

生物資訊與分析

PRS

陳佳煒 (Jia-Wei Chen)

2025/04/15

Genetic architecture

Common Disease Common Variant

An early hypothesis in GWAS that multiple common variants collectively contribute to disease susceptibility. However, CDCV cannot fully explain the missing heritability - the unaccounted genetic contribution to disease risk.

Infinitesimal Model

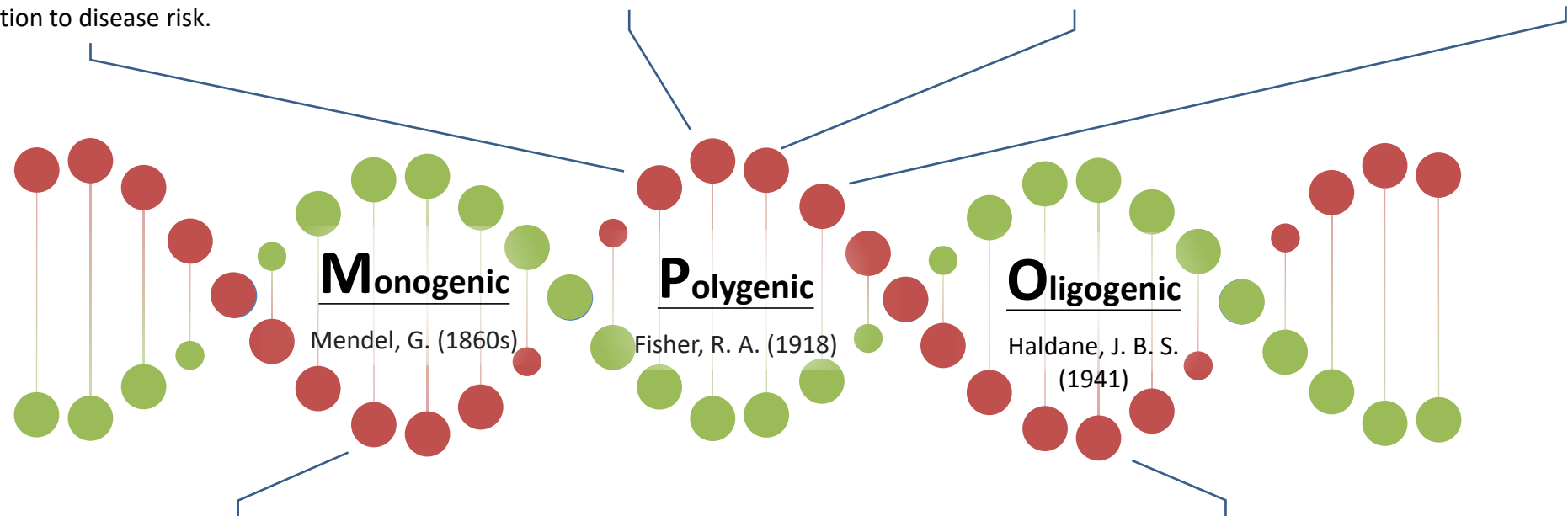
A purely statistical framework that complex traits are controlled by an infinite number of loci, each with very small, additive effects.

Broad Sense Heritability Model

A concept that neither common nor rare variants alone explain the missing heritability. It considers about GxG, GxE, and epigenetic effects

Omnigenic Model

Extends the polygenic concept by incorporating biological networks to explain the contribution of both **core** and **peripheral** genes to complex traits.



Rare Mendelian Disorders

Mutations in a single gene, following Mendelian inheritance patterns (dominant, recessive, or sex-linked), leading to the disorder.

Rare Alleles of Major Effect

An alternative concept to CDCV that a small number of rare variants (with MAF < 0.01) can significantly influence disease development.

Association and prediction

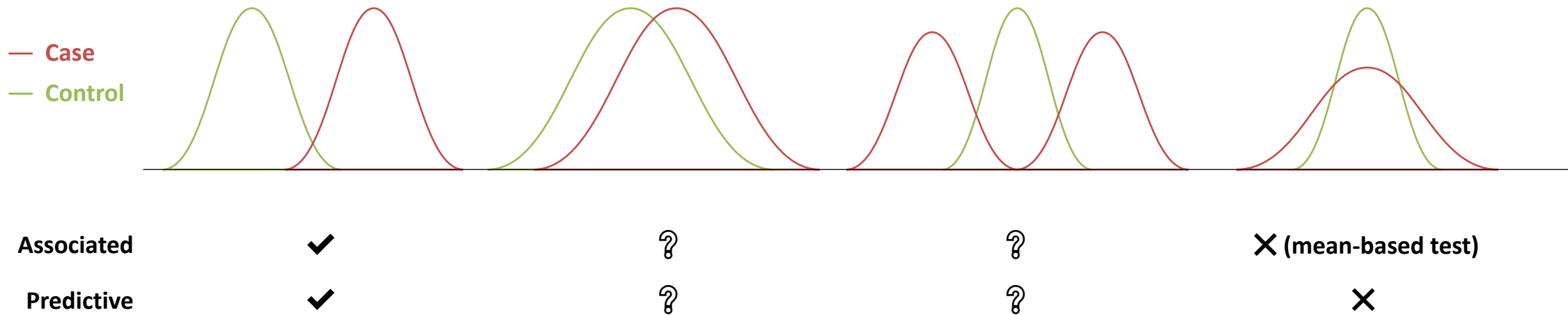
Association

Group (population)-based concept focuses on statistical relationships between variables at the group level, which informs us about broader patterns and relationships within populations.

Even **weakly predictive SNPs** can have **strong association** if the **sample size is large**.

Prediction

Individual-based concept focus on personalized outcomes for specific individuals, and considers unique characteristics, medical history, and relevant factors



GWAS summary statistic

GWAS Catalog 📄 <https://www.ebi.ac.uk/gwas/>

HuGeAMP 📄 <https://kp4cd.org/>

Biobank GWAS summary statistic

China 📄 <https://pheweb.ckbiobank.org/>

Finnish 📄 <https://pheweb.sph.umich.edu/FinMetSeq/>

Japan 📄 <https://pheweb.jp/> (hum0197.v18, hum0014.v32)

Korean 📄 <https://koges.leelabsg.org/>

UK 📄 https://github.com/Nealelab/UK_Biobank_GWAS

Taiwan 📄 <https://taiwanview.twbiobank.org.tw/pheweb.php>

	w_1	w_2	...	w_j	...	w_S	
	SNP 1	SNP 2	...	SNP j	...	SNP S	PRS
Ind 1							PRS_1
Ind 2							PRS_2
...							...
ind i				g_{ij}			PRS_i
...							...
ind N							PRS_N

$$PRS_i = \sum_{j=1}^S w_j \cdot g_{ij}$$



SNP weights (PGS)

PGS Catalog 📄 <https://www.pgscatalog.org/>

Cancer-PRSweb 📄 <https://prsweb.sph.umich.edu:8443/>

LDpred2
PLINK
PRSice
lassosum
PRS-CSx
PRS-CS

Set of SNPs



$$w_j = \beta_j, j = 1..s$$

Genome-wide SNPs



$$w_j = \beta'_j, j = 1..S$$

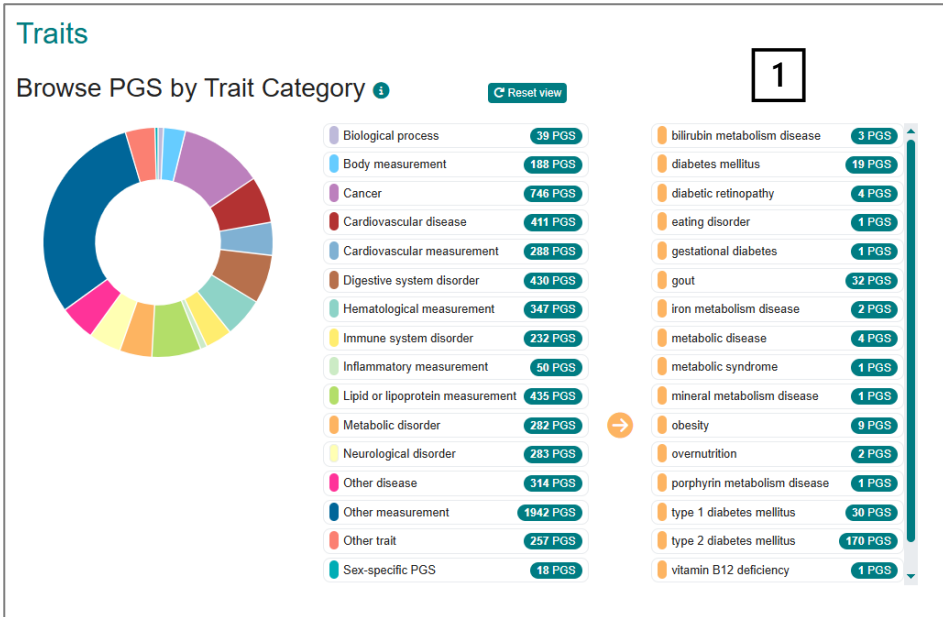
PRS, a composite measure derived from multiple genetic variants, can be simply treated as a linear combination of genotypes.

In regression modeling (linear combination), incorporating more predictors may enhance the R^2 (the proportion of variation explained by the model). However, it may cause overfitting due to the of model complexity, noise, or large betas (sensitive to minor changes).

Beta shrinkage (small beta) can help make beta toward zero that decreases model complexity and sensitivity to minor changes.

Method – External PRS

PGS Catalog



Associated Polygenic Score(s)

Filter PGS by Participant Ancestry i

Individuals included in:

- Any Stage [G, D, E]
- Any Stage [G, D, E]
- All Stages combined [G + D + E]
- Development [G, D]
- GWAS [G]
- Score development [D]
- PGS Evaluation [E]

2

List of ancestries includes:

-
-
- African
- East Asian
- South Asian
- Additional Asian Ancestries
- European
- Greater Middle Eastern
- Hispanic or Latin American
- Additional Diverse Ancestries
- Not Reported

3

Ancestry legend i

- Multi-ancestry (including European)
- Multi-ancestry (excluding European)
- African
- East Asian
- South Asian
- Additional Asian Ancestries
- European
- Greater Middle Eastern
- Hispanic or Latin American
- Additional Diverse Ancestries
- Not Reported

4

Show only Multi-ancestry data i

5

Polygenic Score ID & Name	PGS Publication ID (PGP)	Reported Trait	Mapped Trait(s) (Ontology)	Number of Variants
---------------------------	--------------------------	----------------	----------------------------	--------------------

Performance Metrics i

Disclaimer: The performance metrics are displayed as reported by the source studies. It is important to note that metrics are not necessarily comparable with each other. For example, (described by the PGS Catalog Sample Set [PSS] ID), phenotyping, and statistical modelling. Please refer to the source publication for additional guidance on performance.

6

PGS Performance Metric ID (PPM)	Evaluated Score	PGS Sample Set ID (PSS)	Performance Source	Trait	PGS Effect Sizes (per SD change)	Classification Metrics	Other Metrics
---------------------------------	-----------------	-------------------------	--------------------	-------	----------------------------------	------------------------	---------------

Method – External PRS

PGS Catalog

The PGS Catalog website interface. At the top, it says 'The Polygenic Score (PGS) Catalog' and 'An open database of polygenic scores and the relevant metadata required for accurate application and evaluation.' Below this is a search bar with the text 'Search the PGS Catalog' and a search button. A pink banner indicates 'Availability from: pgsc_eu_eu' and 'A repository website to capture both PGS Catalog and custom polygenic scores'. Below this is a section 'Explore the Data' with three buttons: 'Polygenic Scores' (4,650), 'Traits' (651), and 'Publications' (596). The 'Downloads' section is highlighted, showing a 'Table of Contents' with links to 'Available PGS Catalog downloads', 'PGS Catalog FTP structure', 'PGS Catalog Metadata', 'PGS Scoring Files', and 'Harmonized Files'. The 'PGS Scoring Files' section is expanded, showing links to 'Formatted Files' (Header, Columns, Format changes, Previous formats) and 'Harmonized Files' (File name, Header, Additional Columns).

Available PGS Catalog downloads

The PGS Catalog downloads section. It contains four main categories: 'PGS Scoring Files & Metadata' (Individual PGS variants scoring and metadata files) with a link to 'View PGS Score Directories (FTP)'; 'PGS Catalog Metadata' (Available PGS global metadata files) with a link to 'Bulk Metadata Downloads (FTP)'; 'PGS Catalog REST API' (Programmatic access to the PGS Catalog metadata) with a link to 'REST API endpoint documentation'; and 'Python package pgscatalog_utils' (A collection of tools, such as scoring files download) with a link to 'Python package documentation'.

The PGS Catalog FTP directory structure. The root path is 'ftp://ftp.ebi.ac.uk/pub/databases/spot/pgs'. The directory contains 'pgs_scores_list.txt' (list of Polygenic Score IDs), 'metadata/' (containing 'pgs_all_metadata.xlsx', 'pgs_all_metadata_[sheet_name].csv' (7 files), 'pgs_all_metadata.tar.gz' (xlsx + csv files), 'publications/' (metadata for large studies), and 'previous_releases/'), 'scores/' (containing 'PGS000001/' and 'PGS000002/'), and 'PGS00XXXX/'.

Using reported SNP weights (e.g., [PGS catalog](#) and [Cancer-PRSweb](#)) to calculate the PRS on the target data

```
###PGS CATALOG SCORING FILE - see https://www.pgscatalog.org/downloads/#dl_ftp_scoring for additional information
#format_version=2.0
##POLYGENIC SCORE (PGS) INFORMATION
#pgs_id=PGS000001
#pgs_name=PRS77_BC
#trait_reported=Breast cancer
#trait_mapped=breast carcinoma
#trait_efo=EFO_0000305
#genome_build=NR
#variants_number=77
#weight_type=NR
##SOURCE INFORMATION
#pgp_id=PGP000001
#citation=Mavaddat N et al. J Natl Cancer Inst (2015). doi:10.1093/jnci/djv036
##HARMONIZATION DETAILS
#HmPOS_build=GRCh38
#HmPOS_date=2022-07-29
#HmPOS_match_chr={"True": null, "False": null}
#HmPOS_match_pos={"True": null, "False": null}
```

rsID	chr_name	effect_allele	other_allele	effect_weight	locus_name	OR	hm_source	hm_rsID	hm_chr	hm_pos	hm_inferOtherAllele
rs78540526	11	T	C	0.16220388	CCND1	1.1761	ENSEMBL	rs78540526	11	69516650	
rs75915166	11	A	C	0.023618866	CCND1	1.0239	ENSEMBL	rs75915166	11	69564393	
rs554219	11	G	C	0.1167158	CCND1	1.1238	ENSEMBL	rs554219	11	69516874	
rs7726159	5	A	C	0.035270614	TERT	1.0359	ENSEMBL	rs7726159	5	1282204	
rs10069690	5	T	C	0.02391182	TERT	1.0242	ENSEMBL	rs10069690	5	1279675	
rs2736108	5	T	C	-0.064111945	TERT	0.9379	ENSEMBL	rs2736108	5	1297373	
rs2588809	14	T	C	0.064569771	RAD51L1	1.0667	ENSEMBL	rs2588809	14	68193711	
rs999737	14	T	C	-0.079151438	RAD51L1	0.9239	ENSEMBL	rs999737	14	68567965	

Method – External PRS

```
$ plink --bfile dat_auto_qc \
--score PGS000001_hmPOS_GRCh38.txt 9 3 5 \
--out dat_auto_qc

$ plink --bfile dat_auto_qc \
--score PGS000001_hmPOS_GRCh38.txt 9 3 5 sum \
--out dat_auto_qc
```

PGS000001_hmPOS_GRCh38.txt

```
###PGS CATALOG SCORING FILE - see https://www.pgscatalog.org/downloads/#dl_ftp_scoring for additional information
#format_version=2.0
##POLYGENIC SCORE (PGS) INFORMATION
#pgs_id=PGS000001
#pgs_name=PRS77_BC
#trait_reported=Breast cancer
#trait_mapped=breast carcinoma
#trait_eto=EFO_0000305
#genome_build=NR
#variants_number=77
#weight_type=NR
##SOURCE INFORMATION
#pgp_id=PGP000001
#citation=Mavaddat N et al. J Natl Cancer Inst (2015). doi:10.1093/jnci/djv036
##HARMONIZATION DETAILS
#HmPOS_build=GRCh38
#HmPOS_date=2022-07-29
#HmPOS_match_chr={"True": null, "False": null}
#HmPOS_match_pos={"True": null, "False": null}
rsID      chr_name  effect_allele  other_allele  effect_weight  locus_name  OR      hm_source  hm_rsID  hm_chr  hm_pos  hm_inferOtherAllele
rs78540526 11        T              C              0.16220388    CCND1      1.1761  ENSEMBL   rs78540526 11      69516650
rs75915166 11        A              C              0.023618866   CCND1      1.0239  ENSEMBL   rs75915166 11      69564393
rs554219    11        G              C              0.1167158     CCND1      1.1238  ENSEMBL   rs554219    11      69516874
rs7726159   5         A              C              0.035270614   TERT       1.0359  ENSEMBL   rs7726159   5       1282204
rs10069690  5         T              C              0.02391182    TERT       1.0242  ENSEMBL   rs10069690  5       1279675
rs2736108   5         T              C              -0.064111945  TERT       0.9379  ENSEMBL   rs2736108   5       1297373
rs2588809   14        T              C              0.064569771   RAD51L1    1.0667  ENSEMBL   rs2588809   14      68193711
rs999737    14        T              C              -0.079151438  RAD51L1    0.9239  ENSEMBL   rs999737    14      68567965
```

3

5

9

dat_auto_qc.profile

FID	IID	PHENO	CNT	CNT2	SCORE

FID	IID	PHENO	CNT	CNT2	SCORSUM

$$PRS_i = \sum_{j=1}^S \frac{w_j \cdot g_{ij}}{2 \cdot N_i}$$

N_i = non-missing SNPs in sample i

$$PRS_i = \sum_{j=1}^S w_j \cdot g_{ij}$$

Preparation

Base data (summary statistic)

ID	SNP ID, same representation as in target data, usually rsnumber
CHR	chromosome, same genome build as in target data
BP	physical position, same genome build as in target data
A1	effect allele
A2	other alleles
OR/BETA	estimate
SE	standard error of BETA
P	p-value
N	sample size

LD information

PLINK / PLINK2	depend on target data
PRSize2	depend on target data
LDpred2	GRCh37 & 38 HapMap3 LD blocks and LD matrix
Lassosum	GRCh37 & 38 1000 genomes project Phase I LD blocks
PRS-CS / PRS-CSx	GRCh37 1000 genomes project LD UK Biobank LD

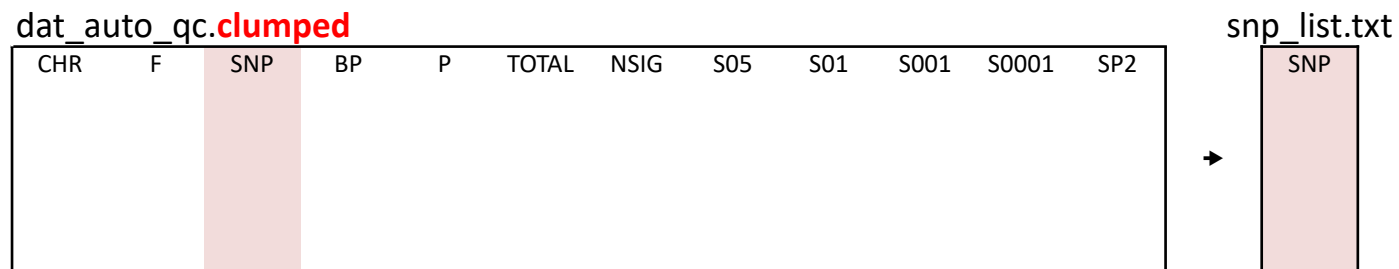
Target data

Genotype	.bed, .bim, .fam
covariate	plink-formatted file
phenotype	plink-formatted file

Method – C+T (Clumping + Thresholding)

```
$ plink --bfile dat_auto_qc \  
  --clump sumStat.txt \  
  --clump-snp-field ID \  
  --clump-field P \  
  --clump-p1 1 --clump-p2 1 --clump-r2 0.2 --clump-kb 500 \  
  --out dat_auto_qc
```

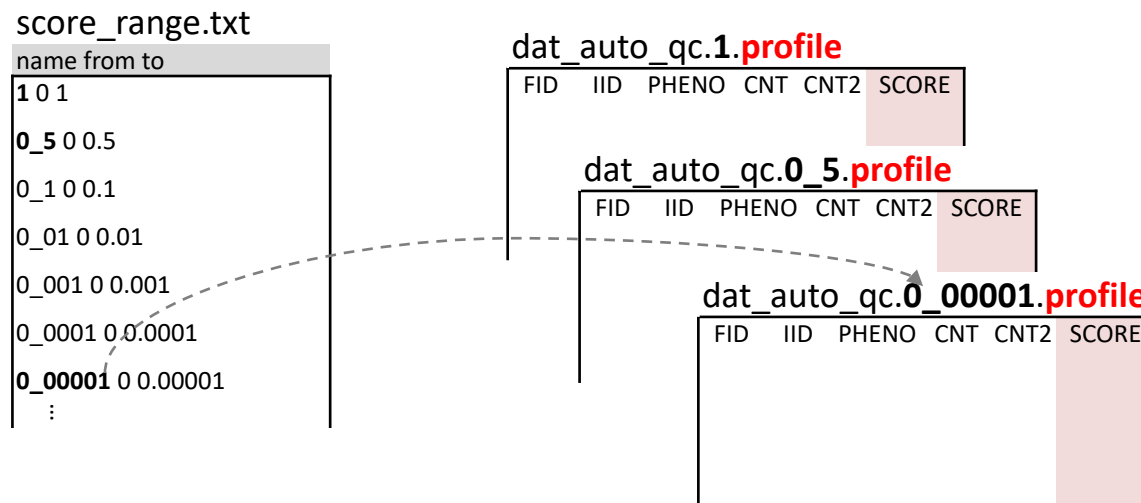
```
$ awk '{print $3}' dat_auto_qc.clumped > snp_list.txt
```



Each row is a clump of markers
indexed by the 'SNP' column (smallest p-value)

Method – C+T (Clumping + Thresholding)

```
$ plink --bfile dat_auto_qc \  
--extract snp_list.txt \  
--score sumStat.txt #id #a1 #b \  
--q-score-range score_range.txt sumStat.txt #id #p \  
--out dat_auto_qc
```



Method – PRSice2

- <https://choishingwan.github.io/PRSice/>

Home

QUICK START

PRSet

DETAIL GUIDES

PRSet

PRSet

Available Commands

DEVELOPERS

Compile from Source

Development Decisions

Useful Resources

MISC

Additional Steps for MAC and Window users

Frequently Asked Questions

Archive

Update Log

Docs » Home

Edit on GitHub

We're hiring!!

We are hiring!!

We are looking for several people to join our team at Mount Sinai in New York City - Postdoc and Faculty positions available! We need people with a strong background in computing/statistics interested in the themes of our lab (see our [lab website](#))

Please email paul.oreilly@mssm.edu if interested!

PRSice-2: Polygenic Risk Score software

PRSice (pronounced 'precise') is a Polygenic Risk Score software for calculating, applying, evaluating and plotting the results of polygenic risk scores (PRS) analyses. Some of the features include:

1. High-resolution scoring (PRS calculated across a large number of P-value thresholds)
2. Identify Most predictive PRS
3. Empirical P-values output (not subject to over-fitting)
4. Genotyped (PLINK binary) and imputed (Oxford bgen v1.2) data input
5. Biobank-scale genotyped data can be analysed within hours
6. Incorporation of covariates
7. Application across multiple target traits simultaneously
8. Results plotted in several formats (bar plots, high-res plots, quantile plots)
9. PRSet: function for calculating PRS across user-defined pathways / gene sets

Executable downloads DOI: 10.5281/zenodo.3703335

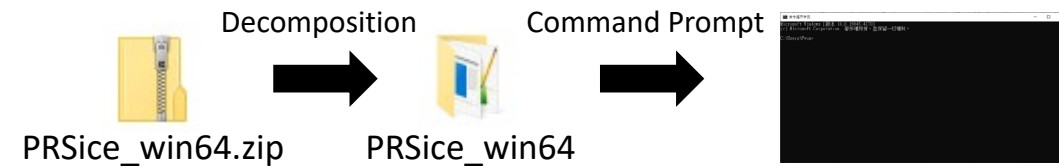
Coverage Status

Operating System	Link
Linux 64-bit	v2.3.5
OS X 64-bit	v2.3.5
Windows 32-bit	Not available
Windows 64-bit	v2.3.5

- **Linux**

```
$ wget https://github.com/choishingwan/PRSice/releases/download/2.3.5/PRSice_linux.zip
$ unzip PRSice_linux.zip -d PRSice
$ cd PRSice
$ Rscript PRSice.R --prsice PRSice_linux
```

- **Windows**



```
> cd PRSice_win64
> Rscript PRSice.R --prsice PRSice_win64.exe
```

Method – PRSice2

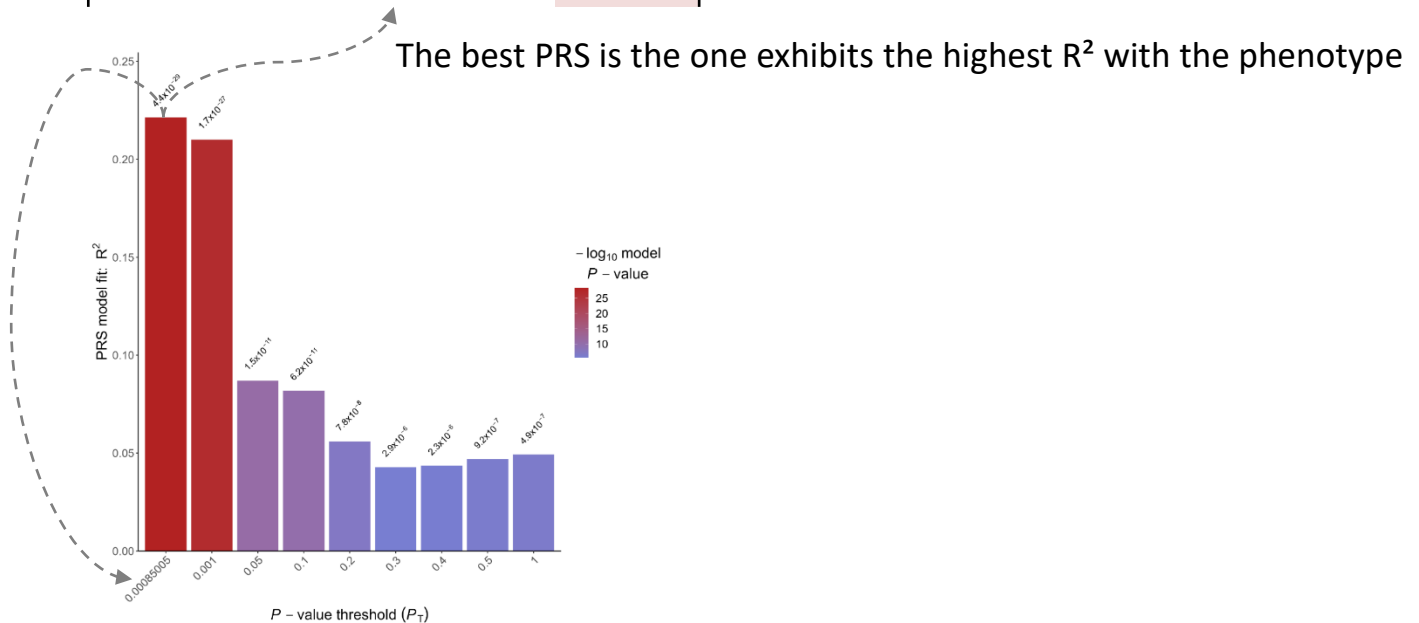
```
$ Rscript PRSice.R \
  --prsice PRSice_linux \
  --target dat_auto_qc \
  --base sumStat.txt \
  --binary-target T \
  --snp ID \
  --A1 A1 \
  --stat BETA \
  --pvalue P \
  --beta \
  --pheno pheCov.txt \
  --pheno-col bt_1 \
  --cov pheCov.txt \
  --cov-col age,sex,@pc[1-10] \
  --out dat_auto_qc
```

dat_auto_qc.summary

Phenotype	Set	Threshold	PRS.R2	Full.R2	Null.R2	Prevalence	Coefficient	Standard.Error	P	Num_SNP
-----------	-----	-----------	--------	---------	---------	------------	-------------	----------------	---	---------

dat_auto_qc.best

FID	IID	In_Regression	PRS
-----	-----	---------------	-----



Method – lassosum

```
> install.packages(c("devtools","RcppArmadillo", "data.table", "Matrix"), dependencies=TRUE)
> devtools::install_github("tshmak/lassosum")

> library(lassosum); library(data.table)
> phecov <- fread("pheCov.txt")
> ss <- fread("sumStat.txt")
> cor <- p2cor(p = ss$P, n = N, sign = log(ss$OR)) # sign = ss$BETA
> out <- lassosum.pipeline(cor = cor, chr = ss$CHR, pos = ss$BP, A1 = ss$A1, A2 = ss$A2, ref.bfile = "dat_auto_qc", test.bfile = "dat_auto_qc", LDblocks = "ASN.hg19")
```

Method – lassosum

```
> result <- validate(out, pheno = phecov$qt_1, covar = phecov[, c("age","sex",paste0("pc",1:10))])  
> ss_ <- data.table(ss[out$sumstats$order][,sbeta:= result$best.beta])
```

ID	CHR	BP	A1	A2	OR	SE	P	N	sbeta

shrinkage beta

```
> result$results.table
```

FID	IID	pheno	best.prs

Method – PRS-CS / PRS-CSx

```
$ git clone https://github.com/getian107/PRScs.git
```

```
$ git clone https://github.com/getian107/PRScsx.git
```

```
$ mkdir LD_ref
```

```
$ wget -O LD_ref/ldblk_1kg_amr.tar.gz https://www.dropbox.com/s/uv5ydr4uv528lca/ldblk_1kg_amr.tar.gz?dl=0
```

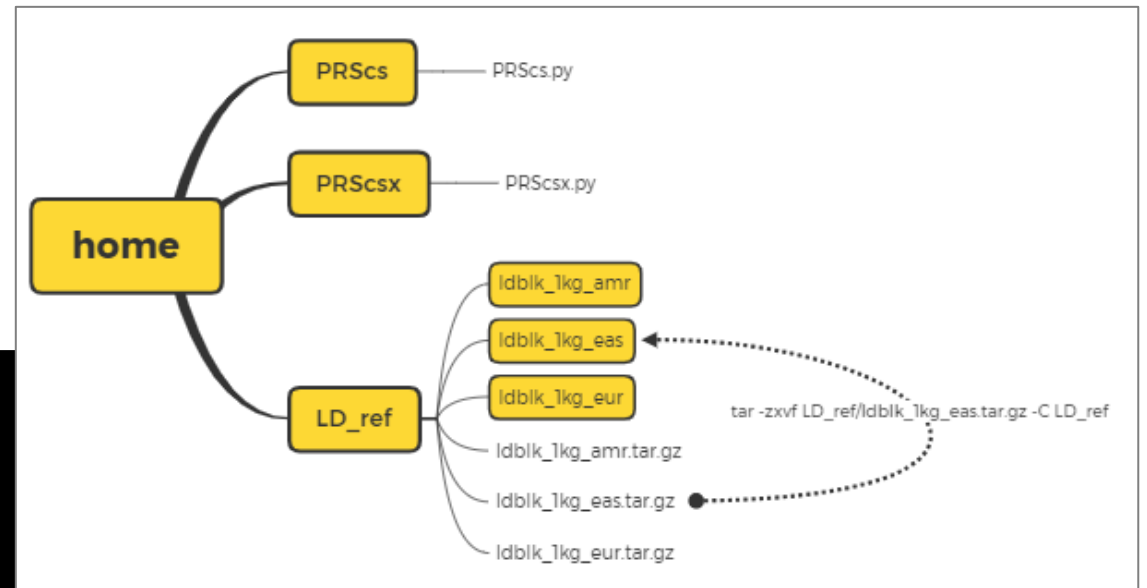
```
$ wget -O LD_ref/ldblk_1kg_eas.tar.gz https://www.dropbox.com/s/7ek4lwwf2b7f749/ldblk_1kg_eas.tar.gz?dl=0
```

```
$ wget -O LD_ref/ldblk_1kg_eur.tar.gz https://www.dropbox.com/s/mt6var0z96vb6fv/ldblk_1kg_eur.tar.gz?dl=0
```

```
$ tar -zxvf LD_ref/ldblk_1kg_amr.tar.gz -C LD_ref
```

```
$ tar -zxvf LD_ref/ldblk_1kg_eas.tar.gz -C LD_ref
```

```
$ tar -zxvf LD_ref/ldblk_1kg_eur.tar.gz -C LD_ref
```



Method – PRS-CS / PRS-CSx

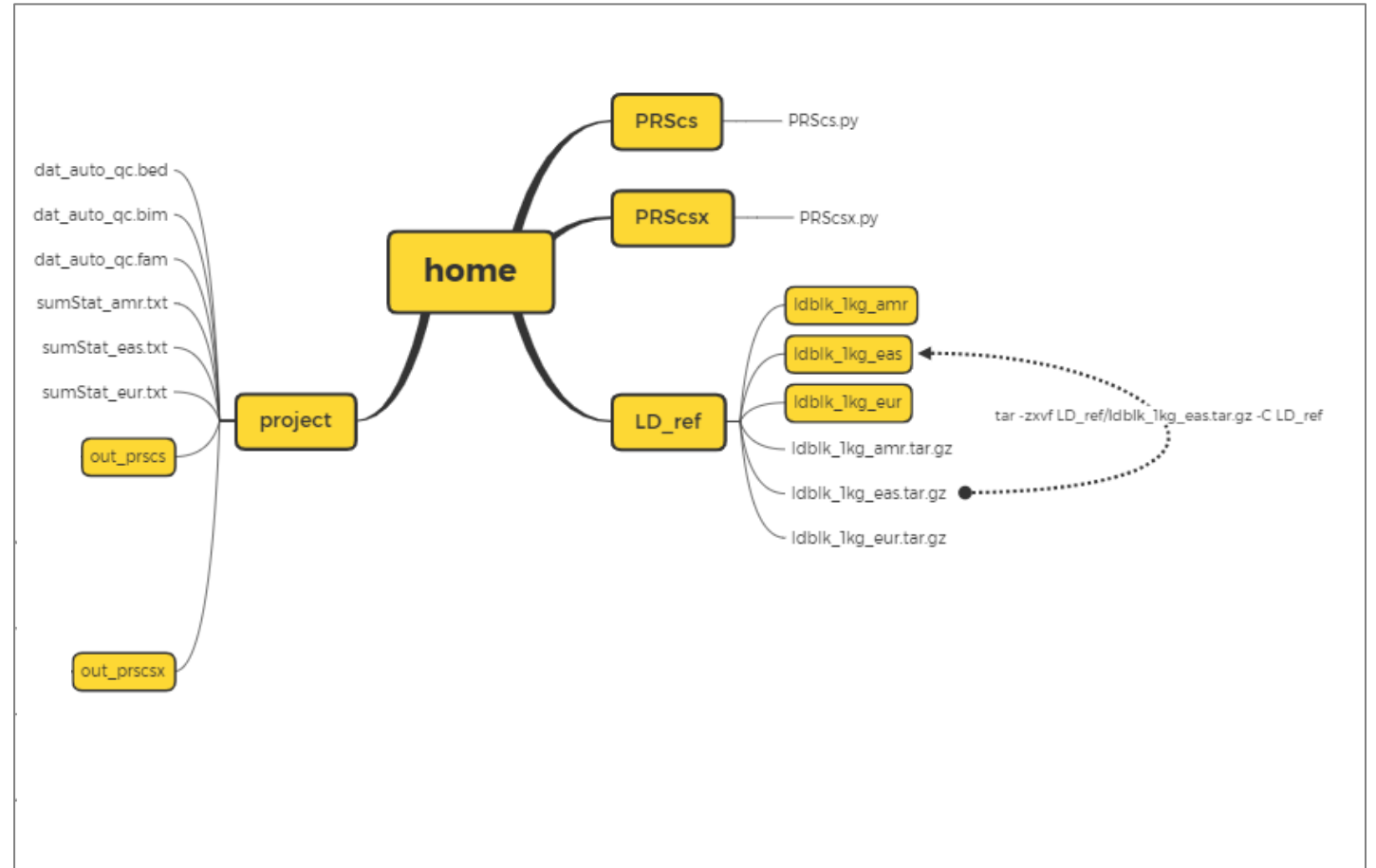
sumStat_*.txt **recommended**

SNP	A1	A2	BETA/OR	SE

sumStat_*.txt

SNP	A1	A2	BETA/OR	P

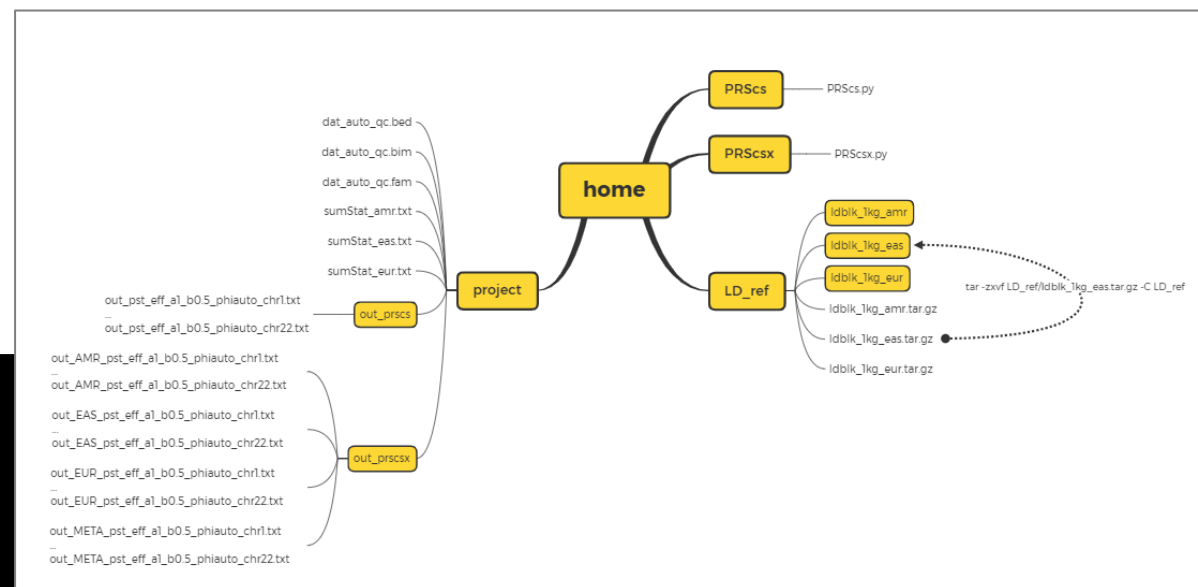
SNP column is **rsID**, because the representation for SNP in provided LD information is rsID



Method – PRS-CS / PRS-CSx

```
$ cd PRScs
$ python3 PRScs.py --ref_dir=../LD_ref/ldblk_1kg_eas \
--bim_prefix=../project/dat_auto_qc \
--sst_file=../project/sumStat_eas.txt \
--n_gwas=N --seed=1 --out_dir=../project/out_prscs/out
```

```
$ mkdir dis_prscsx
$ cd PRScsx
$ python3 PRScsx.py --ref_dir=../LD_ref \
--bim_prefix=../project/dat_auto_qc \
--sst_file=../project/sumStat_amr.txt,../project/sumStat_eas.txt,../project/sumStat_eur.txt \
--n_gwas=N_amr,N_eas,N_eur --pop=AMR,EAS,EUR \
--seed=1 --meta=TRUE --out_dir=../project/out_prscsx --out_name=out
```



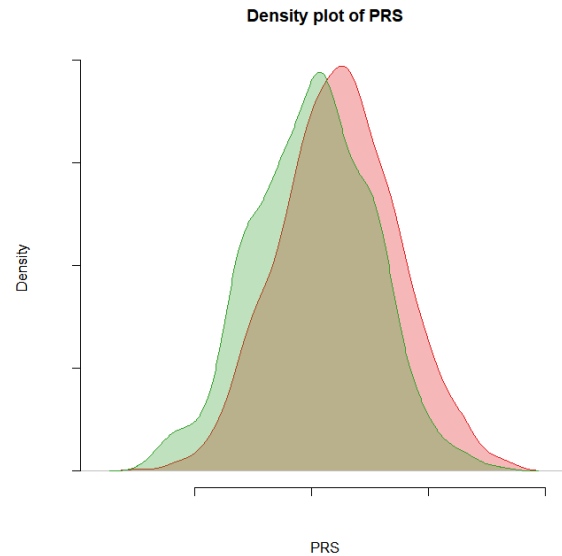
dat_auto_qc_pst_eff_a[1]_b[0.5]_phiauto_chr*.txt

CHR	RSID	BP	A1	A2	Weight

PRS – Figures

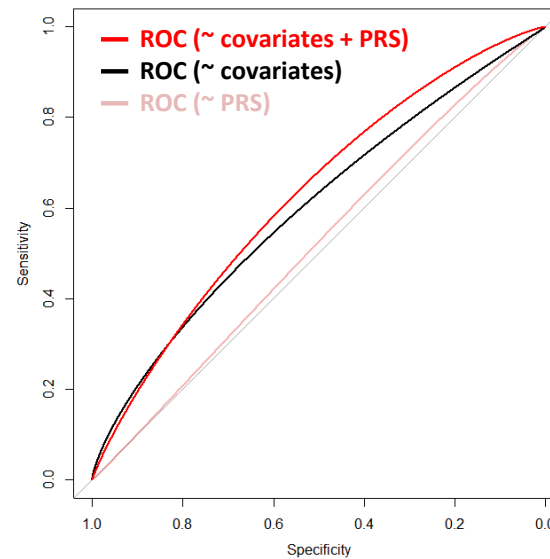
Density plot

Use it to understand patterns, trends, and the underlying structure of numeric data among groups.



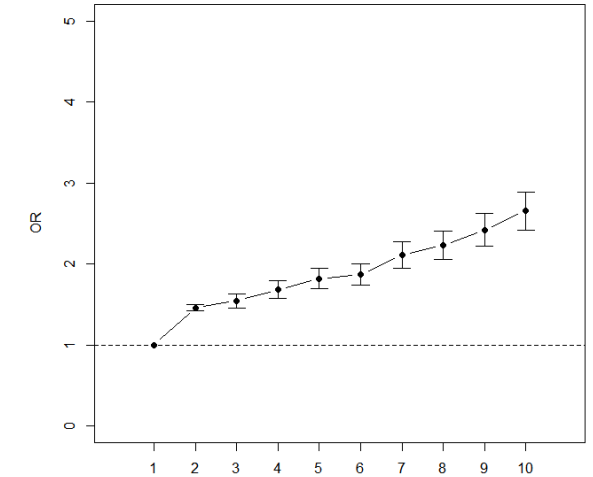
ROC curve

Use it to evaluate binary classifiers and understand their discrimination ability.



OR decile plot

It categorizes large data sets into 10 equally sized subsections (deciles) based on a given metric (e.g., OR), and then fit a logistic regression model with a binary trait and a categorized *PRS* as the predictor



95% CI_{delta} of OR = [OR \pm 1.96 \times OR \times se(BETA)]
95% CI_{MLE} of OR = [exp(BETA \pm 1.96 \times se(BETA))]

log(OR) = BETA

Note

- **QC for base data**

- Duplicated SNPs: it occurs an error when using `plink --score` to calculate PRS
- Ambiguous SNPs: if there is no information about strands of base and target data, exclude them!
- Mismatch SNPs: when using `plink --score` to calculate PRS, it treats flipped alleles of a SNP as distinct

- **QC for target data**

- Array data: GWAS QC
- Imputation data: Sample QC (based on array sample QC) + Variant QC (infoscore, CR, MAF)

- **Software usage**

- Consistence of genome builds: base data, target data, and **reference data**
- Practice <https://choishingwan.github.io/PRS-Tutorial/>
<https://privefl.github.io/bigsnp/articles/>

Software	LD resource	LD genome build
PLINK	target data	
PRSice2	target data	
LDpred2	HapMap3	hg18, GRCh37 (hg19), GRCh38 (hg38)
lassosum	1000 genomes project Phase I	GRCh37 (hg19), GRCh38 (hg38)
PRS-CS/PRS-CSx	1000 genomes project Phase 3 UK Biobank	GRCh37 (hg19)

Note

- **Enhancing disease risk prediction beyond PRS**

- As we know, **DNA is relatively stable throughout an individual's life**. Therefore, **relying solely on a PRS prediction model is insufficient**. To more accurately assess disease risk, we must consider additional factors or covariates: demographic variables (age and gender), environmental variables (abc-covariates), etc.

- metaGRS (Meta-Genomic Risk Score)

$$\begin{array}{c} \text{Disease status} \end{array} = \begin{array}{c} \text{PRS for disease} \end{array} + \begin{array}{c} \text{PRS for subtypes} \end{array} + \begin{array}{c} \text{PRS for exposures} \end{array} + \begin{array}{c} \text{PRS for risk factors} \end{array} + \dots$$

- **Healthcare application**

- Hingorani, Aroon D., et al. (2023) Performance of polygenic risk scores in screening, prediction, and risk stratification: secondary analysis of data in the Polygenic Score Catalog. BMJ medicine 2.1 – **Disapproval of PRS**

Imputation server

Taiwan Biobank Imputation Server

TWB Imputation Server @ NCHC

Home Help Contact


Sign up Login

Taiwan Biobank Imputation Server


Free Next-Generation Genotype Imputation Service

Sign up now Login


The easiest way to impute genotypes



Upload your **genotypes** to our server
located in Taiwan.
All interactions with the server are **secured**.



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.

Wide-range of reference panels supported

HapMap
Phase 2

Taiwan Biobank 1.5k

1000 Genomes
Phase 3

Michigan Imputation Server 2

Michigan Imputation Server 2

Home Help Contact

Sign up Login

Michigan Imputation Server 2

Free Next-Generation Genotype Imputation Platform

Sign up now Login

120.3M
Imputed Genomes

13350
Registered Users

2
Running Jobs

15 September 2024

We have successfully migrated to a new architecture and released Michigan Imputation Server 2!

Breaking changes

- We've updated our Quality Control process to include allele swap checks, a necessary change to align with recent updates in Minimac4 for improved data accuracy. Please click [here](#) for more details.
- The Michigan Imputation Server now requires Imputation Bot version 2.x.x or higher to function correctly. If you are using version 1.0.0 or lower, please update and reinstall Imputation Bot to avoid errors. Please click [here](#) for more details.
- We have updated the API. Please click [here](#) for more details.

[More News](#)

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

TOPMed Imputation Server

 **BioData CATALYST**
TOPMed Imputation Server

Home About Help Contact

Sign up Login

Users are suggested to use professional email addresses. Email addresses from free mail providers (such as gmail.com) will soon be unable to register for accounts. Some accounts associated with free providers have already been disabled. If you have any questions or need access, please contact us at imputationserver@umich.edu.

TOPMed Imputation Server

Free Next-Generation Genotype Imputation Service


Sign up now Login

76.8M
Imputed Genomes


5723
Registered Users

17
Active Jobs


The easiest way to impute genotypes



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



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.

The TOPMed Imputation Server is powered by software invented and developed by the [University of Michigan](#) and driven by data provided by the investigators of the [TOPMed Program](#).

20

Imputation server

Taiwan Biobank Imputation Server

The image displays three overlapping screenshots of the Taiwan Biobank Imputation Server interface.

Leftmost Screenshot: Shows the main landing page. The header includes "TWB Imputation Server @ NCHC" and navigation links (Home, Help, Contact). The main heading is "Taiwan Biobank Imputation Server" with the subtitle "Free Next-Generation Genotype Imputation Service". Below this are "Sign up now" and "Login" buttons. The main content area features the text "The easiest way to impute genotypes" and two icons: "Upload your genotypes to our server" and "Choose a reference panel". At the bottom, it lists "Wide-range of reference panels supported" with three options: HapMap Phase 2, Taiwan Biobank 1.5k, and 1000 Genomes Phase 3.

Middle Screenshot: Shows the "Sign in" modal. It includes fields for "Username:" (containing "jiawei") and "Password:" (masked with dots). A "Sign in" button is present. Below the modal, there are links for "New user? Sign up for free" and "Forgotten your password? Reset your password".

Rightmost Screenshot: Shows the "Genotype Imputation (Minimac4)" page. The header includes "TWB Imputation Server @ NCHC" and navigation links (Home, Run, Jobs, Help, Contact). The main heading is "Taiwan Biobank Imputation Server" with the subtitle "Free Next-Generation Genotype Imputation Service". Below this is a "Run" button. The main content area features the text "The easiest way to impute genotypes" and three icons: "Upload your genotypes to our server", "Choose a reference panel", and "Download the results". At the bottom, it lists "Wide-range of reference panels supported" with three options: HapMap Phase 2, Taiwan Biobank 1.5k, and 1000 Genomes Phase 3.

TWB Imputation Server @ NCHC

HomeRunJobsHelpContact

jiawei

Genotype Imputation (Minimac4) 1.6.7

This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found [here](#).

If your input data is **GRCh37/hg19** please ensure chromosomes are encoded without prefix (e.g. **20**).
If your input data is **GRCh38hg38** please ensure chromosomes are encoded with prefix 'chr' (e.g. **chr20**). <https://imputationserver.readthedocs.io>

Run

Name

optional job name

Reference Panel (Details)

-- select an option --

-- select an option --

1000G Phase 3 v5

HapMap 2

TWB hg38 1.5k

Input Files (VCF)

Select Files

Multiple files can be selected by using the **ctrl** / **cmd** or **shift** keys.

Array Build

GRCh37/hg19

Please note that the final SNP coordinates always match the reference build.

rsq Filter

off

Phasing

Eagle v2.4 (phased output)

Population

-- select an option --

Mode

Quality Control & Imputation

☐ AES 256 encryption

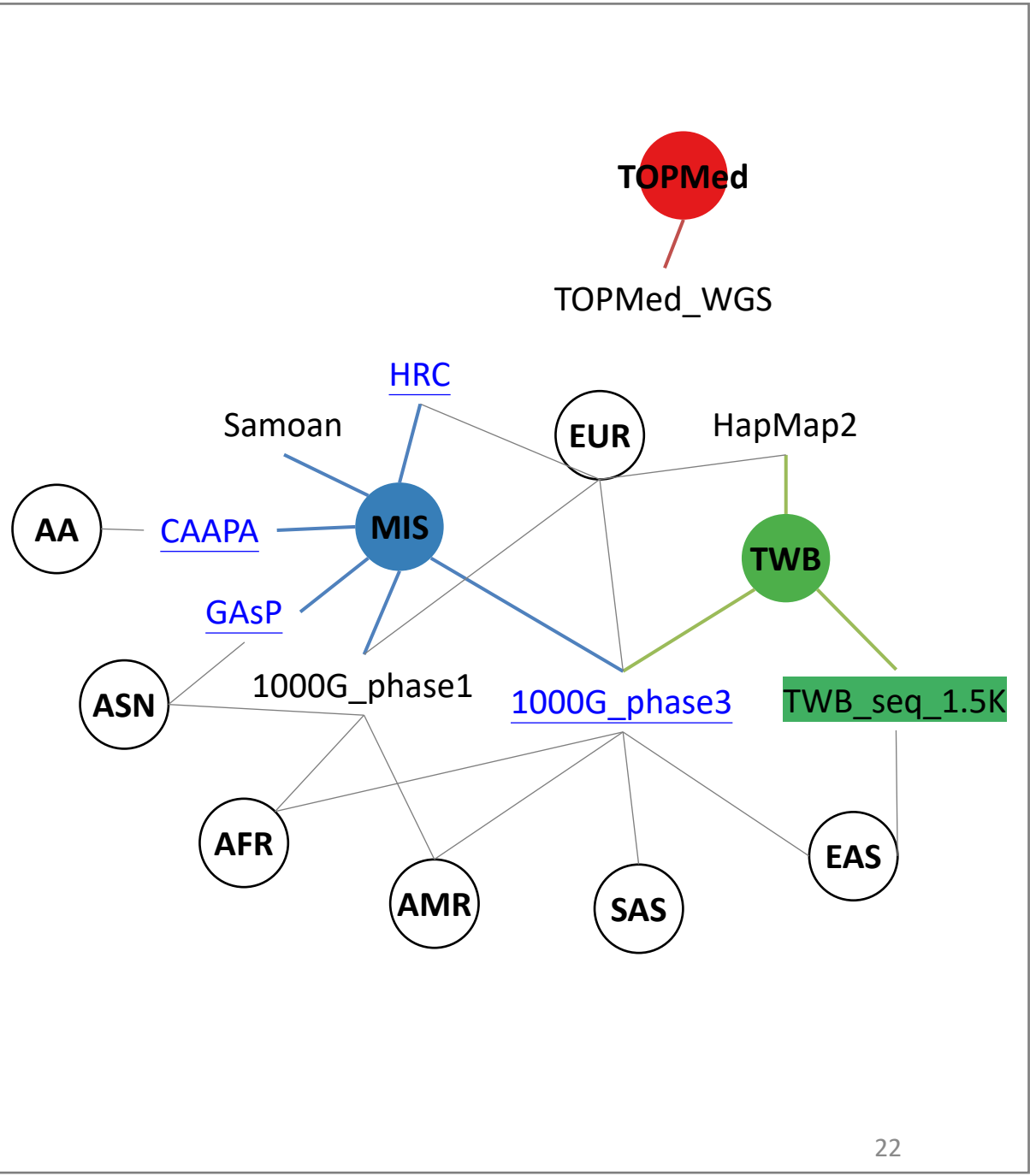
Imputation Server encrypts all zip files by default. Please note that AES encryption does not work with standard unzip programs. Use 7z instead.

☐ Generate Meta-imputation file

☐ I will not attempt to re-identify or contact research participants.

☐ I will report any inadvertent data release, security breach or other data management incident of which I become aware.

Submit Job



TWB Imputation Server @ NCHC Home Run Jobs Help Contact jiawei

Genotype Imputation (Minimac4) 1.6.7

This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found [here](#).

If your input data is **GRCh37/hg19** please ensure chromosomes are encoded without prefix (e.g. **20**).
If your input data is **GRCh38hg38** please ensure chromosomes are encoded with prefix 'chr' (e.g. **chr20**). <https://imputationserver.readthedocs.io>

Run

Name

optional job name

Reference Panel [\(Details\)](#)

-- select an option --

Input Files [\(VCF\)](#)

File Upload

File Upload

URLs (HTTP)

Secure File Transfer Protocol (SFTP)

S3 Bucket

Select Files

Multiple files can be selected by using the **ctrl** / **cmd** or **shift** keys.

Array Build

GRCh37/hg19

Please note that the final SNP coordinates always match the reference build.

rsq Filter

off

Phasing

Eagle v2.4 (phased output)

Population

-- select an option --

Mode

Quality Control & Imputation

☒ AES 256 encryption

Imputation Server encrypts all zip files by default. Please note that AES encryption does not work with standard unzip programs. Use 7z instead.

☒ Generate Meta-imputation file

☒ I will not attempt to re-identify or contact research participants.

☒ I will report any inadvertent data release, security breach or other data management incident of which I become aware.

Submit Job

- 1 chromosome representation in GRCh37 and GRCh38 are different
- 2 variants are ordered by genomic positions
- 3 save as **.vcf.gz** by chromosomes

Genotype Imputation (Minimac4) 1.6.7

This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found [here](#).

If your input data is **GRCh37/hg19** please ensure chromosomes are encoded without prefix (e.g. **20**).

If your input data is **GRCh38hg38** please ensure chromosomes are encoded with prefix 'chr' (e.g. **chr20**). <https://imputationserver.readthedocs.io>

Run

Name

optional job name

Reference Panel [\(Details\)](#)

-- select an option --

Input Files [\(VCF\)](#)

File Upload

Select Files

Multiple files can be selected by using the **ctrl** / **cmd** or **shift** keys.

Array Build

GRCh37/hg19

-- select an option --

GRCh37/hg19

GRCh38/hg38

rsq Filter

Phasing

Eagle v2.4 (phased output)

Population

-- select an option --

Mode

Quality Control & Imputation

☐ AES 256 encryption

Imputation Server encrypts all zip files by default. Please note that AES encryption does not work with standard unzip programs. Use 7z instead.

☐ Generate Meta-imputation file

☐ I will not attempt to re-identify or contact research participants.

☐ I will report any inadvertent data release, security breach or other data management incident of which I become aware.

Submit Job

TWB Imputation Server @ NCHC

HomeRunJobsHelpContact

jiawei

Genotype Imputation (Minimac4) 1.6.7

This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found [here](#).

If your input data is **GRCh37/hg19** please ensure chromosomes are encoded without prefix (e.g. **20**).
If your input data is **GRCh38/hg38** please ensure chromosomes are encoded with prefix 'chr' (e.g. **chr20**). <https://imputationserver.readthedocs.io>

Run

Name

optional job name

Reference Panel [\(Details\)](#)

-- select an option --

Input Files [\(VCF\)](#)

File Upload

Select Files

Multiple files can be selected by using the **ctrl** / **cmd** or **shift** keys.

Array Build

GRCh37/hg19

Please note that the final SNP coordinates always match the reference build.

rsq Filter

off

-- select an option --

off

0.001

0.1

0.2

0.3

Phasing

Population

Mode

Generate Meta-imputation file

I will not attempt to re-identify or contact research participants.

I will report any inadvertent data release, security breach or other data management incident of which I become aware.

Submit Job

1 0.3 is recommended, refer to https://genome.sph.umich.edu/wiki/MaCH_FAQ

Genotype Imputation (Minimac4) 1.6.7

This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found [here](#).

If your input data is **GRCh37/hg19** please ensure chromosomes are encoded without prefix (e.g. **20**).

If your input data is **GRCh38hg38** please ensure chromosomes are encoded with prefix 'chr' (e.g. **chr20**). <https://imputationserver.readthedocs.io>

[Run](#)

Name

Reference Panel [\(Details\)](#)Input Files [\(VCF\)](#)[Select Files](#)

Multiple files can be selected by using the **ctrl** / **cmd** or **shift** keys.

Array Build

Please note that the final SNP coordinates always match the reference build.

rsq Filter

Phasing

Population

Mode

☐ AES 256 encryption

Imputation Server encrypts all zip files by default. Please note that AES encryption does not work with standard unzip programs. Use 7z instead.

☐ Generate Meta-imputation file☐ I will not attempt to re-identify or contact research participants.☐ I will report any inadvertent data release, security breach or other data management incident of which I become aware.[Submit Job](#)

TWB Imputation Server @ NCHC

HomeRunJobsHelpContact

jiawei

Genotype Imputation (Minimac4) 1.6.7

This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found [here](#).

If your input data is **GRCh37/hg19** please ensure chromosomes are encoded without prefix (e.g. **20**).
If your input data is **GRCh38/hg38** please ensure chromosomes are encoded with prefix 'chr' (e.g. **chr20**). <https://imputationserver.readthedocs.io>

Run

Name

Reference Panel [\(Details\)](#)

TWB hg38 1.5k

Input Files [\(VCF\)](#)

File Upload

Select Files

Multiple files can be selected by using the **ctrl** / **cmd** or **shift** keys.

Array Build

GRCh37/hg19

Please note that the final SNP coordinates always match the reference build.

rsq Filter

off

Phasing

Eagle v2.4 (phased output)

Population

-- select an option --

Mode

-- select an option --

TWB hg38 1.5k

EAS

Other/Mixed

Please note that AES encryption does not work with standard unzip programs: use 7z instead.

☐ Generate Meta-imputation file

☐ I will not attempt to re-identify or contact research participants.

☐ I will report any inadvertent data release, security breach or other data management incident of which I become aware.

Submit Job

1 The process compares AFs between target data and the reference panel, filtering out variants with large discrepancies

2 Even if certain variants are excluded at this stage, imputed results for them remain based on the reference panel

3 'Other/Mixed' indicates the check is not performed; choose it when the population of target data is not listed

TWB Imputation Server @ NCHC

HomeRunJobsHelpContact

jiawei

Genotype Imputation (Minimac4) 1.6.7

This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found [here](#).

If your input data is **GRCh37/hg19** please ensure chromosomes are encoded without prefix (e.g. **20**).
If your input data is **GRCh38/hg38** please ensure chromosomes are encoded with prefix 'chr' (e.g. **chr20**). <https://imputationserver.readthedocs.io>

Run

Name

Reference Panel [\(Details\)](#)

1000G Phase 3 v5

Input Files [\(VCF\)](#)

File Upload

Select Files

Multiple files can be selected by using the **ctrl** / **cmd** or **shift** keys.

Array Build

GRCh37/hg19

Please note that the final SNP coordinates always match the reference build.

rsq Filter

off

Phasing

Eagle v2.4 (phased output)

Population

-- select an option --

Mode

-- select an option --

1000G Phase 3 v5

AFR

AMR

EAS

SAS

EUR

Other/Mixed

☐ I will not attempt to re-

☐ I will report any inadvertent data release, security breach or other data management incident of which I become aware.

Submit Job

1 The process compares AFs between target data and the reference panel, filtering out variants with large discrepancies

2 Even if certain variants are excluded at this stage, imputed results for them remain based on the reference panel

3 'Other/Mixed' indicates the check is not performed; choose it when the population of target data is not listed

TWB Imputation Server @ NCHC

HomeRunJobsHelpContact

jiawei

Genotype Imputation (Minimac4) 1.6.7

This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found [here](#).

If your input data is **GRCh37/hg19** please ensure chromosomes are encoded without prefix (e.g. **20**).
If your input data is **GRCh38/hg38** please ensure chromosomes are encoded with prefix 'chr' (e.g. **chr20**). <https://imputationserver.readthedocs.io>

Run

Name

optional job name

Reference Panel [\(Details\)](#)

HapMap 2

Input Files [\(VCF\)](#)

File Upload

Select Files

Multiple files can be selected by using the **ctrl** / **cmd** or **shift** keys.

Array Build

GRCh37/hg19

Please note that the final SNP coordinates always match the reference build.

rsq Filter

off

Phasing

Eagle v2.4 (phased output)

Population

-- select an option --

Mode

-- select an option --

HapMap 2

EUR

Mixed

Please note that AES encryption does not work with standard unzip programs: use 7z instead.

☐ Generate Meta-imputation file

☐ I will not attempt to re-identify or contact research participants.

☐ I will report any inadvertent data release, security breach or other data management incident of which I become aware.

Submit Job

1 The process compares AFs between target data and the reference panel, filtering out variants with large discrepancies

2 Even if certain variants are excluded at this stage, imputed results for them remain based on the reference panel

3 'Other/Mixed' indicates the check is not performed; choose it when the population of target data is not listed

Genotype Imputation (Minimac4) 1.6.7

This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found [here](#).

If your input data is **GRCh37/hg19** please ensure chromosomes are encoded without prefix (e.g. **20**).

If your input data is **GRCh38hg38** please ensure chromosomes are encoded with prefix 'chr' (e.g. **chr20**). <https://imputationserver.readthedocs.io>

[Run](#)

Name

Reference Panel [\(Details\)](#)Input Files [\(VCF\)](#)[Select Files](#)

Multiple files can be selected by using the **ctrl** / **cmd** or **shift** keys.

Array Build

Please note that the final SNP coordinates always match the reference build.

rsq Filter

Phasing

Population

Mode

Please note that AES encryption does not work with standard

☐ I will not attempt to re-identify or contact research participants.☐ I will report any inadvertent data release, security breach or other data management incident of which I become aware.[Submit Job](#)

Genotype Imputation (Minimac4) 1.6.7

This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found [here](#).


If your input data is **GRCh37/hg19** please ensure chromosomes are encoded without prefix (e.g. **20**).

If your input data is **GRCh38hg38** please ensure chromosomes are encoded with prefix 'chr' (e.g. **chr20**). <https://imputationserver.readthedocs.io>

 Run

Name

Reference Panel ([Details](#))

Input Files ([VCF](#))
 Select Files

Multiple files can be selected by using the **ctrl** / **cmd** or **shift** keys.

Array Build

Please note that the final SNP coordinates always match the reference build.

rsq Filter

Phasing

Population


Mode
☒ AES 256 encryption

Imputation Server encrypts all zip files by default. Please note that AES encryption does not work with standard unzip programs. Use 7z instead.

☒ Generate Meta-imputation file

☒ I will not attempt to re-identify or contact research participants.

☒ I will report any inadvertent data release, security breach or other data management incident of which I become aware.

 Submit Job

Thanks for your attention!!

<(_ _)>

Q & A

Q: rsID vs. hm_id in harmonized file of PGS Catalog

A: PGS Catalog 使用 Ensembl 以 chr/pos 做 genome build 的轉換，因此 hm_rsID 是在轉換後該位點的 rsnumber (有可能 merge 到新的 ID)

rsID	chr_name	chr_position	hm_rsID	hm_chr	hm_pos	
rs3795818	1	228685999		1	228498298	Note: rs3795818 is unobserved
rs28468602	13	80259570	rs3064655	13	79685435	Note: rs28468602 was merged into rs3064655
rs10744926	13	86912082		NA	NA	Note: rs10744926 is located on 13:2401 in GRCh38

Q: Using imputation data to calculate PRS by C+T

A: SOP 限制了我的想像，習慣將 imputation data 用 shrinkage beta 來計算 PRS；但用 C+T 也是可以，建議先用 infoscore 過濾位點後再使用 C+T，應該能期待得到比 array data 做 C+T 還不錯的結果

Q: PRSice2 shows the message "There are 1 region(s) with p-value less than 1e-5. Please note that these results are inflated due to the overfitting inherent in finding the best-fit PRS (but it's still best to find the best-fit PRS!). You can use the --perm option (see manual) to calculate an empirical P-value."

A: 在尋找 best p-value threshold 時，是以不同 threshold 所得到的 score 與 phenotype 做 association 後得到 p-value 來決定 (取最小者)；但擔心 overfitting 的問題 (套用到同樣 phenotype 但不同 dataset 時結果並不好)，建議用 permuted p-value 來決定 best p-value threshold

Q: PRS-CSx --meta

A: PRS-CSx 的 --meta 主要採 inverse-variance-weighted meta-analysis 於各族群估出的 weights (posterior effect sizes)，若有位點只存在於某一族群，其估計出的 weight 將與該族群相同；一般而言，--meta 得到的 variants 會是分析族群位點的聯集，附檔「生物資訊與分析_PRS_20250415.html」中，PRS-CSx 的分析結果可看到 EAS、EUR、META 各有 957101、1000352、1035617 個位點，且預測結果較 PRS-CS 為佳 (雖然 pattern 與理想情況仍有差距)