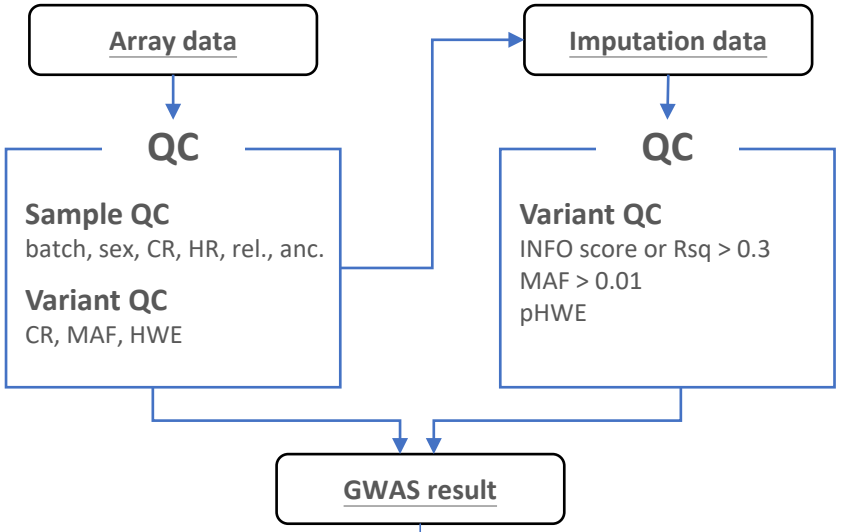
After the GWAS



COJO

COnditional & JOint analysis identifies independent signals within a locus by stepwise conditional analysis

MR

Mendelian Randomization uses individual variants as "instruments" to test if an exposure causally influences an outcome

GBA

Gene-Based Association aggregates variant-level signals into a single gene-level test, boosting power for polygenic genes

GSA

Gene-Set-based Association aggregates variant-level signals into a single gene-set-level test, revealing the biology underlying polygenic signal

FM

Fine-Mapping narrows a GWAS locus from hundreds of correlated SNPs to a credible set of likely causal variants by modeling LD structure

PRS

Polygenic Risk Score Sums genome-wide effect sizes into a per-individual genetic liability score

TWAS

Transcriptome-Wide Association Study links genetic risk to gene expression by testing whether predicted expression associates with a trait

GSE

Gene Set Enrichment analysis tests whether genes implicated by GWAS are statistically over-represented in biological pathways or gene sets

FA

Functional Annotation integrates genomic features to prioritize variants by biological plausibility, boosting power and interpretability

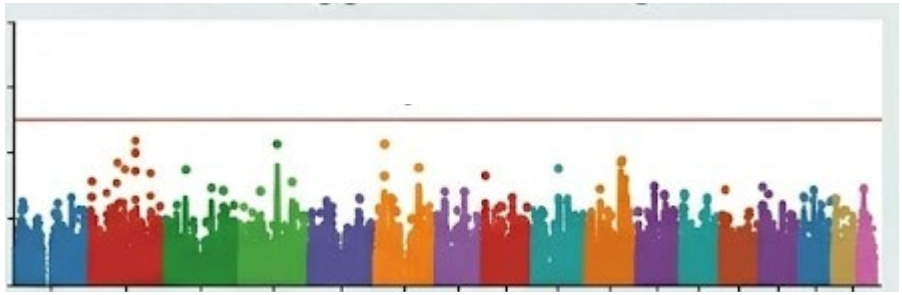
Coloc

Colocalization tests whether two traits share the same causal variant in a locus

GWAS result

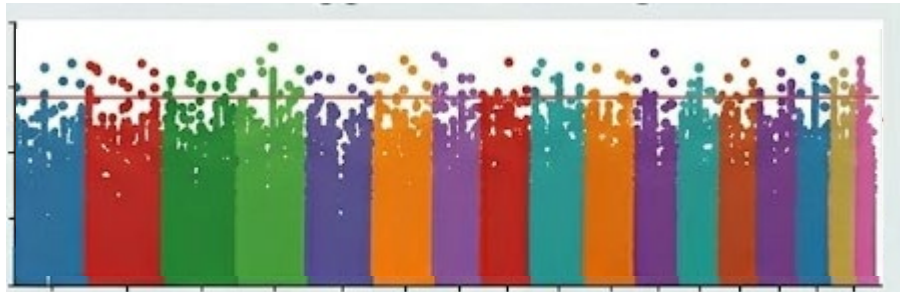
Although some peaks are observed but below the significance threshold, expanding the sample size or performing a meta-analysis with other GWAS cohorts could enable their identification as significant

Underpowered

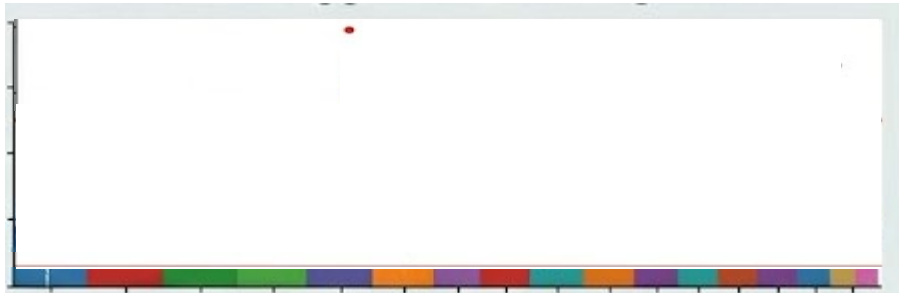


It can result from population stratification, cryptic relatedness, or technical artifacts; therefore, it is essential to review QC measures and confirm the association model.

Genomic inflation

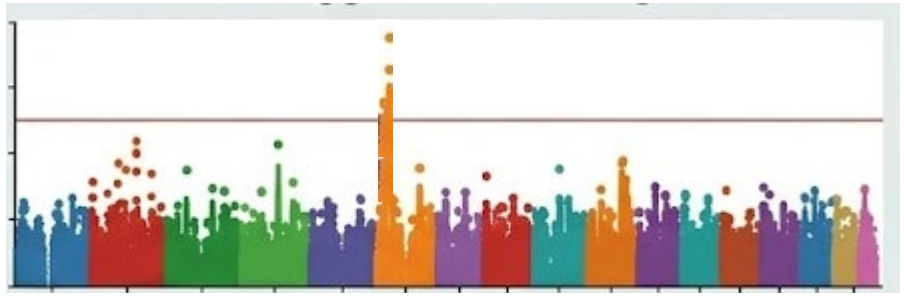


Winner's Curse



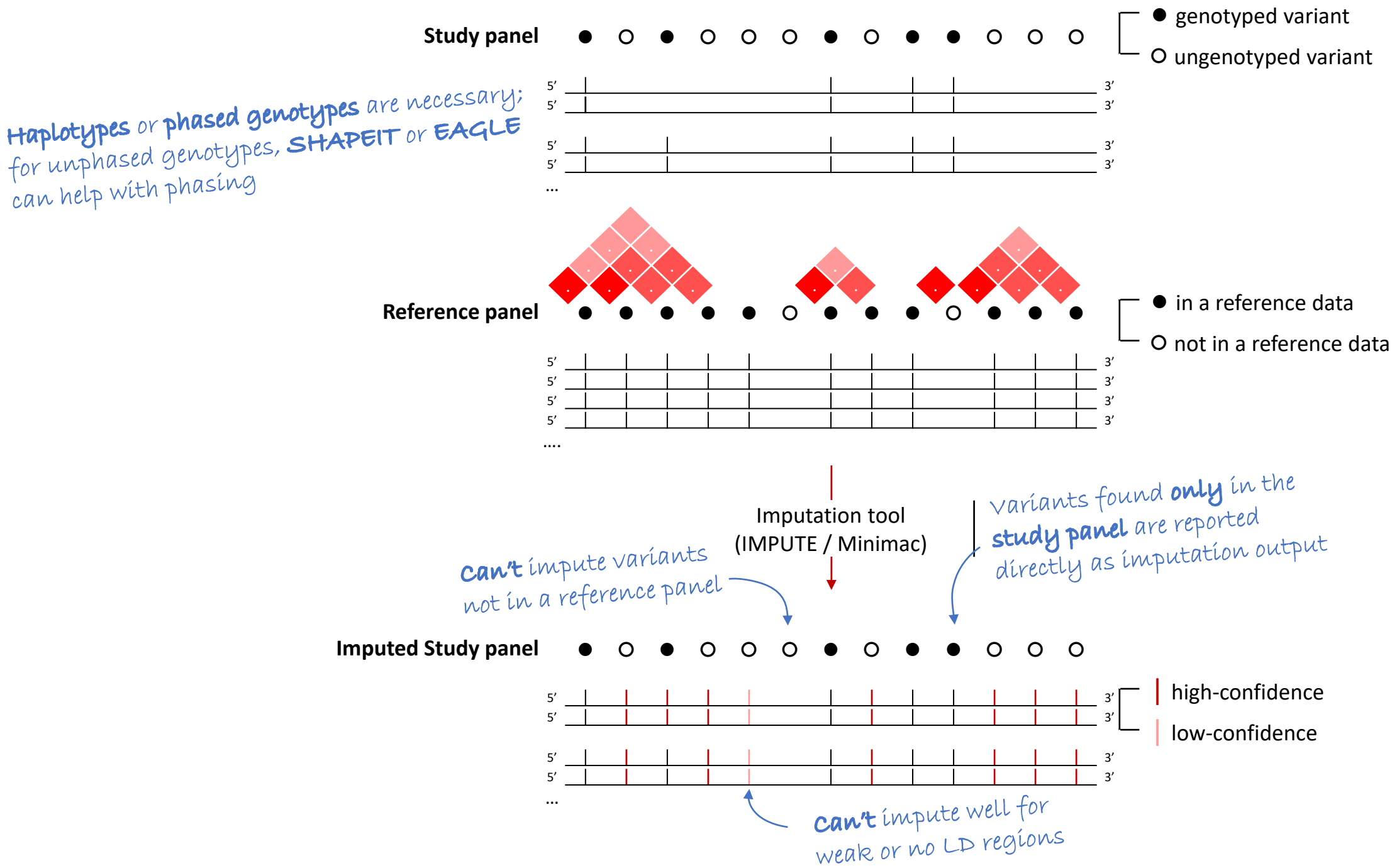
A situation of effect size is overestimated, check for sample QC (batch) and variant QC (CR and HWE) or find an independent dataset for validation; if it's based on array data, may try imputation data

HLA region



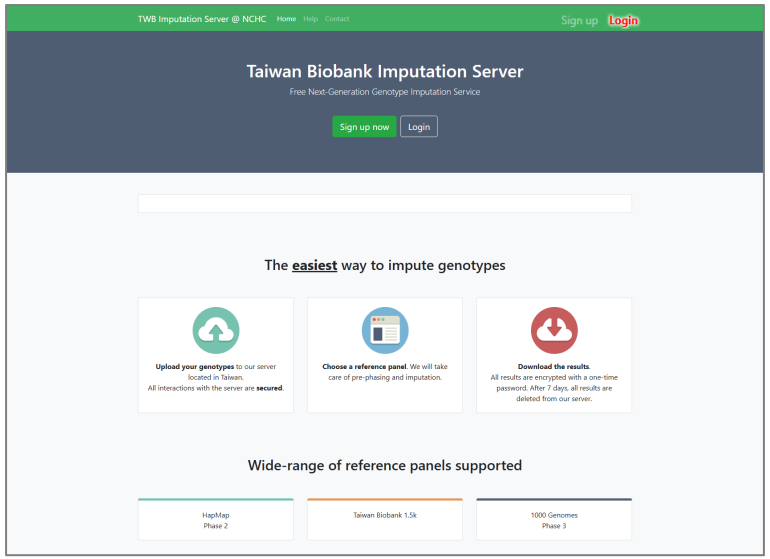
It looks fine but make sure your disease is a kind of ones (e.g., autoimmune) that related to MHC region

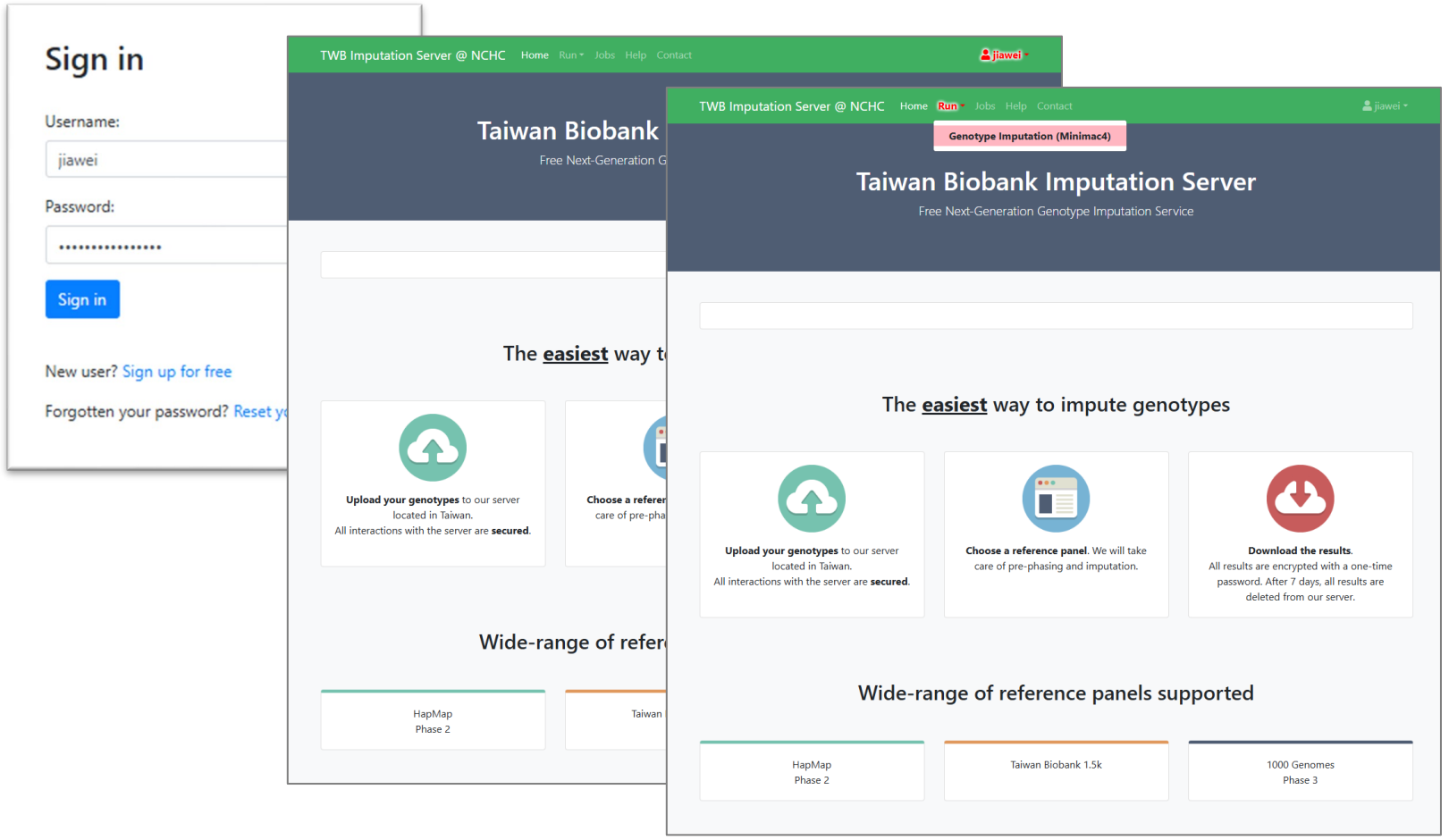
Imputation


Haplotypes or phased genotypes are necessary; for unphased genotypes, SHAPEIT or EAGLE can help with phasing
Study panel
genotyped variant
ungenotyped variant
5′
3′
Reference panel
in a reference data
not in a reference data
Imputation tool
(IMPUTE / Minimac)
Can't impute variants not in a reference panel
Variants found only in the study panel are reported directly as imputation output
Imputed Study panel
high-confidence
low-confidence
Can't impute well for weak or no LD regions

Imputation server

Taiwan Biobank Imputation Server





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 --
- -- 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

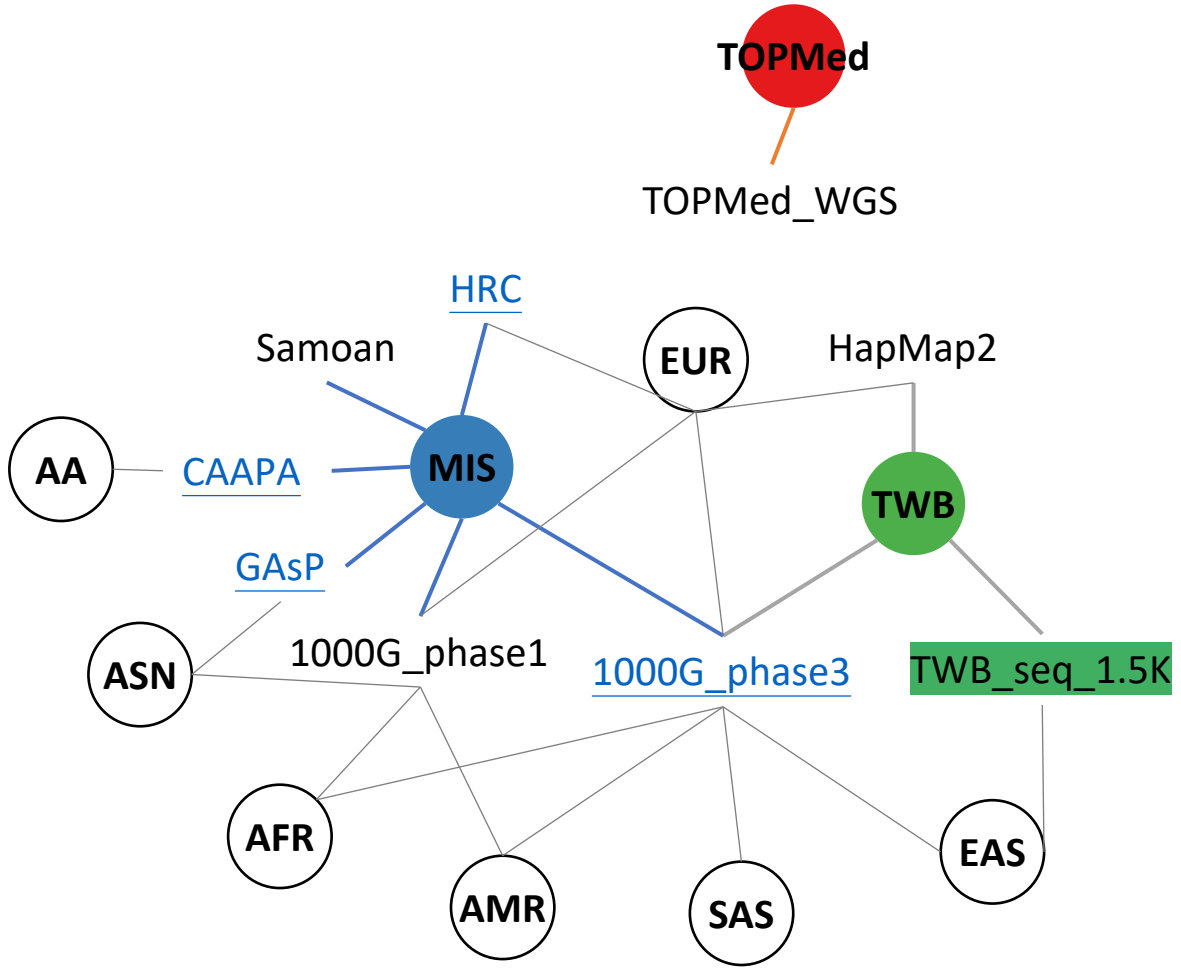
Imputation Server encrypts all zip files by default. Please note that AES encryption does not work with standard unzip programs. Use 7z instead.

Generate Meta-imputation file

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

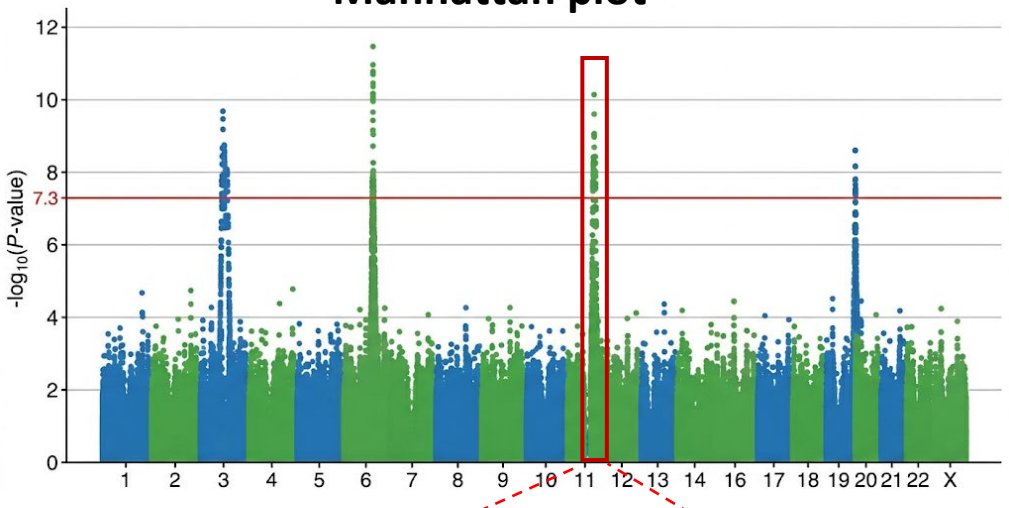
I will report any inadvertent data release, security breach or other data management incident of which I become aware.

Submit Job

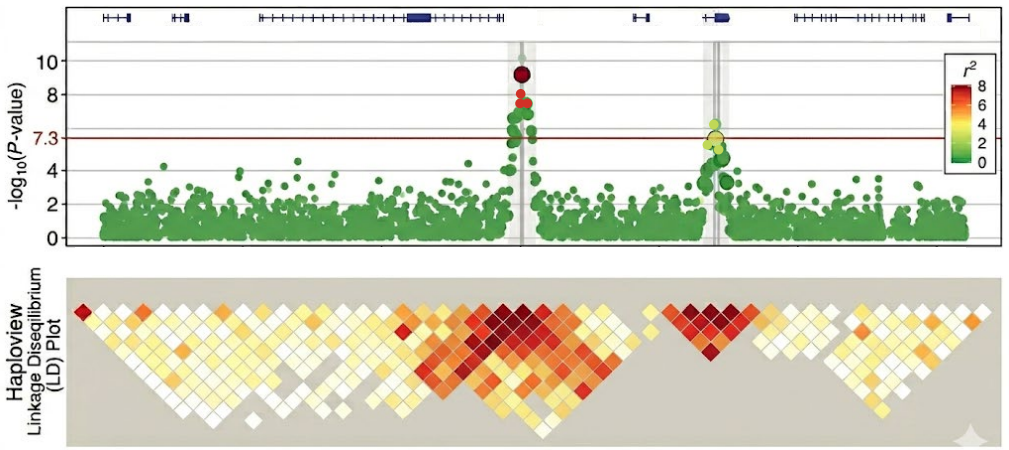


Fine-mapping

Manhattan plot



Regional plot



Single causal variant assumption

causal ●

Stepwise / COJO

causal ● & ●

Bayesian

credible set ●●●●●●●●

SuSiE (Sum of Single Effects Linear Regression)

credible sets ●●●● & ●●●●

$\boldsymbol{z} = (z_1, z_2, \dots, z_p)$: summary statistics
$\boldsymbol{R_{p\times p}}$: LD matrix

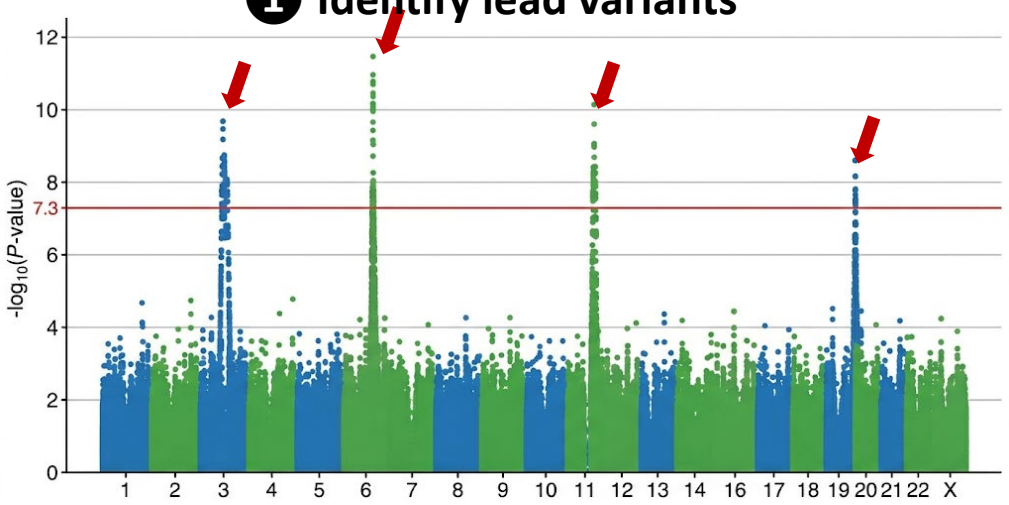
$z \approx \sum_l^L R\,\alpha^{(l)} + \epsilon$
observed signal explained by L latent single-effect signals

$\alpha^{(l)} = \left(\alpha_1^{(l)}, \dots, \alpha_p^{(l)}\right)$: probability vector for signal l over all SNPs
$\sum_l^L \alpha^{(l)} = 1$: each signal corresponds to one causal SNP

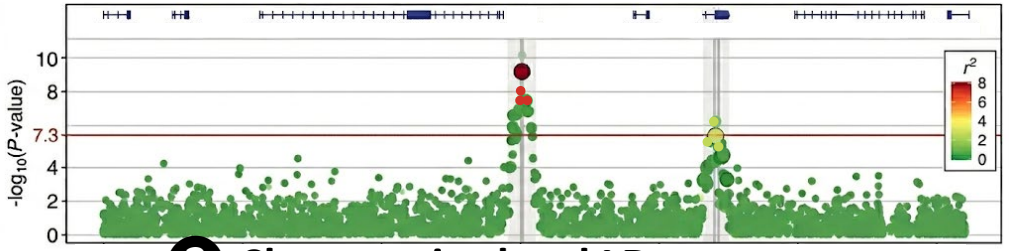
$r^{(l)} = \boldsymbol{z} - \sum_l^L \alpha^{(l)}$
residual: isolate signal l by removing all other signals

$\alpha_j^{(l)} \propto similarity\left(r^{(l)}, R_{.j}\right)$
$PIP_j = 1 - \prod_l^L \left(1 - \alpha_j^{(l)}\right)$ probability variant j is causal in at least one signal

❶ Identify lead variants



❷ Define genomic windows



❸ Characterize local LD structure

❹ Perform multi-signal fine-mapping

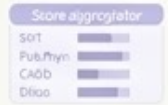
R vignette

```
> # install.packages(c("susieR", "data.table"))
> library(susieR)
> library(data.table)

> # Assuming a file with columns: SNP, BETA, SE (or Z)
> sumstats <- fread("path/to/sumstats.txt")

> # Should be a square matrix (p x p) matching the SNPs in sumstats
> R_matrix <- as.matrix(fread("path/to/ld_matrix.ld"))

> fit_rss <- susie_rss( z = sumstats, R = R_matrix,
>                       n = n, # sample size
>                       estimate_residual_variance = TRUE)

> summary(fitted_rss1)$cs
> #  cs | cs_log10bf | cs_avg_r2 | cs_min_r2 | variable
```

Functional annotation

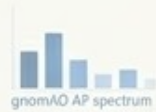
1000 Genomes Project

Population freq.



dbNSFP

Pathogenicity score



gnomAD

Population freq.

MutationTaster

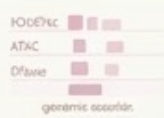
Pathogenicity score

Reactome

Pathway / function

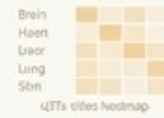
CADD

Pathogenicity score



ENCODE

Regulatory



GTEx

Pathway / function

OMIM

Clinical database

RegulomeDB

Regulatory

ClinPred

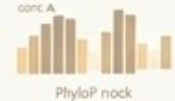
Pathogenicity score

ExAC

Population freq.

GWAS Catalog

Clinical database



PhyloP

Conservation

REVEL

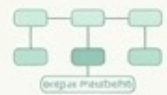
Pathogenicity score

ClinVar

Clinical database

FATHMM

Pathogenicity score

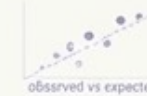


KEGG

Pathway / function

pLI

Pathogenicity score

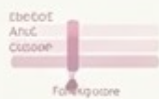


RVIS

Pathogenicity score

COSMIC

Clinical database



FunSeq2

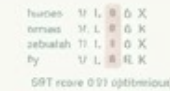
Pathogenicity score

LRT

Pathogenicity score

PolyPhen-2

Pathogenicity score



SIFT

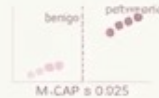
Pathogenicity score

DANN

Pathogenicity score

GDI

Pathogenicity score



M-CAP

Pathogenicity score

PROVEAN

Pathogenicity score

SpliceAI

Regulatory

Population Frequency (AFs)

1000 Genomes Project reference population allele frequencies from diverse global populations

gnomAD large population AF resource; very low AF supports rarity

ExAC exome-based population AF database

Pathway

KEGG pathway-level database of many species, often used for functional interpretation and enrichment

Reactome reaction-level (detailed molecular reaction steps) database mainly of Human

Conservation

PhyloP base-level conservation score:
> 0 (conserved); < 0 (accelerated); ≈ 0 (neutral)

pLI probability of Loss-of-function intolerance:
0 (LoF-tolerant) → 1 (highly LoF-intolerant)

RVIS genes with more extreme values are less tolerant to functional variation:
> 0 (LoF-tolerant) → < 0 (LoF-intolerant)

GDI gene-level disease burden index:
100 (LoF-tolerant) → 0 (highly LoF-intolerant)

Missense-specific predictors

SIFT missense tolerance score:
0-0.05 Damaging (D); > 0.05 Tolerated (T)

PolyPhen missense impact predictor:
0 Benign (D) → 1 Probably Damaging (D)

Clinical database

ClinVar clinical variant interpretations; pathogenic or likely pathogenic submissions are direct evidence for clinical relevance

OMIM gene-disease catalog for Mendelian disorders; strong support for known causal genes

Ensemble Scores

CADD integrative deleteriousness score:
PHRED-scaled 1 → 99 Deleterious

DANN deep-learning pathogenicity predictor :
0 → 1 Likely Deleterious

Supporting evidence databases

COSMIC somatic mutation cancer database; supports oncogenic relevance but not definitive

GWAS-Catalog research-only evidence of trait-associated loci from association studies; useful for linking variants/genes to common traits

Regulation

GTEx tissue-specific expression and eQTL resource; helps link variants to tissue-relevant function

ENCODE Experimental regulatory annotation resource for chromatin, TF binding, and accessible regions

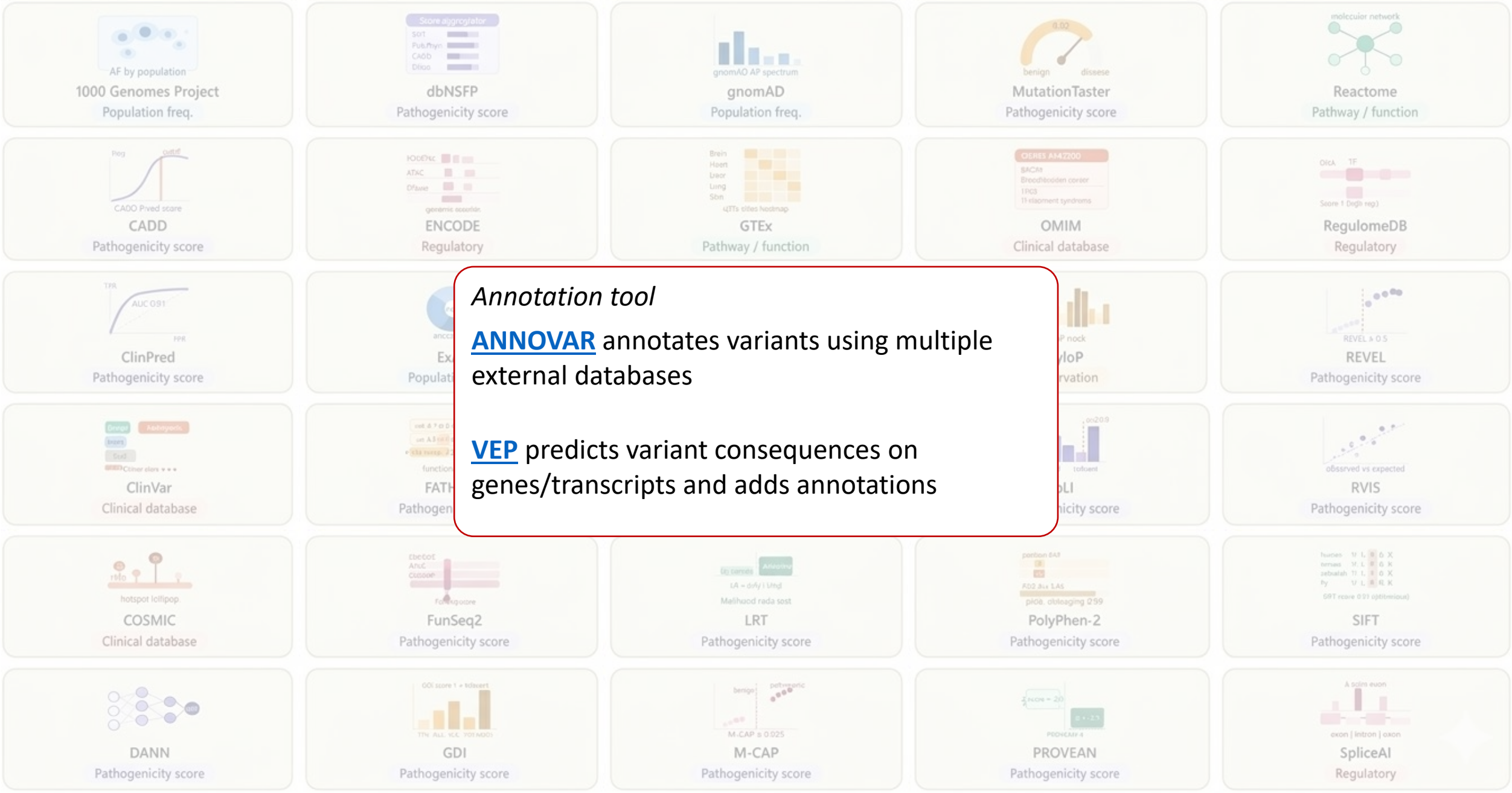
RegulomeDB: Integrative database scoring noncoding regulatory evidence; lower-numbered/stronger categories indicate more support for regulatory function

SpliceAI: splicing impact predictor; higher scores indicate stronger predicted splice disruption

Annotation tool

ANNOVAR annotates variants using multiple external databases

VEP predicts variant consequences on genes/transcripts and adds annotations



Polygenic risk score

	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

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

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

Base Data
GWAS summary statistic
(OR, beta, p-value)

Independent!

Target Data
Individual genotype data
(array, imputation)

SNP weights (PGS)

PGS Catalog https://www.pgscatalog.org/

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

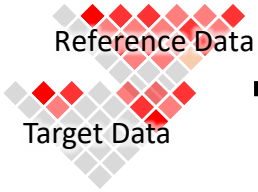
LDpred2
PLINK
PRSice
PRS-CSx
lassosum
PRS-CS

Set of SNPs

clumping / pruning

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

Genome-wide SNPs



beta shrinkage

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

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.

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

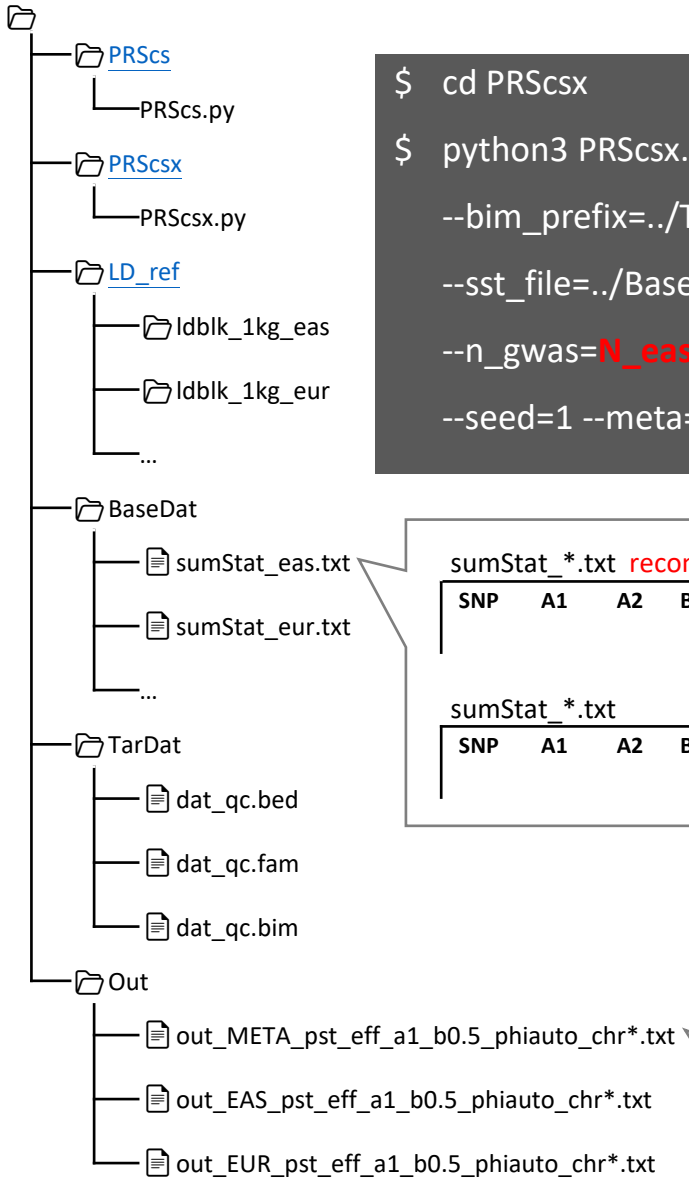
PLINK / PLINK2	depend on target data
PRSice2	depend on target data
LDpred2	**GRCh37 & 38** HapMap3 LD blocks and LD matrix
Lassosum	**GRCh37 & 38** 1000 genomes project Phase I LD blocks
PRS-CS / PRS-CSx	**GRCh37** 1000 genomes project LD UK Biobank LD

Target data (individual-level data)

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

PRS-CSx example

Manually calculate SNP weights by published tools




```
$ cd PRScsx
$ python3 PRScsx.py --ref_dir=../LD_ref \
    --bim_prefix=../TarDat/dat_qc \
    --sst_file=../BaseDat/sumStat_eas.txt,../BaseDat/sumStat_eur.txt \
    --n_gwas=N_eas,N_eur --pop=EAS,EUR \
    --seed=1 --meta=TRUE --out_dir=../Out/out_prscsx --out_name=out
```

sumStat_*.txt recommended

SNP	A1	A2	BETA/OR	SE

sumStat_*.txt

SNP	A1	A2	BETA/OR	P

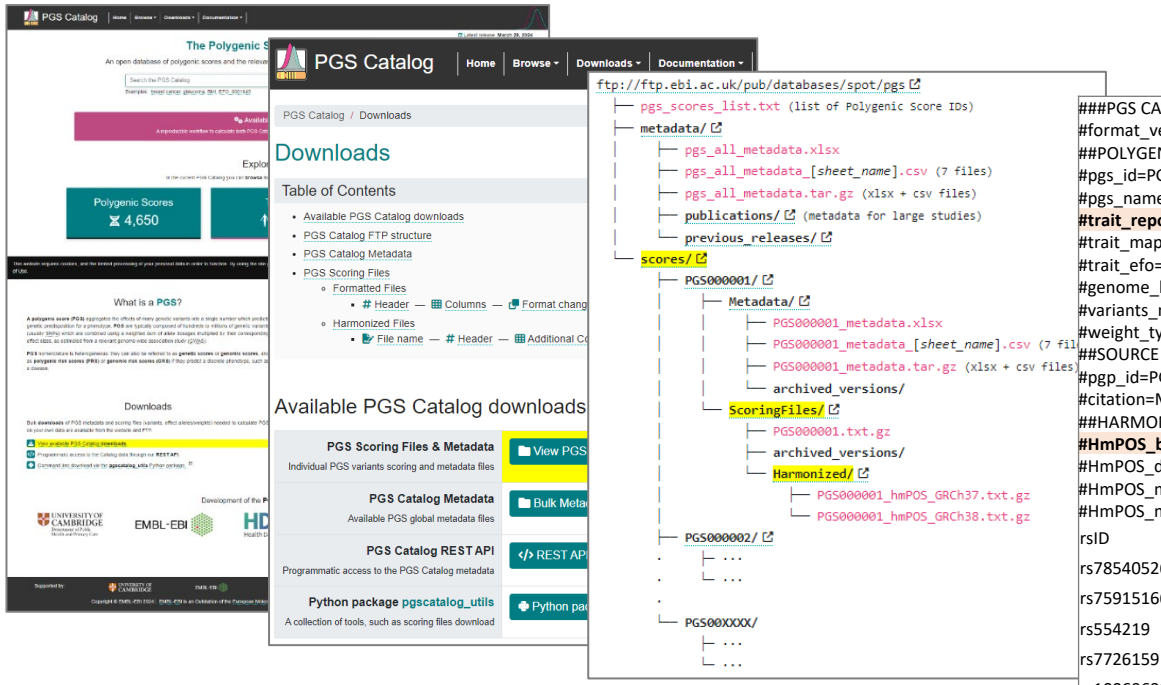
out_*_pst_eff_a1_b0.5_phiauto_chr*.txt

CHR	RSID	BP	A1	A2	Weight

External PRS example

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

PGS Catalog




```
###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	

3 5 9

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

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

dat_qc.**profile**

FID	IID	PHENO	CNT	CNT2	**SCORE**

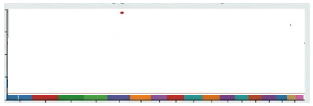
FID	IID	PHENO	CNT	CNT2	**SCORSUM**

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

N_i = non-missing SNPs in sample i

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

Q & A

Q:  **The red dot is like an outlier (in statistic meaning), can I remove it directly?**

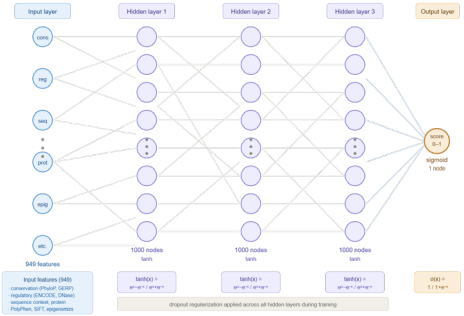
A: 睜一隻閉一隻眼說是可以，但任何分析在刪除資料點時都還是希望能有明確的理由去移除掉，在 GWAS 中移除位點方式主要在 QC 時檢查 CR、MAF、HWE 或其它有關品質的額外訊息，

Q: Is imputation necessary?

A: Imputation 最直觀的好處是能提升解析度補齊未在 array 上的位點 (可能找到真正的 causal variants)，現在多數 GWAS papers 可能會要求 imputation，沒被要求到而不想做我覺得沒關係；但若有考慮 post-GWAS 如 fine-mapping、beta-shrinkage PRS 等那就一定要做

Q: How does DANN work? How is deep learning applied on it?

A: 請 Claude 依據 (Quan et al. 2015) 的 Method 建構模型架構如右圖，其實沒有很「深」，使用與 CADD 相同的資料，但將 CADD 整合 score 的方法 (SVM) 改成 neural network，將 949 種 scores 作為 input 得到 output 為一個整合的 score



Q: With different base data, we may have different weights for PRS calculation, how to choose?

A: PRS 是 prediction 工具，因此以簡單的結果論方式，看哪種 weights 得到的個人 score 在預測疾病的表現比較好 (e.g., logistic regression 看 AUC 的表現)，另外，在使用 PRS 預測疾病情況時，常常會將 covariates (age、gender、PCs 等)考慮進去

Q: If only one dataset is available, can it be split into training, validation, and testing sets for PRS?

A: 可以，但可能不是 PRS 分析會先考量的方式 (明顯的失去很多 samples 來評估)；如果能夠從外部找到所研究疾病的 GWAS summary statistics，那只需將資料切成 training 和 testing 兩部分，training set 求 weights，testing set 評估結果；當然，最理想情況還是希望有個「真正」的 independent dataset 來做評估

Thanks for your attention!!
<(_ _)>