THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Introduction to GWAS summary statistics & application in polygenic prediction

Jian Zeng

j.zeng@uq.edu.au

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | Institute for Molecular Bioscience

Program in Complex Trait Genomics

Outline

What are GWAS summary statistics (SumStats)?

Where to download?

What should we check?

How to use them?

- Example: polygenic score (PGS) prediction using SBayesR

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

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

nature genetics

Asking for more

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

—Nat Genet editorial, July 2012

Cell Genomics

CellPress
OPEN ACCESS

2021

Perspective

Workshop proceedings: GWAS summary statistics standards and sharing

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

Table 1. Recommended standard reporting elements for GWAS SumStats

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

PRIMER

Check for updates

Genome-wide association studies

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

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

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

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

Critical information from GWAS SumStats

- SNP name/position
- Effect allele and alternate allele (A1 and A2)
- Effect allele frequency
- Marginal SNP effect
- Standard error
- P-value
- (Per-SNP) GWAS sample size

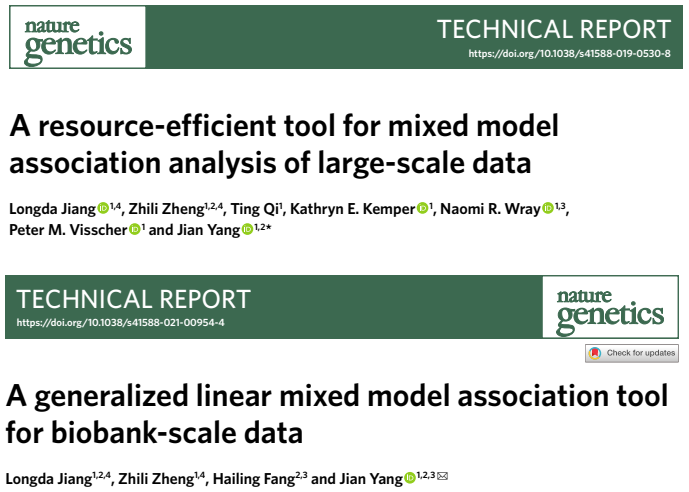
```
SNP A1 A2 freq b se p N
rs1001 A G 0.8493 0.0024 0.0055 0.6653 129850
rs1002 C G 0.0306 0.0034 0.0115 0.7659 129799
rs1003 A C 0.5128 0.0045 0.0038 0.2319 129830
```

.ma file

Where to download the UK Biobank SumStats?

fastGWA

https://yanglab.westlake.edu.cn/software/gcta/#DataResource


nature genetics

TECHNICAL REPORT

https://doi.org/10.1038/s41588-019-0530-8

A resource-efficient tool for mixed model association analysis of large-scale data

Longda Jiang[1,4], Zhili Zheng[1,2,4], Ting Qi[1], Kathryn E. Kemper[1], Naomi R. Wray[1,3], Peter M. Visscher[1] and Jian Yang[1,2*]

TECHNICAL REPORT

https://doi.org/10.1038/s41588-021-00954-4

nature genetics

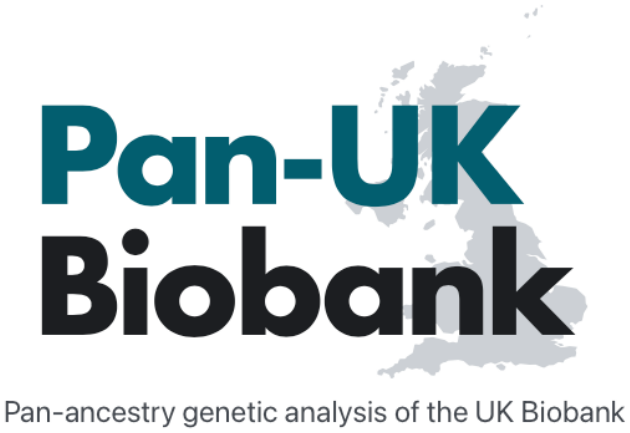
Check for updates

A generalized linear mixed model association tool for biobank-scale data

Longda Jiang[1,2,4], Zhili Zheng[1,4], Hailing Fang[2,3] and Jian Yang[1,2,3]✉

Pan-UKB

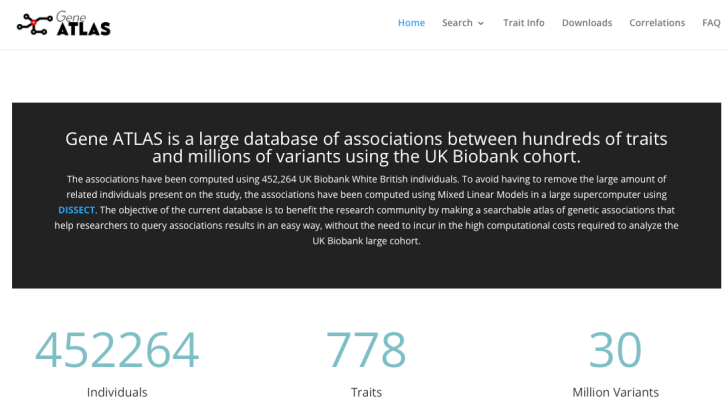
https://pan.ukbb.broadinstitute.org



Pan-ancestry genetic analysis of the UK Biobank

GeneAtlas

http://geneatlas.roslin.ed.ac.uk



GCTA
a tool for Genome-wide Complex Trait Analysis

Data Resource

UK Biobank GWAS results

We developed two resource-efficient tools, fastGWA and fastGWA-GLMM, for mixed model-based association analysis in large-scale data. We first applied fastGWA to 2,173 traits on 456,422 array-genotyped as well as 49,960 whole-exome-sequenced individuals of European ancestry in the UK Biobank (UKB) (Jiang et al. 2019 Nat Genet). We then applied fastGWA-GLMM to 2,989 binary traits on 456,348 array-genotyped individuals of European ancestry in the UKB (Jiang et al. 2021 Nat Genet). See below for detailed instructions to query and download the data. You can also query or visualize the GWAS summary data using the fastGWA data portal.

- fastGWA summary statistics using the imputed data (Jiang et al. 2019 Nat Genet): 456,422 individuals of European ancestry; 8,531,416 variants (MAF > 0.01 and missingness rate < 0.1); 2,173 traits.
  - Summary table: UKB_impute_v1.1.csv
  - Online tool: https://yanglab.westlake.edu.cn/data/ukb_fastgwa/imp/
  - Linux command to download all the summary statistics (2,173 files; 454 GB in total):

Imputed data

```
mkdir ukb && cd ukb && wget https://yanglab.westlake.edu.cn/software/gcta/res/UKB_impute_v1.1.list && wget -i UKB_impute_v1.1.list
```

- fastGWA summary statistics using the whole-exome sequence (WES) data (Jiang et al. 2019 Nat Genet): 46,191 individuals of European ancestry; 152,327 variants (MAF > 0.01 and missingness rate < 0.1); 2,048 valid traits.
  - Summary table: UKB_WES_v1.1.csv
  - Online tool: https://yanglab.westlake.edu.cn/data/ukb_fastgwa/wes/
  - Linux command to download all the summary statistics (2,048 files; 8 GB in total):

Whole-exome sequence data

```
mkdir wes && cd wes && wget https://yanglab.westlake.edu.cn/software/gcta/res/UKB_WES_v1.1.list && wget -i UKB_WES_v1.1.list
```

- fastGWA-GLMM summary statistics using the imputed data (Jiang et al. 2021 Nat Genet): 456,422 individuals of European ancestry; 11,842,647 variants (MAF > 0.0001 and missingness rate < 0.1); 2,989 binary traits.
  - Summary table: UKB_binary_v1.11.csv
  - Online tool: https://yanglab.westlake.edu.cn/data/ukb_fastgwa/imp_binary/
  - Linux command to download all the summary statistics (2,989 files; 1.2 TB in total):

Imputed data for binary traits

```
mkdir ukb_binary && cd ukb_binary && wget https://yanglab.westlake.edu.cn/software/gcta/res/UKB_binary_v1.11.list  && wget -i UKB_bi
nary_v1.11.list
```

Data format from fastGWA

Summary table (UKB_impute_v1.1.csv):

ID	Description	Data_type	Method	N	Ncase	Gender_specific	URL
100001	Food weight	Continuous	MLM	64029	NA		https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKB/100001.v1.1.fastGWA.gz
100002	Energy	Continuous	MLM	64029	NA		https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKB/100002.v1.1.fastGWA.gz
100003	Protein	Continuous	LR	64029	NA		https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKB/100003.v1.1.fastGWA.gz
100004	Fat	Continuous	MLM	64029	NA		https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKB/100004.v1.1.fastGWA.gz
100005	Carbohydrate	Continuous	MLM	64029	NA		https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKB/100005.v1.1.fastGWA.gz
100006	Saturated fat	Continuous	MLM	64029	NA		https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKB/100006.v1.1.fastGWA.gz
100007	Polyunsaturated fat	Continuous	MLM	64029	NA		https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKB/100007.v1.1.fastGWA.gz
100008	Total sugars	Continuous	MLM	64029	NA		https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKB/100008.v1.1.fastGWA.gz
100009	Englyst dietary fibre	Continuous	MLM	64029	NA		https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKB/100009.v1.1.fastGWA.gz
100010	Portion size	Ordered_Categorical	MLM	64001	NA		https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKB/100010.v1.1.fastGWA.gz
100011	Iron	Continuous	MLM	64029	NA		https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKB/100011.v1.1.fastGWA.gz

Summary statistics (100001.v1.1.fastGWA.gz):

CHR	SNP	POS	A1	A2	N	AF1	BETA	SE	P
1	rs144155419	717587	G	A	62687	0.989495	-0.00553627	0.0270775	0.837994
1	rs58276399	731718	T	C	61089	0.888139	-0.00324112	0.00889488	0.715574
1	rs141242758	734349	T	C	61284	0.888185	-0.00148931	0.00888356	0.86686
1	rs28544273	751343	T	A	62869	0.877778	-0.00614735	0.00842396	0.465546

Format of the summary statistics from fastGWA:

```
CHR: chromosome
SNP: SNP ID
POS: SNP position
A1: the effect allele
A2: the other allele
N: per allele sample size
AF1: frequency of A1
BETA: SNP effect
SE: standard error
P: p value
```

Data format from fastGWA-GLMM

Summary table (UKB_binary_v1.1.csv):

ID	Description	N	N_case	N_control	ratio	URL
785_PheCode	Abdominal pain	456348	35233	421115	0.083665982	https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKBbin/785_PheCode.v1.0.fastGWA.gz
418_PheCode	Nonspecific chest pain	456348	29202	427146	0.068365383	https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKBbin/418_PheCode.v1.0.fastGWA.gz
208_PheCode	Benign neoplasm of colon	456348	19026	437322	0.043505701	https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKBbin/208_PheCode.v1.0.fastGWA.gz
535_PheCode	Gastritis and duodenitis	456348	17420	438928	0.039687603	https://yanglab.westlake.edu.cn/resources/fastgwa_data/UKBbin/535_PheCode.v1.0.fastGWA.gz

Summary statistics (785_PheCode.v1.0.fastGWA.gz):

CHR	SNP	POS	A1	A2	N	AF1	T	SE_T	P_noSPA	BETA	SE	P	CONVERGE			
1	rs2531267	69569	C	T	456285	0.000151221	0.44236	3.10901	0.886856		0.0457649		0.321646		0.886856	1
1	rs12238997	693731	G	A	418020	0.0936486	-9.26915		70.4227	0.895284		-0.00186902		0.0142	0.895284	1
1	rs144155419	717587	A	G	446865	0.0105065	5.90265	25.6125	0.817734		0.00899798		0.0390435		0.817734	1
1	rs189787166	723329	T	A	455943	0.00132034	4.55647	9.18687	0.619911		0.0539875		0.108851		0.619911	1
1	rs148120343	730087	C	T	432816	0.0439633	3.02531	50.5154	0.952244		0.00118556		0.0197959		0.952244	1
1	rs58276399	731718	C	T	436309	0.111288	24.933	78.0154	0.749278		0.00409652		0.012818		0.749278	1
1	rs141242758	734349	C	T	437558	0.111188	19.3493	78.0909	0.804305		0.00317296		0.0128056		0.804305	1

Format of the summary statistics from fastGWA-GLMM:

```
CHR:  chromosome
SNP: SNP ID
POS: SNP position
A1: the effect allele
A2: the other allele
N: per allele sample size
AF1: the allele frequency of A1
T: GLMM score statistic
SE_T: standard error of the score statistic
P_noSPA: raw p-value
BETA: SNP effect or log(odds ratio)
SE: standard error for the estimated effect size after the SPA correction
P: p-value after the SPA correction
CONVERGE: to indicate whether the SPA correction is converged for the variant
```

Other useful data resource in the public domain

GWAS catalog (https://www.ebi.ac.uk/gwas/)

- 59,946 studies

FinGenn (https://www.finngen.fi/en)

- Freeze 10: >412,000 individuals, 2,408 disease endpoints

PGC (https://pgc.unc.edu)

- Multiple waves with increasing sample sizes for psychiatric disorders

Global Biobank Engine (https://biobankengine.stanford.edu)

- > 750,000 individuals across three population cohorts: UK Biobank, Million Veterans Program and Biobank Japan.

What can we do with them?

- **Meta-analysis**: METAL, MTAG
- **Finding independent association loci**: PLINK-clumping, GCTA-COJO
- **Fine-mapping causal variants**: FINEMAP, SuSiE
- **Exploring pleiotropic effects** (PheWAS)
- **Gene-based test**: MAGMA, fastBAT, mBAT-combo
- **Integrating with functional data**: coloc, SMR, TWAS, OPERA
- **Inferring trait-relevant tissues/cell types**: LDSE-SEG, MAGMA-gene-set, scDRS
- **Estimating SNP-based heritability**: LDSC, SBayesR
- **Estimating genetic correlation**: Popcorn, MiXeR
- **Predicting polygenic score (PGS/PRS)**: PRScie, LDpred2, PRScs, SBayesR
- **Inferring causal relationship between traits**: GSMR, LCV
- …

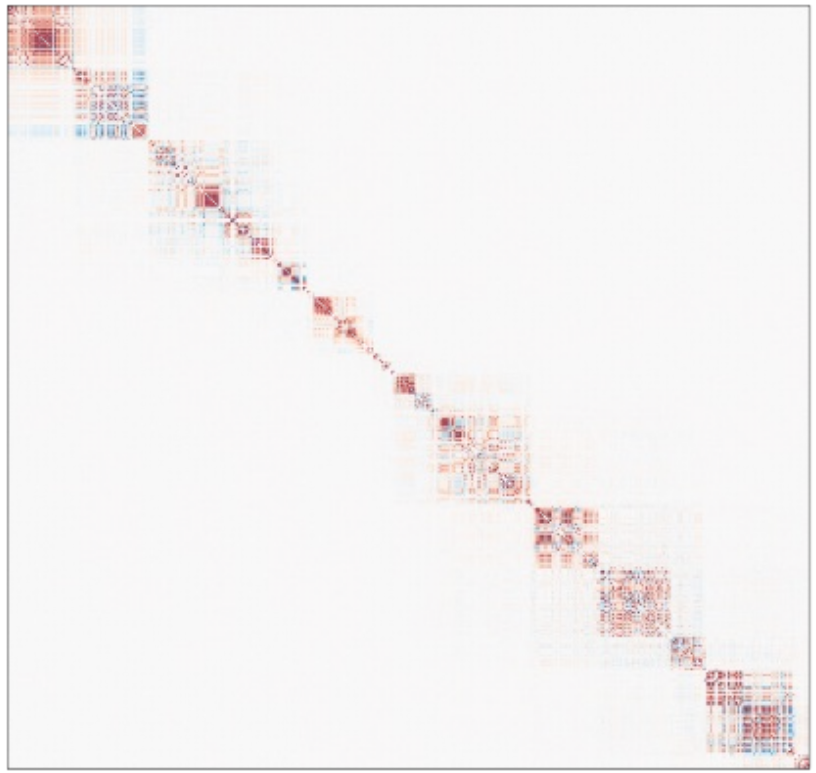
Linkage disequilibrium (LD) correlations

Usually obtained from a reference population

LD correlation matrix

$$\mathbf{R} = \frac{1}{n}\mathbf{X}'\mathbf{X}$$

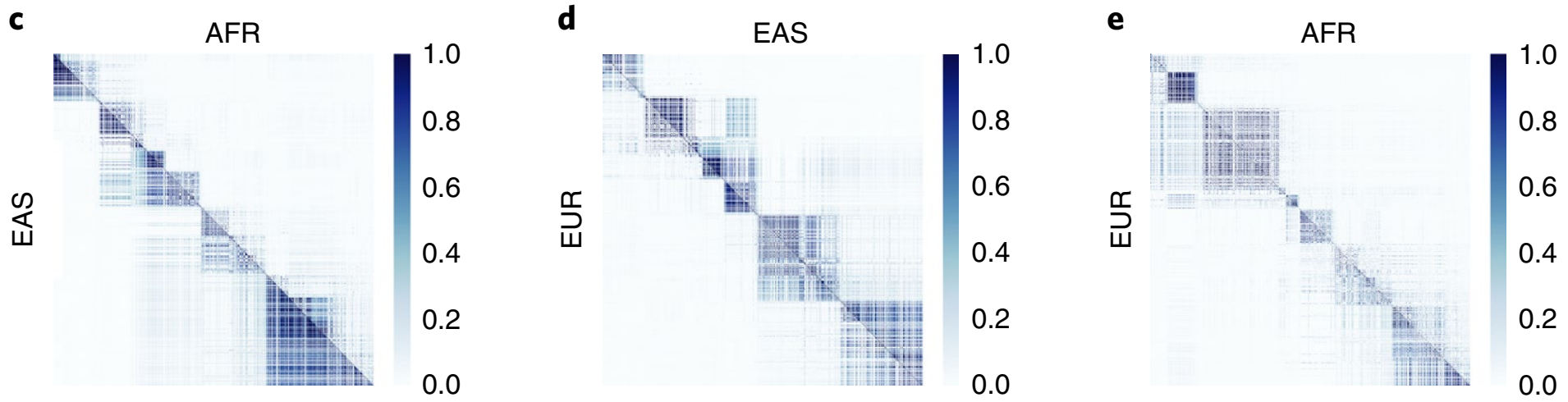
assuming $\mathbf{X}$ is standardised with mean zero and variance one



Match in ancestry

LD reference needs to match with GWAS sample in genetics

- No systematic differences in LD → same ancestry and population structure
- Minimum sampling variance in LD → LD ref sample size cannot be too small



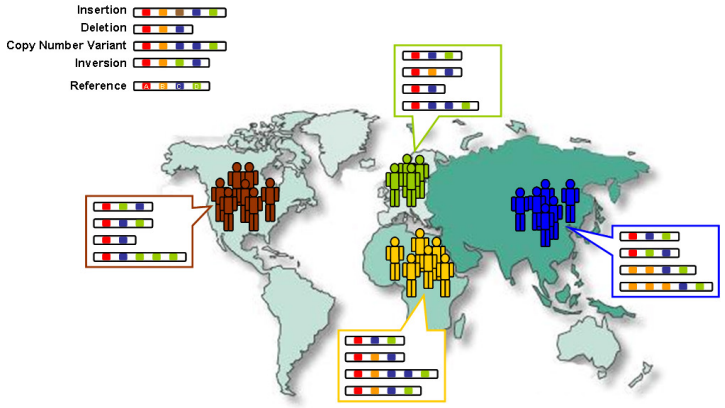
Martin et al 2019 Nature Genetics

Where to find LD reference data?

1000 Genomes Project (1KGP)

Individual sequence data

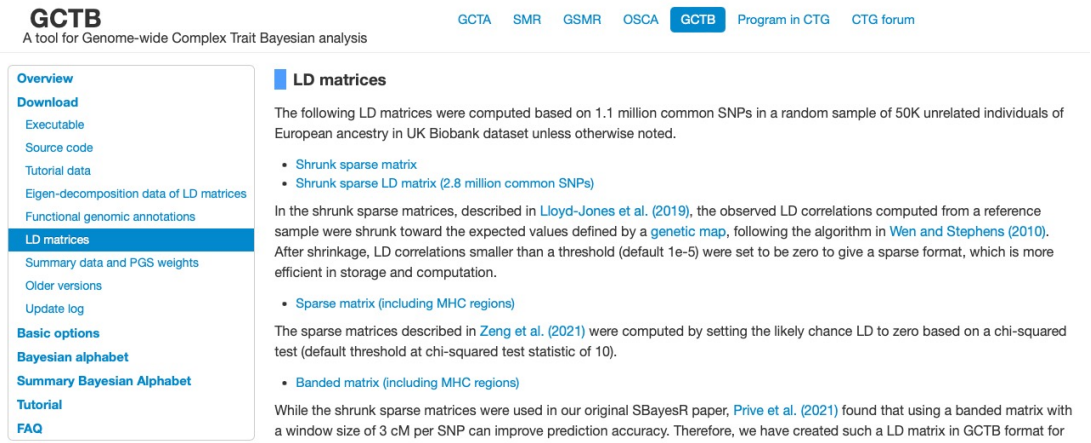
https://www.internationalgenome.org



UK Biobank (UKB)

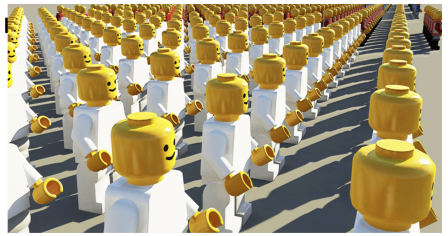
We provide LD matrices computed from a subset of UKB samples

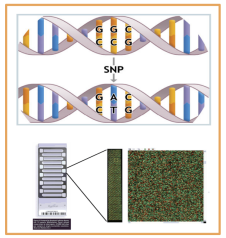
https://cnsgenomics.com/software/gctb/#LDmatrices



Individual-data model

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

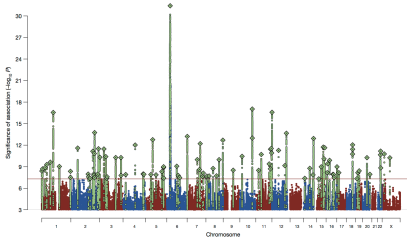


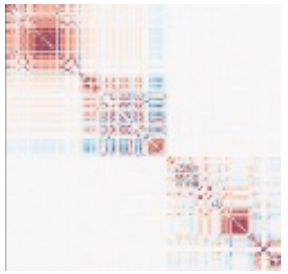


→

Summary-data model

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





The key is that $b_j = \frac{1}{n}\mathbf{X}_j'\mathbf{y}$ where $\mathbf{X}_j'\mathbf{y}$ is the sufficient statistic for many analyses.

What should we check prior to the analysis?

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Raw data file

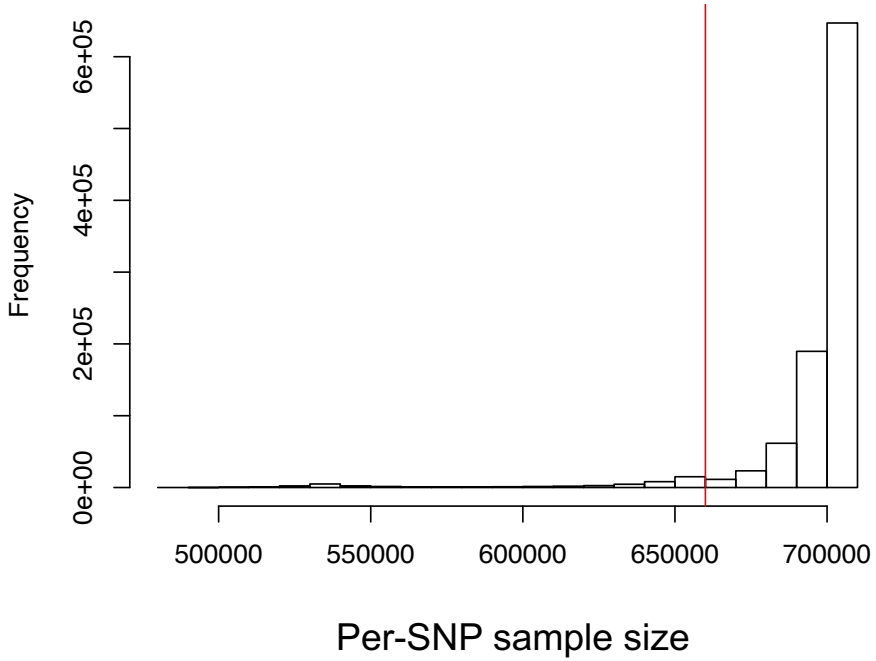
Item	What could be wrong?	How to fix?
Genome build	Inconsistent coordinates among GWAS summary data and LD reference.	Lift up to the same genome build using *liftover*
SNP ID	rsID not provided.	Use chromosome and position information to find their rsID (from LD reference file).
Alleles	Lower/upper case.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

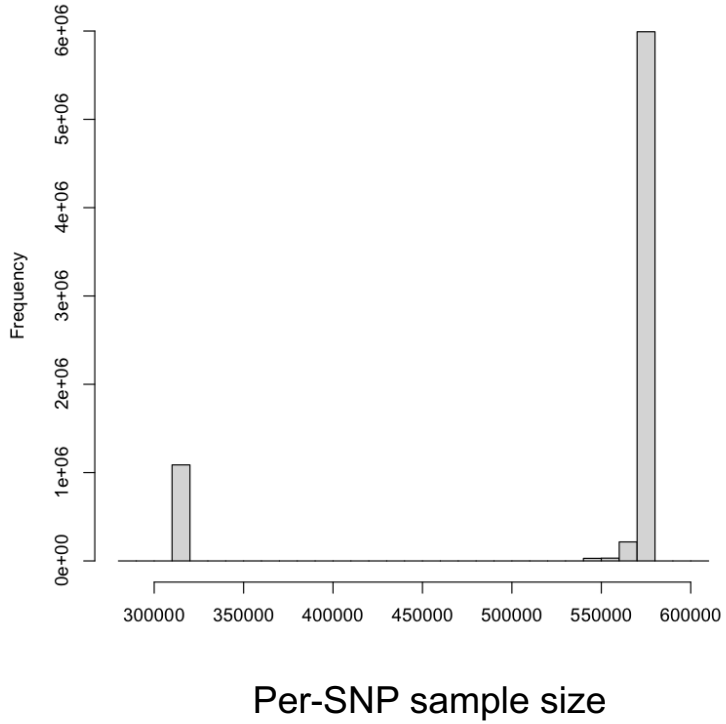
What should we check prior to the analysis? (cont')

Quality control (QC)

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

Heterogeneity in per-SNP sample size

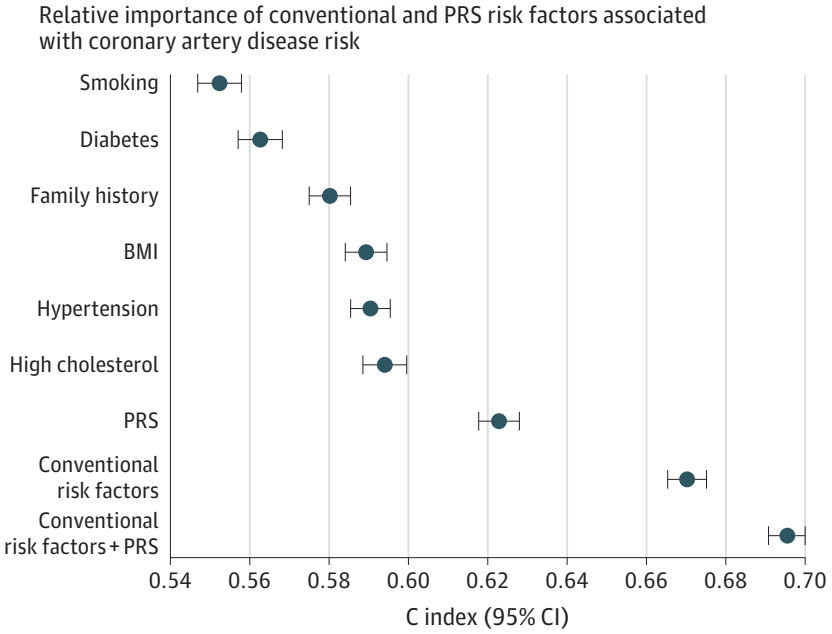


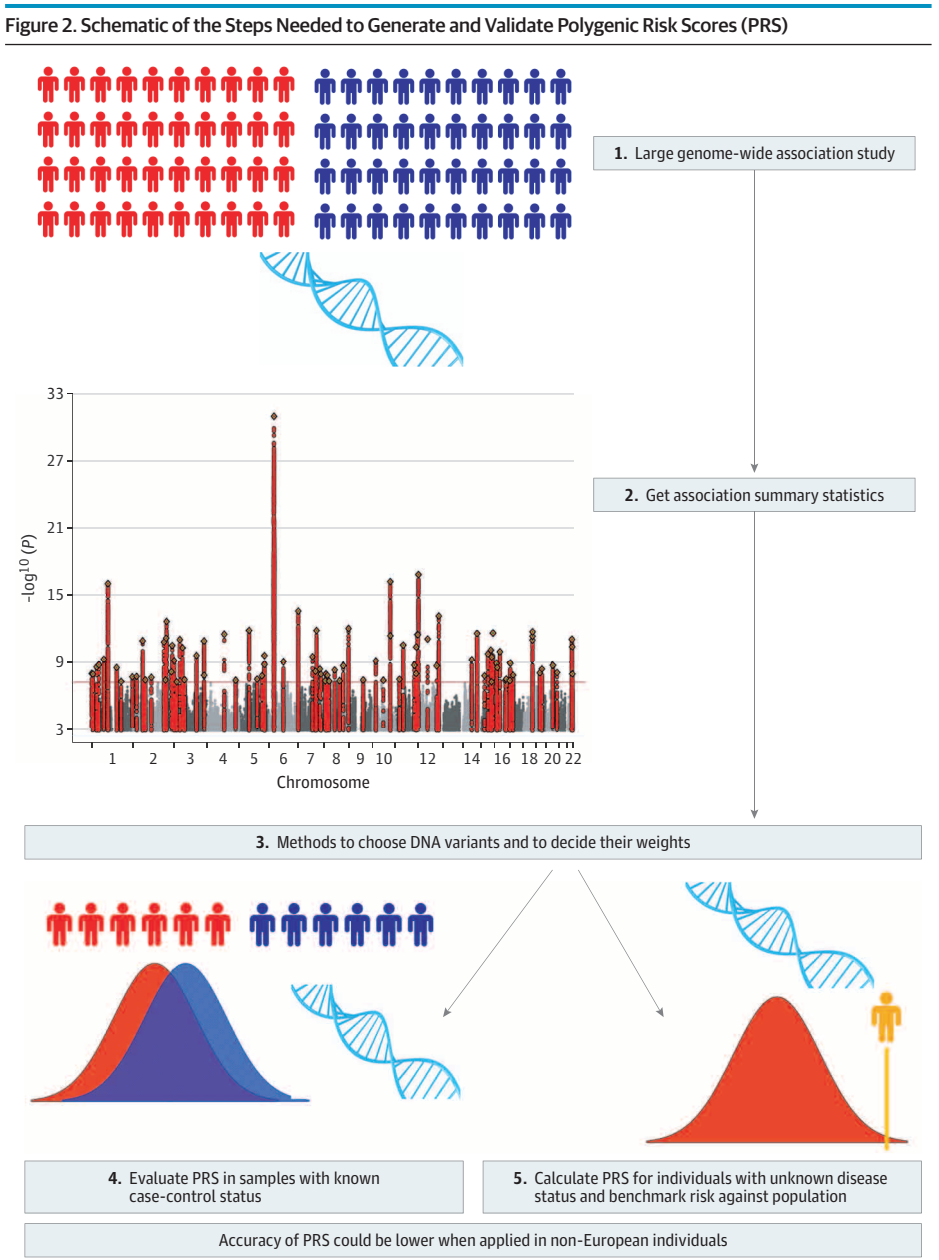


Example

Polygenic scores (PGS)

Polygenic risk score (PRS)





Wray et al, *JAMA Psychiatry* 2021; Inouye et al, *J Am Coll Cardiol.* 2018

Polygenic score (PGS)

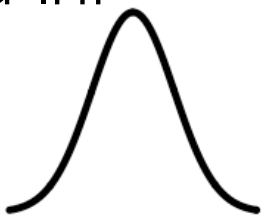
A weighted sum of the count of risk alleles

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

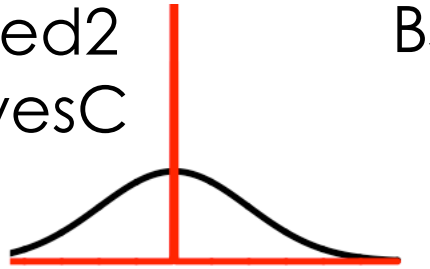
How many SNPs?
Which SNPs?
What weights?

New methods model genetic architecture:

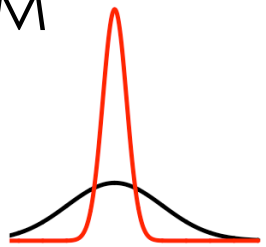
LDpred-Inf
SBLUP



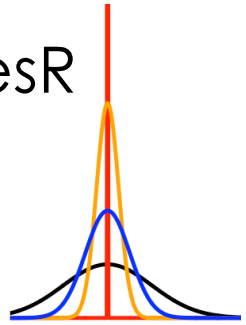
LDPred2
SBayesC



BSLMM



SBayesR



SBayesR (Lloyd-Jones and Zeng et al 2019)

The University Of Queensland AUSTRALIA

Model:

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

GWAS SNP marginal effects	LD correlation matrix	SNP joint effects	Residuals

nature COMMUNICATIONS

ARTICLE

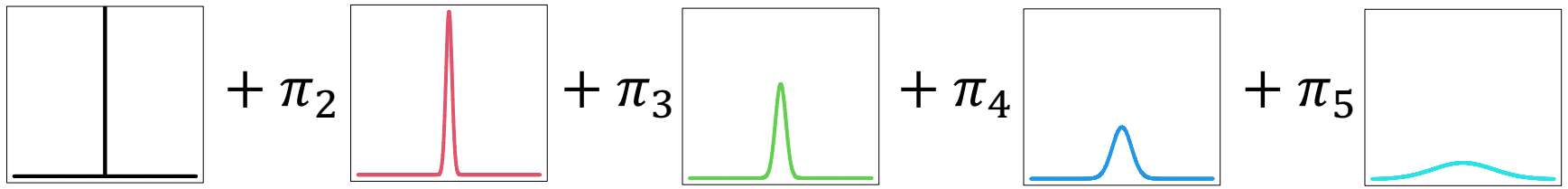
https://doi.org/10.1038/s41467-019-12653-0 OPEN

Improved polygenic prediction by Bayesian multiple regression on summary statistics

Luke R. Lloyd-Jones[1,9*], Jian Zeng[1,9*], Julia Sidorenko[1,2], Loïc Yengo[1], Gerhard Moser[3,4], Kathryn E. Kemper[1], Huanwei Wang[1], Zhili Zheng[1], Reedik Magi[2], Tõnu Esko[2], Andres Metspalu[2,5], Naomi R. Wray[1,6], Michael E. Goddard[7], Jian Yang[1,8*] & Peter M. Visscher[1*]

Each SNP effect has a mixture prior distribution:

$$\beta_j \sim \pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5$$

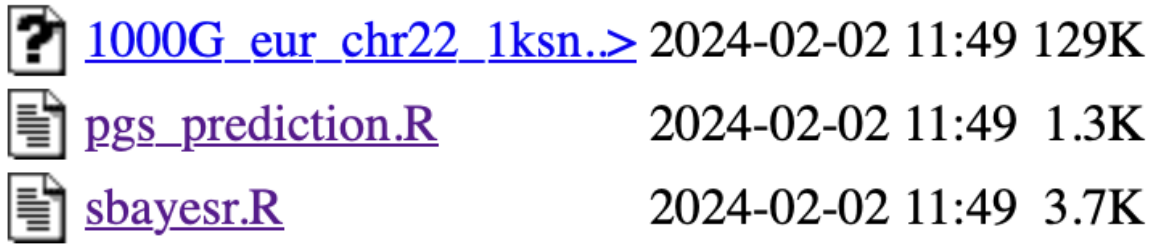


can accommodate various genetic architectures

Practical

Download the example data and R scripts at

https://cnsgenomics.com/data/teaching/AusUKB2024/session1/



For real trait analysis, we use GCTB to perform the analysis.

https://cnsgenomics.com/software/gctb/

Genetics & Genomics Winter School

June 24 - 28, 2024

Registration open soon!