GWAS stands for **Genome-Wide Association Study.**

- It is a method used in genetics and genomics research to identify **associations between genetic variations and traits or diseases.**
- In a GWAS, researchers examine a large number of genetic markers, such as single nucleotide polymorphisms (SNPs), across the entire genome of individuals from a population.
- By comparing the presence or absence of specific genetic variations in individuals with a particular trait or disease to those without the trait or disease, researchers can **identify potential genetic associations**.
- GWAS studies have been instrumental in discovering genetic variants that are associated with various complex traits and diseases, including common conditions like diabetes, heart disease, and certain types of cancer. The findings from GWAS can provide insights into the underlying genetic basis of these traits and diseases, as well as potential targets for further research and therapeutic interventions.

PRS stands for **Polygenic Risk Score**.

It is a numerical score calculated using data from GWAS studies to **estimate an individual's genetic predisposition to a particular trait or disease**.

The PRS is based on the **cumulative effects of multiple genetic variants identified in GWAS**, each of which may have a small effect on the trait or disease.

To calculate a PRS, researchers assign a weight to each genetic variant based on its strength of association with the trait or disease in the GWAS. These weights are then multiplied by the number of risk alleles (i.e., the specific genetic variant associated with the trait or disease) an individual carries at each relevant locus. The weighted scores are summed to generate an overall PRS for an individual, indicating their genetic risk profile for the trait or disease of interest.

PRS can be used in various applications, including

Genome-wide association studies (GWAS)

- GWAS aims to identify associations of genotypes with phenotypes.
- GWAS can consider copy-number variants or sequence variations in the human genome, although the most commonly studied genetic variants in GWAS are single-nucleotide polymorphisms (SNPs).
- It typically reports blocks of correlated SNPs that all show a statistically significant association with the trait of interest, known as genomic risk loci

Genome-wide association studies (GWAS) have become increasingly popular to identify associations between single nucleotide polymorphisms (SNPs) and phenotypic traits.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/

Common and rare variants

- Genome-wide association studies (GWAS) generally involve targeted genotyping of specific and pre-selected variants using microarrays, whereas
- Whole-exome sequencing (WES) and whole-genome sequencing (WGS) studies aim to capture all genetic variation.
- Declaring a variant as common or rare is population-specific and cannot be generalized across populations.
- Generally, common variants are those with a minor allele frequency above 10%, although as population sizes grow this threshold can be as low as 1% as researchers typically adhere to a minimum minor allele count; for example, at least 100 individuals who carry at least one copy of the minor allele.

- With WGS and WES studies just beginning to mature, current analysis protocols may need to be extended to also cover specific issues that arise when analysing rare variants, for example, when controlling for population stratification, or imputing missing genotypes.

https://www.nature.com/articles/s43586-021-00056-9

Genetics basics when understanding -GWAS

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/

The aim of genome-wide association studies (GWAS) is to identify **single nucleotide polymorphisms** of which the allele frequencies vary systematically as a function of phenotypic trait values .

Single nucleotide polymorphism (SNP): This is a variation in a single nucleotide (i.e., A, C, G, or T) that occurs at a specific position in the genome. A SNP usually exists as two different forms (e.g., A vs. T). These different forms are called alleles. A SNP with two alleles has three different genotypes (e.g., AA, AT, and TT).

Heterozygosity: This is the carrying of two different alleles of a specific SNP. The heterozygosity rate of an individual is the proportion of heterozygous genotypes. High levels of heterozygosity within an individual might be an indication of low sample quality whereas low levels of heterozygosity may be due to inbreeding.

The Hardy–Weinberg (dis)equilibrium (HWE) law: This concerns the relation between the allele and genotype frequencies. It assumes an indefinitely large population, with no selection, mutation, or migration. The law states that the genotype and the allele frequencies are constant over generations. Violation of the HWE law indicates that genotype frequencies are significantly different from expectations (e.g., if the frequency of allele A = 0.20 and the frequency of allele T = 0.80; the expected frequency of genotype AT is 2*0.2*0.8 = 0.32) and the observed frequency should not be significantly different. In GWAS, it is generally assumed that deviations from HWE are the result of genotyping errors. The HWE thresholds in cases are often less stringent than those in controls, as the violation of the HWE law in cases can be indicative of true genetic association with disease risk.

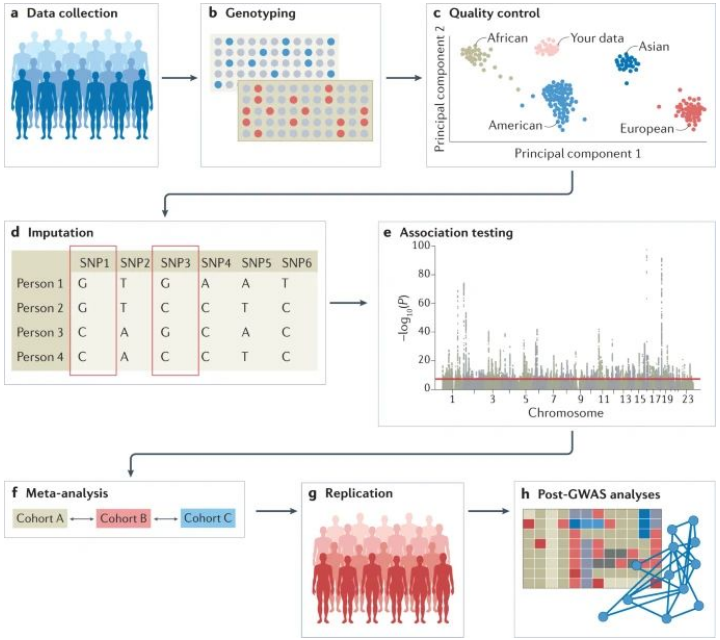
Population stratification: This is the presence of multiple subpopulations (e.g., individuals with different ethnic background) in a study. Because allele frequencies can differ between subpopulations, population stratification can lead to false positive associations and/or mask true associations. An excellent example of this is the chopstick gene, where a SNP, due to population stratification, accounted for nearly half of the variance in the capacity to eat with chopsticks (Hamer & Sirota, 2000).

For rest of the important definitions see - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/#mpr1608-blk-0001

Conducting GWAS

https://www.nature.com/articles/s43586-021-00056-9

Overview of steps for conducting GWAS



a | **Selecting study populations:** Data can be collected from study cohorts etc.
b | **Genotyping:** Genotypic data can be collected using microarrays to capture common variants, or NGS methods for WGS or whole-exome sequencing WES.
Data processing

c | Quality control includes steps at the wet-laboratory stage, such as genotype calling and DNA switches, and dry-laboratory stages on called genotypes, such as deletion of bad SNPs and individuals, detection of population strata in the sample and calculation of principal components. Figure depicts clustering of individuals according to genetic substrata.

d | Genotypic data can be phased, and untyped genotypes imputed using information from matched reference populations from repositories such as 1000 Genomes Project or TopMed.

e | **Testing for associations:** Genetic association tests are run for each genetic variant, using an appropriate model (for example, additive, non-additive, linear or logistic regression). Output is inspected for unusual patterns and summary statistics are generated.

Accounting for false discovery

Genome-wide association meta-analysis

f | Results from multiple smaller cohorts are combined using standardized statistical pipelines.
g | Results can be replicated using internal replication or external replication in an independent cohort.
h | In silico analysis of genome-wide association studies (GWAS), using information from external resources. This can include in silico fine-mapping, SNP to gene mapping, gene to function mapping, pathway analysis, genetic correlation analysis, Mendelian randomization and polygenic risk prediction. After GWAS, functional hypotheses can be tested using experimental techniques such as CRISPR or

GWAS, PRS tutorial

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/

A basic understanding of the theory behind genetic analysis (e.g., GWAS and PRS), the essential QC steps, and the use of appropriate software and methods, along with practical experience are imperative to be able to conduct a genetic study with reliable and reproducible results.
This tutorial highlights important concepts to successfully conduct a GWAS and PRS analysis. We presented a tutorial based on commonly used, open-source, freely available software tools, that are accessibly for novice users. In addition, we made scripts and a simulated data set available to provide hands-on practice at https://github.com/MareesAT/GWA_tutorial/.

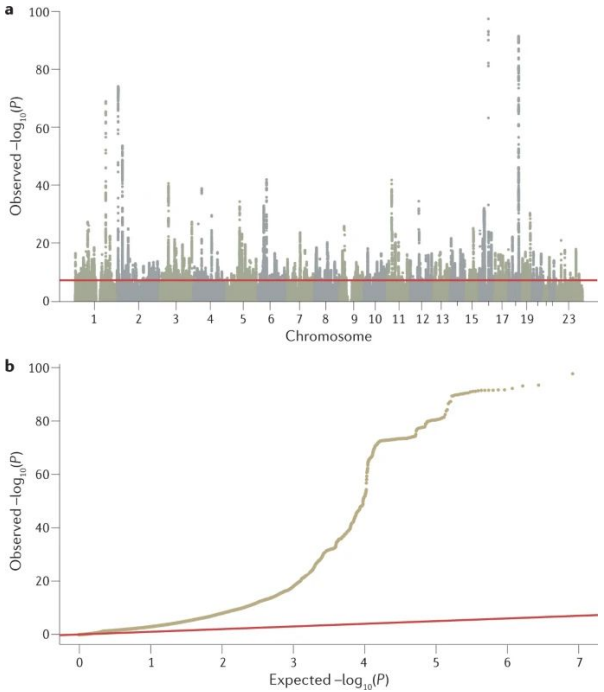
Output of GWAS

Summary statistics: These are the results obtained after conducting a GWAS, including information on chromosome number, position of the SNP, SNP(rs)-identifier, MAF, effect size (odds ratio/beta), standard error, and *p* value. Summary statistics of GWAS are often freely accessible or shared between researchers.

The primary output of a GWAS analysis is a list of *P* values, effect sizes and their directions generated from the association tests of all tested genetic variants with a phenotype of interest.

The data is routinely visualized using Manhattan plots and quantile–quantile plots

a | **Manhattan plot** showing significance of each variant's association with a phenotype (body mass index in this case[77]). Each dot represents a single-nucleotide polymorphism (SNP), with SNPs ordered on the *x* axis according to their genomic position. *y* axis represents strength of their association measured as $-\log_{10}$ transformed *P* values. Red line marks genome-wide significance threshold of $P < 5 \times 10^{-8}$.



Important databases

GWAS relies strongly on in-depth knowledge of the genetic architecture of the human genome, which is provided by two important research initiatives, namely, the International HapMap Project and the 1000 Genomes project.

The **International HapMap Project** (http://hapmap.ncbi.nlm.nih.gov/; Gibbs et al., 2003) described the patterns of common SNPs within the human DNA sequence whereas

The **1000 Genomes (1KG) project** (http://www.1000genomes.org/; Altshuler et al., 2012) provided a map of both common and rare SNPs.

Also

The **database of Genotypes and Phenotypes (dbGaP)** (https://www.ncbi.nlm.nih.gov/gap/) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.

GWAS tutorial (very good and detailed)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/#mpr1608-sec-0006title

SOFTWARE (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/#mpr1608-blk-0001)

QC OF GENETIC DATA (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/#mpr1608-sec-0009title)

CONTROLLING FOR POPULATION STRATIFICATION (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/#mpr1608-sec-0013title)

STATISTICAL TESTS OF ASSOCIATION FOR BINARY AND QUANTITATIVE TRAITS

(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/#mpr1608-sec-0014title)

After QC and calculation of MDS components, the data are ready for subsequent association tests.

Various types of association that are suitable for

- Binary traits (e.g., alcohol dependent patients vs. healthy controls) or
- Quantitative traits (e.g., the number of alcoholic beverages consumed per week).

Correction for multiple testing

Polygenic risk score (PRS)

Because GWAS results showed that effect sizes of individual SNPs are small, researchers in the psychiatric field developed an interest in methods that aggregate the effect of SNPs.

E.g. **Polygenic risk score (PRS)** analysis as is one of the most relevant method: it is relatively easy to conduct while it can be applied to target samples with relatively modest sample sizes. **PRS** combines the effect sizes of multiple SNPs into a single aggregated score that can be used to predict disease risk. Thus

PRS is an estimate of an individual's genetic liability to a trait or disease, calculated according to their

Genotype profile and

Relevant genome-wide association study (GWAS) data.

Very simplistic introduction : https://www.youtube.com/watch?v=-A44pRFrbwc

The PRS is an individual-level score that is calculated based on

The number of risk variants that a person carries, weighted by SNP *effect sizes* that are derived from an independent large-scaled discovery GWAS.

As such, the score is an indication of the total genetic risk of a specific individual for a particular trait, which can be used for clinical prediction or screening , e.g., breast cancer.

PRS has been further used to investigate whether **genetic effect sizes obtained from a GWAS of a specific phenotype of interest can be used to predict the risk of another phenotype…**

Polygenic Risk Score (PRS) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/#mpr1608-sec-0018title

Single variant association analysis has been the primary method in GWAS but requires very large sample sizes to detect more than a handful of SNPs for many complex traits.

In contrast, PRS analysis does not aim to identify individual SNPs but instead aggregates genetic risk across the genome in a single individual polygenic score for a trait of interest.

PRSs are commonly used to predict the risk of disease in a target cohort using the GWAS summary statistics of an independent discovery cohort (Fig. 4). PRSs can be used to identify individuals at a high risk of disease for clinical interventions and provide additional information over traditional clinical risk scores for stratified screening.

- To conduct PRS analysis, trait-specific weights (beta's for continuous traits and the log of the odds ratios for binary traits) are obtained from a discovery GWAS.
- In the target sample, a PRS is calculated for each individual based on the weighted sum of the number of risk alleles that he or she carries multiplied by the trait-specific weights.
- Although in principle all common SNPs could be used in a PRS analysis, it is customary to first clump (see **clumping**) the GWAS results before computing risk scores. *p* value thresholds are typically used to remove SNPs that show little or no statistical evidence for association (e.g., only keep SNPs with *p* values <0.5 or <0.1. Usually, multiple PRS analyses will be performed, with varying thresholds for the *p* values.

See next slide

Polygenic Risk Score (PRS)

Discovery GWAS

	Weight*	Risk Allele
SNP1	0.2	A
SNP2	-0.3	C
SNP3	0.1	G

Individual	Alleles SNP1	Alleles SNP2	Alleles SNP3
1	T	AA	CG
2	AA	CA	GG
3	TT	AC	CG
4	TT	AA	GG
5	TA	CA	GC
6	AT	CA	CG
7	AA	AA	GG
8	AA	CC	CG
9	TA	CC	GC
10	AT	AA	CG

PRS:

Individual	SNP 1	SNP 2	SNP 3	**PRS**
1	0.2+0.0	0.0+0.0	0.0+0.1	**0.3**
2	0.2+0.2	-0.3+0.0	0.1+0.1	**0.3**
3	0.0+0.0	0.0-0.3	0.0+0.1	**-0.2**
4	0.0+0.0	0.0+0.0	0.1+0.1	**0.2**
5	0.0+0.2	-0.3+0.0	0.1+0.0	**0.0**
6	0.2+0.0	-0.3+0.0	0.0+0.1	**0.0**
7	0.2+0.2	0.0+0.0	+0.1+0.1	**0.6**
8	0.2+0.2	-0.3-0.3	0.0+0.1	**-0.1**
9	0.0+0.2	-0.3-0.3	0.1+0.0	**-0.3**
10	0.2+0.0	0.0+0.0	0.0+0.1	**0.3**

Working example of three single nucleotide polymorphisms (SNPs) aggregated into a single individual polygenic risk score (PRS). *The weight is either the beta or the log of the odds-ratio, depending on whether a continuous or binary trait is analysed

- Once PRS have been calculated for all subjects in the target sample, the scores can be used in a (logistic) regression analysis to predict any trait that is expected to show genetic overlap with the trait of interest.
- The prediction accuracy can be expressed with the (pseudo-)R 2 measure of the regression analysis.
- It is important to include at least a few MDS components as covariates in the regression analysis to control for population stratification.
- A convenient program to perform PRS analysis is PRSice (see http://prsice.info/; Euesden, Lewis, & O'Reilly, 2015).
- It takes care of clumping, p value thresholds, MDS components, and plots attractive graphs. We refer to https://github.com/MareesAT/GWA_tutorial/ (4_PRS.doc) for a tutorial on how to perform your own PRS analysis using PRSice. Other programs for the application of PRS are, for example, PLINK (`--score`) and LDpred (Purcell et al., 2007; Vilhjalmsson et al., 2015).

Conducting polygenic risk prediction analyses

Key terms and definitions

Classic PRS method: the method—commonly known as the C+T method—for calculating PRSs applied in the key early PRS empirical studies, theoretical evaluations and software implementations.

The method involves computing PRSs based on a subset of partially independent (clumped) SNPs exceeding a specific GWAS association *P* value threshold. PRS analyses can be characterized by the two key input data sets that they require:

Base data: the GWAS summary statistics (e.g., effect sizes or *P* values) on which the PRS calculation is based. The base trait is the phenotype of study in the GWAS.

Target data: the genotype-phenotype data, in, for example, PLINK binary format, of individuals in whom PRSs are calculated.

The PRSs infer genetic liability of the base trait and are tested for association with the target trait.

Conducting polygenic risk prediction analyses

PRS analyses can be characterized by the two key input data sets that they require:

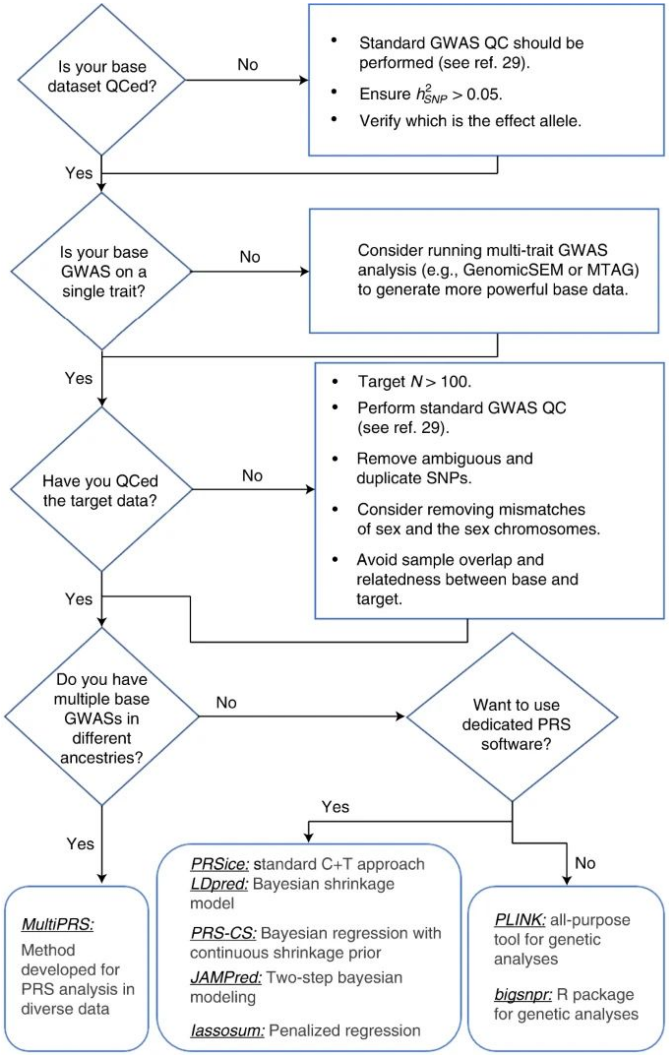
(i) **base data (GWAS)**, consisting of summary statistics (e.g., betas and *P* values) of genotype-phenotype associations at genetic variants (hereafter SNPs) genome-wide, typically made available online in text format by the investigators who performed the GWAS; and

(ii) **target data**, consisting of genotypes, and usually also phenotype(s), in individuals from a sample to which the researchers performing the PRS analysis have access (often not publicly available), which should be independent of the GWAS sample .

Conducting polygenic risk prediction analyses

A flow chart of suggested analytical steps that can be followed to perform QC and select software for PRS analyses.
https://www.nature.com/articles/s41596-020-0353-1#Sec2

This figure illustrates a PRS analysis pipeline, highlighting QC steps and some of the main software programs presently available to users as options, which may be selected according to scientific question, data, estimated accuracy and speed of PRS computation method

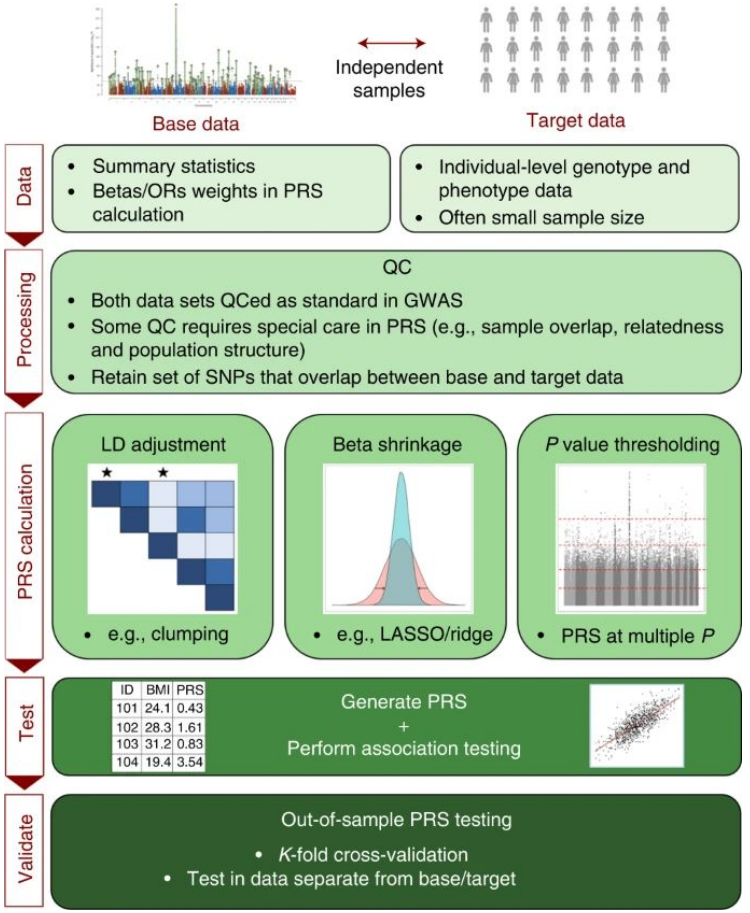


The PRS analysis process

https://choishingwan.github.io/PRS-Tutorial/

The first step in Polygenic Risk Score (PRS) analyses is to generate or obtain the **base data (GWAS summary statistics)**. Ideally these will correspond to the **most powerful GWAS results available on the phenotype under study**.

Target data consist of **individual-level genotype-phenotype data**, usually generated within your lab/department/collaboration.



The PRS analysis process -Steps https://choishingwan.github.io/PRS-Tutorial/

QC of base and target data

The power and validity of PRS analyses are dependent on the quality of the base and target data.

Calculating and Analysing PRS

The programs are

- PLINK
- PRSice-2
- LDPred-2
- lassosum

Plotting the Results

Nextflow pipelines for GWAS

(1) **nf-core/gwas** (currently in development and does not yet have any stable releases)

https://nf-co.re/gwas

(2) **H3AGWAS** : A portable workflow for Genome Wide Association Studies
https://www.biorxiv.org/content/10.1101/2022.05.02.490206v1.full.pdf

Docker images

https://quay.io/organization/h3abionet_org/

https://github.com/h3abionet/h3agwas-docker

(3) **nf-gwas-pipeline**: A Nextflow Genome-Wide Association Study Pipeline

https://joss.theoj.org/papers/10.21105/joss.02957

https://github.com/montilab/nf-gwas-pipeline

nf-core/gwas

H3AGWAS : A portable workflow for Genome Wide Association Studies: Using Nextflow and containerisation

A powerful, scalable and portable workflow implementing

1. **Pre-association analysis**
   - Producing PLINK data The call2plink workflow converts Illumina genotyping reports into PLINK format.
   - Quality control

2. **Implementation of various association testing methods**

3. **Post-association analysis of results**

Use genotype data and results of association testing in order to

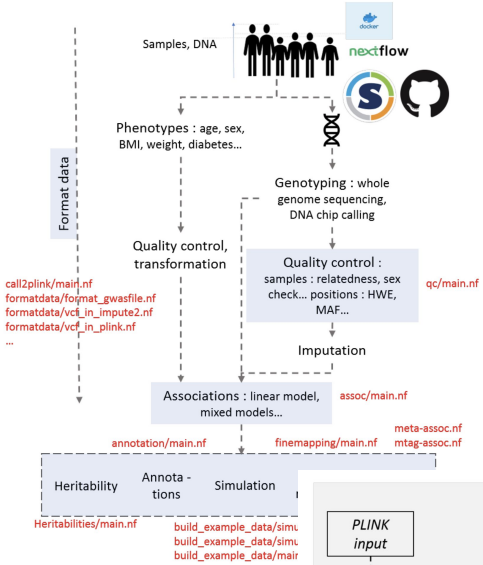
(1) find putative causal variants

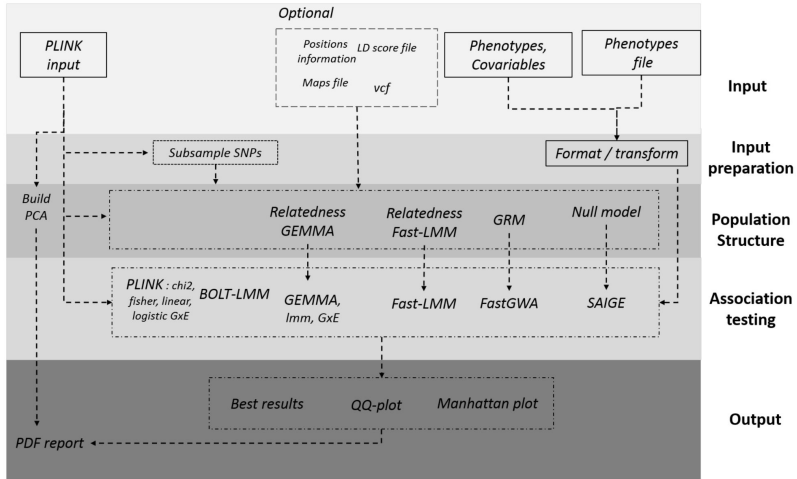
(2) perform a meta-analysis or multi-trait genomewide association study using summary statistics

(3) estimate global heritability; and

(4) annotate positions

The workflow is scalable — laptop to cluster to cloud (e.g., SLURM, AWS Batch, Azure). All required software is containerised and can run under Docker on Singularity.



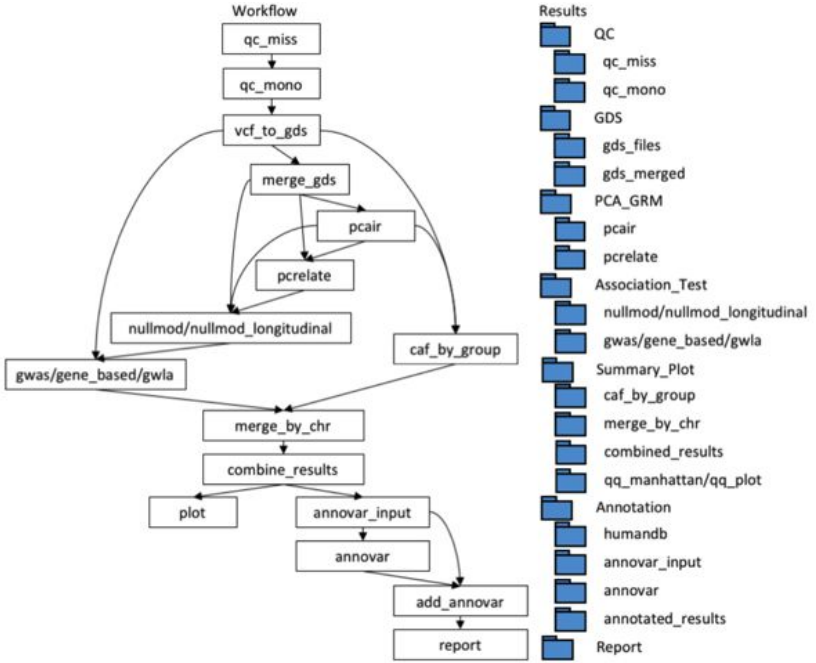


nf-gwas-pipeline: A Nextflow Genome-Wide Association Study Pipeline

- Combines multiple analysis tools – including bcftools, vcftools, the R packages SNPRelate/GENESIS/GMMAT and ANNOVAR – through Nextflow.

- The GWAS pipeline integrates the following steps
  - Data **quality control**
  - Principal component analysis and genetic relationship inference
  - Assessment and **genetic association analyses,** includi analysis of cross-sectional and longitudinal studies with either single variants or gene-based tests, into a unified analysis workflow. **Multiple ways**
  - Output: **Visualization** (Manhattan and QQ-plots) and **annotation** (ANNOVAR and MAF)

- The pipeline is implemented in Nextflow, dependencies are distributed through Docker, and the code is publicly available on Github.

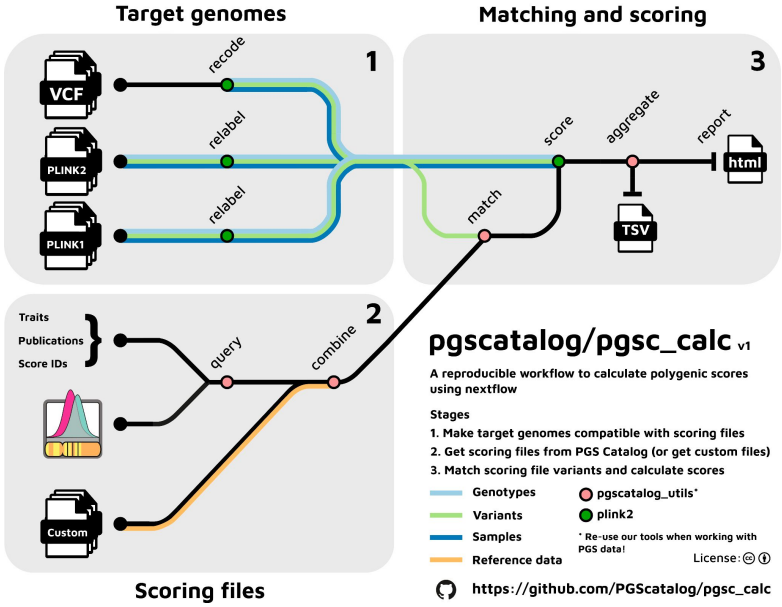


The Polygenic Score Catalog Calculator (pgsc_calc) : A nextflow pipeline

A bioinformatics best-practice analysis pipeline for calculating polygenic [risk] scores on samples with imputed genotypes using existing scoring files from the Polygenic Score (PGS) Catalog and/or user-defined PGS/PRS.

The Polygenic Score (PGS) Catalog (https://www.pgscatalog.org/) is an open database of published polygenic scores (PGS).

Each PGS in the Catalog is consistently annotated with relevant metadata; including scoring files (variants, effect alleles/weights), annotations of how the PGS was developed and applied, and evaluations of their predictive performance.



(1) Data compatibility

- Automatically combines and creates scoring files for efficient parallel computation of multiple PGS
- Matching variants in the scoring files against variants in the target dataset (in plink bfile/pfile or VCF format)

(2) **Base data** : GWAS

- Downloading scoring files using the PGS Catalog API in a specified genome build (GRCh37 and GRCh38).
- **Reading custom scoring files** (and performing a liftover if genotyping data is in a different build).
-

(3) Calculate PRS

- Calculates PGS for all samples (linear sum of weights and dosages)
- Creates a summary report to visualize score distributions and pipeline metadata (variant matching QC)

References

Genome-wide association studies

Nature Reviews methods primers Published: 26 August 2021 (https://www.nature.com/articles/s43586-021-00056-9)

Tutorial: a guide to performing polygenic risk score analyses

Nature Protocols Published: 24 July 2020 (https://www.nature.com/articles/s41596-020-0353-1)

A youtube video (not yet watched)

AG2PI Workshop #4 - A Practical Guide to Genome-Wide Association Studies (GWAS)

http://www.transplantdb.eu/sites/transplantdb.eu/files/HandsOnTutorialtoGWAS_Seren-030715.pdf

https://hastie.su.domains/ElemStatLearn/

Tutorial: a guide to performing polygenic risk score analyses

https://www.nature.com/articles/s41596-020-0353-1

The Polygenic Score Catalog Calculator (`pgsc_calc`)

https://github.com/PGScatalog/pgsc_calc

`pgsc_calc` is a bioinformatics best-practice analysis pipeline for calculating polygenic [risk] scores on samples with imputed genotypes using existing scoring files from the Polygenic Score (PGS) Catalog and/or user-defined PGS/PRS.