



THE UNIVERSITY  
OF QUEENSLAND  
AUSTRALIA

CREATE CHANGE

# GWAS summary statistics

*Genetics & Genomics Winter School*  
*Module 5*

# 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

Perspective

# Workshop proceedings: GWAS summary statistics standards and sharing

2021

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

**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 <sup>a</sup>
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

## Genome-wide association studies

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

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

Database	Content
GWAS Catalog <sup>110</sup>	GWAS summary statistics and GWAS lead SNPs reported in GWAS papers
GeneAtlas <sup>8</sup>	UK Biobank GWAS summary statistics
Pan UKBB	UK Biobank GWAS summary statistics
GWAS Atlas <sup>273</sup>	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.<sup>13</sup>. 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
```

## What are the minimum data required?

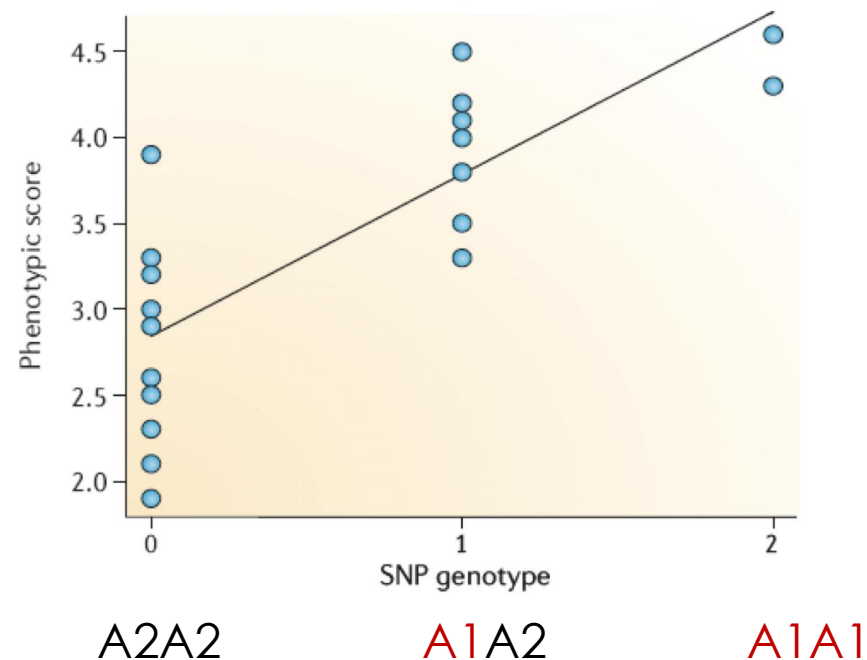
Given the standard GWAS with genotypes being allelic counts (0/1/2), the minimum data required include:

- SNP marginal effect estimates
  - Standard errors
  - GWAS sample size
- 
- GWAS sumstats
- 
- LD correlations among SNPs
- 
- LD matrix

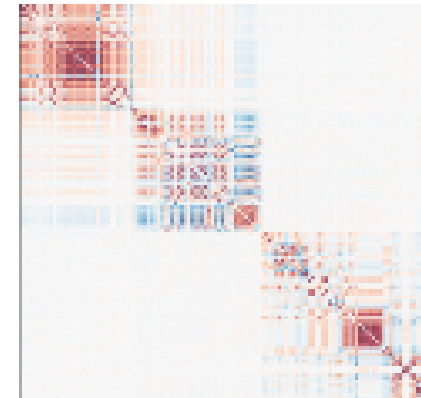
## Other information critical to quality control (QC)

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

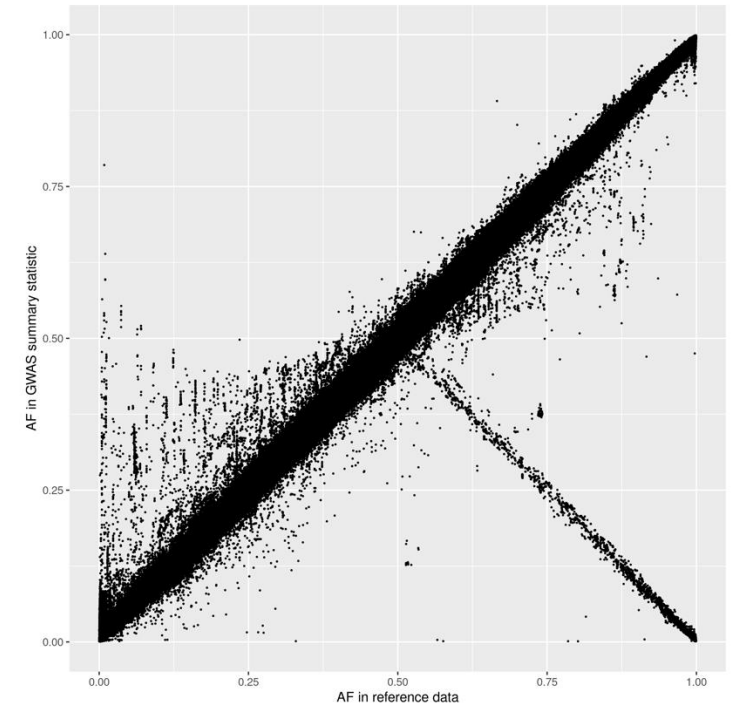
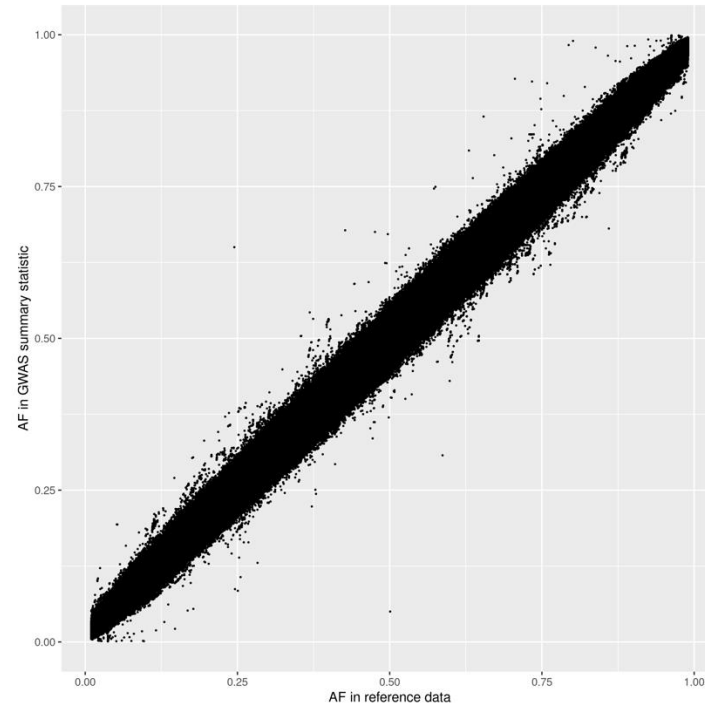
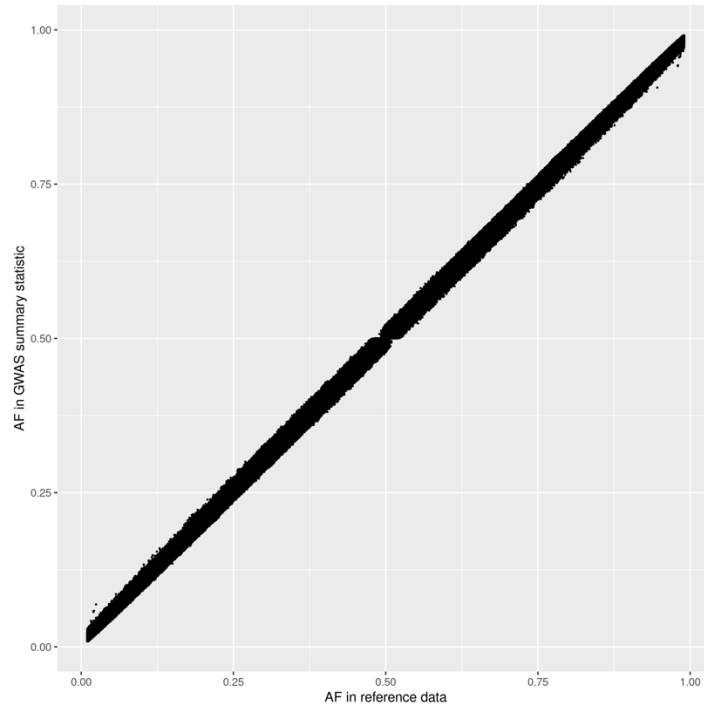
e.g., A1 allele



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)

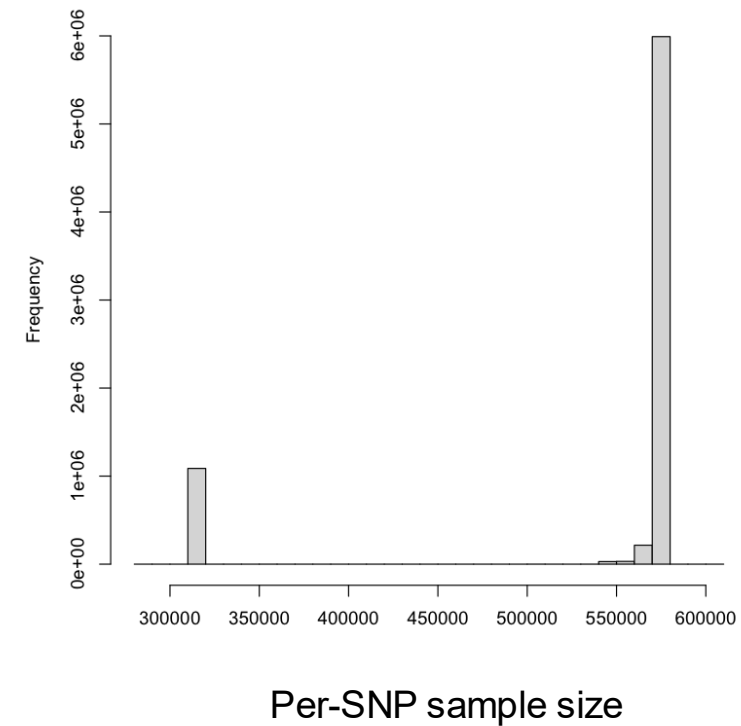
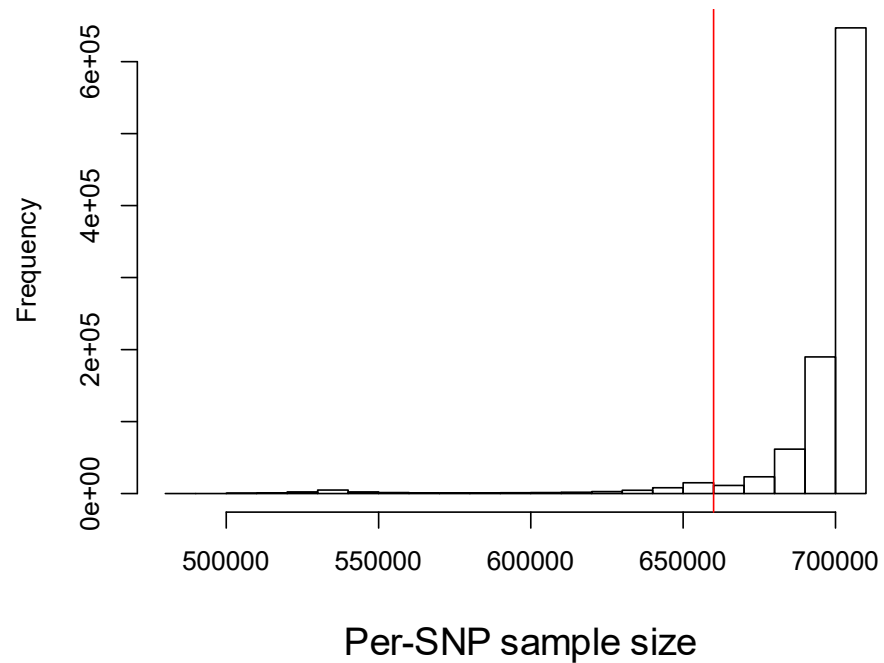
### **Per-SNP sample size**

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?

## 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 <i>liftover</i>
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.

## 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., <i>ImpG</i> ).
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 ( <i>DENTIST</i> ).
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., $> 3 \times \text{SD}$ . Impute $N = 1/(2pq(\text{SE} + b^2))$ if necessary.
Sample size for disease	Total sample size ( $N_{\text{case}} + N_{\text{ctrl}}$ ) or effective sample size - which one to use?	For <i>SBayes</i> , we recommend using the total sample size.

# Practical6\_SumStat\_QC.html



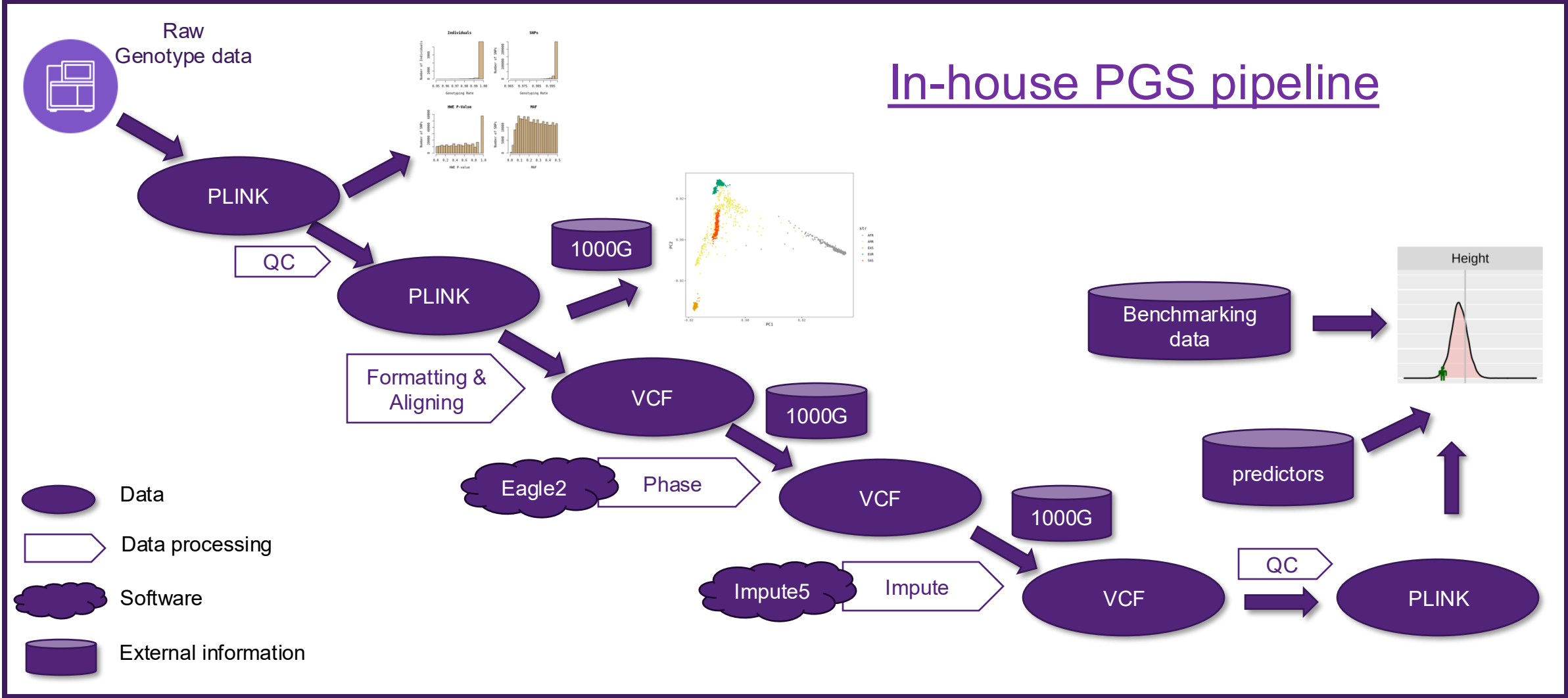
THE UNIVERSITY  
OF QUEENSLAND  
AUSTRALIA

CREATE CHANGE

# 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

# Why a raw data is not ready for PGS profiling?

## ➤ Quality

## ➤ Coverage

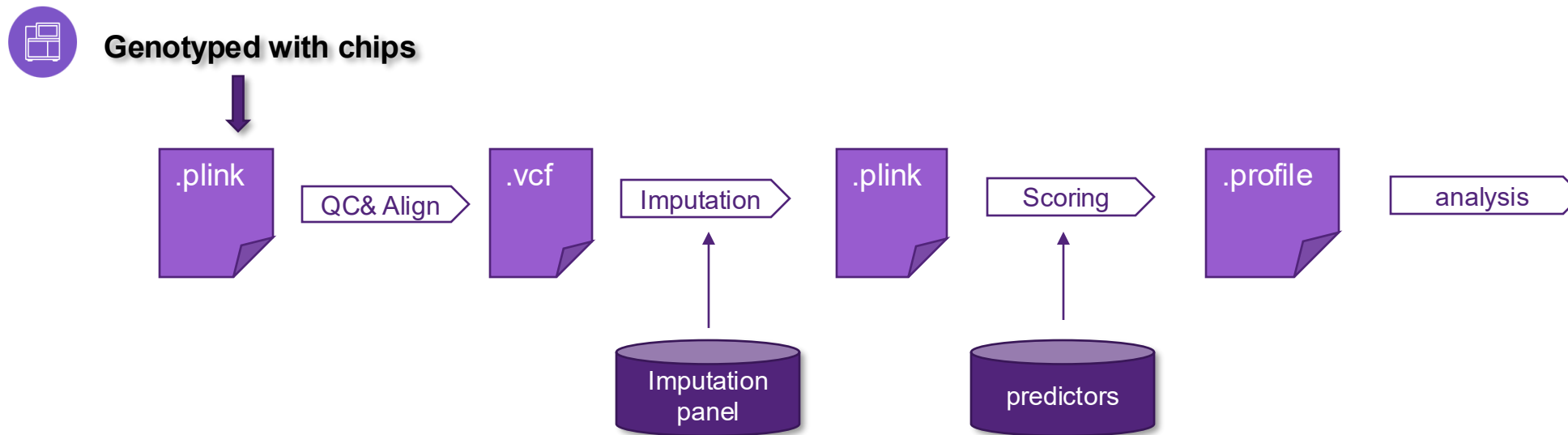
A high density  
SBayesRC Predictor  
– 7.3M SNPs



	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

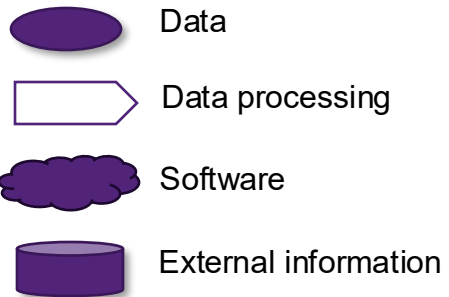
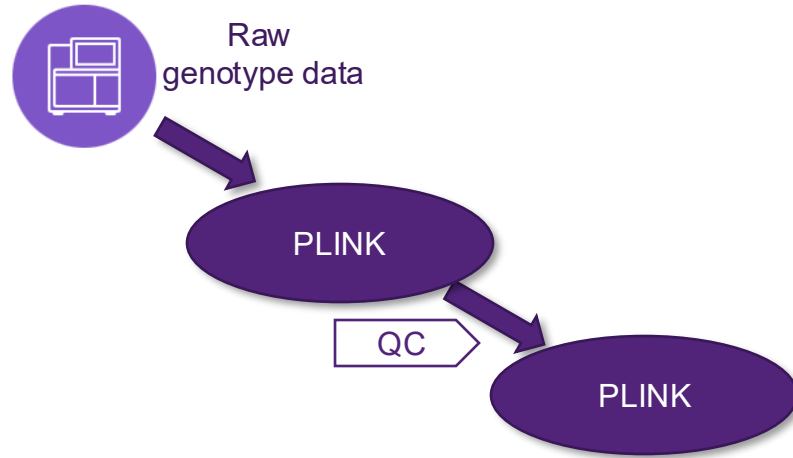


# Overview of PGS pipeline



# 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>

## Human genomes

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

- Fileserver (bigBed, maf, fa, etc) annotations [Telomere-to-Telomere](#)
- Standard genome sequence files and select annotations (2bit, GTF, GC-content, etc)
- LiftOver files
- Pairwise alignments ▶ [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 ▶ [hg19ToHg38.over.chain.gz](#)
- SNP-masked fasta files ▶ [hg38ToHg19.over.chain.gz](#)
- LiftOver files
- Pairwise alignments ▶ [hs1ToHg38.over.chain.gz](#)
- Multiple alignments ▶ [hs1ToHg19.over.chain.gz](#)
- Patches ▶
- Data archive [hg38ToHs1.over.chain.gz](#)

Very similar / Same.

GRCh37 names them `chr1`, `chr2`, `chr3`, etc, while hg19 just has `1`, `2`, `3`.

Different Mitochondria contigs.

### 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 ▶

# Liftover plink files

Best solution: recommend realigning the manifest files with BCFtools/gtc2vcf (<http://github.com/freeseek/gtc2vcf>)

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`

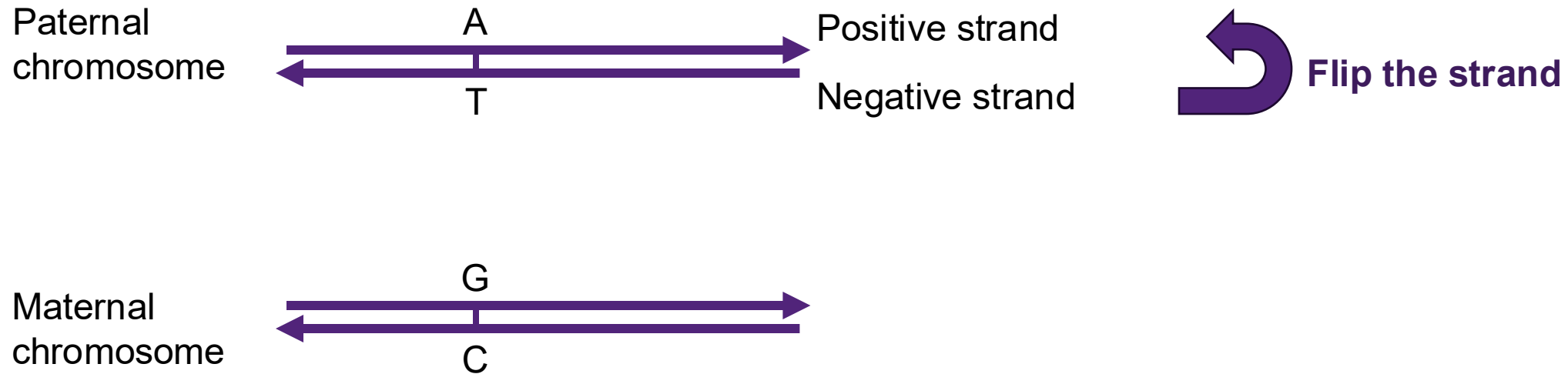
- Available: <https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle>



# 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



# Strand resource

<https://www.strand.org.uk>

Strand Files

[Top Strand](#)
[Source Strand](#)
[ILMN Strand](#)
[Affymetrix Arrays](#)
[AB Alleles](#)
[Ref/Alt](#)

## ILMN Strand

These files assume the data are aligned to the ILMN Strand.

Content: Choose of the array of interest to view/download the data on the different genome builds

GSA-24v1-0\_A2

GSA-24v1-0\_A2  
 ILMN Strand  
 NCBI35  
 GSA-24v1-0\_A2

Usage is:

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

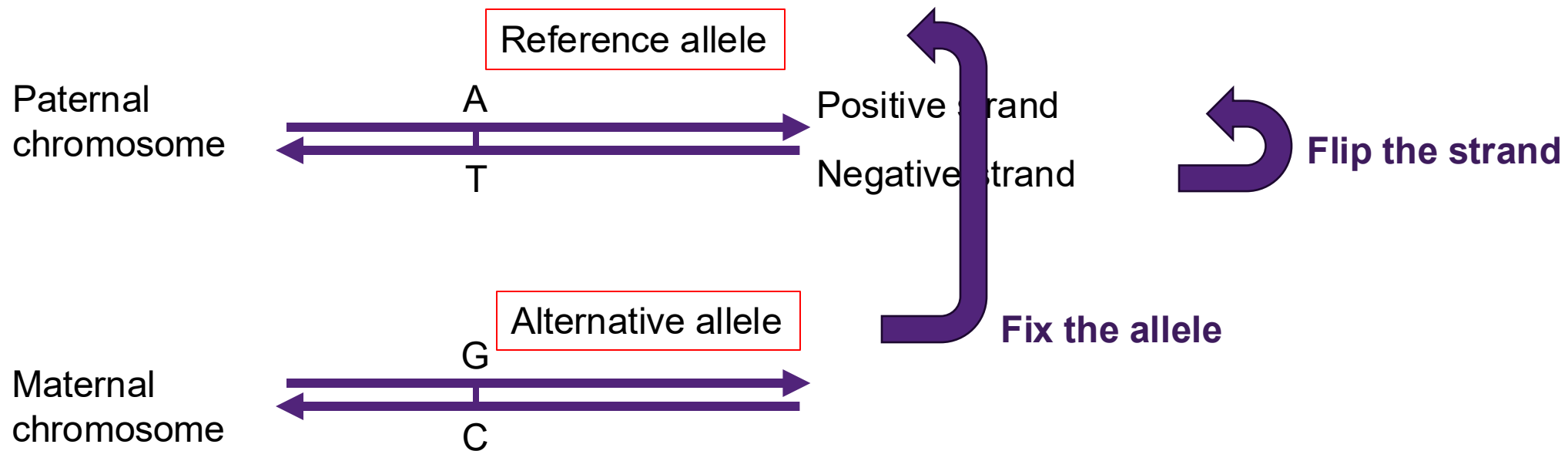
where:

- `<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

GSA-24v3-0\_A2

GSAMD-24v1-0\_20011747\_A1

# 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_glk_v37.fasta).
ref2fix=${refpath}/human_glk_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

## Example

**VCF header**

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

**Mandatory header lines**

**Optional header lines** (meta-data about the annotations in the VCF body)

**Body**

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0/1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1/0:77	1/1:95
1	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

**Reference alleles (GT=0)**

**Alternate alleles (GT>0 is an index to the ALT column)**

**Deletion**

**SNP**

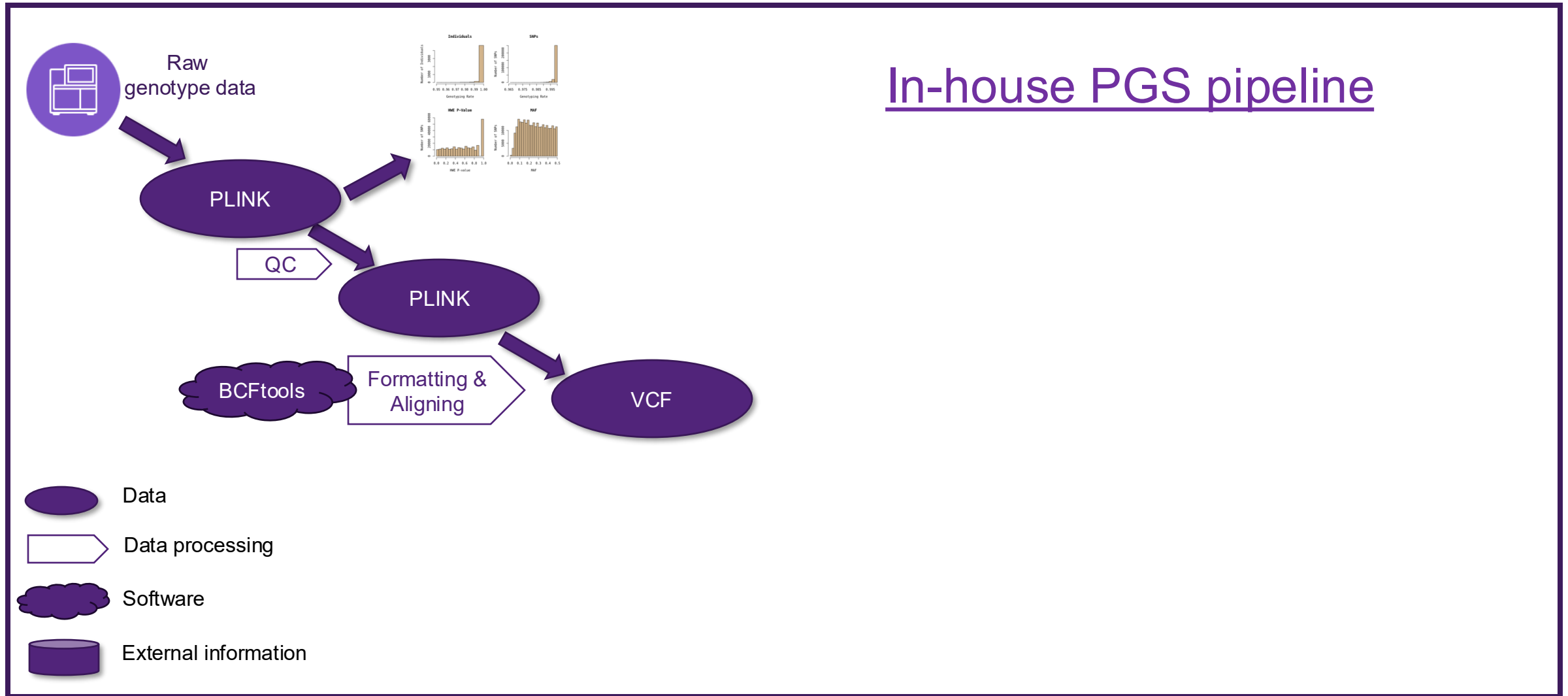
**Large SV**

**Insertion**

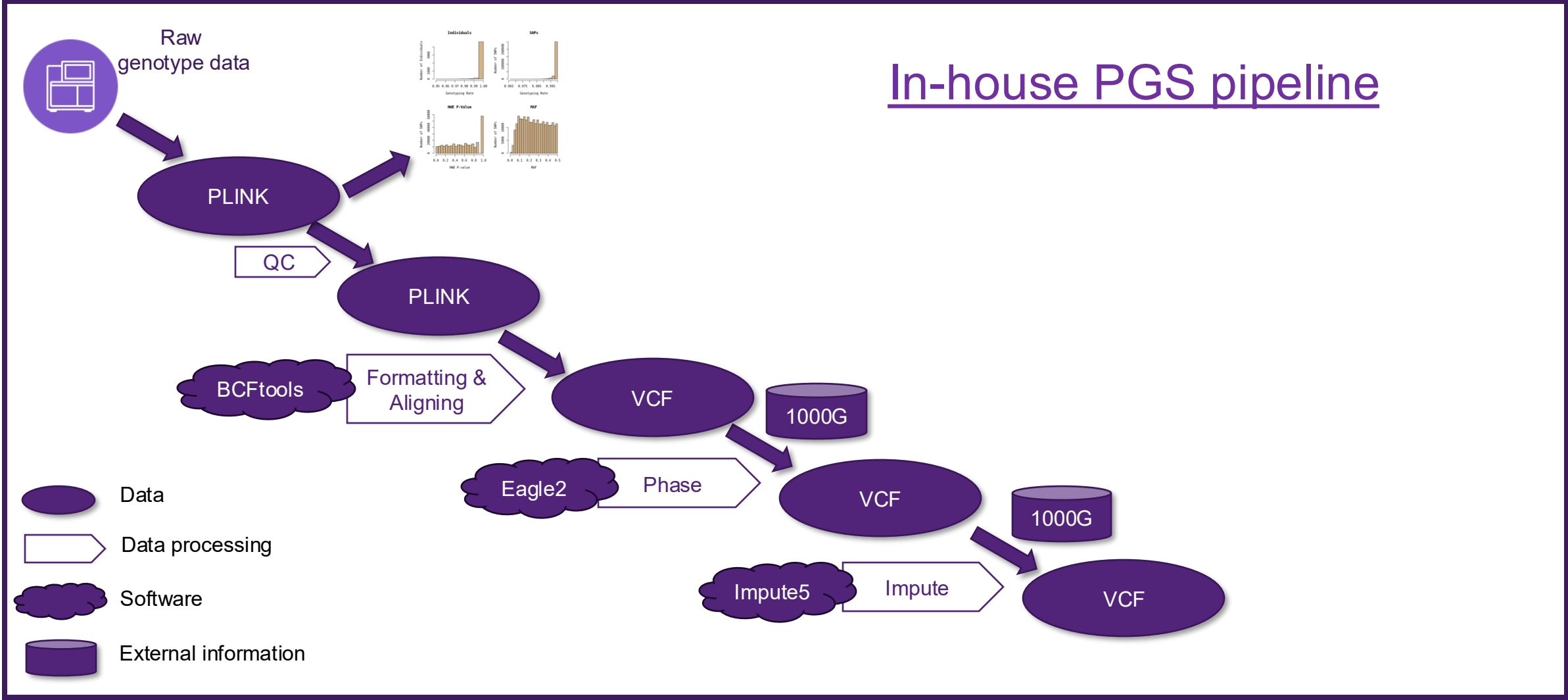
**Other event**

**Phased data** (G and C above are on the same chromosome)

# schematic of technical pipeline



# schematic of technical pipeline





# phasing

**a** Genotype data with missing data at untyped SNPs (grey question marks)

1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

**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

0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	0	0	1	0	0	1	1	1
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0

# 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

# Sanger Imputation Service

This is a free genotype **imputation** and **phasing** service provided by the [Wellcome Sanger Institute](#). You can upload GWAS data in VCF or 23andMe format and receive imputed and phased genomes back. Click [here](#) to learn more and [follow us on Twitter](#).

## Before you start

Be sure to [read through the instructions](#).

You will need to set up a free account with [Globus](#) and have [Globus Connect](#) running at your institute or on your computer to transfer files to and from the service.

## Ready to start?


If you are ready to upload your data, please fill in the details below to **register an imputation and/or phasing job**. If you need more information, see the [about](#) page. See also our [Privacy and Security](#) statement.

What is this ?

Next

The Sanger Imputation Service is developed by the [Vertebrate Resequencing Group](#) at the Wellcome Sanger Institute  
Copyright © 2015-2018 Genome Research Limited (reg no. 2742969) is a charity registered in England with number 1021457.  
[Terms and conditions](#) | [Cookies policy](#).

Most user friendly

 **BioData CATALYST**

Home About Help Contact Sign up Login

Dec 2023 release: updated TOPMed r3 panel and server security enhancements

# TOPMed Imputation Server

Free Next-Generation Genotype Imputation Service

Sign up now

Login

62.2M

Imputed Genomes


4834

Registered Users


12

Active Jobs


## The easiest way to impute genotypes



**Upload your genotypes to our secured service.**



**Choose a reference panel.** We will take care of pre-phasing and imputation.



**Download the results.** All results are encrypted with a one-time password. After 7 days, all results are deleted from our server.

Largest panel

Michigan Imputation Server

Home Help Contact

Sign up Login

# Michigan Imputation Server

Free Next-Generation Genotype Imputation Platform

Sign up now

Login

112.6M

Imputed Genomes

12111

Registered Users

24

Running Jobs

## Genotype Imputation

You can upload genotyping data and the application imputes your genotypes against different reference panels.

Run

Learn more

## HLA Imputation

Enables accurate prediction of human leukocyte antigen (HLA) genotypes from your uploaded genotyping data using multi-ancestry reference panels.

Run

Learn more

## Polygenic Score Calculation

You can upload genotyping data and the application imputes your genotypes, performs ancestry estimation and finally calculates Polygenic Risk Scores.

Run

Learn more

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} \
```

 A chromosome can be imputed as chunks

```
--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
```

# 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.

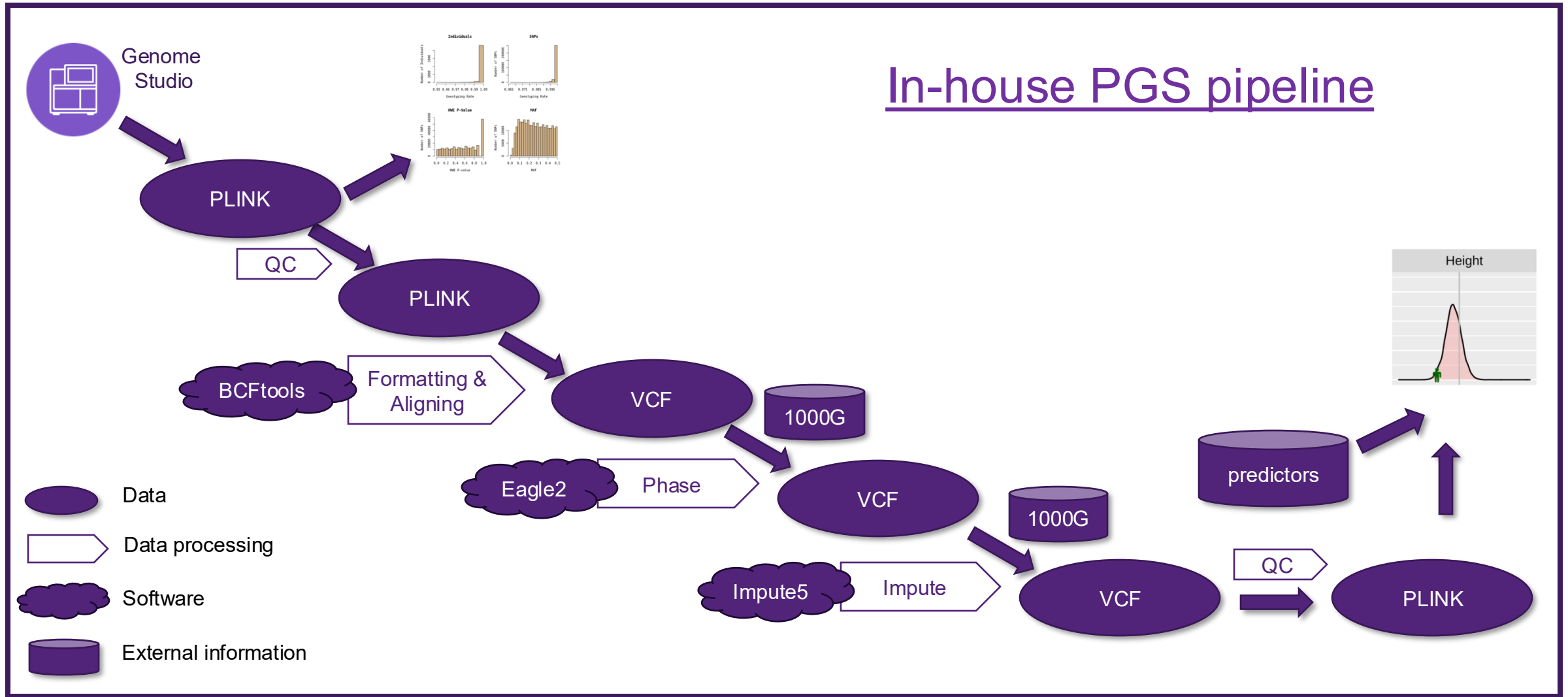
## ➤ SNP ID

- 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
```

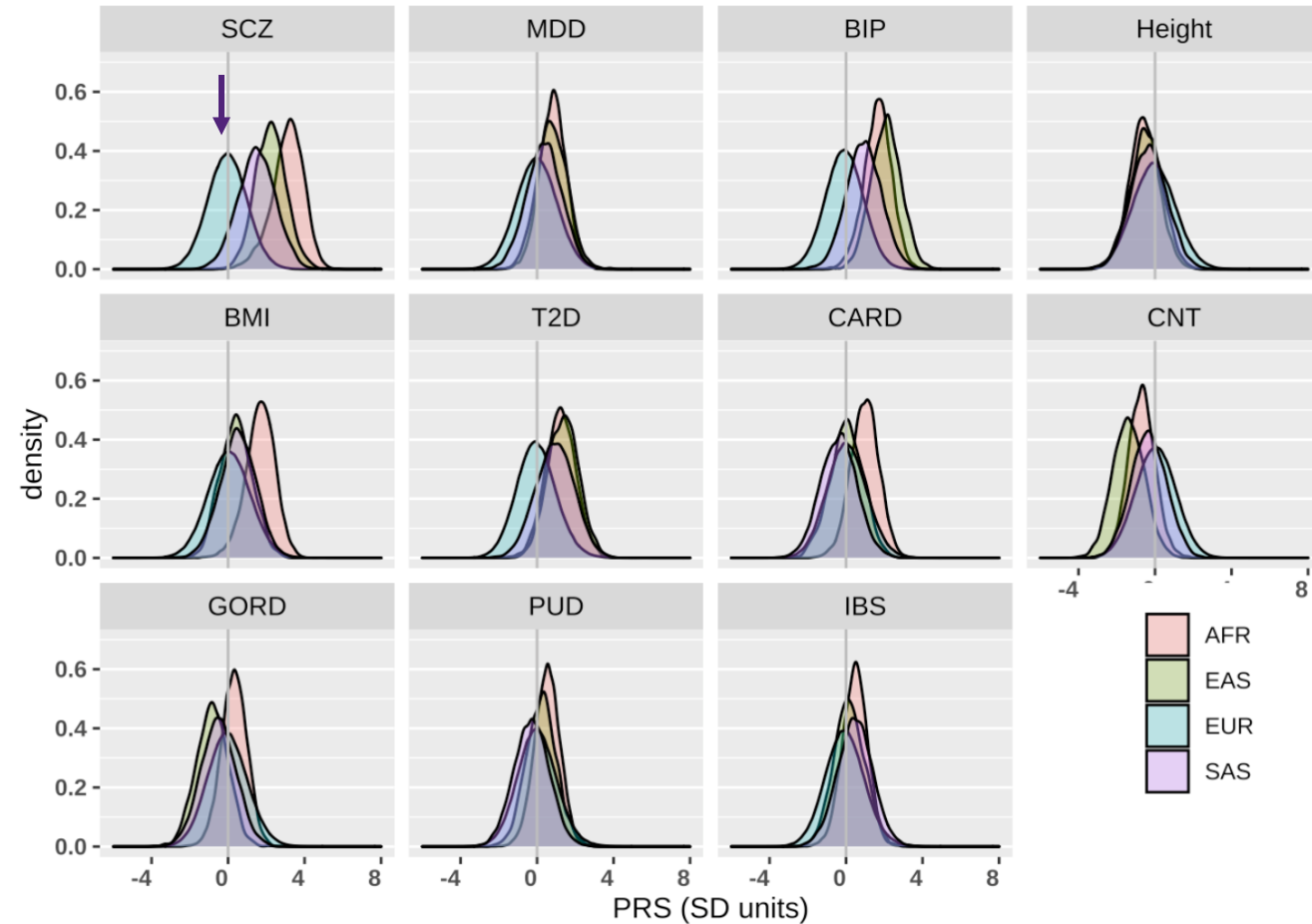
The parameters after your predictor file means

- 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

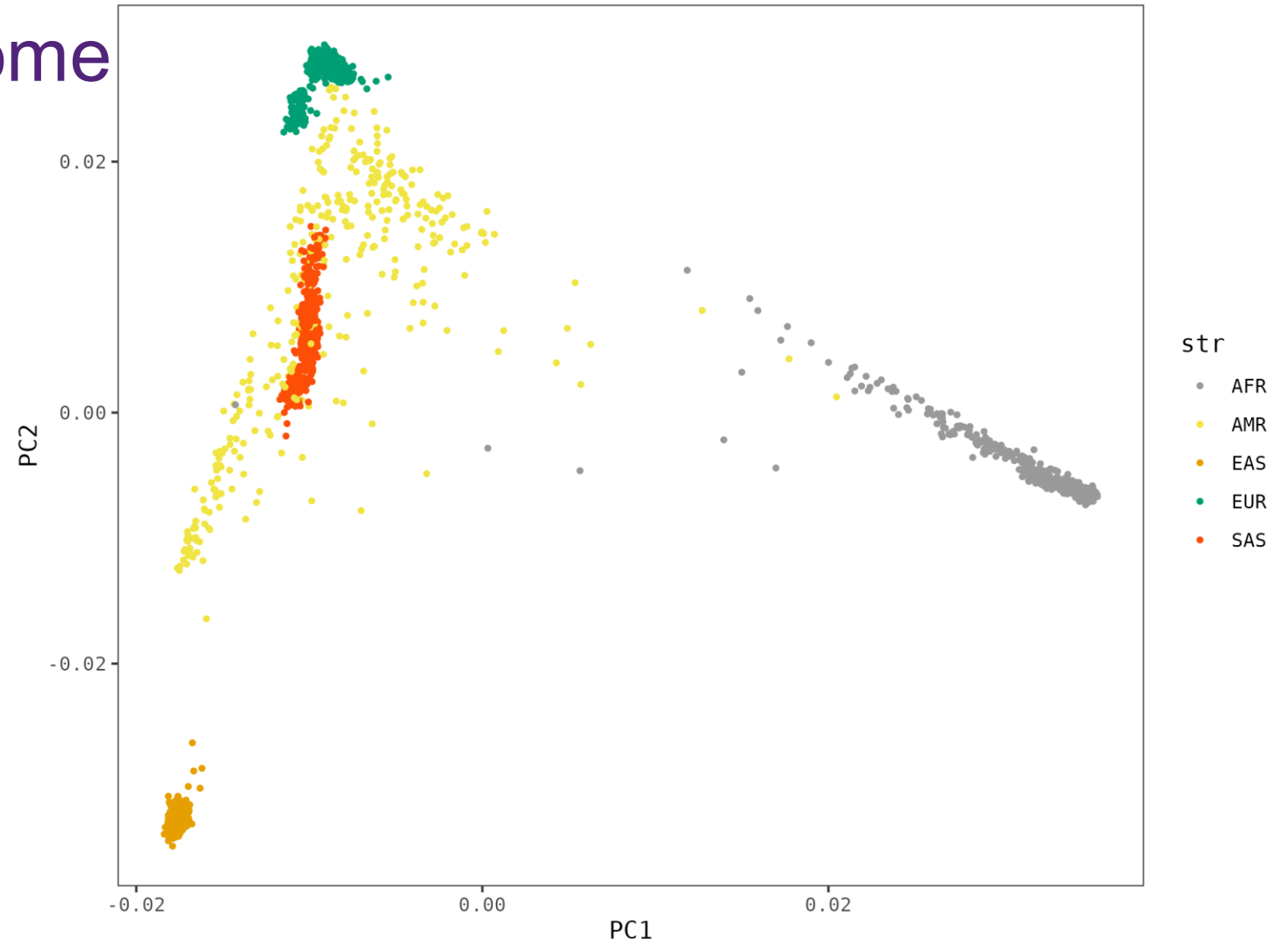
```
### generate GRM of reference data
```

```
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



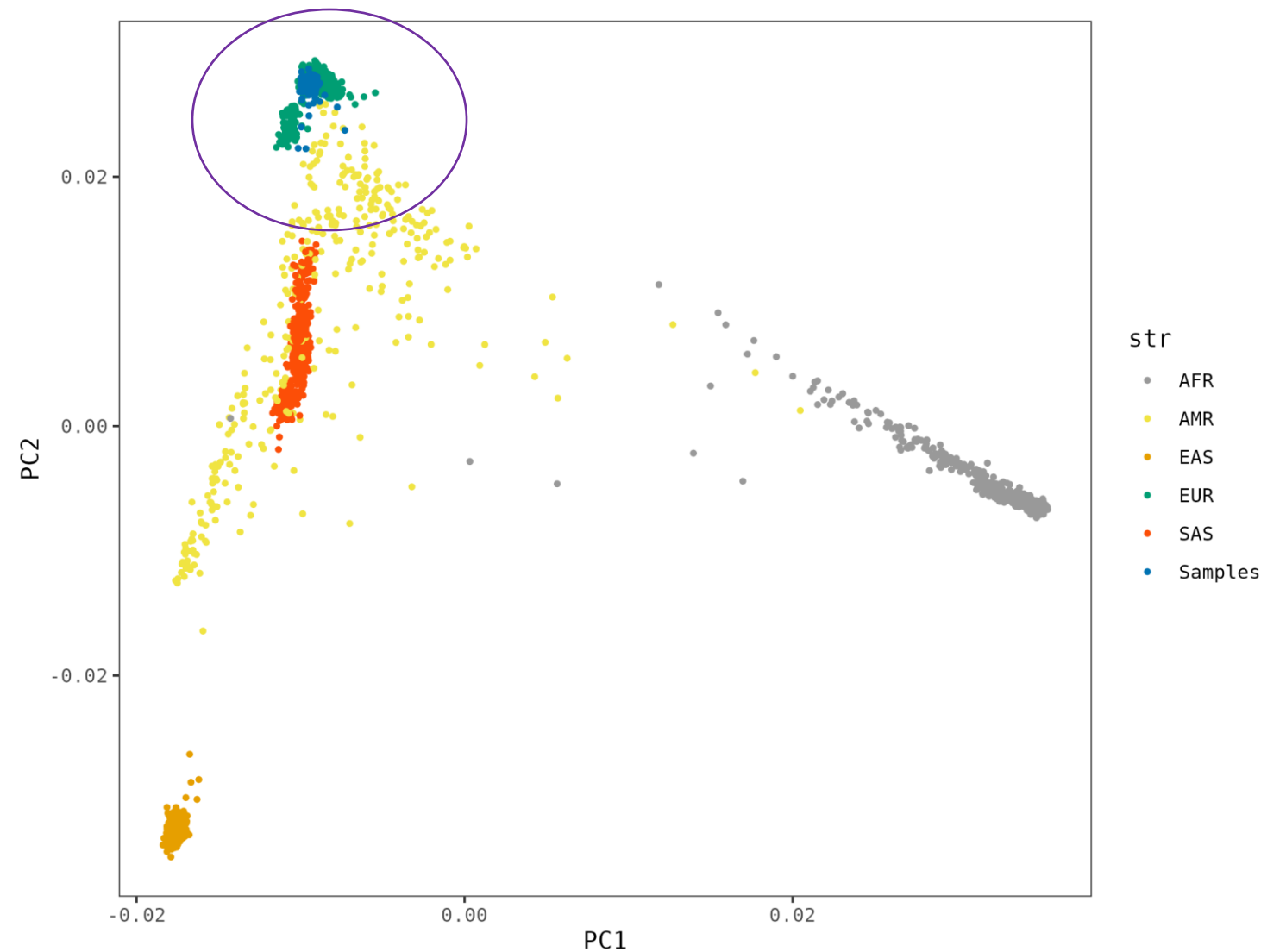
# PC projection using GCTA

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

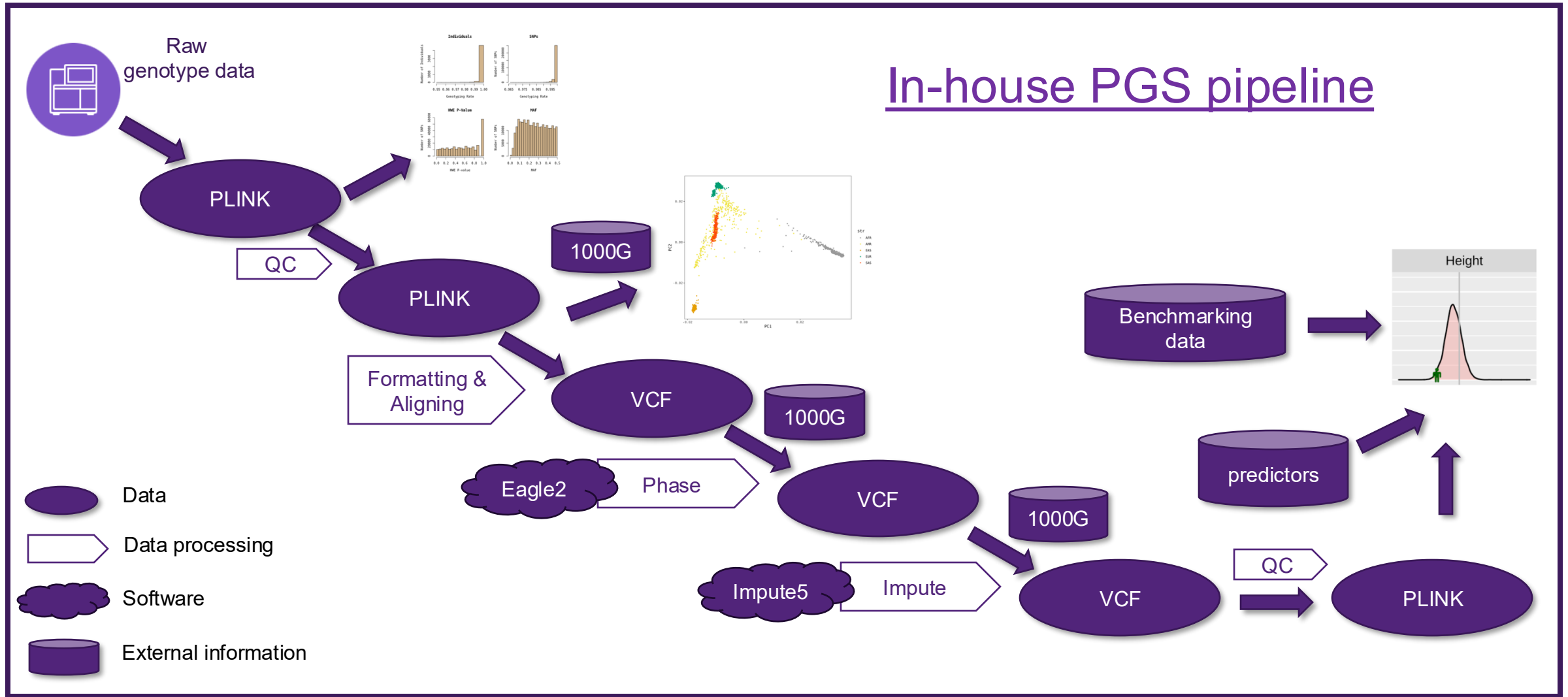
### PC projection
gcta \
--bfile ${data} \
--extract common.SNPs.txt \
--project-loading ${pcref}.05.common_pca3_snp_loading 3 \
--out ${data}_05.common_pca3
```



# PC projection



# schematic of technical pipeline



# Questions and Wrap Up

# Survey

<https://form.jotform.com/251872762920866>

Thank you!!