Installing R & R packages

- R can be downloaded for Windows, Mac & Linux (https://cran.r-project.org/)
- RStudio (https://posit.co/download/rstudio-desktop/) is also available for these platforms. Users may want to download RStudio in addition to R because it provides a user-friendly integrated development environment (IDE) with features like syntax highlighting, project management, and advanced visualization tools.
- Code for installing necessary R packages:
  - install.packages(c("data.table", "bigsnpr", "devtools", "dplyr", "tidyr", "ggplot2", "ggpubr", "GGally", "pROC", "patchwork"))
  - devtools::install_github("kaustubhad/fastman")

GWAS & PRS Considerations

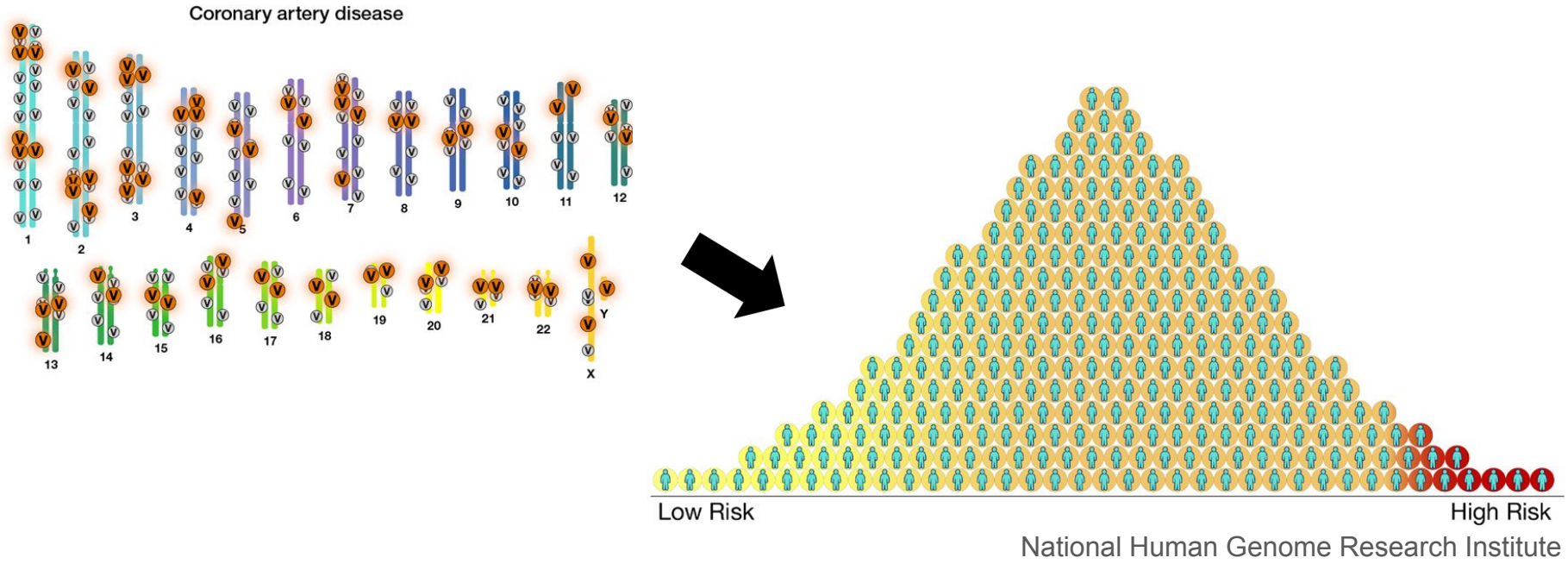
IGES/ASHG Shared Session

Christopher H. Arehart & Iain R. Konigsberg, Ph.D.

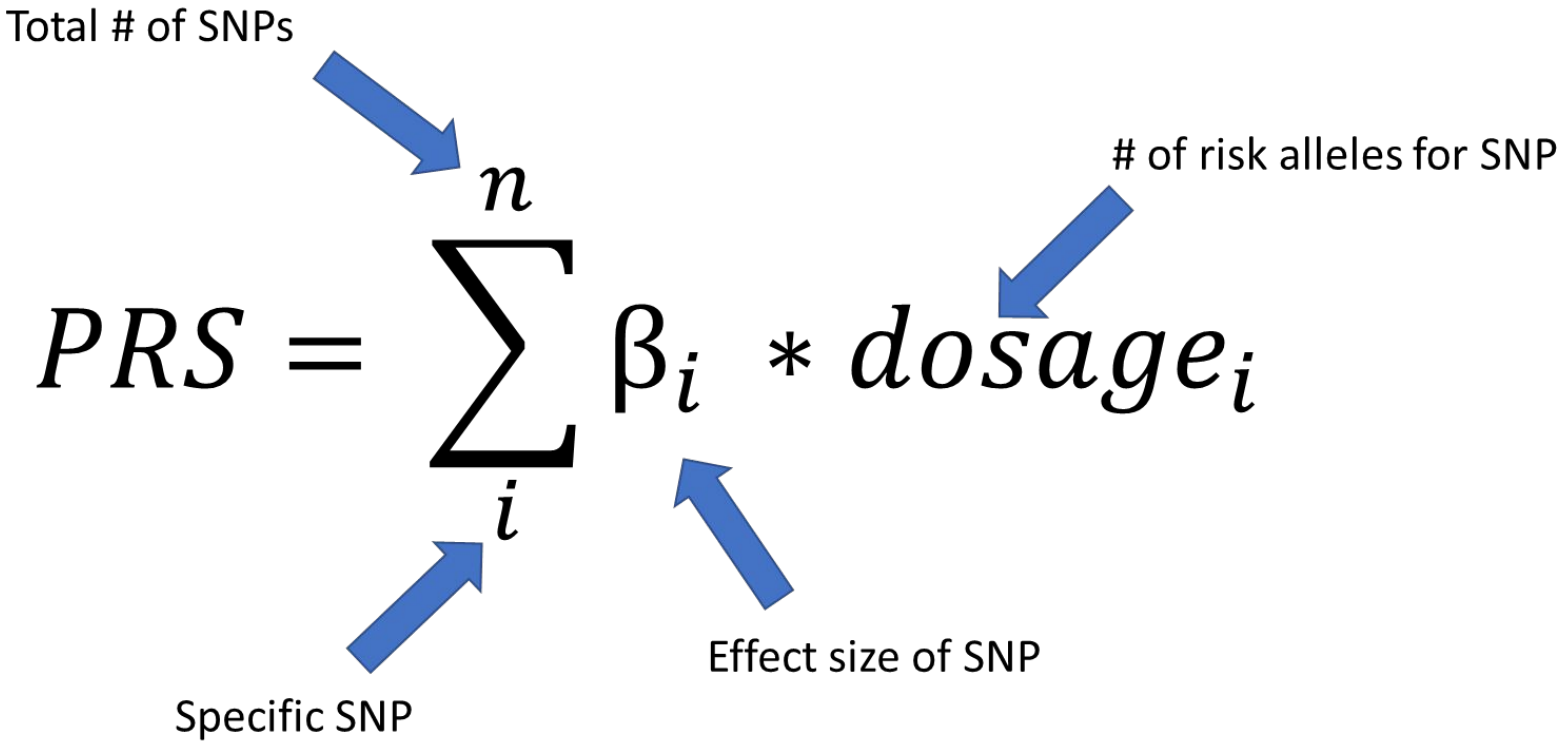
Polygenic Risk Scores (PRS)

Linear summations of genetic risk from variants across the genome

PRS are often useful in stratifying populations by risk of disease (especially those with high genetic heritability)

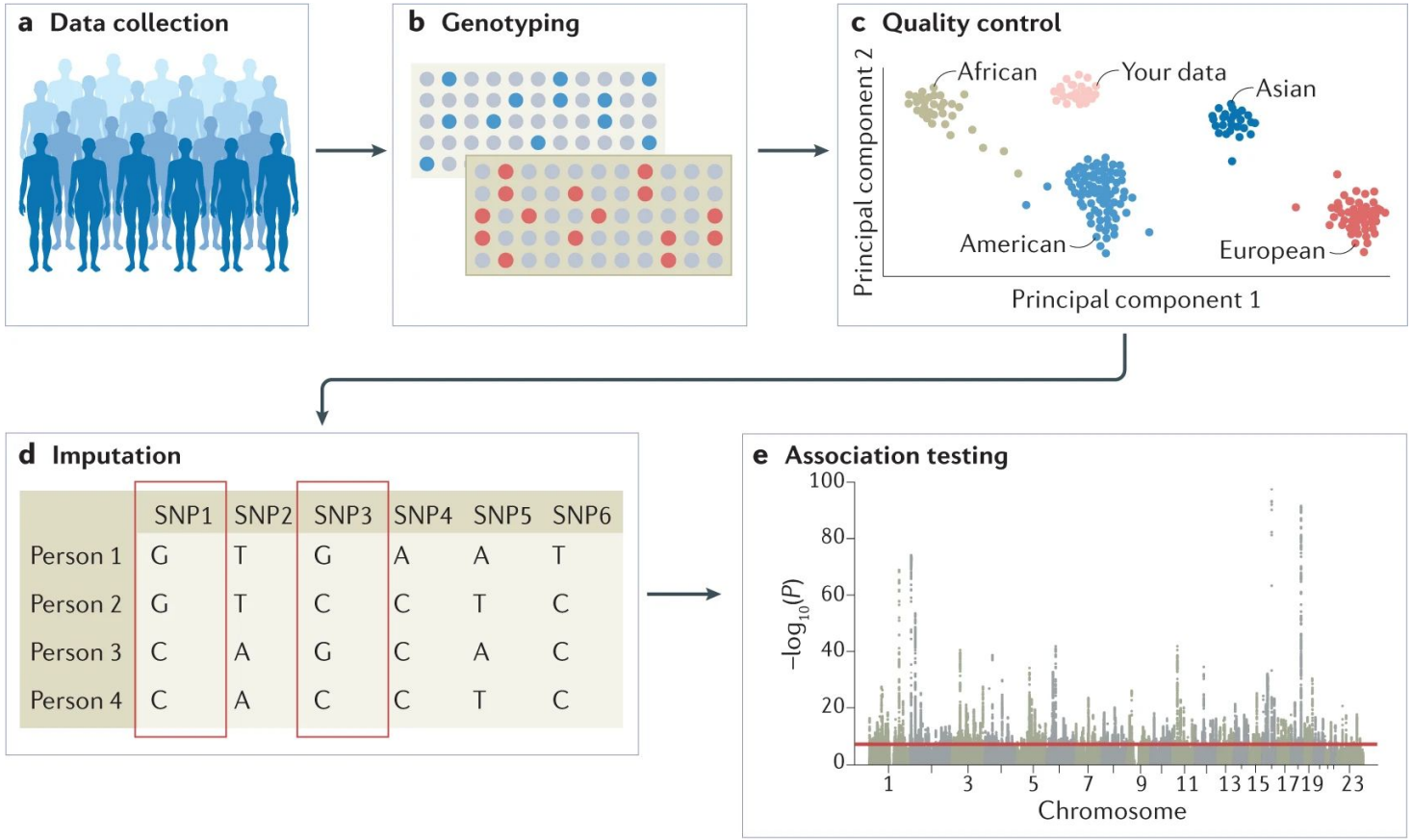


Calculating PRS



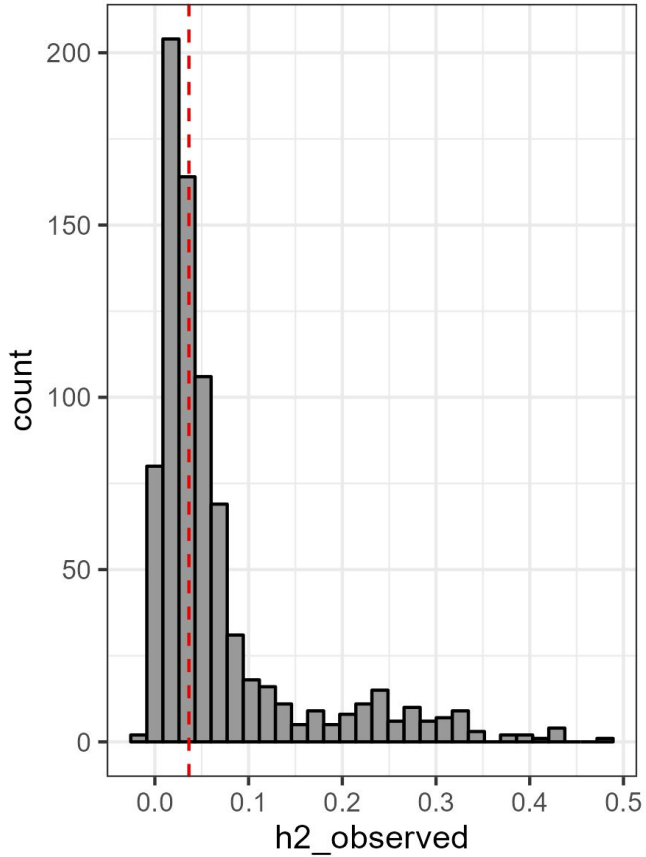
PRS weights are derived from GWAS summary statistics

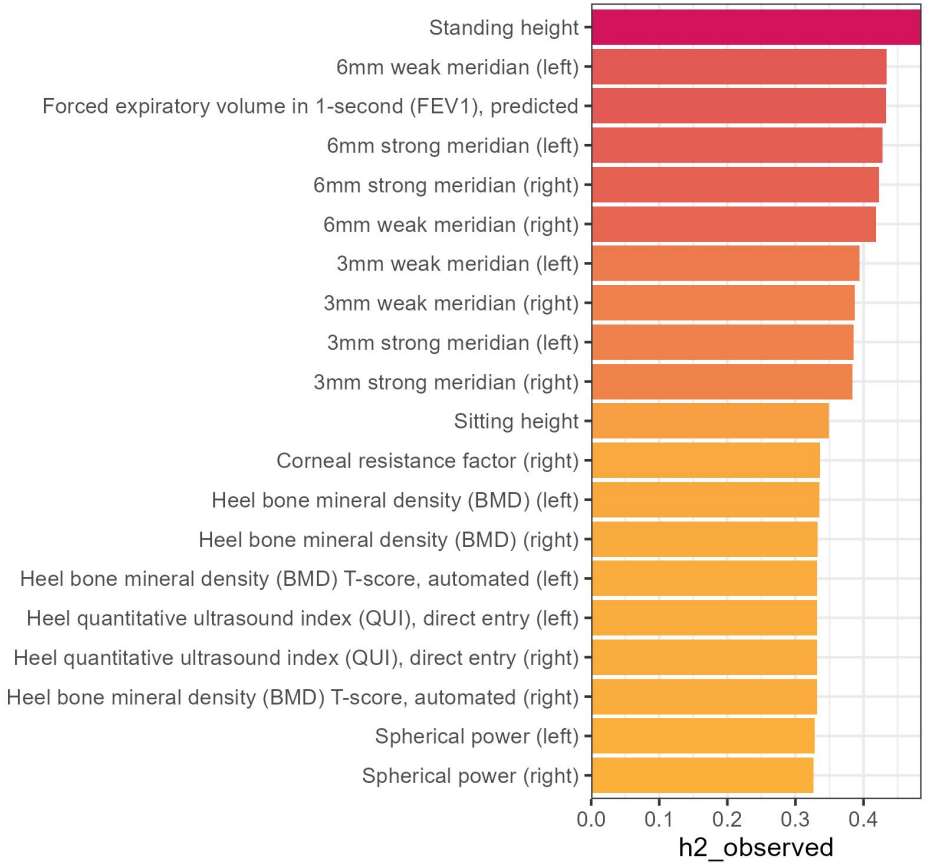
Genome-Wide Association Studies (GWAS)



Uffelmann et al. *Nat Rev Methods Primers* 2021

Genetic (SNP-based) heritability





Results downloaded from nealelab.github.io/UKBB_ldsc

GWAS Quality Control

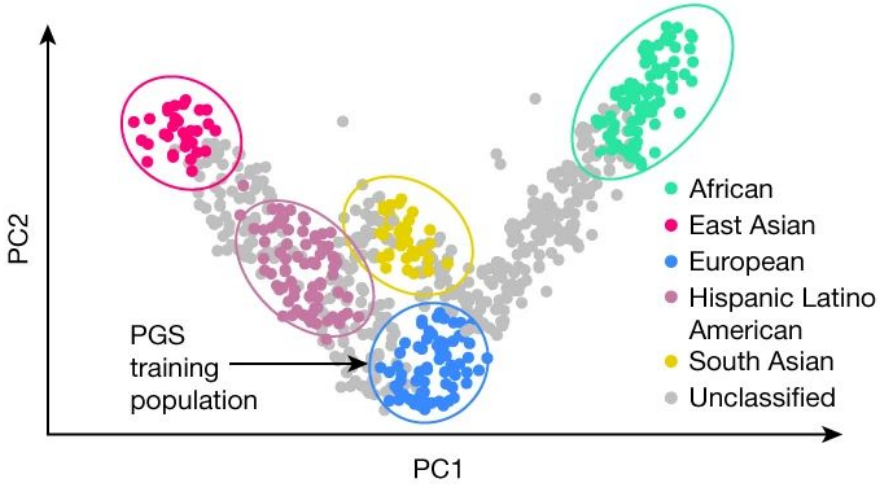
Variant-level QC:

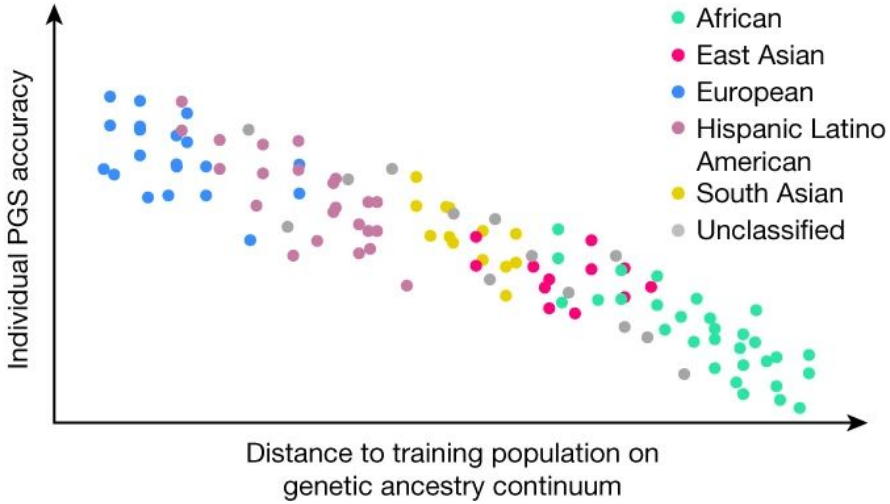
- Call rate
- Hardy-Weinberg Equilibrium
- Heterozygosity
- Minor allele frequency

Sample-level QC:

- Call rate
- Reported vs. observed sex
- Heterozygosity
- Relatedness

GWAS considerations: genetic similarity





- Environment (exposures, SDoH, etc.) likely to also differ b/w populations

Ding et al. *Nature* 2024.

GWAS Summary Statistics

SNP	effect_allele	non_effect_allele	effect	pvalue	chr	POS	SE	N
rs9999981	A	G	0.00013	0.9721	4	139575905	0.00364	157048
rs9999982	G	A	0.00842	0.03031	4	122776933	0.00389	163367
rs9999983	C	T	0.00006	0.9927	4	151115333	0.00645	163684
rs9999985	G	A	0.02017	0.07694	4	38779091	0.01140	163689
rs9999987	T	C	-0.00094	0.9108	4	4936161	0.00836	161350
rs9999993	T	A	0.00317	0.3622	4	98562671	0.00348	163657
rs9999995	G	A	-0.00255	0.6851	4	185171608	0.00629	163176
rs9999996	C	A	-0.00340	0.5113	4	69782467	0.00518	163467
rs9999997	G	A	0.00351	0.3168	4	163870478	0.00351	161565
rs9999998	C	T	0.00090	0.8512	4	117161848	0.00480	161921

Datasets to Calculate PRS

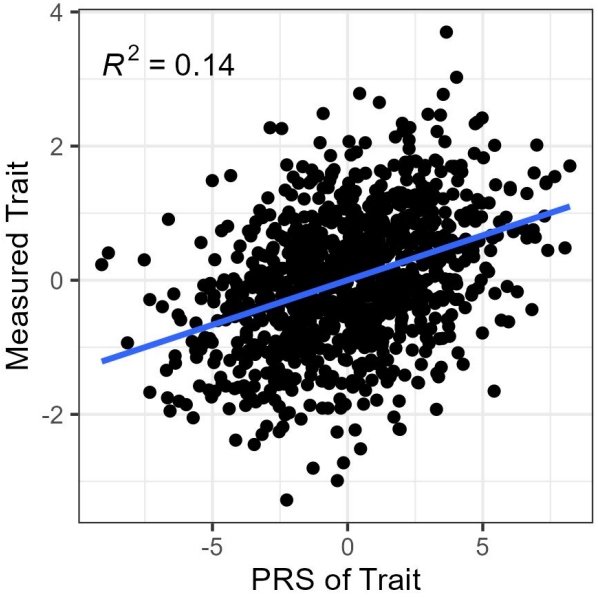
- To calculate a PRS, you need a dataset with 1) individual-level genetic information and 2) accompanying phenotype information

Example PLINK .fam file

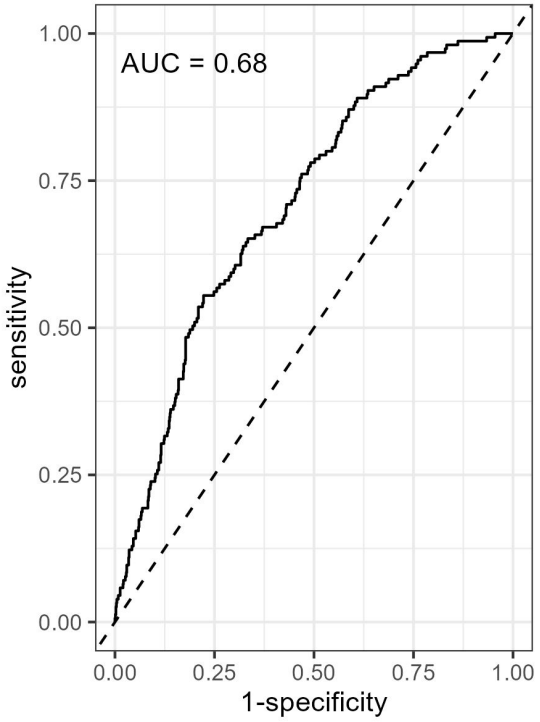
FID	IID	paternalID	maternalID	sex	phenotype
1	1	0	0	2	0
2	2	0	0	2	0
3	3	0	0	2	1
4	4	0	0	1	0
5	5	0	0	2	0
6	6	0	0	2	1

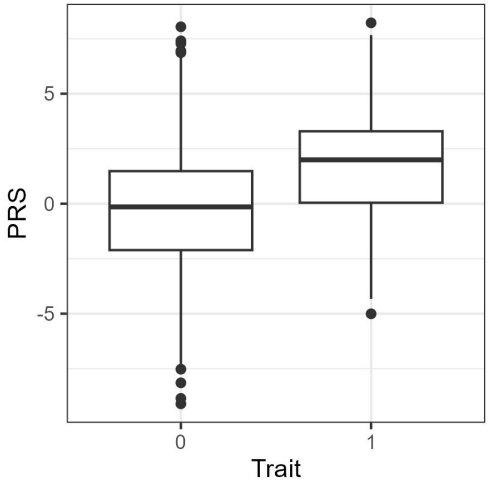
Continuous vs binary phenotypes

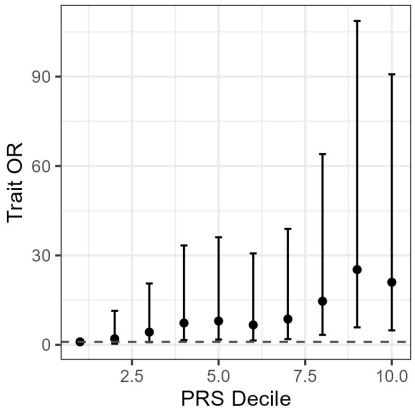
Continuous



Binary







QC considerations: target data

Sample overlap between the base data (training GWAS) and target data can result in substantial inflation of the association between the PRS and the tested trait

The following plink command applies some of these QC metrics to the target data:

```
plink --bfile myData \        ⇐ plink genotype file prefix
      --maf 0.01 \            ⇐ Remove SNPs MAF < 0.01
      --hwe 1e-6 \            ⇐ Remove SNPs with low Hardy-Weinberg Equilibrium p-value in controls
      --geno 0.01 \           ⇐ Remove SNPs with high missingness across individuals
      --mind 0.01 \           ⇐ Exclude individuals with high genotype missingness
      --out myData.QC         ⇐ plink genotype output file prefix
```

Remove individuals with high or low heterozygosity rates

- Use plink to remove individuals beyond 3 standard deviations from the mean

Closely related individuals can lead to overfitted results

- Remove individuals with KING kinship coefficients < 0.0884

Choi et al., 2020
Nature Protocols

QC considerations: heritability, genome build

Recommended to ensure $h_{SNP}^2 > 0.05$ before running PRS analyses (e.g., LD Score Regression [LDSC])

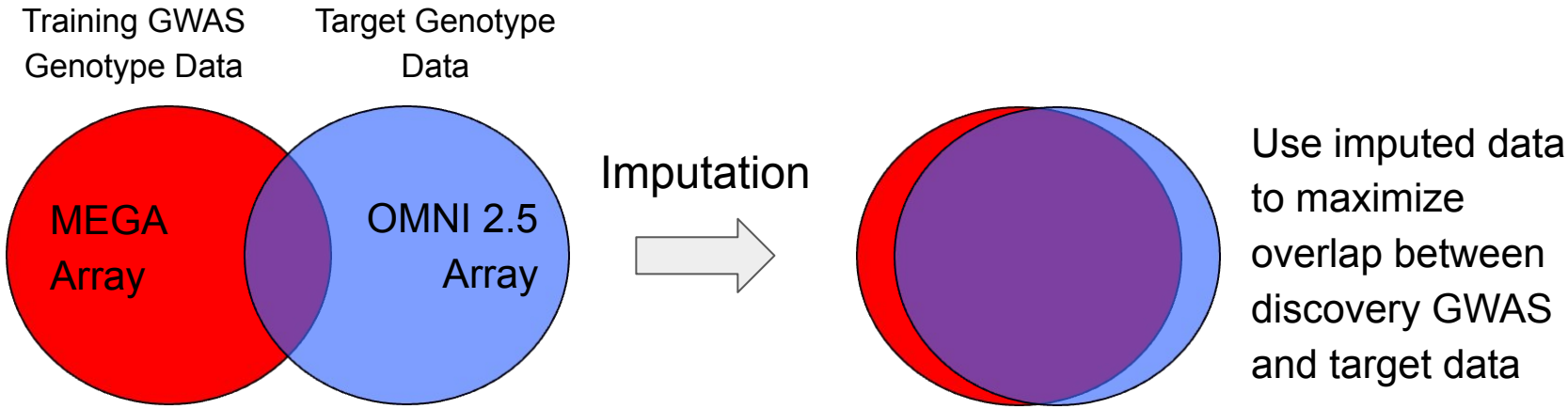
SNP-based heritability is an upper threshold for potential PRS accuracy

Ensure that the genome build matches between the training data and the target data

liftOver is a command line tool that can convert between GRCh37-hg19 & GRCh38-hg38

```
liftOver input.bed hg19ToHg38.over.chain.gz output.bed unlifted.bed
```

QC considerations: Single Nucleotide Polymorphisms (SNPs)



- Recommend removing rare SNPs with MAF < 1% and INFO < 0.8
- Attempt strand-flipping for SNPs with mismatching alleles (A ↔ T & C ↔ G)
- Attempt allele reversal (swap effect vs non-effect alleles; multiply GWAS beta by -1)
- Remove non-resolvable mismatching SNPs and ambiguous SNPs
- Generally PRSs are constructed using chr1-22 and exclude sex chromosomes

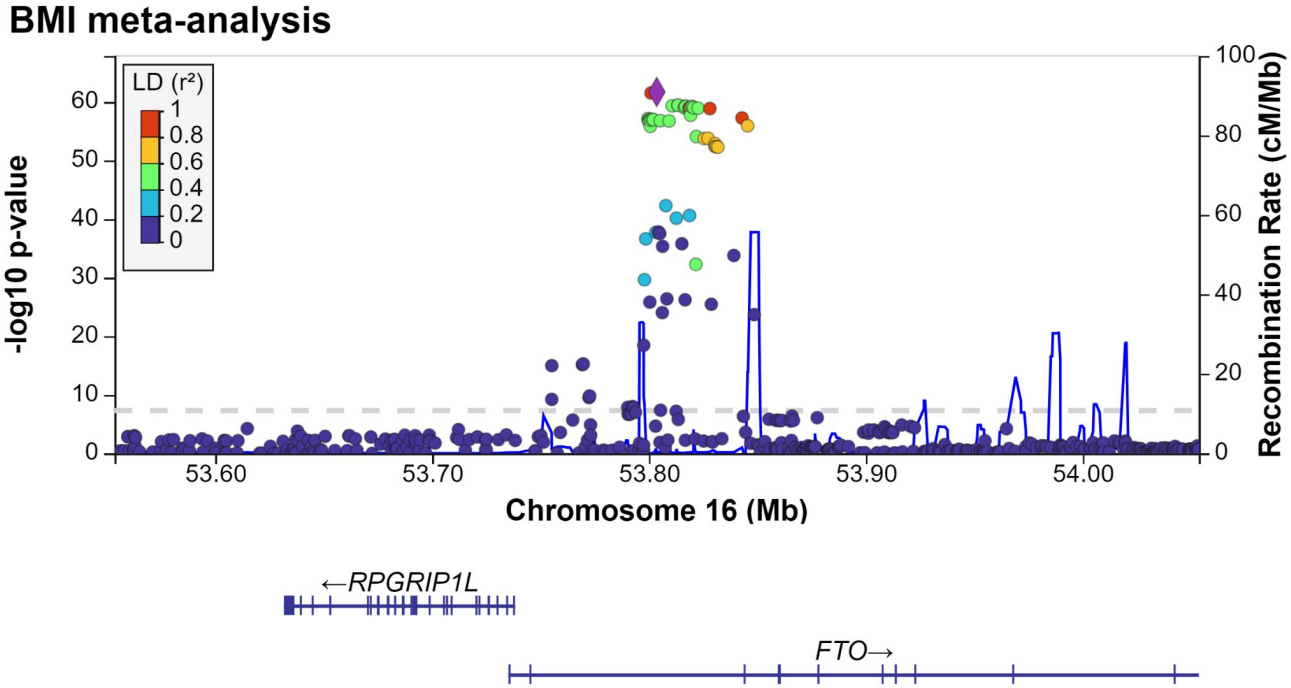
Linkage Disequilibrium

Linkage Disequilibrium (LD): the correlation between the genotypes of genetic variants across the genome

Causal variants inflate the associations of nearby SNPs

Without accounting for LD, the PRS could overestimate the effect of some regions by ‘double-counting’ the contribution across correlated SNPs

An LD reference panel (N>1000 & matching the GWAS population) allows us to account for these inflated associations



LD adjustment via pruning and clumping (P+T, C+T)

LD clumping are used to keep a subset of SNPs that are nearly uncorrelated with each other (keep only one representative SNP per region of LD)

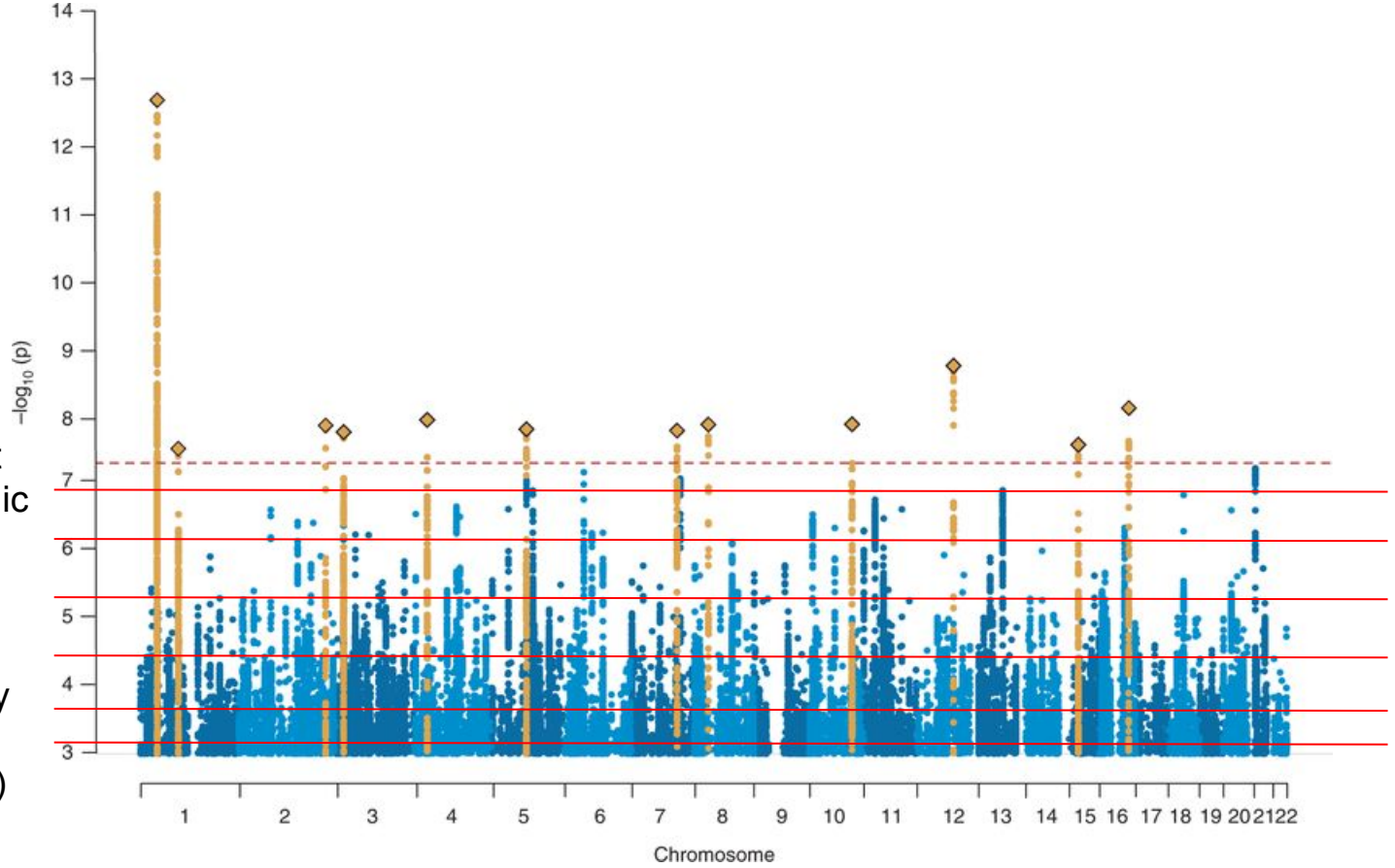
LD clumping is used to preferentially leave in strongest GWAS signal per region of LD and remove nearby weaker signals

LD adjustment aims to approximately capture the causal signal

P-value thresholding

"best-fit" PRS for a trait that is not very polygenic

"best-fit" PRS for a highly polygenic trait (e.g., infinitesimal model)



Demontis et al., 2018, Nat Gen

LD-clumping using PLINK

```
plink --bfile myData.QC \
      --clump-p1 1 \
      --clump-r2 0.1 \
      --clump-kb 250 \
      --clump sumstats \
      --clump-snp-field SNP \
      --clump-field P \
      --out myData
```

⇐ P-value threshold for a SNP to be included as an index SNP

⇐ Remove SNPs with >0.1 LD r^2 with the index SNPs

⇐ Consider SNPs within 250k of each index SNP

⇐ GWAS summary statistics file with p-values

⇐ Column label specifying SNP IDs

⇐ Column label specifying p-values

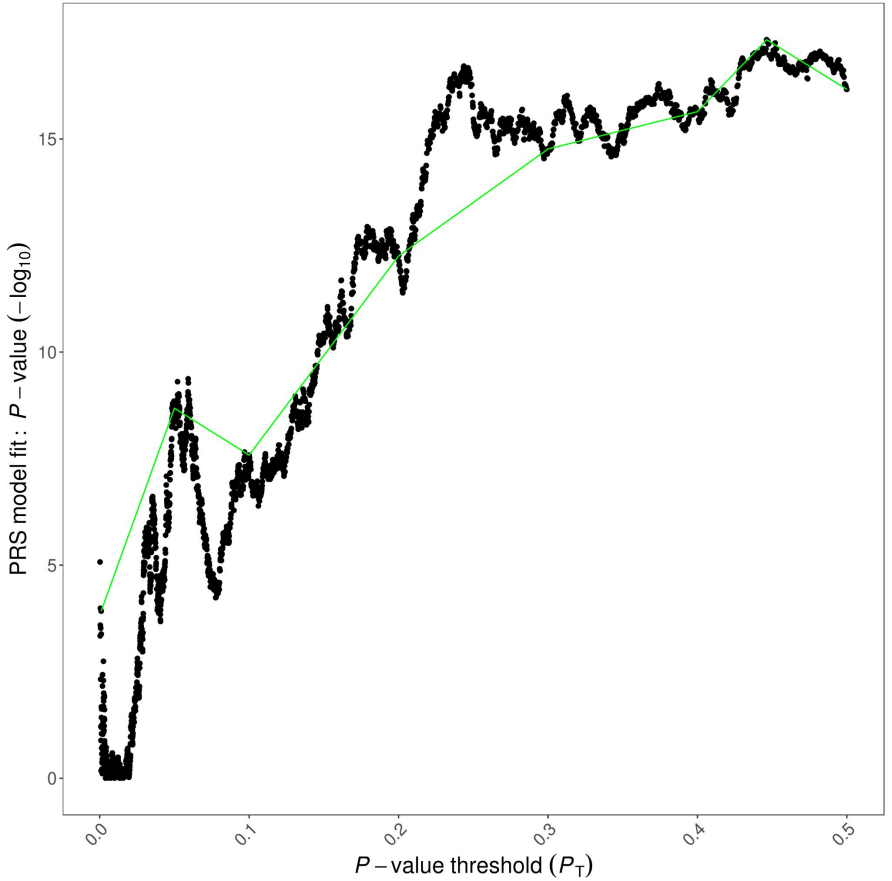
Output: myData.clumped file describing the index SNPs

The GWAS weights for these SNPs (that pass a p-value threshold) are then multiplied by the genotypes in our target dataset and summed to generate polygenic scores

PRS weights: More elegant approaches (PRSice-2)

Automated model fitting (C+T) across possible *p*-value thresholds

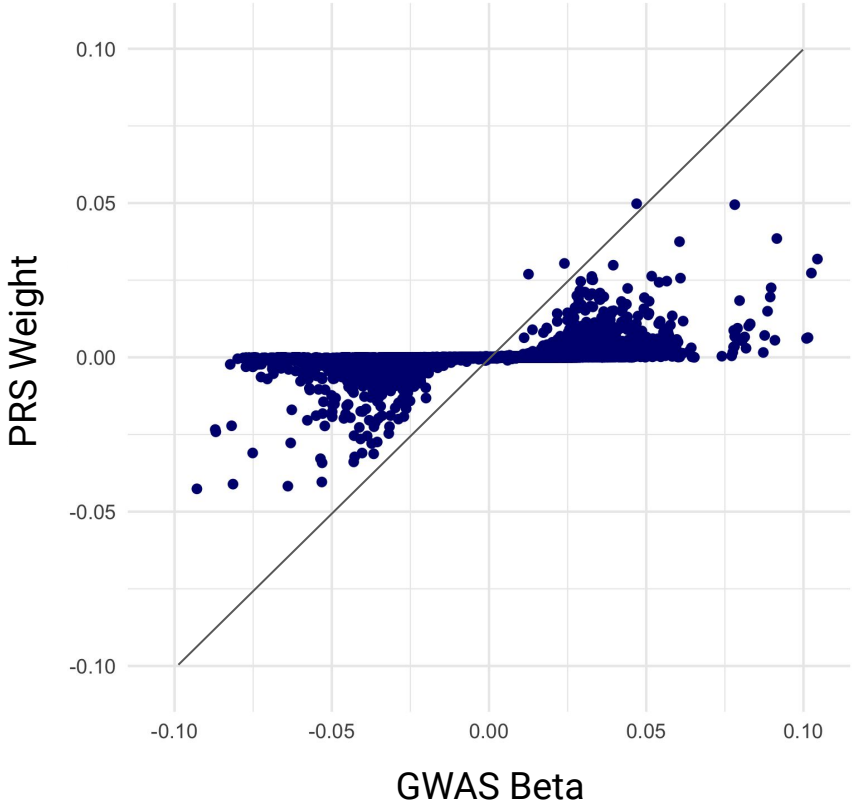
For complex polygenic traits, generally including more variants leads to better performance



Comparing GWAS beta (x-axis) to PRS weights (y-axis)

More sophisticated PRS methods shrink PRS weights based on the underlying LD structure and the standard errors of the GWAS betas.

In the end, PRS SNP weights are generally attenuated compared to the original GWAS SNP effect sizes.



Arehart et al., 2024, medRxiv

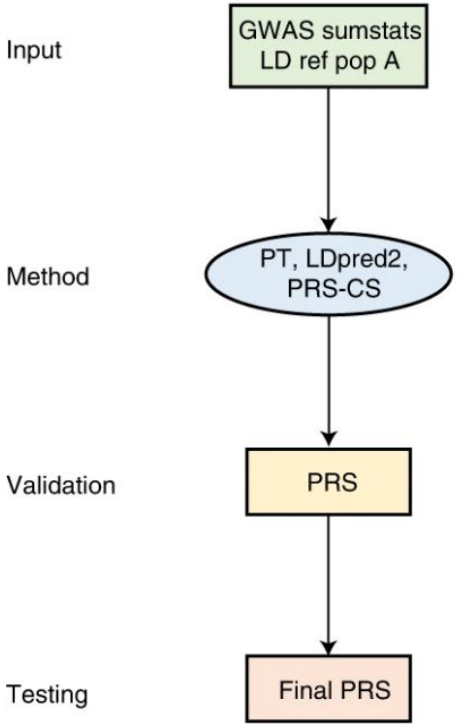
PRS weights: More sophisticated approaches

LDPred2 -- Bayesian regression is used to estimate posterior mean (adjusted weights) based on:

- Prior belief in the proportion of causal variants
- GWAS effect estimates
- LD structure of the GWAS sample

Lassosum -- LASSO regression is used to shrink and select SNP effect sizes based on:

- Sparsity and shrinkage parameters λ and s
- GWAS effect estimates
- LD structure of the GWAS sample



Ruan et al., 2022, Nat Gen

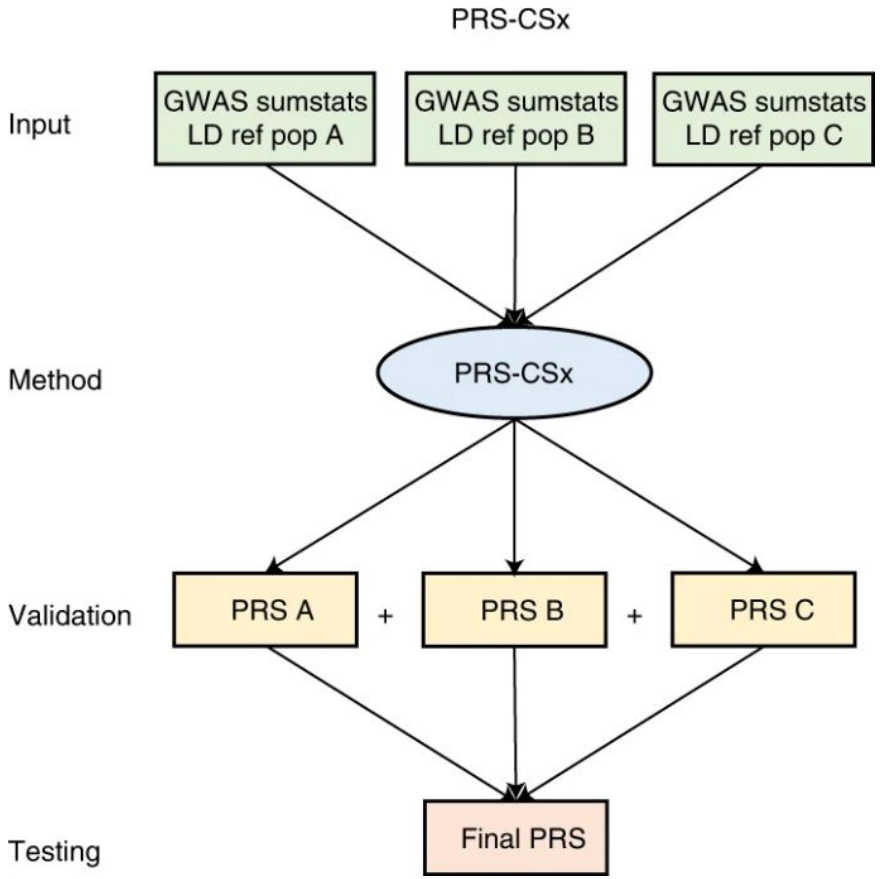
Tuning the Parameters

How do we select the optimal p-value threshold, or proportion of causal variants (to maximize the predictive power of the PRS)?

- Careful cross-validation of target dataset (phenotype & genotype data)
  - Ideally, an independent validation sample is also used to ensure the generalizability of results
- Splitting the individuals in the target dataset
  - Validation set to choose best-performing hyper-parameters (e.g., 10k individuals)
  - Test set to evaluate final polygenic scores
- LDpred2-auto is a method free of hyper-parameters
  - Averages the posterior weights across converged models
  - Automatically estimates the sparsity p and the SNP heritability

More sophisticated approaches (PRS-CSx - cross ancestry)

- Combines GWAS summary statistics from various ancestral groups
- Uses population-specific LD reference panels to adjust for LD in each population.
- Uses bayesian regression with continuous shrinkage priors to adjust SNP effect sizes based on GWAS effect estimates and LD structure
- SNPs with weaker evidence for association are more heavily shrunk toward zero



Ruan et al., 2022, Nat Gen

Tutorial I: Generating PRS Weights from GWAS Data

- Goal: Use GWAS summary statistics to calculate clumping + thresholding (C+T) PRS for individuals in a 1000 Genomes sample.

- Inputs:
  - GWAS summary statistics for height from the GIANT consortium
  - Chromosome 9 variants from the 1000 Genomes project

- Outputs:
  - 3 C+T PRS for height using chr9 variants (3 different *p*-value thresholds: 5×10^{-8}, 1×10^{-5}, & 1