THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

GWAS summary statistics

Tian Lin
tian.lin@imb.uq.edu.au

Consensus of sharing GWAS summary statistics (in human genetics research community)

Has Become a standard to share and make publicly available the summary-level data when publishing a GWAS study.

nature genetics

Asking for more

Because of the usefulness of genome-wide association study (GWAS) data for mapping regulatory variation in the human genome, the journal now asks authors to report the co-location of trait-associated variants with gene regulatory elements identified by epigenetic, functional and conservation criteria. We also ask that authors publish or database the genotype frequencies or association *P* values for all SNPs investigated, whether or not they reached genome-wide significance.

—Nat Genet editorial, July 2012

Cell Genomics

CellPress OPEN ACCESS

2021

Perspective

Workshop proceedings: GWAS summary statistics standards and sharing

Jacqueline A.L. MacArthur,[1,2,*] Annalisa Buniello,[1] Laura W. Harris,[1] James Hayhurst,[1] Aoife McMahon,[1] Elliot Sollis,[1] Maria Cerezo,[1] Peggy Hall,[3] Elizabeth Lewis,[1] Patricia L. Whetzel,[1] Orli G. Bahcall,[4] Inês Barroso,[5] Robert J. Carroll,[6] Michael Inouye,[7,8,9] Teri A. Manolio,[3] Stephen S. Rich,[10] Lucia A. Hindorff,[3] Ken Wiley,[3] and Helen Parkinson[1,*]

Table 1. Recommended standard reporting elements for GWAS SumStats

Data element	Column header	Mandatory/Optional
variant id	variant_id	One form of variant ID is mandatory, either rsID or chromosome, base pair location, and genome build[a]
chromosome	chromosome	
base pair location	base_pair_location	
p value	p_value	Mandatory
effect allele	effect_allele	Mandatory
other allele	other_allele	Mandatory
effect allele frequency	effect_allele_frequency	Mandatory
effect (odds ratio or beta)	odds_ratio or beta	Mandatory
standard error	standard_error	Mandatory
upper confidence interval	ci_upper	Optional
lower confidence interval	ci_lower	Optional

PRIMER

Check for updates

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] ✉

Table 3 | **Databases of GWAS summary statistics**

Database	Content
GWAS Catalog[110]	GWAS summary statistics and GWAS lead SNPs reported in GWAS papers
GeneAtlas[8]	UK Biobank GWAS summary statistics
Pan UKBB	UK Biobank GWAS summary statistics
GWAS Atlas[273]	Collection of publicly available GWAS summary statistics with follow-up in silico analysis
FinnGen results	GWAS summary statistics released from FinnGen, a project that collected biological samples from many sources in Finland
dbGAP	Public depository of National Institutes of Health-funded genomics data including GWAS summary statistics
OpenGWAS database	GWAS summary data sets
Pheweb.jp	GWAS summary statistics of Biobank Japan and cross-population meta-analyses

For a comprehensive list of genetic data resources, see REF.[13]. GWAS, genome-wide association studies; SNP, single-nucleotide polymorphism.

Critical information from GWAS summary data

- SNP ID
- Effect alleles and alternate alleles (A1 and A2)
- Effect allele frequencies
- Marginal SNP effects
- Standard errors
- P value
- (Per-SNP) sample sizes

COJO file (.ma)

```
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
```

Yang et al., 2012

What are the minimum data required?

Given the standard GWAS with genotypes being allelic counts (0/1/2), the minimum data required include:

- SNP marginal effect estimates
- Standard errors
- GWAS sample size

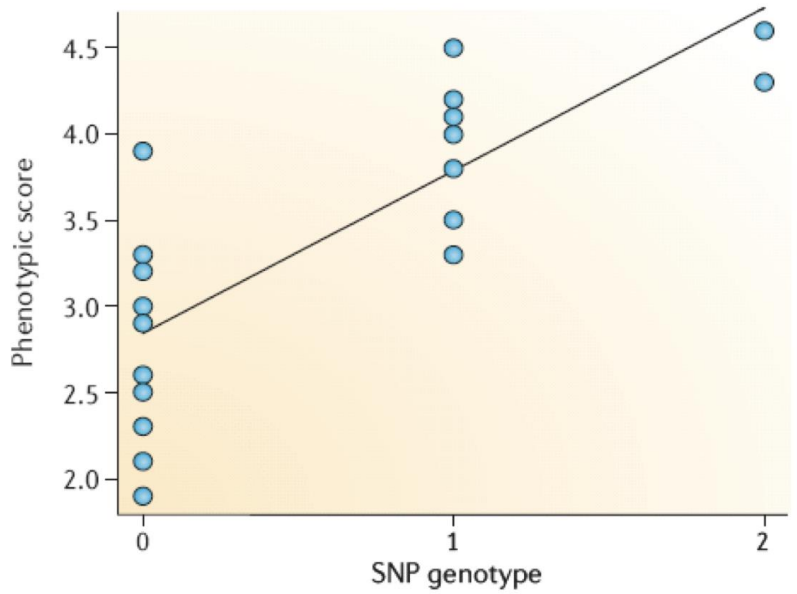
} GWAS sumstats

- LD correlations among SNPs → LD matrix

Other information critical to quality control (QC)

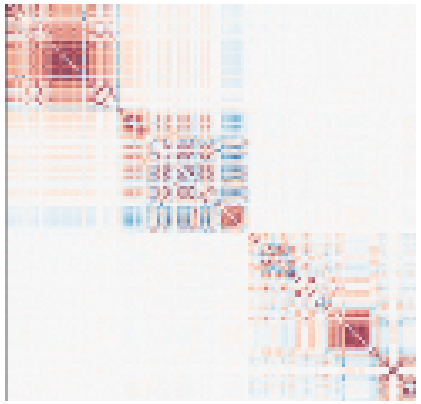
Which allele is the **effect allele** in GWAS?

e.g., A1 allele

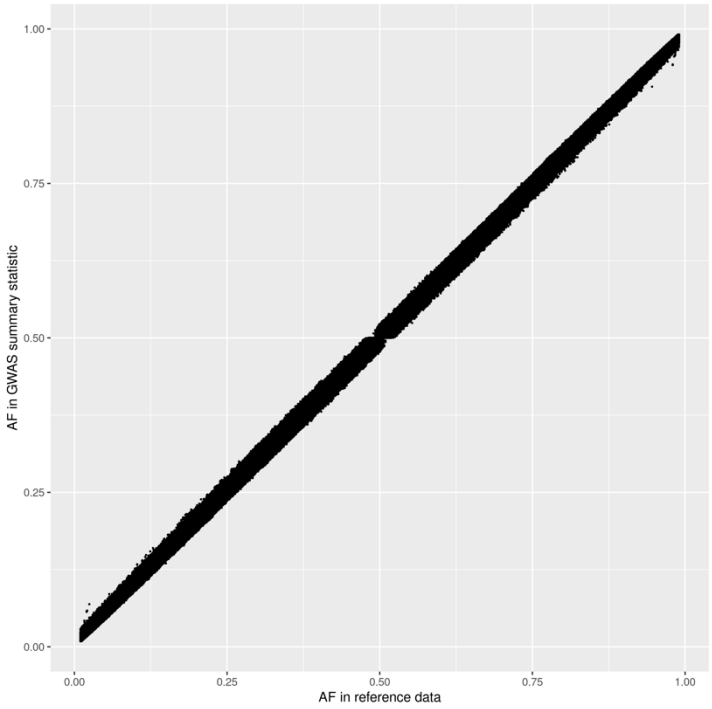


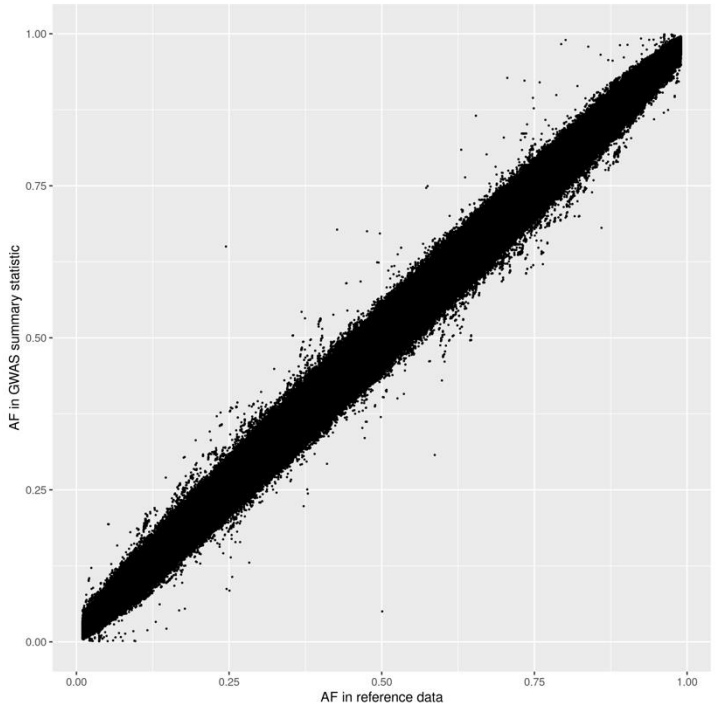
A2A2 A1A2 A1A1

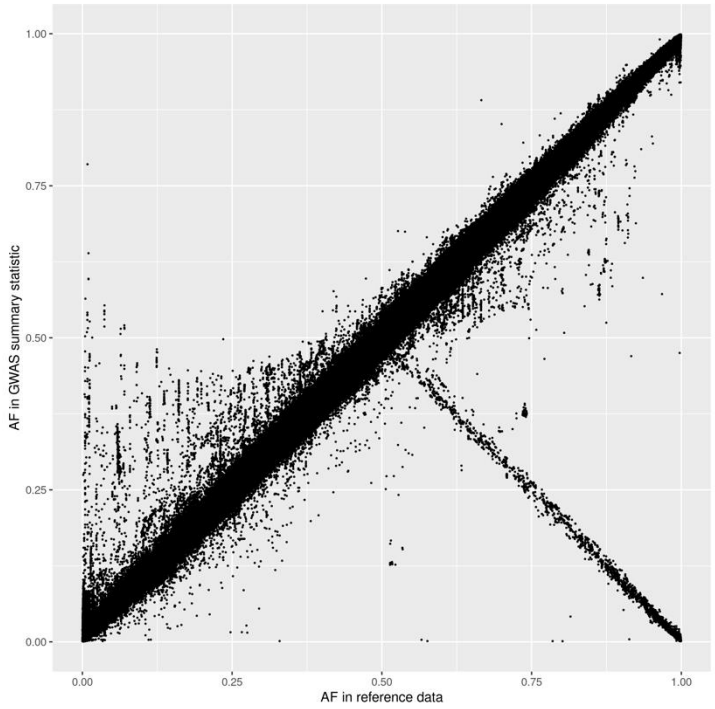
Need to match with the allele used to calculate the LD matrix in the reference sample



Allele Frequency







AF in LD reference ⟶

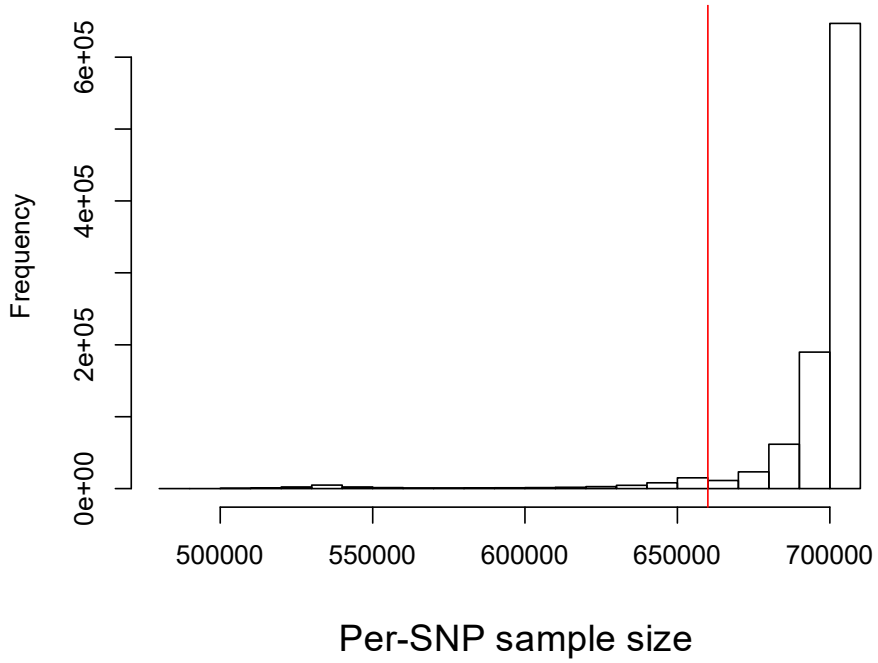
Other information critical to quality control (QC)

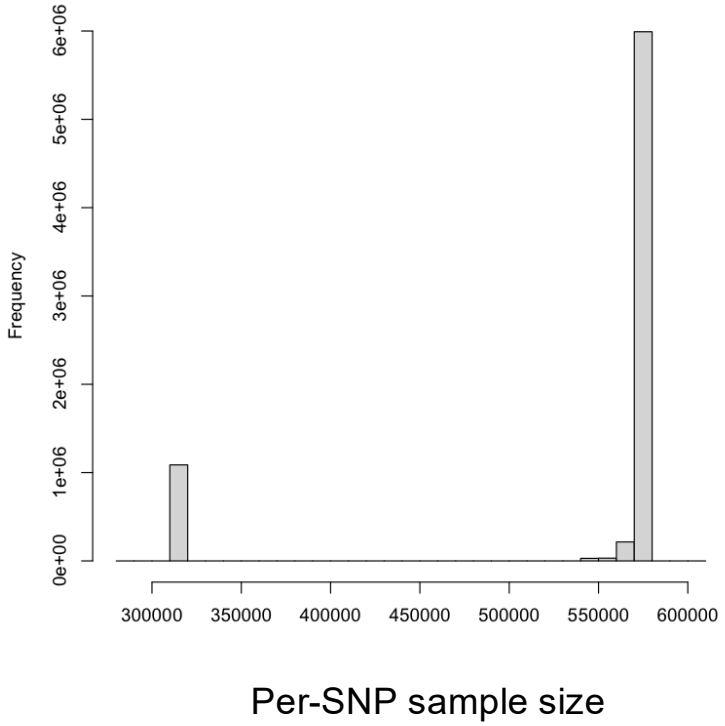
Per-SNP sample size

Heterogeneity in per-SNP sample size (usually due to meta-analysis) may result in a convergence problem in MCMC.

We recommend to visualise the per-SNP sample size distribution and remove the outliers.

Heterogeneity in per-SNP sample size





What should we check prior to the analysis?

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Raw data file

Item	What could be wrong?	How to fix?
Genome build	Inconsistent coordinates among GWAS summary data and LD reference.	Lift up to the same genome build using *liftover*
SNP ID	rsID not provided.	Use chromosome and position information to find their rsID (from LD reference file).
Alleles	Lower/upper case. Unknown effect allele (A1/A2, REF/ALT).	Check ReadMe file. Check if the predictor is negatively correlated with the phenotype.
Effect allele frequency (p)	Missing data. Provided data are minor allele frequency (MAF). Separate values in cases and controls.	Use data from LD reference. Impute by summary data $2pq = 1/(N * SE + N * b^2)$. Compute $p = \frac{N_{case}\, p_{case} + N_{ctrl}\, p_{ctrl}}{N_{case} + N_{ctrl}}$.
Marginal effect (b)	Provided data are Z-score or odds ratio (OR).	b = Z/SE if SE is provided, or $b = Z/\sqrt{2p(1-p)(N+Z^2)}$ given unit variance. b = log(OR).
Standard error (SE)	Missing data.	SE = b/Z if b is provided, or $SE = 1/\sqrt{2p(1-p)(N+Z^2)}$ given unit variance.
Sample size (N)	Missing data. Separate values in cases and controls.	Check publication/ReadMe file. Some methods require total sample size, while some requires effective sample size.
Incorrect data field format.	Some data field has NA and is non-numeric.	Convert to correct format and filter/impute missing data.

What should we check prior to the analysis? (cont)

Quality control (QC)

Item	What could be wrong?	How to fix?
Missing data	Some SNPs have missing data.	Impute the missing data or remove SNPs.
Mismatched SNPs	SNPs in GWAS are missing in the LD reference, or in reverse.	For applications requiring a perfect match, filter SNPs or impute their marginal effects (e.g., *ImpG*).
Allele discordance	Discordant alleles between data sets, e.g., A/T in GWAS but T/A in LD reference.	Flip the alleles in GWAS and take the opposite sign of the marginal effect size.
Allele frequency differences	Large differences between GWAS and LD reference data.	Remove SNPs with large difference, e.g., > 0.2.
LD differences	LD reference does not match LD in the GWAS sample.	Choose a better LD reference. Remove SNPs with LD heterogeneity (*DENTIST*).
Variable per-SNP sample sizes	Dispersed/skewed/multimodal distribution. Only overall sample size provided in meta-analysis.	Visualise the distribution. Remove long tail/minor mode/ outliers, e.g., > 3*SD. Impute $N = 1/(2pq(SE+b^2))$ if necessary.
Sample size for disease	Total sample size ($N_{case} + N_{ctrl}$) or effective sample size - which one to use?	For *SBayes*, we recommend using the total sample size.

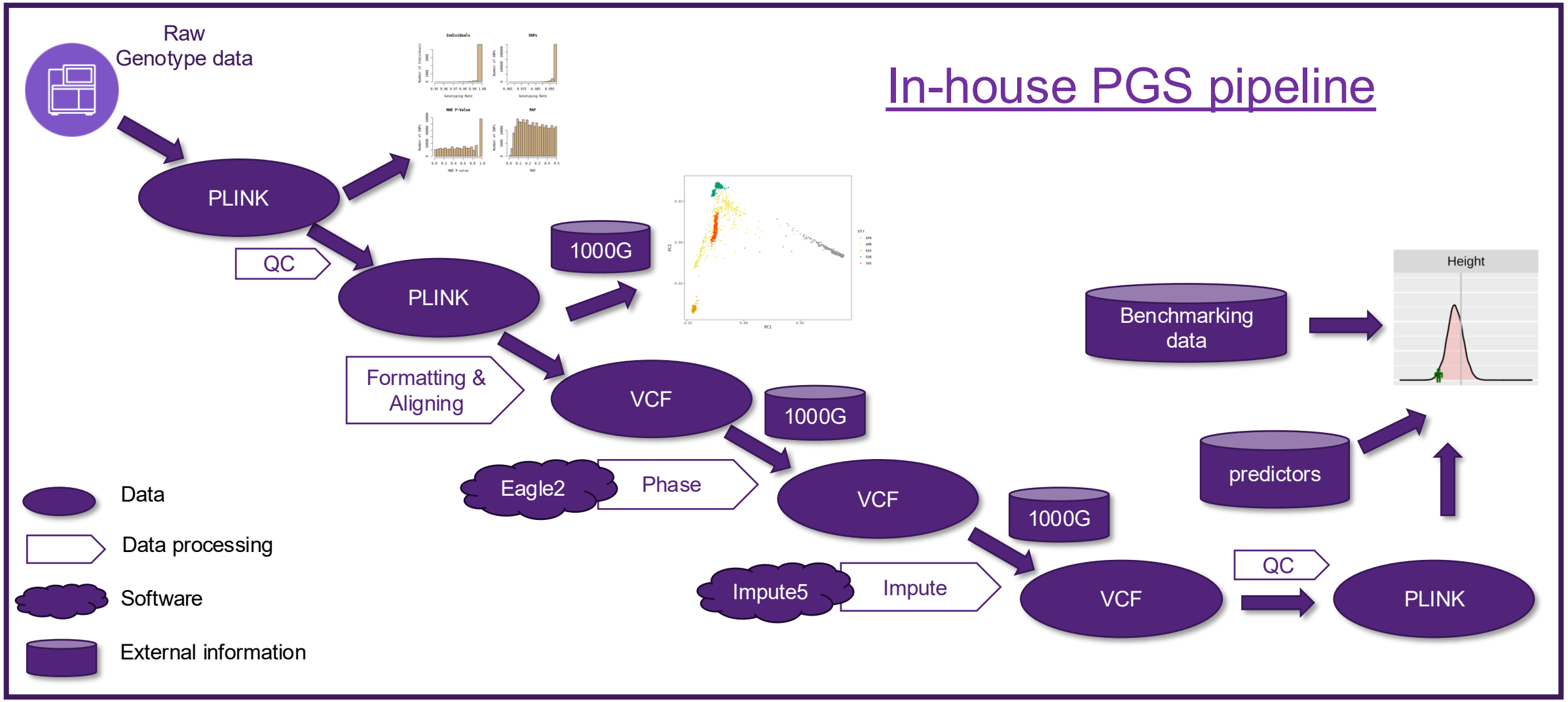
Practical6_SumStat_QC.html

An in-house PGS pipeline

Genetics & Genomics Winter School

Module 5

schematic of technical pipeline



Genotype data from arrays

- Can assay ~1M SNPs per individual with 'SNP chips'
- Data is typically 'counts' of a reference allele

genotype file:

	SNP1	SNP2	SNP3	SNP4
Bob	0	1	0	1
Fred	1	2	0	0
Jose	1	2	2	2
Andy	2	1	1	1

map file:

	chr	position	ref	alt
SNP1	1	52196307	A	T
SNP2	1	52462094	C	T
SNP3	1	52736008	A	G
SNP4	1	53010891	T	C

[borrowed from Keth]

Why a raw data is not ready for PGS profiling?

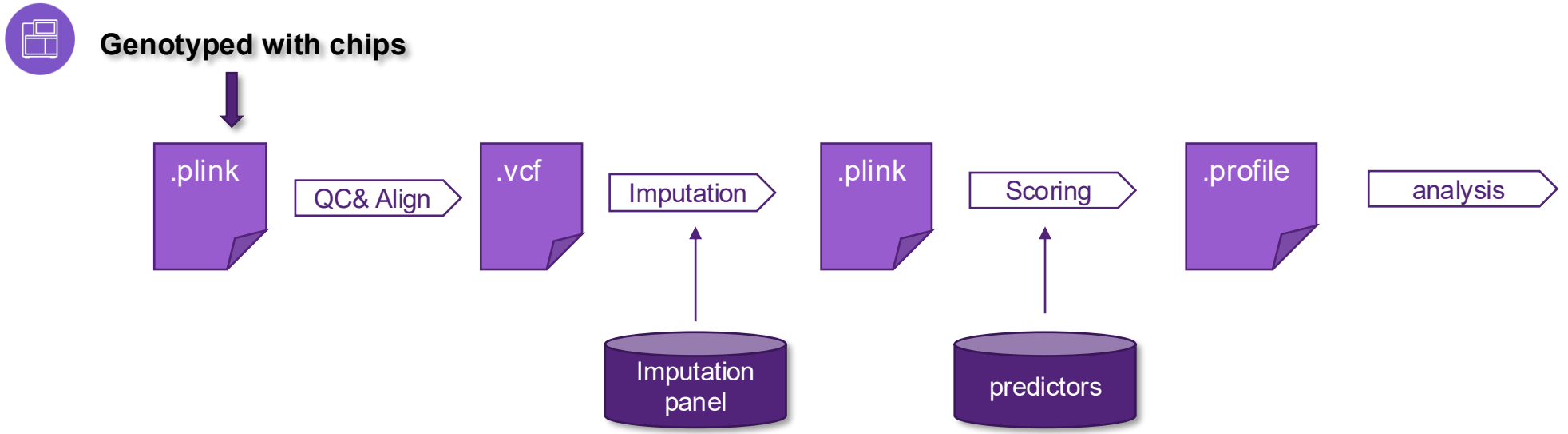
- Quality

- Coverage

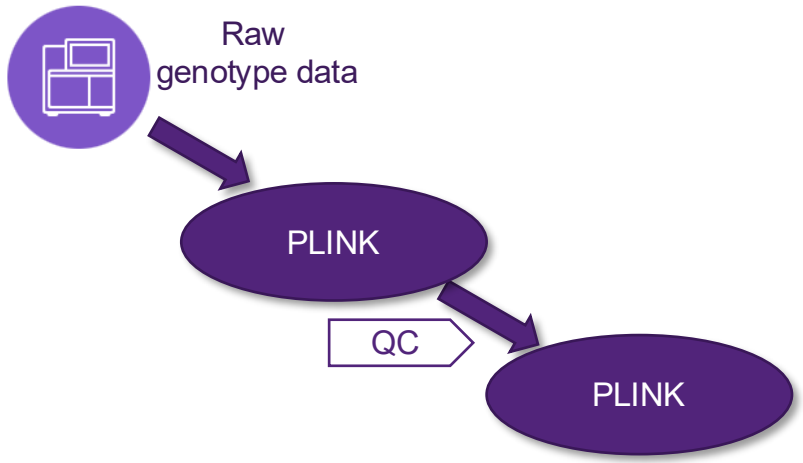
A high density SBayesRC Predictor – 7.3M SNPs

	Number of Nucleotide/Variants
Whole human genome haplotype	3 billion
TopMed	445 million
1000G	80 million
HRC	40 million
HapMap3	1 million
Illumina GSA chip	654 thousand

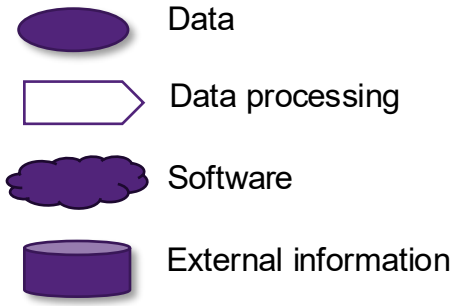
Overview of PGS pipeline



schematic of technical pipeline



In-house PGS pipeline



Revisit Genotype data QC for a GWAS study

- Per Individual QC
  - *1) removal of individuals with excess missing genotypes*
  - *2) removal of individuals with outlying homozygosity values*
  - *3) remove of samples showing a discordant sex*
  - *4) removal of related or duplicate samples, and*
  - *5) removal of ancestry outliers*

- Per SNP QC
  - *1) removal of SNPs with excess missing genotypes*
  - *2) removal of SNPs that deviate from Hardy-Weinberg equilibrium*
  - *3) removal of SNPs with low minor allele frequency*
  - *4) comparing allele frequency to known values*

Extra consideration in practice

- Large number of SNPs with MAF = 0
  - Missing Alleles

- Replicates and relatives can exist

Genotype data QC

- Per Individual QC
  - *1) removal of individuals with excess missing genotypes*
  - *2) removal of individuals with outlying homozygosity values*
  - *3) remove of samples showing a discordant sex*
  - *4) removal of related or duplicate samples, and*
  - *5) removal of ancestry outliers*

- Per SNP QC
  - *1) removal of SNPs with excess missing genotypes*
  - *2) removal of SNPs that deviate from Hardy-Weinberg equilibrium*
  - *3) removal of SNPs with low minor allele frequency*
  - *4) comparing allele frequency to known values*

Extra consideration in practice

- Large number of SNPs with MAF = 0
  - Missing Alleles
- Replicates and relatives can exist
- Different genome build between raw data and imputation panel

Human Genome Assemblies

https://hgdownload.soe.ucsc.edu/downloads.html

Very similar / Same.
GRCh37 names them `chr1`, `chr2`, `chr3`, etc, while hg19 just has `1`, `2`, `3`.
Different Mitochondria contigs.

Human genomes

Jan. 2022 (T2T-CHM13 v2.0/hs1)

- Fileserver (bigBed, maf, fa, etc) annotations
- Standard genome sequence files and select annotations (2bit, GTF, GC-content, etc)
- LiftOver files
- Pairwise alignments

Telomere-to-Telomere

A haploid human genome without gaps

Dec. 2013 (GRCh38/hg38)

- Genome sequence files and select annotations (2bit, GTF, GC-content, etc)
- Sequence data by chromosome
- Annotations
- SNP-masked fasta files
- LiftOver files
- Pairwise alignments
- Multiple alignments
- Patches
- Data archive

hg19ToHg38.over.chain.gz

hg38ToHg19.over.chain.gz

hs1ToHg38.over.chain.gz

hs1ToHg19.over.chain.gz

hg38ToHs1.over.chain.gz

Feb. 2009 (GRCh37/hg19)

- Genome sequence files and select annotations (2bit, GTF, GC-content, etc)
- Sequence data by chromosome
- Annotations
- GC percent data
- Protein database for hg19
- SNP-masked fasta files
- LiftOver files
- Pairwise alignments (primates)
- Pairwise alignments (other mammals)
- Pairwise alignments (other vertebrates)
- Multiple alignments
- Patches
- Data archive

Mar. 2006 (NCBI36/hg18)

- Data and annotations

GRCh = Genome Reference Consortium Human Build

Liftover

We recommend realigning against the manifest files if possible!

Option 1. https://www.strand.org.uk

- update_build.sh <bed-file-stem> <strand-file> <output-file-stem>

Option 2. https://genome.sph.umich.edu/wiki/LiftOver

- python liftMap.py -m data_recoded.map -p data_recoded.ped -o data_recoded_lifted

Option 3. LiftOverPlink

- https://github.com/sritchie73/liftOverPlink/blob/master/README.md

Option 4. use reference file to update dbSNP locations in bim file or GWAS statistics

- Hg38 dbsnp_146.hg38.vcf.gz
- Hg19 dbsnp_138.hg19.vcf.gz

- Available: https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle

Best liftover tool: BCFtools/liftover

bcftools +liftover --no-version

```
-Ou ALL.wgs.phase3_shapeit2_mvncall_integrated_v5c.20130502.sites.vcf.gz -- \
-s   GRCh37/human_g1k_v37.fasta \
-f   GRCh38/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna \
-c   GRCh38/hg19ToHg38.over.chain.gz \
--reject ALL.wgs.phase3_shapeit2_mvncall_integrated_v5c.20130502.sites.reject.bcf \
--reject-type b \
--write-src | \
```

-Ou: --output-type (u = uncompressed BCF)

-s: --src-fasta-ref <file> source reference sequence in fasta format

-f, --fasta-ref <file> destination reference sequence in fasta forma

-c, --chain <file> UCSC liftOver chain file

--reject <file> output variants that cannot be lifted over

--reject-type (b = compressed BCF)

--write-src write the source contig/position/alleles for lifted variants

```
bcftools sort -Ob | tee ALL.wgs.phase3_shapeit2_mvncall_integrated_v5c.20130502.sites.hg38.bcf | \
 bcftools index --force \
        --output ALL.wgs.phase3_shapeit2_mvncall_integrated_v5c.20130502.sites.hg38.bcf.csi
```

Extra consideration in practice

- Large number of SNPs with MAF = 0
  - Missing Alleles
- Replicates and relatives can exist
- Different genome build between raw data and imputation panel
- SNPs alleles from negative strand

Chromosomes, strands and SNP alleles

Paternal chromosome

Positive strand

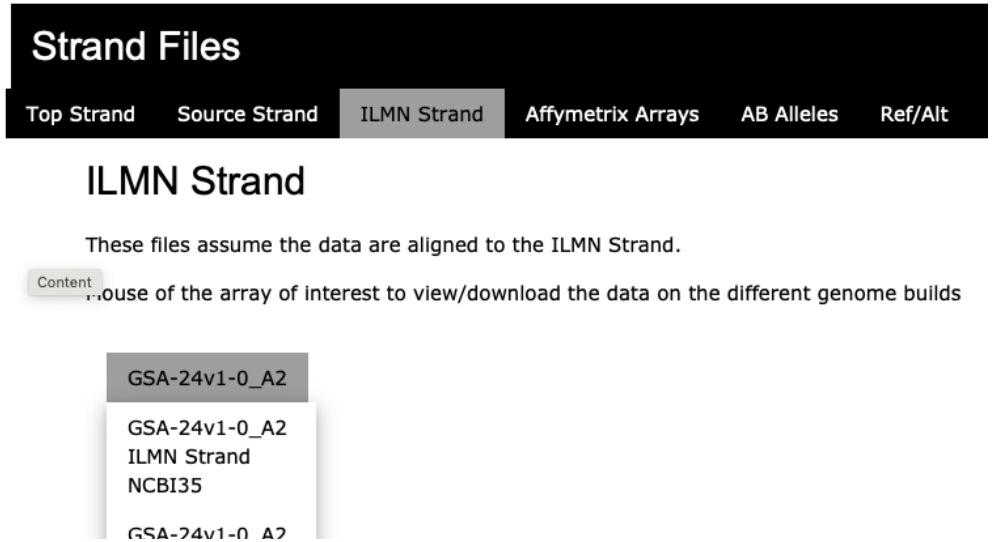
Negative strand

Flip the strand

Maternal chromosome

Strand resource

https://www.strand.org.uk



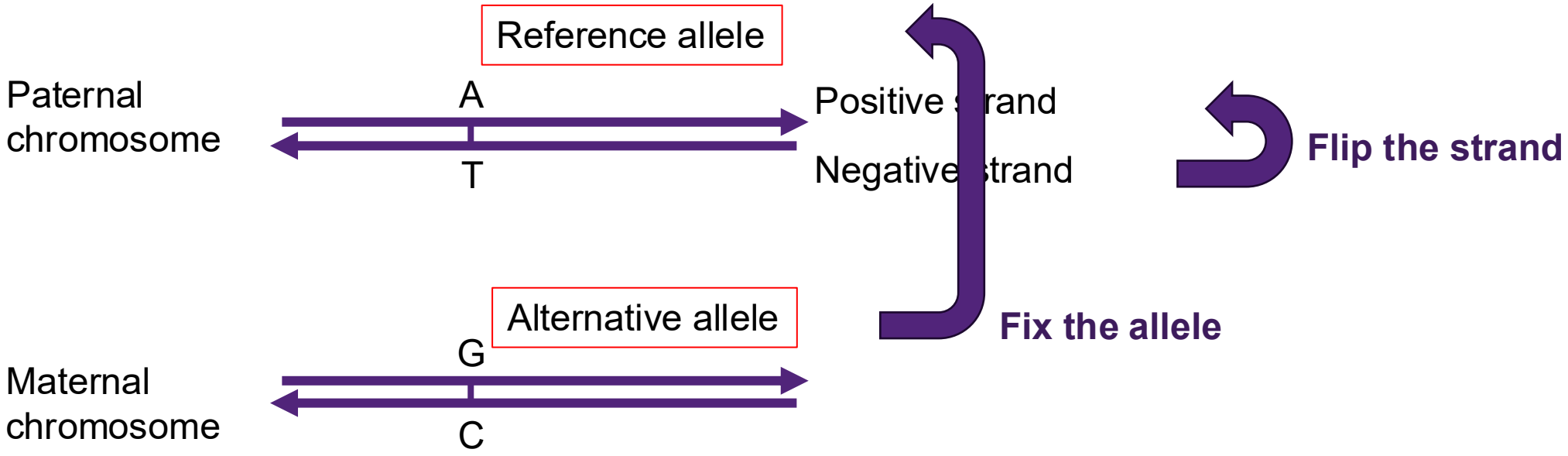
Usage is:

update_build.sh <bed-file-stem> <strand-file> <output-file-stem>

where:

<bed-file-stem>	is the name of your binary ped set minus the .bed, .bim or .fam extension
<strand-file>	is appropriate strand file for you chip and current strand orientation (TOP, SOURCE, ILMN)
<output-file-stem>	is the name of the new output file to create again minus the .bed, .bim or .fam extension

Chromosomes, strands and SNP alleles

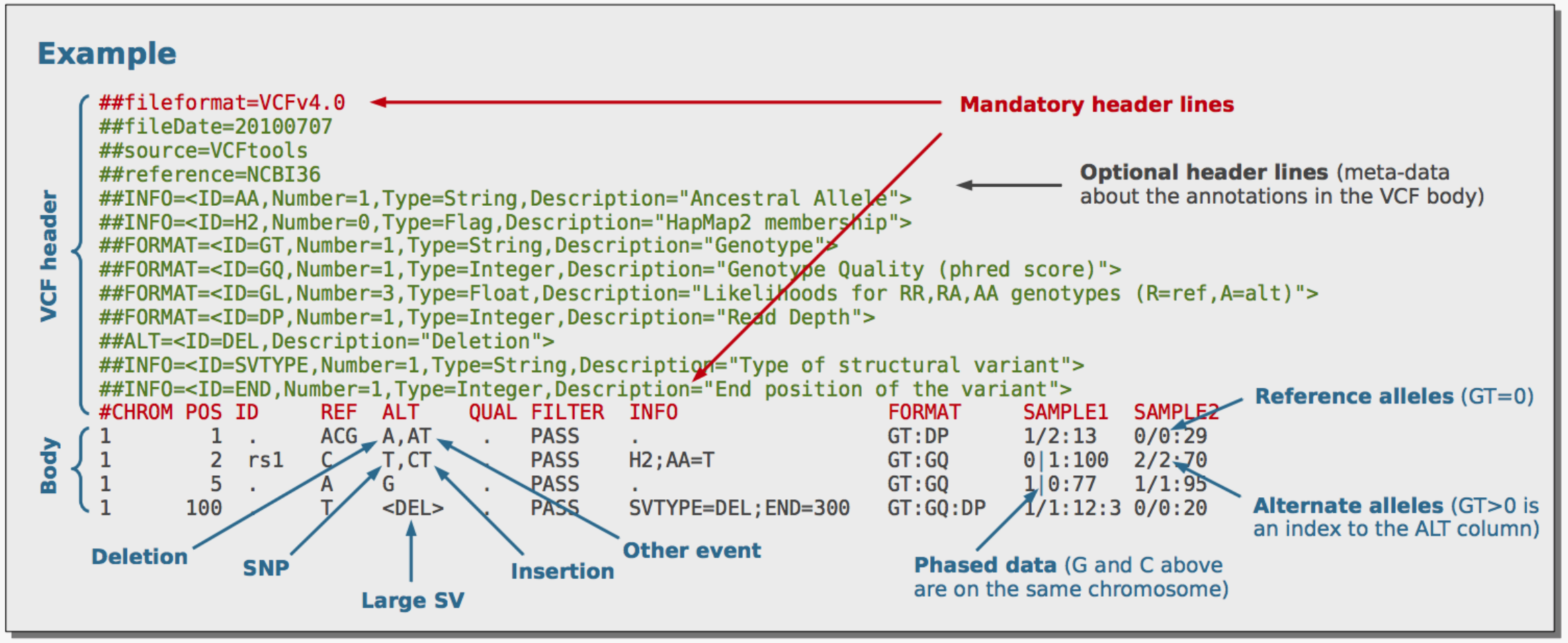


Example script to fix ref allele

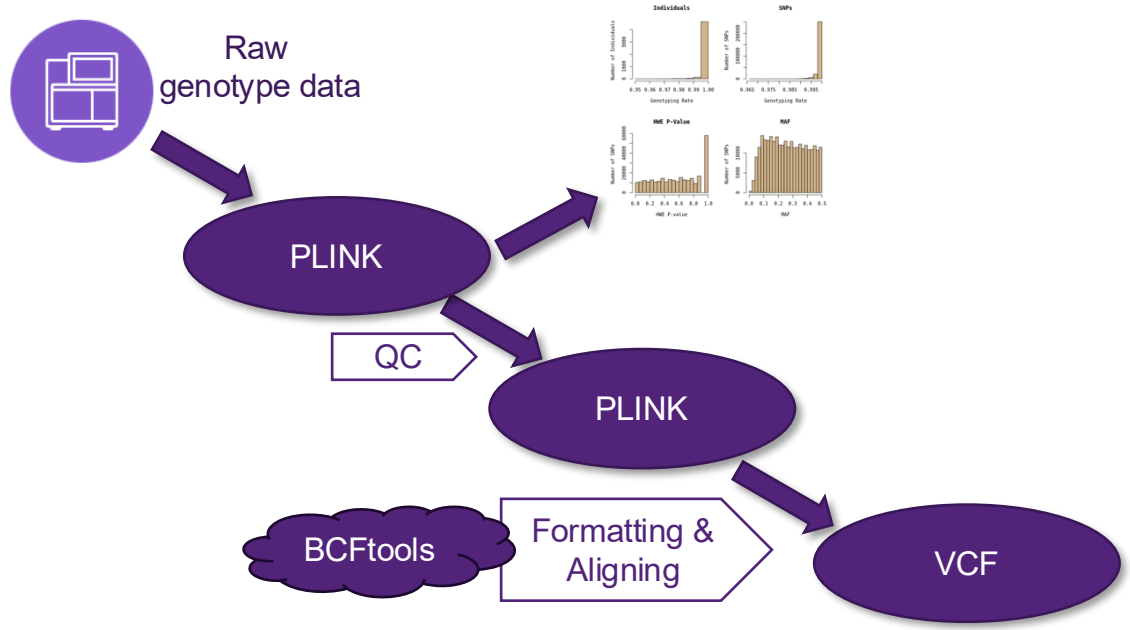
```
chr=22

# Pull out data for relevant chromosome and convert to VCF.
plink --bfile ${data}_chr${chr}  --recode vcf --out  ${data}_chr${chr}

# Sort and compress the VCF file
vcf-sort ${data}_chr${chr}.vcf | bgzip -c > ${data}_chr${chr}.vcf.gz

# Fix the reference allele to match the GRCh37 reference fasta (human_g1k_v37.fasta).
ref2fix=${refpath}/human_g1k_v37.fasta
BCFTOOLS_PLUGINS=/software/bin/
bcftools \
  +fixref \
  ${data}_chr${chr}.vcf.gz \
  -Oz \
  -o fixed_${data}_chr${chr}.vcf.gz  \
  -- -d \
  -f ${ref2fix} \
  -m flip

zcat fixed_${data}_chr${chr}.vcf.gz | bgzip -c > indexed_fixed_${data}_chr${chr}.vcf.gz

# create index file.
tabix indexed_fixed_${data}_chr${chr}.vcf.gz
```

BCFtools
VCFtools

Example VCF files

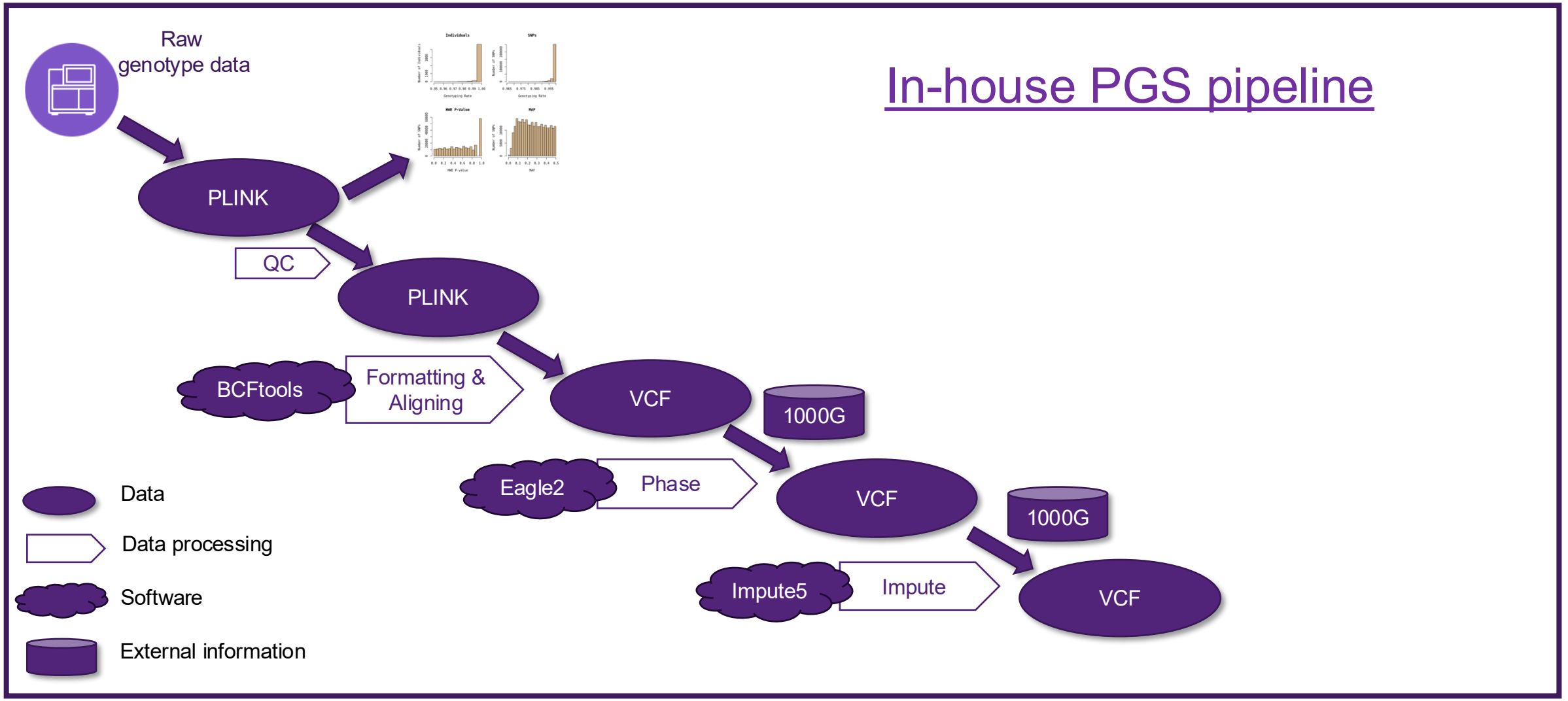


schematic of technical pipeline



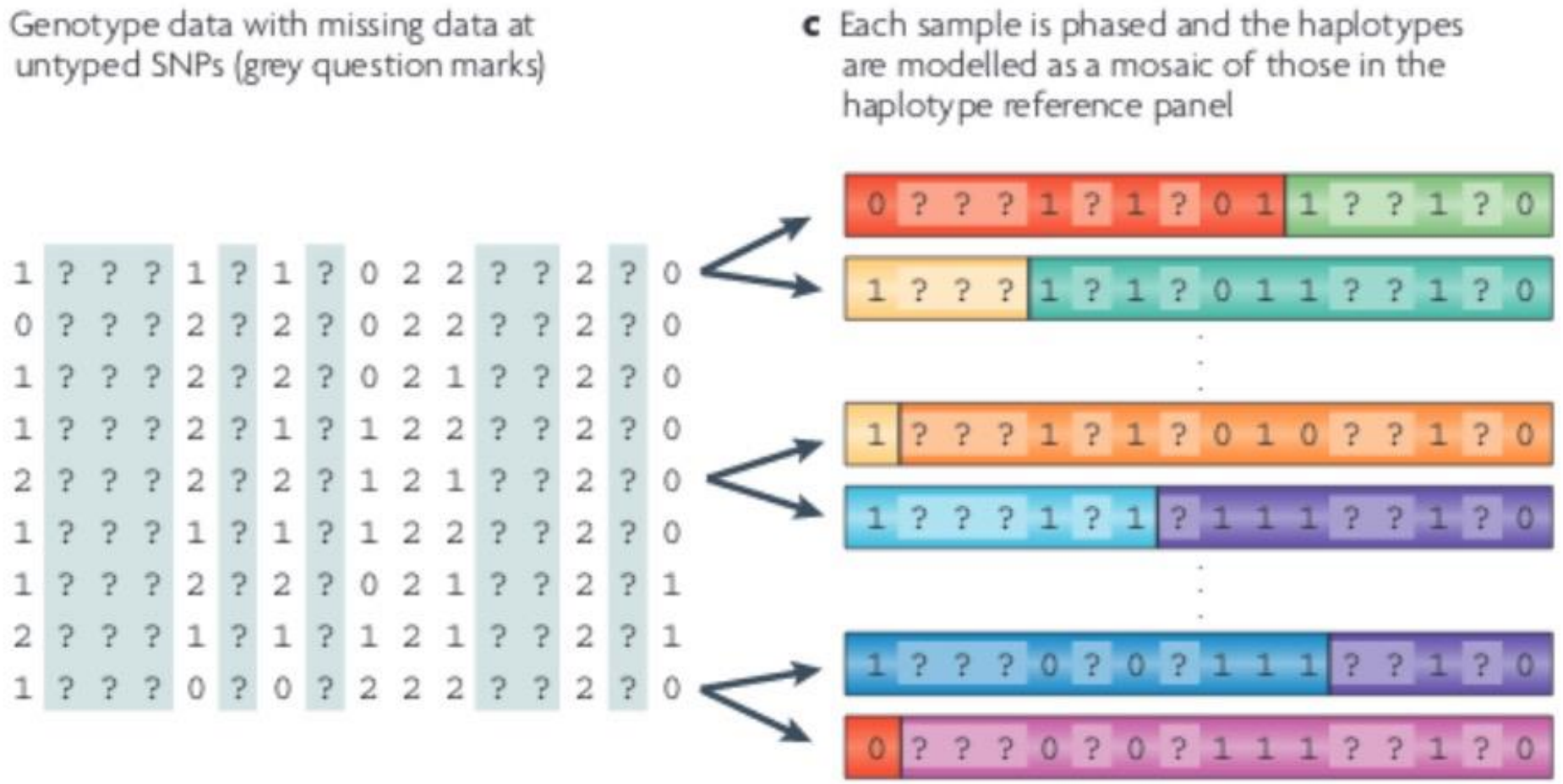
In-house PGS pipeline

schematic of technical pipeline



In-house PGS pipeline

phasing


a Genotype data with missing data at untyped SNPs (grey question marks)
c Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel

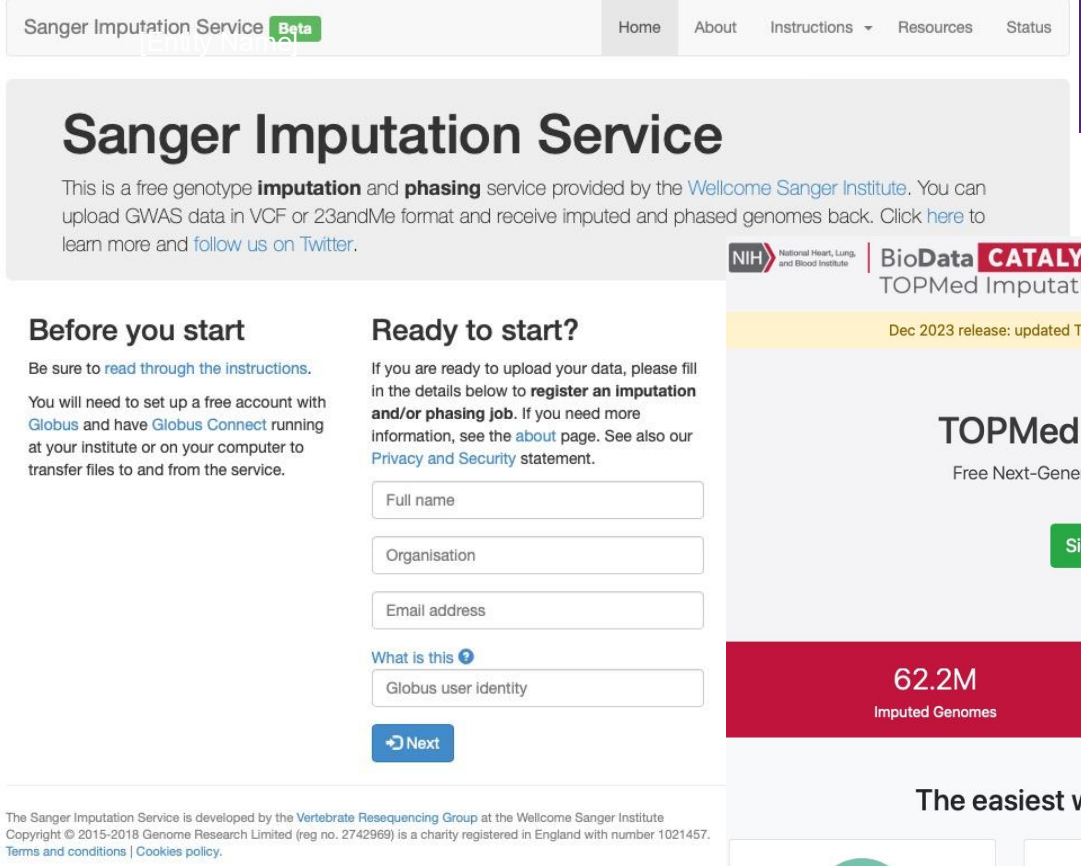
Imputation

Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel

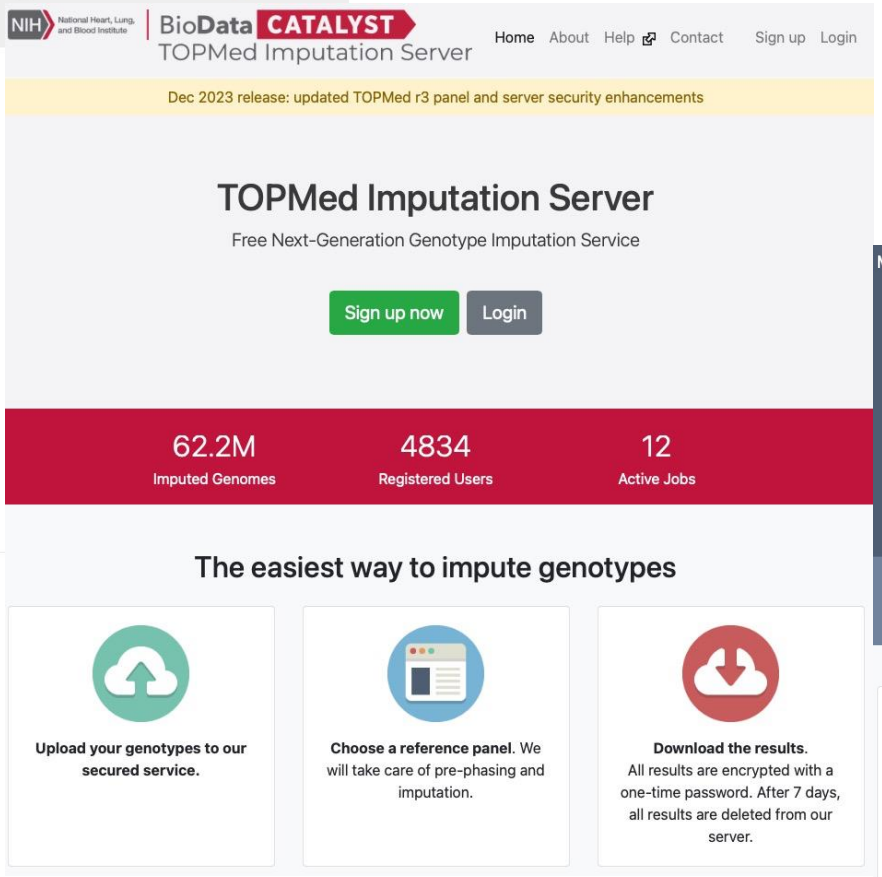
d Reference set of haplotypes, for example, HapMap

Panel options

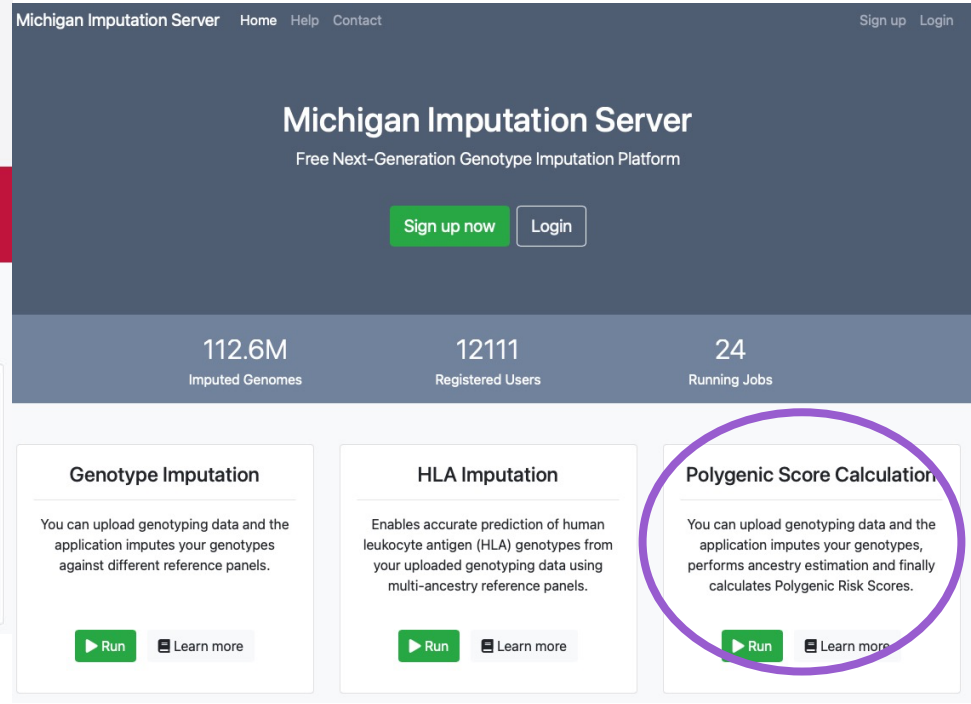
Reference Panel	Number of Individuals	Number of Variants	Population Focus
Haplotype Reference Consortium (HRC)	~32,000	~40 million	European
1000 Genomes Project	2,504	~88 million	Global, diverse
TOPMed	~62,000	>300 million	Diverse, underrepresented
UK10K	~3,800	~30 million	UK, European
GoT2D	~2,657	~20 million	Type 2 Diabetes, Metabolic



Most user friendly



Largest panel



Multiple features

Imputation servers

In house Phasing with Eagle2

```
## Example script
geneticmap=genetic_map_chr${chr}_combined_b37.txt
reference=ALL.chr${chr}.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz

eagle \
 --vcfRef=$reference \
 --vcfTarget=indexed_fixed_${data}_chr${chr}.vcf.gz \
 --geneticMapFile=$geneticmap \
 --vcfOutFormat=z \
 --outPrefix=phased_chr${chr} > phasing.log
```

Alternative: SHAPEIT4

In house Imputation with Impute5

```
## Example script
impute5_1.1.5_static \
 --m $geneticmap \
 --h $reference \
 --g phased_chr${chr}.vcf.gz \
 --r ${chr}:${intstart}-${intend} \
 --ne 20000 \  ## effective sample size, default 10k~20k for human
 --threads 1 \
 --o imputed_chr${chr}_chunk.vcf.gz \
 --l imputed_chr${chr}_chunk.log
```

→ A chromosome can be imputed as chunks

After imputation

- Format

- Imputed data is output as a zipped VCF file. We usually change the format back to PLINK for following analysis.

Example script

```
## convert the format using plink
plink --vcf  imputed_chr${chr}.vcf.gz   \
            --id-delim '_'  \
            --keep-allele-order \
            --make-bed \
            --out  imputed_chr${chr}
```

After imputation

- Format

- Imputed data is output as a zipped VCF file. We usually change the format back to PLINK for following analysis.

- SNP ID

- Plink does not like duplicate and missing IDs.
  - Fill in dbSNP ID if it's not used, as in files from Michigan and TopMed server
  - Replace missing SNP IDs with "chr_pos"
  - Rename duplicate SNP IDs with "_dup"

After imputation

- Format

- Imputed data is output as a zipped VCF file. We usually change the format back to PLINK for following analysis.

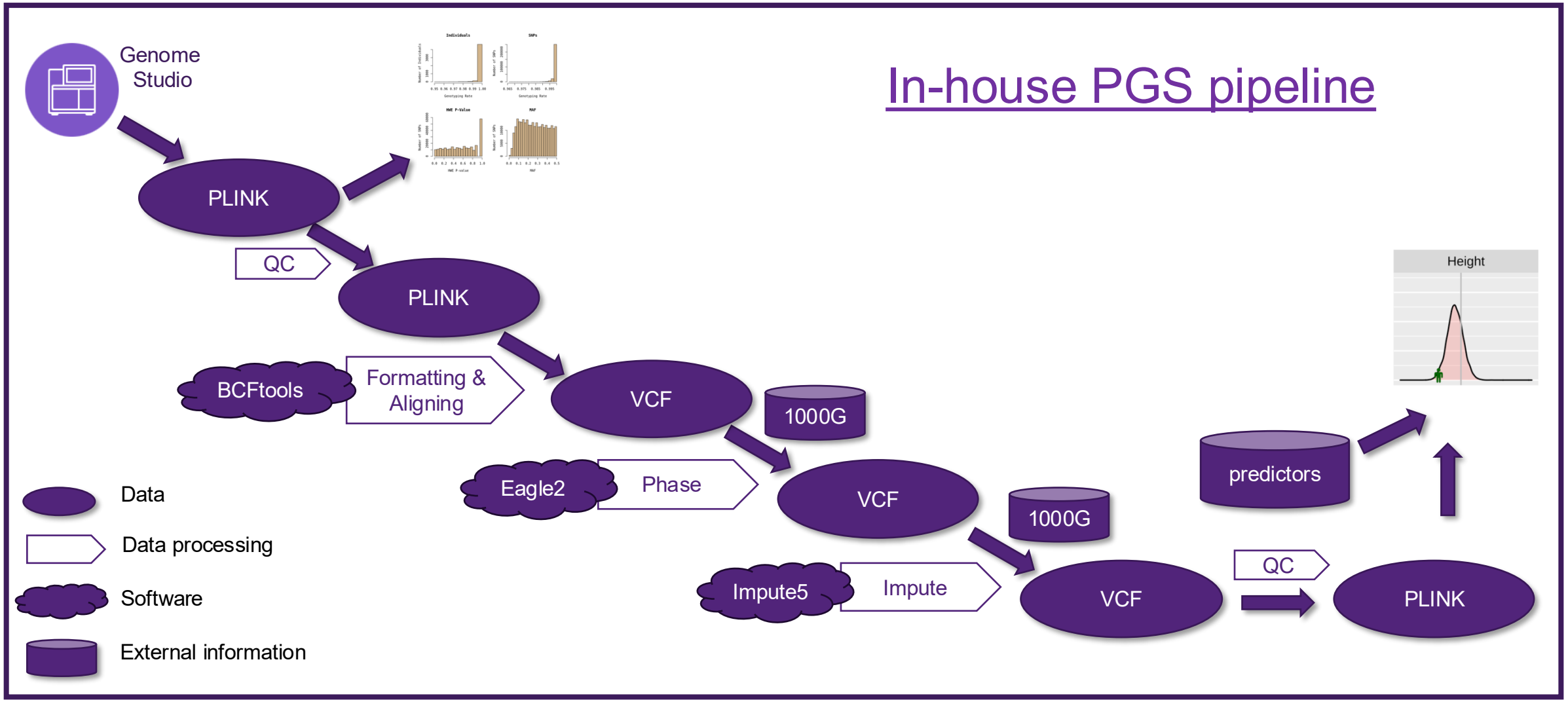
- SNP ID

- Plink does not like duplicate and missing IDs

- Quality

- We suggest to keep all the SNPs regardless of the info score and allele frequency for PGS profiling

schematic of technical pipeline



PGS profiling

```
## Example script
plink \
 --bfile  ${target}_chr${i} \
 --extract overlap_SNPs.txt \
 --score  ${trait}_SBayesRC_predictor.txt 2 5 8 header sum \
 --out ${target}_${trait}_SBayesRC
```

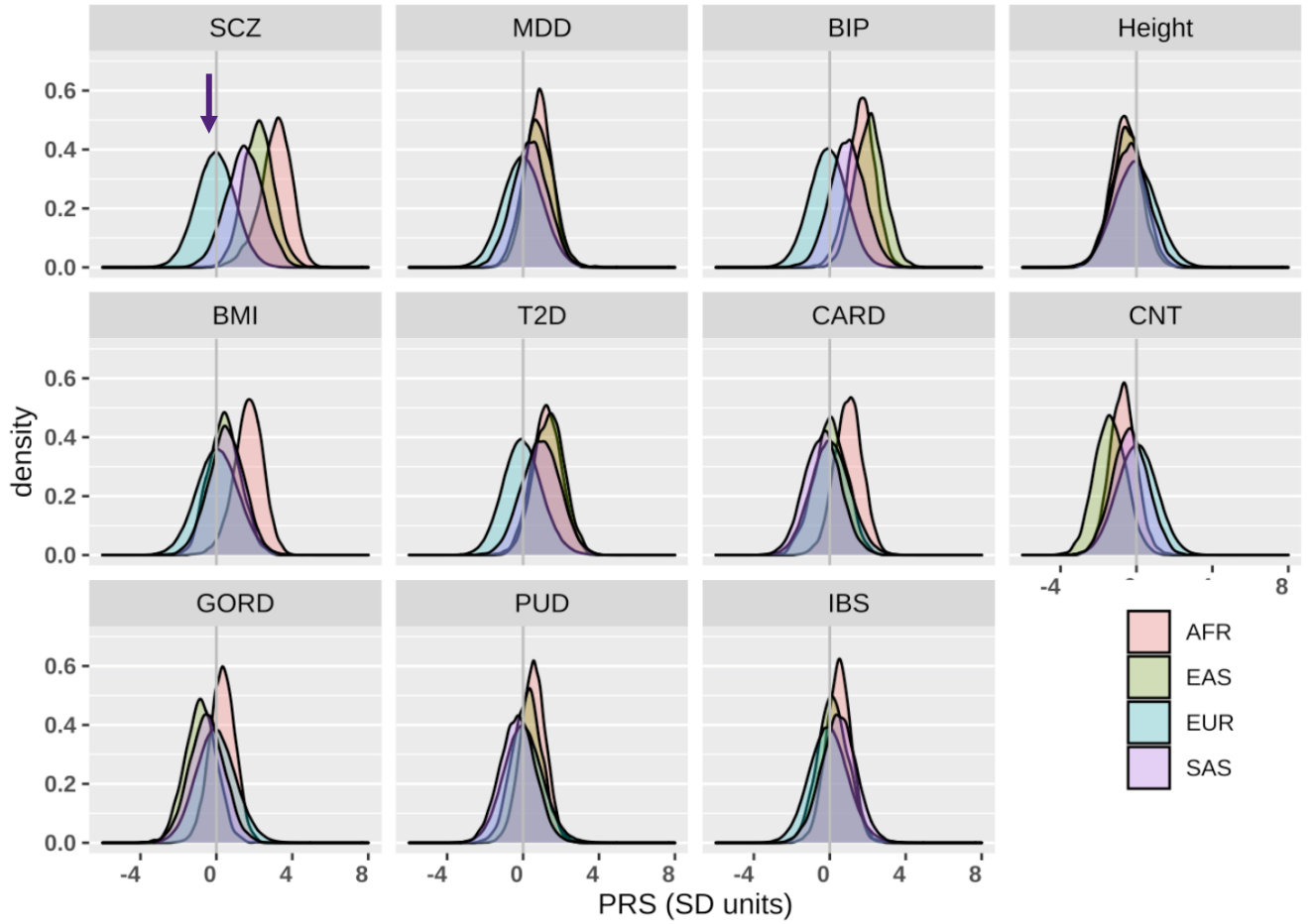
The parameters after your predictor file means

- 2 5 8: Take only the first three columns in the predictor file. The order should be columns of SNP, A1, Effect.
- header: The predictor file has a header row.
- sum: PLINK prefers to divide the score by the number of SNPs in predictor. Using “sum” will prevent the division step.
- It’s suggested to use overlap SNPs if you are going to compare two sets of data.

Interpret PGS

- Case vs. Control?
- Benchmark with population-wise scores

Match ancestry when benchmarking the PGS

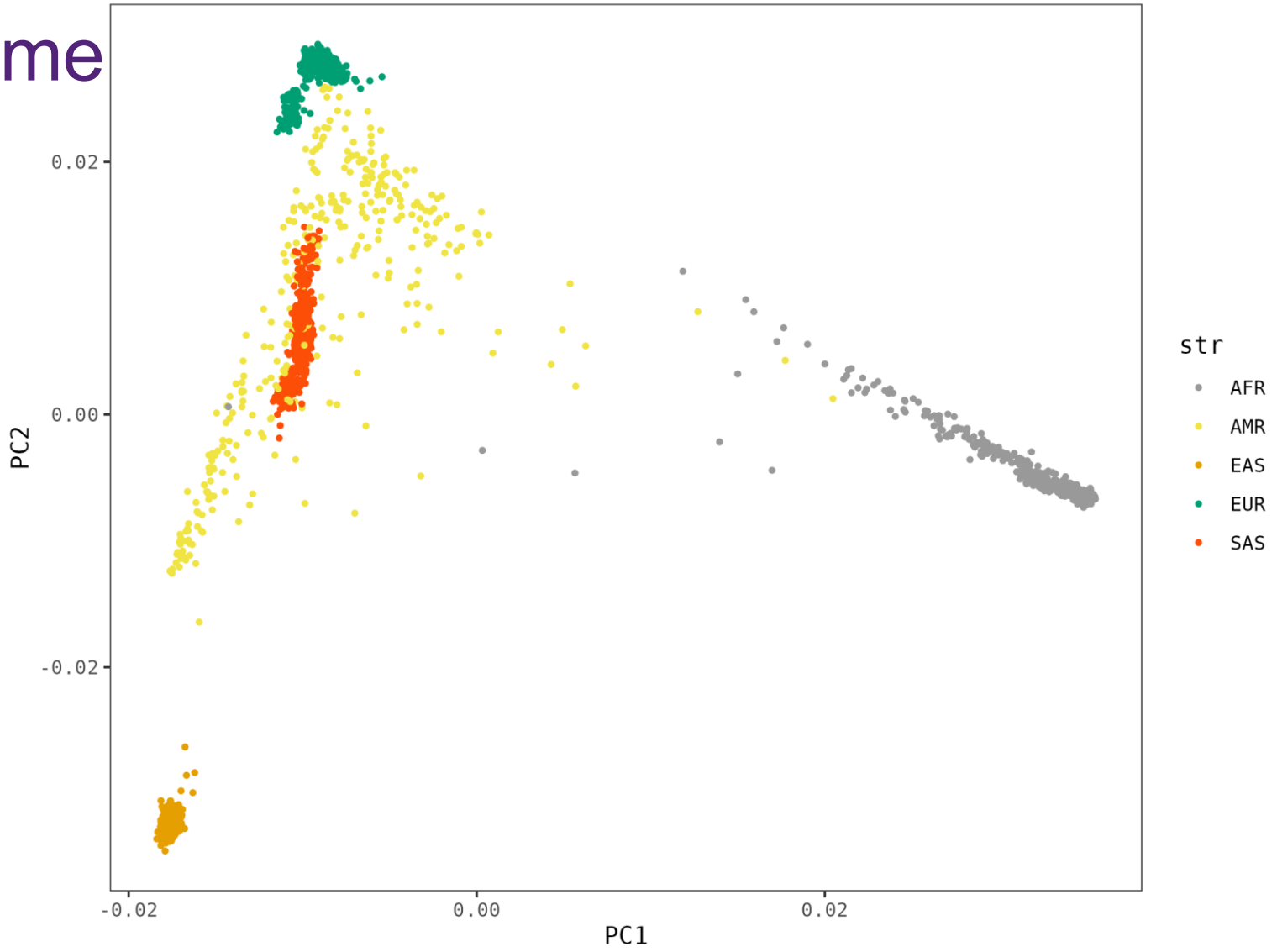


PC calculation using GCTA

- 1000G is the most widely used reference data

```
### generate GRM of reference data
     gcta --bfile ${refpath}/${pcref}.05  \
    --extract common.SNPs.txt \
    --make-grm \
    --out ${pcref}.05.common

### calculate PC of reference data
     gcta --grm     ${pcref}.05.common  --pca 3   --out  ${pcref}.05.common_pca3
```

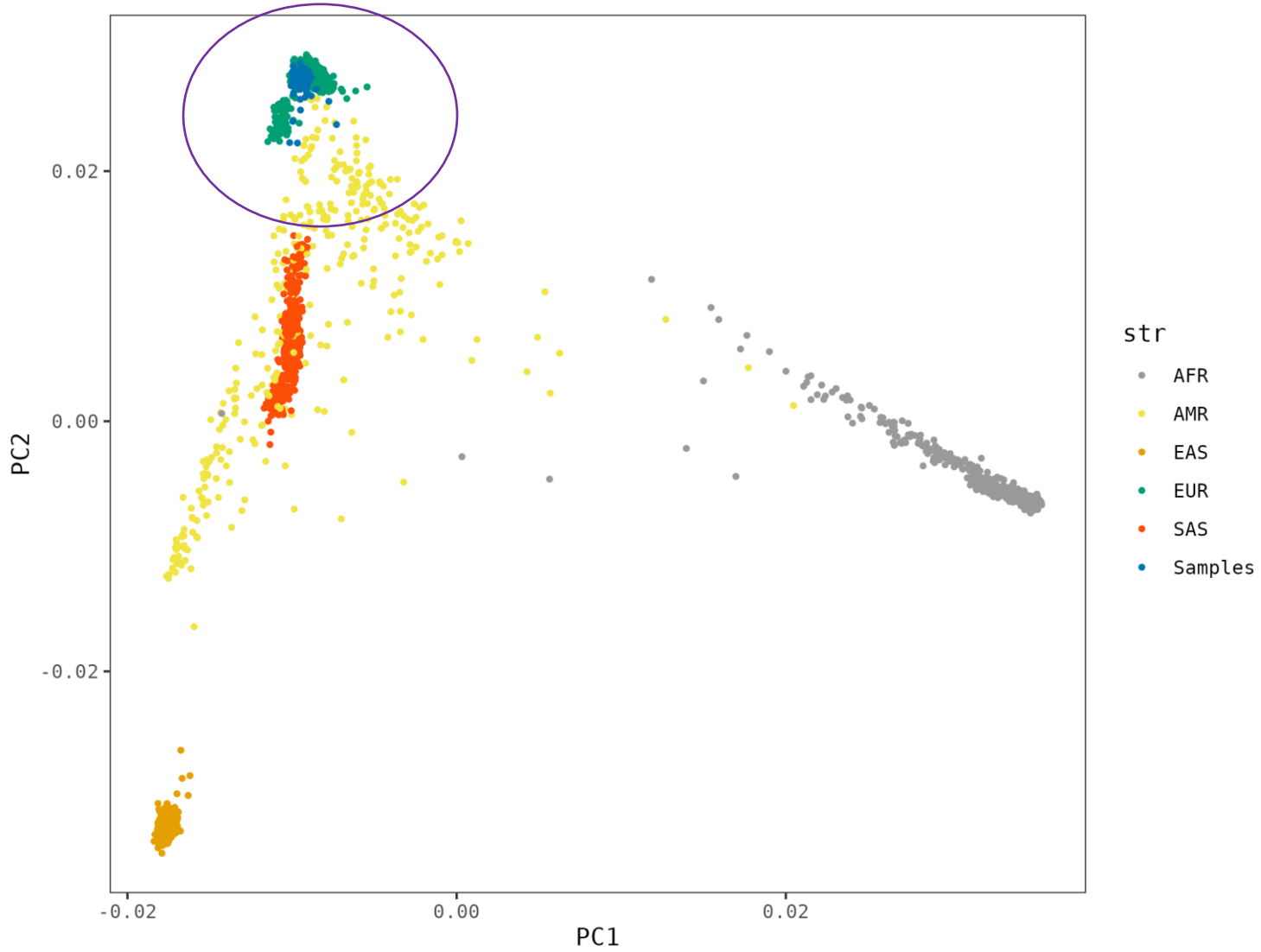
PC plot of 1000Genome



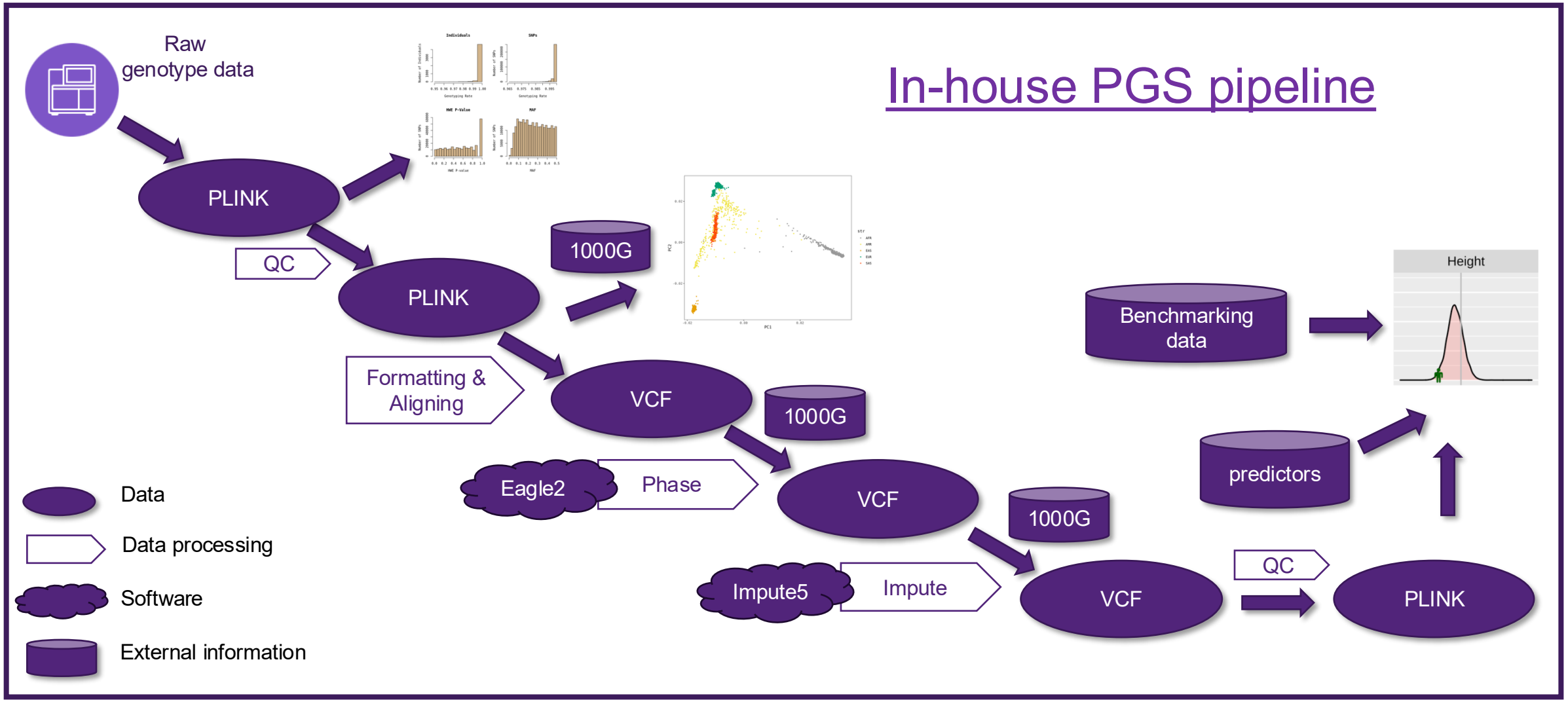
PC projection using GCTA

```
### PC loading
     gcta \
    --bfile  ${refpath}/${pcref}.05    \
    --extract  common.SNPs.txt  \
    --pc-loading   ${pcref}.05.common_pca3   \
    --out  ${pcref}.05.common_pca3_snp_loading

### PC projection
     gcta  \
    --bfile   ${data}  \
    --extract  common.SNPs.txt  \
    --project-loading   ${pcref}.05.common_pca3_snp_loading  3  \
    --out  ${data}_05.common_pca3
```

PC projection



schematic of technical pipeline



Questions and Wrap Up

Polygenic scores

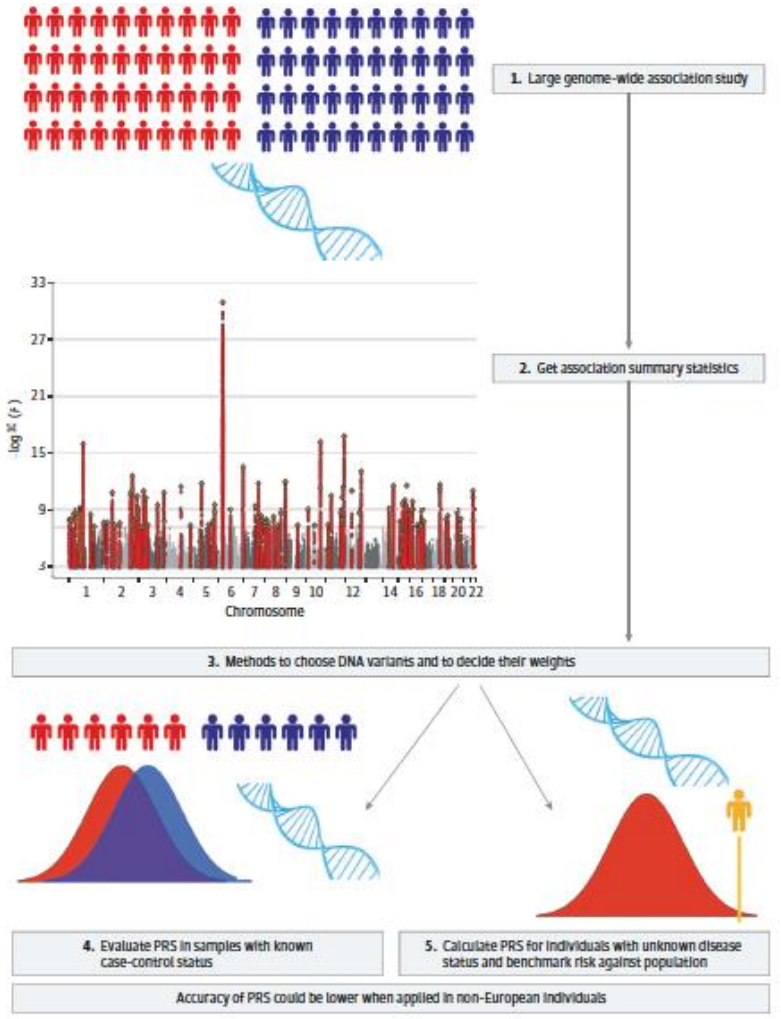
Polygenic score (PGS) is a weighted count of risk alleles

$$PGS = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

0, 1 or 2
Risk alleles

Which SNPs?

What weights?



- Don't need to know causal variants for prediction!
- Prediction can be based on correlated variants.

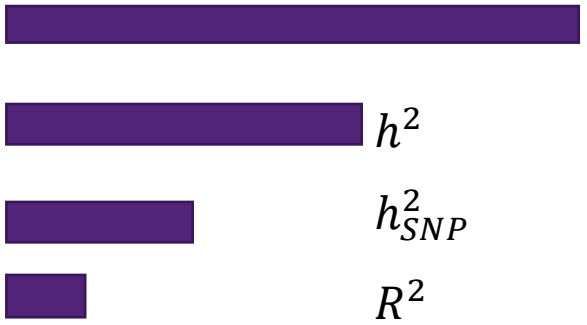
Limitations in prediction accuracy

- PGS have a **theoretical** upper limit dependent on the **heritability of the trait** (how much of the variance of trait values between people is attributed to genetic factors).

- PGS have a **technical** upper limit associated with the proportion of **variance tagged** by the DNA variants measured.

- PGS have a **practical** upper limit dependent on the **sample size of the discovery sample** used to estimate effect sizes of risk alleles, and the **quality** of the discovery sample.

- PGS can be pushed closer to the technical upper limit by the **statistical methodology** used to generate the optimal weighting given to the risk alleles, and new methods integrate new biological data.



Schizophrenia

Max:
25% Liability
AUC 0.84

Current:
11% Liability
AUC 0.74

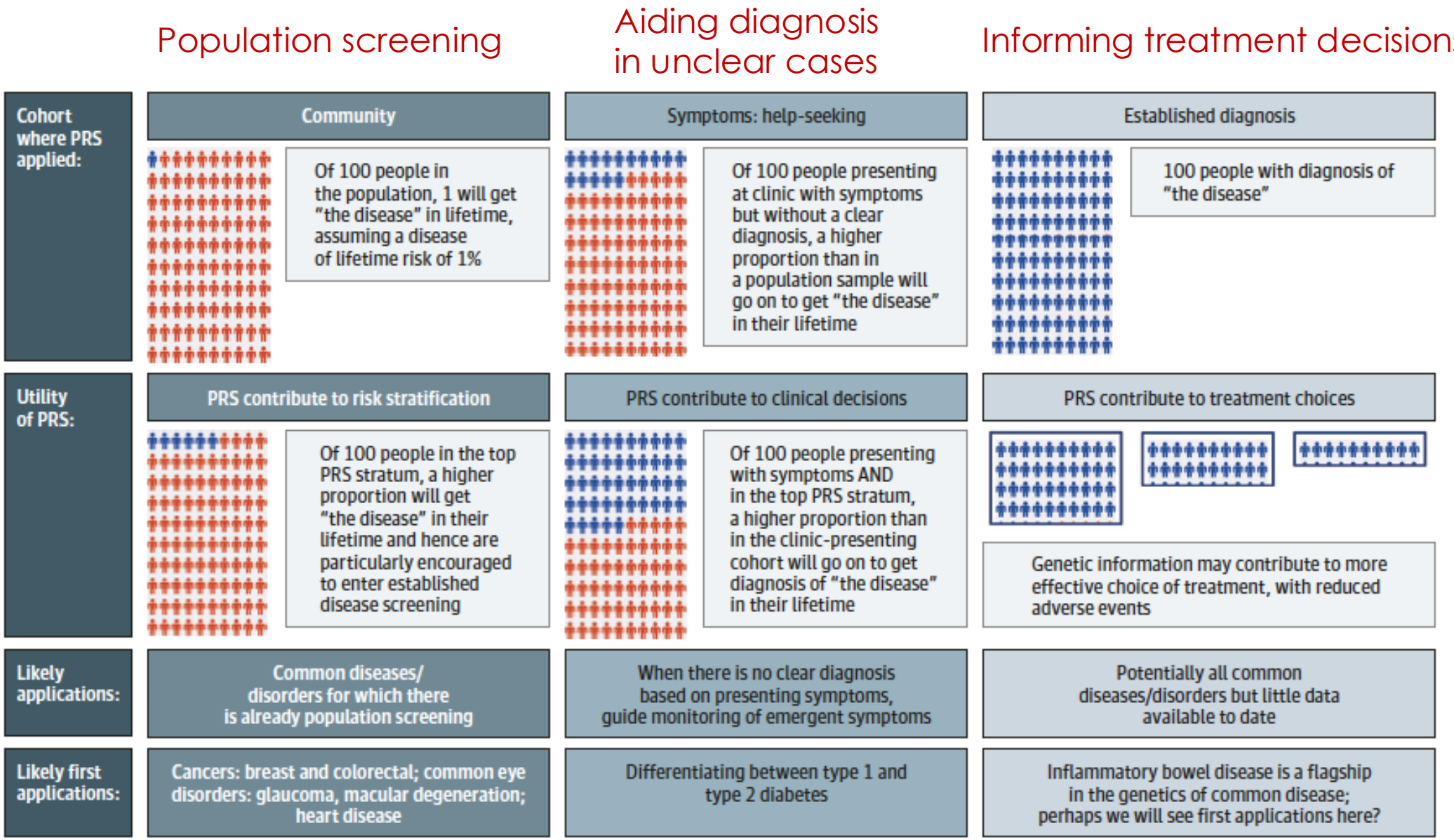
Polygenic scores cannot be highly accurate predictors of phenotypes

Clinical applications

JAMA Psychiatry | Review

From Basic Science to Clinical Application of Polygenic Risk Scores
A Primer

Naomi R. Wray, PhD; Tian Lin, PhD; Jehannine Austin, PhD; John J. McGrath, MD, PhD; Ian B. Hickie, MD; Graham K. Murray, MD, PhD; Peter M. Visscher, PhD



Two solutions

Clumping and P-value thresholding (C+PT)

Include only the most strongly associated SNP from each LD block (Purcell et al., 2009)

$$\sum_{j=1}^{n_{SNP}} \hat{\beta}_j x_{ij}$$

Whole-genome regression approaches

Include all SNPs but adjust the effect sizes for LD

Weights: Set equal to GWAS coefficients.

Loci: Selected by

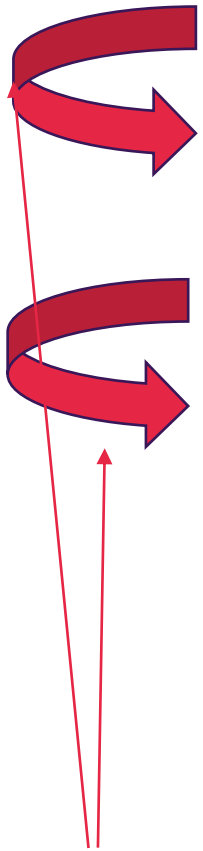
1. using a **clumping** algorithm that ensures the included SNPs are all approximately independent of each other
2. omitting SNPs whose *P* value for association with the phenotype is above a certain **threshold**

Weights: Set to GWAS coefficients **adjusted for LD** → from a random-effect model regressing the phenotype on all SNPs

Loci: Include **all SNPs**, no LD-based pruning

Examples: BLUP (Meuwissen et al. 2001), LDpred (Vilhjalmsson et al. 2015, Prive et al. 2020), PRS-CS (Ge et al. 2019), SBayesR (Lloyd-Jones et al. 2019)

Polygenic prediction



- **<u>Discovery</u>/Training/Derivation**
  - Estimate the effect sizes ($\hat{b}$) of SNPs on a trait (y) – GWAS
- **<u>Tunning</u>/Validation**
  - Further estimate some parameters (depends on methods; not all methods require it)
- **<u>Target</u>/Testing/Validation**
  - Build a polygenetic risk score (PRS) ($\hat{y}$):
  - Evaluate the prediction performance/accuracy

Should be independent; no overlap; out-of-sample prediction

GENOME-WIDE ASSOCIATION STUDIES — OPINION

Pitfalls of predicting complex traits from SNPs

Naomi R. Wray, Jian Yang, Ben J. Hayes, Alkes L. Price, Michael E. Goddard and Peter M. Visscher

More advanced methods

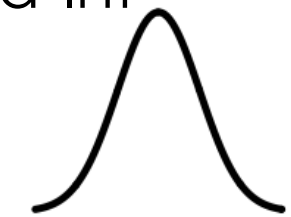
A weighted sum of the count of risk alleles

$$\text{PRS} = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

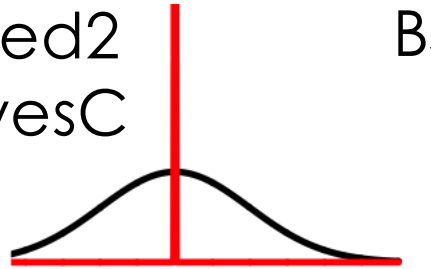
How many SNPs?
Which SNPs?
What weights?

New methods model genetic architecture

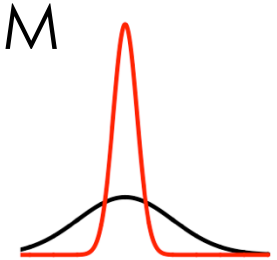
LDpred-Inf
SBLUP



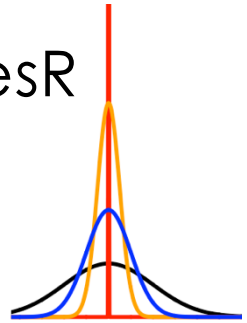
LDPred2
SBayesC



BSLMM



SBayesR



Best linear unbiased prediction (BLUP)

Linear mixed model

$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

BLUP solutions

$$\begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n}'\mathbf{1_n} & \mathbf{1_n'X} \\ \mathbf{X'1_n} & \mathbf{X'X} + \mathbf{I}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \end{bmatrix}$$

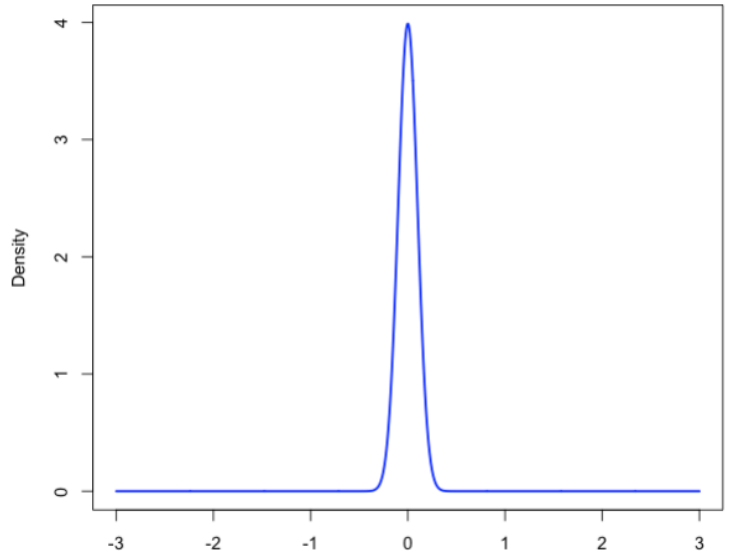
I = identity matrix (dimensions *m* x *m*)

$\lambda = \sigma_e^2 \,/\, \sigma_\beta^2$

- Need to specify the shrinkage parameter λ

- Assumes SNPs effects are (*infinitesimal* model):
  - all non-zero
  - very small
  - normally distributed

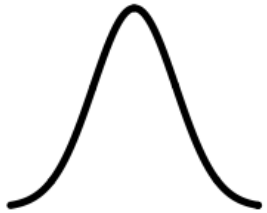
How realistic they are?



Bayesian methods in human genetics

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

PRS-CS



Continuous shrinkage

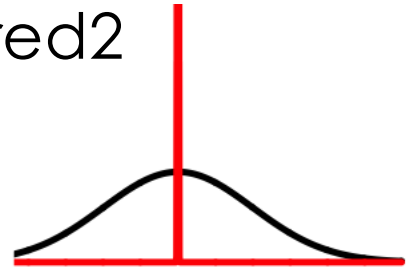
$$\beta_j \sim N(0, \phi\psi_j)$$

$$\psi_j \sim Ga(a, \delta_j)$$

$$\delta_j \sim Ga(b, 1)$$

Parameters a and b determine how aggressively to shrink small estimates and how much you don't shrink large ones

LDPred2

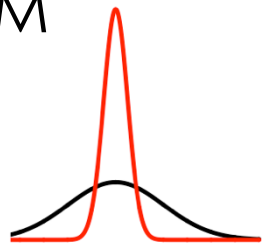


Spike-and-Slab

$$\beta_j \sim \begin{cases} N(0, \tau^2), & \pi \\ 0 & 1-\pi \end{cases}$$

π can be estimated from data, sparsity allowed, $\tau^2 = h^2/M\pi$

BSLMM

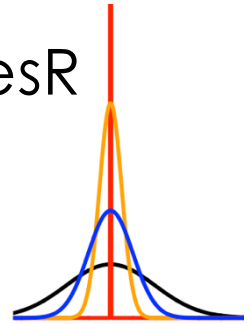


Normal mixture

$$\beta_j \sim \begin{cases} N(0, \sigma_b^2 + \sigma_u^2), & \pi \\ N(0, \sigma_u^2), & 1-\pi \end{cases}$$

σ_b^2 captures "sparse effects" as in the spike-and-slab model, and σ_u^2 captures small polygenic effects.

SBayesR



Flexible finite mixture of normal distributions, sparsity allowed

$$\beta_j \sim \begin{cases} 0, & \pi_1 \\ N(0, \gamma_2\sigma_b^2), & \pi_2 \\ \cdots & \\ N(0, \gamma_C\sigma_b^2), & 1-\sum_{c=1}^{C-1}\pi_c \end{cases}$$

Bayes theorem

$$P(x \mid y) \propto P(y \mid x)P(x)$$

Probability of parameters x given the data y (posterior)

Is proportional to

Probability of data y given the x (likelihood of data)

Prior probability of x

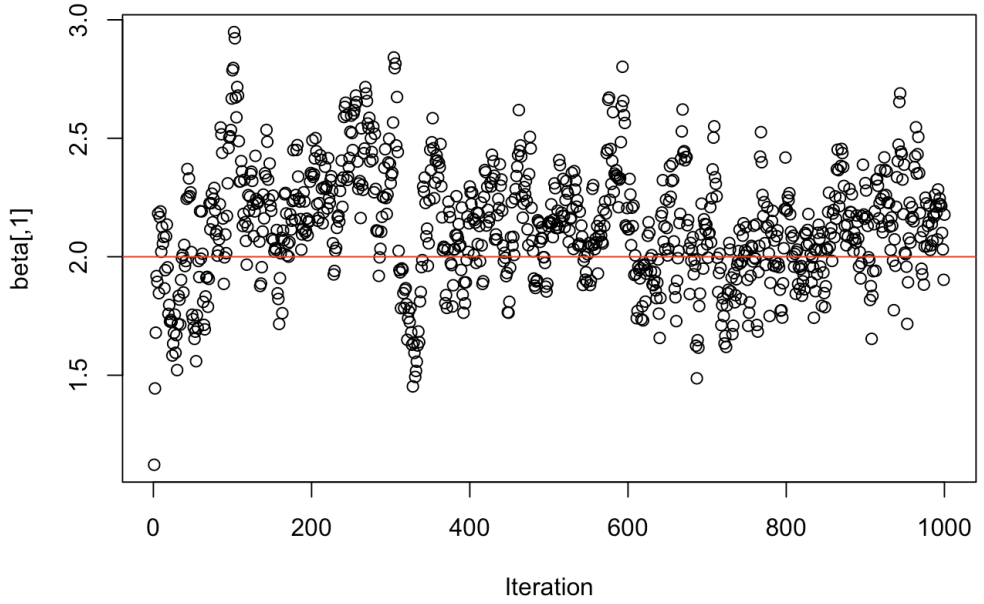
Gibbs sampling

- Set starting values for $(\mu, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma_\beta^2, \pi, \sigma_e^2)$
- Then (for many iterations)
  - For each SNP, sample δ_j, β_j conditional on other parameters
  - Sample $\mu, \sigma_\beta^2, \pi, \sigma_e^2$ with updated $\boldsymbol{\delta}, \boldsymbol{\beta}$

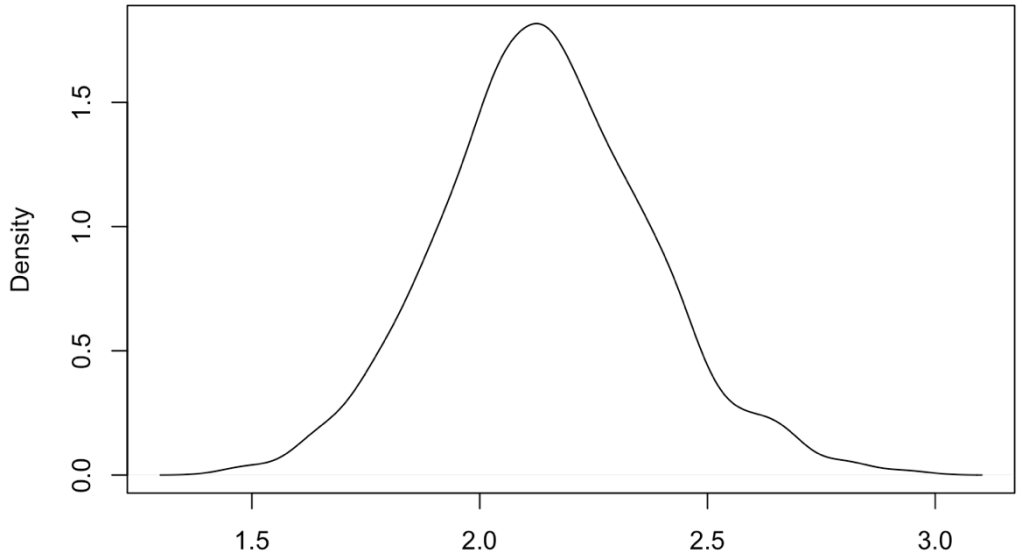
Samples reconstruct posterior distributions of parameters

Gibbs sampling

Trace plot

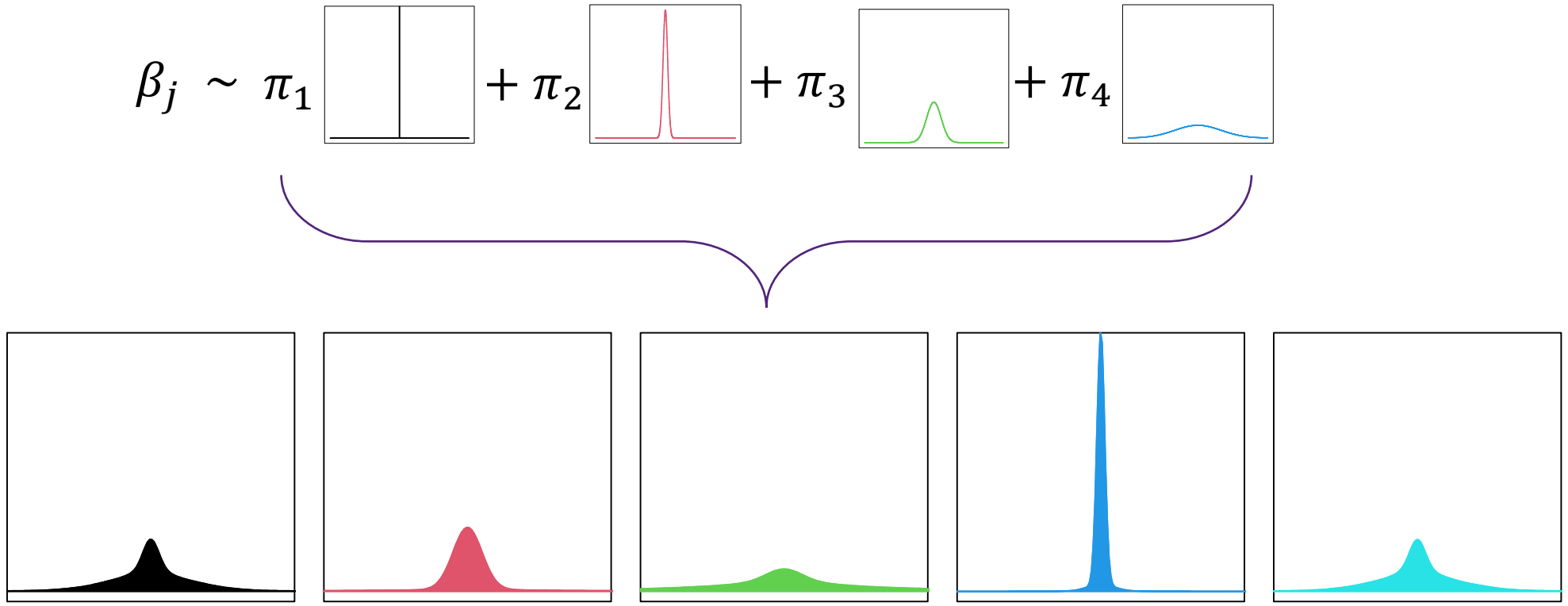


Posterior distribution



Posterior mean is used as the point estimate of the SNP effect

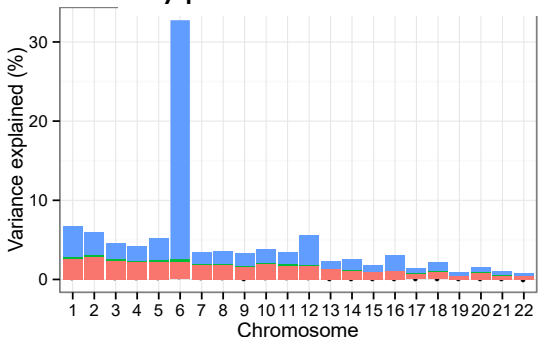
Why use multi-normal mixture?



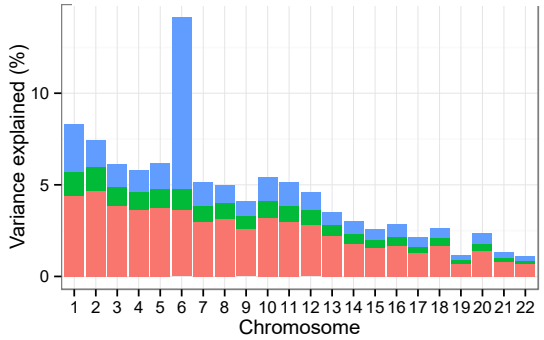
Account for almost any distribution!

Many polygenic genetic architectures

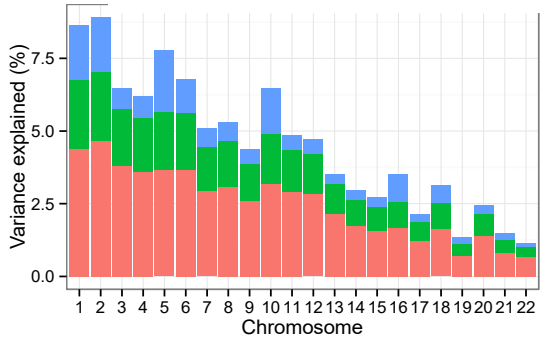
Type 1 Diabetes



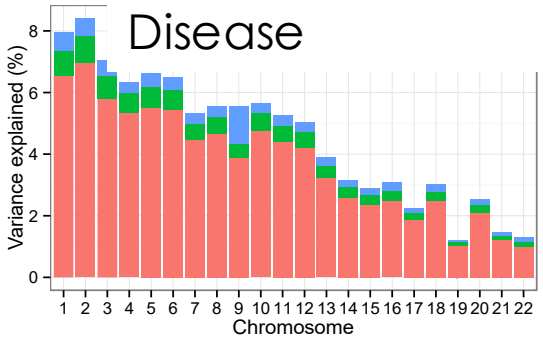
Rheumatoid Arthritis



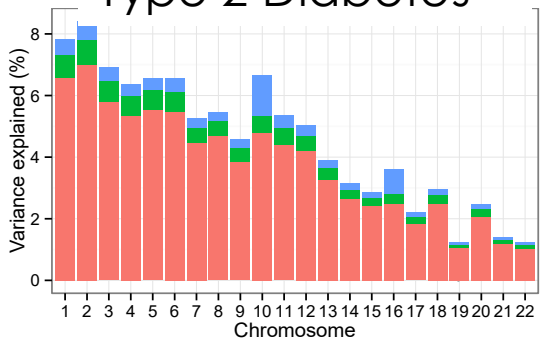
Crohn's Diseasse



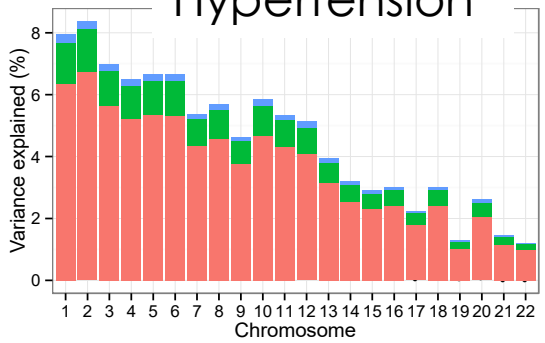
Coronary Artery Disease



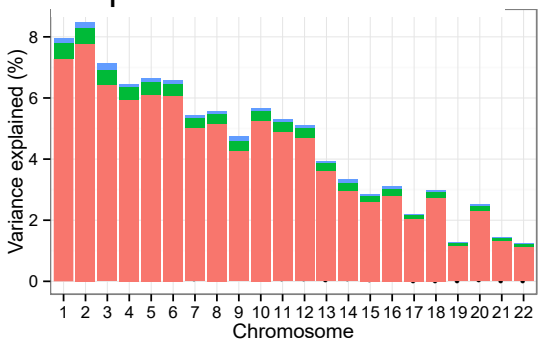
Type 2 Diabetes



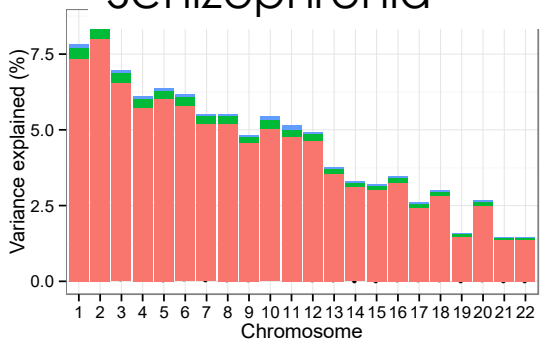
Hypertension



Bipolar Disorder



Schizophrenia



Mixture component

$10^{-4} \times \sigma^2_g$ small

$10^{-3} \times \sigma^2_g$ middle

$10^{-2} \times \sigma^2_g$ big

PLOS | GENETICS

RESEARCH ARTICLE

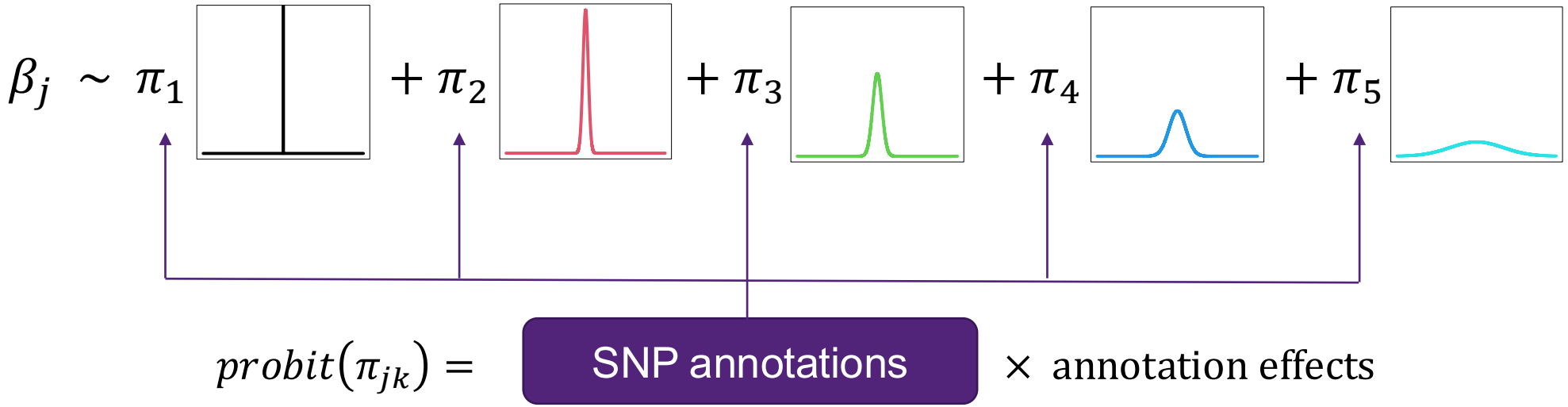
Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model

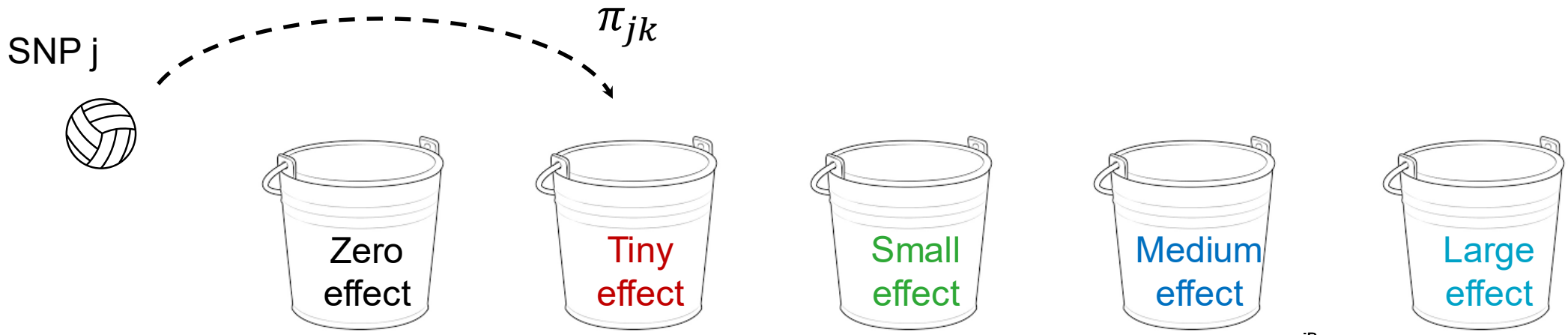
Gerhard Moser[1]*, Sang Hong Lee[1], Ben J. Hayes[2,3], Michael E. Goddard[2,4], Naomi R. Wray[1], Peter M. Visscher[1,5]

Many DNA variants contribute to genetic risk, and most have very small effects.

SBayesRC

Incorporate functional annotations through a hierarchical prior:



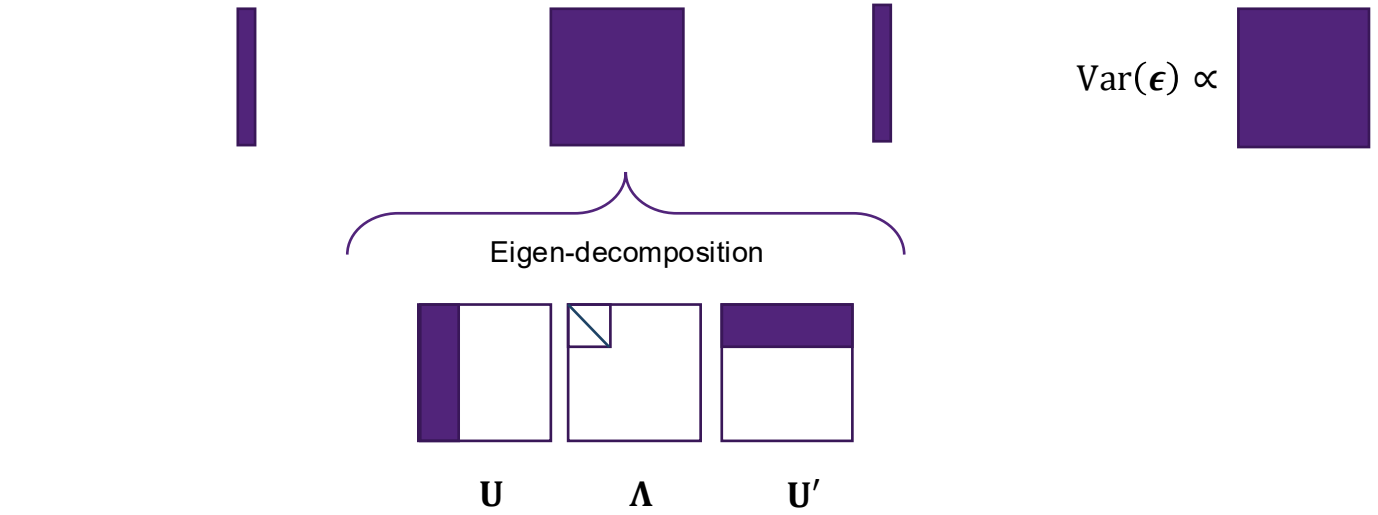


Low-rank model (fits 7M SNPs or more)

In each quasi-independent LD block:

$$\mathbf{b} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

GWAS SNP marginal effects | LD correlation matrix | SNP joint effects | Residuals

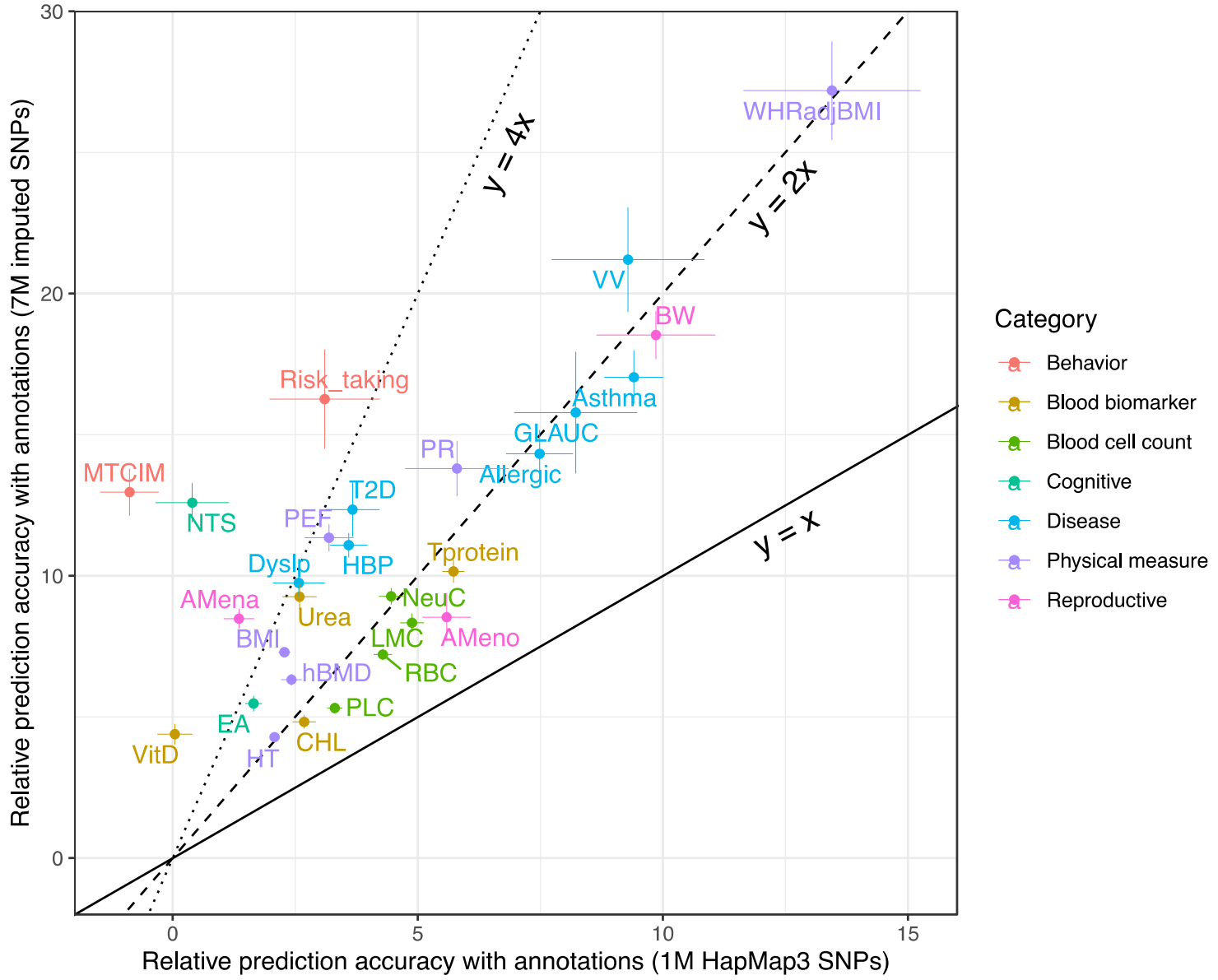


$$\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\mathbf{b} = \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}'\boldsymbol{\beta} + \boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\boldsymbol{\epsilon}$$

$$\mathbf{w} = \mathbf{Q}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

It only requires the top 20% PCs to explain 99.5% of the variance in LD!

Interaction between SNP density and annotation information



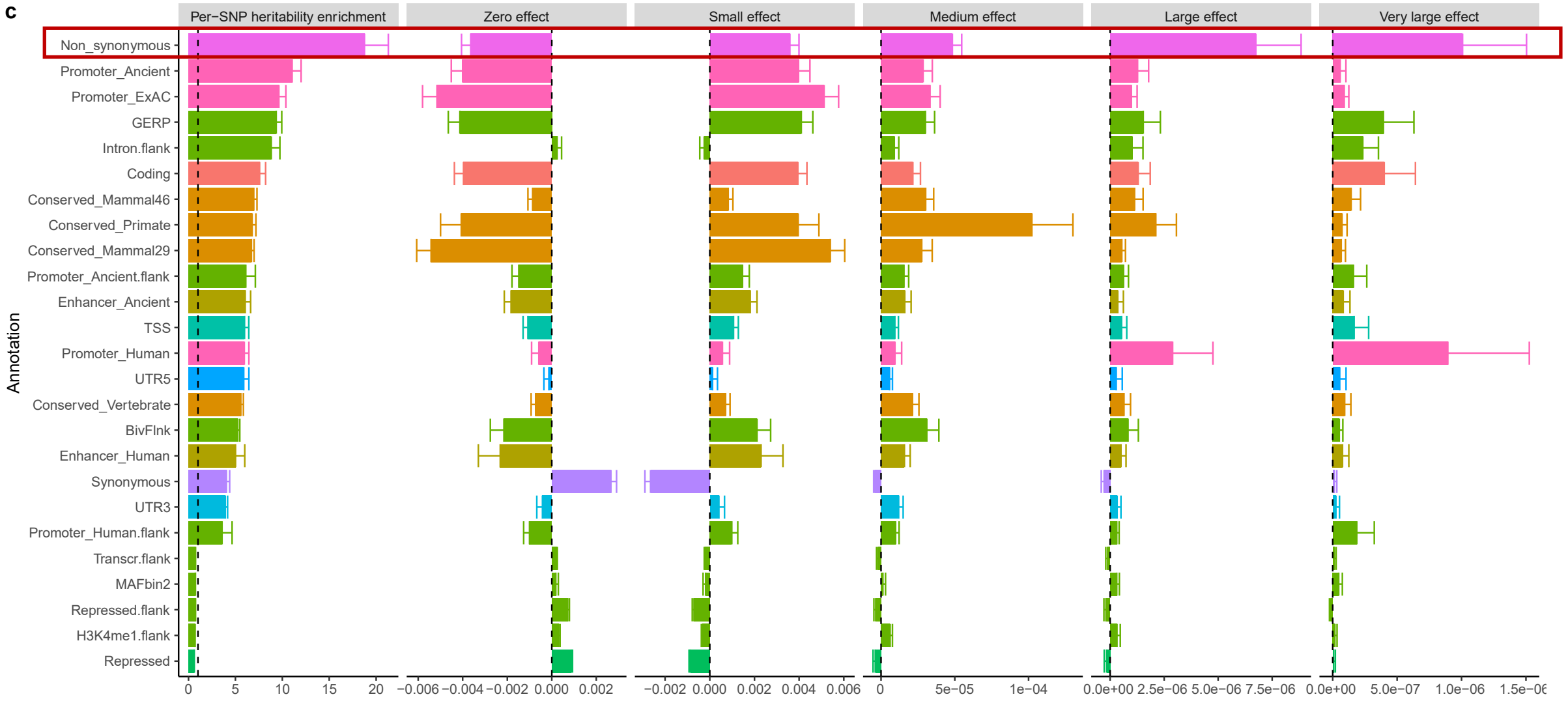
Improvement (%) in prediction accuracy with vs. without annotations:

$$\frac{R^2_{\text{annot}} - R^2_{\text{wo}}}{R^2_{\text{wo}}}$$

using 7M imputed SNPs (y-axis) or 1M HapMap3 SNPs (x-axis).

Annotations help more with increased SNP density

Functional genetic architecture



PGS evaluation in diseases (binary traits)

Statistics to measure prediction accuracy

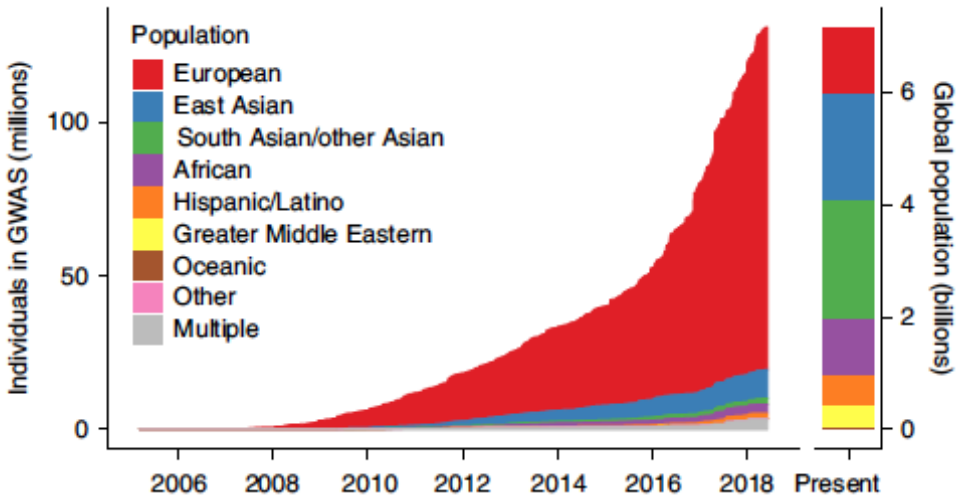
- Pseudo R^2 from logistic regression
- AUC (area under the ROC curve)
- Variance explained on liability scale
- Decile odds ratio (OR)
- Risk stratification

Ancestry

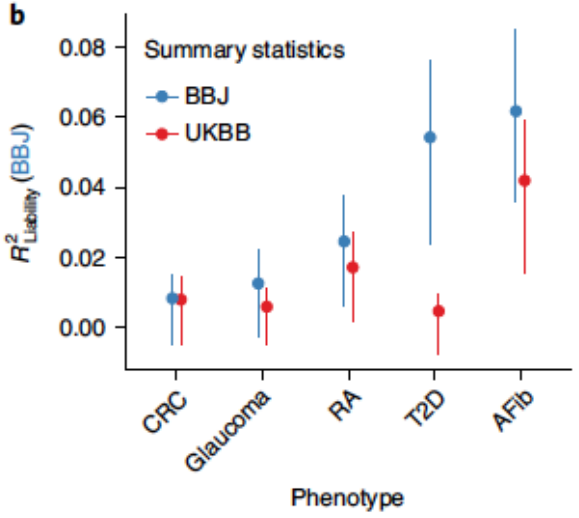
PERSPECTIVE
https://doi.org/10.1038/s41588-019-0379-x
nature genetics

Clinical use of current polygenic risk scores may exacerbate health disparities

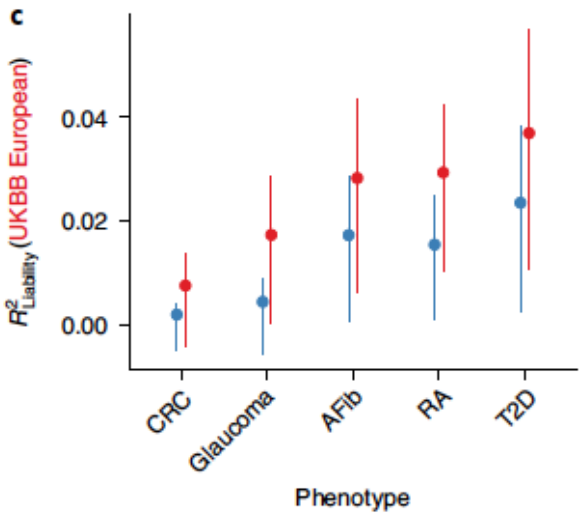
Alicia R. Martin[1,2,3*], Masahiro Kanai[1,2,3,4,5], Yoichiro Kamatani[5,6], Yukinori Okada[5,7,8], Benjamin M. Neale[1,2,3] and Mark J. Daly[1,2,3,9]

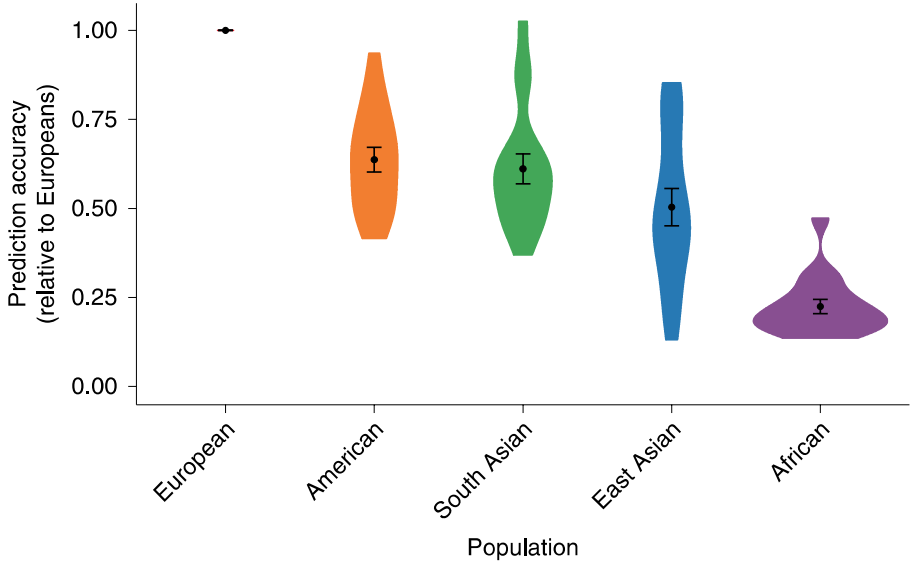


Predicted into Japanese

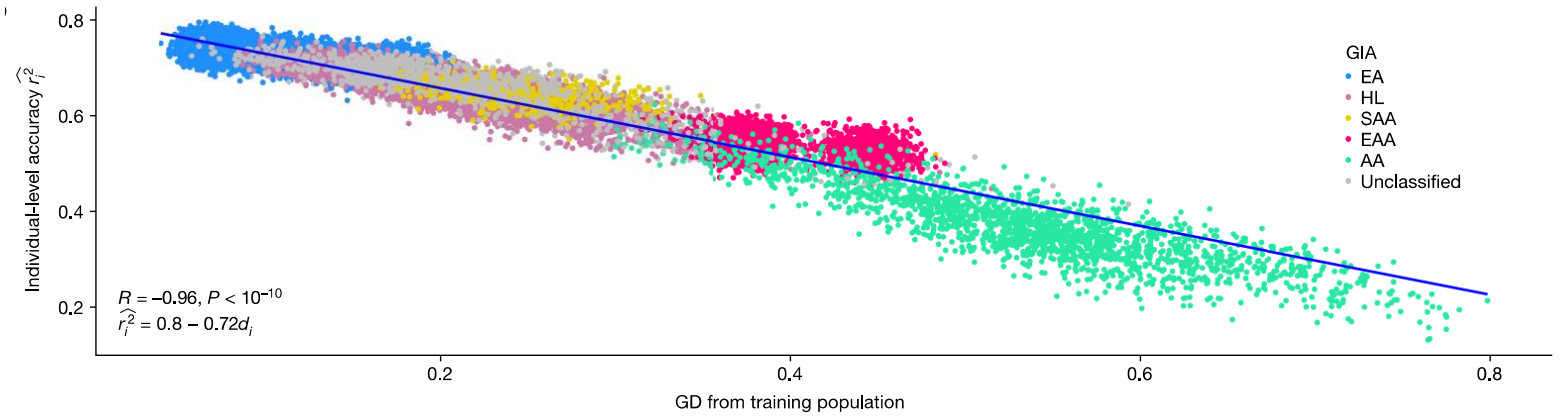


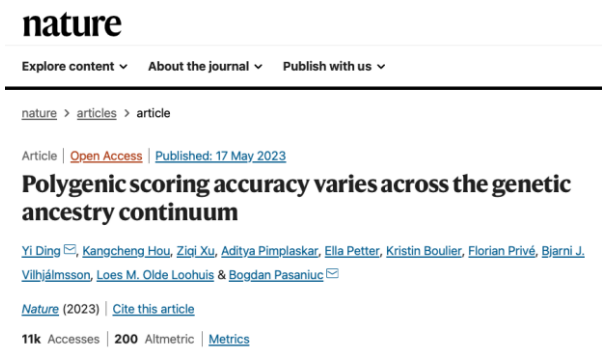
Predicted into European





Ancestry



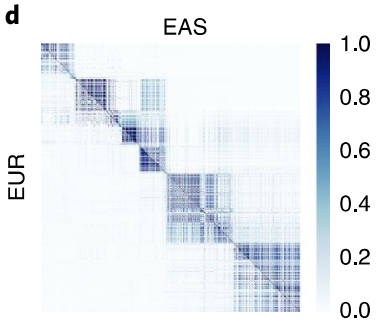


Issues

- Same causal variants
  - Different allele frequencies
  - LD differences
  - Different effect sizes
- Different causal variants
  - GxE
  - Different phenotype

In *general:*
We expect common causal variants to be shared across ancestries

But correlation structure differs



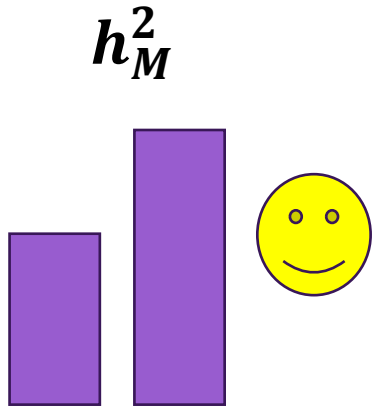
How about whole genome sequencing?

Maximum depends on maximising h_m^2

We use GWAS data so the maximum h_m^2 is the SNP-based heritability

Theoretical maximum depends on the heritability of the trait

$$R^2 \approx \frac{h_m^2}{1 + \frac{m}{Nh_m^2}}$$

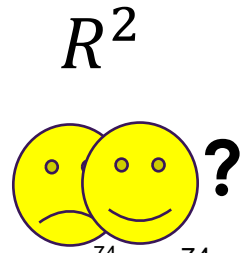


With whole genome sequencing the variance captured by all measured SNPs will increase

But the number of SNPs that we have estimate effect sizes for increases much more

Need MASSIVE discovery sample sizes for WGS associations

Also… rare variants are less likely to be shared across populations



Wray et al (2019) Complex trait prediction from genome data. Genetics

74