THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

CREATE CHANGE

GWAS Summary Statistics

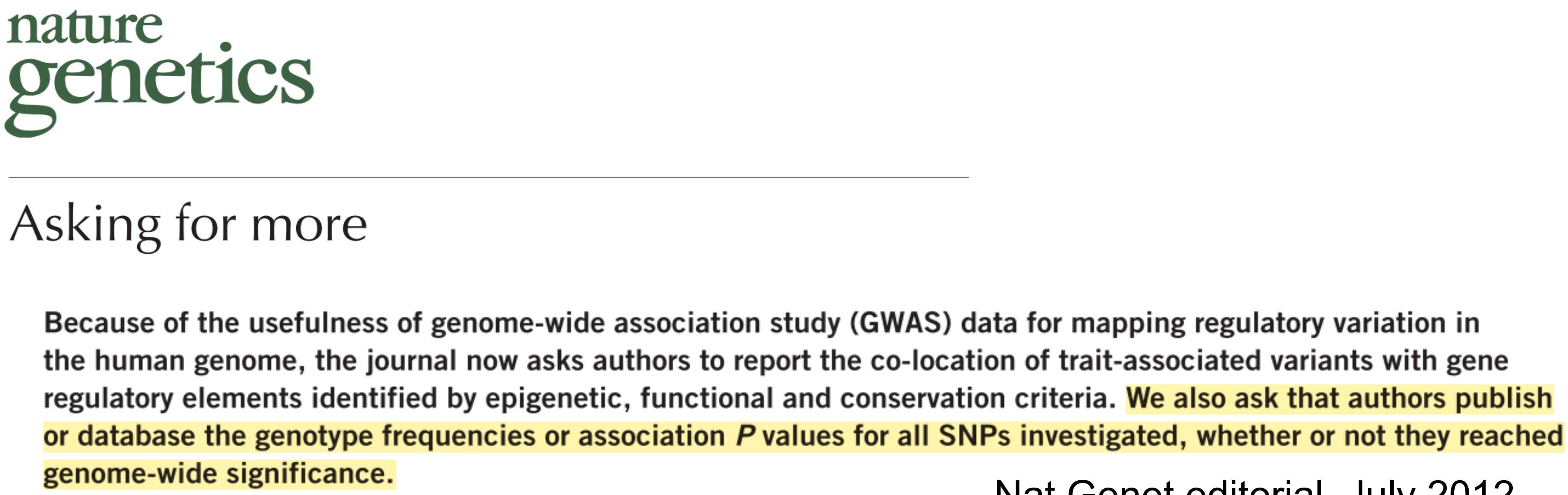
Slides adapted from Jian Zeng

Outline

- Why use GWAS summary statistics (SumStats)?
- Where to download?
- What should we check?
    - About the study?
    - About the data?
- What can we do with them?

Sharing of GWAS summary statistics

There is a consensus within the human genetics research community that it is standard to publicly share 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.

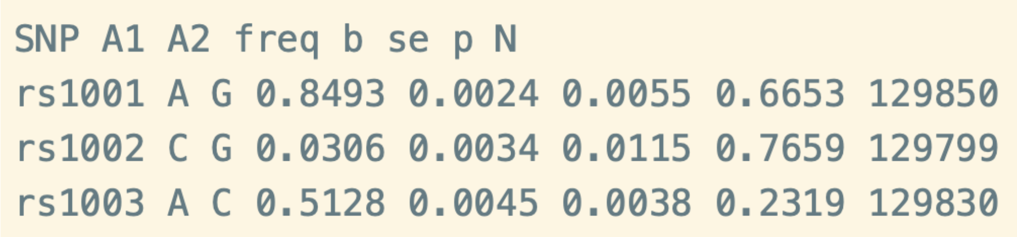
—Nat Genet editorial, July 2012

Why use GWAS SumStats?

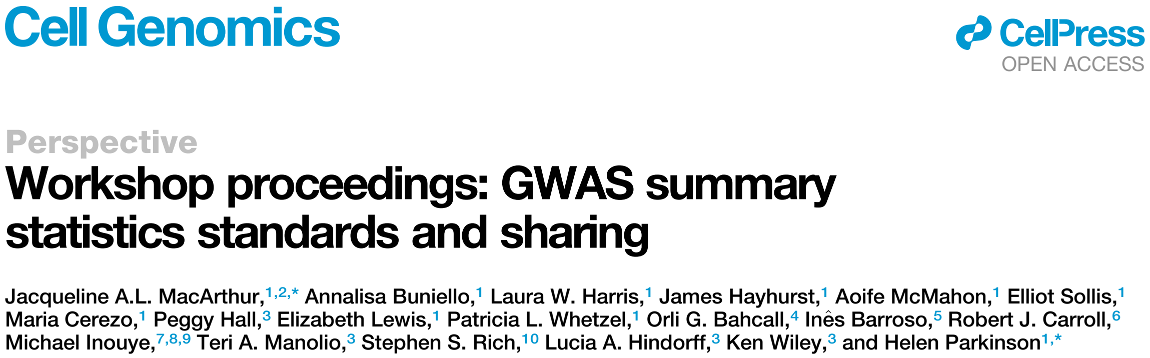
- Access to large sample of individual level data is rare but publishing the summary statistics is a standard
- Unless your phenotype is novel, it is likely a GWAS has already been performed
- Allows us to harness much larger sample sizes

What are GWAS SumStats?

The aggregate association data for every SNP analysed in a GWAS



```
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
```



Cell Genomics

CellPress
OPEN ACCESS

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,*]

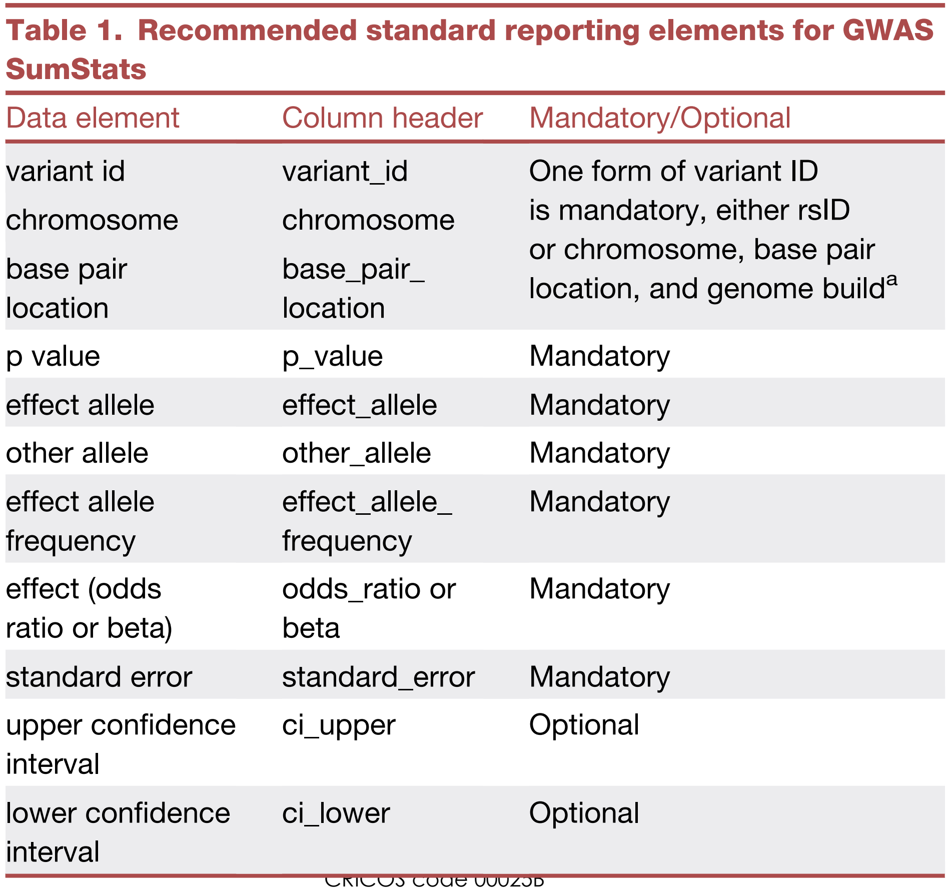


Table 1. Recommended standard reporting elements for GWAS SumStats

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

Where to download GWAS SumStats?

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] ✉

Database	Content
GWAS Catalog https://www.ebi.ac.uk/gwas/	GWAS summary statistics and GWAS lead SNPs reported in GWAS papers
GeneAtlas http://geneatlas.roslin.ed.ac.uk/	UK Biobank GWAS summary statistics
Pan UKBB https://pan.ukbb.broadinstitute.org/	UK Biobank GWAS summary statistics
GWAS Atlas https://atlas.ctglab.nl/	Collection of publicly available GWAS summary statistics with follow-up in silico analysis
FinnGen results https://www.finngen.fi/en/access_results	GWAS summary statistics released from FinnGen, a project that collected biological samples from many sources in Finland
dbGAP https://www.ncbi.nlm.nih.gov/gap/	Public depository of National Institutes of Health-funded genomics data including GWAS summary statistics
OpenGWAS database https://gwas.mrcieu.ac.uk/	GWAS summary data sets
Pheweb.jp https://pheweb.jp/	GWAS summary statistics of Biobank Japan and cross-population meta-analyses

Large GWAS Consortia

There are lots of consortia..

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

- Psychiatric disorders

GIANT (https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files)

- Anthropometric traits

ENIGMA (http://enigma.ini.usc.edu/research/download-enigma-gwas-results/)

- Subcortical brain and hippocampal volumes

GLGC (http://lipidgenetics.org/)

- Global lipids genetics consortium

SSGAC (https://www.thessgac.org/data)

- Social Sciences Genetic Association Consortium - social and psychological traits

EGG (https://egg-consortium.org/)

- Traits related to early growth.

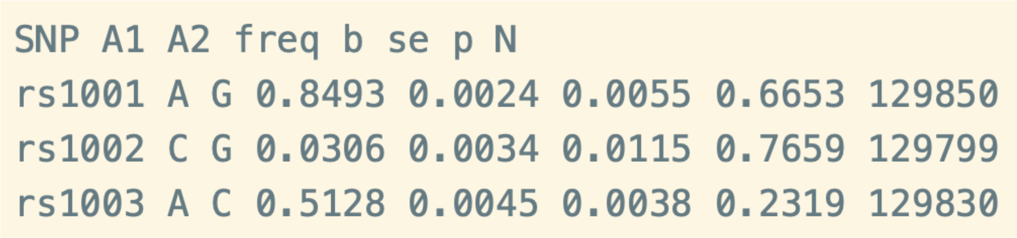
Critical information from the study

- What is the phenotype?
  - How was it measured?
  - How was it treated e.g. transformed?
- What QC has been done? Covariates?
- What sample was this performed in?
  - Sample size
  - Genetic ancestry
- If you plan to use sumstats from more than one study, is there sample overlap?

Critical information from GWAS SumStats

Is there a ReadMe?

- 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
```

What should we check prior to the analysis?

Raw data file

Item	What could be wrong?	How to fix?
Genome build	Inconsistent coordinates among GWAS summary data and LD reference.	Lift up to the same genome build using *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 $p = \frac{N_{case}\, p_{case} + N_{ctrl}\, p_{ctrl}}{N_{case} + N_{ctrl}}$.

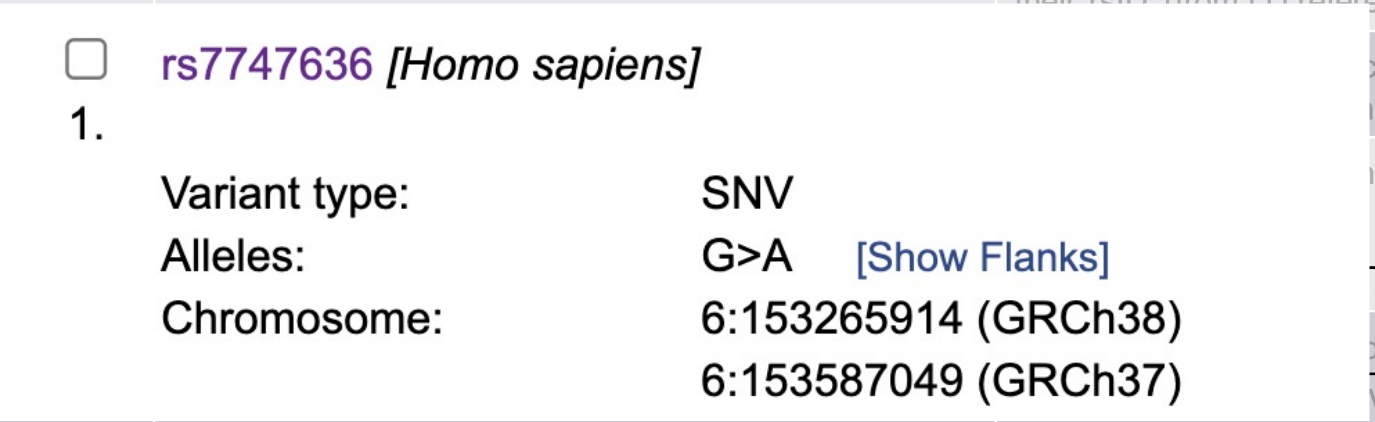
Marginal effect (b)	Provided data are Z-score or odds ratio (OR).	b = Z/SE if SE is provided,
		or $b = Z/\sqrt{2p(1-p)(N+Z^2)}$ given unit variance.
		b = log(OR).

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?

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		k if the predictor is n the phenotype.
Effect allele freq (p)		nce. $2pq = 1/(N * SE + N * b^2)$. $\frac{+ N_{ctrl} p_{ctrl}}{+ N_{ctrl}}$.
Marginal effect (		d, $N + Z^2)$ 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.



□ rs7747636 *[Homo sapiens]*

1.

Variant type: SNV

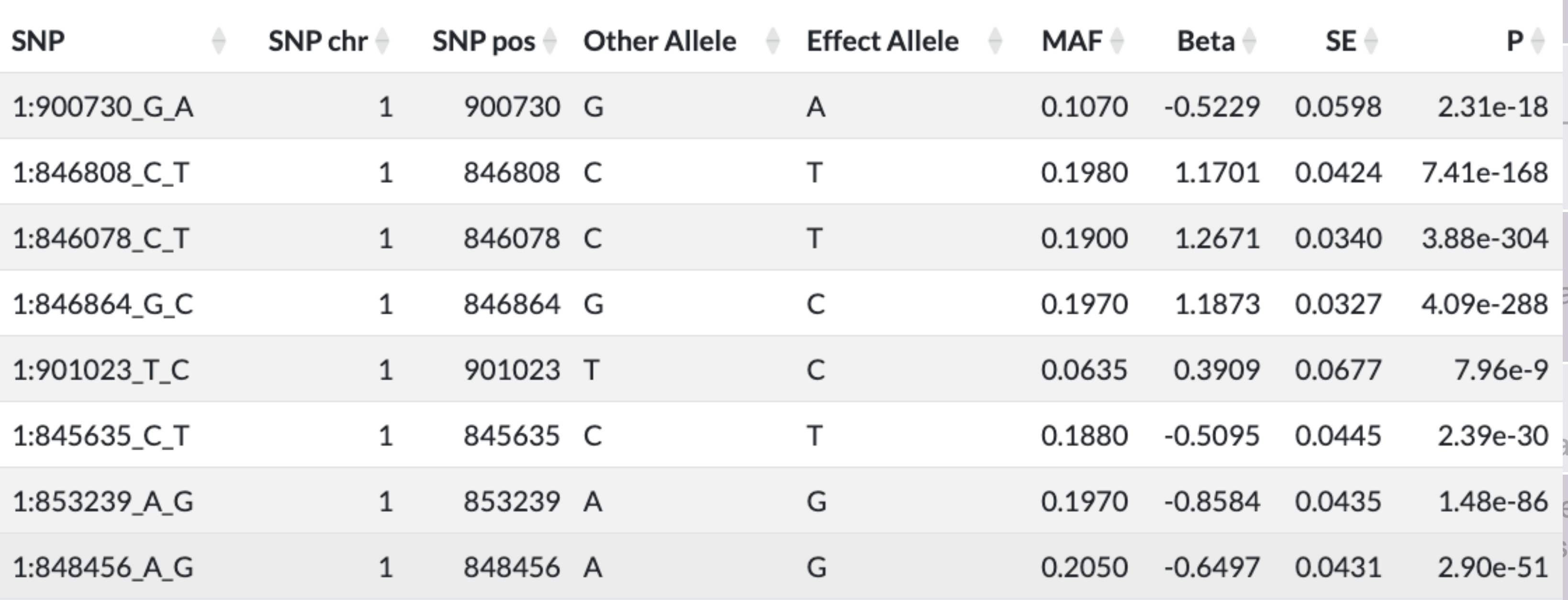
Alleles: G>A [Show Flanks]

Chromosome: 6:153265914 (GRCh38)
6:153587049 (GRCh37)

What should we check prior to the analysis?

Raw data file

Item	What could be wrong?	How to fix?
Genome build	Inconsistent coordinates among GWAS summary data and LD reference.	Lift up to the same genome build using *liftover*
SNP ID	rsID not provided.	Use chromosome and position information to find their rsID (from reference file).
Alleles	Lower/upper case.	Check ReadMe file. Check if the predictor is
Effect allele frequenc (p)		$+ N * b^2)$.
Marginal effect (b)		ance.
Standard error (SE)		ariance.
Sample size (N)		equire sample size.
Incorrect data field format.	Some data field has NA and is non-numeric.	Convert to correct format and filter/impute missing data.

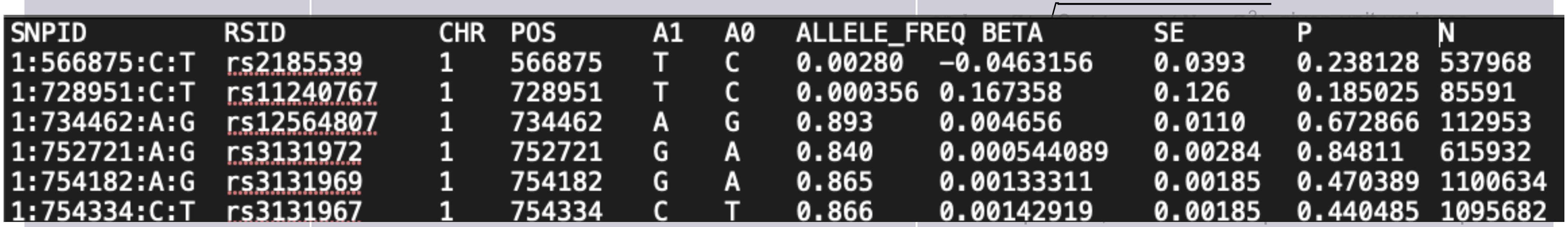


SNP	SNP chr	SNP pos	Other Allele	Effect Allele	MAF	Beta	SE	P
1:900730_G_A	1	900730	G	A	0.1070	-0.5229	0.0598	2.31e-18
1:846808_C_T	1	846808	C	T	0.1980	1.1701	0.0424	7.41e-168
1:846078_C_T	1	846078	C	T	0.1900	1.2671	0.0340	3.88e-304
1:846864_G_C	1	846864	G	C	0.1970	1.1873	0.0327	4.09e-288
1:901023_T_C	1	901023	T	C	0.0635	0.3909	0.0677	7.96e-9
1:845635_C_T	1	845635	C	T	0.1880	-0.5095	0.0445	2.39e-30
1:853239_A_G	1	853239	A	G	0.1970	-0.8584	0.0435	1.48e-86
1:848456_A_G	1	848456	A	G	0.2050	-0.6497	0.0431	2.90e-51

What should we check prior to the analysis?

Raw data file

Item	What could be wrong?	How to fix?
Genome build	Inconsistent coordinates among GWAS summary data and LD reference.	Lift up to the same genome build using *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,
Incorrect data field format.	Some data field has NA and is non-numeric.	Convert to correct format and filter/impute missing data.



```
SNPID          RSID          CHR  POS     A1   A0   ALLELE_FREQ BETA        SE       P        N
1:566875:C:T   rs2185539     1    566875  T    C    0.00280   -0.0463156   0.0393   0.238128 537968
1:728951:C:T   rs11240767    1    728951  T    C    0.000356  0.167358     0.126    0.185025 85591
1:734462:A:G   rs12564807    1    734462  A    G    0.893     0.004656     0.0110   0.672866 112953
1:752721:A:G   rs3131972     1    752721  G    A    0.840     0.000544089  0.00284  0.84811  615932
1:754182:A:G   rs3131969     1    754182  G    A    0.865     0.00133311   0.00185  0.470389 1100634
1:754334:C:T   rs3131967     1    754334  C    T    0.866     0.00142919   0.00185  0.440485 1095682
```

What should we check prior to the analysis?

Raw data file

Item	What could be wrong?	How to fix?
Genome build	Inconsistent coordinates among GWAS summary data and LD reference.	Lift up to the same genome build using *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.

What can we do with them?

- **Meta-analysis**: METAL, MTAG
- **Finding independent association loci**: PLINK-clumping, GCTA-COJO
- **Fine-mapping causal variants**: SuSiE, FINEMAP

These will be covered on Tuesday

- **Variant annotation**: ANNOVAR
- **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
- …

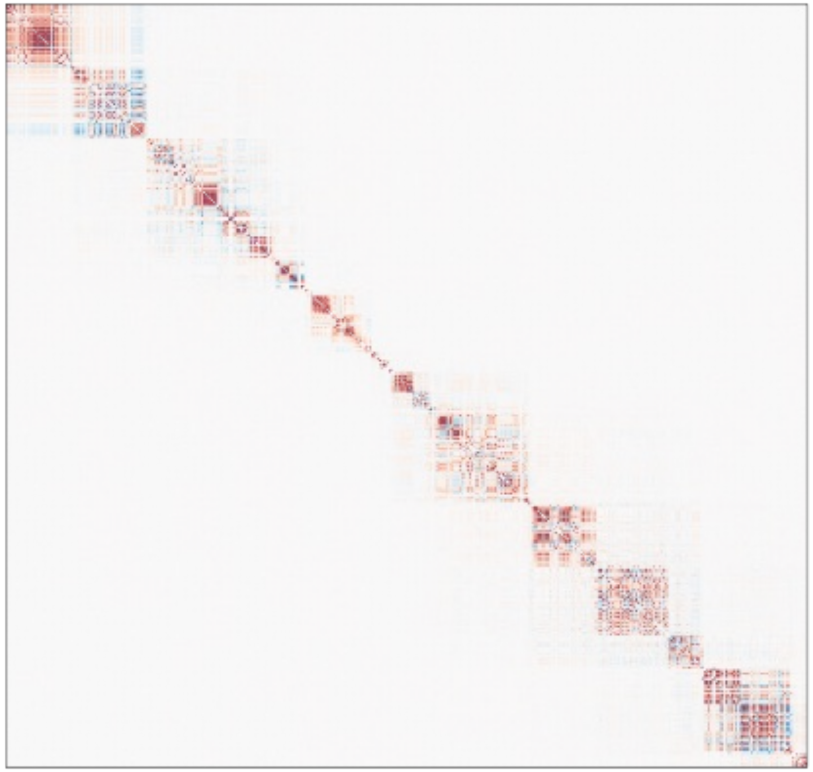
Linkage disequilibrium (LD) correlations

Usually obtained from a reference population

LD correlation matrix

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

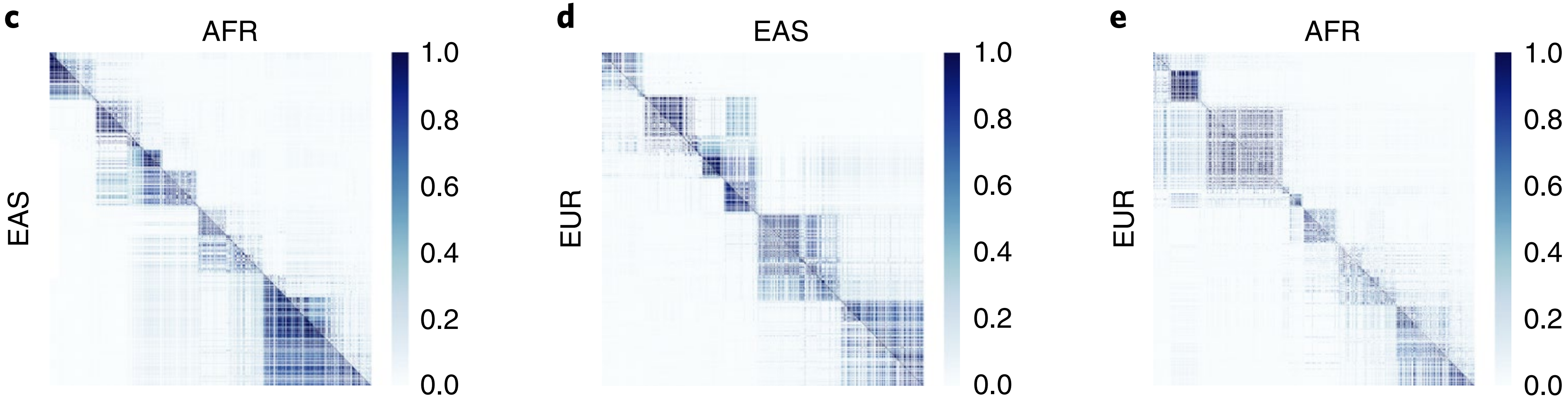
assuming $\mathbf{X}$ is standardised with mean zero and variance one



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



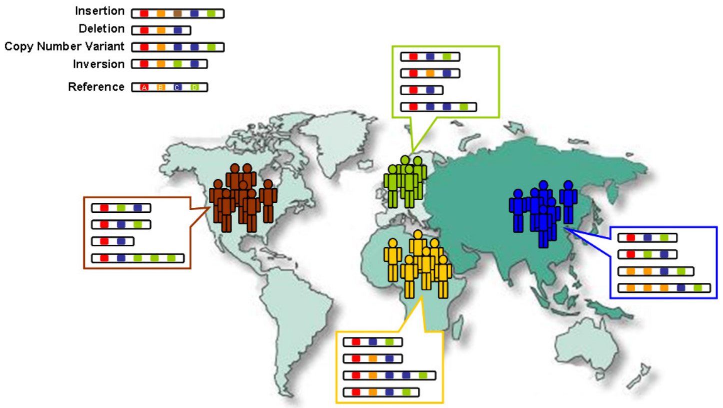
Martin et al 2019 Nature Genetics

Where to find LD reference data?

1000 Genomes Project (1KGP)

Individual sequence data

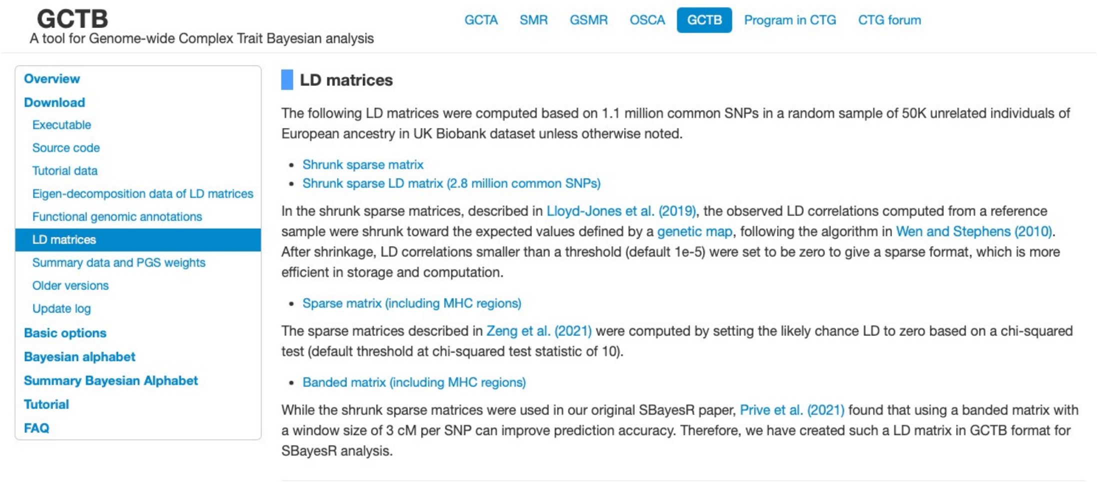
https://www.internationalgenome.org



UK Biobank (UKB)

We provide LD matrices computed from a subset of UKB samples

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


GCTB
A tool for Genome-wide Complex Trait Bayesian analysis
GCTA
SMR
GSMR
OSCA
GCTB
Program in CTG
CTG forum
Overview
Download
Executable
Source code
Tutorial data
Eigen-decomposition data of LD matrices
Functional genomic annotations
LD matrices
Summary data and PGS weights
Older versions
Update log
Basic options
Bayesian alphabet
Summary Bayesian Alphabet
Tutorial
FAQ

LD matrices

The following LD matrices were computed based on 1.1 million common SNPs in a random sample of 50K unrelated individuals of European ancestry in UK Biobank dataset unless otherwise noted.

- Shrunk sparse matrix
- Shrunk sparse LD matrix (2.8 million common SNPs)

In the shrunk sparse matrices, described in Lloyd-Jones et al. (2019), the observed LD correlations computed from a reference sample were shrunk toward the expected values defined by a genetic map, following the algorithm in Wen and Stephens (2010). After shrinkage, LD correlations smaller than a threshold (default 1e-5) were set to be zero to give a sparse format, which is more efficient in storage and computation.

- Sparse matrix (including MHC regions)

The sparse matrices described in Zeng et al. (2021) were computed by setting the likely chance LD to zero based on a chi-squared test (default threshold at chi-squared test statistic of 10).

- Banded matrix (including MHC regions)

While the shrunk sparse matrices were used in our original SBayesR paper, Prive et al. (2021) found that using a banded matrix with a window size of 3 cM per SNP can improve prediction accuracy. Therefore, we have created such a LD matrix in GCTB format for SBayesR analysis.

Summary

- GWAS summary statistics are publicly available for almost every trait you could think of
- Before using publicly available data make sure you understand how it was created and what it is comprised of
  - The checks you will want to do will depend on what you plan to do with the data