Day 2
Introduction to PRS

Computing PRS, performing QC, cross-trait analyses, visualising results

Jones Gyamfi
Gloria Kirabo

Polygenic Risk Score Analysis, 22 - 27 June 2025,
Makerere University, Uganda

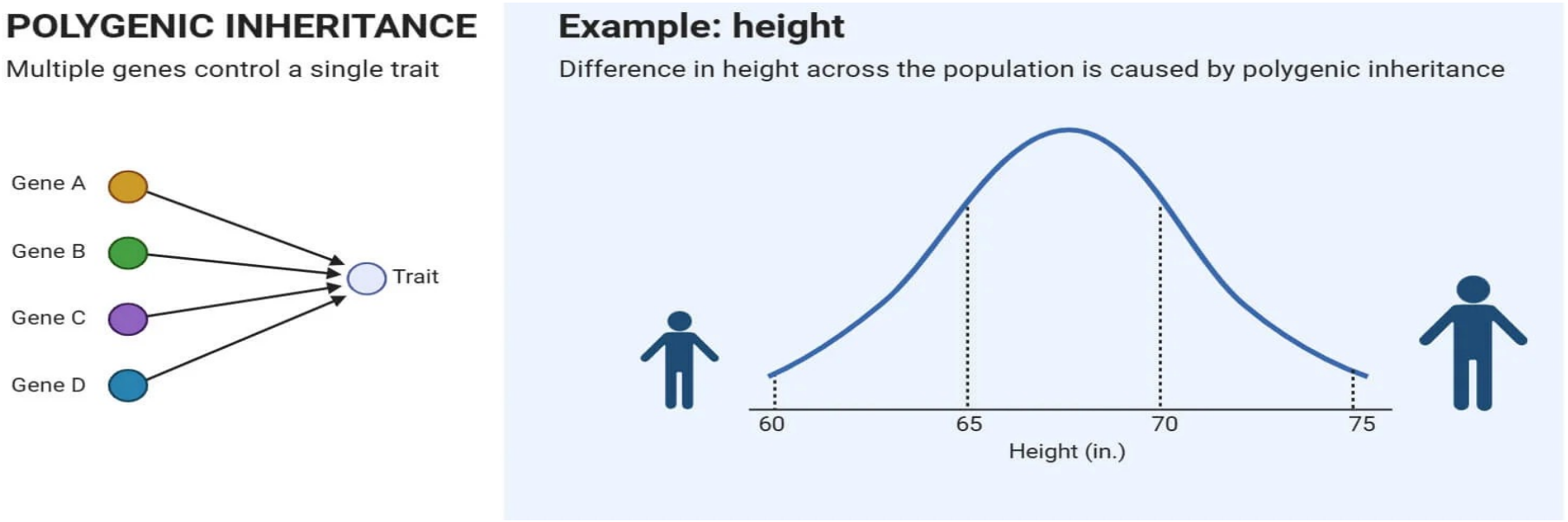
MAKERERE UNIVERSITY
IDI INFECTIOUS DISEASES INSTITUTE
AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES
wellcome connecting science

Learning Objectives

- Define the concept of polygenic risk scores (PRS) and explain how PRS are calculated
- Outline key quality control parameters for a PRS analysis
- Perform basic PRS analyses using PRSice2
- Interpret results generated from PRS analyses
- Customize visualization of results

What is PRS?

- **Monogenic traits vs Polygenic traits**



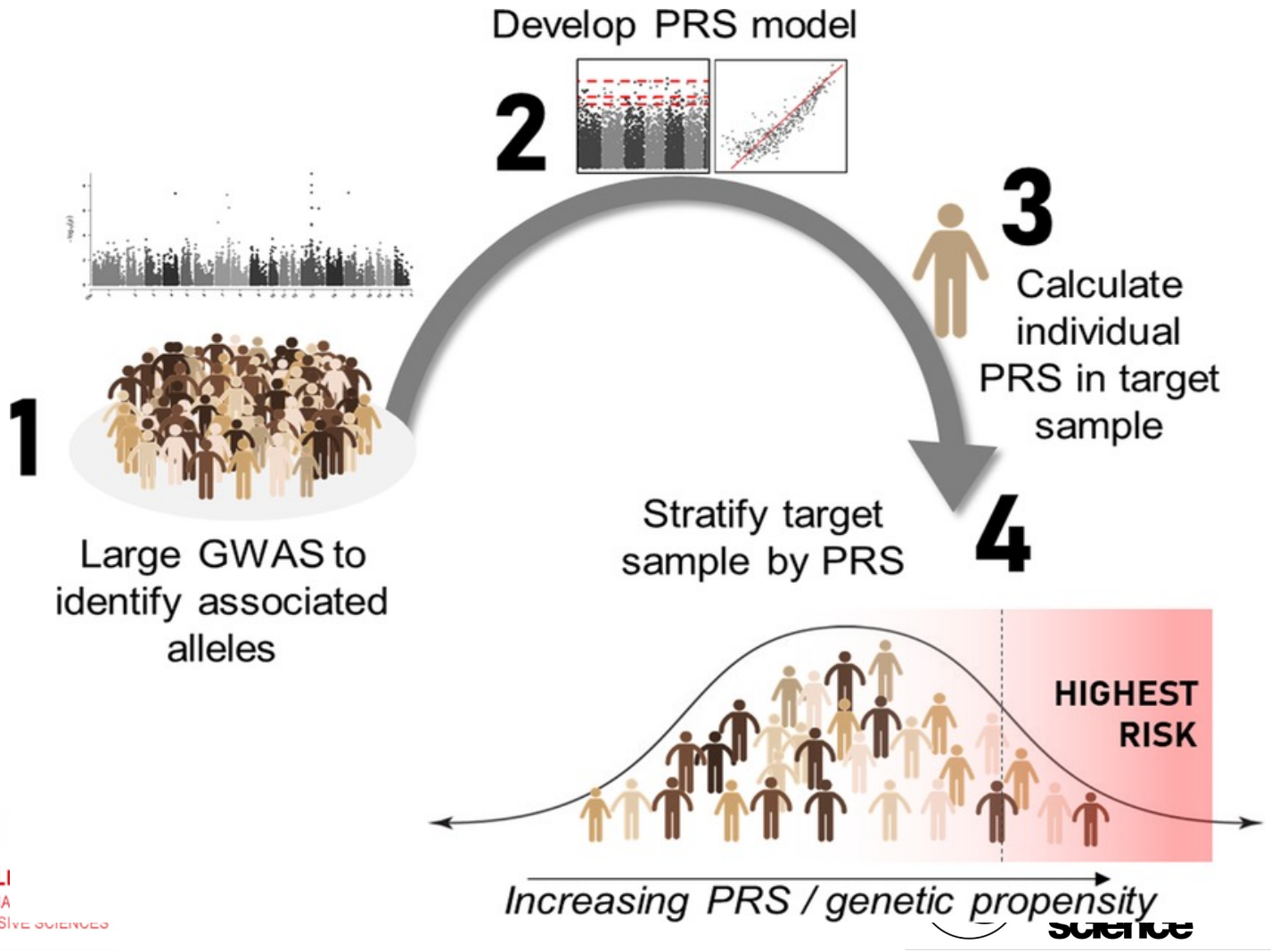
- Complex traits are polygenic!
- There are many genetic variants that contribute to the phenotype.

What is PRS?

- **Polygenic risk scores, or polygenic scores**, as a single value estimate of an individual's common genetic propensity to a phenotype, calculated as a sum of their genome wide genotypes, weighted by corresponding genotype effect size estimates (or Z-scores) derived from summary statistic GWAS data.

PRS

- Polygenic risk scores are calculated by;
  - computing the **sum of risk alleles** corresponding to a phenotype of interest in each individual, **weighted by the effect size estimate** of the most powerful GWAS on the phenotype.

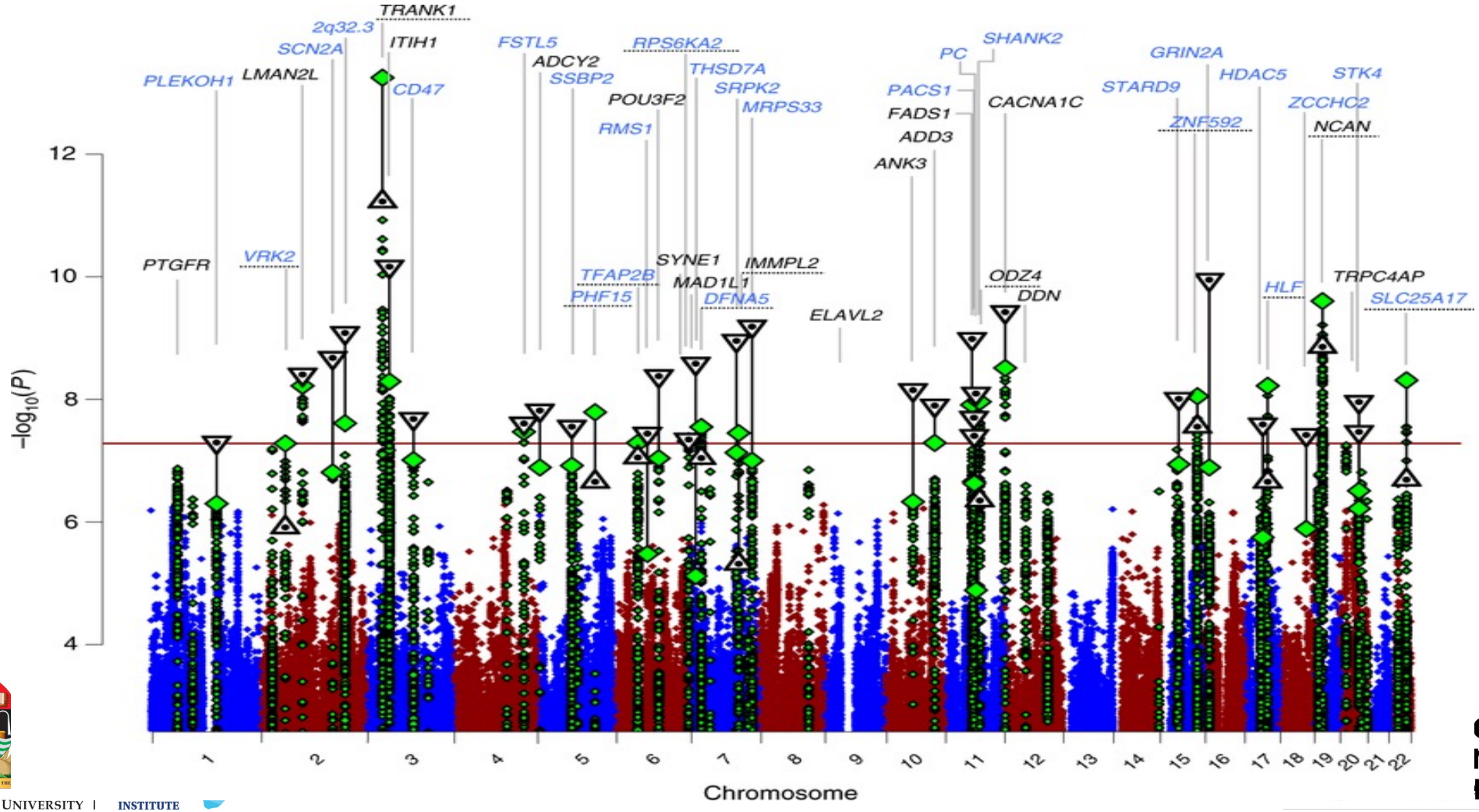


MAKERERE UNIVERSITY

IDI INFECTIOUS DISEASES INSTITUTE

AFRICAN CENTER OF EXCELLI IN BIOINFORMA & DATA INTENSI

Starting point is your GWAS Data



MAKERERE UNIVERSITY | INSTITUTE

come necting nce

PRS

Polygenic Risk Scores

Summarizes the whole genome-wide data into a number based on variation in multiple loci and their associated weights

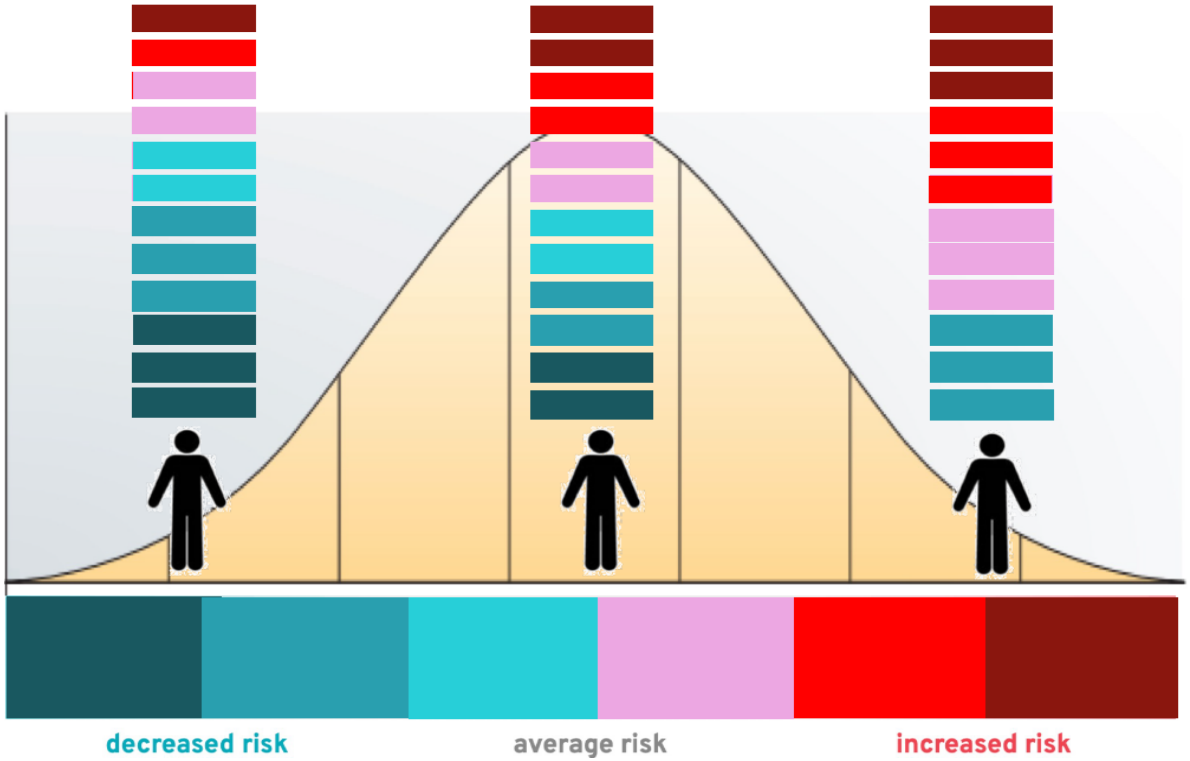
$$PRS_i = \sum_{j=1}^{m} G_{ij} * \beta_j$$

$$PRS = \beta_1 SNP_1 + \beta_2 SNP_2 + \cdots + \beta_n SNP_n$$

Effect size

Number of risk alleles

Number of variants



PRS is the sum of trait-associated risk alleles across many genetic loci, typical weighted by effect sizes from a GWAS

MAKERERE UNIVERSITY

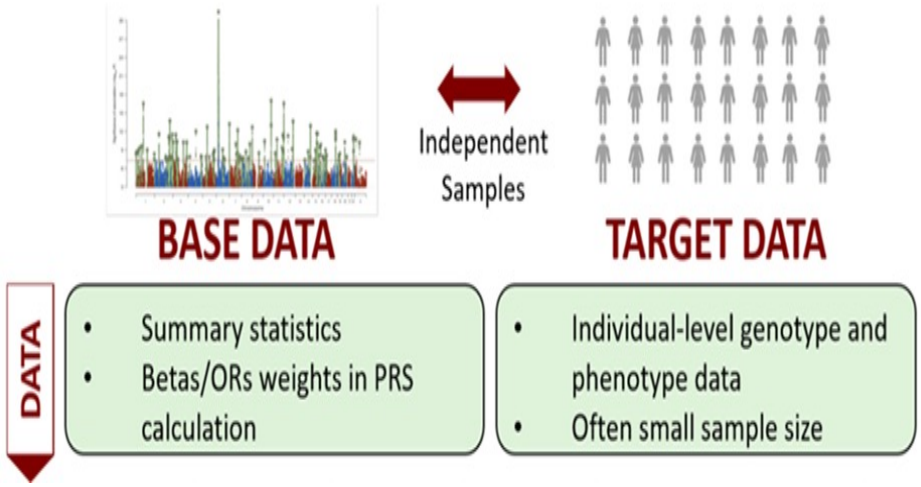
IDI INFECTIOUS DISEASES INSTITUTE

AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES

wellcome connecting science

Performing PRS – Data required

- PRS analyses require two input data sets:
  - **(i) base (GWAS) data:** summary statistics (e.g. betas, P-values) of genotype-phenotype associations at genetic variants (hereafter SNPs) genome-wide, and
  - **(ii) target data:** genotypes and phenotype(s) in individuals of the target sample.



Term	Contains	Purpose	Example
Base	GWAS summary stats	Provides SNP weights for PRS	GIANT summary stats for height
Target	Individual-level genotype + phenotype	Apply PRS, evaluate prediction	Your cohort with SNPs and phenotypes
Discovery	Raw data used to run GWAS	Identify associated SNPs (→ base data)	UK Biobank GWAS for CAD
Validation	New cohort for independent testing	Test generalizability of PRS or GWAS	New dataset from another population

Base Data: GWAS SumStats

- **Summary statistics** are defined as the aggregate p-values and association data for every variant analysed in a genome-wide association study (GWAS).

- The aggregate association data for every SNP analysed in a GWAS

Is there a ReadMe?

- SNP name/position
- Effect allele and alternate allele (A1 and A2)
- Effect allele frequency
- Marginal SNP effect
- Standard error
- P-value
- (Per-SNP) GWAS sample size

```
SNP A1 A2 freq b se p N
rs1001 A G 0.8493 0.0024 0.0055 0.6653 129850
rs1002 C G 0.0306 0.0034 0.0115 0.7659 129799
rs1003 A C 0.5128 0.0045 0.0038 0.2319 129830
```

Chr	SNP	bp	A1	A2	Freq	b	se	p	K
1	rs2286139	761732	C	T	0.1379	-0.0104056	0.00732416	0.155397	0.5
1	rs12562034	768448	A	G	0.10475	-0.00627592	0.00827054	0.447955	0.5
1	rs4970383	838555	A	C	0.247975	0.00946201	0.00587444	0.107243	0.5
1	rs1806509	853954	C	A	0.3912	0.0152744	0.00523012	0.00349507	0.5
1	rs13302982	861808	A	G	0.018025	-0.0180122	0.0189517	0.341895	0.5
1	rs28576697	870645	C	T	0.29355	0.0116486	0.00556379	0.0362916	0.5
1	rs2340582	882803	A	G	0.05465	0.0119371	0.0111055	0.282426	0.5
1	rs3748594	886384	A	G	0.025975	-0.01244	0.0158797	0.433401	0.5
1	rs28504611	908414	T	C	0.022225	0.00388796	0.0171623	0.820781	0.5
1	rs9777939	929190	A	G	0.03345	-0.00446279	0.0141522	0.752502	0.5
1	rs1891910	932457	A	G	0.2295	-0.00647527	0.00605864	0.285175	0.5
1	rs35940137	940203	A	G	0.050575	-0.10689	0.0115935	2.97533e-20	0.5
1	rs6657048	957640	T	C	0.011825	0.0892934	0.0233322	0.000129691	0.5
1	rs9803031	987200	C	T	0.083875	-0.00434284	0.0091306	0.634334	0.5

MAKERERE UNIVERSITY

IDI INFECTIOUS DISEASES INSTITUTE

The Target Data

1	1:729679	0	729679	C	G
1	rs58276399	0	731718	C	T
1	rs141242758	0	734349	C	T
1	rs79010578	0	736289	A	T
1	rs139221807	0	746189	G	A
1	1:752566	0	752566	G	A
1	1:752721	0	752721	A	G
1	1:753405	0	753405	C	A
1	rs2073813	0	753541	A	G
1	1:754182	0	754182	A	G

TAR.bim Columns (no header):

Chr	SNP	CM	BP	A1	A2
Chromosome	SNP ID	Genetic distance	Base pair position	Allele 1	Allele 2

1. **Phenotype File** (TAR.cad or TAR.height): This file contains the phenotype or trait you want to predict using PRS.

2. **Covariate File** (TAR.covariate): Contains covariates for adjusting PRS models (e.g., age, sex, principal components).

3. **Genotype Files** (PLINK binary format) Files: TAR.bed, TAR.bim, TAR.fam.bed: binary genotype data (no readable headings).bim: variant annotation

MAKERERE UNIVERSITY

IDI INFECTIOUS DISEASES INSTITUTE

AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES

wellcome connecting science

The GWAS SumStats

- GWAS on quantitative trait, the effect size is typically given as a **beta coefficient (β)** from a **linear regression** with SNP genotypes as predictor of phenotype.
    - The β coefficient estimates the increase in the phenotype for each copy of the effect allele.
    - For example, if the effect allele of a SNP is G and the non-effect allele is A , then the genotypes AA, AG and GG will be coded as 0, 1 and 2 respectively.
    - In this scenario, the β **coefficient reflects how much the phenotype changes for each G allele present**

The GWAS SumStats

- If GWAS is performed on a binary trait (e.g. case-control study), the effect size is usually reported as an **Odd Ratios (OR)** and generated from a logistic regression.
    - *Using the same example, if the OR from the GWAS is 2 with respect to the G allele, then the OR of AG relative to AA is 2, and the OR of GG relative to AA is 4.*
    - An individual with the GG genotype are estimated* to be 4 times more likely to be a case than someone with the AA genotype (*an Odds Ratio is itself an estimate of a Risk Ratio, which cannot be calculated from a case/control study)

The GWAS SumStats

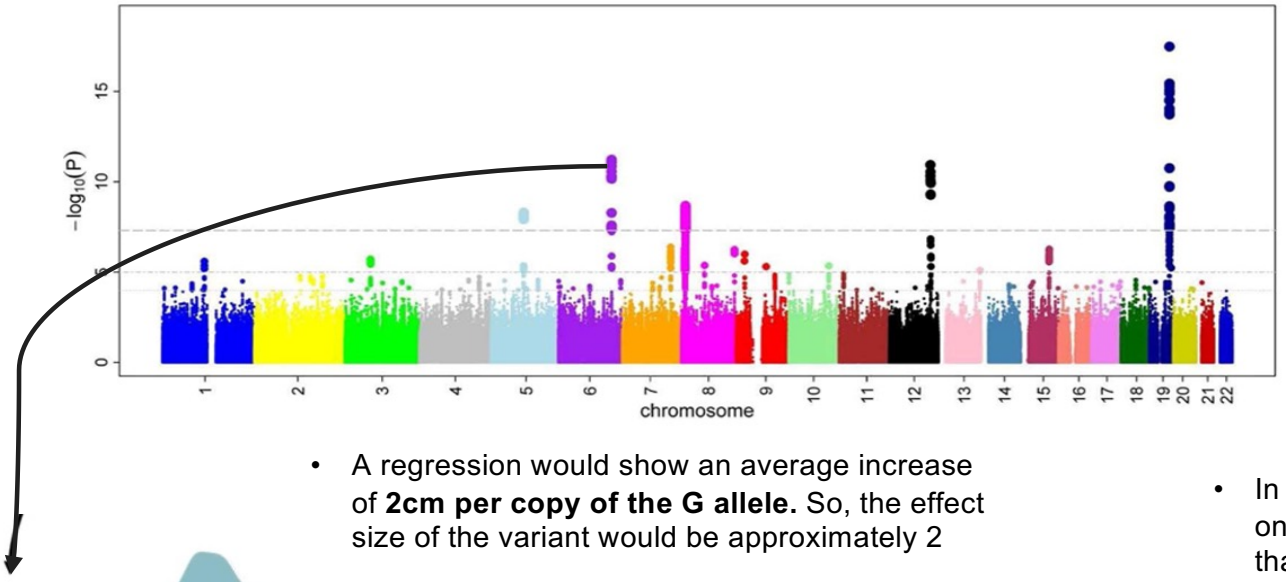
- The relationship between the β coefficient and the OR is: OR = e^{β} and $\log_e(OR) = \beta$.

- While GWAS usually convert from the β to the OR when reporting results, most PRS software convert OR back to β's($log_e(OR)$) to allow simple addition.

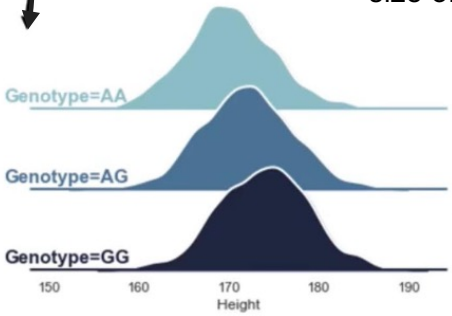
$$PRS = \beta_1 SNP_1 + \beta_2 SNP_2 + \cdots + \beta_n SNP_n$$

Effect size — Number of risk alleles — Number of variants

Revision Practical Example



- A regression would show an average increase of **2cm per copy of the G allele.** So, the effect size of the variant would be approximately 2



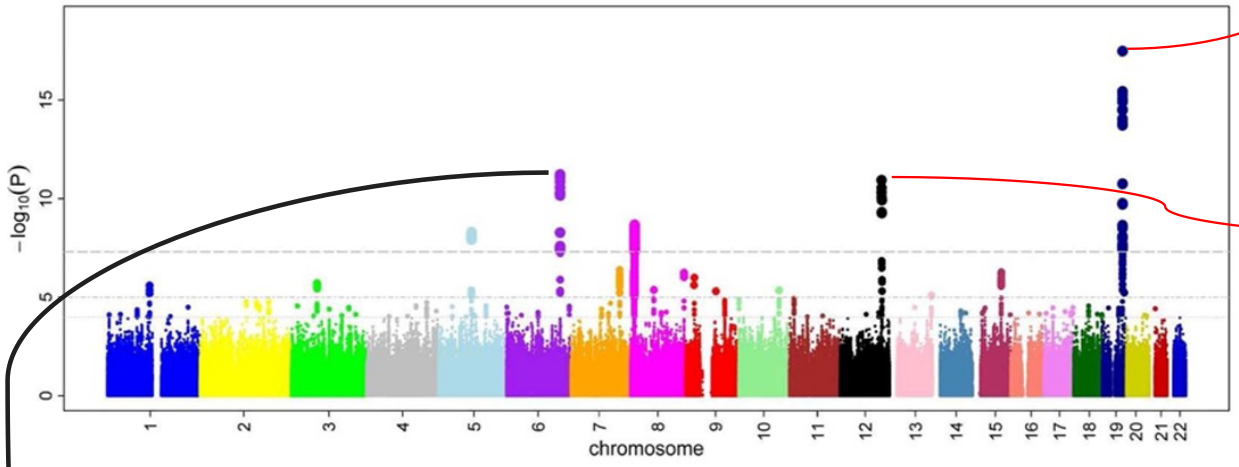
- In a new sample we would expect AG individuals to be on average 2 cm taller than the AA and 2cm shorter than GG

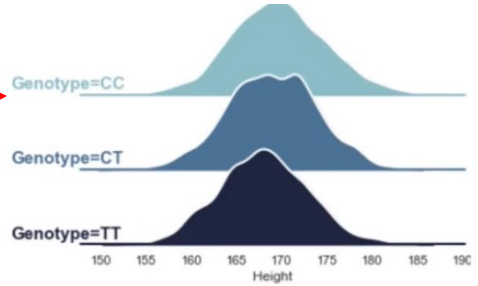
MAKERERE UNIVERSITY

IDI INFECTIOUS DISEASES INSTITUTE

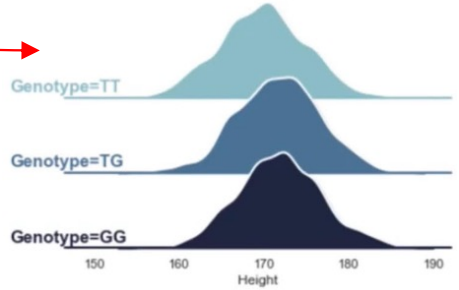
wellcome connecting science

Practical example



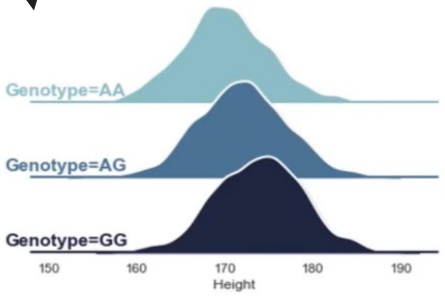


- Effect size of -1cm per T allele

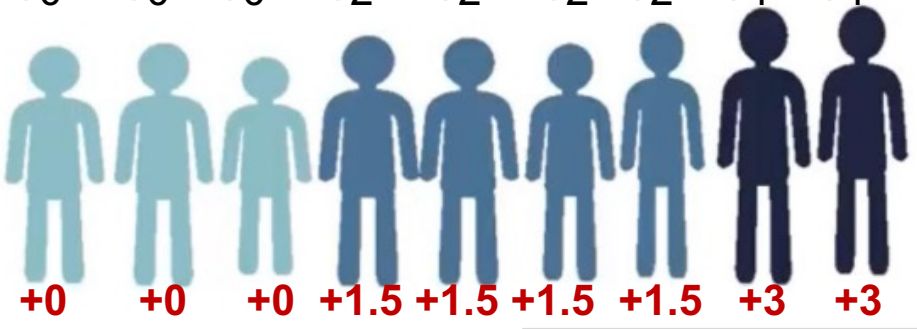


- Effect size of +0.5cm per G allele

- Effect size of 2cm per G allele



+0	+0	+0	+0.5	+0.5	+0.5	+0.5	+1	+1
+0	+0	+0	-1	-1	-1	-1	-2	-2
				2*1				2*2
								(Copies of G)
+0	+0	+0	+2	+2	+2	+2	+4	+4



+0	+0	+0	+1.5	+1.5	+1.5	+1.5	+3	+3

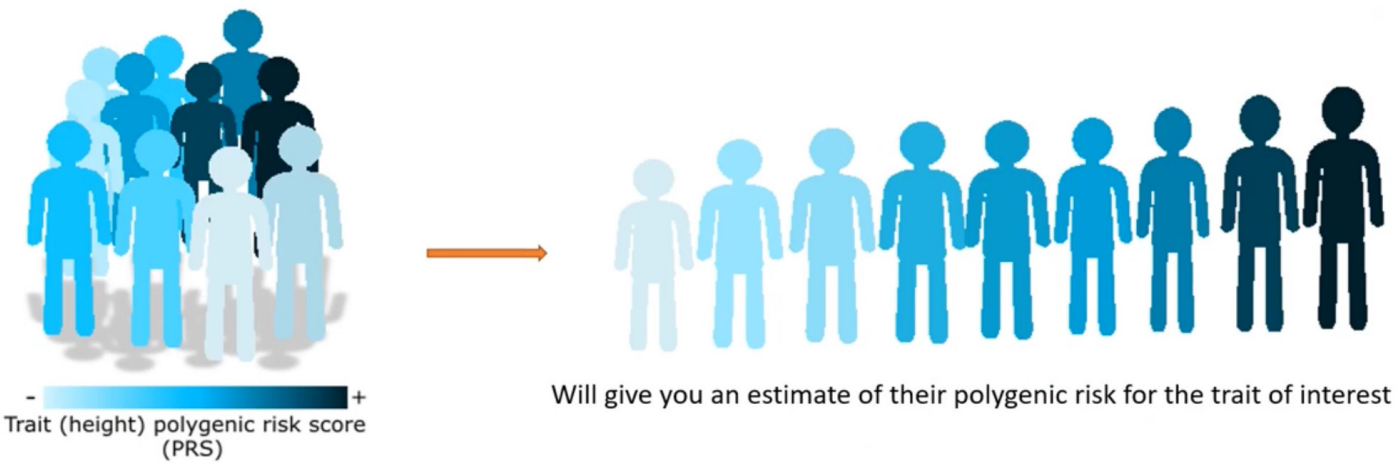
MAKERERE UNIVERSITY

IDI INFECTIOUS DISEASES INSTITUTE

AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES

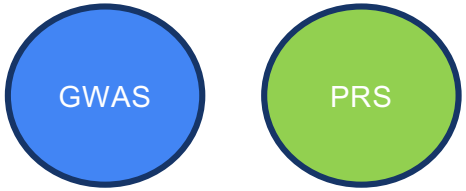
Practical Example

Repeat including the other variants and sum across all loci



Polygenic risk score – Weighted sum of alleles which quantify the effect of several genetic variants on an individual's phenotype.

Caution! The sample for which the PRS will be calculated should be independent from that of the discovery GWAS. Sample overlap will bias your results



MAKERERE UNIVERSITY

IDI INFECTIOUS DISEASES INSTITUTE

AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES

wellcome connecting science

Where to download GWAS SumStats?

Genome-wide association studies

Emil Uffelmann[1], Qin Qin Huang[2], Nchangwi Syntia Munung[3], Jantina de Vries[3], Yukinori Okada[4,5], Alicia R. Martin[6,7,8], Hilary C. Martin[2], Tuuli Lappalainen[9,10,12] and Danielle Posthuma[1,11] ✉

Database	Content
GWAS Catalog https://www.ebi.ac.uk/gwas/	GWAS summary statistics and GWAS lead SNPs reported in GWAS papers
GeneAtlas http://geneatlas.roslin.ed.ac.uk/	UK Biobank GWAS summary statistics
Pan UKBB https://pan.ukbb.broadinstitute.org/	UK Biobank GWAS summary statistics
GWAS Atlas https://atlas.ctglab.nl/	Collection of publicly available GWAS summary statistics with follow-up in silico analysis
FinnGen results https://www.finngen.fi/en/access_results	GWAS summary statistics released from FinnGen, a project that collected biological samples from many sources in Finland
dbGAP https://www.ncbi.nlm.nih.gov/gap/	Public depository of National Institutes of Health-funded genomics data including GWAS summary statistics
OpenGWAS database https://gwas.mrcieu.ac.uk/	GWAS summary data sets
Pheweb.jp https://pheweb.jp/	GWAS summary statistics of Biobank Japan and cross-population meta-analyses

Where to download Genotype data

1. **1000 Genomes Project:** Whole-genome sequencing data for ~2,500 individuals from diverse populations. Format: VCF; can be converted to PLINK format. Link: https://www.internationalgenome.org/data
2. **UK Biobank (application required):** Genotype and phenotype data for ~500,000 UK individuals. Access: Requires institutional affiliation and application. Link: https://www.ukbiobank.ac.uk/enable-your-research
3. **dbGaP (Database of Genotypes and Phenotypes):** Genotype and phenotype data from NIH-funded studies. Disease-specific PRS analysis. Access: Controlled-access; requires application and data use agreement. Link: https://www.ncbi.nlm.nih.gov/gap/
4. **OpenGWAS and IEU OpenGWAS Project**: Primarily GWAS summary statistics, but links to relevant genotype datasets are sometimes provided. Link: https://gwas.mrcieu.ac.uk/
5. **Simulated Genotype Data for Teaching and Practice:** GitHub: https://github.com/choishingwan/PRSice

1. **Data Conversion Tools:** If the data isn't in PLINK format: Convert VCF to PLINK:
   plink --vcf input.vcf --make-bed --out output
   Impute or filter using tools like Michigan Imputation Server

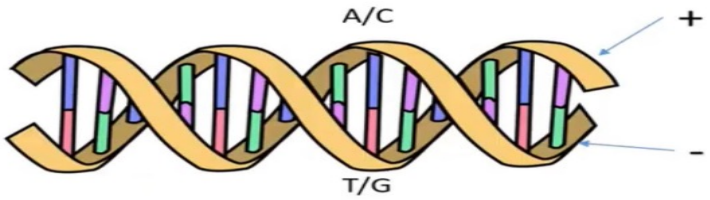
Applications of PRS

- Identifying shared aetiology among traits,
- Test for genome-wide gene*environment and gene*gene interactions
- Perform Mendelian Randomisation studies to infer causal relationships,
- For patient stratification and sub-phenotyping
- Predicting clinical risk for screening/prevention/intervention
- Predicting genetic propensity to actionable traits
- Improving case/control classification (e.g., For clinical trials)
- Pathway analyses: providing individual pathway PRS

Performing PRS – QCs

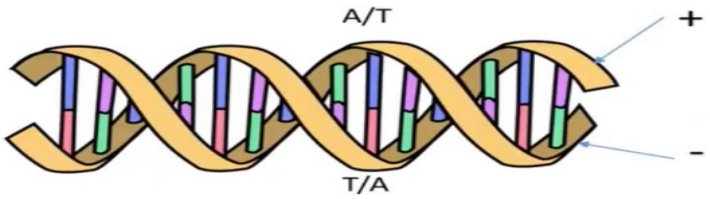
- For SNPs that have complementary alleles, e.g. A/T, G/C, we cannot be certain that the alleles referred to in the target data correspond to those of the base data or whether they are the 'other way around' due to being on the other DNA strand (unless the same genotyping chip was used for all data).
  - These SNPs are known as **ambiguous SNPs**, we remove ambiguous SNPs in PRSice to avoid the possibility of introducing unknown bias.

Note on ambiguous variants



	REF	ALT
rsxxy	A	C
rsxxy	T	G

The variant is not ambiguous
(we can easily tell the strand)



	REF	ALT
rsxxx	A	T
rsxxx	T	A

The variant is ambiguous
(hard to tell the strand)

MAKERERE UNIVERSITY | IDI INFECTIOUS DISEASES INSTITUTE | IN BIOINFORMATICS & DATA INTENSIVE SCIENCES

wellcome connecting science

Performing PRS – QCs

- **Mismatching genotypes [Strand mismatch] —**

Imagine the SNP **rs123456**:

Dataset	Effect Allele	Other Allele	SNP Genotype
Base (GWAS)	A	G	A = increases disease risk
Target	T	C	✗ mismatch

We will expect the target to be A/G

- For example, while the effect allele of a SNP is T in the base data, the effect allele in the target might be G instead.
- When this occurs, **allele flipping** should be performed, where the genotype encoding in the target data is reversed so that TT, TG and GG are coded as 2, 1 and 0.
- Again, this is usually performed automatically by PRS software.

Performing PRS – Quality Control

- Heritability is a measure of how much of the variation in a trait within a population is due to genetic differences, not how much of the trait itself is genetic.

- "**Chip heritability**" (SNP heritability) refers to the proportion of phenotypic variance in a trait that can be explained by the genetic variants included on a genotyping chip (SNPs). Chip heritability focuses on the genetic variants that are specifically included on a genotyping chip (like a DNA microarray). It helps us understand the contribution of common genetic variants (SNPs) to a trait's variation.

- **Heritability check**— Performing PRS analyses that use GWAS data with a chip-heritability estimate $h\ snp^{2} > 0.05$. Avoid PRS on GWAS with $h^2 < 0.05$

Performing PRS – Quality Control

- **Effect allele**—Some GWAS results files do not make clear which allele is the effect allele and which is the non-effect allele. Hence the identity of the effect allele from the base GWAS data must be obtained.

- **File transfer** — Important to ensure that files have not been corrupted during transfer (e.g. using md5sum). PRS calculation errors are often due to corrupt files.

- **Genome Build** — Ensure that the base and target data SNPs have genomic positions assigned on the same genome build. **LiftOver** is an excellent tool for standardizing genome build across different data sets.

- **Duplicate SNPs** — Ensure that there are no duplicated SNPs in either the base or target data. Remove, or software may crash

Performing PRS – Quality Control

- **Sex chromosomes** —remove individuals for which there is a difference between reported sex and that indicated by the sex chromosomes. A sex check can be performed in PLINK. And remove X/Y SNPs to avoid confounding-by-sex

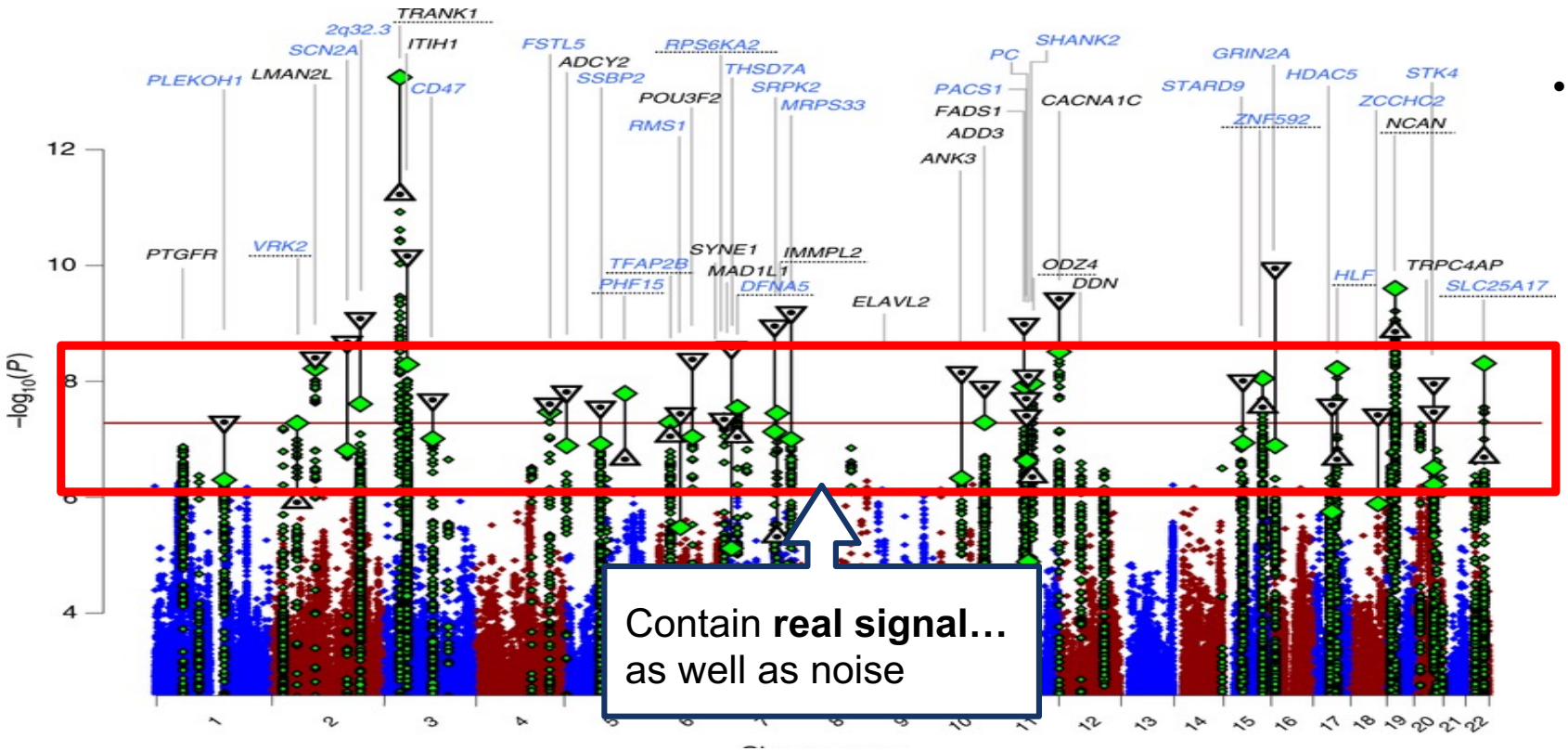
- **Sample overlap** — Sample overlap between the base and target data can result in substantial inflation of the association between the PRS and the trait tested in the target data and so must be eliminated.

- **Relatedness** — A high degree of relatedness among individuals between the base and target data can also generate inflation of the association between the PRS and target phenotype.

Performing PRS – Calculation of Polygenic Risk Scores

- The key factors in the development of methods for calculating PRS are:
    - (i) the potential adjustment of GWAS estimated effect sizes via e.g. shrinkage and incorporation of their uncertainty,
    - (ii) the task of dealing with Linkage Disequilibrium .

Which SNPs do we include in our PRS?



Contain **real signal…** as well as noise

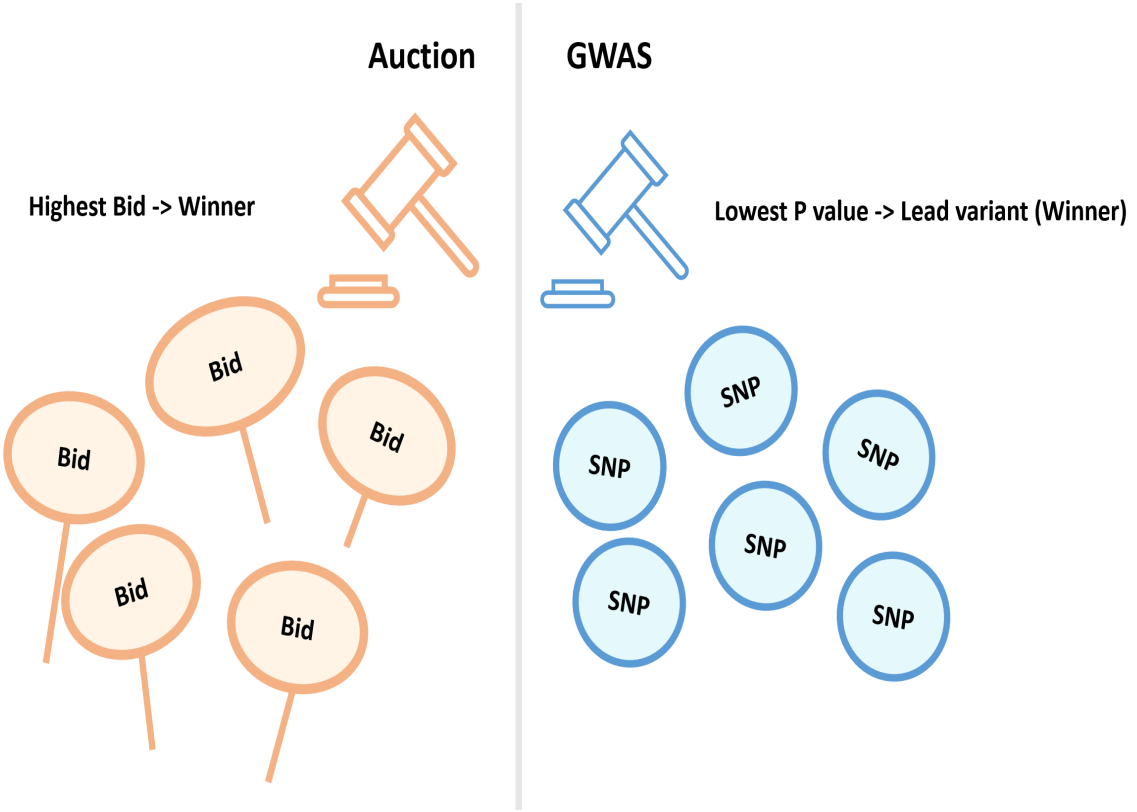
- **Key concepts for PRS:** Including non-significant SNPs may increase predictive power

- The use of unadjusted effect size estimates of all SNPs could generate poorly estimated PRS with high standard error.
- Two broad shrinkage strategies have been adopted:
- (i) shrinkage of the effect estimates of all SNPs via standard or tailored statistical techniques, and
- (ii) use of P-value selection thresholds as inclusion criteria for SNPs into the score.

Performing PRS – Shrinkage of GWAS effect size

- **Winners Curse:** Winner's curse refers to the phenomenon that genetic effects are systematically overestimated by thresholding or selection process in genetic association studies.

- In GWAS, the winners (most significant SNPs), wins because its effect size estimate is inflated



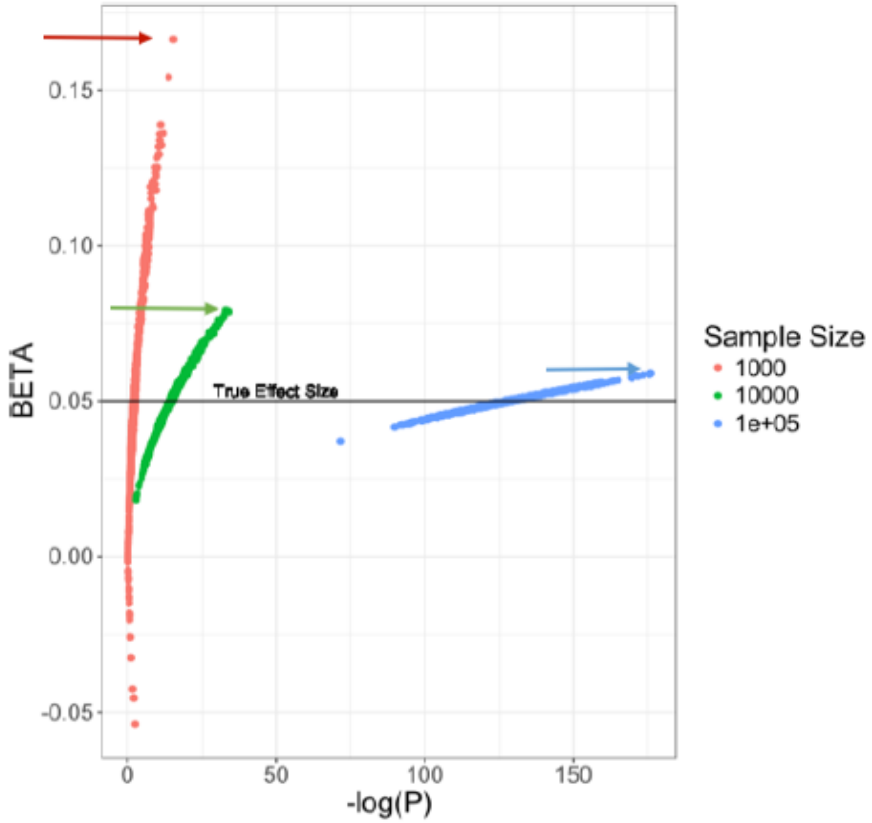
MAKERERE UNIVERSITY

IDI INFECTIOUS DISEASES INSTITUTE

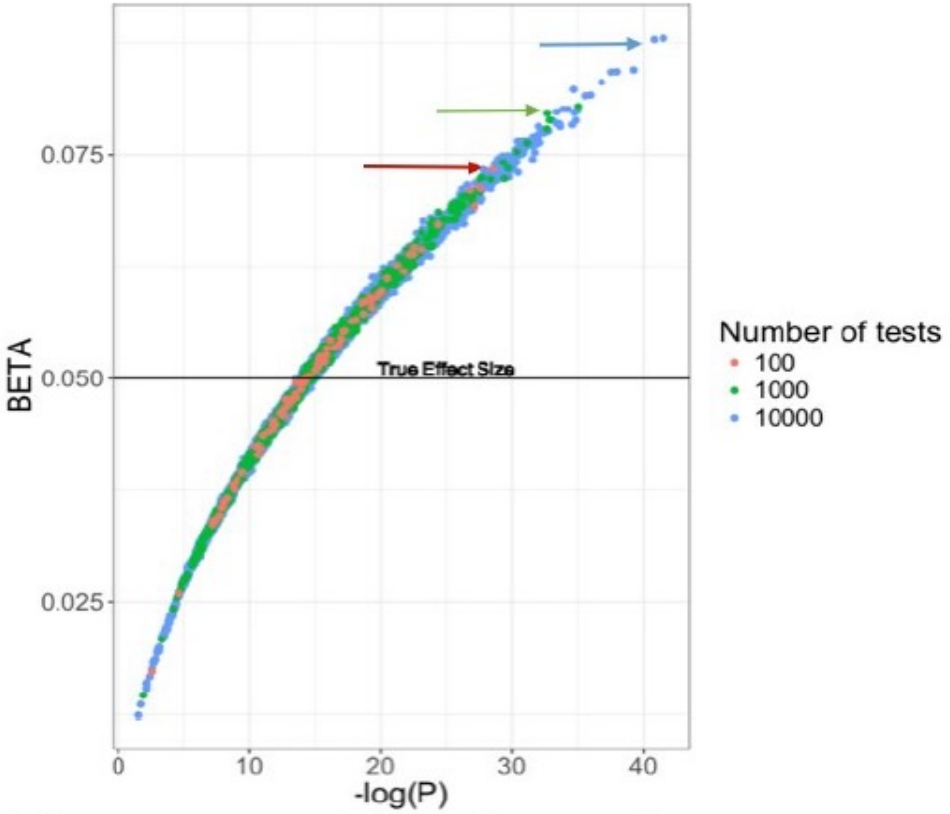
AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES

wellcome connecting science

Performing PRS – Winners Curse



The smaller sample size, the more inflated the effect size estimates



The more tests performed, the more inflated the top results are

MAKERERE UNIVERSITY
INFECTIOUS DISEASES INSTITUTE
IN BIOINFORMATICS & DATA INTENSIVE SCIENCES
science

Performing PRS – Shrinkage of Effect size

- Shrinkage of all SNPs: LDpred / Lassosum employ Bayesian and penalized regression (respectively) to include all SNPs but to shrink all their effects.
  - Some force most effect estimates to zero or close to zero, some mostly shrink small effects, while others shrink the largest effects most.

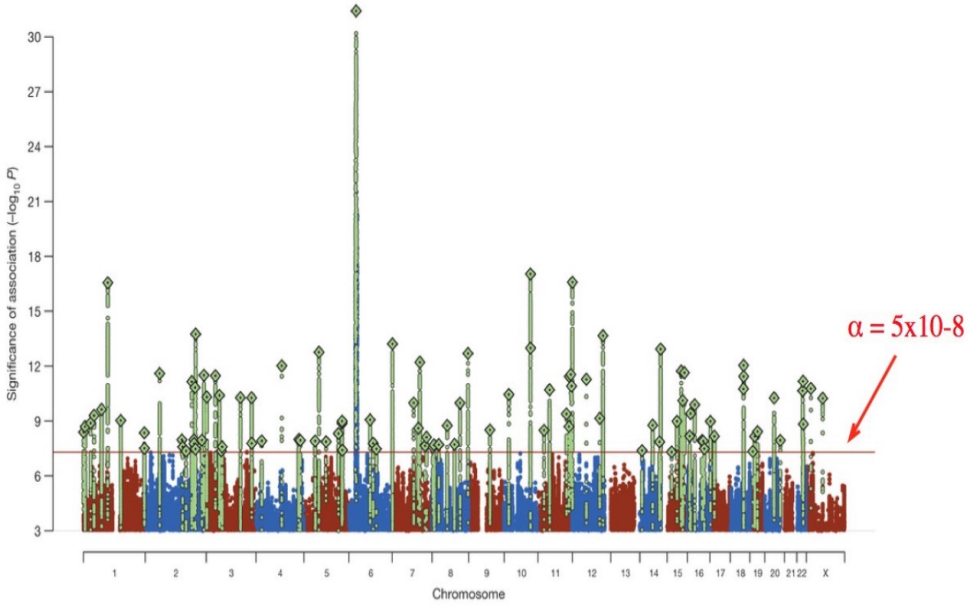
Performing PRS – Thresholding

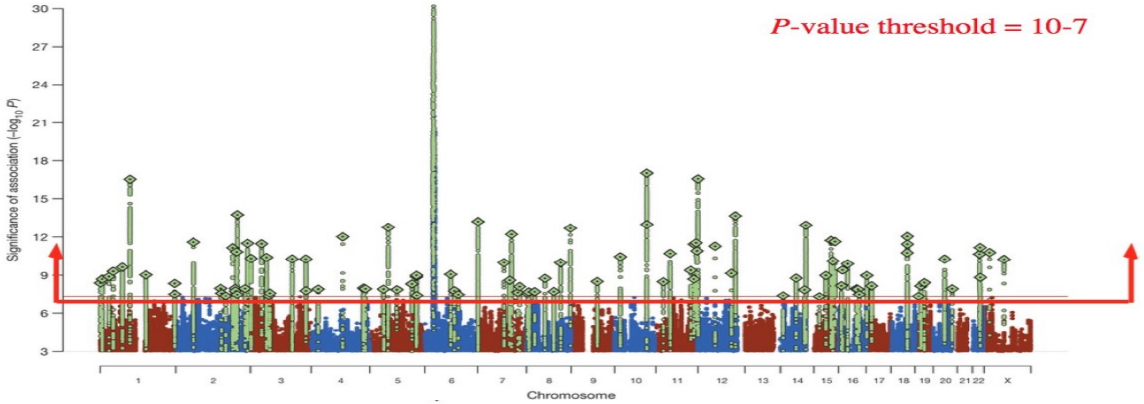
- **P-value selection threshold approach**
- only those SNPs with a GWAS association p-value below a certain threshold (e.g. $p < 1x10^{-5}$) are included in the calculation of the PRS, while all other SNPs are excluded.

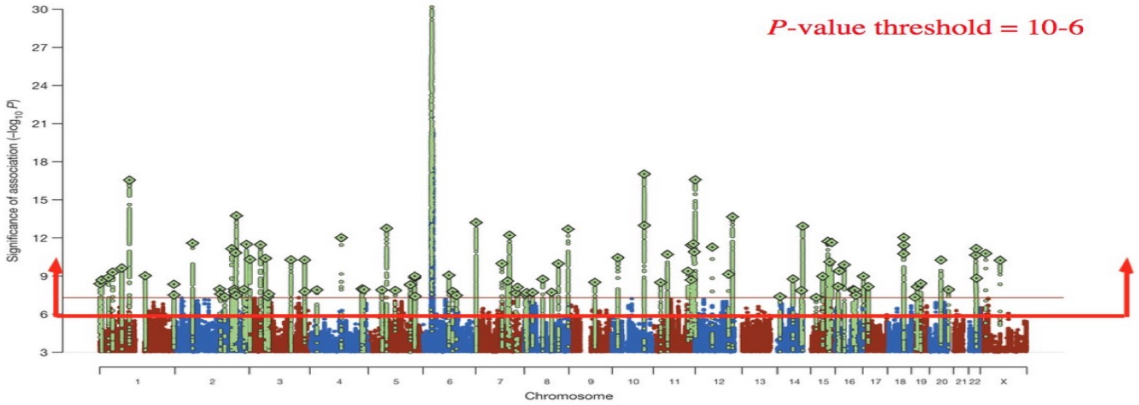
- The **'optimal threshold'** compute PRS from only a subset of the SNPs expected to be more predictive of the target trait.

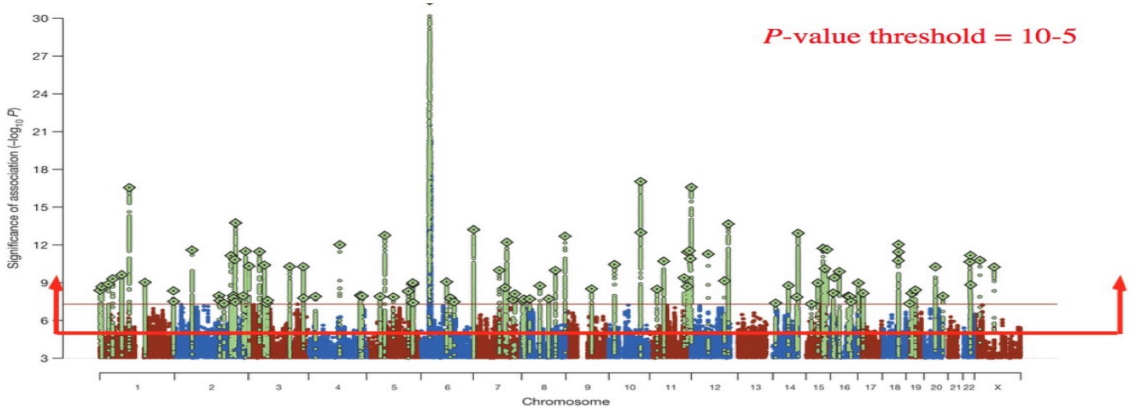
- The standard PRS approach retains only a subset of the more significant SNPs (PLINK, PRSice, PRSice2)
  → all unselected SNPs effectively shrunk to 0

Performing PRS – Thresholding









MAKERERE UNIVERSITY

IDI INFECTIOUS DISEASES INSTITUTE

AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES

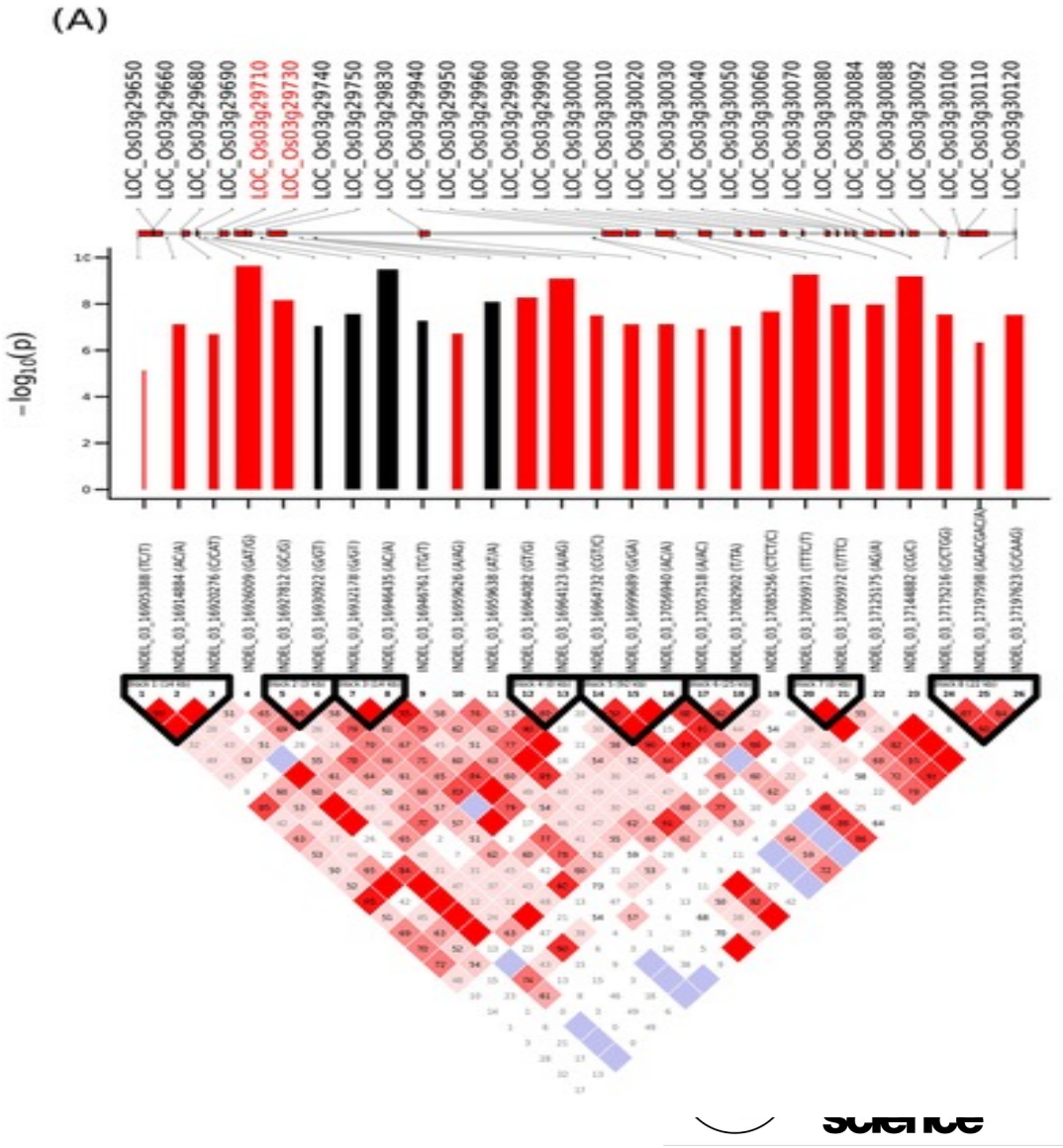
Controlling for Linkage Disequilibrium

- Linkage disequilibrium (LD), or simply LD, refers to the non-random association of alleles at different gene loci in a population.

- There are two main options for approximating the PRS that would have been generated from full conditional GWAS:
  - 1. SNPs are clumped so that the retained SNPs are largely independent of each other, allowing their effects to be summed, assuming additive effects,
  - 2. all SNPs are included and the LD between them is accounted for.

Clumping

- The standard approach ‘controls’ for LD via **clumping**
- Clumping thins down the SNP set - retaining most associated SNPs - so that those remaining are ~independent
- In the ‘standard approach’ to PRS calculation, option (1) is combined with P-value thresholding and called the **C+T (clumping + thresholding)**



MAKERERE UNIVERSITY

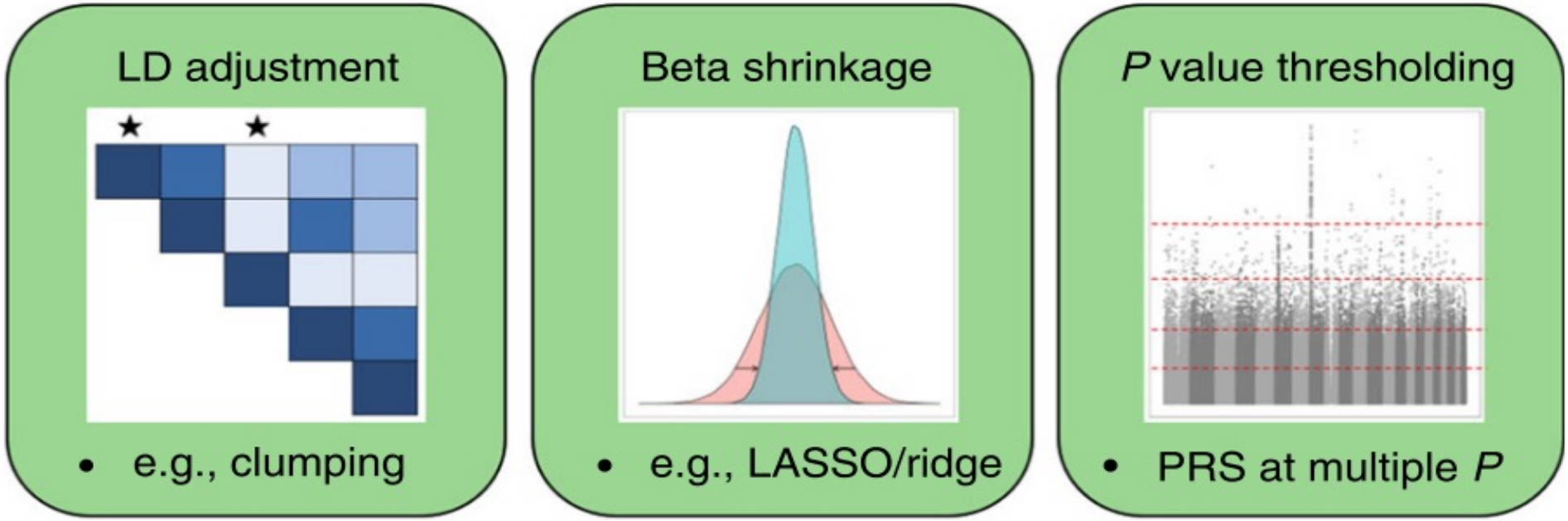
IDI INFECTIOUS DISEASES INSTITUTE

AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES

WHY HAS THE C+T REMAINED POPULAR?

- Clumping can detect several signals in the same region
- Infeasible to search entire space of predictors (thresholding a fast variable selection method)
- C+T approach highly interpretable
- Much faster than other methods (LDpred, Lassosum) but with similar performance

Shrinkage of GWAS effect size



What units should PRS be measured in?

- The units of the GWAS effect sizes determine the units of the PRS; e.g. if calculating a height PRS using effect sizes from a height GWAS that are reported in centimetres (cm), then the resulting PRS will also be in units of cm.

- PRS on a binary (case/control) phenotype, the effect sizes used as weights are typically reported as log Odds Ratios (log(ORs)).

PRS – Portability Problem?

MAKERERE UNIVERSITY

IDI INFECTIOUS DISEASES INSTITUTE

AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES

wellcome connecting science

PRS – Portability Problem

- Likely some combination of differences in LD, genetic drift, natural selection, demographic histories and the environment
- Probably mostly ‘statistical’ - exacerbated by aggregation across genome wide SNPs (inc. null SNPs)
- This can cause a deflation in PRS power when applied across populations
- However, PRS can also be inflated by ‘local’ effects..

Results and Interpretation

- Once PRS have been calculated, a regression is then performed in the target sample, with the PRS as a predictor of the target trait, and covariates included as appropriate.

- A typical PRS study involves testing evidence for an association between a PRS and a trait(s) in the target data and evaluating its potential effect.

Results and Interpretation

- The association between PRS and outcome can be measured with:
  - Standard association or goodness-of-fit metrics, such as the **P-value** to test a null hypothesis of no association,
  - **Phenotypic variance explained (R^2),**
  - Effect size estimate (beta or OR) per unit of PRS or between specific strata e.g. high vs low risk individuals, and
  - Measures of discrimination in disease prediction, such as area under the curve (AUC).

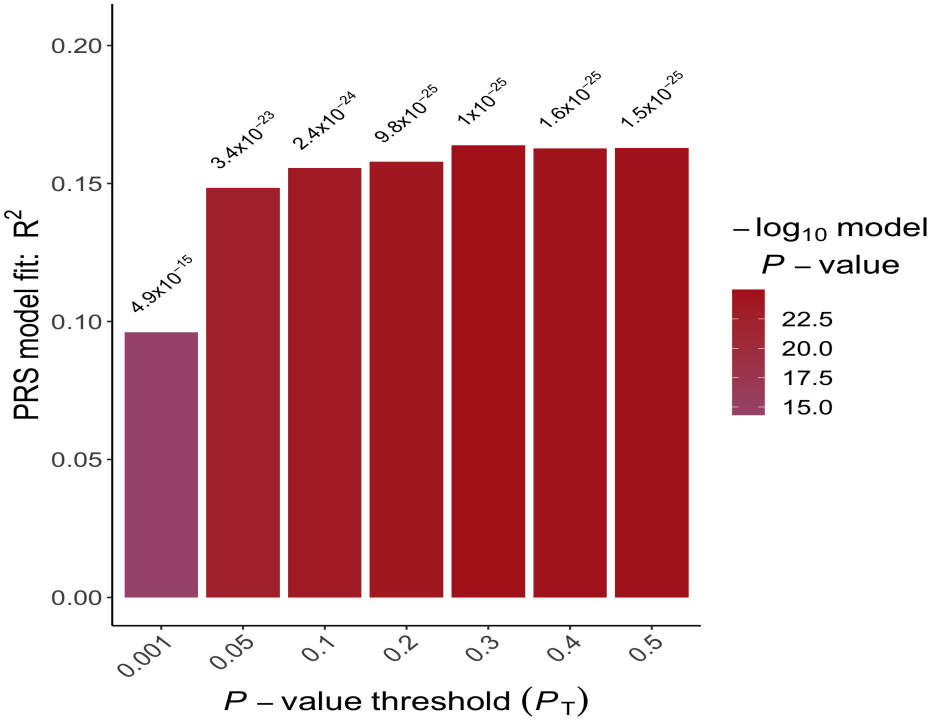
Results and Interpretation

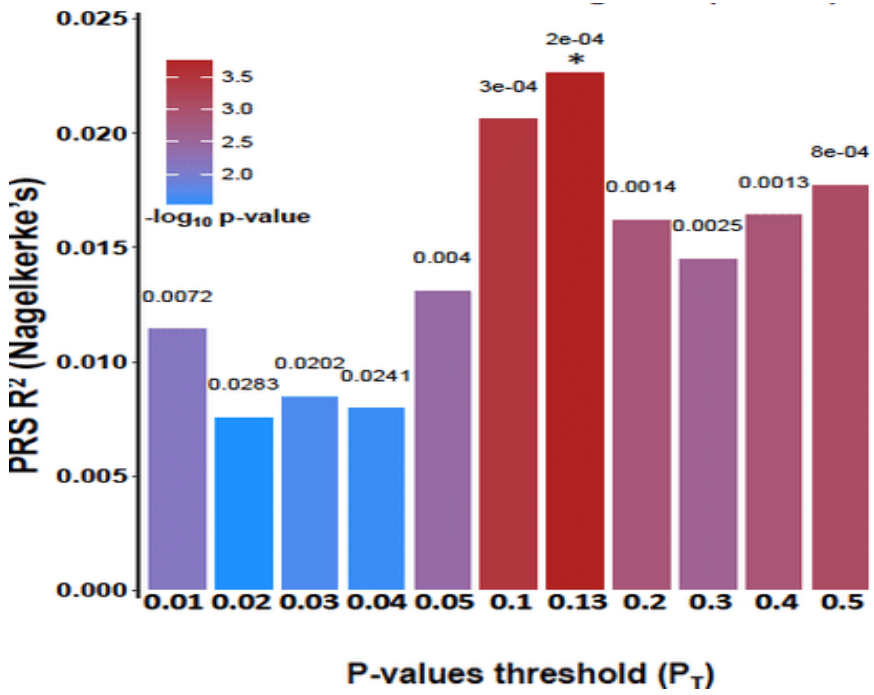
1. **Construct PRS**: Sum of weighted alleles across many SNPs, where weights come from GWAS effect sizes.
2. **Prepare phenotype and covariates** for individuals in your target dataset.
3. **Run regression**:
    - **Logistic regression** if the outcome is **binary** (e.g., case/control)
    - **Linear regression** if the outcome is **quantitative** (e.g., height, blood pressure)
4. **Interpret the association**:
    - **p-value**: Is the PRS significantly associated with the trait?
    - **β coefficient or OR**: What is the effect size?
    - **R^2 or AUC**: How well does the PRS explain the trait?

Results and Interpretation

- Variance explained (R^2) for continuous trait outcomes, but approximate measures ("pseudo-R^2") for case/control outcomes.
- The Nagelkerke R^2 is the most popular approach (but still produce biases).
- R^2 on the liability scale estimates the proportion of variance explained by the PRS of a hypothetical normally distributed latent variable that underlies and causes case/control status.
- Lee R^2 is an alternate pseudo-R^2 metric that accounts for case/control ratio R2.

PLOTTING RESULT: BAR PLOT





- Most predictive threshold included, flexible choice of bars in PRSice

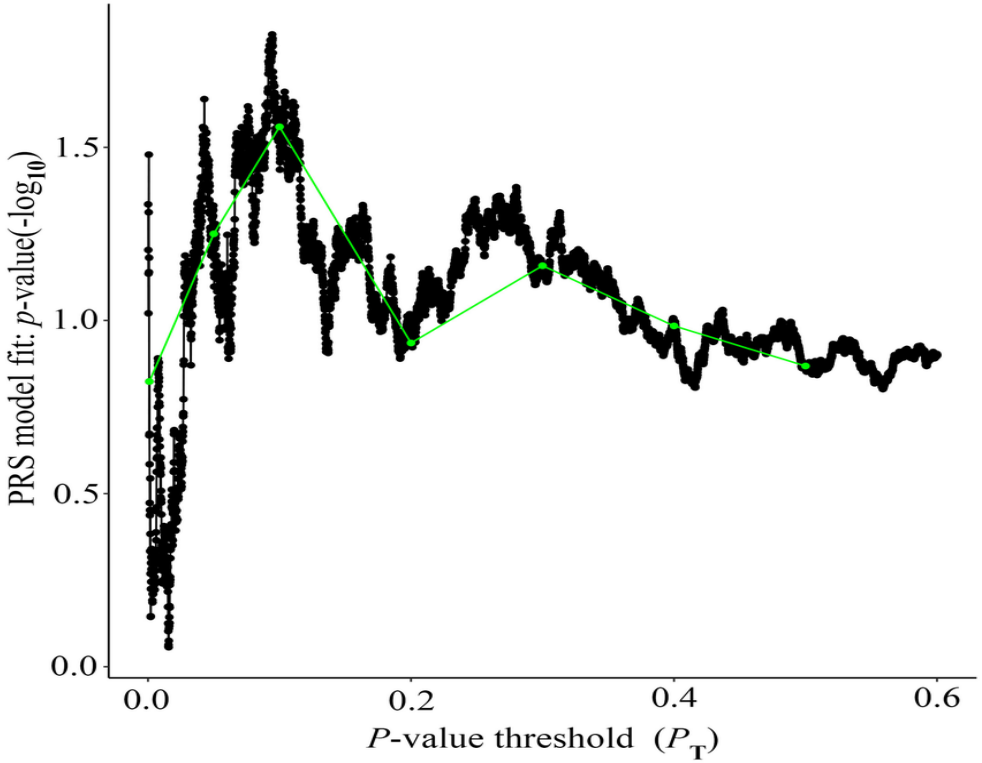
MAKERERE UNIVERSITY

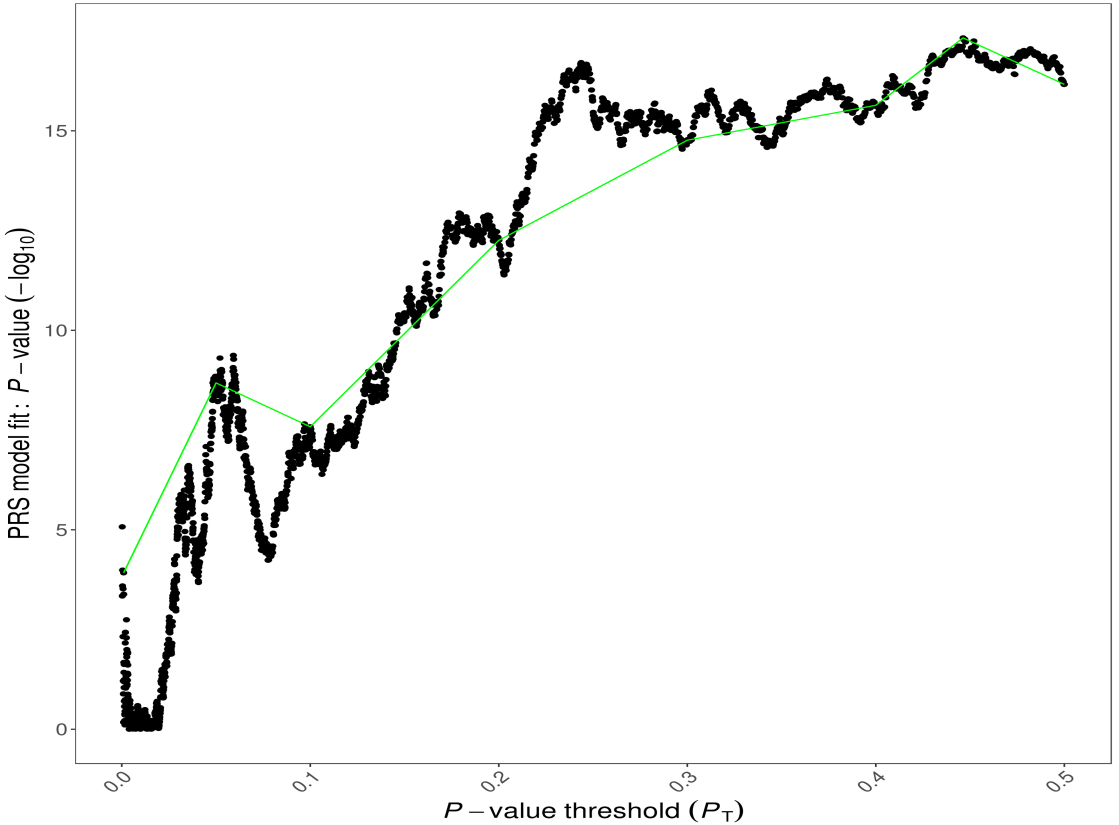
IDI INFECTIOUS DISEASES INSTITUTE

AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES

wellcome connecting science

PLOTTING RESULT: HIGH RESOLUTION PLOT

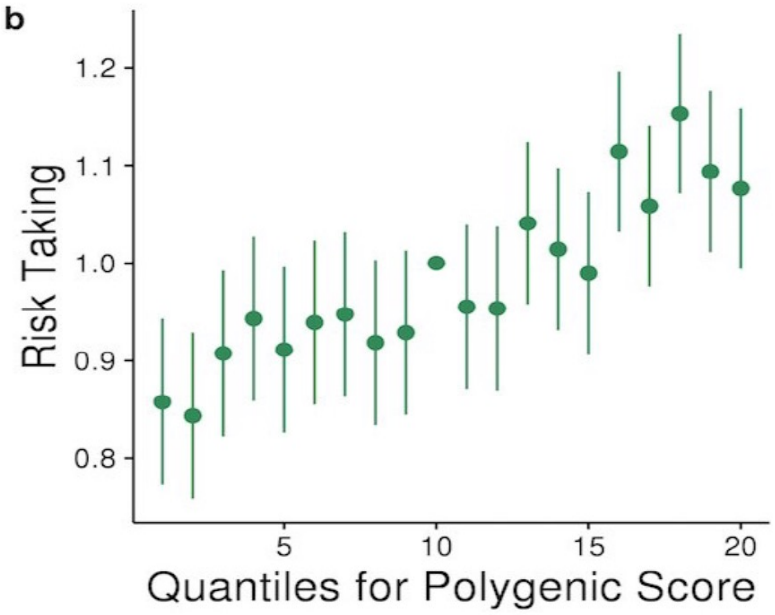


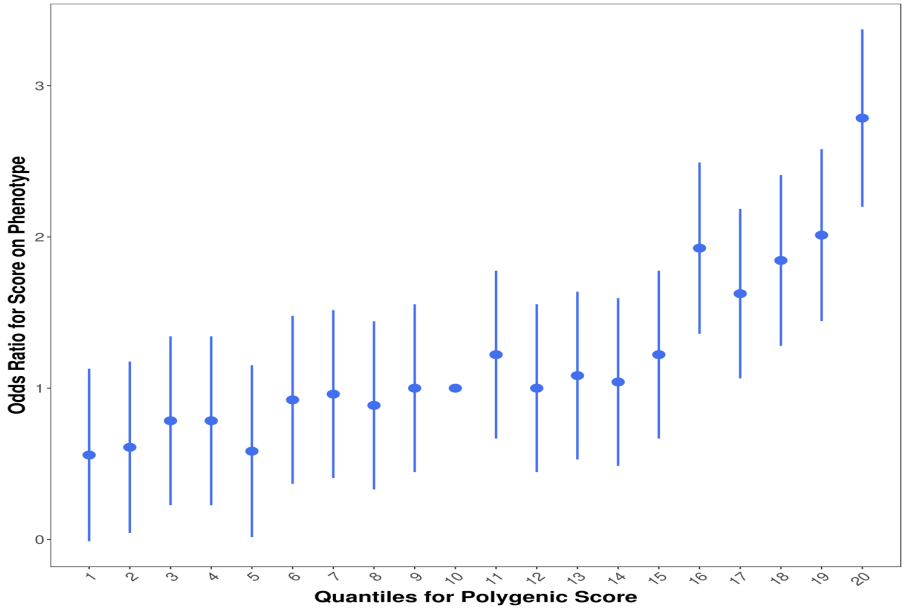


- Performs tests at almost all P-value thresholds, and search to optimise prediction

MAKERERE UNIVERSITY
IDI INFECTIOUS DISEASES INSTITUTE
AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES
wellcome connecting science

PLOTTING RESULT: QUANTILE PLOT





- Individuals grouped into PRS quantiles and then regression performed.

MAKERERE UNIVERSITY

IDI INFECTIOUS DISEASES INSTITUTE

AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES

wellcome connecting science

OVERFITTING vs UNDERFITTING

- Most predictive PRS – based on testing at many P-value thresholds – overfits to the target data and thus produces inflated results and false conclusions.

- Performing no optimisation of parameters – e.g. selecting a single arbitrary P-value threshold (such as P < x10-8 or P = 1), may lead to serious underfitting, and can lead to false conclusions.

OVERFITTING vs UNDERFITTING

- First, parameters are optimised using a training sample and then the optimised model is tested in a **test or validation dataset to assess performance.**
    - This means that a third data set is required for out-of-sample prediction.
- In the absence of an independent data set, the target sample can be subdivided into training and validation data sets, and this process can be repeated with different partitions of the sample, e.g. performing 10-fold cross-validation
- ***A true out-of-sample, and thus not overfit, assessment of performance can only be achieved via final testing on a sample entirely separate from data used in training.***

What is PRSICE-2

- A software tool to calculate, evaluate, and visualize PRS efficiently.
- Supports high-throughput scoring and clumping.
- **Why use PRSice-2?**
  - Fast and user-friendly.
  - Compatible with PLINK files.
  - Allows permutation testing, covariate adjustment, and automatic p-value thresholding.

PRSice-2 Workflow

- Quality Control: Matching SNPs between base and target datasets.
- Filtering by MAF, INFO score, missingness, etc.
- Clumping: Removes correlated SNPs (LD pruning) to avoid redundancy.
- Scoring: Calculates PRS at multiple p-value thresholds. Weights alleles using base effect sizes.
- Model Evaluation: Regresses phenotype on PRS ± covariates. Reports R^2, p-values, and best-fit threshold.

What is PRSICE-2

- PRSice performs strand flipping and clumping automatically and generates the **Height.gws.summary** f ile.

- The summary file contains the following columns:

1. **Phenotype** - Name of Phenotype.
2. **Set** - Name of Gene Set. Default is *Base*
3. **Threshold** - Best P-value Threshold
4. **PRS.R2** - Variance explained by the PRS
5. **Full.R2** - Variance explained by the full model (including the covariates)
6. **Null.R2** - Variance explained by the covariates (none provided here)
7. **Prevalence** - The population disease prevalence as indicated by the user (not provided here due to testing continuous trait)
8. **Coefficient** - The β coefficient corresponding to the effect estimate of the best-fit PRS on the target trait in the regression. A one unit increase in the PRS increases the outcome by β
9. **Standard.Error** - The standard error of the best-fit PRS β coefficient (see above)
10. **P** - The P -value relating to testing the null hypothesis that the best-fit PRS β coefficient is zero.
11. **Num_SNP** - Number of SNPs included in the best-fit PRS
12. **Empirical-P** - Only provided if permutation is performed. This is the empirical P -value corresponding to the association test of the best-fit PRS - this controls for the over-fitting that occurs when multiple thresholds are tested.

MAKERERE UNIVERSITY

IDI INFECTIOUS DISEASES INSTITUTE

AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES

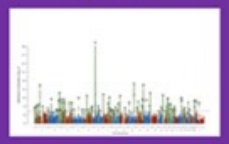
wellcome connecting science

The Standard Approach

PRS SUBSTRATE

Discovery dataset

GWAS



GWAS summary statistics

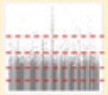
SNP	Risk allele	β
rs1234	C	0.456
rs1345	G	0.215
...	T	0.345
....	....	0.132
.....	....	....

PRS CALCULATION

LD adjustment

i.e. clumping

P-value thresholding



i.e. multiple P thresholds

Clumping & Thresholding scores

Beta shrinkage

i.e. LDpred, lassosum

Genome-wide scores

PRS TRAINING

Training dataset

Generate multiple PRS

ID	Age	PRS
101	45	0.41
102	52	3.23
103	47	1.45

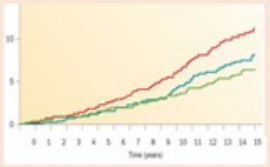
Perform association testing

Identify "best" score parameters

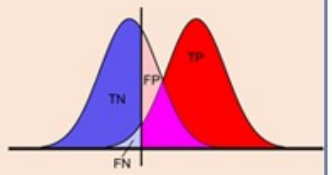
PRS VALIDATION

Validation dataset

Test "best" score



Report performance metrics



i.e. AUC, p-value, HR

Our Task for the DAY

Predicting Genetic Risk for Height and Coronary Artery Disease in a Target Cohort Using Polygenic Risk Scores.

- Background: In this exercise, we will explore how to use Polygenic Risk Scores (PRS) to estimate genetic predisposition for two phenotypes: Height and Coronary Artery Disease (CAD).

- The Base Data comprises GWAS summary statistics from large consortia studies, while the Target Data consists of genotype and phenotype information from a simulated cohort.

MAKERERE UNIVERSITY

IDI INFECTIOUS DISEASES INSTITUTE

AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES

wellcome connecting science

Our Dataset

- **Base Data:**
  - (1) Height GWAS summary statistics - Provided by the GIANT Consortium, this dataset contains **effect sizes (beta coefficients)** and p-values from a large-scale GWAS of height measured in over 250,000 individuals. Height is a quantitative trait, and the effect sizes represent the estimated increase or decrease in height per copy of the effect allele.

  - (2) CAD GWAS summary statistics - Provided by the CARDIoGRAMplusC4D Consortium, this dataset includes GWAS results from approximately 60,000 CAD cases and 120,000 controls. CAD is a **binary trait (case/control)**, and effect sizes are reported as odds ratios (converted to betas for PRS calculation).

- **Target Data** - The Target Data contains genotype information (in PLINK format) and simulated phenotype data for a cohort of individuals.

A Quick Peak at our Data

- Add script:
- head Target_Data/TAR.bim
- nano Base_Data/GIANT_Height.txt
- nano Base_Data/cad.add.txt

Exercise Context and Aim

- The goal is to use the Base Data (GWAS summary statistics) as a reference to compute Polygenic Risk Scores for the Target cohort.

- ***Specifically:***
  - Use the Height GWAS summary statistics to calculate PRS for height in the target individuals.

  - Use the CAD GWAS summary statistics to calculate PRS for CAD risk in the target individuals.

- Finally, evaluate how well the PRS predicts the phenotypes in the target cohort, such as by examining the correlation between PRS and measured height or by comparing PRS distributions between CAD cases and controls.

MAKERERE UNIVERSITY | IDI INFECTIOUS DISEASES INSTITUTE | AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES | wellcome connecting science

Key Learning Outcomes

- By completing this case study, students will:
- Understand how large-scale GWAS summary statistics are used as base data for PRS calculation.
- Learn practical steps to harmonize and prepare base and target data.
- Gain experience running PRS software (PRSice, PLINK) to compute scores.
- Interpret PRS results in the context of both quantitative and binary phenotypes.
- Appreciate the complexities involved in polygenic prediction, such as LD structure and allele alignment.

Practical Exercises

- Open the Height GWAS file (**GIANT_Height.txt**)
- Inspect the SNPs at the top of the file
- Consider the **SNPs *rs4747841*** and ***rs878177***
- What will the 'PRS' of an individual with genotypes **AA** and **TC**, respectively, be?
- And what about for an individual with **AG** and **CC**, respectively?
- What do these PRS values mean in terms of the height of those individuals?

- Column names are not standardised across reported GWAS results; thus it is important to check which column is the effect (coded) allele and which is the non-effect allele.
- For example, in the height GWAS conducted by the GIANT consortium, the effect allele is in the column Allele1, while Allele2 represents the non-effect allele.

MAKERERE UNIVERSITY
IDI INFECTIOUS DISEASES INSTITUTE
AFRICAN CENTER OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCES
wellcome connecting science

Checking the GIANT_Height.txt file

To ensure that the next step run without errors we need to check that the first line in the base_data file (GIANT_Height.txt) have the correct heading.

Run

```
head Base_Data/GIANT_Height.txt
```

Check that the first line is not empty

```
head -n 1 Base_Data/GIANT_Height.txt
```

If this comes out blank then it means the first line does not have the heading required for PRSice to run, hence we need to correct this. Check that the first two lines of the Base_Data file

```
head -n 2 Base_Data/GIANT_Height.txt
```

Remove the empty line by running

```
sed -i '1d' Base_Data/GIANT_Height.txt
```

Recheck to see this error is corrected

Check that the first line is not empty

```
head -n 1 Base_Data/GIANT_Height.txt
```