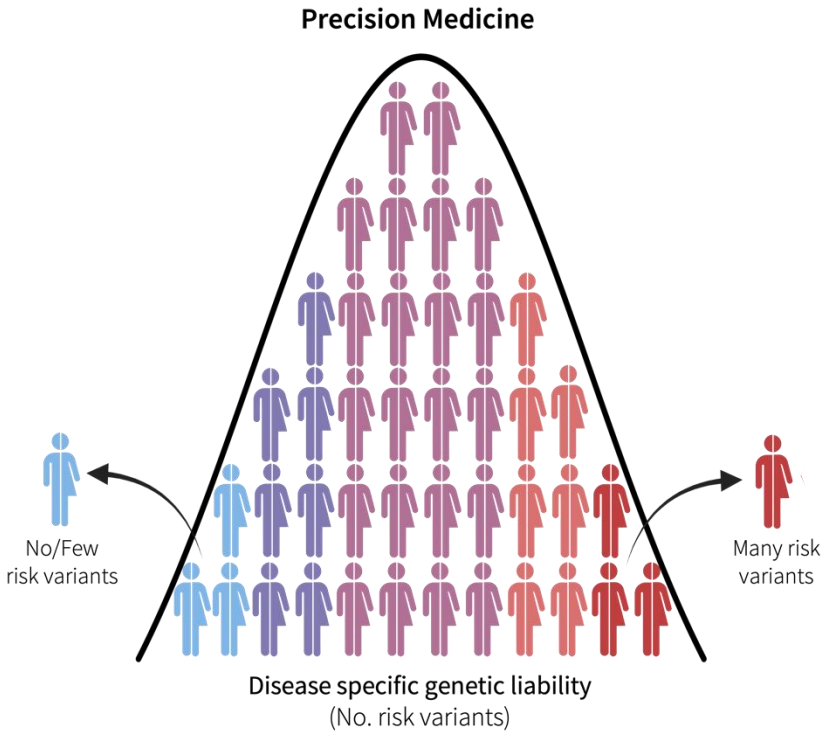
- The first polygenic score
- What you need is…
- Commonly used scoring algorithms
- Workflow
- Current challenges with PGS?

THE FIRST PGS

Idea originating back to animal and plant breeding - finding the ´best´ animals that should establish the next generation.

In 2007, Wray proposed to ‘predict’ human disease traits from SNP data.

Wray, Goddard & Visscher (2007) Genome Research, 17:1520–1528

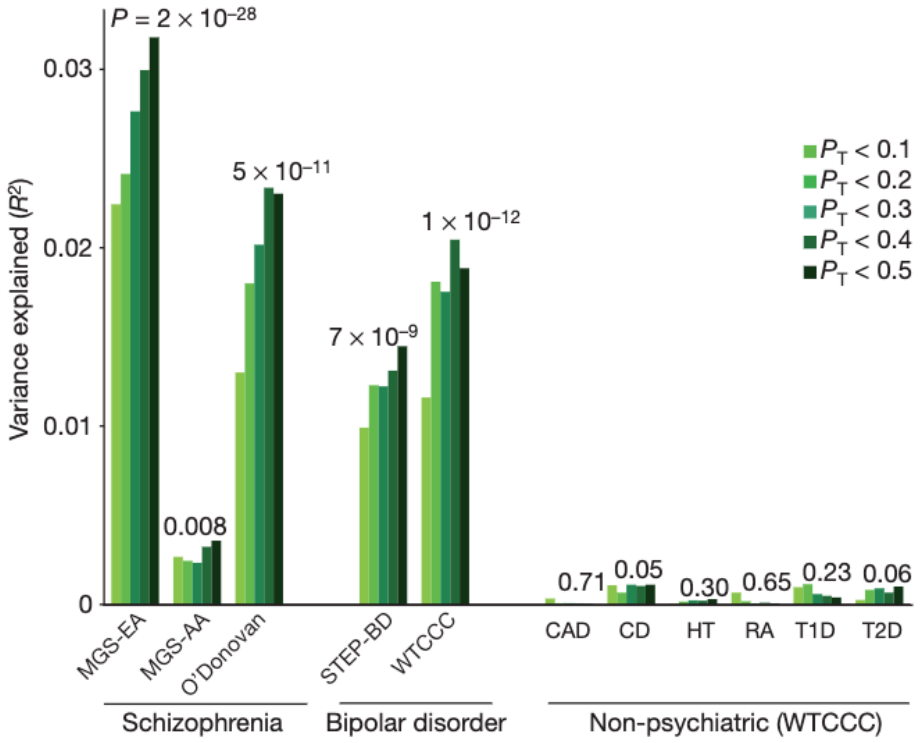


THE FIRST PGS

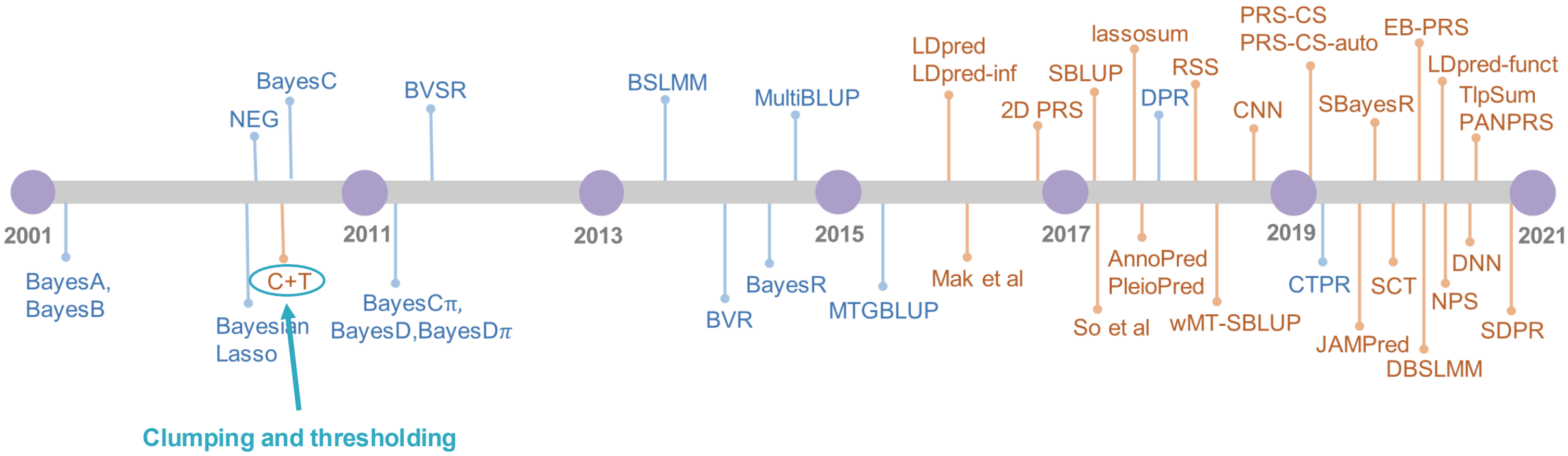
In 2009, Purcell constructed the first polygenic score for a human disease:

'*We summarized variation across nominally associated loci into quantitative scores*'

The International Schizophrenia Consortium (2009) Nature, 460



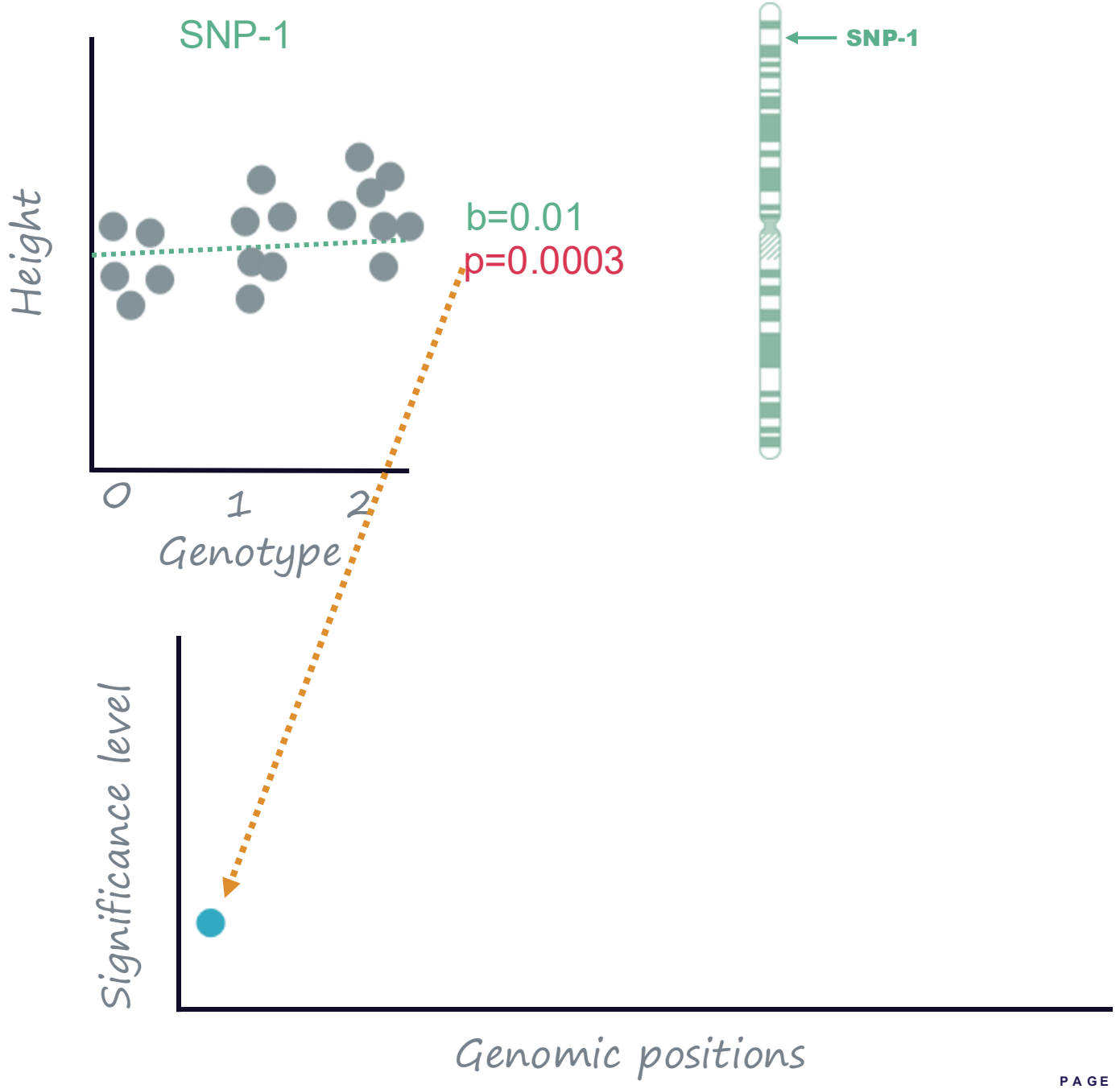
A LARGE PALETTE OF PGS METHODS



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

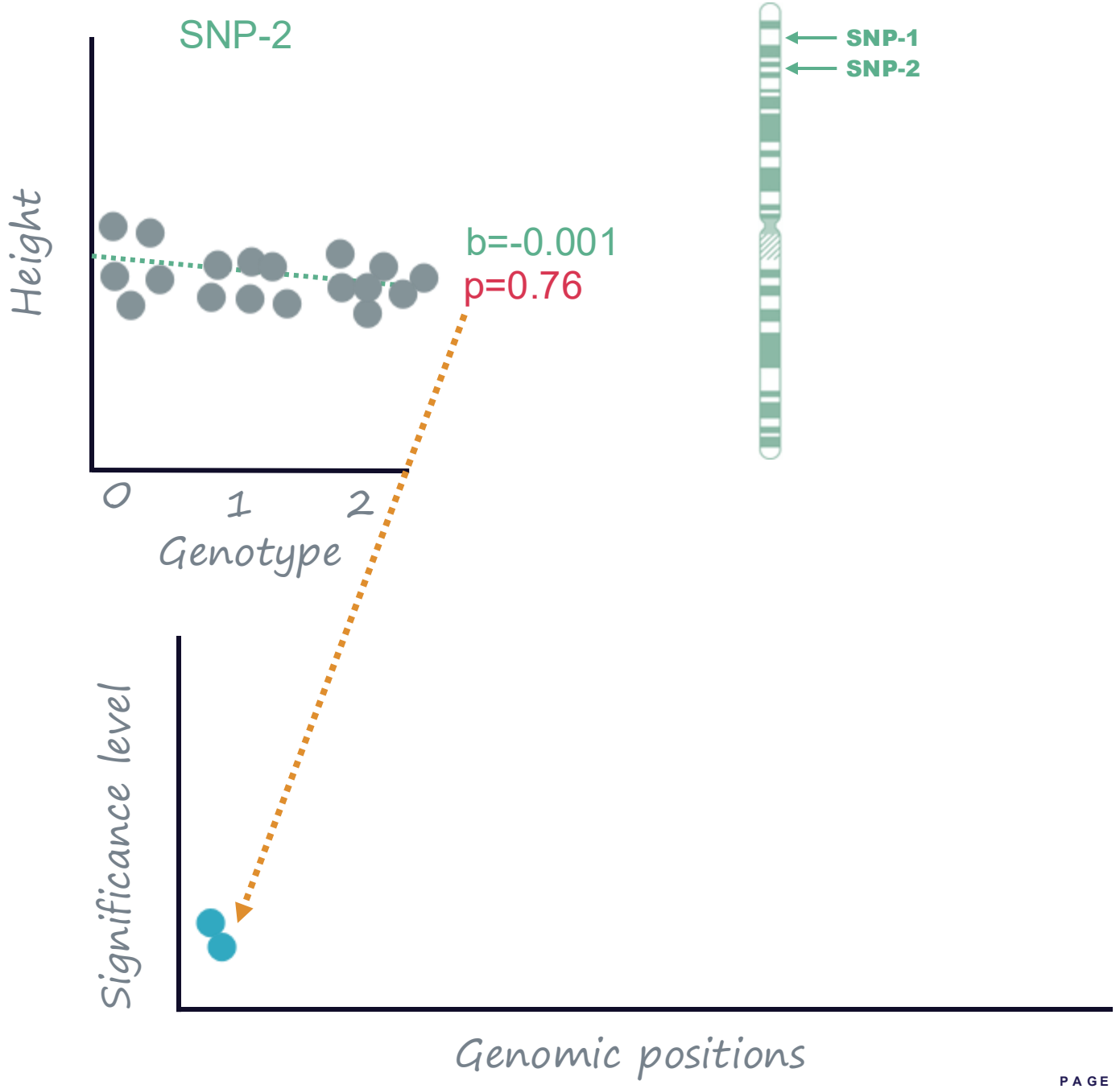
GWAS RECAP

Which SNPs associate with height?



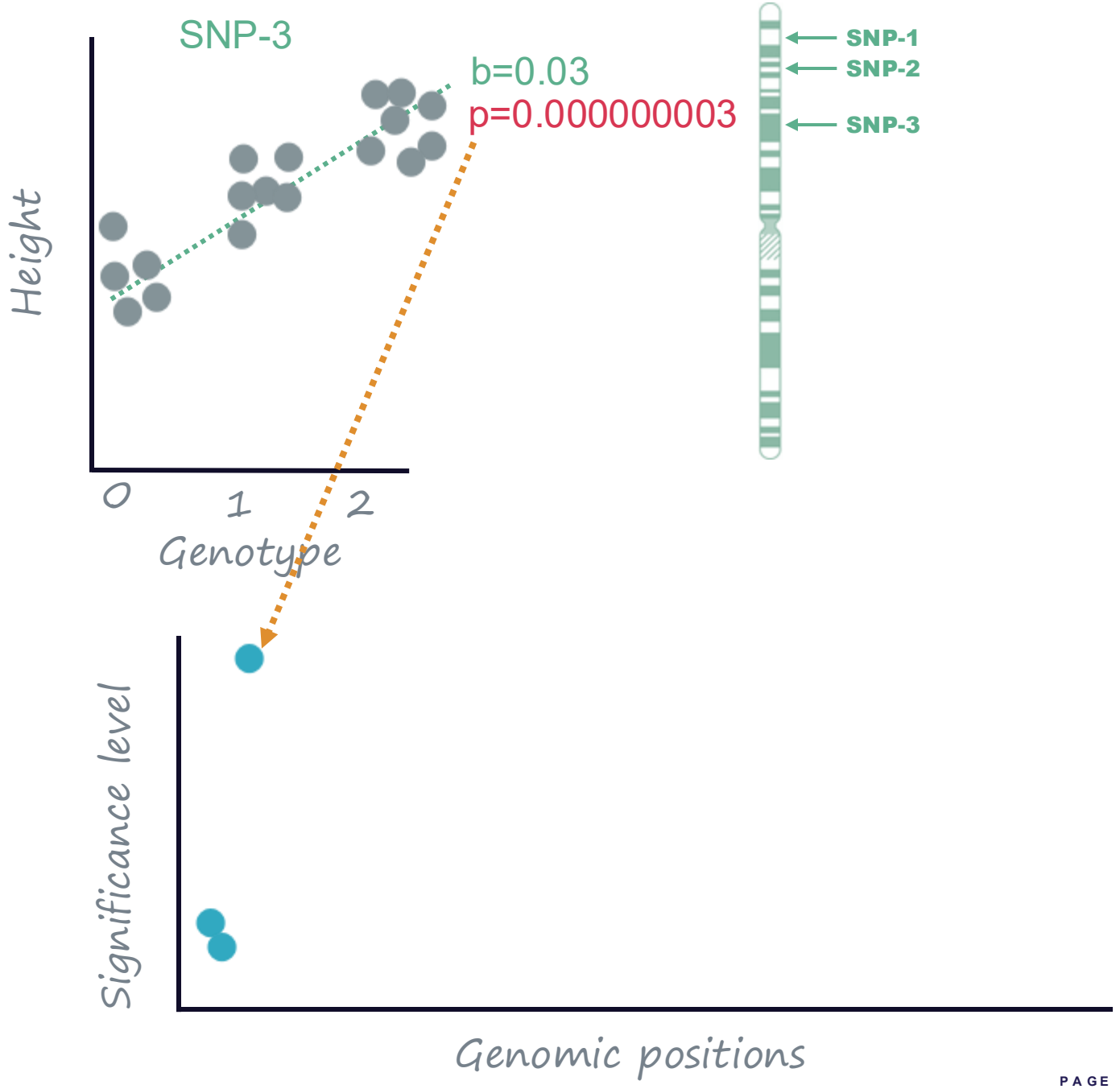
GWAS RECAP

Which SNPs associate with height?



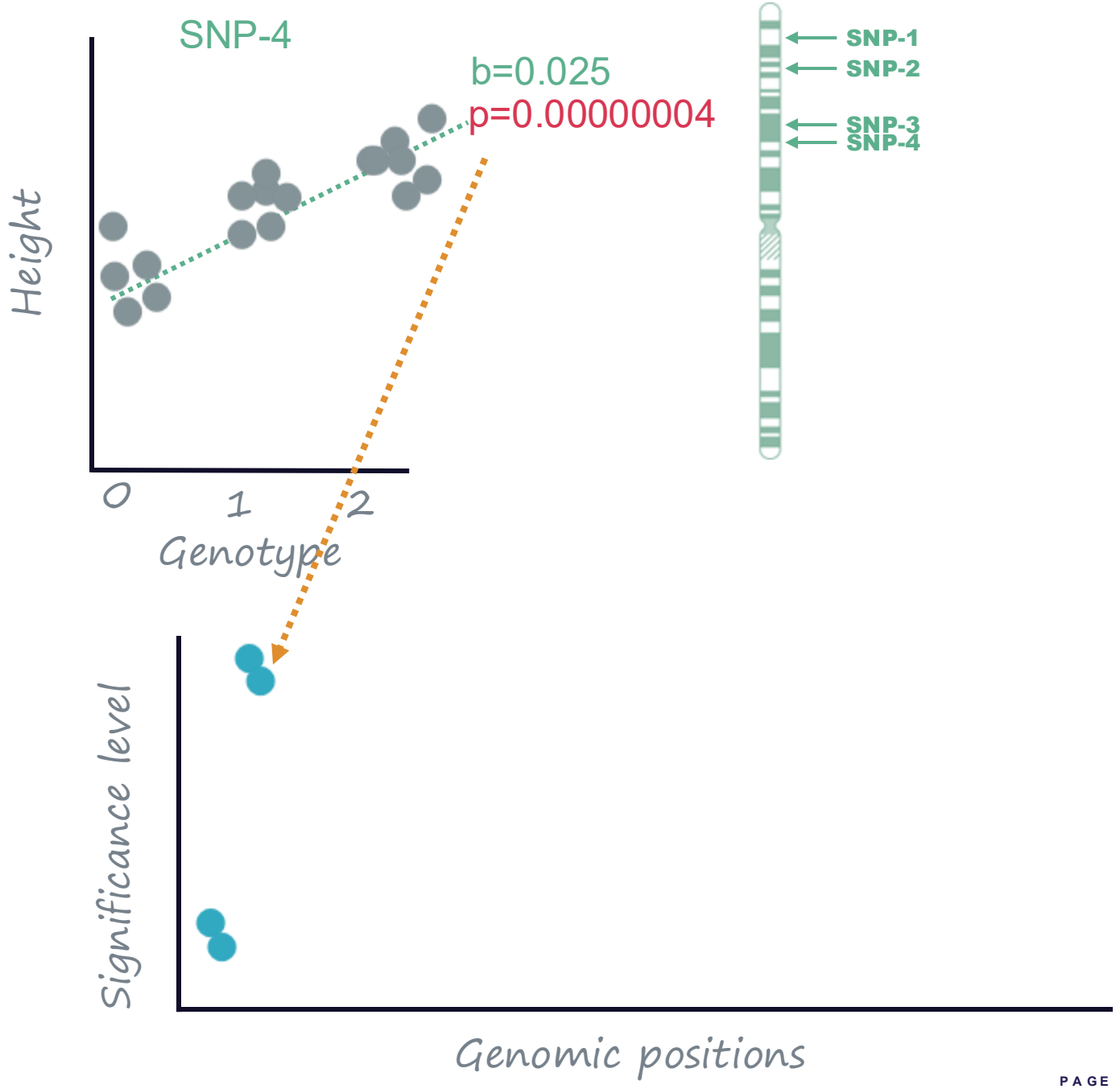
GWAS RECAP

Which SNPs associate with height?



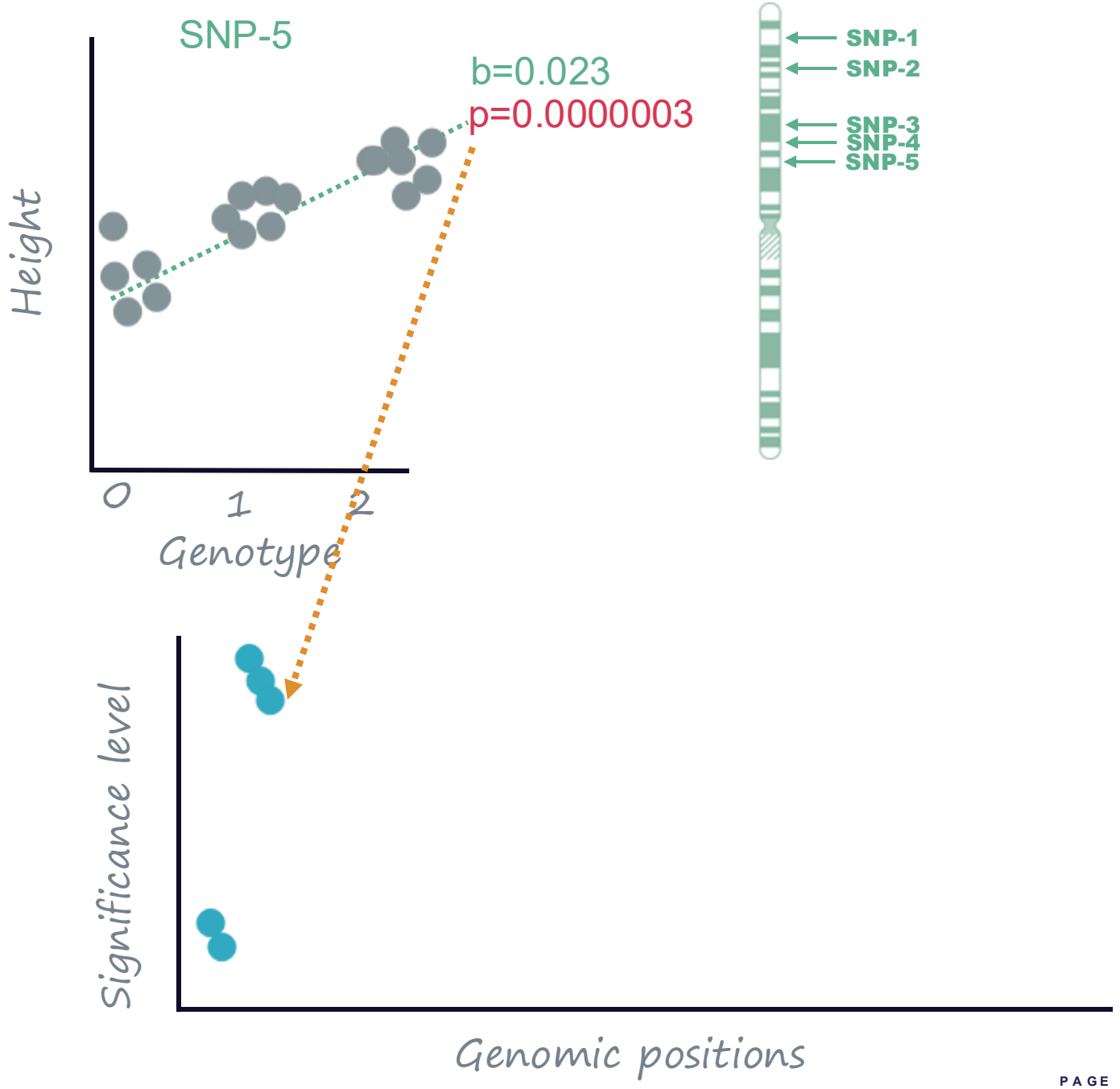
GWAS RECAP

Which SNPs associate with height?



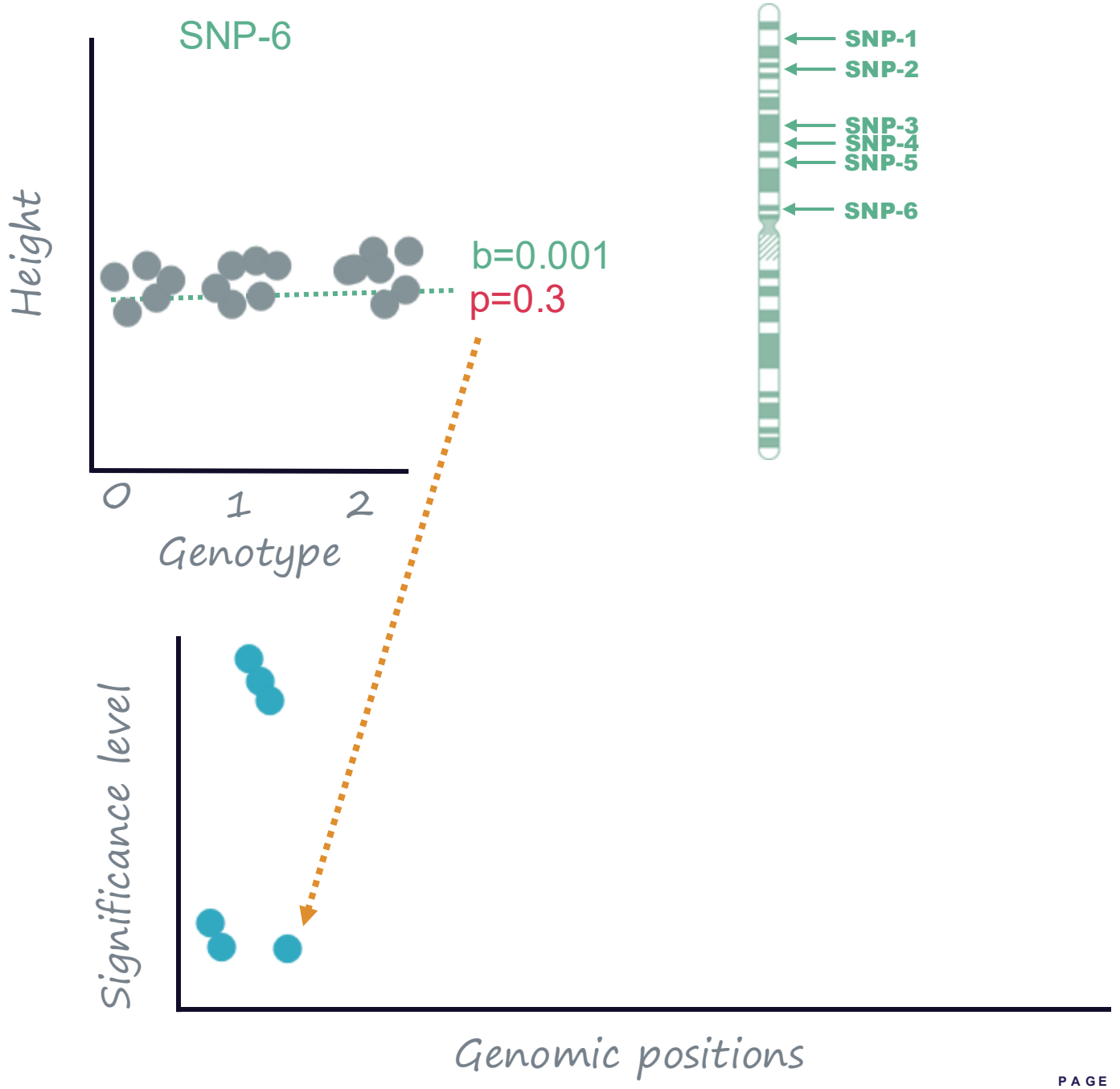
GWAS RECAP

Which SNPs associate with height?



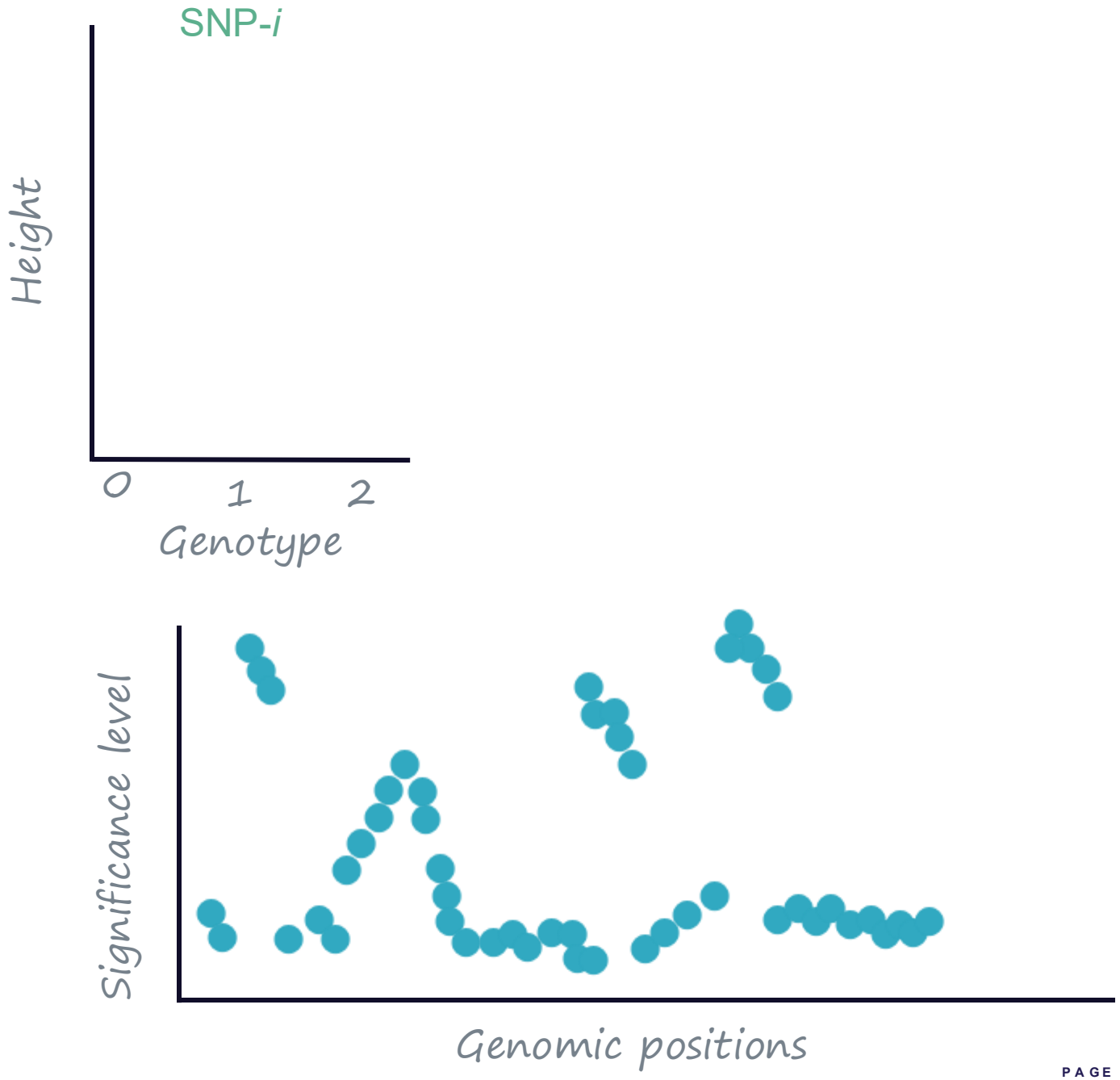
GWAS RECAP

Which SNPs associate with height?



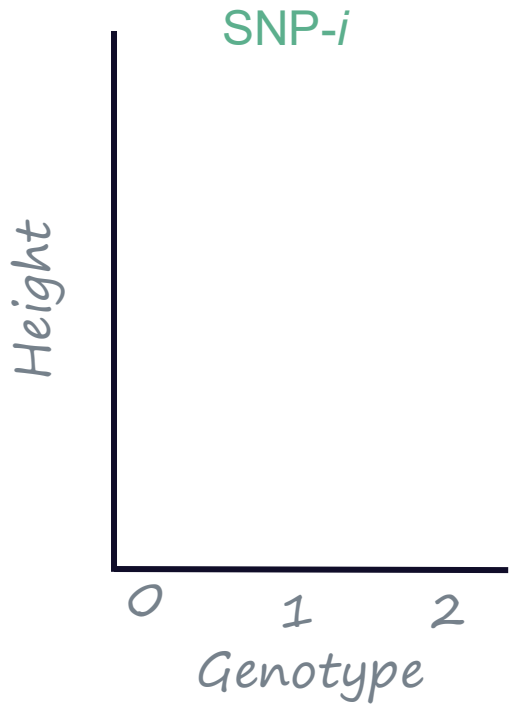
GWAS RECAP

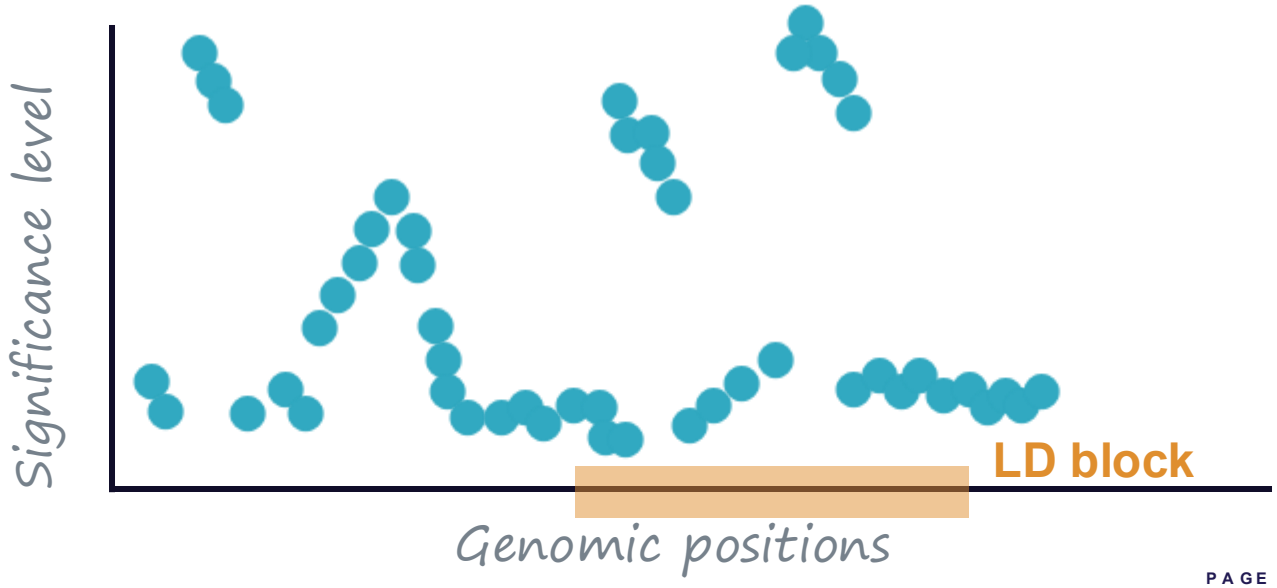
Which SNPs associate with height?



GWAS RECAP

Which SNPs associate with height?





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

$$PGS = \sum X_i b_i$$

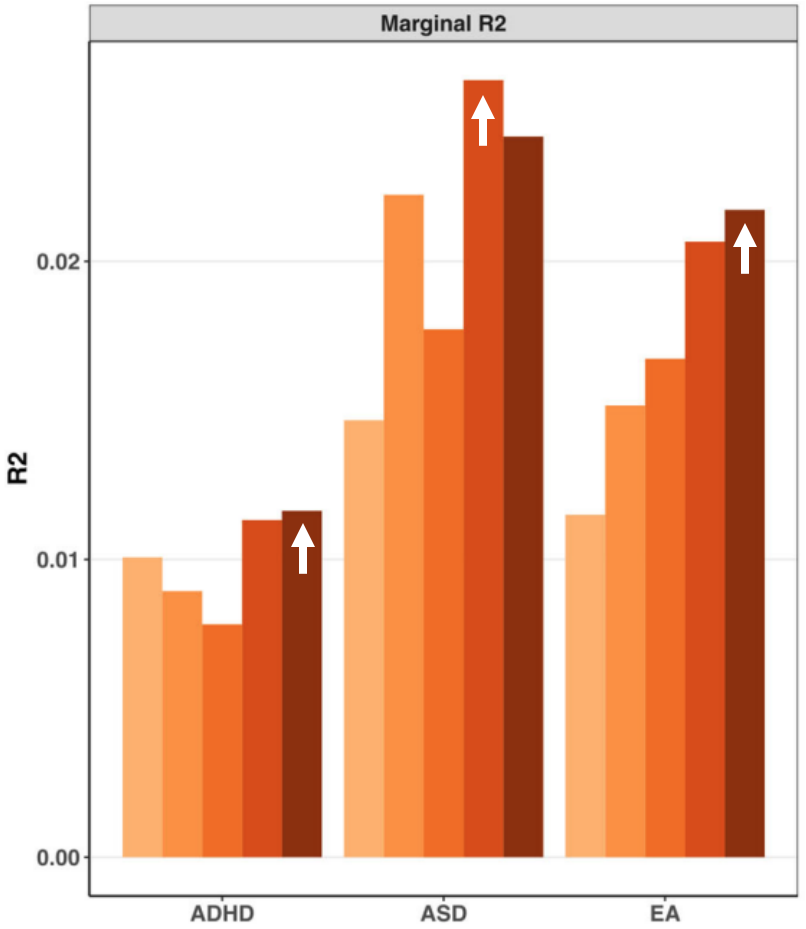
$$= \boldsymbol{X}_{11} \times 0.21 + \boldsymbol{X}_{9} \times 0.14$$

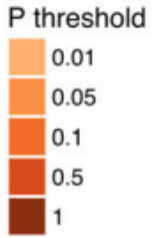
CLUMPING AND THRESHOLDING (C+T)

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

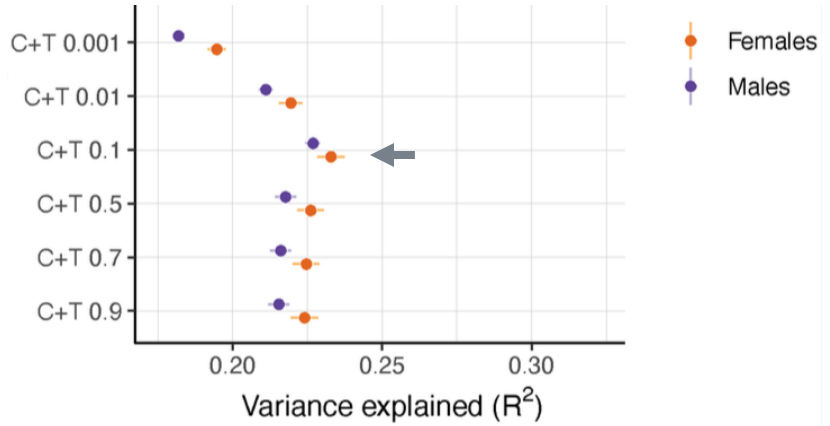
How does the PGS associate with the disease

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





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

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

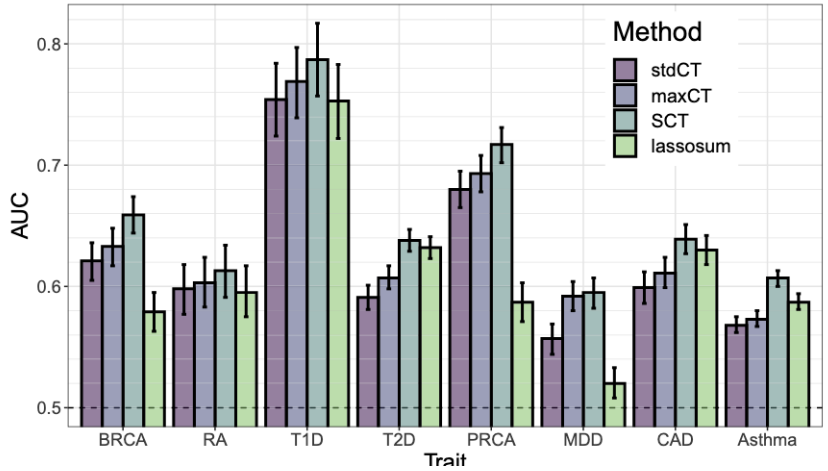


Table 2. Optimal Choices of C+T Parameters

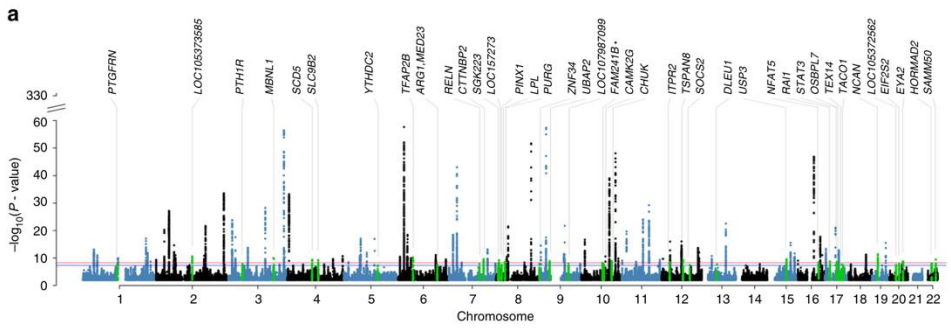
Trait	w_c	r_c^2	INFO_T	p_T
Breast cancer (BRCA)	2,500	0.2	0.95	2.2×10^{-4}
Rheumatoid arthritis (RA)	200	0.5	0.95	7.5×10^{-2}
Type 1 diabetes (T1D)	10K–50K	0.01	0.90	2.6×10^{-5}
Type 2 diabetes (T2D)	625	0.8	0.95	1.1×10^{-2}
Prostate cancer (PRCA)	10K–50K	0.01	0.90	4.2×10^{-6}
Depression (MDD)	625	0.8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

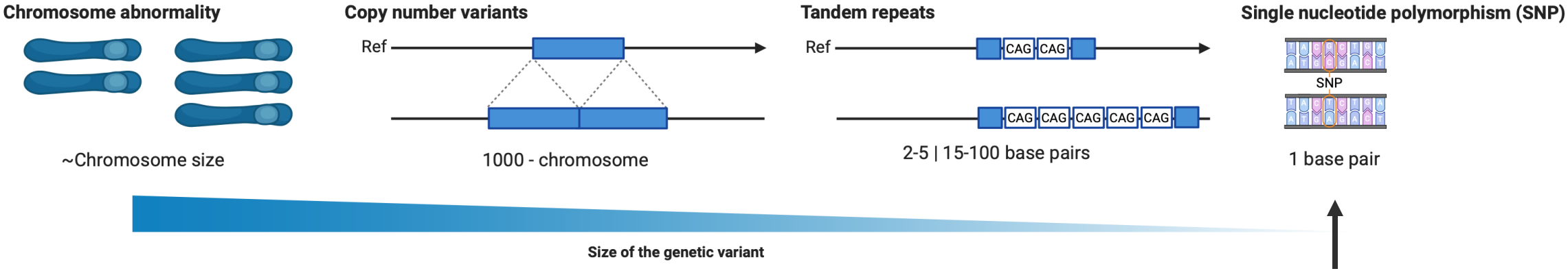
1. A large well-powered GWAS for your trait of interest



2. An independent cohort that has been genotyped

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

GENETIC DATA / GENETIC VARIATION

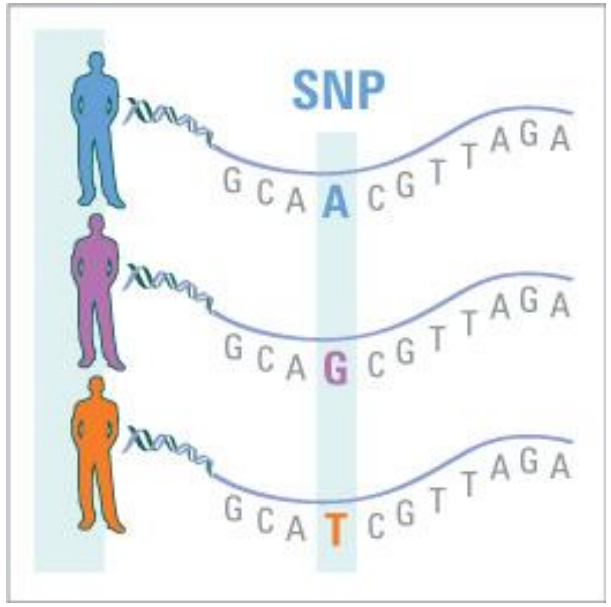


For PGS (GWAS) we are interested in SNP/SNV data

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

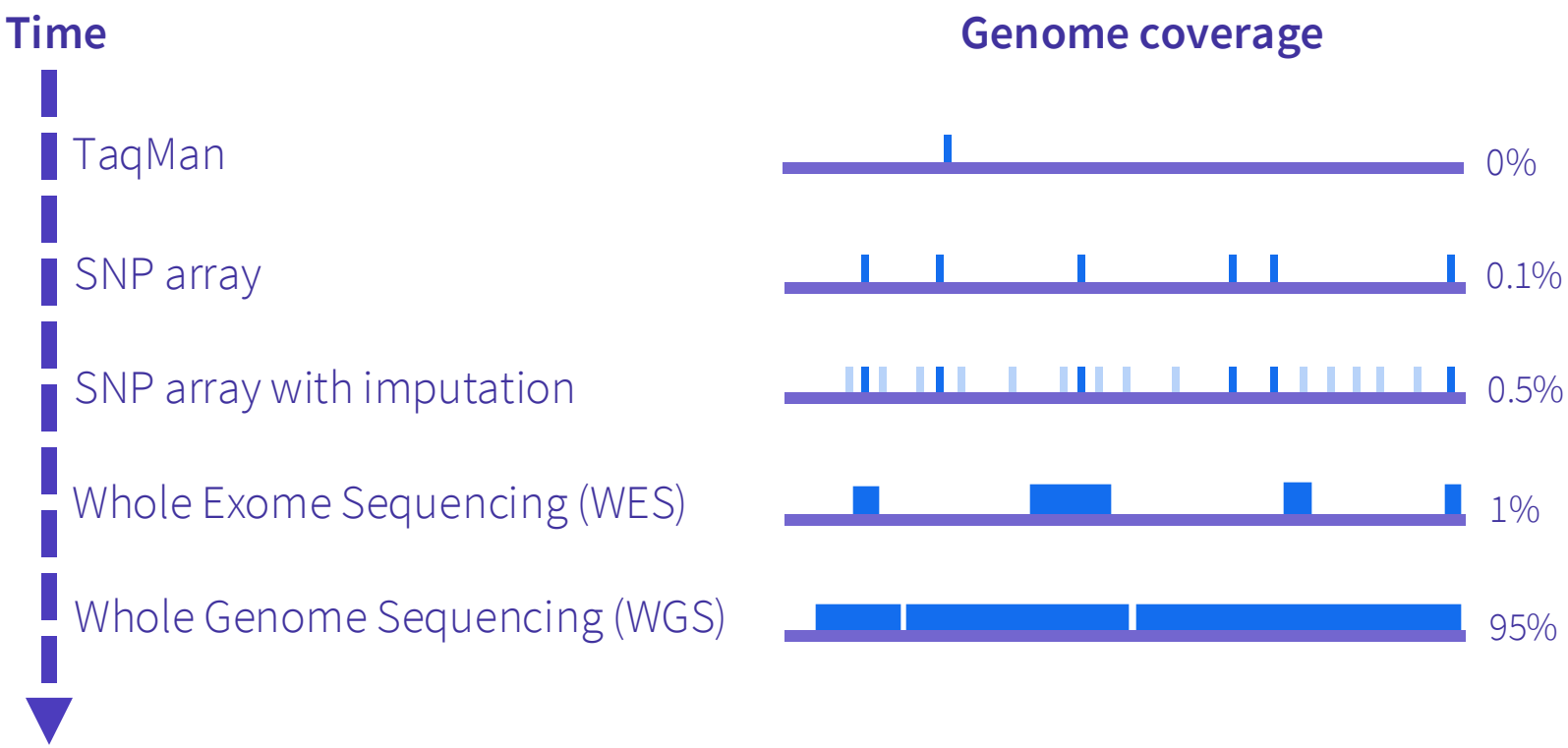


All (known) SNPs has a unique identifier

(independent of alleles)

rsXXX – Ref-SNP cluster ID number

GENOMIC COVERAGE

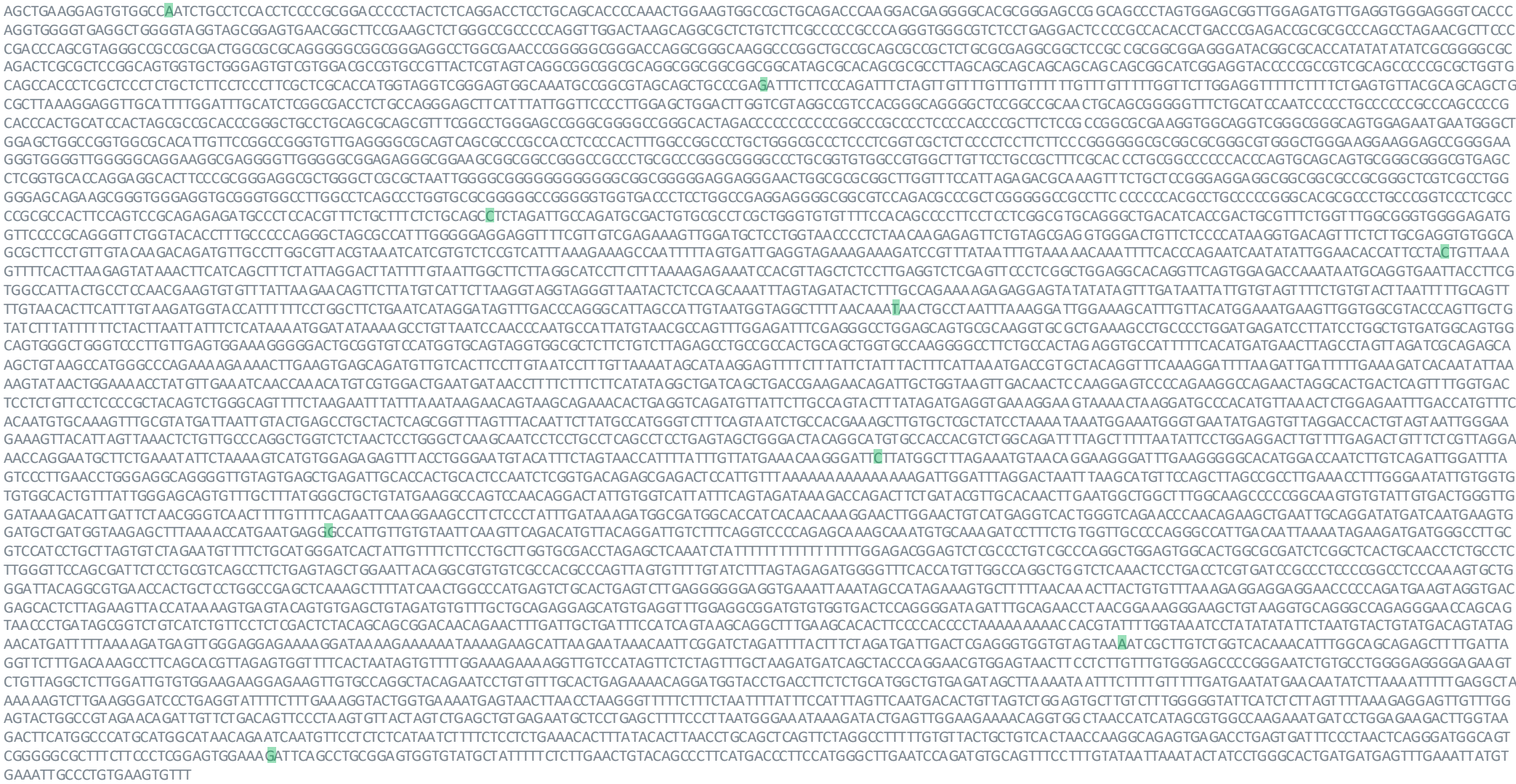


Adapted from Uitterlinden A. (2016) An Introduction to Genome-Wide Association Studies: GWAS for Dummies. Seminars in Reproductive Medicine, 34(4): 196-204.

GENOTYPING VS SEQUENCING

GENOTYPING VS SEQUENCING

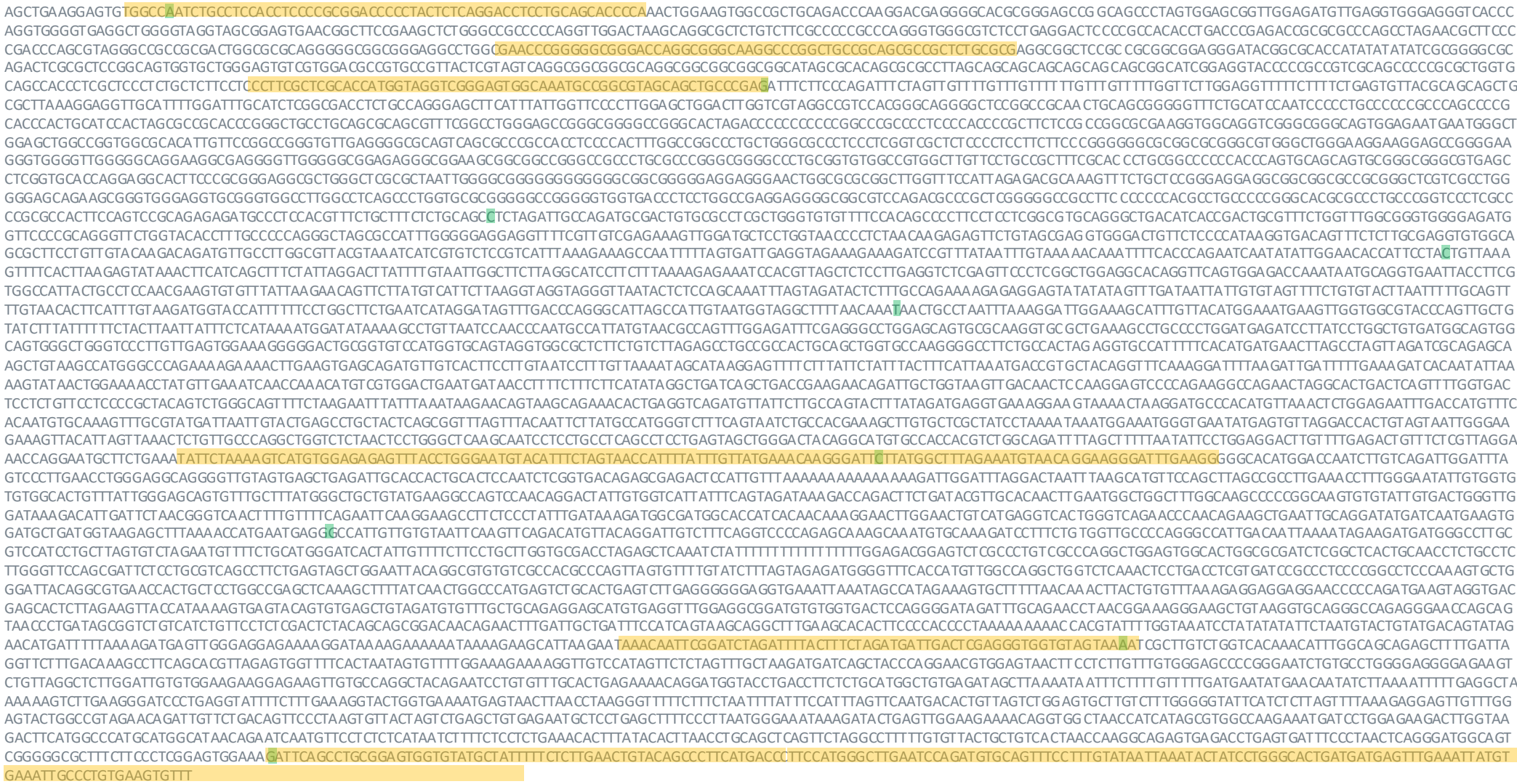
Genotyping



GENOTYPING VS SEQUENCING

Genotyping

WES

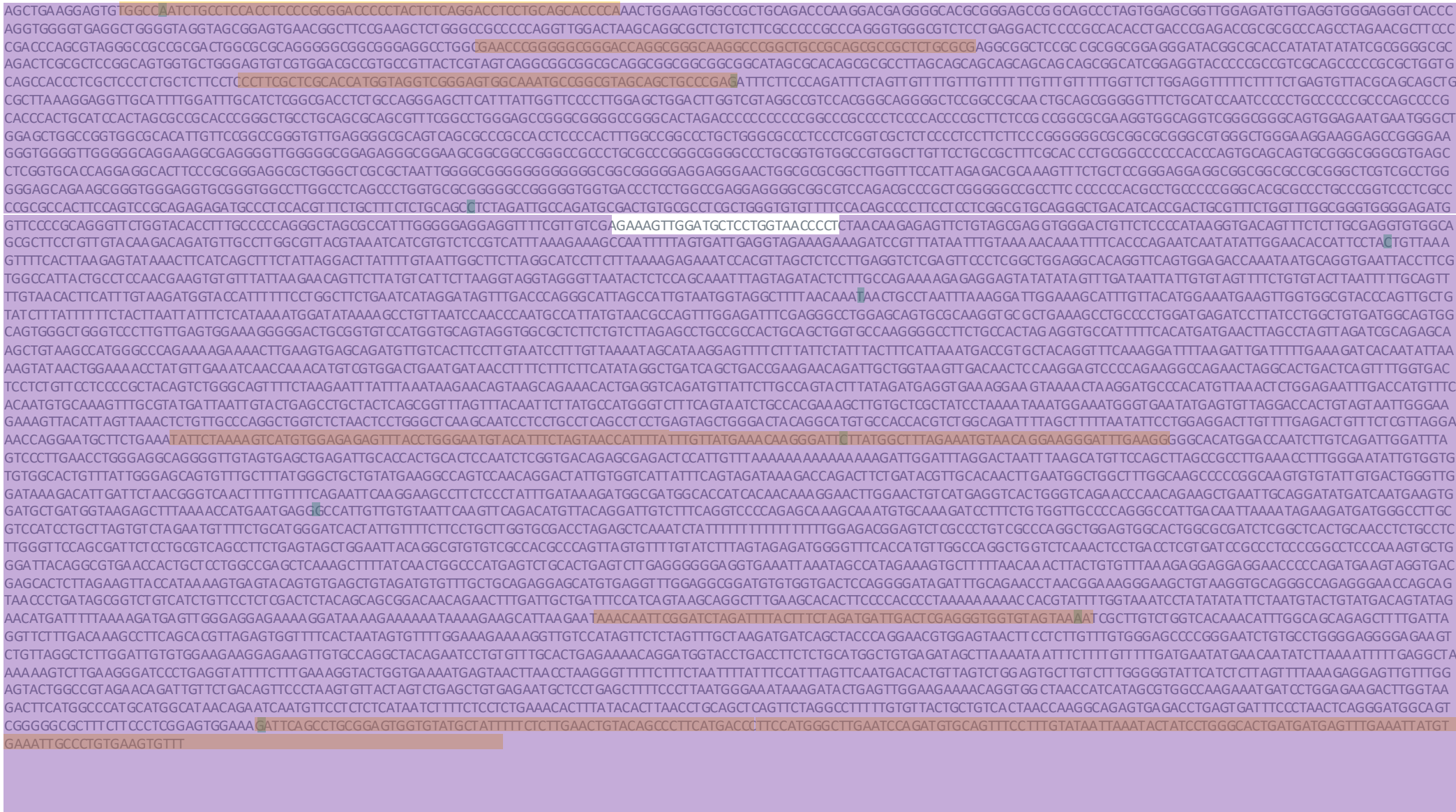


GENOTYPING VS SEQUENCING

Genotyping

WES

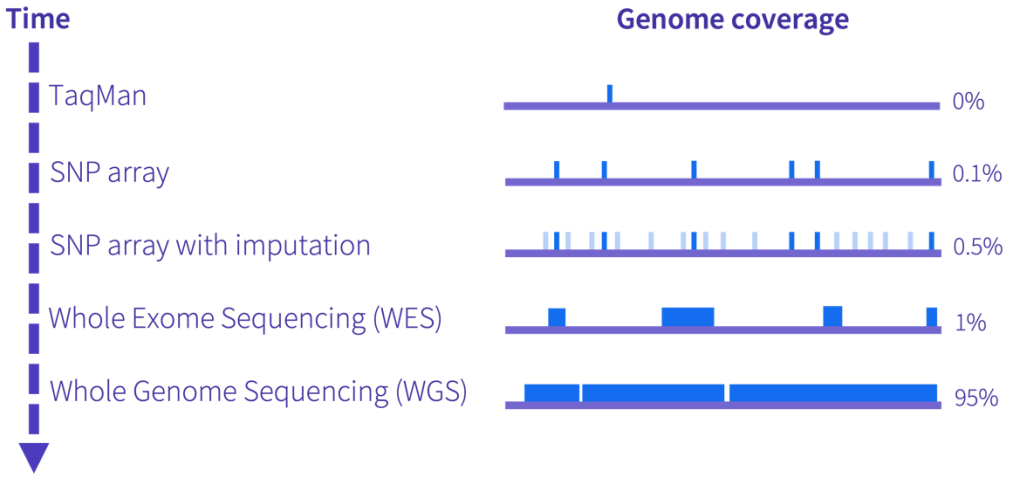
WGS



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

Because of LD you do not have to analyse all 3,000,000,000 variants in the genome.

Typically, we genotype ½ - 1 million variants

Because of LD we can impute (“guess” what variants are next to the genotyped variant) up til 10 million common genetic variants.

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

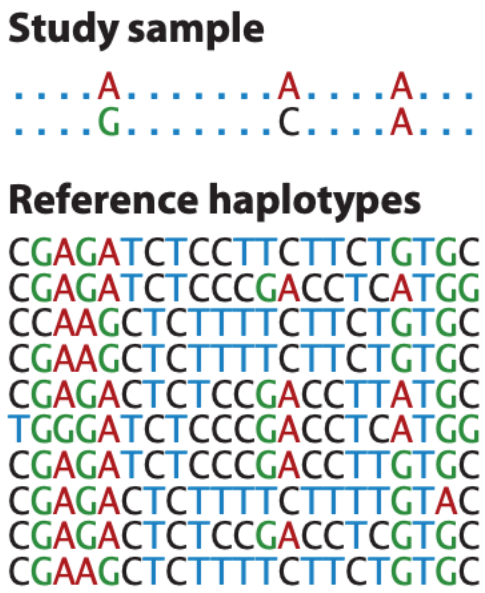
Different haplotypes

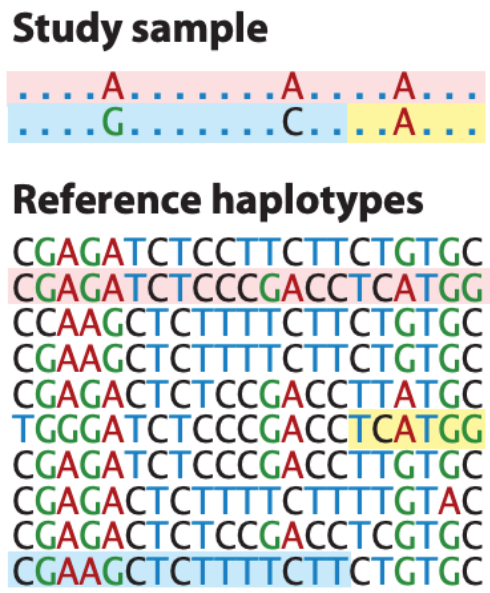
A	C	A
A	C	C
A	T	A
A	C	A
G	C	A
G	C	C
G	T	A
G	T	C

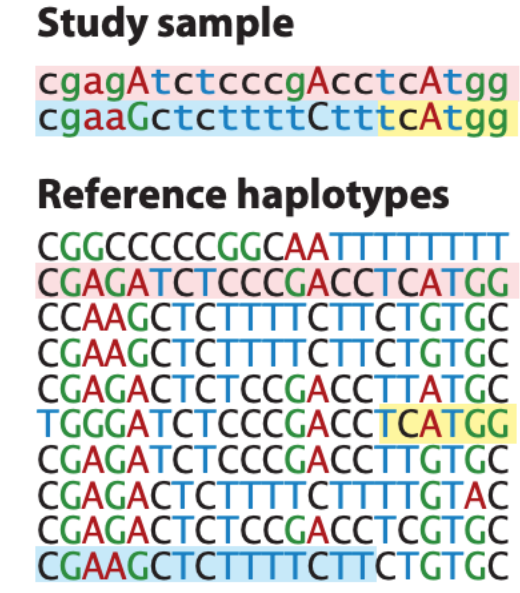
Phasing = estimate the most likely haplotypes

IMPUTATION

Genotypes





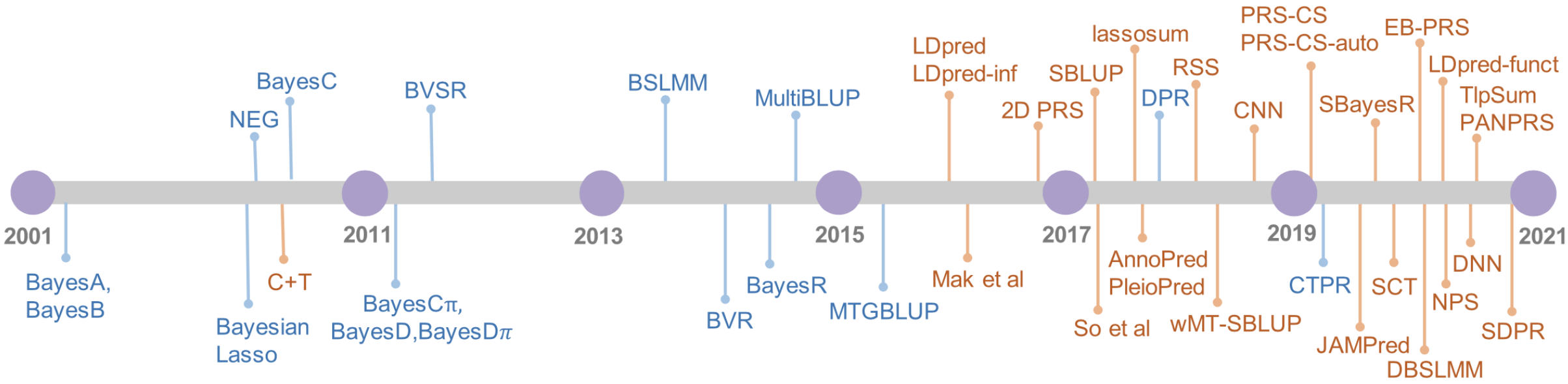


From a sequencing study

A LARGE PALETTE OF PGS METHODS

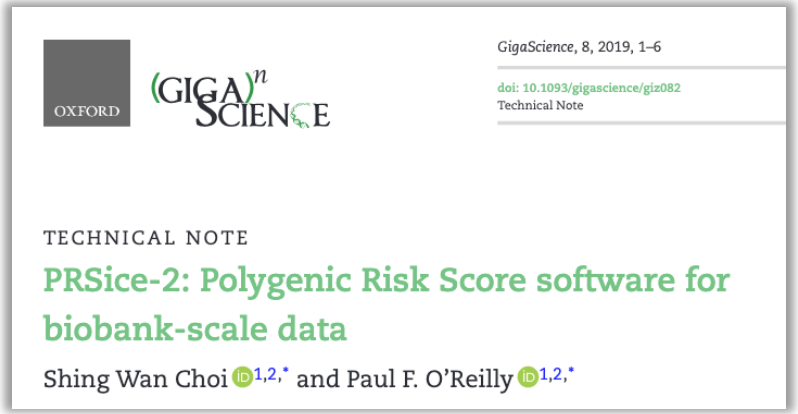
-- Require individual level data --
(can be tricky because of GDPR)

-- Require summary data --



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

COMMONLY USED METHODS



OXFORD (GIGA)n SCIENCE

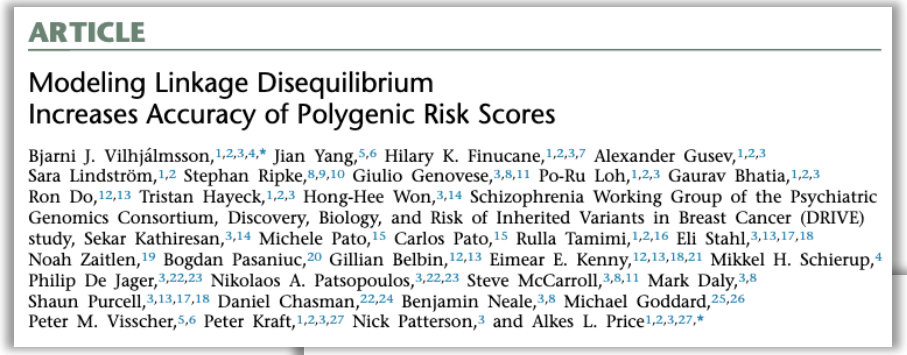
GigaScience, 8, 2019, 1–6

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]

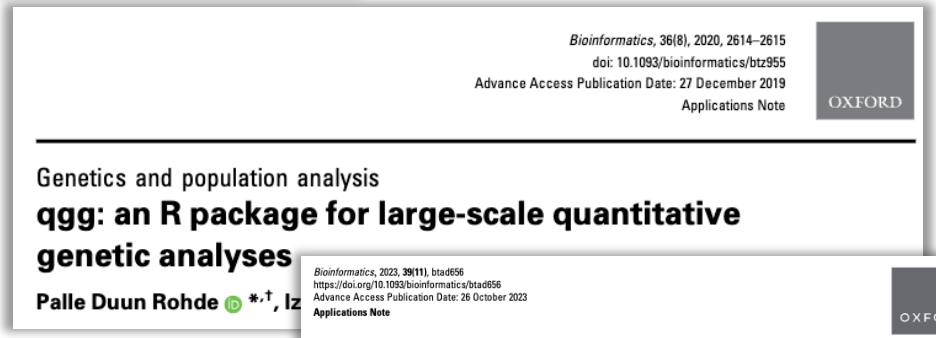
nature COMMUNICATIONS

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

COMMONLY USED METHODS

Different schrinkage methods

Clumping and thresholding (C+T)

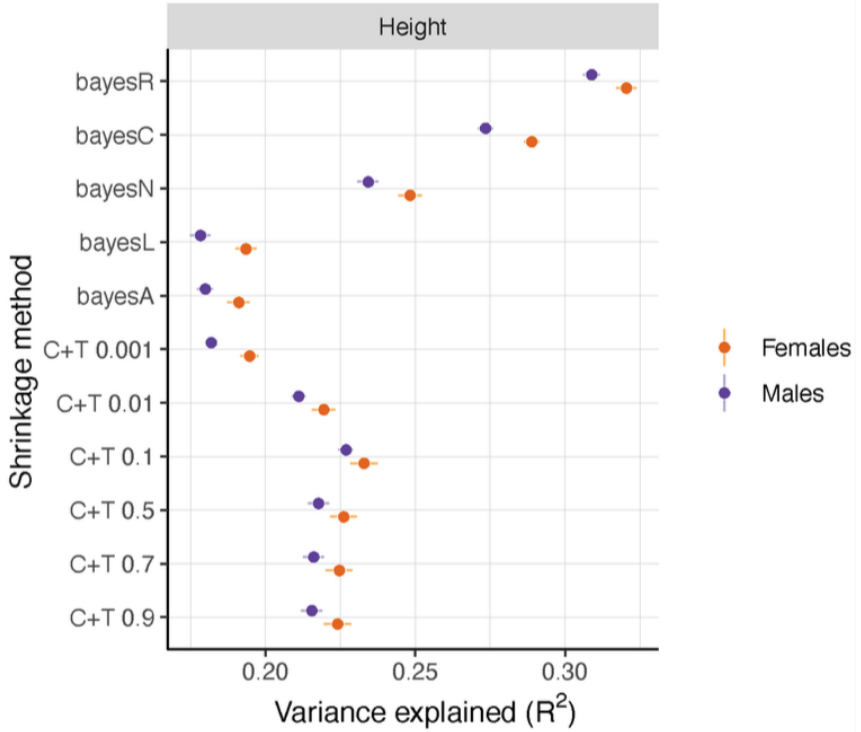
Bayes-N; $\beta \sim N(0, \sigma^2_\beta)$

Bayes-L; $f(\beta_j|\tau^2_j, \sigma^2_e) \sim N(\beta_j|0, \tau^2_j \times \sigma^2_e)$

Bayes-A; $\beta_j \sim N(0, \sigma^2_{\beta_j})$

Bayes-C; $\beta_j \sim N(0, \sigma^2_{\beta_j})$ with probability π, and $\beta_j = 0$ with probability $(1 - \pi)$, where π is assumed to follow a beta-distribution.

Bayes-R; $\beta_j \sim N(0, \gamma_C \sigma^2_{\beta_j})$, where C defines number of classes (e.g., C=4,$\gamma = (0, 0.01, 0.1, 1.0)$)




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

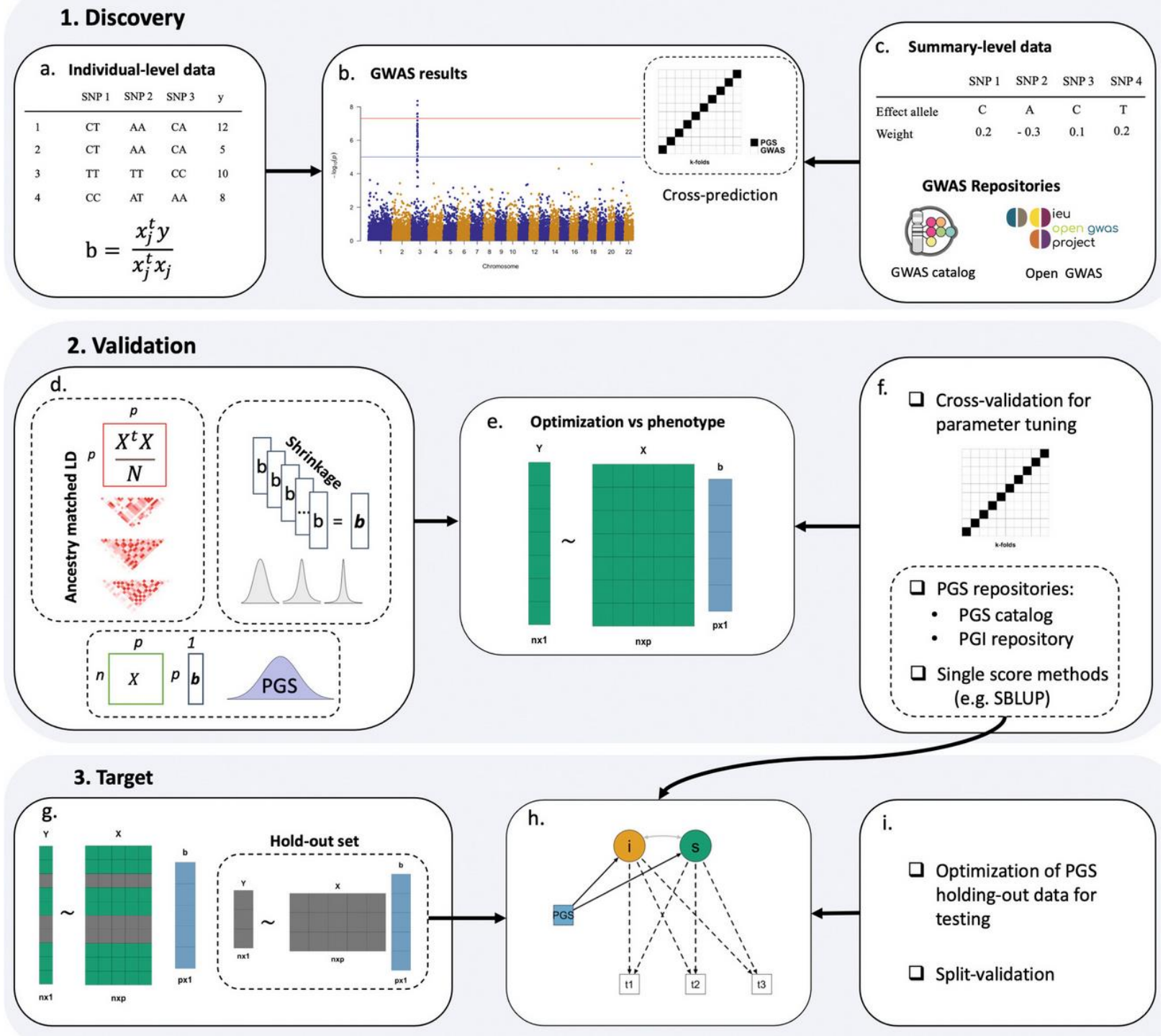
OXFORD

Genetics and population analysis

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

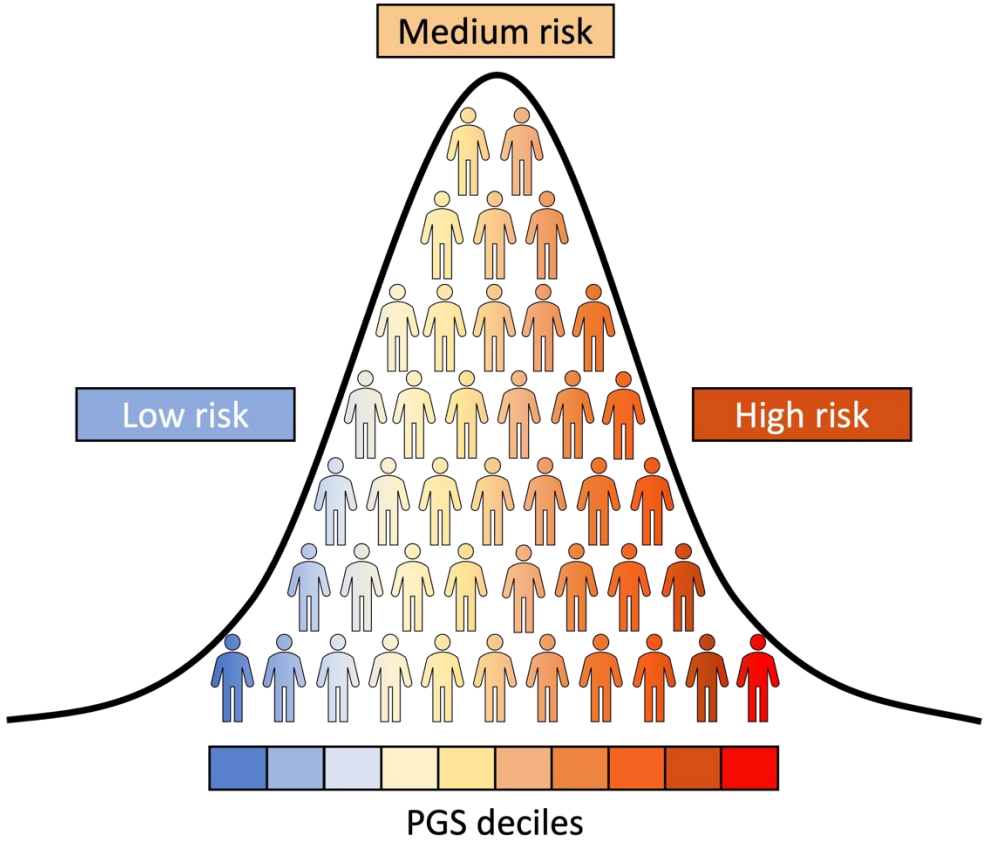
Palle Duun Rohde [1,*], **Izel Fourie Sørensen**[2], **Peter Sørensen**[2,*]

https://pdrohde.github.io/

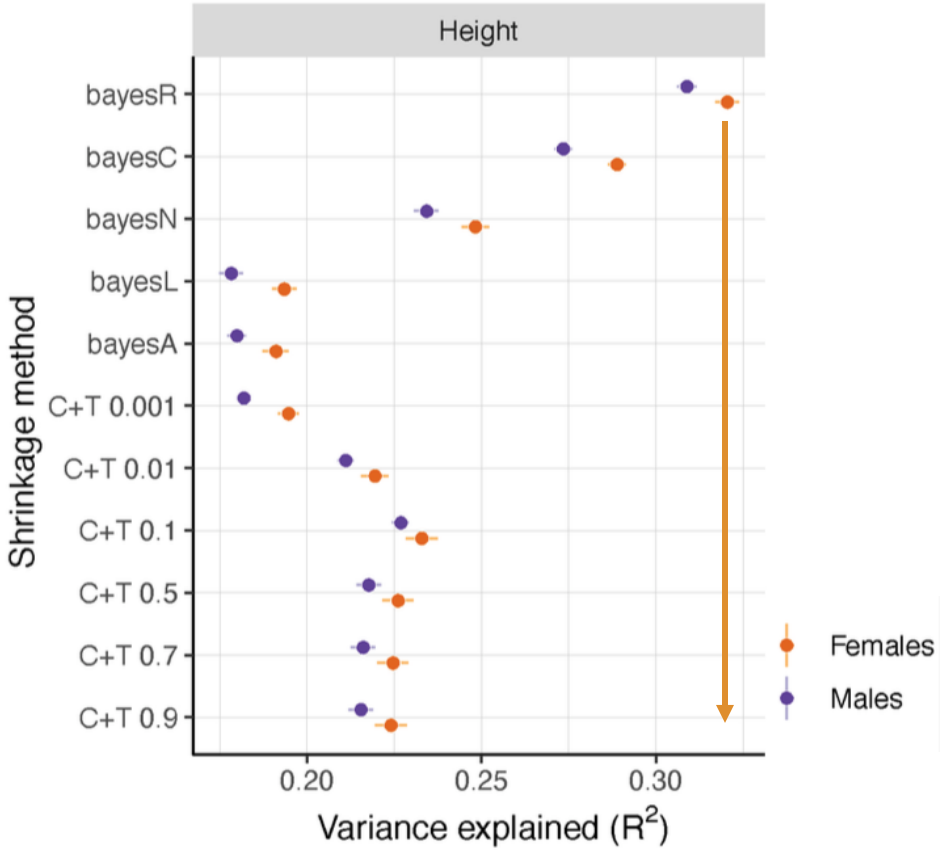


Allegrini et al (2022) J Child Psych, 63:1111-1124

CHALLENGES WITH PGS



HOW GOOD ARE PGS?



For human height the best PGS has an accuracy of ~40%.

Human height has a heritability (h^2) of ~65%.

The heritability is the boundary of how accurately we can predict the phenotype, thus, for human height the maximum accuracy is expected to be 65%.

Thus, for human height we currently lack 30% variance explained

Several reasons/challenges for lack of predictive ability

CHALLENGE 1

TOO MANY WITH ”AVERAGE” SCORE

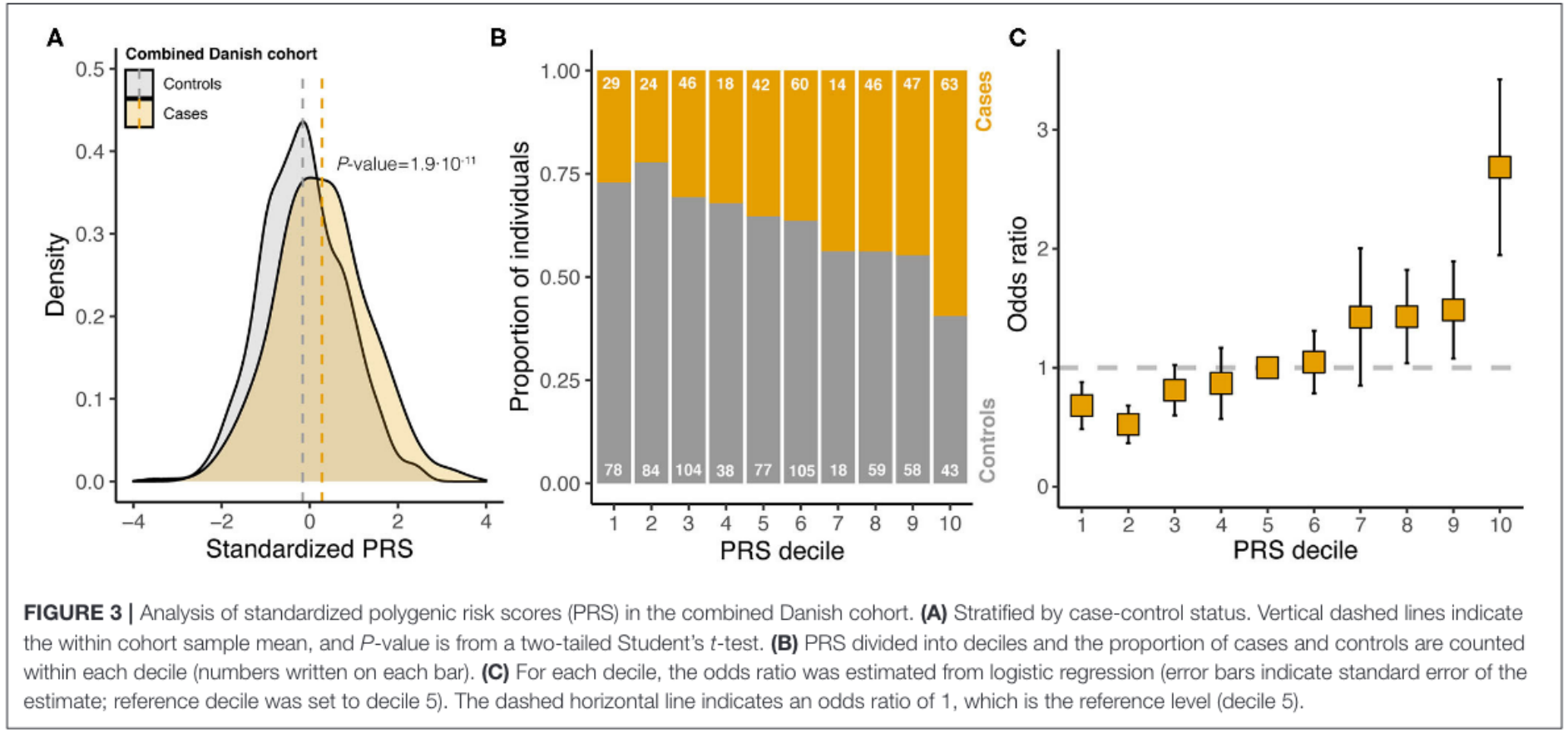


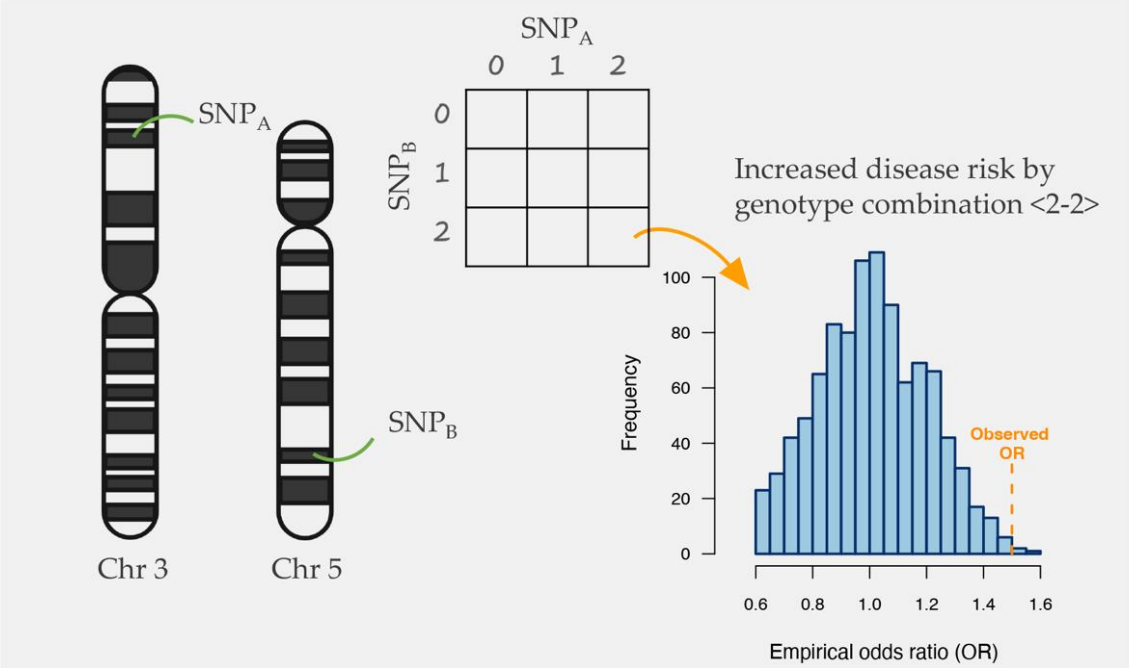
FIGURE 3 | Analysis of standardized polygenic risk scores (PRS) in the combined Danish cohort. **(A)** Stratified by case-control status. Vertical dashed lines indicate the within cohort sample mean, and *P*-value is from a two-tailed Student's *t*-test. **(B)** PRS divided into deciles and the proportion of cases and controls are counted within each decile (numbers written on each bar). **(C)** For each decile, the odds ratio was estimated from logistic regression (error bars indicate standard error of the estimate; reference decile was set to decile 5). The dashed horizontal line indicates an odds ratio of 1, which is the reference level (decile 5).

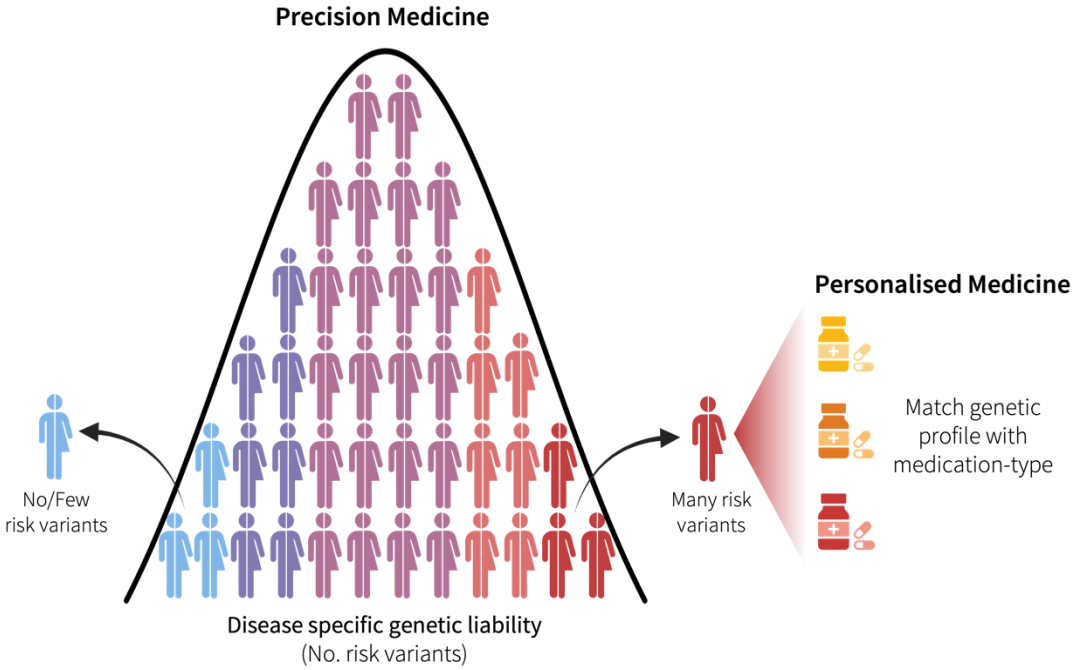
Kloeve-mogensen, Rohde, et al 2021, Front Repro Health

$$PGS = \sum X_i b_i$$

CHALLENGE 2

PGS ARE BASED ON ADDITIVE EFFECTS

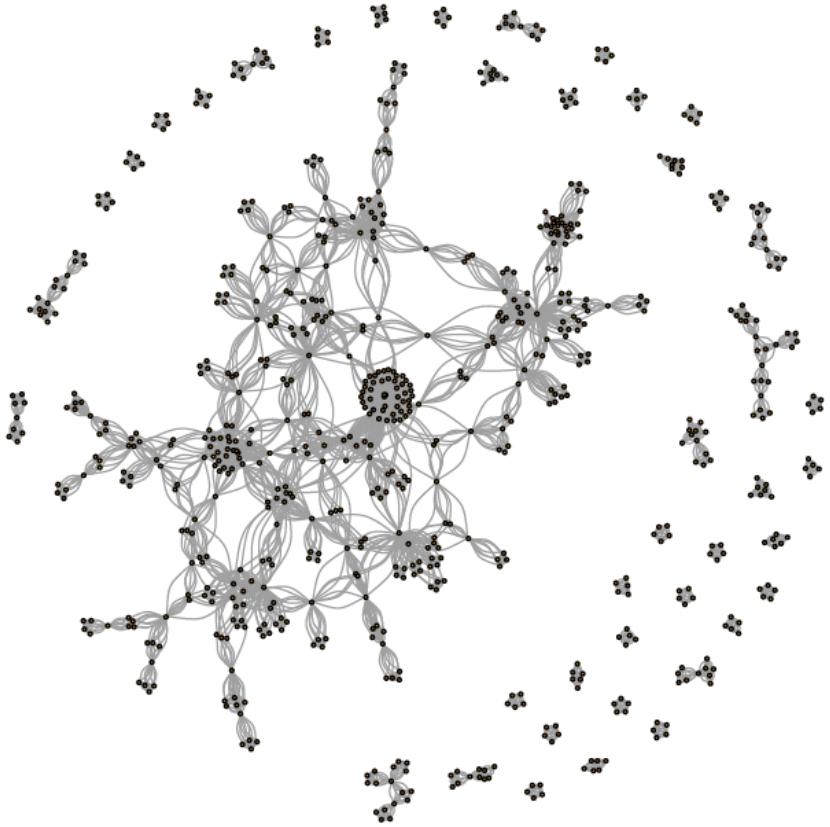


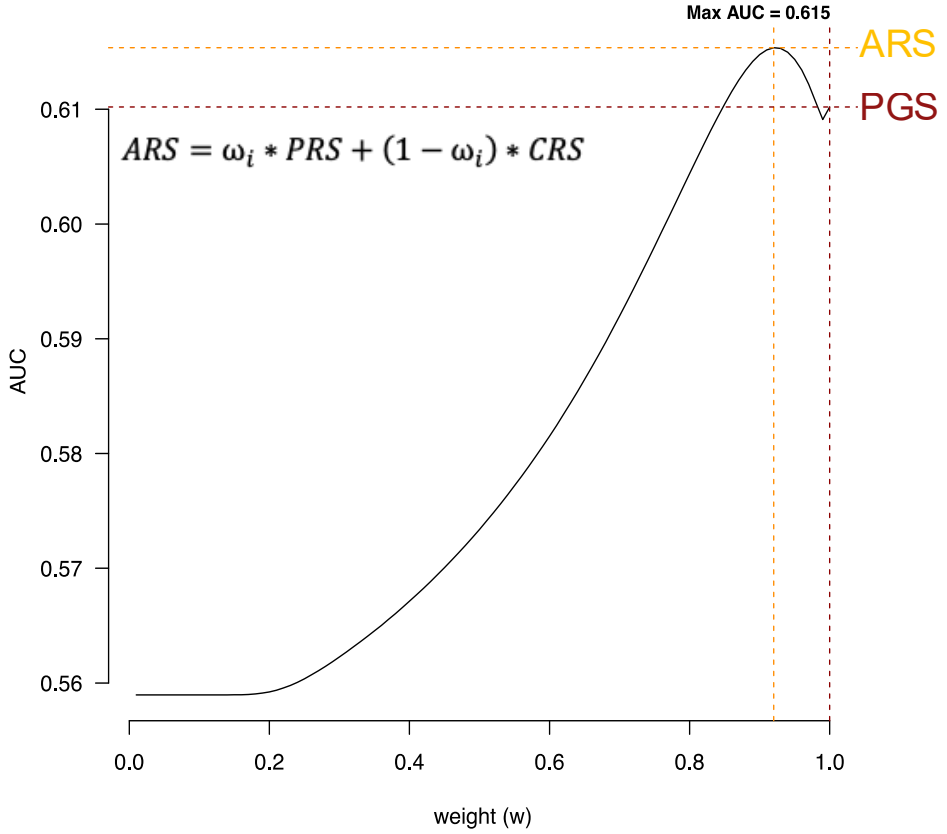


$$PGS = \sum X_i b_i$$

CHALLENGE 2

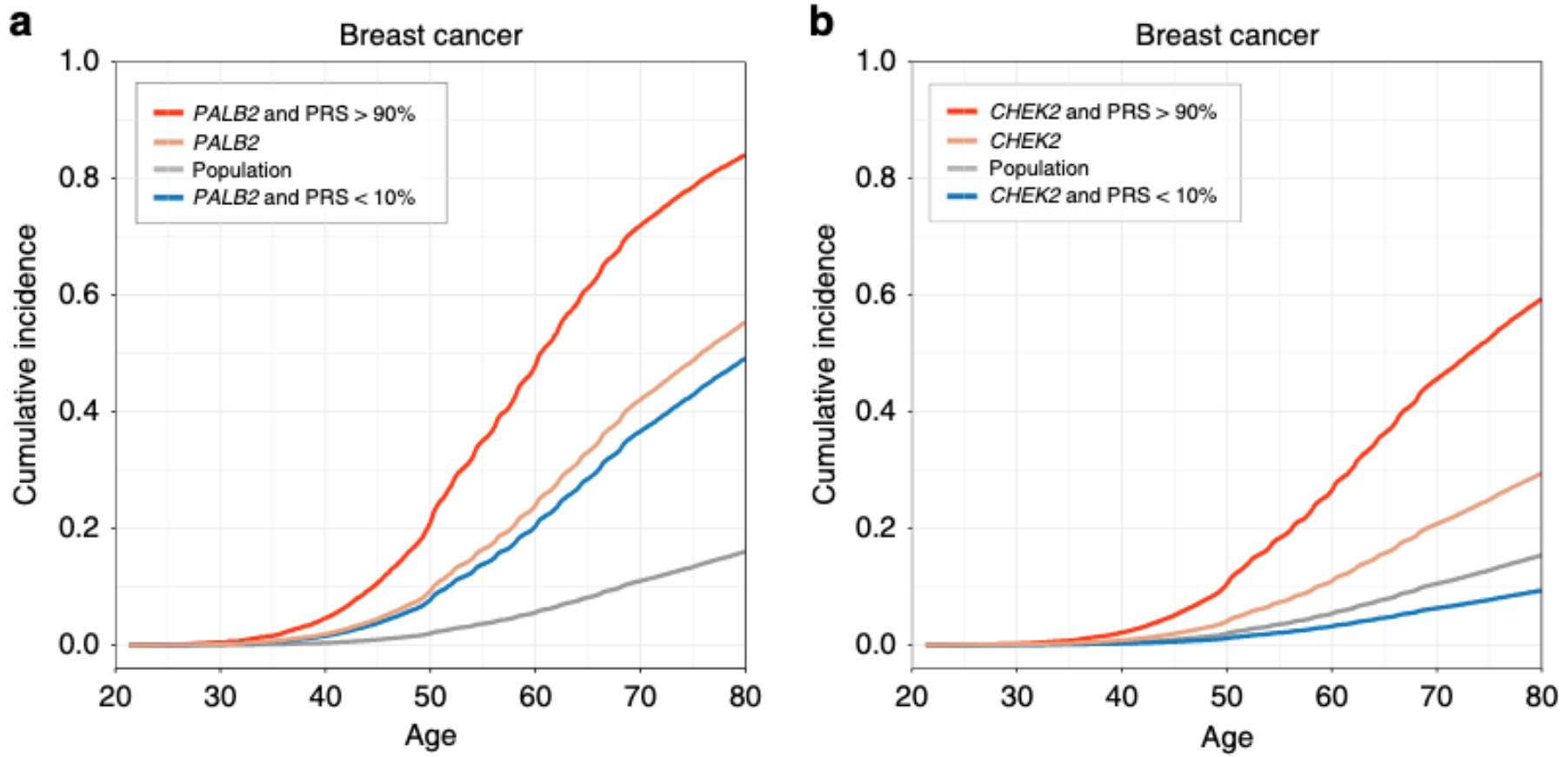
PGS ARE BASED ON ADDITIVE EFFECTS





CHALLENGE 3

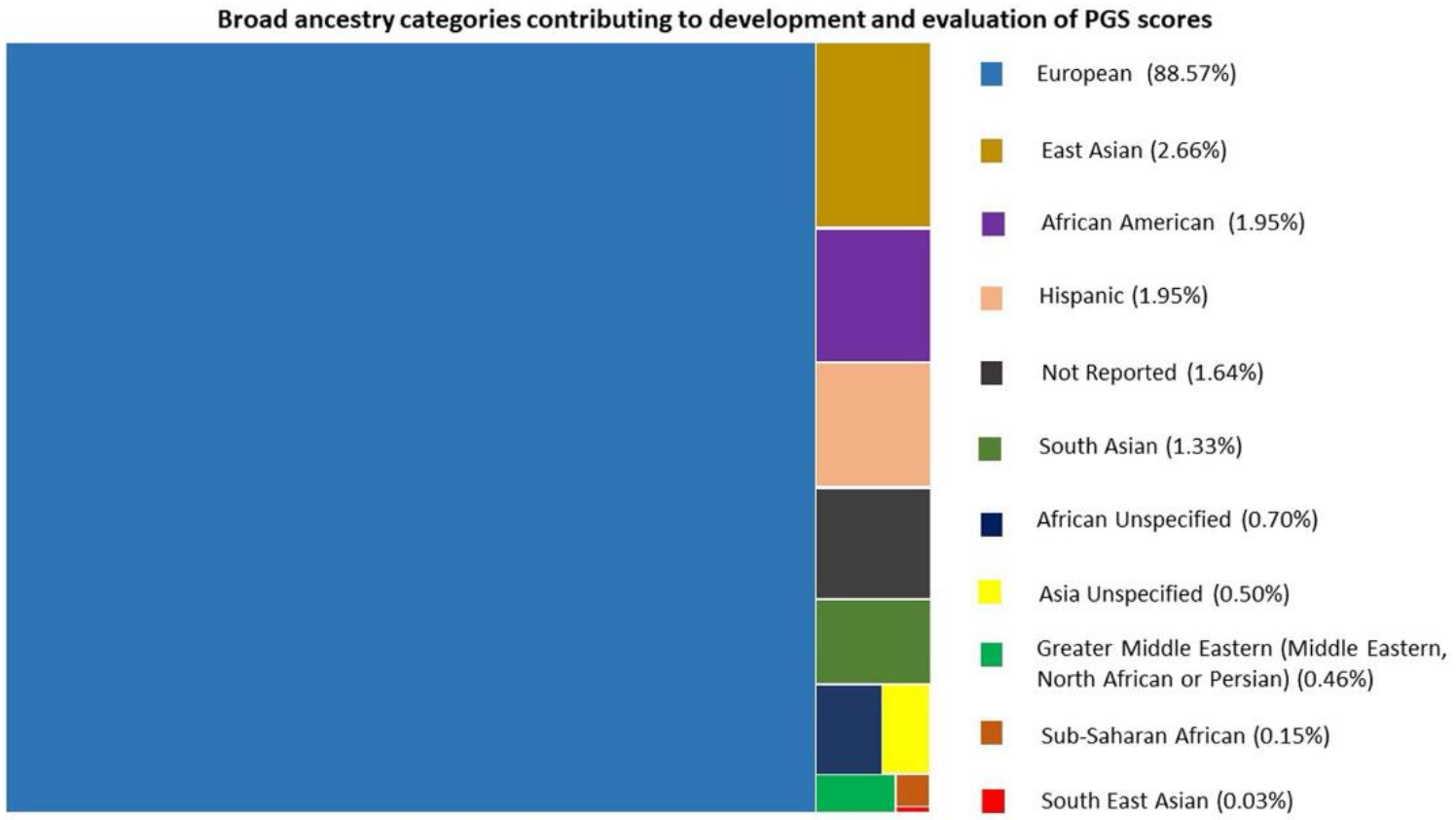
PGS ARE BASED ON COMMON SNPs



Mars et al 2020, Nat Com, 11:6383

CHALLENGE 4

POPULATION SPECIFIC



Prediction of T2D

Population	Liability R^2	Covariates-adjusted AUC
European	9.2%	0.66
African	2.8%	0.58
Hispanic	8.0%	0.63

Fatumo et al 2023, Genome Medicine
Ge et al 2022, Genome Medicine

SESSION 2

- The first polygenic score
- What you need is…
- Commonly used scoring algorithms
- Workflow
- Current challenges with PGS?

AALBORG UNIVERSITY


BREAK

AGENDA

08:00 – 08:30	Welcome and common introductions
08:30 – 09:10	Session 1: Introduction to Polygenic Scores (PGS)
09:10 – 09:20	Break
09:20 – 10:00	Session 2: Data Sources and Computational Methods
10:00 – 10:10	Break
10:10 – 10:40	Session 3: Evaluating and Interpreting Polygenic Scores
10:40 – 11:00	Break
11:00 – 11:45	Session 4: Advanced Applications and Future Directions
11:45 – 12:30	Lunch and short walk
12:30 – 15:30	Identification of 2-3 projects of common interest
15:30 – 16:00	Next steps and thank you for today

SESSION 3

- How to measure ‘accuracy’?
- Interpretability and risk communication
- Lack of transferability
- Applications…?

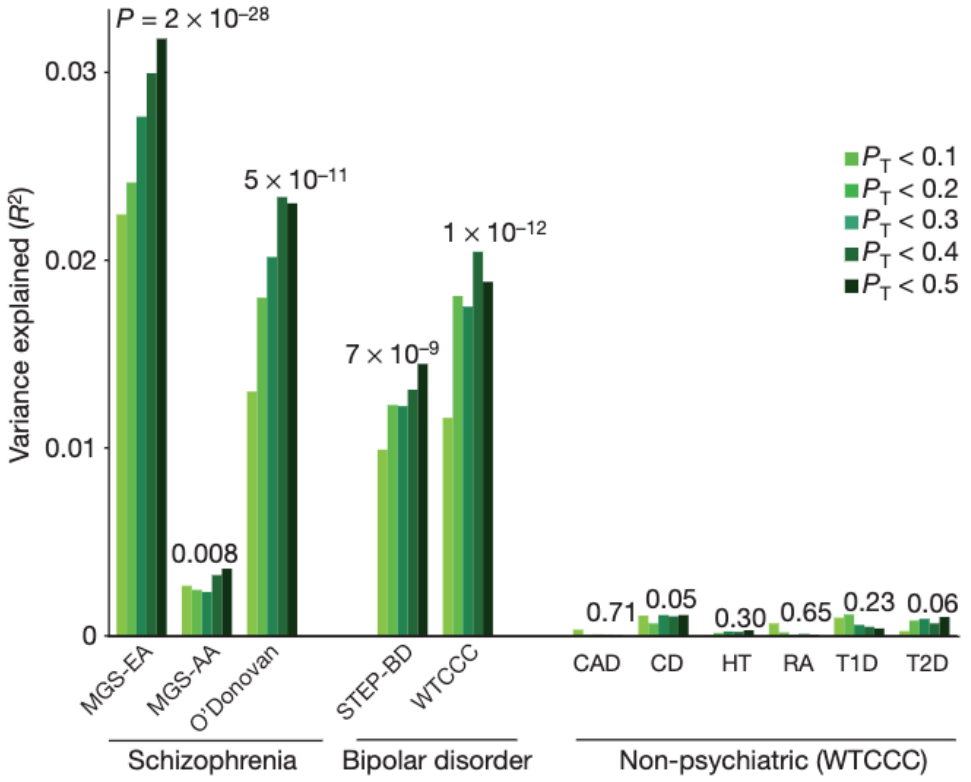
EVALUATE POLYGENIC PROFILES

$$y = Xb + Zc + e$$

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

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

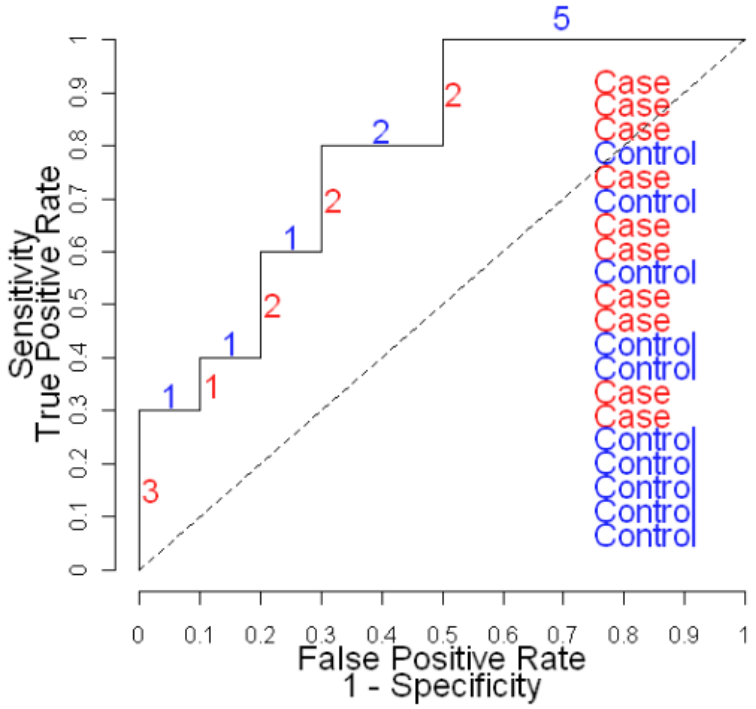
Variance explained (R^2) for quantitative traits, and Nagelkerke's R^2 for binary traits (**however, NagR2 is biased with disease prevalence!**)



The International Schizophrenia Consortium (2009) Nature, 460

EVALUATE POLYGENIC PROFILES

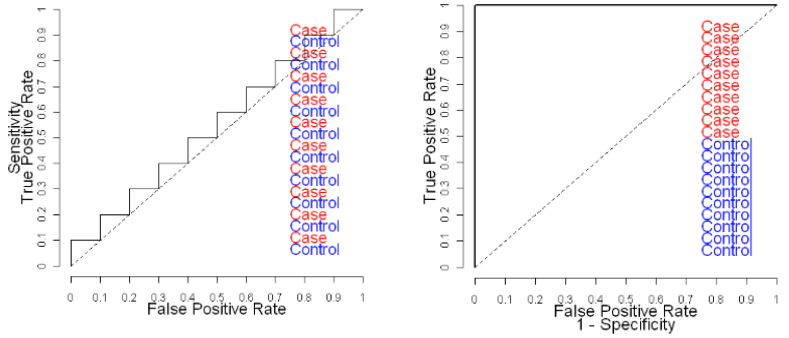
Rank individuals on score from highest ranked to lowest



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

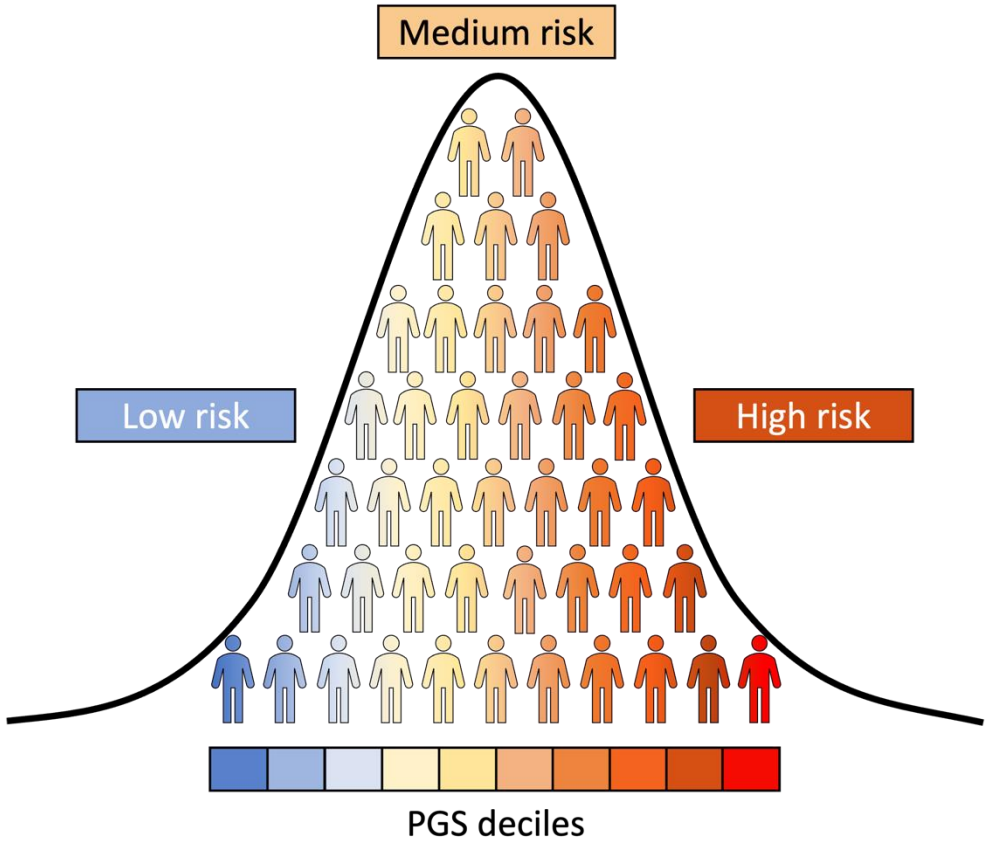
Odds ratio (OR)

Cut the distribution into deciles

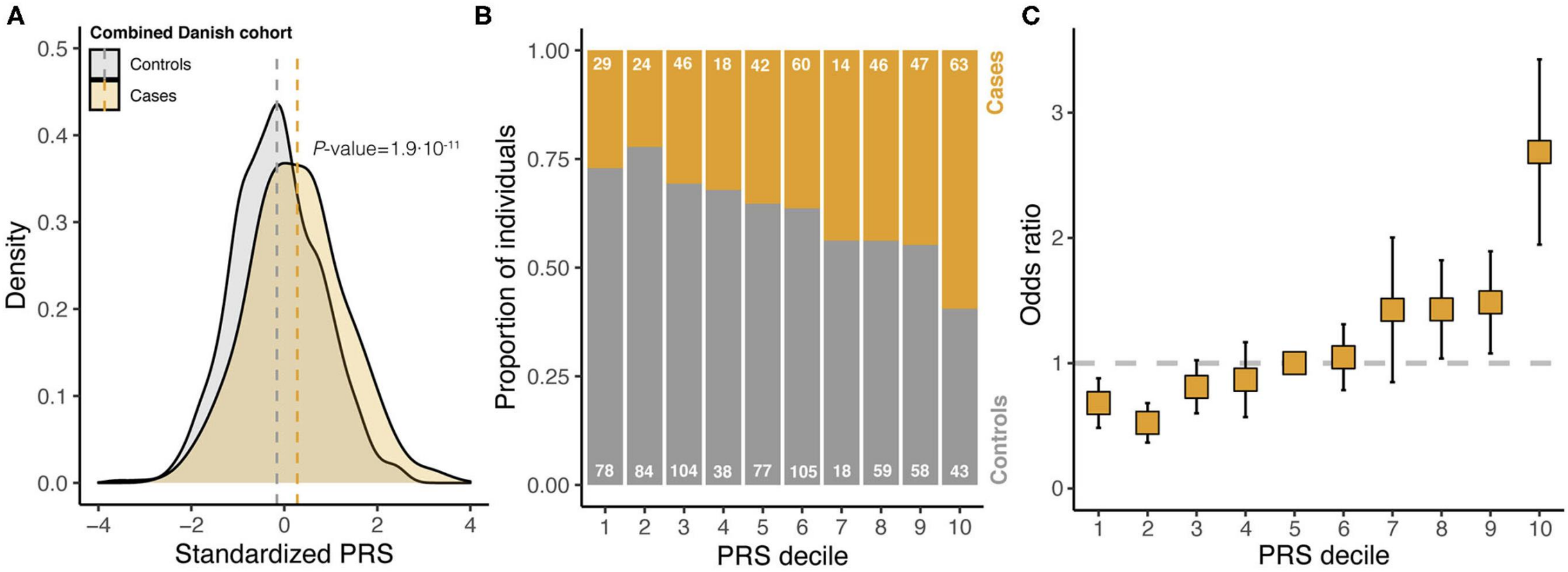
Each decile will include both cases and controls

Odds of being a case in each decile

Odds ratio for each decile compared to the first decile



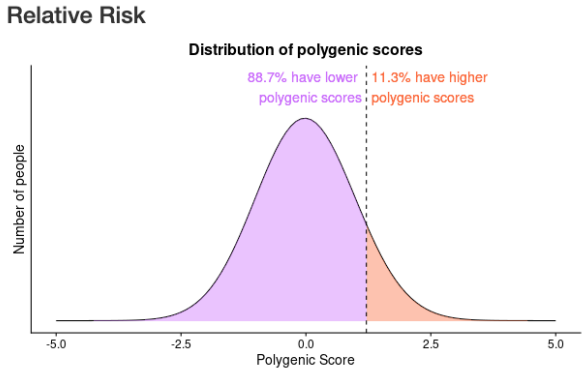
A 14-SNP PGS

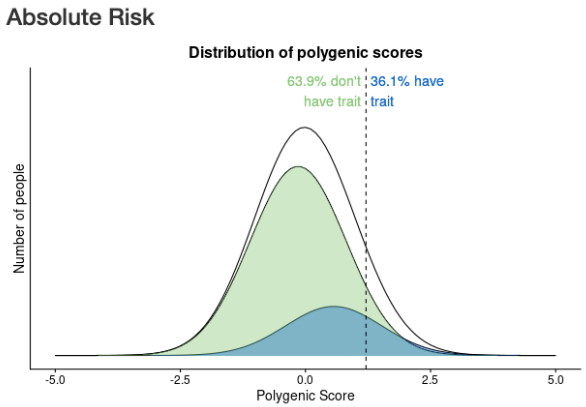


Kloeve-mogensen, Rohde, et al 2021, Front Repro Health

INTERPRETABILITY & RISK COMMUNICATION

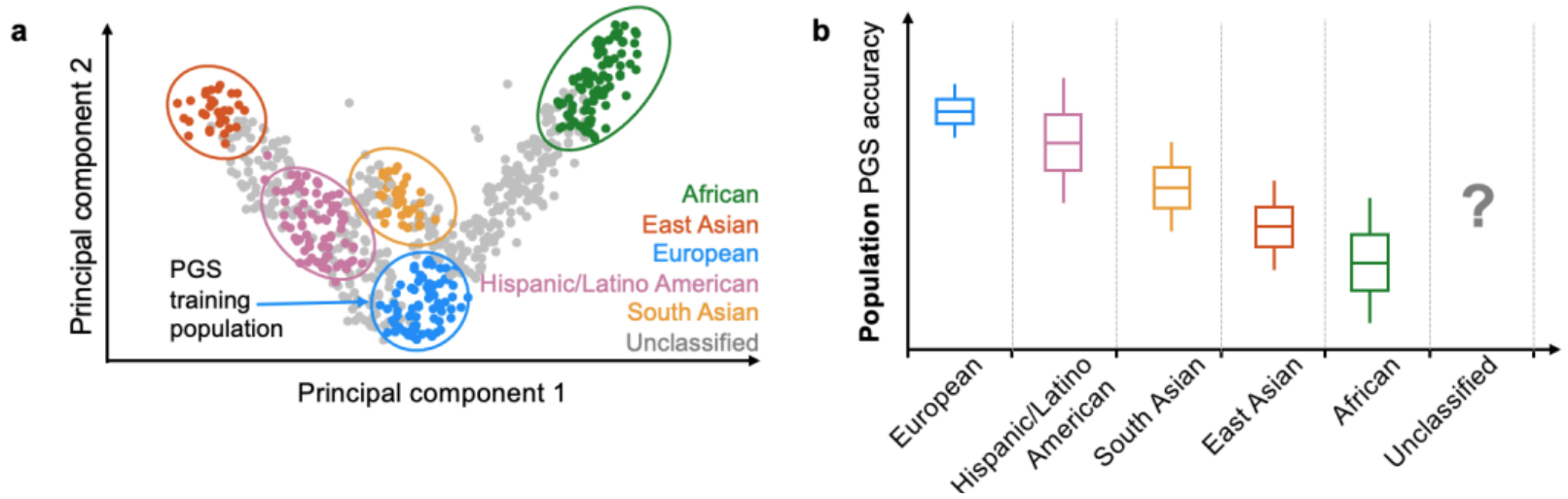
- The risk associated with the PGS is a relative risk
  - People have suggested methods to convert relative PGS risk to absolute risks (https://opain.github.io/GenoPred/PRS_to_Abs_tool.html)
- The relative risk may sound high, but the absolute risk is low
- Hard to use meaningfully in clinical decisions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

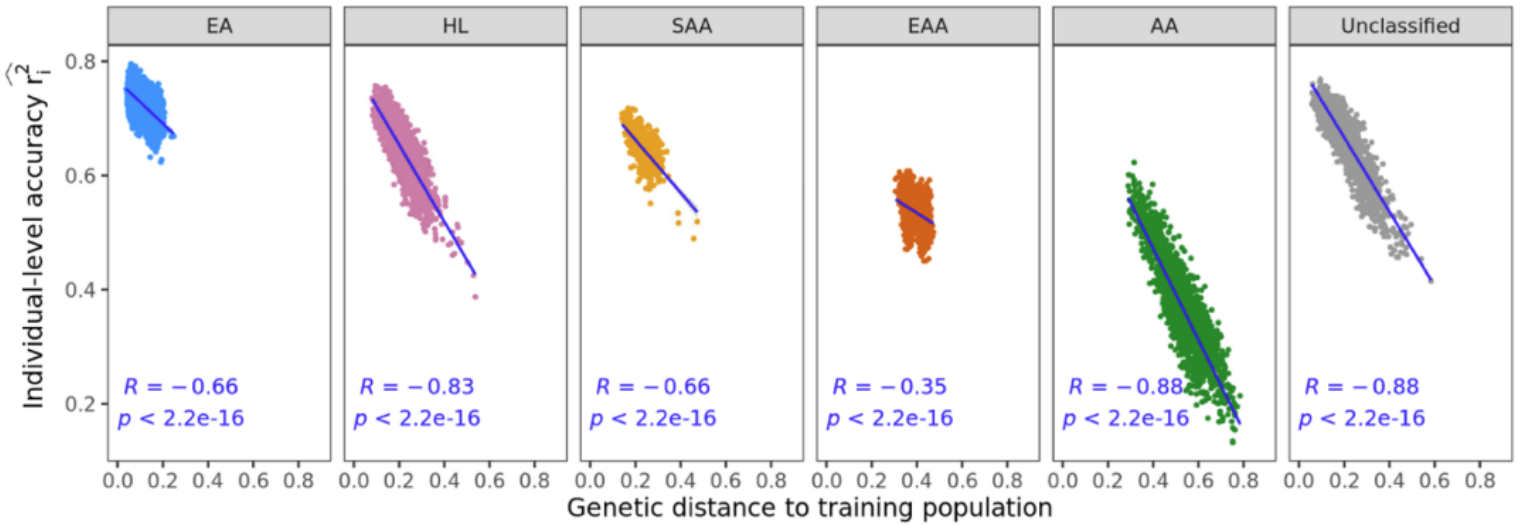


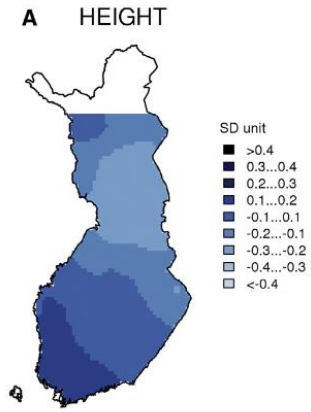


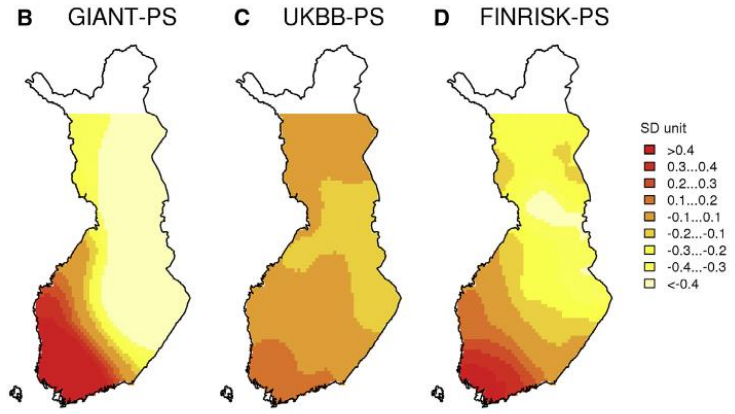
Pain et al 2022, EJHG, 3:339-348

LACK OF TRANSFERABILITY









Ding et al 2022, Nature, 618

APPLICATIOS?

PGS FOR CAD

Table 3 | Prevalence and clinical impact of a high GPS

High GPS definition	Reference group	Odds ratio	95% CI	*P* value
CAD				
Top 20% of distribution	Remaining 80%	2.55	2.43–2.67	$<1\times10^{-300}$
Top 10% of distribution	Remaining 90%	2.89	2.74–3.05	$<1\times10^{-300}$
Top 5% of distribution	Remaining 95%	3.34	3.12–3.58	6.5×10^{-264}
Top 1% of distribution	Remaining 99%	4.83	4.25–5.46	1.0×10^{-132}
Top 0.5% of distribution	Remaining 99.5%	5.17	4.34–6.12	7.9×10^{-78}

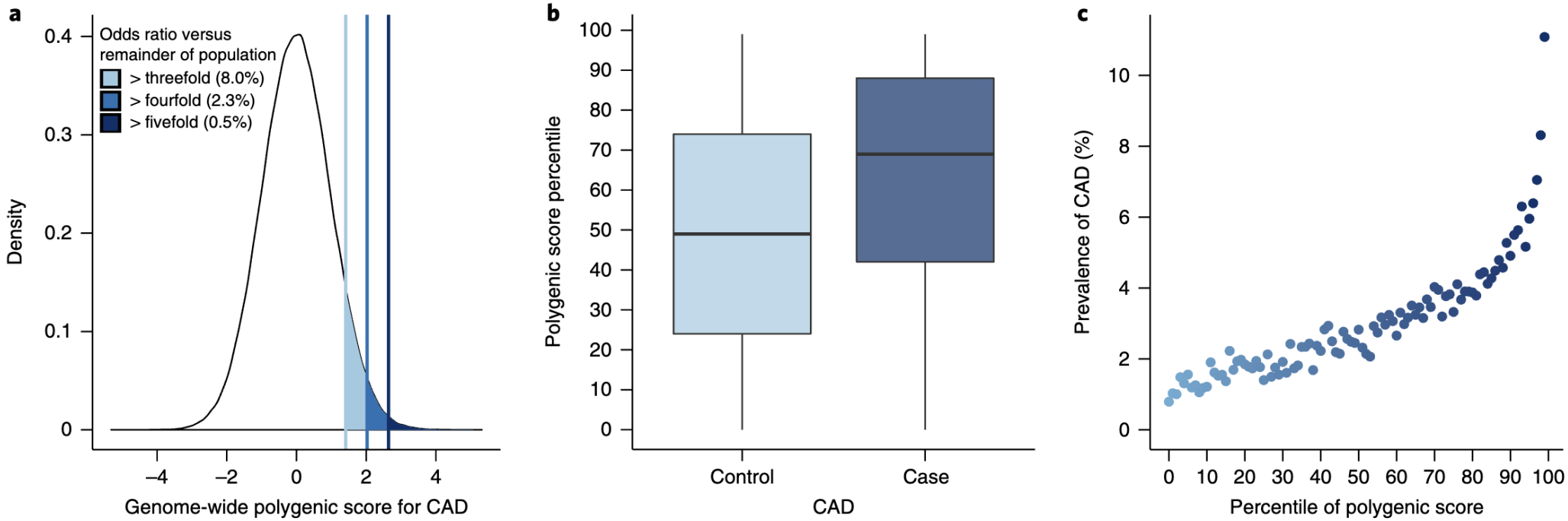


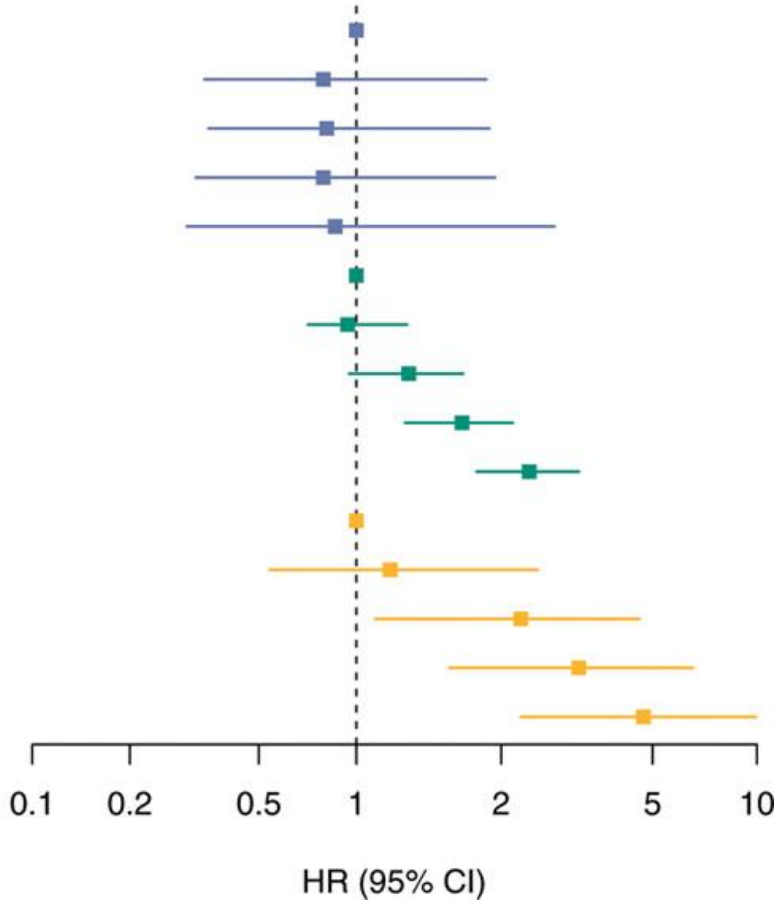
Table 2 | Proportion of the population at three-, four- and fivefold increased risk for each of the five common diseases

High GPS definition	Individuals in testing dataset (*n*)	% of individuals
Odds ratio ≥3.0		
CAD	23,119/288,978	8.0

Khera et al (2018), Nat Gen, 50:1219–1224

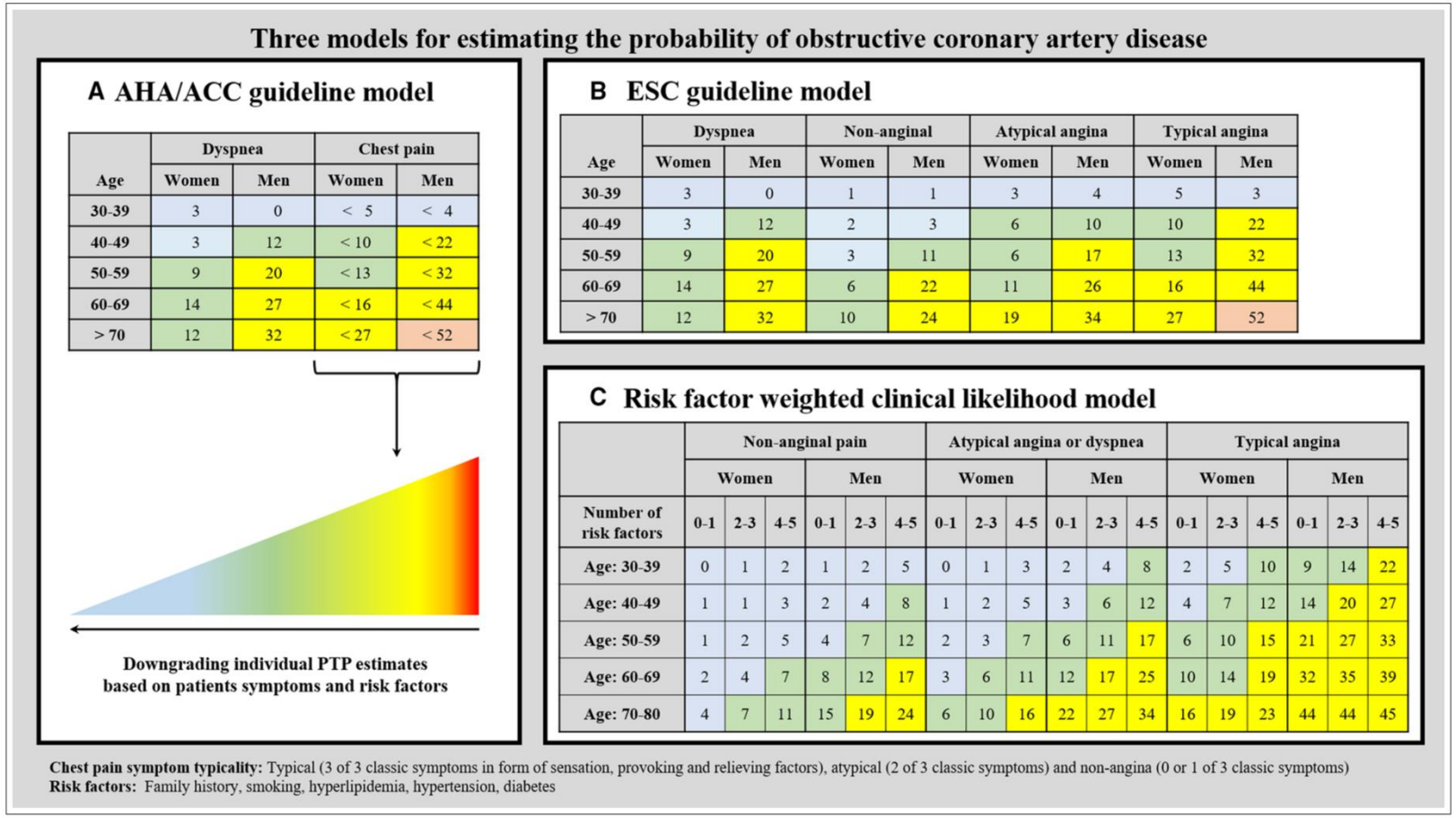
STRATIFIED USE OF PGS

Group	LDL–C (mg/dL)	HR (95%CI)	*P* value	Total Inds / Cases
Low PRS	<100	1	[Reference]	1142 / 7
	100 – <130	0.79 (0.34–1.84)	0.58	4194 / 26
	130 – <160	0.81 (0.35–1.87)	0.62	4570 / 38
	160 – <190	0.79 (0.32–1.93)	0.6	2264 / 21
	≥ 190	0.86 (0.3–2.71)	0.86	701 / 7
Int PRS	<100	1	[Reference]	6824 / 61
	100 – <130	0.94 (0.71–1.24)	0.66	28352 / 283
	130 – <160	1.25 (0.95–1.63)	0.11	38352 / 615
	160 – <190	1.62 (1.23–2.13)	<0.001	21576 / 483
	≥ 190	2.34 (1.75–3.14)	<0.001	7867 / 246
High PRS	<100	1	[Reference]	624 / 8
	100 – <130	1.15 (0.54–2.46)	0.72	3082 / 44
	130 – <160	2.23 (1.08–4.59)	0.03	4808 / 151
	160 – <190	3.14 (1.52–6.5)	0.002	3091 / 144
	≥ 190	4.71 (2.23–9.94)	<0.001	1267 / 81



Bolli et al (2021), Circulation, 143:1452–1454

RISK ASSESSMENT FOR OBSTRUCTIVE CAD



Three models for estimating the probability of obstructive coronary artery disease

A AHA/ACC guideline model

Age	Dyspnea		Chest pain	
	Women	Men	Women	Men
30-39	3	0	< 5	< 4
40-49	3	12	< 10	< 22
50-59	9	20	< 13	< 32
60-69	14	27	< 16	< 44
> 70	12	32	< 27	< 52

Downgrading individual PTP estimates based on patients symptoms and risk factors

B ESC guideline model

Age	Dyspnea		Non-anginal		Atypical angina		Typical angina	
	Women	Men	Women	Men	Women	Men	Women	Men
30-39	3	0	1	1	3	4	5	3
40-49	3	12	2	3	6	10	10	22
50-59	9	20	3	11	6	17	13	32
60-69	14	27	6	22	11	26	16	44
> 70	12	32	10	24	19	34	27	52

C Risk factor weighted clinical likelihood model

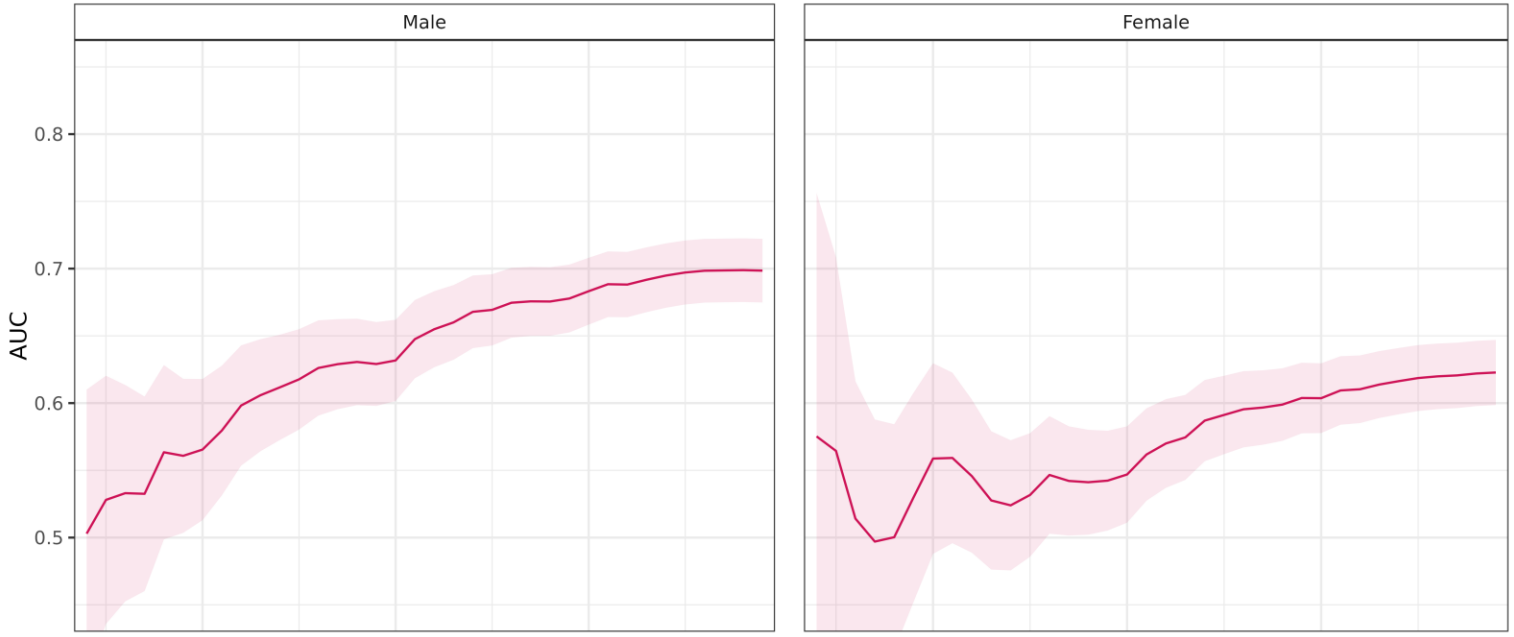
	Non-anginal pain						Atypical angina or dyspnea						Typical angina					
	Women			Men			Women			Men			Women			Men		
Number of risk factors	0-1	2-3	4-5	0-1	2-3	4-5	0-1	2-3	4-5	0-1	2-3	4-5	0-1	2-3	4-5	0-1	2-3	4-5
Age: 30-39	0	1	2	1	2	5	0	1	3	2	4	8	2	5	10	9	14	22
Age: 40-49	1	1	3	2	4	8	1	2	5	3	6	12	4	7	12	14	20	27
Age: 50-59	1	2	5	4	7	12	2	3	7	6	11	17	6	10	15	21	27	33
Age: 60-69	2	4	7	8	12	17	3	6	11	12	17	25	10	14	19	32	35	39
Age: 70-80	4	7	11	15	19	24	6	10	16	22	27	34	16	19	23	44	44	45

Chest pain symptom typicality: Typical (3 of 3 classic symptoms in form of sensation, provoking and relieving factors), atypical (2 of 3 classic symptoms) and non-angina (0 or 1 of 3 classic symptoms)
Risk factors: Family history, smoking, hyperlipidemia, hypertension, diabetes

0-5: Low risk
5-15: Intermediate risk
>15 high risk

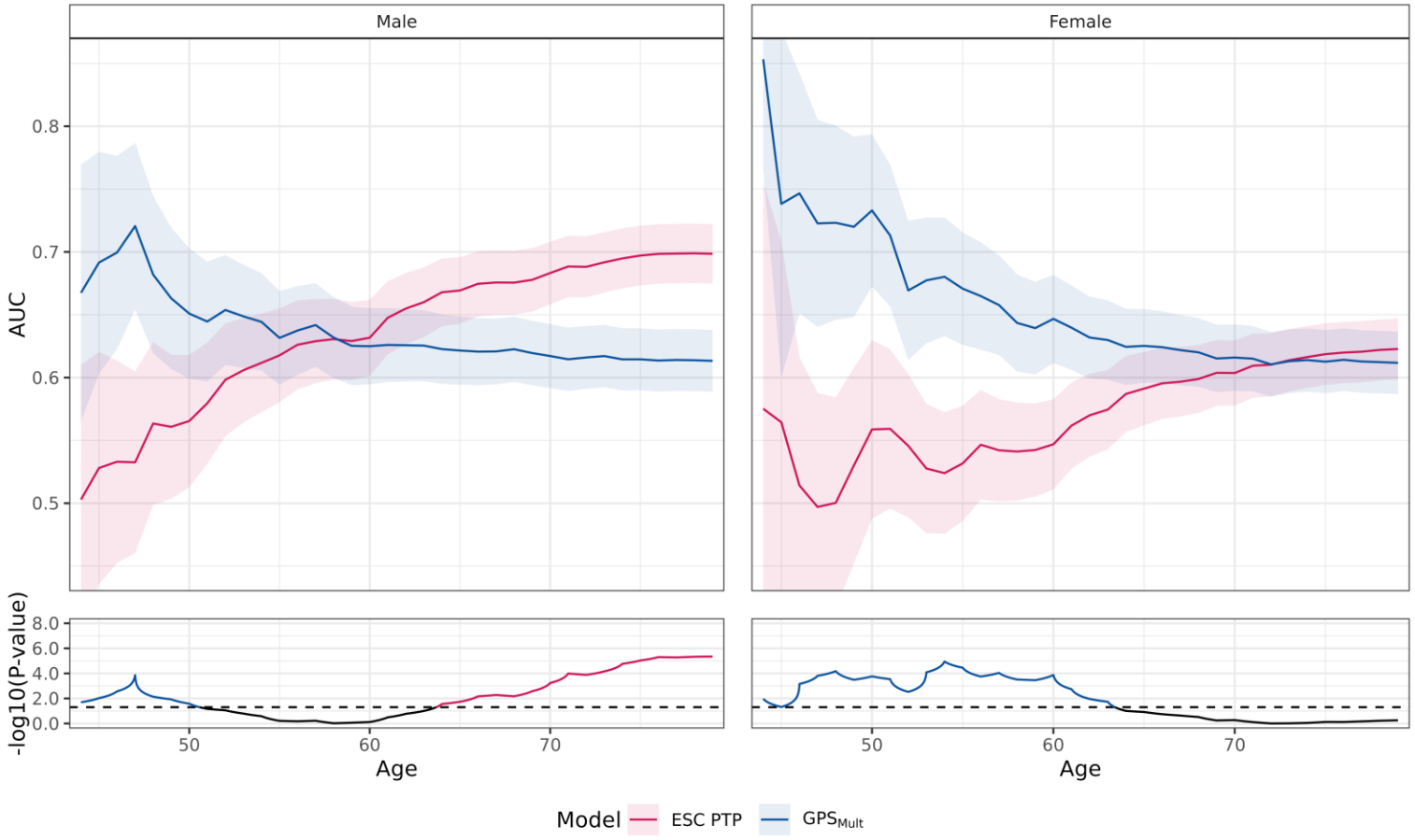
Winther et al. 2022, J Am Heart Ass, 11:24

STRATIFIED USE OF PGS



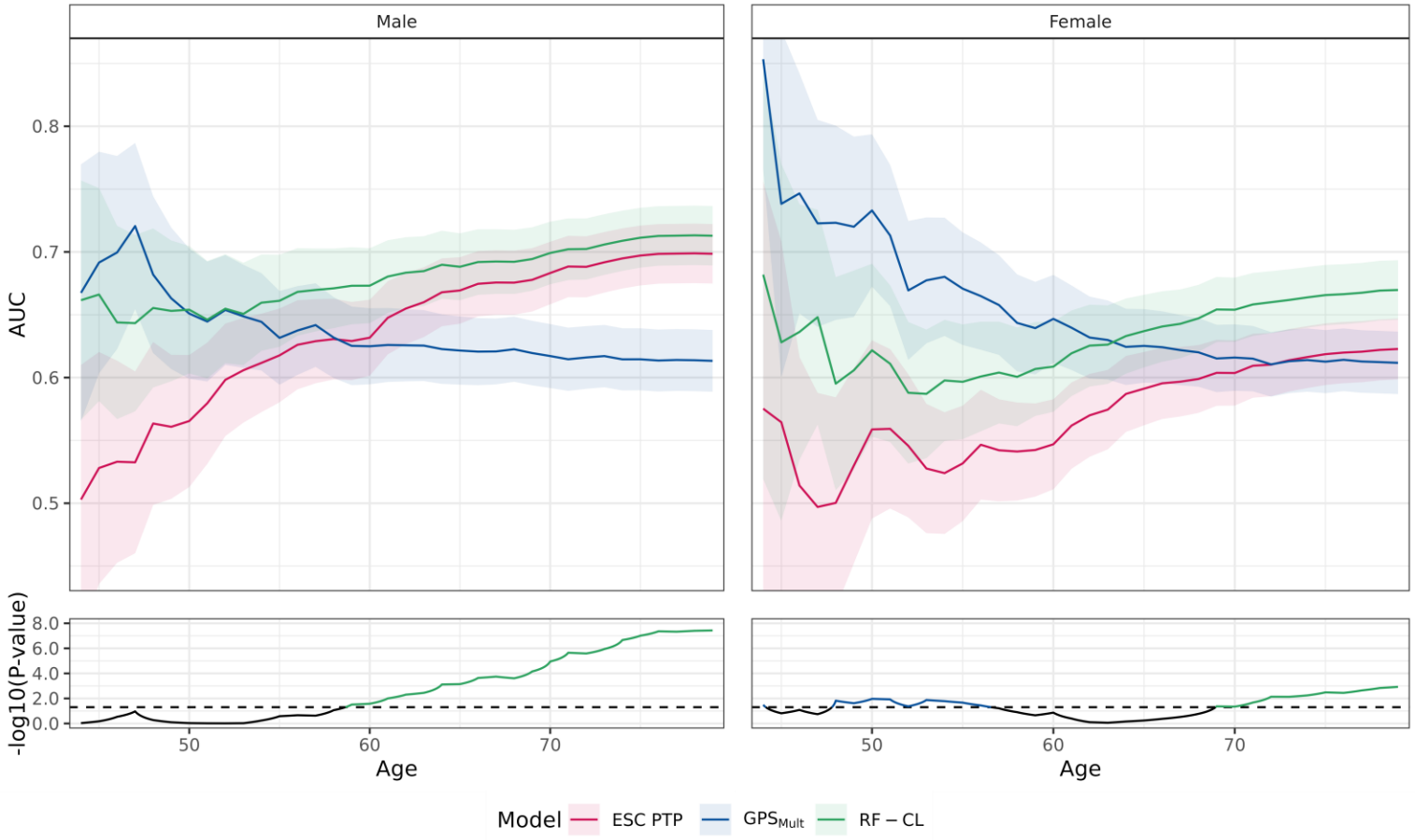
Møller, et al., Rohde, in prep

STRATIFIED USE OF PGS



Møller, et al., Rohde, in prep

STRATIFIED USE OF PGS



Møller, et al., Rohde, in prep

STRATIFIED USE OF PGS

ESC PTP

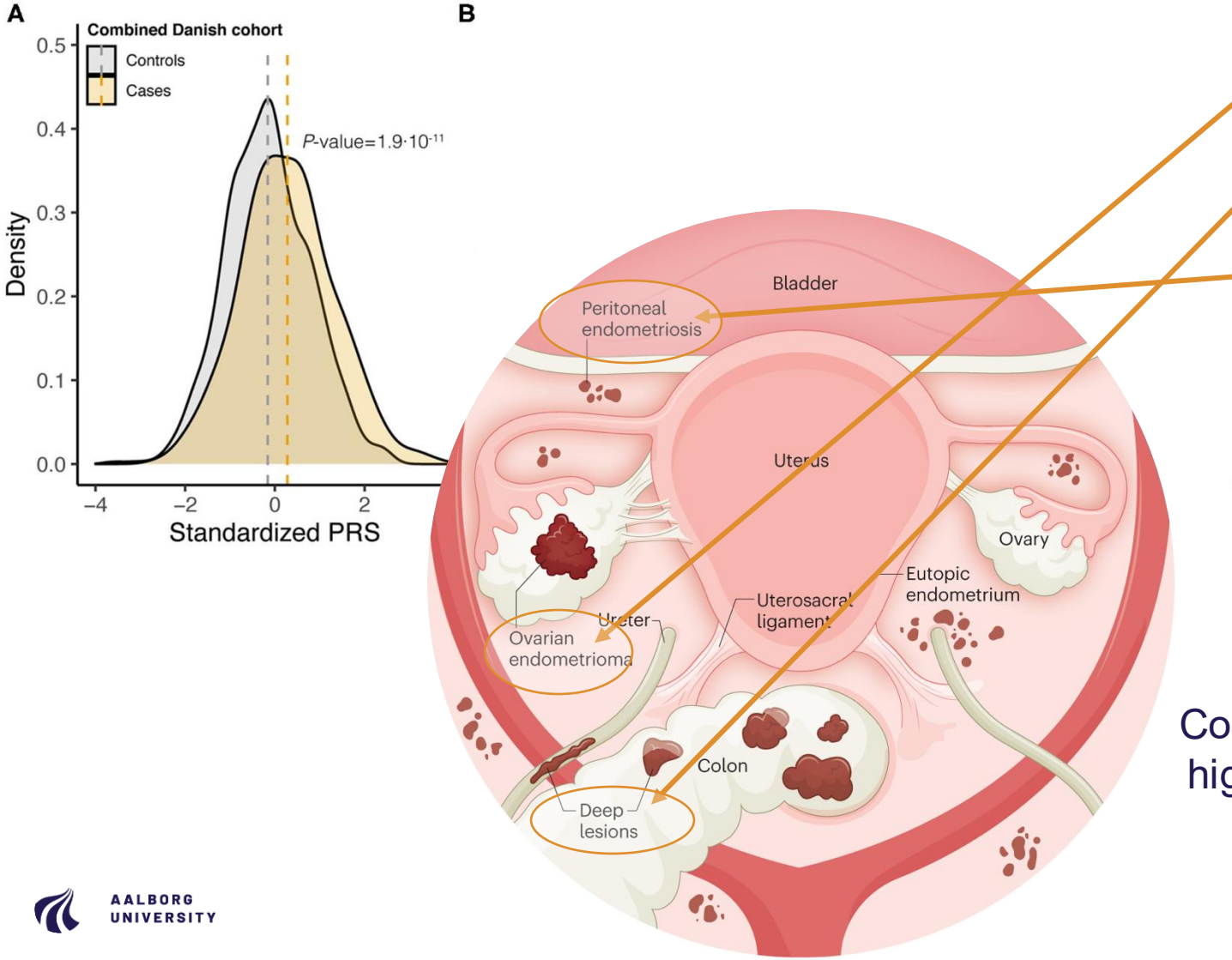
Group	Sample size	Cases	Referrals	Correct referrals	Cases referred	Cases missed
Women < 55	282 (13%)	27 (10%)	0 (0%)	0 (0%)	0%	27 (15%)
Women ≥ 55	689 (33%)	146 (21%)	156 (23%)	40 (26%)	27%	106 (59%)
Men < 55	395 (19%)	73 (18%)	189 (48%)	42 (22%)	58%	31 (17%)
Men ≥ 55	733 (35%)	298 (41%)	661 (90%)	283 (43%)	95%	15 (8%)
Total	2099 (100%)	544 (26%)	1006 (48%)	365 (36%)	67%	179 (100%)

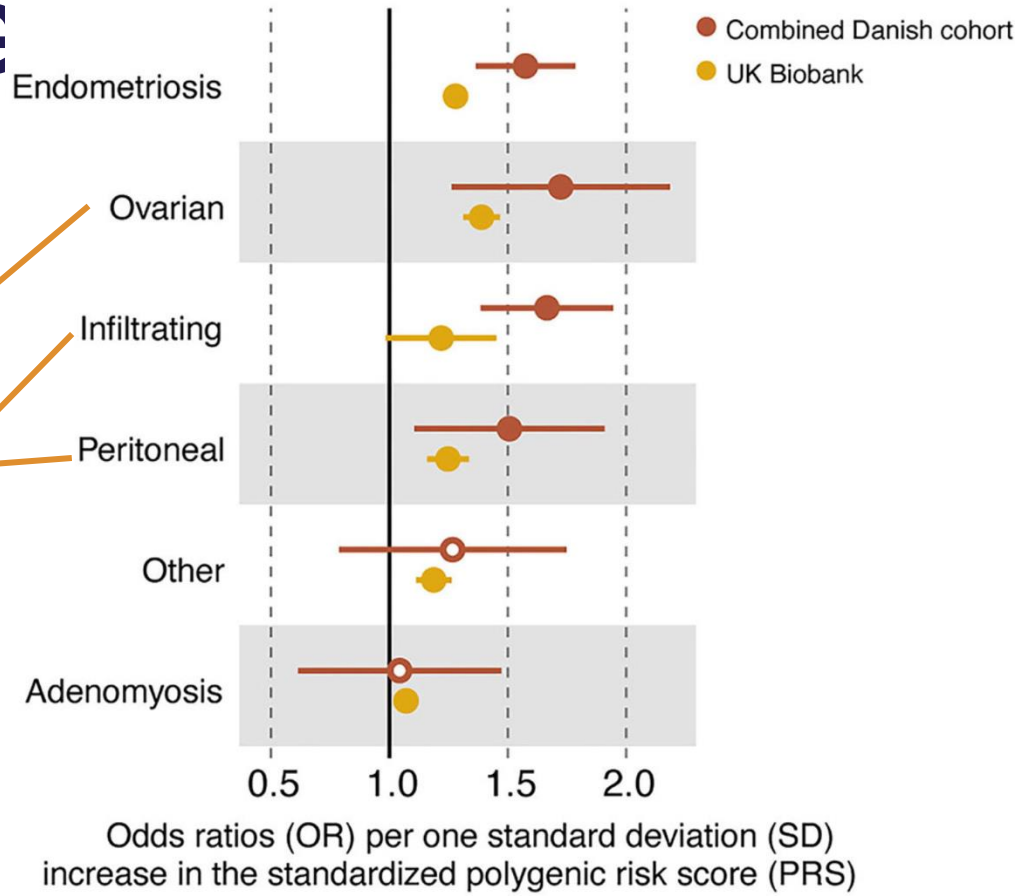
ESC PTP + PGS

Group	Sample size	Cases	Referrals	Correct referrals	Cases referred	Cases missed
Women < 55	282 (13%)	27 (10%)	10 (4%)	2 (20%)	7%	25 (15%)
Women ≥ 55	689 (33%)	146 (21%)	153 (22%)	48 (31%)	33%	98 (58%)
Men < 55	395 (19%)	73 (18%)	179 (45%)	47 (26%)	64%	26 (15%)
Men ≥ 55	733 (35%)	298 (41%)	620 (85%)	279 (45%)	94%	19 (11%)
Total	2099 (100%)	544 (26%)	962 (46%)	376 (39%)	69%	168 (100%)

Møller, et al., Rohde, in prep

LOOK INTO THE BIOLOG





Could indicate that ovarian endometriosis has a higher genetic burden than the other subtypes.

Kloeve-mogensen, Rohde, et al 2021, Front Repro Health

SESSION 3

- How to measure ‘accuracy’?
- Interpretability and risk communication
- Lack of transferability
- Applications…?


BREAK

AGENDA

08:00 – 08:30	Welcome and common introductions
08:30 – 09:10	Session 1: Introduction to Polygenic Scores (PGS)
09:10 – 09:20	Break
09:20 – 10:00	Session 2: Data Sources and Computational Methods
10:00 – 10:10	Break
10:10 – 10:40	Session 3: Evaluating and Interpreting Polygenic Scores
10:40 – 11:00	Break
11:00 – 11:45	Session 4: Advanced Applications and Future Directions
11:45 – 12:30	Lunch and short walk
12:30 – 15:30	Identification of 2-3 projects of common interest
15:30 – 16:00	Next steps and thank you for today

SESSION 4

- Many challenges – how to circumvent these?
- Enhancing biological understanding?
- PGS in combination with proteomics

CHALLENGES WITH PGS

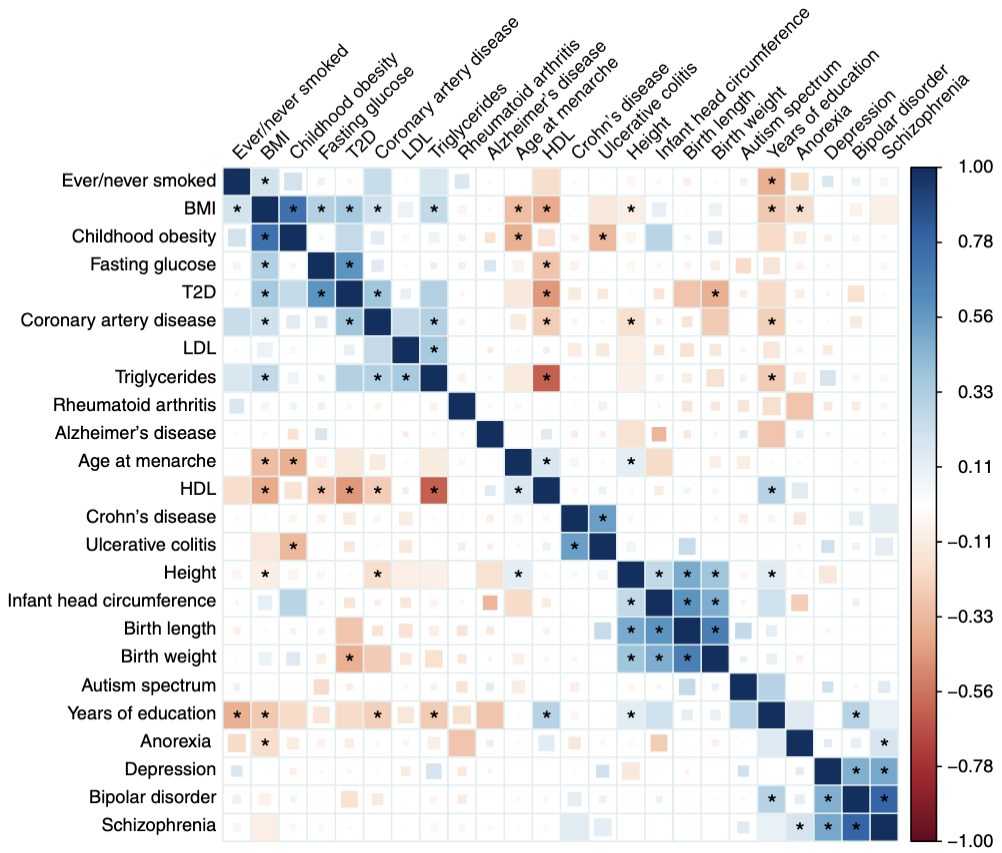
1) Too many individuals have ‘average’ PGS
2) Only capture additive effects
3) Only common variants
4) Transferability
5) Understanding biology

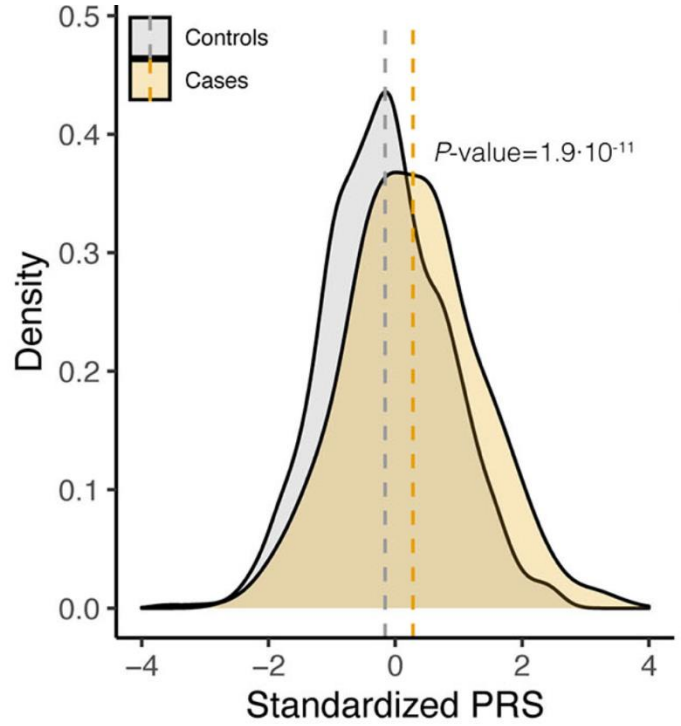
Improving PGS and Future directios

→ Improve discriminative ability using multi-trait models and functional annotation
→ Ancestry-aware models
→ Integration with single cell and spatial transcriptomics
→ Integrative genomics (combining PGS with other ‘omics)

IMPROVING PGS

BETTER DISCRIMINATIVE ABILITY





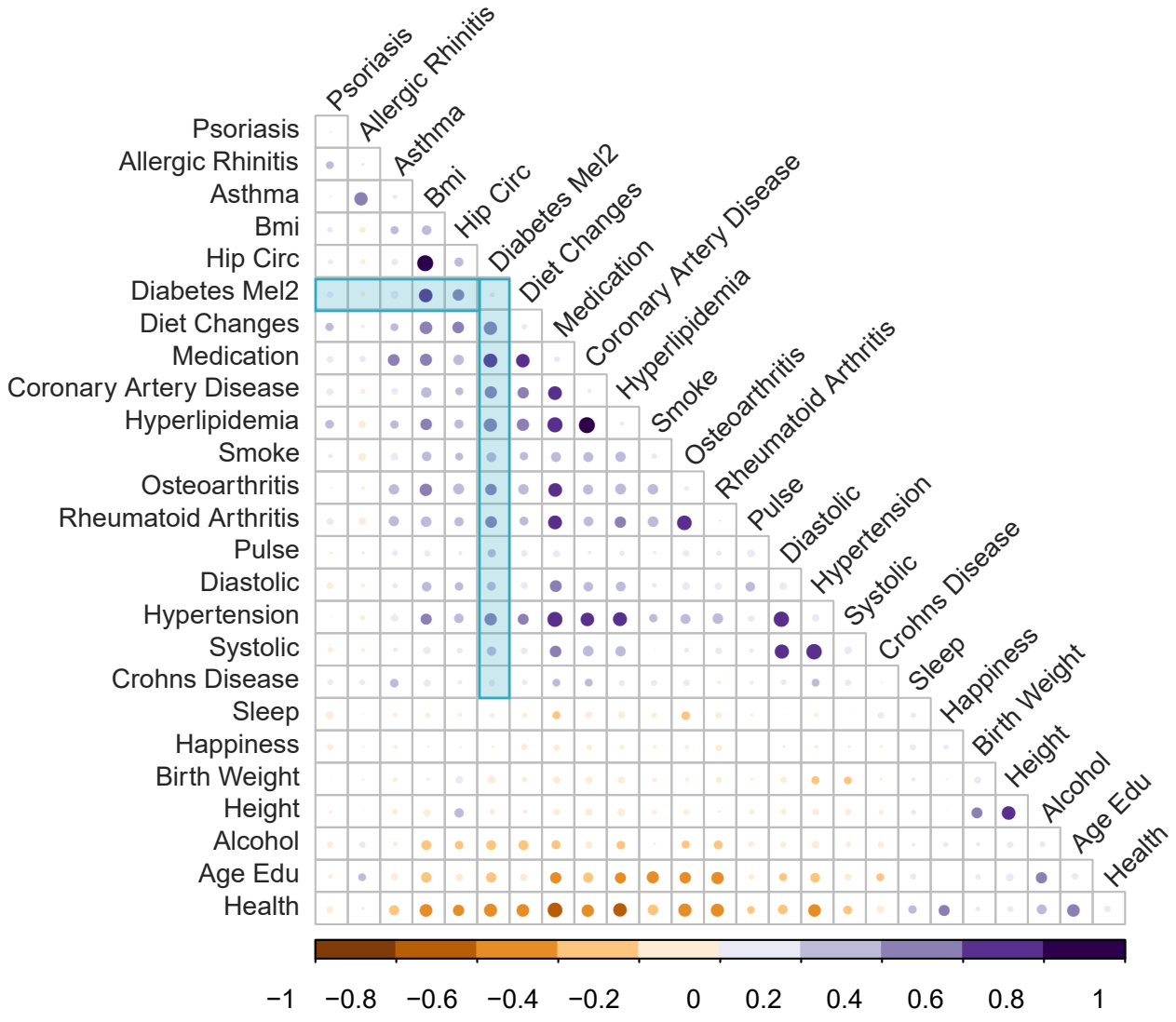
Identifying genetic correlations between complex traits and diseases can provide useful etiological insights and help prioritize likely causal relationships

Kloeve-Mogensen, Rohde et al. (2023). *Front Rep Med*, *3*, 793226.
Bulik-Sullivan et al. (2015). *Nat Gen*, *11*(47), 1236-1241.

MULTI-TRAIT PGS
FOR TYPE 2 DIABETES

T2D is strongly correlated with a range of complex diseases and traits, such as overweight, cardiovascular diseases, hypertension, chronic kidney disease etc.

Can we improve prediction accuracy by leveraging shared genetic signatures?



Rohde et al. (2021) Frontiers in Medicine, 8, 711208
Rohde et al. unpublished

MULTI-TRAIT PGS
FOR TYPE 2 DIABETES

$$\text{PGS} = \sum X_i \hat{\beta}_i$$

$$\hat{\beta}_i$$

Each marker effect ($\hat{\beta}$) is weighted by w, which is found by using selection index theory

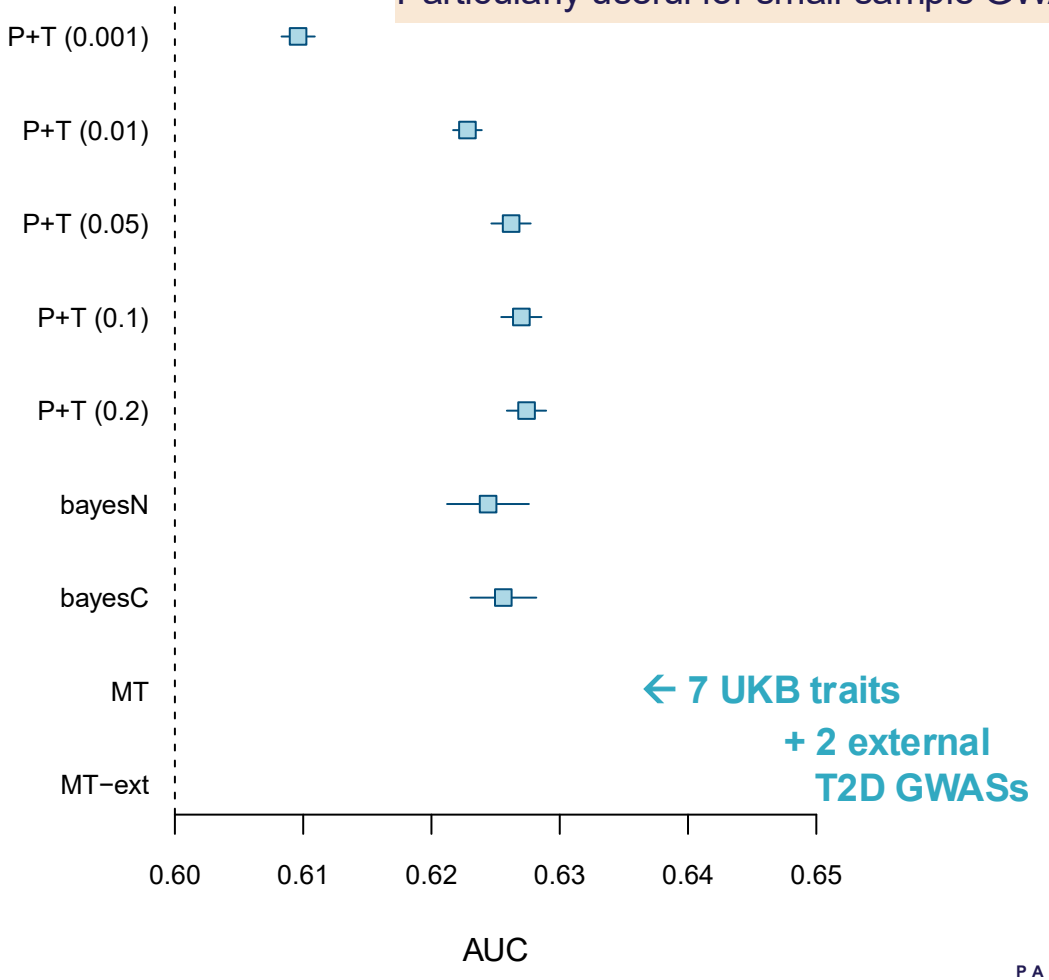
$$w = \begin{bmatrix} \frac{h_1^2}{M} + \frac{1}{N_1} & \cdots & \frac{r_g h_1 h_k}{M} \\ \vdots & \ddots & \vdots \\ \frac{r_g h_k h_1}{M} & \cdots & \frac{h_k^2}{M} + \frac{1}{N_k} \end{bmatrix}^{-1} \begin{bmatrix} \frac{h_1^2}{M} \\ \vdots \\ \frac{r_g h_1 h_k}{M} \end{bmatrix}$$

$$\text{MT}-\text{PGS} = \sum_{i=1}^{m} X_i \hat{\beta}_{wMTi}$$

The PGS accuracy depends on the power of the GWAS (=sample size).

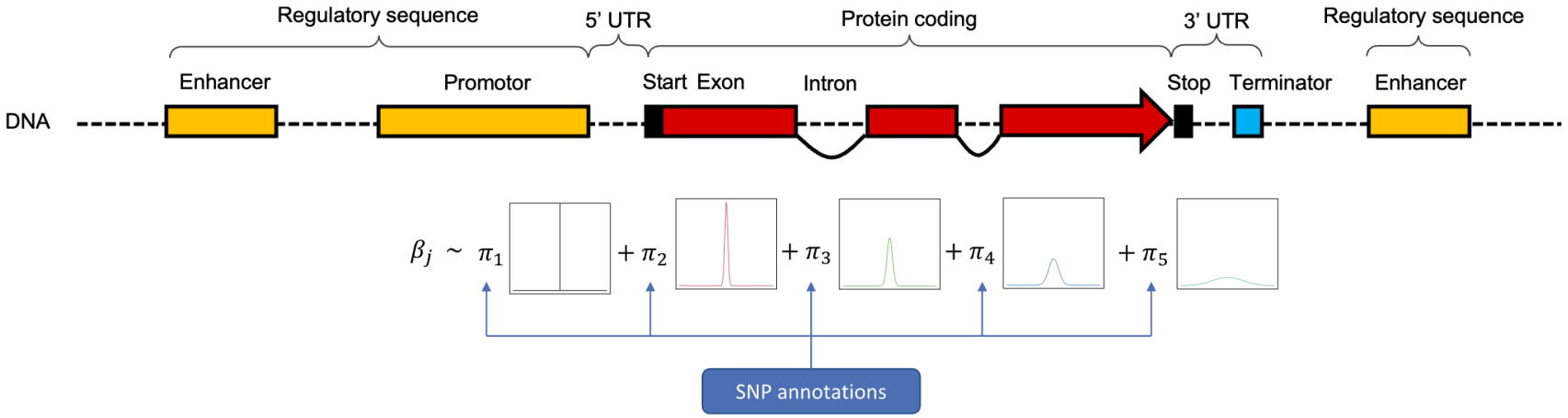
Utilising the genetic correlation among phenotypes is equivalent to increasing GWAS sample size.

Particularly useful for small-sample GWAS.

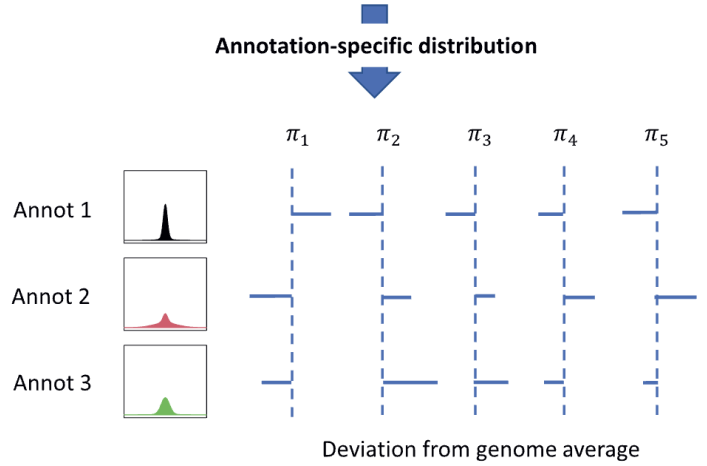


Rohde et al. (2021) Frontiers in Medicine
Rohde et al. (2023). *Bioinformatics, 11*(39), btad656.

ENHANCING PGS BY ANNOTATION

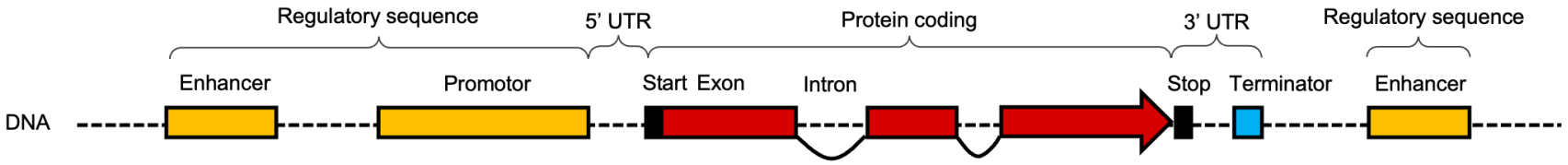


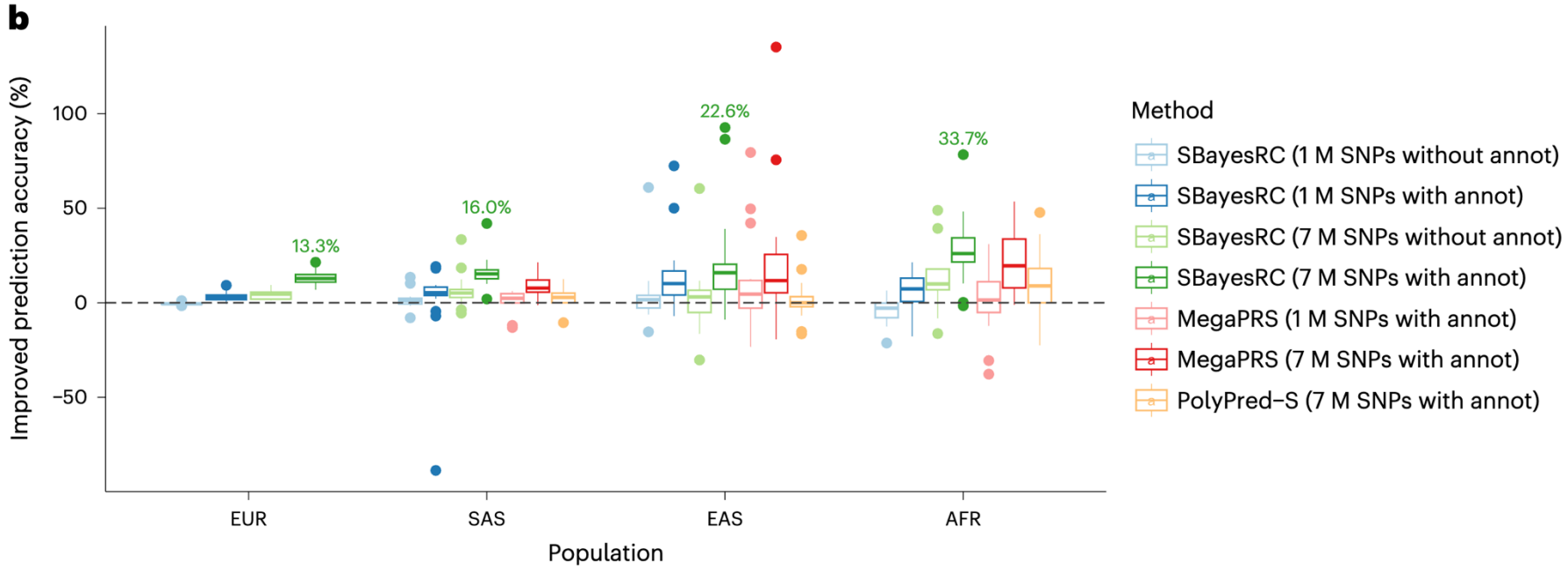
$$f(\pi_{jk}) = \text{Intercept} + \sum \text{SNP annotation} \times \text{annotation effect}$$



Zheng et al. (2024). *Nat Gen*, *56*:767-777

ENHANCING PGS BY ANNOTATION





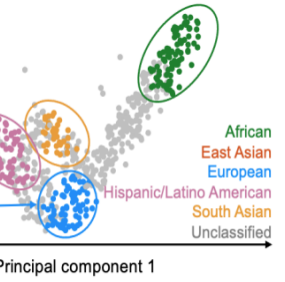
Zheng et al. (2024). *Nat Gen, 56*:767-777

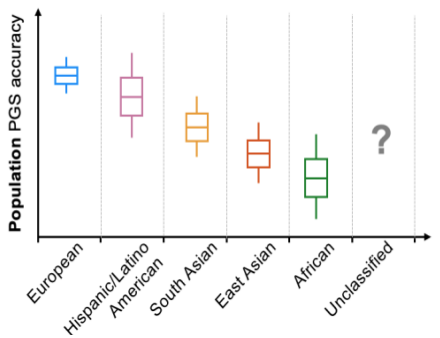
CROSS-ANCESTRY PGS

$$b = \left(X'X + I\frac{\sigma_e^2}{\sigma_b^2} \right)^{-1} X'y$$

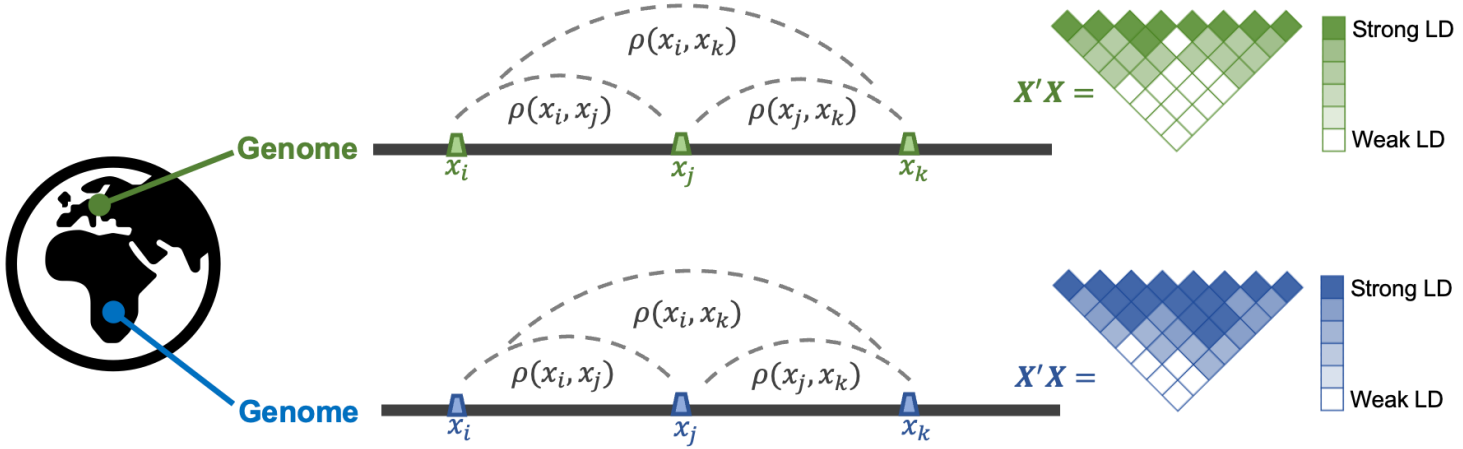
$$X'X = D^{0.5} B D^{0.5}$$

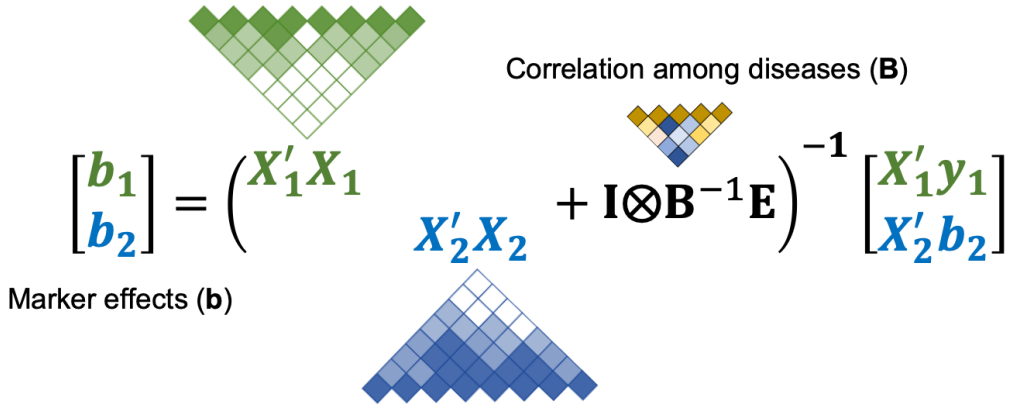
$$X'y = Db_m$$





Computation of ancestry-specific LD





MT-BLR

$$PGS_1 = \sum X_1 b_1$$

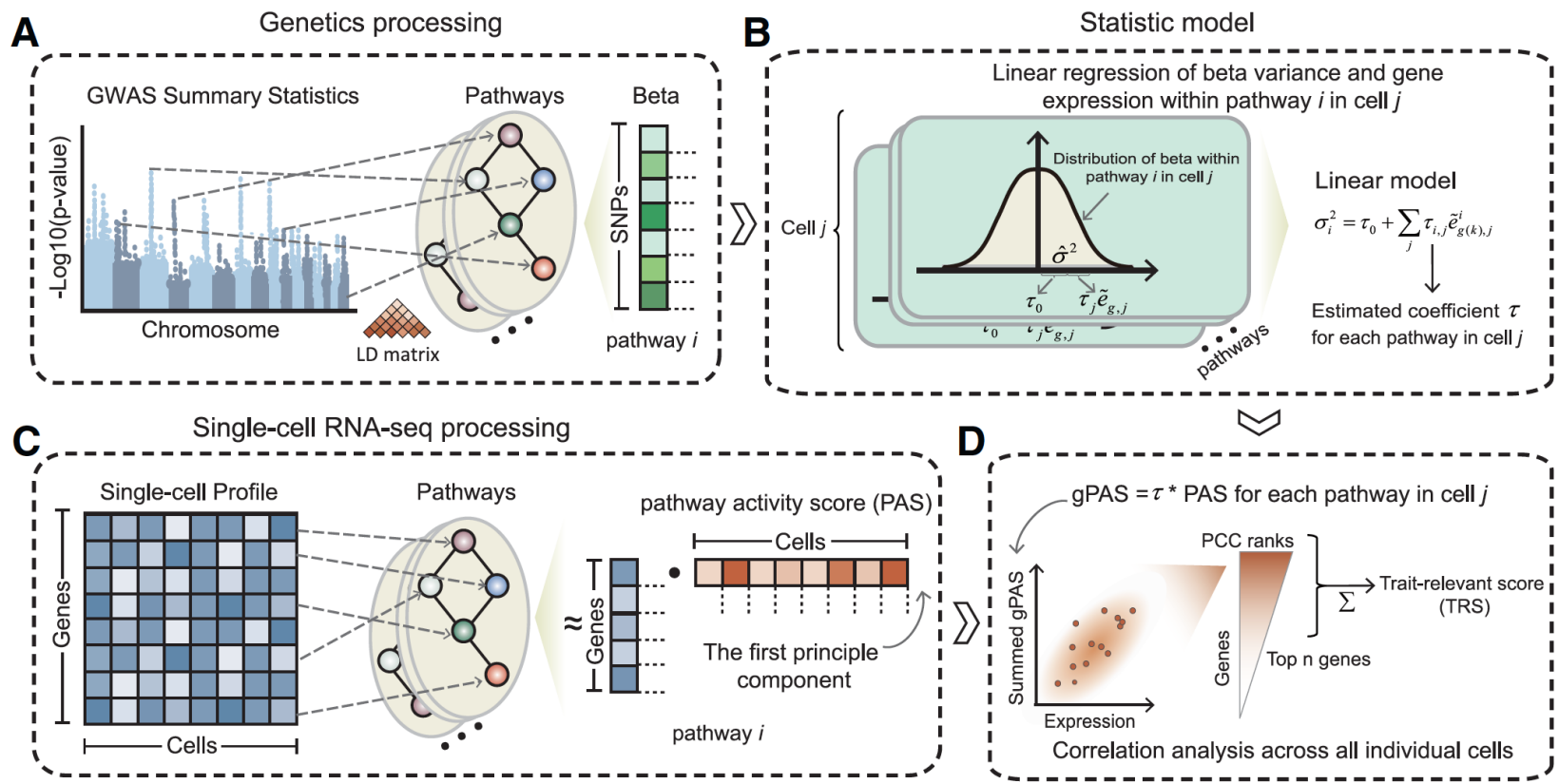
$$PGS_2 = \sum X_2 b_2$$

Polygenic scores (PGS) combining ancestry-aware genetic risk factors.

Just one of my ideas for a new project

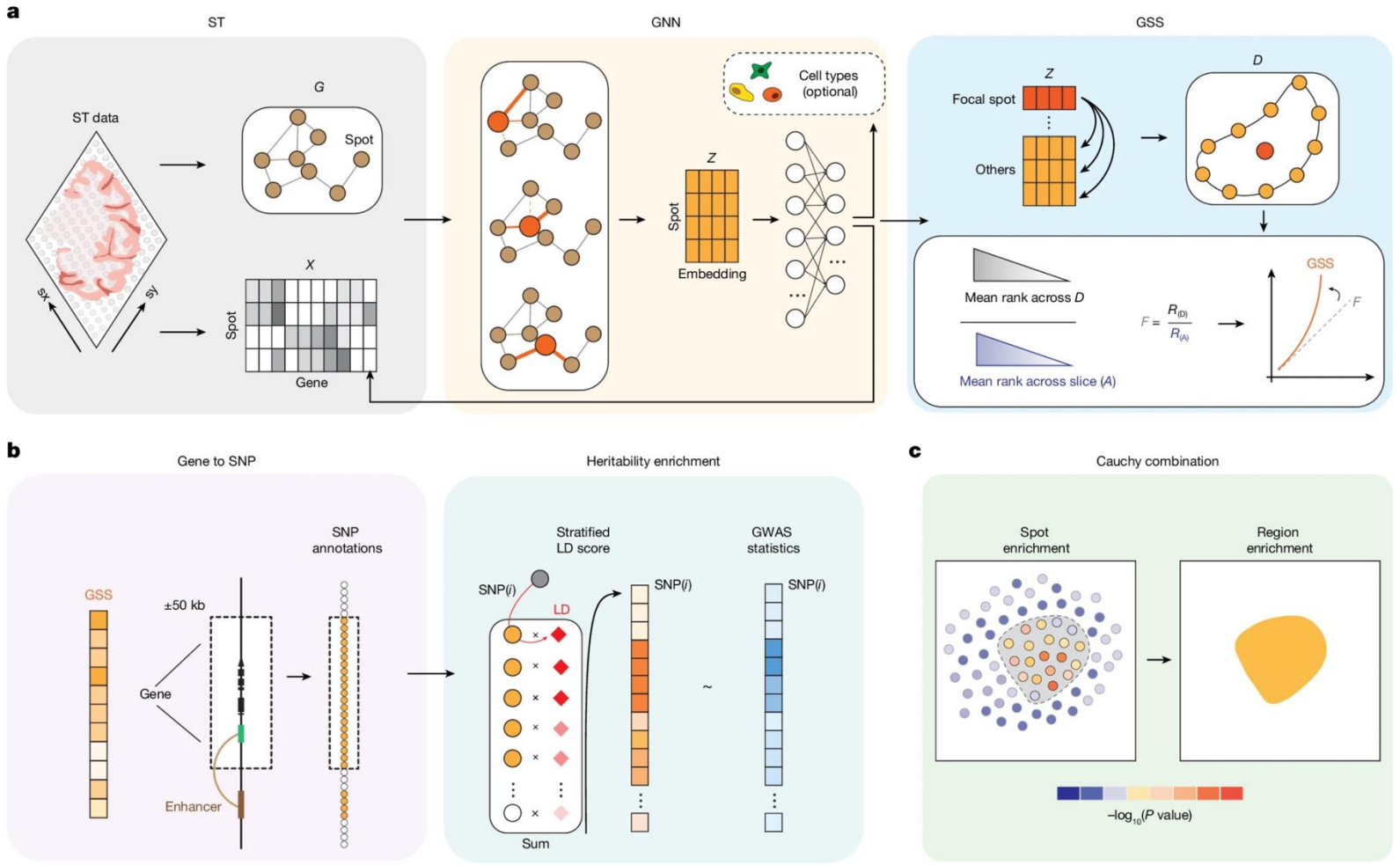
UTILISING PGS WITH SINGLE CELL DATA

Integrate scRNA-seq data with GWAS data to discover cellular context for complex diseases



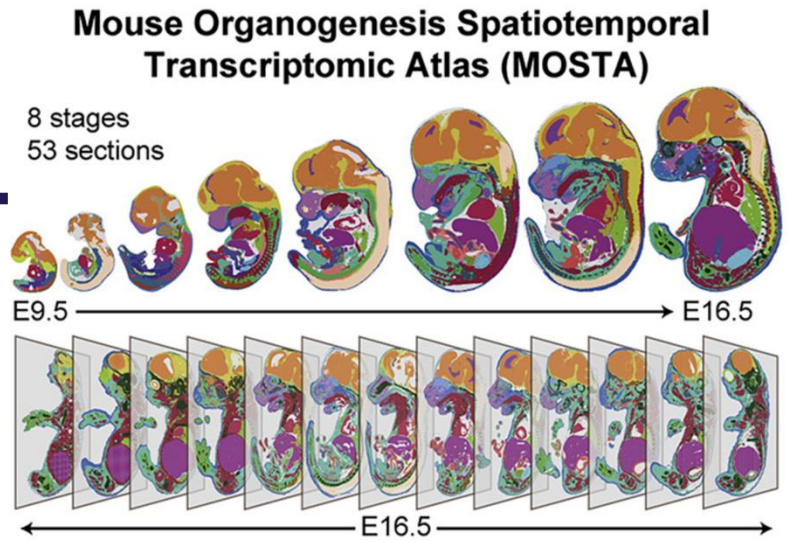
Ma et al. (2023). *Cell Genomics*, *18*:100383

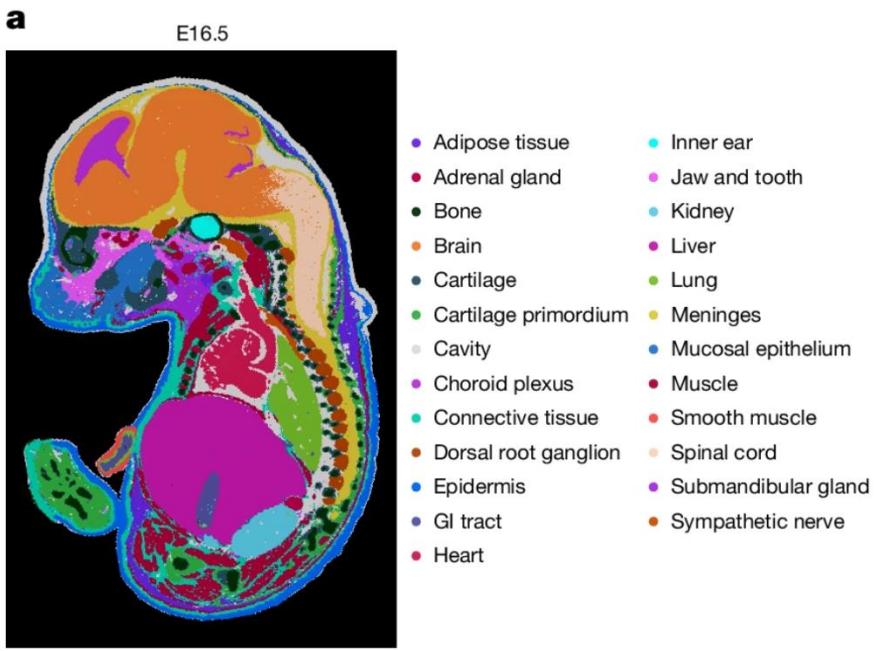
UTILISING PGS WITH SPATIAL TRANSCRIPTOMICS

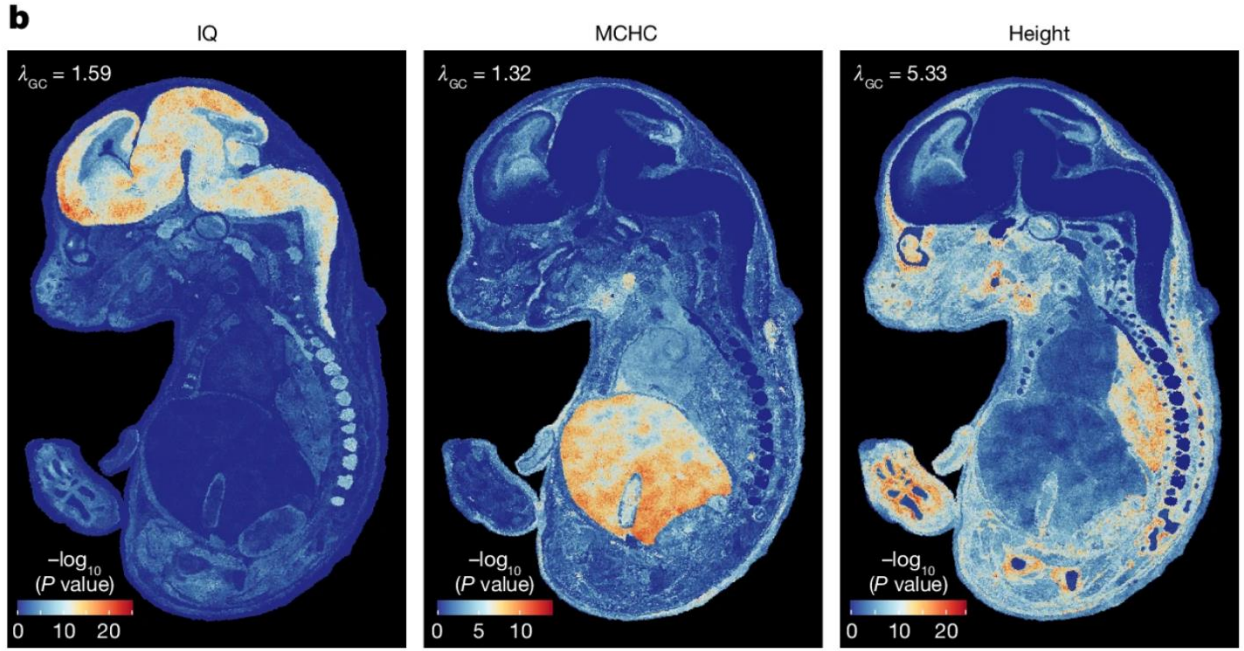


Song et al. (2025). *Nature*, *doi:10.1038/s41586-025-08757-x*

UTILISING PGS WITH SPATIAL TRANSCRIPTOMICS

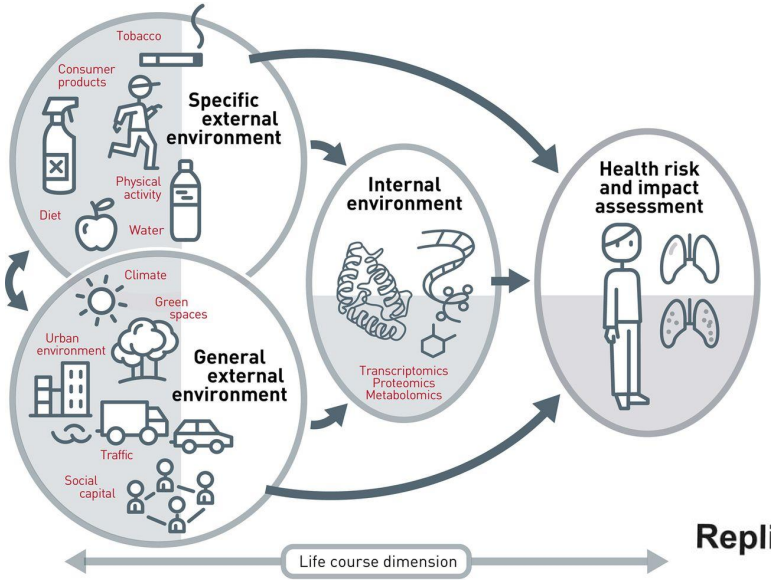






Song et al. (2025). *Nature*, *doi:10.1038/s41586-025-08757-x*

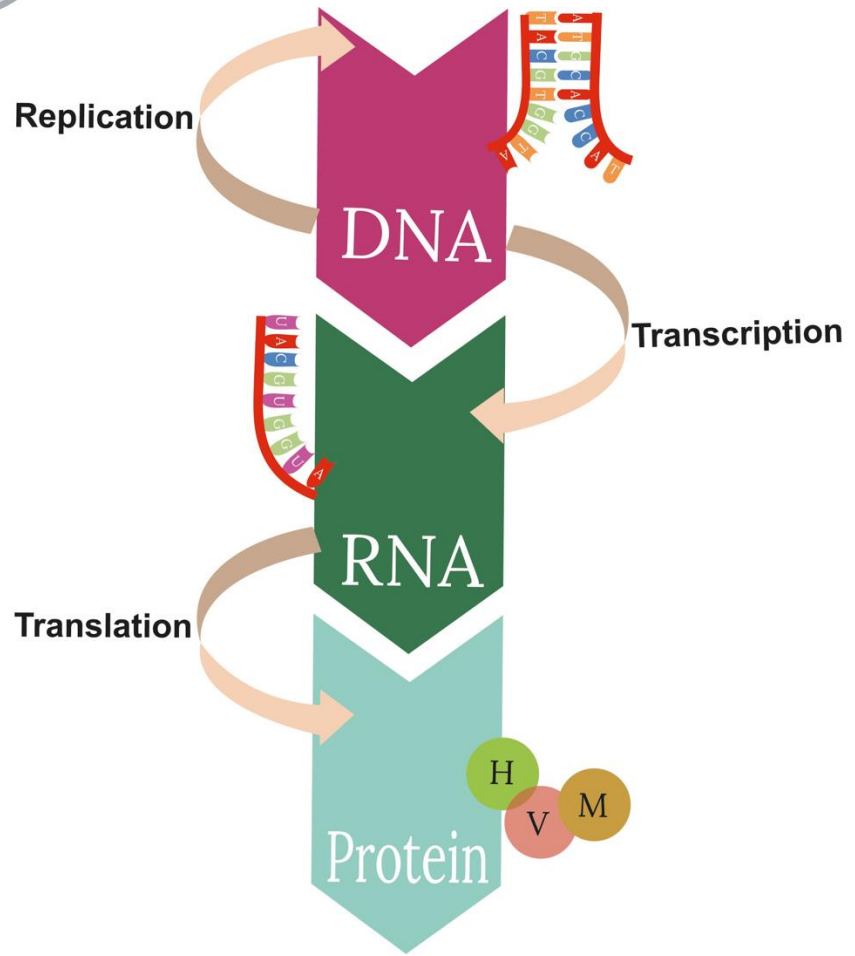
Multifactorial traits have a **genetic component** and an **environmental component**



The **genetic profile is stable** throughout life – thus, a polygenic score may inform about our inherent disease risk.

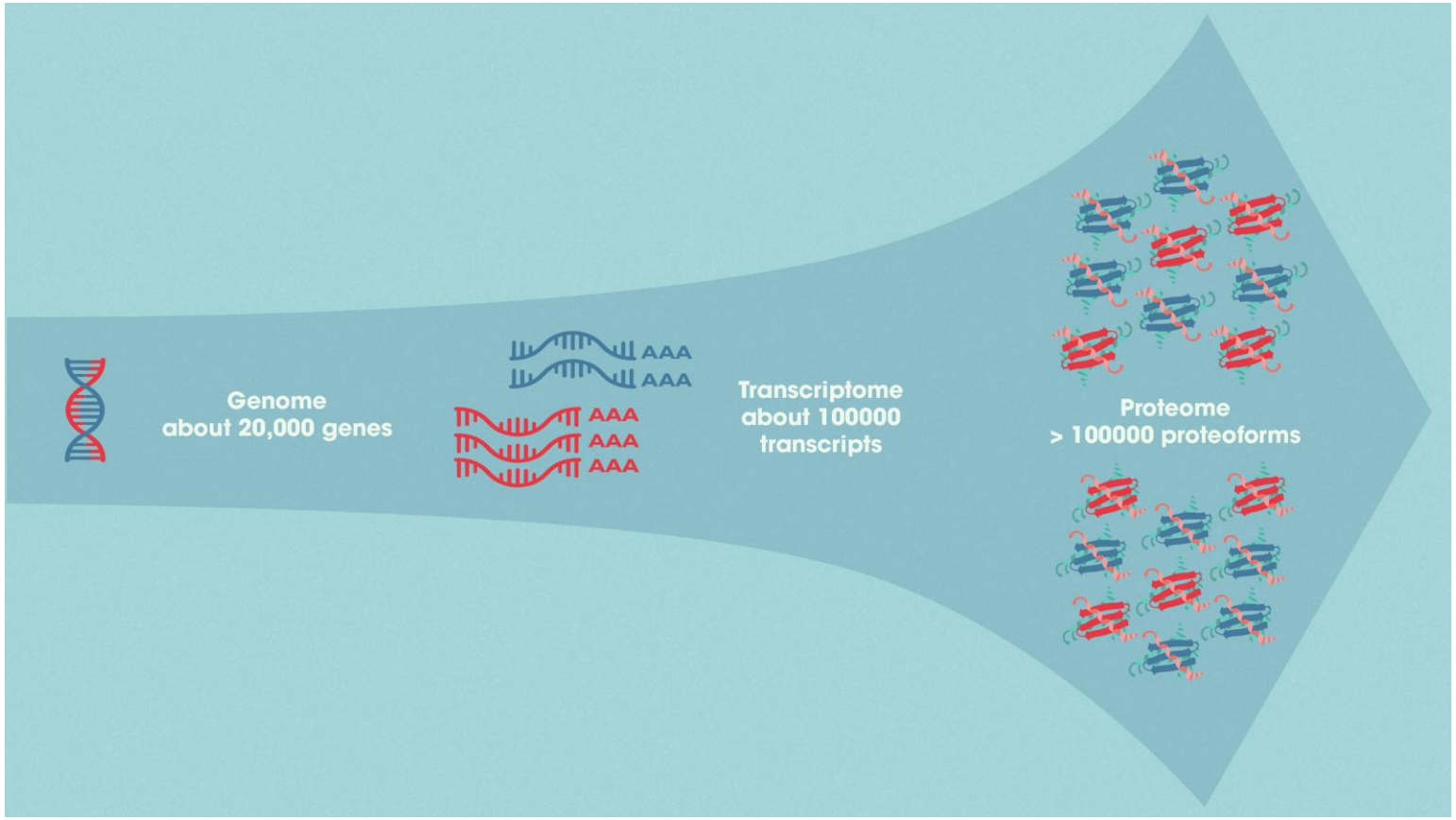
Environmental exposures affect **transcription and translation during life**, which hence might affect disease risk

Therefore, other omic technologies might inform about disease pathology and progression.



GENOME VS PROTEOME

- Unlike the genome, the composition of the proteome is in a constant state of flux over time and throughout the organism
- Environmental exposures affect transcription of DNA, and translation of RNA
- Proteome may help us understand disease pathology



Predicting presence of coronary plaques featuring high-risk characteristics using polygenic risk scores and targeted proteomics in patients with suspected coronary artery disease

Møller, P.L., Rohde, P. D., et al, Circulation: Genomics and Precision medicine, 2023
Møller, P.L., Rohde, P. D., et al, Genome Medicine 2023

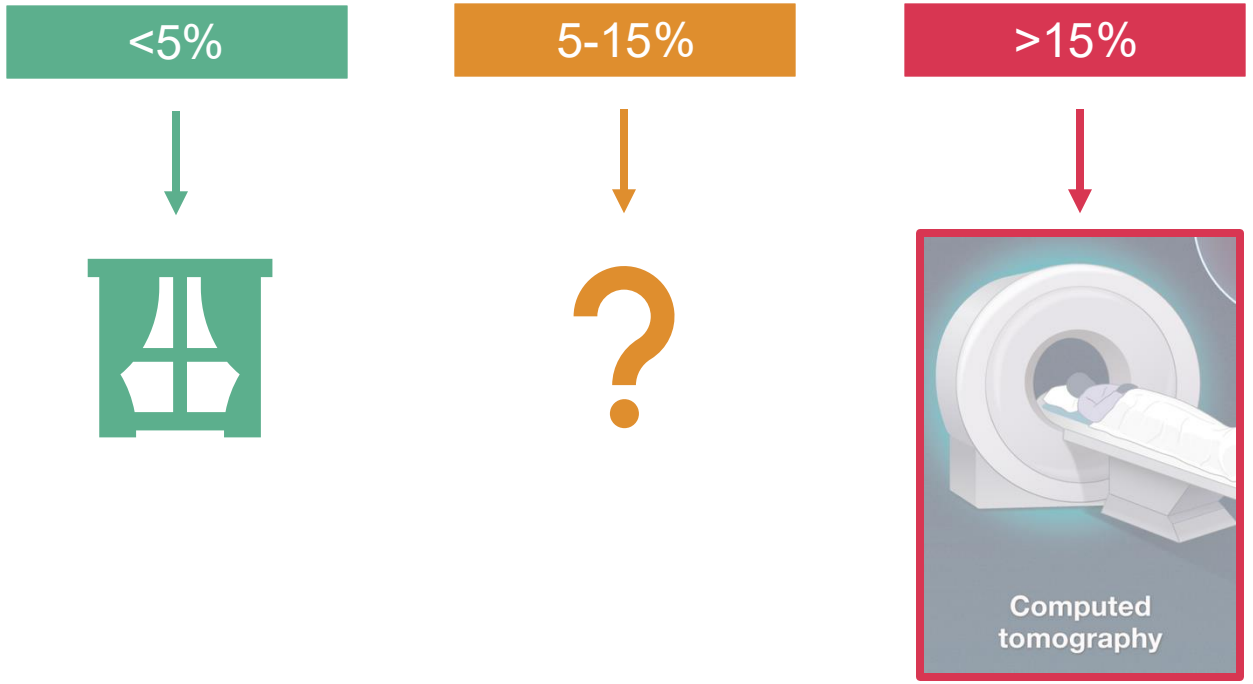
AALBORG
UNIVERSITY

PATIENT WITH CHEST PAIN

Patient with de-novo chest pain

Risk of obstructive **coronary artery disease** (CAD)

Pretest probability (PTP) = gender, age and type of chest pain



From 2019 ESC Guidelines [PMID: 31504439]

Dan-NICAD COHORT

Morten Bøttcher

Simon Winther

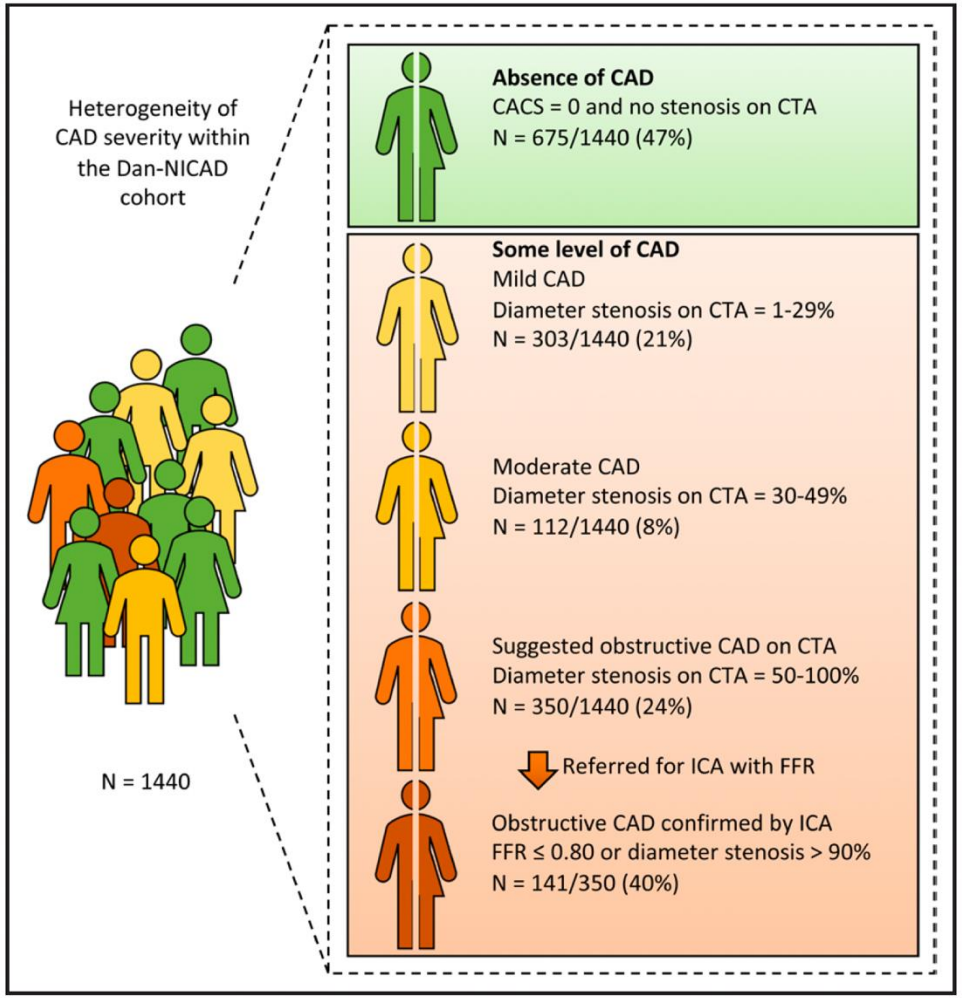
~4000 patients with symptoms of obstructive coronary artery disease (CAD)

All patients undergo coronary CT

1650 patients have **plaque morphology**, **DNA** and **targeted proteomics** [368 proteins, Olink®]

Nissen et al 2016 [PMID: 27225018]
Rasmussen et al 2019 [PMID: 31323454]

CAN WE IDENTIFY *HEALTHY* PATIENTS

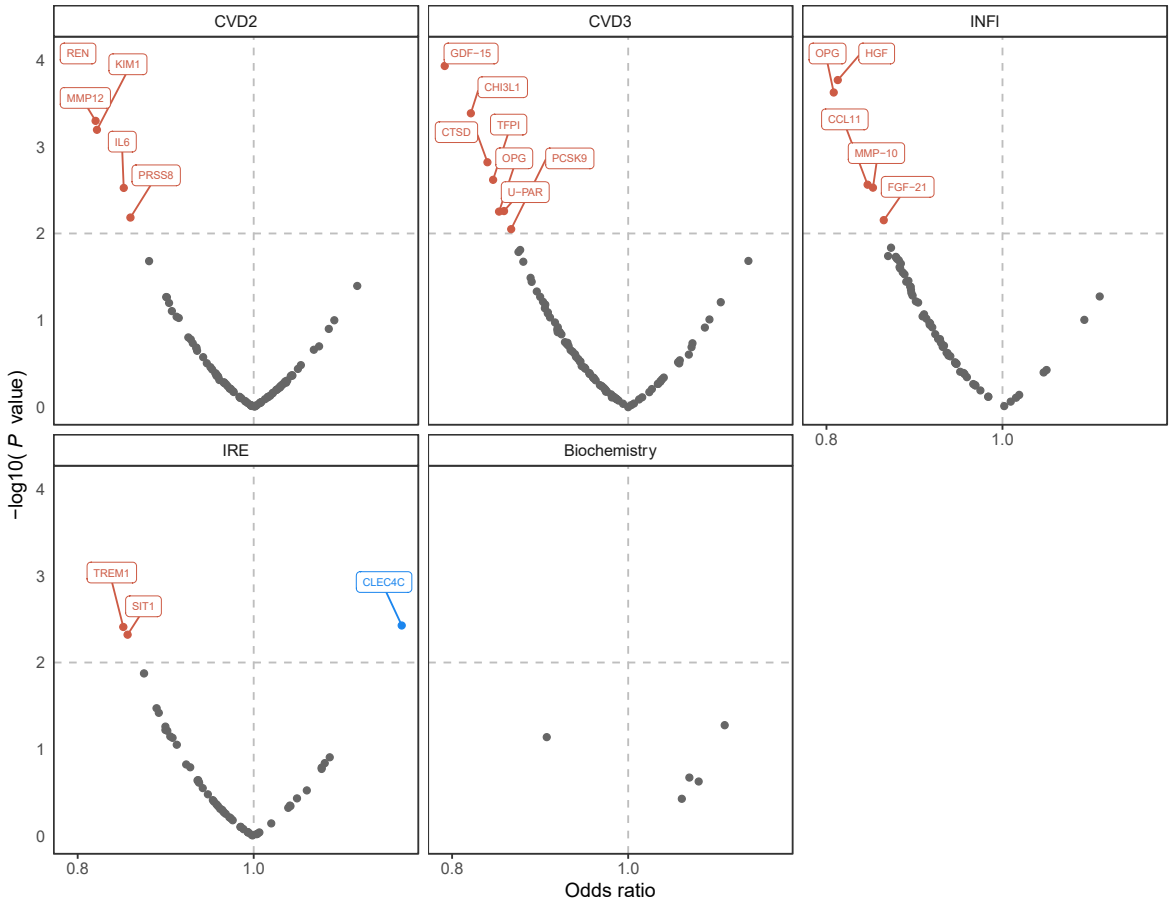


Circulation: Genomic and Precision Medicine

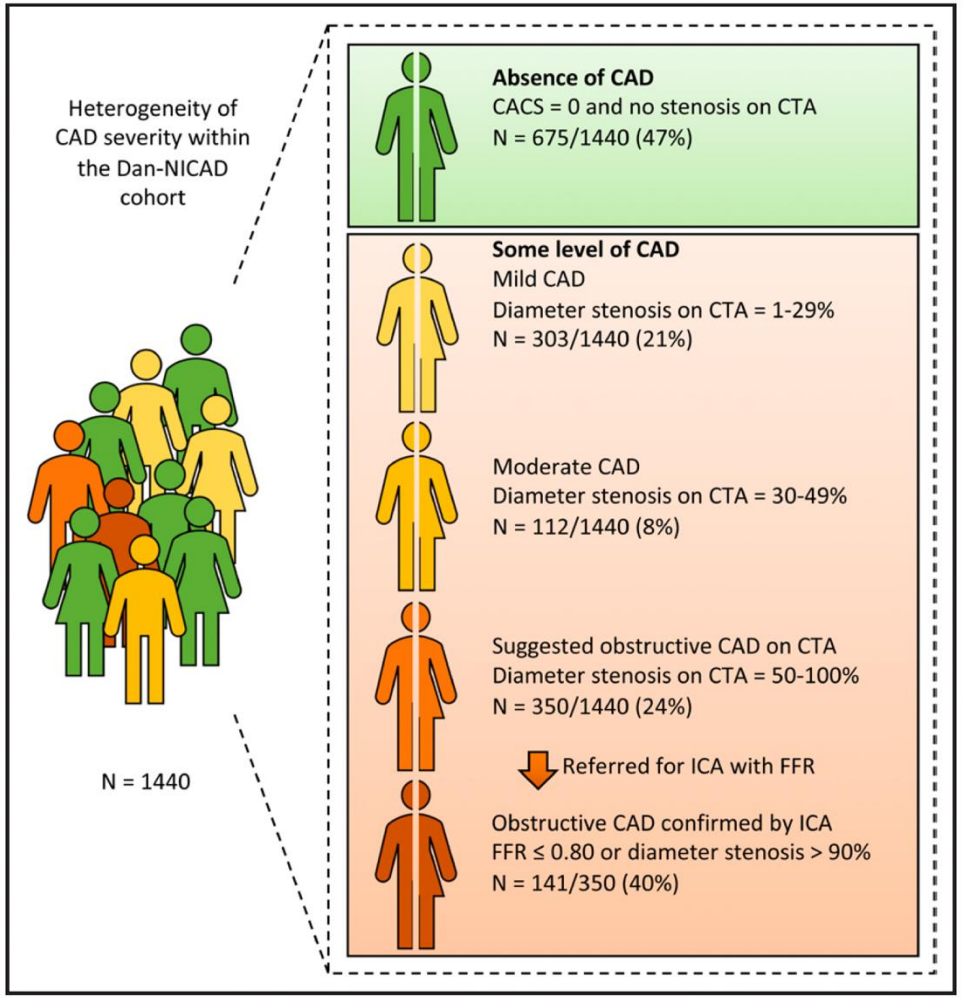
ORIGINAL ARTICLE

Combining Polygenic and Proteomic Risk Scores With Clinical Risk Factors to Improve Performance for Diagnosing Absence of Coronary Artery Disease in Patients With de novo Chest Pain

Peter Loof Møller, MSc; Palle Duun Rohde, MSc, PhD; Jonathan Nørtoft Dahl, MD; Laust Dupont Rasmussen, MD, PhD; Samuel Emil Schmidt, MSc, PhD; Louise Nissen, MD, PhD; Victoria McGilligan, PhD; Jacob F. Bentzon, MD, PhD; Daniel F. Gudbjartsson, MSc, PhD; Kari Stefansson, MD, PhD; Hilma Holm, MD; Simon Winther, MD, PhD; Morten Bøttcher, MD, PhD; Mette Nyegaard, MSc, PhD



CAN WE IDENTIFY *HEALTHY* PATIENTS

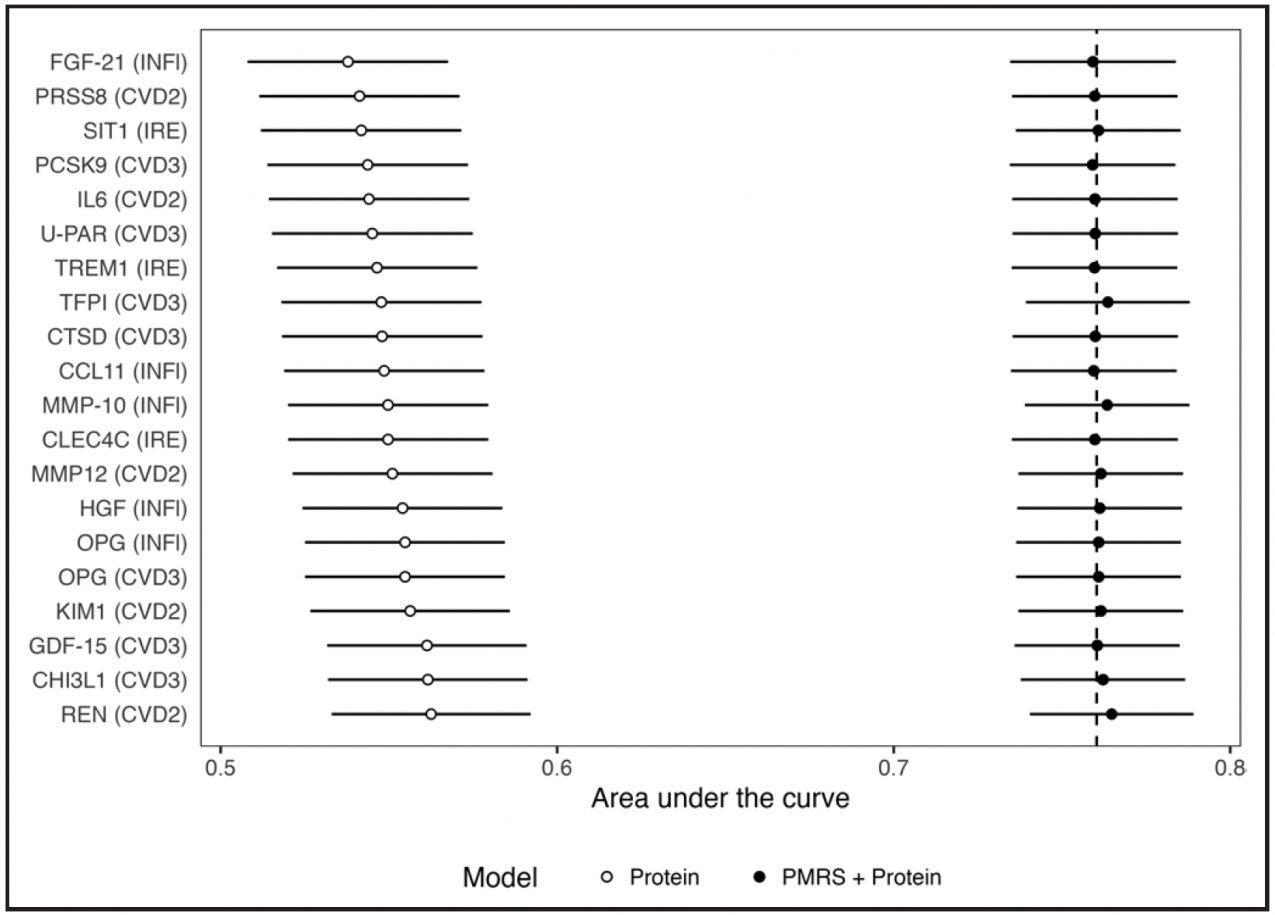


Circulation: Genomic and Precision Medicine

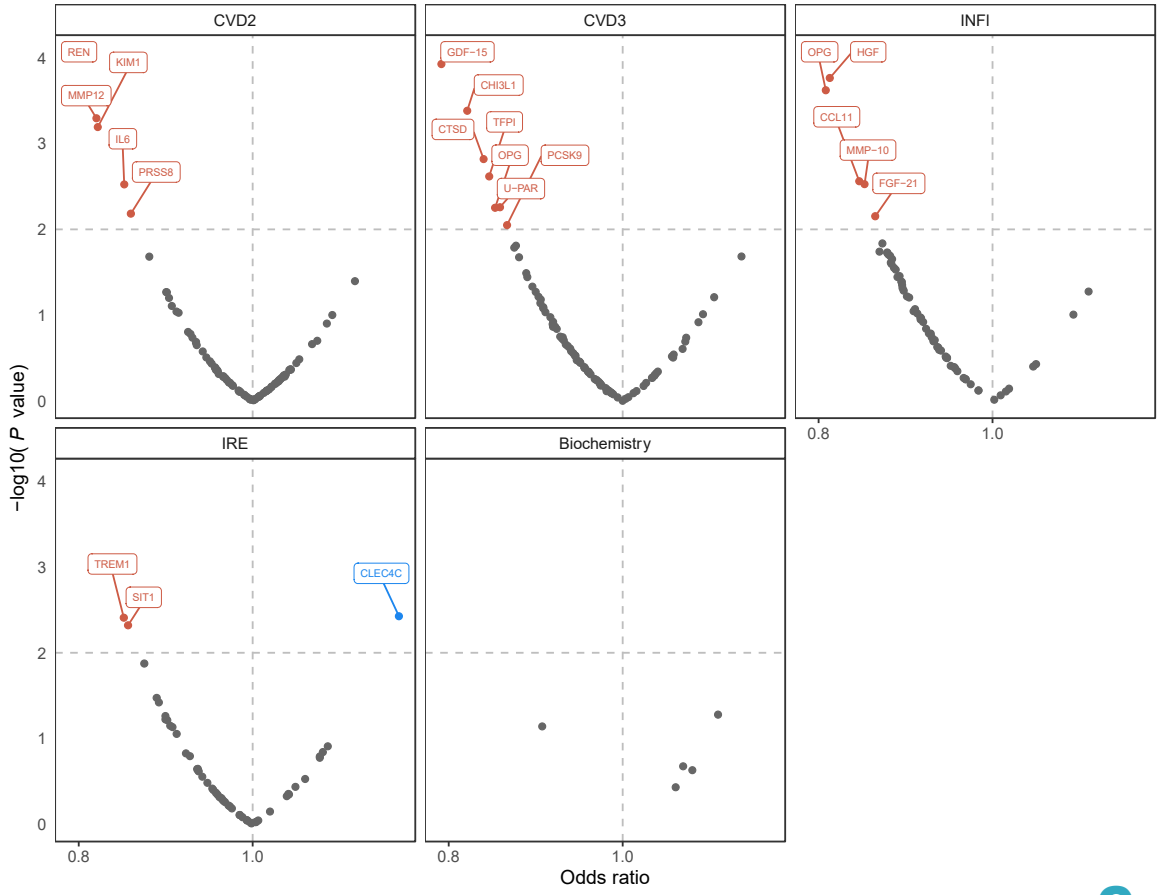
ORIGINAL ARTICLE

Combining Polygenic and Proteomic Risk Scores With Clinical Risk Factors to Improve Performance for Diagnosing Absence of Coronary Artery Disease in Patients With de novo Chest Pain

Peter Loof Møller, MSc; Palle Duun Rohde, MSc, PhD; Jonathan Nørtoft Dahl, MD; Laust Dupont Rasmussen, MD, PhD; Samuel Emil Schmidt, MSc, PhD; Louise Nissen, MD, PhD; Victoria McGilligan, PhD; Jacob F. Bentzon, MD, PhD; Daniel F. Gudbjartsson, MSc, PhD; Kari Stefansson, MD, PhD; Hilma Holm, MD; Simon Winther, MD, PhD; Morten Bøttcher, MD, PhD; Mette Nyegaard, MSc, PhD



CAN WE IDENTIFY *HEALTHY* PATIENTS

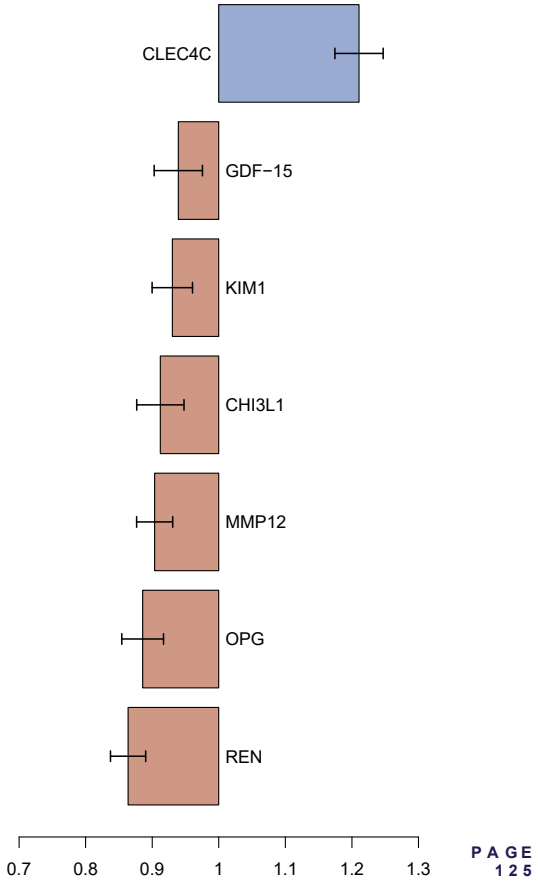


Circulation: Genomic and Precision Medicine

ORIGINAL ARTICLE

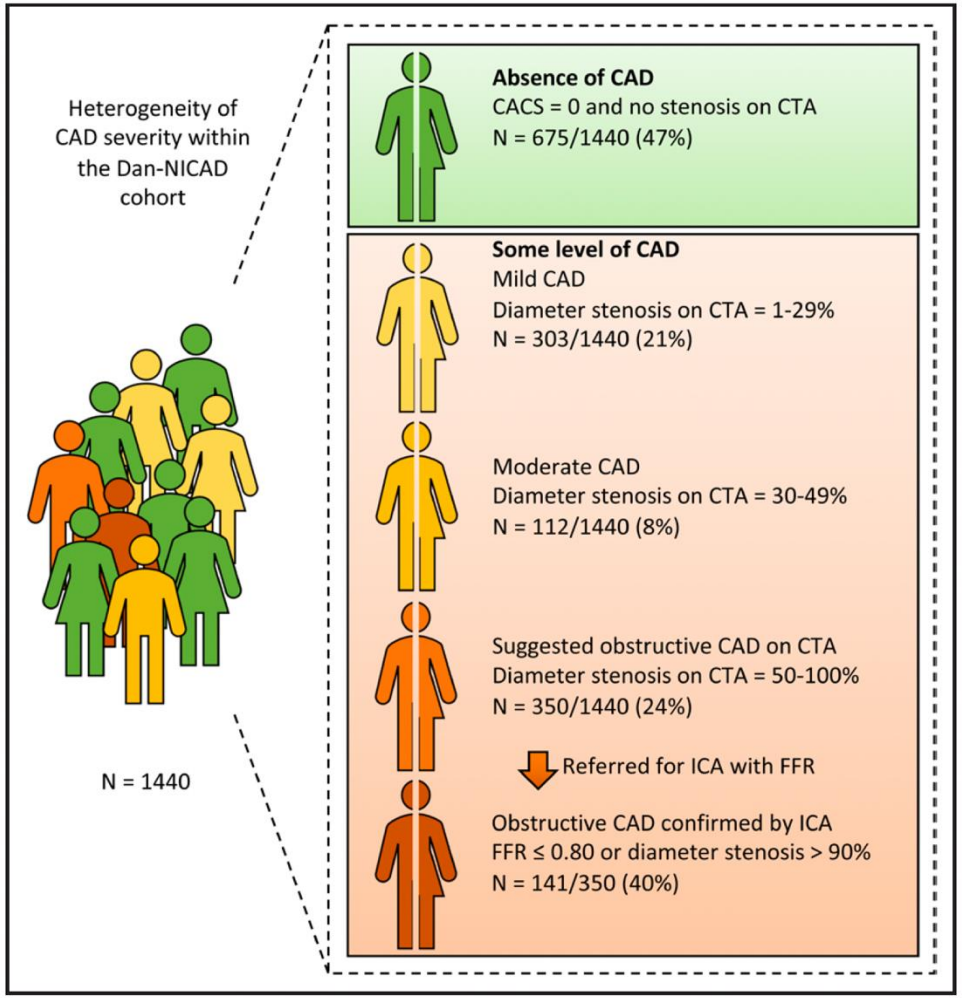
Combining Polygenic and Proteomic Risk Scores With Clinical Risk Factors to Improve Performance for Diagnosing Absence of Coronary Artery Disease in Patients With de novo Chest Pain

Peter Loof Møller, MSc; Palle Duun Rohde, MSc, PhD; Jonathan Nørtoft Dahl, MD; Laust Dupont Rasmussen, MD, PhD; Samuel Emil Schmidt, MSc, PhD; Louise Nissen, MD, PhD; Victoria McGilligan, PhD; Jacob F. Bentzon, MD, PhD; Daniel F. Gudbjartsson, MSc, PhD; Kari Stefansson, MD, PhD; Hilma Holm, MD; Simon Winther, MD, PhD; Morten Bøttcher, MD, PhD; Mette Nyegaard, MSc, PhD



Combining 300+ proteins in ONE model

CAN WE IDENTIFY *HEALTHY* PATIENTS

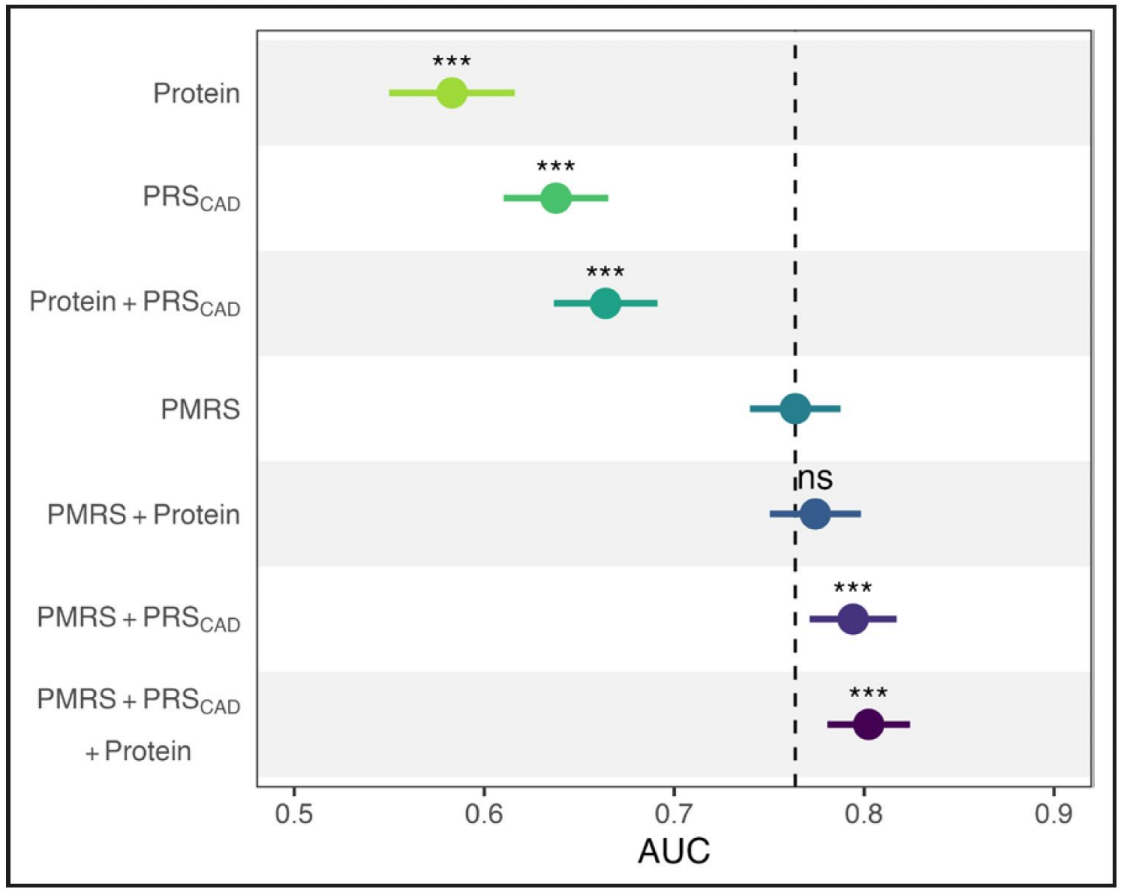


Circulation: Genomic and Precision Medicine

ORIGINAL ARTICLE

Combining Polygenic and Proteomic Risk Scores With Clinical Risk Factors to Improve Performance for Diagnosing Absence of Coronary Artery Disease in Patients With de novo Chest Pain

Peter Loof Møller, MSc; Palle Duun Rohde, MSc, PhD; Jonathan Nørtoft Dahl, MD; Laust Dupont Rasmussen, MD, PhD; Samuel Emil Schmidt, MSc, PhD; Louise Nissen, MD, PhD; Victoria McGilligan, PhD; Jacob F. Bentzon, MD, PhD; Daniel F. Gudbjartsson, MSc, PhD; Kari Stefansson, MD, PhD; Hilma Holm, MD; Simon Winther, MD, PhD; Morten Bøttcher, MD, PhD; Mette Nyegaard, MSc, PhD



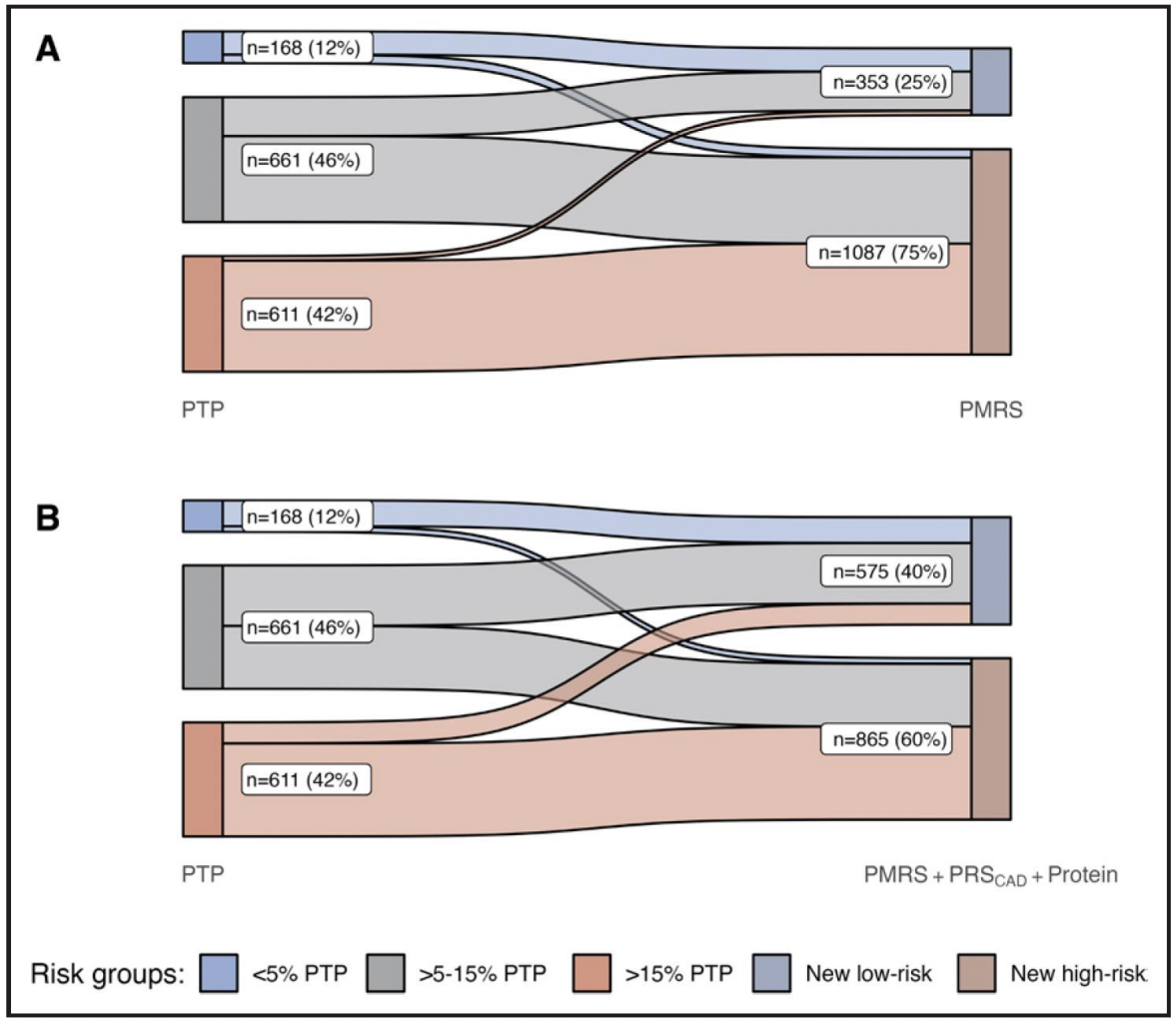
CAN WE IDENTIFY *HEALTHY* PATIENTS

Circulation: Genomic and Precision Medicine

ORIGINAL ARTICLE

Combining Polygenic and Proteomic Risk Scores With Clinical Risk Factors to Improve Performance for Diagnosing Absence of Coronary Artery Disease in Patients With de novo Chest Pain

Peter Loof Møller, MSc; Palle Duun Rohde, MSc, PhD; Jonathan Nørtoft Dahl, MD; Laust Dupont Rasmussen, MD, PhD; Samuel Emil Schmidt, MSc, PhD; Louise Nissen, MD, PhD; Victoria McGilligan, PhD; Jacob F. Bentzon, MD, PhD; Daniel F. Gudbjartsson, MSc, PhD; Kari Stefansson, MD, PhD; Hilma Holm, MD; Simon Winther, MD, PhD; Morten Bøttcher, MD, PhD; Mette Nyegaard, MSc, PhD



Tradition method

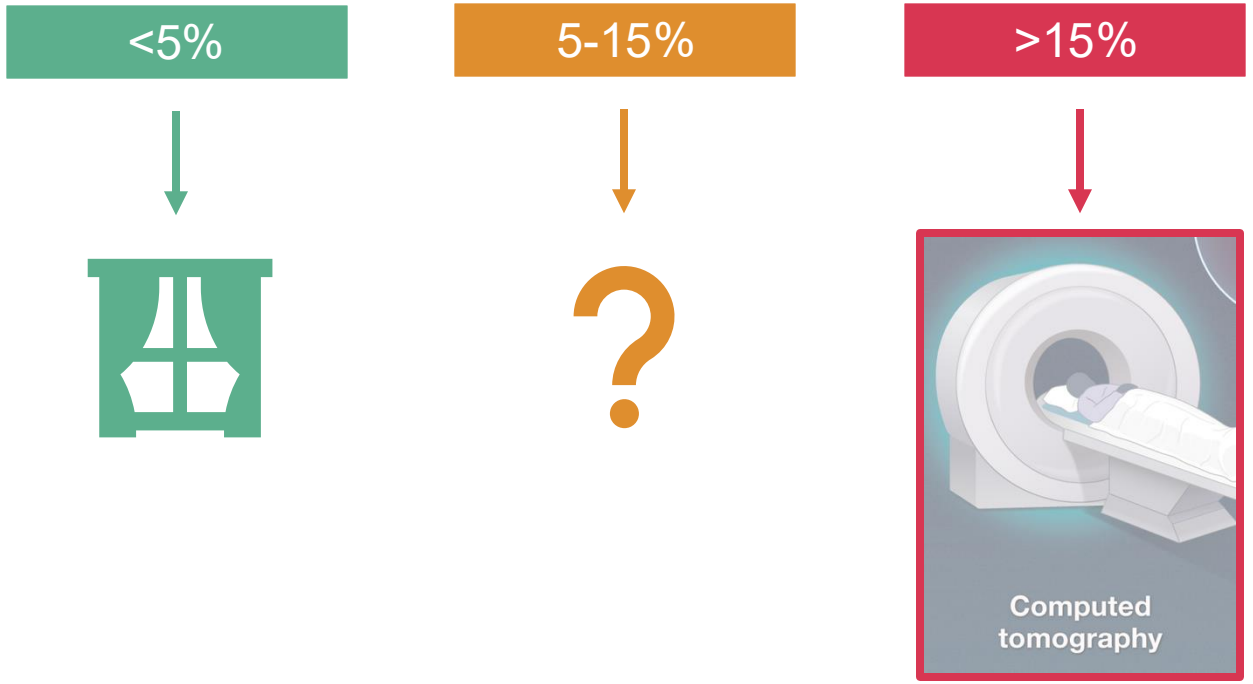
Combining proteomics, genomics and clinical risk factors has better discriminative ability

PATIENT WITH CHEST PAIN

Patient with de-novo chest pain

Risk of obstructive **coronary artery disease** (CAD)

Pretest probability (PTP) = gender, age and type of chest pain

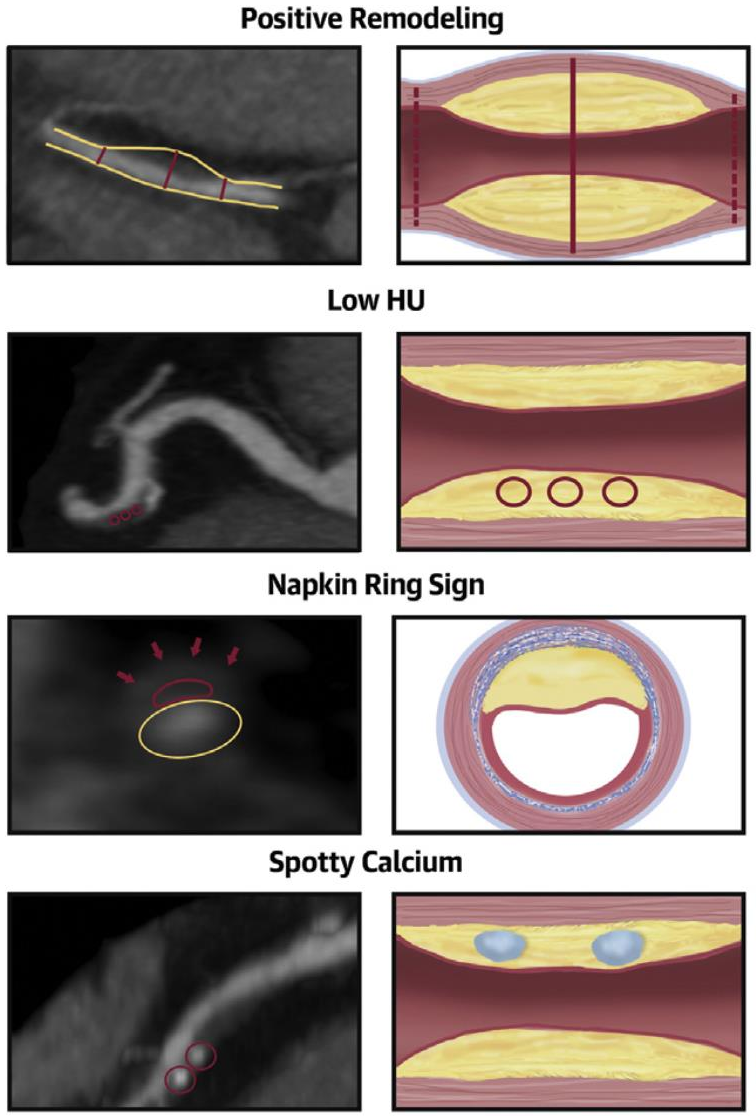


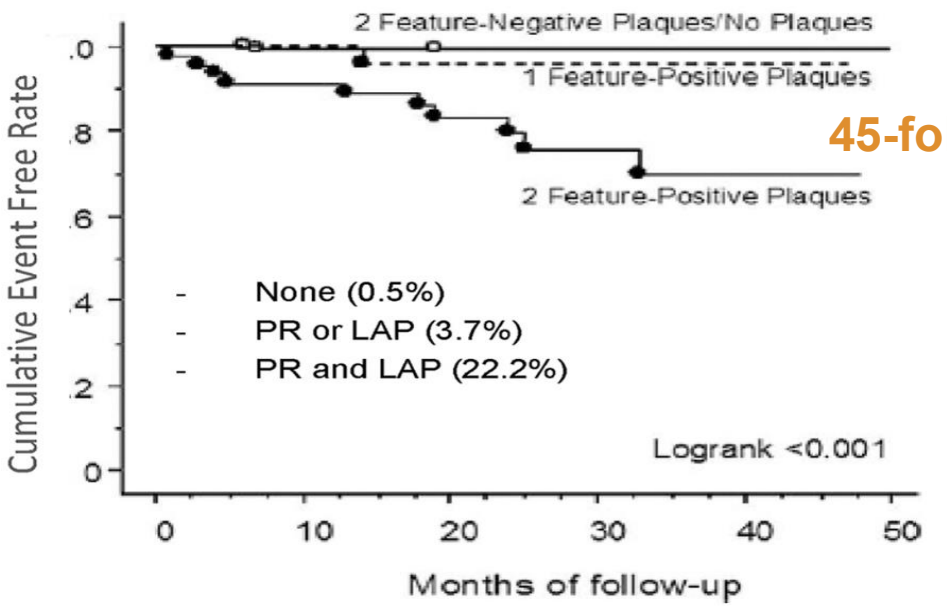
From 2019 ESC Guidelines [PMID: 31504439]

LOW PTP

Some patients have a low PTP-score, but still have chest pain, and are at risk of a cardiac event.

HIGH RISK CORONARY PLAQUES [HRP]





45-fold higher likelihood of cardiac events

Can DNA and targeted proteomics predict the presence of HRP in low risk patients?

Puchner et al 2014 [PMID: 25125300]
Shaw et al 2021 [PMID: 33303383]

PREDICTING HRP

Møller et al. Genome Medicine (2024) 16:40
https://doi.org/10.1186/s13073-024-01313-8

Genome Medicine

RESEARCH Open Access

Predicting the presence of coronary plaques featuring high-risk characteristics using polygenic risk scores and targeted proteomics in patients with suspected coronary artery disease

Peter Loof Møller[1,2], Palle Duun Rohde[2], Jonathan Nørtoft Dahl[3,4], Laust Dupont Rasmussen[3,10], Louise Nissen[3,4], Samuel Emil Schmidt[2], Victoria McGilligan[5], Daniel F. Gudbjartsson[6,7], Kari Stefansson[6,8], Hilma Holm[6], Jacob Fog Bentzon[4,9], Morten Bøttcher[3,4], Simon Winther[3,4] and Mette Nyegaard[2*]

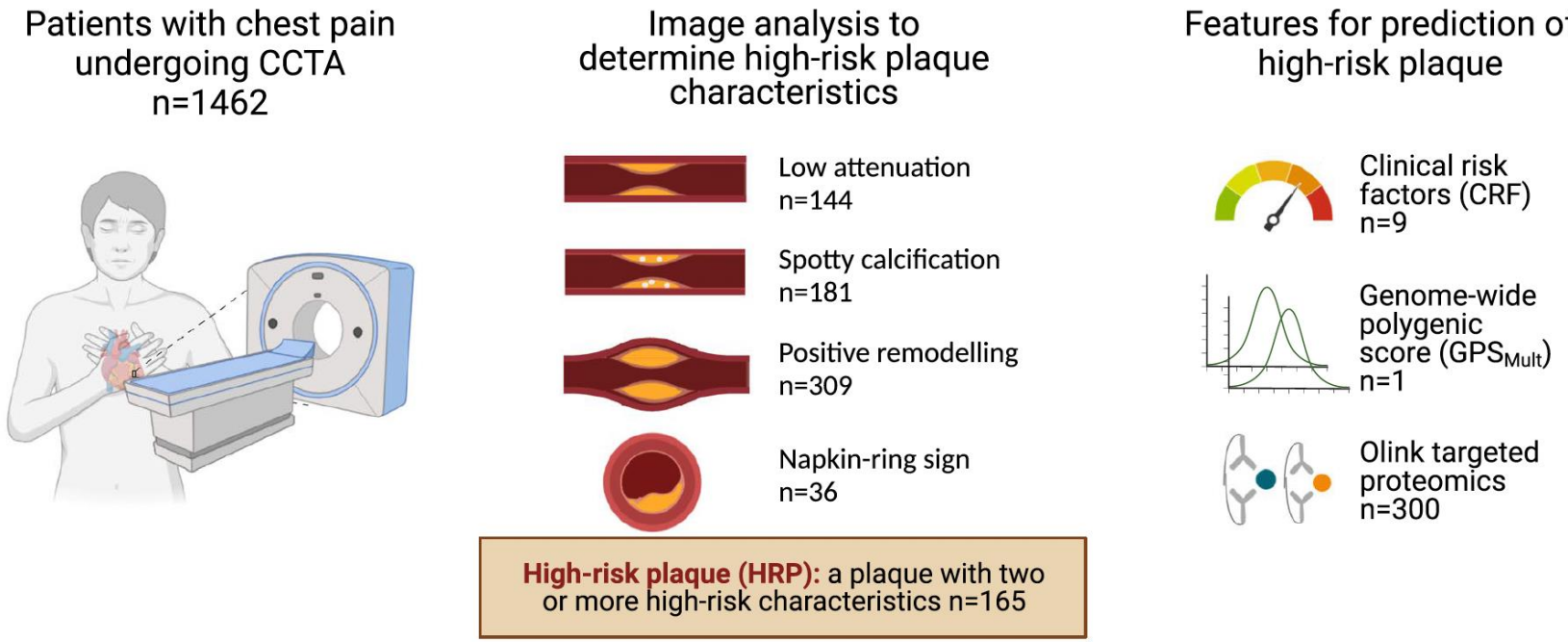


Fig. 1 Study design. 1462 patients underwent coronary computed tomography angiography (CCTA), followed by image analysis of high-risk plaque (HRP) characteristics. Finally, nine clinical risk factors, one multi-trait multi-ancestry genome-wide polygenic score (GPS_{Mult}), and 300 proteins were used to predict HRP presence

PREDICTING HRP

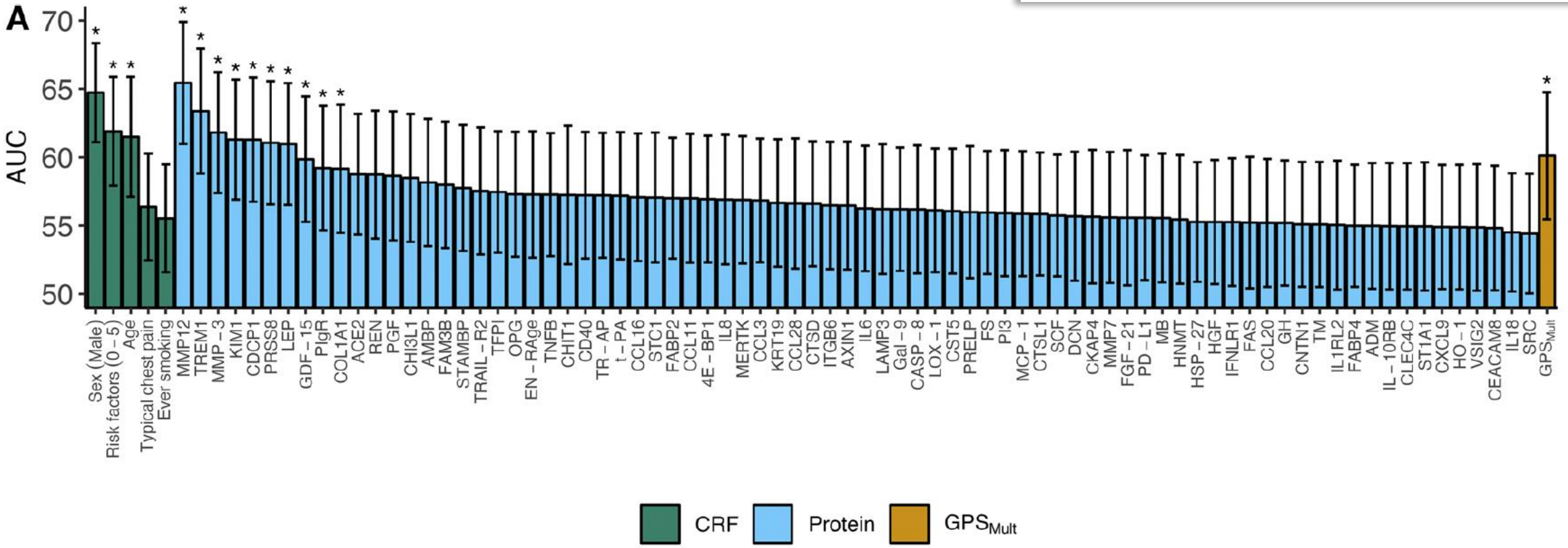
Møller et al. Genome Medicine (2024) 16:40
https://doi.org/10.1186/s13073-024-01313-8

Genome Medicine

RESEARCH Open Access

Predicting the presence of coronary plaques featuring high-risk characteristics using polygenic risk scores and targeted proteomics in patients with suspected coronary artery disease

Peter Loof Møller[1,2], Palle Duun Rohde[2], Jonathan Nørtoft Dahl[3,4], Laust Dupont Rasmussen[3,10], Louise Nissen[3,4], Samuel Emil Schmidt[2], Victoria McGilligan[5], Daniel F. Gudbjartsson[6,7], Kari Stefansson[6,8], Hilma Holm[6], Jacob Fog Bentzon[4,9], Morten Bøttcher[3,4], Simon Winther[3,4] and Mette Nyegaard[2*]



PREDICTING HRP

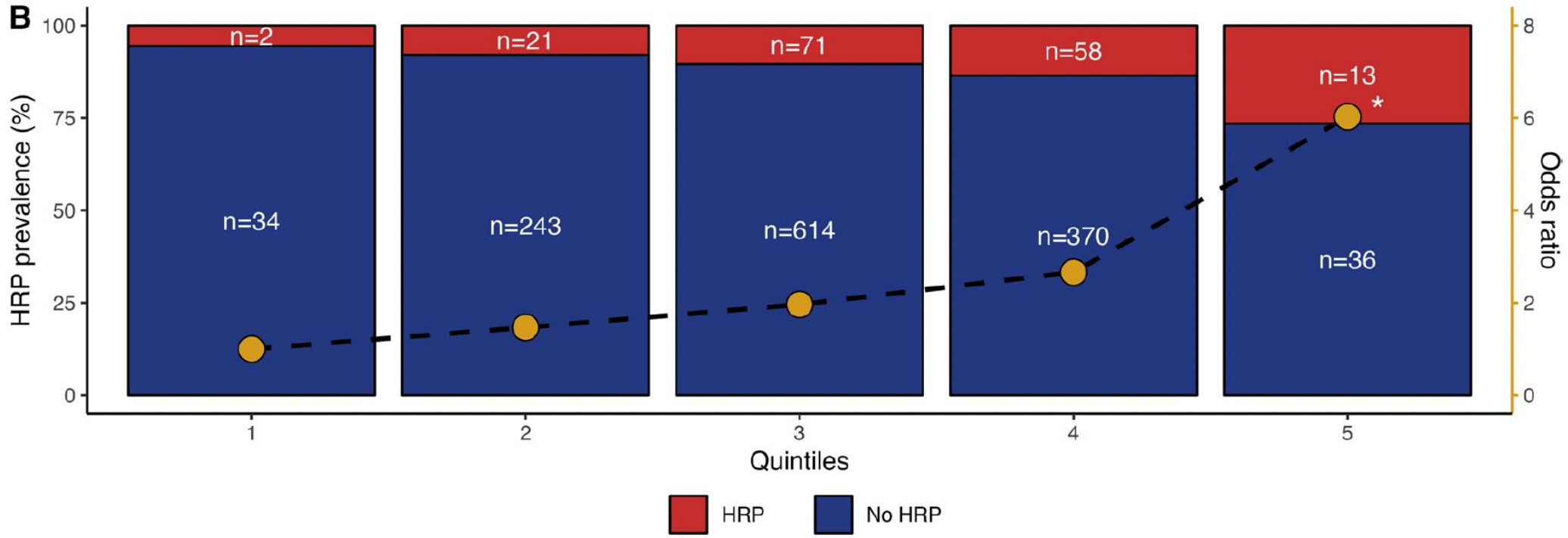
Møller et al. Genome Medicine (2024) 16:40
https://doi.org/10.1186/s13073-024-01313-8

Genome Medicine

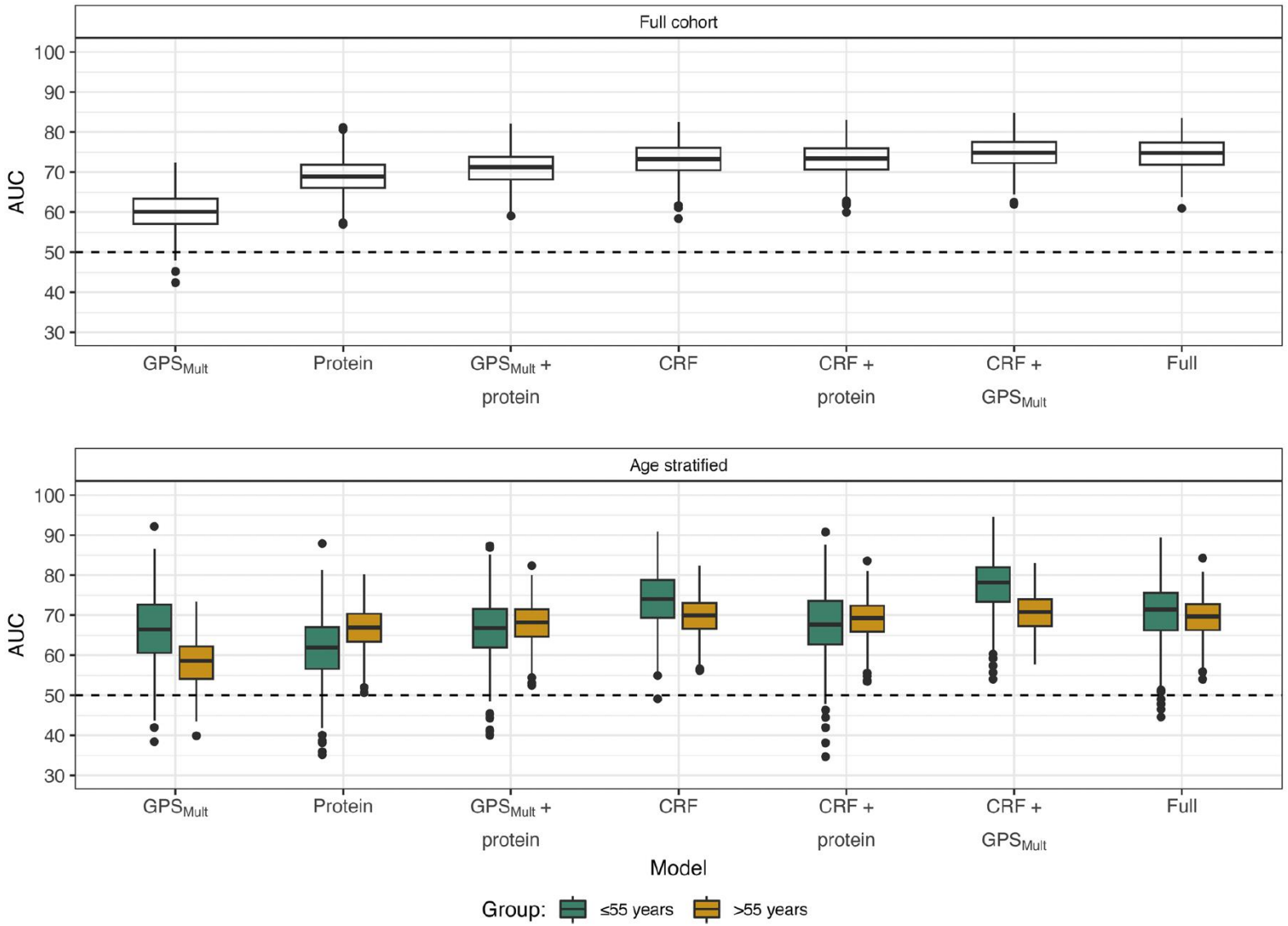
RESEARCH Open Access

Predicting the presence of coronary plaques featuring high-risk characteristics using polygenic risk scores and targeted proteomics in patients with suspected coronary artery disease

Peter Loof Møller[1,2], Palle Duun Rohde[2], Jonathan Nørtoft Dahl[3,4], Laust Dupont Rasmussen[3,10], Louise Nissen[3,4], Samuel Emil Schmidt[2], Victoria McGilligan[5], Daniel F. Gudbjartsson[6,7], Kari Stefansson[6,8], Hilma Holm[6], Jacob Fog Bentzon[4,9], Morten Bøttcher[3,4], Simon Winther[3,4] and Mette Nyegaard[2*]



PREDICTING HRP



Møller et al. Genome Medicine (2024) 16:40
https://doi.org/10.1186/s13073-024-01313-8

Genome Medicine

RESEARCH Open Access

Predicting the presence of coronary plaques featuring high-risk characteristics using polygenic risk scores and targeted proteomics in patients with suspected coronary artery disease

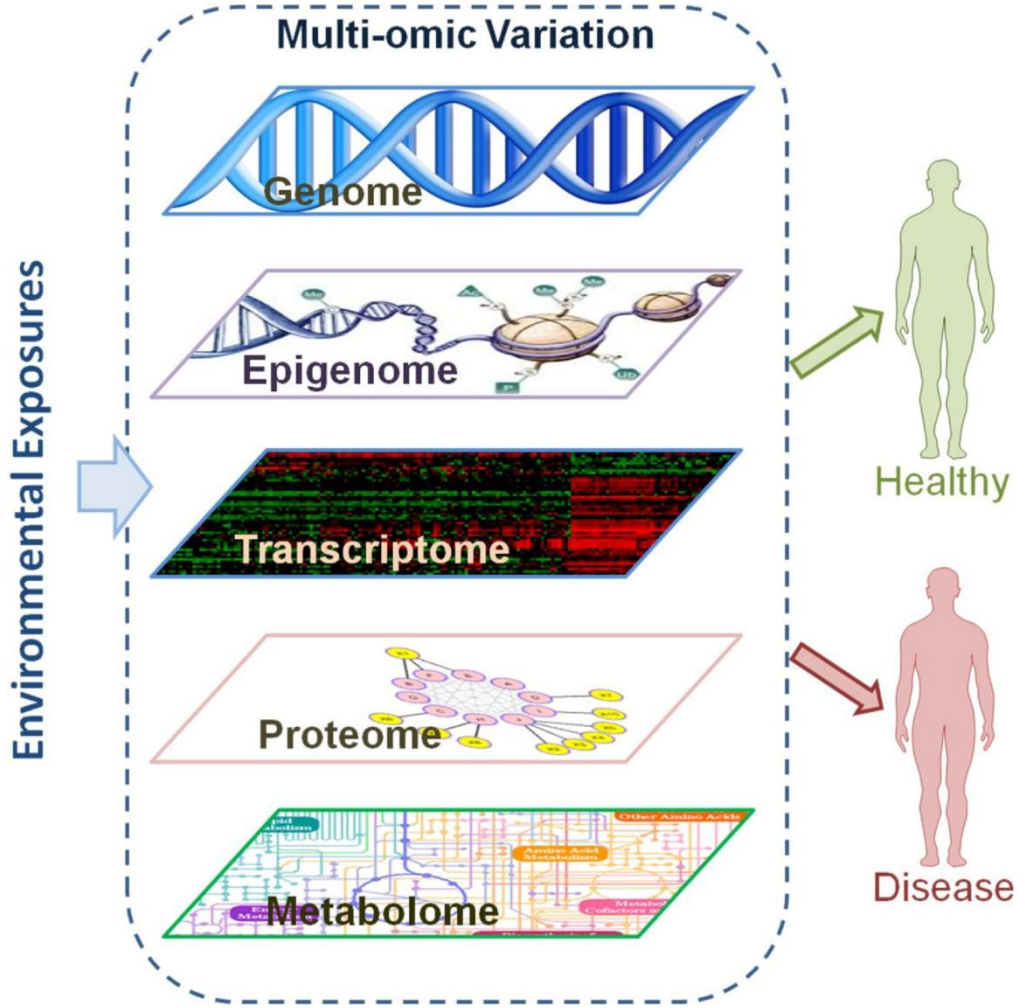
Peter Loof Møller[1,2], Palle Duun Rohde[2], Jonathan Nørtoft Dahl[3,4], Laust Dupont Rasmussen[3,10], Louise Nissen[3,4], Samuel Emil Schmidt[2], Victoria McGilligan[5], Daniel F. Gudbjartsson[6,7], Kari Stefansson[6,8], Hilma Holm[6], Jacob Fog Bentzon[4,9], Morten Bøttcher[3,4], Simon Winther[3,4] and Mette Nyegaard[2*]

Combining CRFs with PGS increased prediction accuracy, while the inclusion of proteomics provided no significant improvement to either the CRF the CRF + PGS

The PGS works best in young individuals – why?

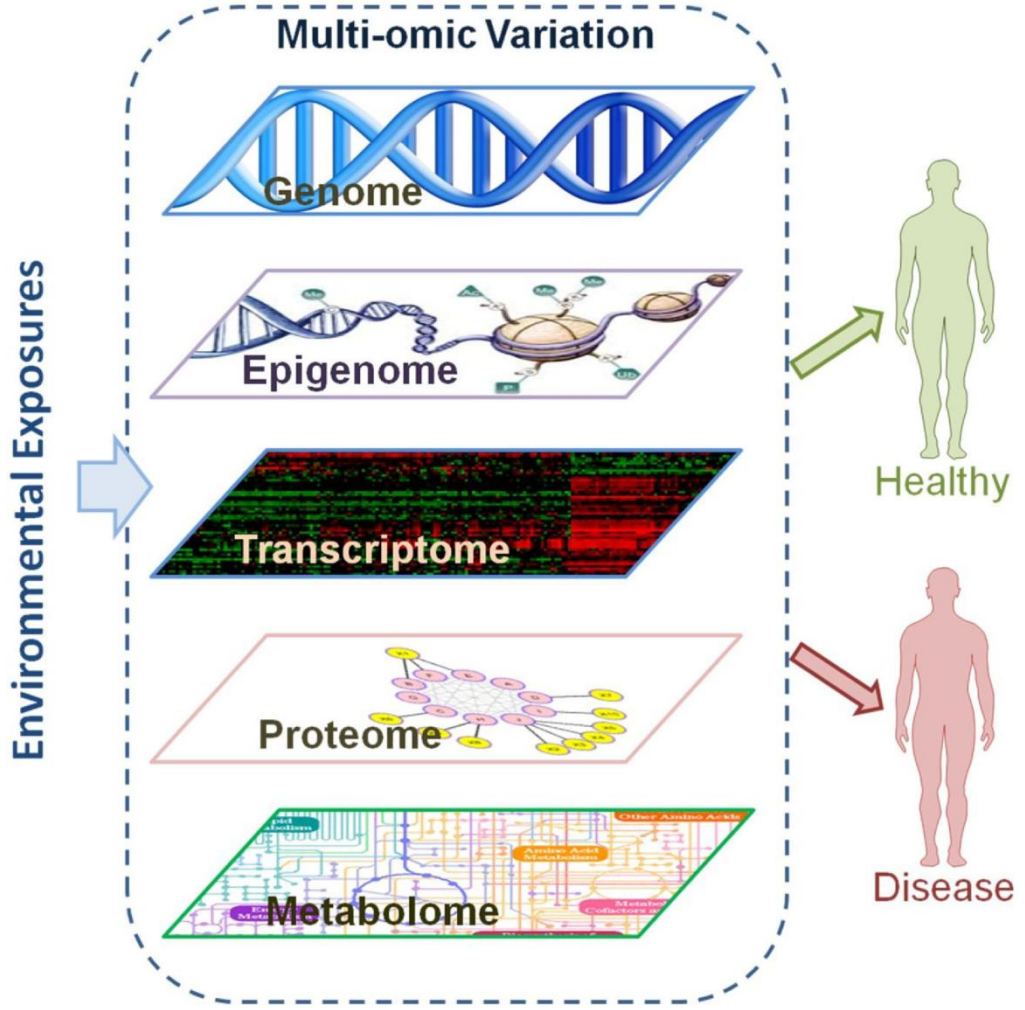
VARIATION AT MULTIPLE LAYERS

- Biological data are now being collected in large scale aross the different layers of molecular organisation.



VARIATION AT MULTIPLE LAYERS

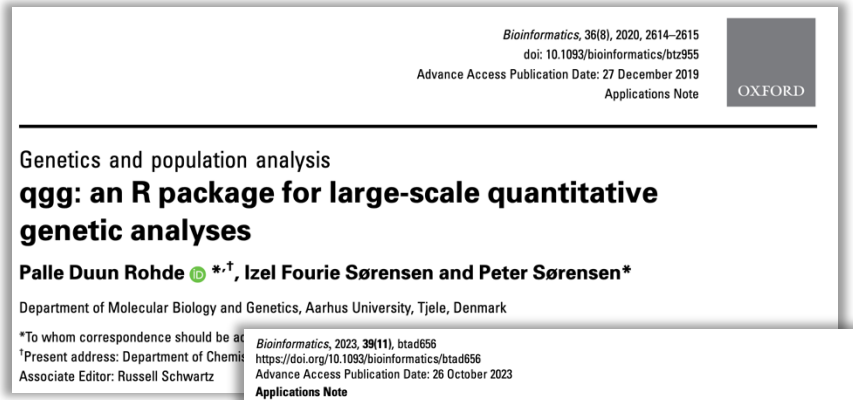
- Biological data are now being collected in large scale aross the different layers of molecular organisation.
- Variation within each layer has generated knowledge about human complex diseases.



SESSION 4

- Many challenges – how to circumvent these?
- Enhancing biological understanding?
- PGS in combination with proteomics

A toolbox for statistical genetic analyses of complex traits



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 *,†, **Izel Fourie Sørensen and Peter Sørensen***

Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark

*To whom correspondence should be ad
†Present address: Department of Chemis
Associate Editor: Russell Schwartz

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

OXFORD

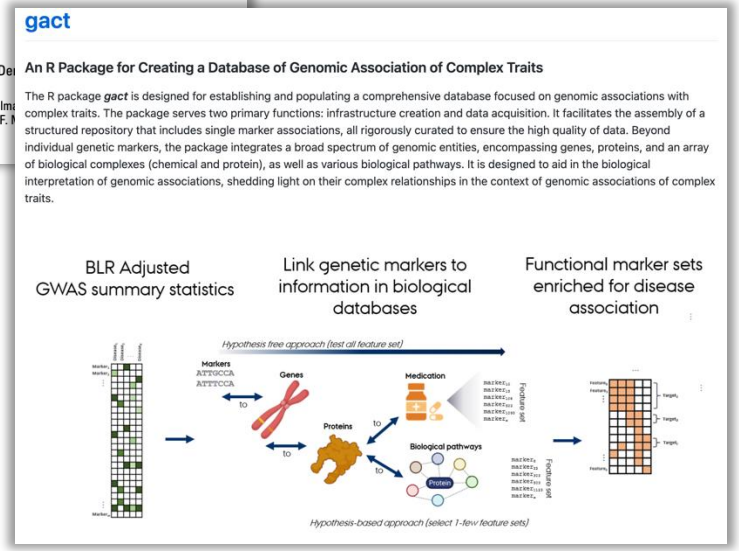
Genetics and population analysis

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, De
[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
Denmark. E-mail: palledr@hst.aau.dk (P.D.R.); Center for Quantitative Genetics and Genomics, Aarhus University, C. F. M
E-mail: pso@qgg.au.dk (P.S.)
Associate Editor: Christina Kendziorski



gact

An R Package for Creating a Database of Genomic Association of Complex Traits

The R package ***gact*** is designed for establishing and populating a comprehensive database focused on genomic associations with complex traits. The package serves two primary functions: infrastructure creation and data acquisition. It facilitates the assembly of a structured repository that includes single marker associations, all rigorously curated to ensure the high quality of data. Beyond individual genetic markers, the package integrates a broad spectrum of genomic entities, encompassing genes, proteins, and an array of biological complexes (chemical and protein), as well as various biological pathways. It is designed to aid in the biological interpretation of genomic associations, shedding light on their complex relationships in the context of genomic associations of complex traits.

Tutorials

Below is a set of tutorials used for the qgg package:

This tutorial provides a brief introduction to R package qgg using small simulated data examples.
Practicals_brief_introduction

This tutorial provides an introduction to R package qgg using 1000G data.
Practicals_1000G_tutorials

This tutorial provide a simple introduction to polygenic risk scoring (PRS) of complex traits and diseases using simulated data. The practical will be a mix of theoretical and practical exercises in R that are used for illustrating/applying the theory presented in the corresponding lecture notes on polygenic risk scoring.
Practicals_human_example

In this tutorial we will be analysing quantitative traits observed in a mice population. The mouse data consist of phenotypes for traits related to growth and obesity (e.g. body weight, glucose levels in blood), pedigree information, and genetic marker data.
Practicals_mouse_example

Notes

Below is a set of notes for the quantitative genetic theory, statistical models and methods implemented in the qgg package:

Quantitative Genetics Theory

Estimation of Genetic Predisposition

Estimation of Genetic Parameters

Linear Mixed Models

Best Linear Unbiased Prediction Models

REstricted Maximum Likelihood Methods

Gene Set Enrichment Analysis

Bayesian Linear Regression Models

AALBORG UNIVERSITY

GitHub https://pdrohde.github.io/

THE PURPOSE OF TODAY

- **Give an introduction to polygenic scores (PGS)**
- **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)
- **Stratify a population/cohort by their inherent genetic load towards common complex diseases**
  - Polygenic scores (PGS)
- **Identify future projects of common interests**

AGENDA

08:00 – 08:30	Welcome and common introductions
08:30 – 09:10	Session 1: Introduction to Polygenic Scores (PGS)
09:10 – 09:20	Break
09:20 – 10:00	Session 2: Data Sources and Computational Methods
10:00 – 10:10	Break
10:10 – 10:40	Session 3: Evaluating and Interpreting Polygenic Scores
10:40 – 11:00	Break
11:00 – 11:45	Session 4: Advanced Applications and Future Directions
11:45 – 12:30	Lunch and short walk
12:30 – 15:30	Identification of 2-3 projects of common interest
15:30 – 16:00	Next steps and thank you for today