LD Score Regression Practical
Part I: SNP Heritability

International Statistical Genetics Workshop

IBG, University of Colorado Boulder

Wednesday, 5th March 2025

MadhurBain Singh

singhm18@vcu.edu

Virginia Institute for Psychiatric and Behavior Genetics

Virginia Commonwealth University, Richmond, VA

Thanks to Benjamin Neale and Michel Nivard.

Practical Working Directory

```
cd  ~/practicals/3.1.SNPHeritability_MichelNivard/final
```

Qualtrics link

This link is also on top of the R script: ***LDSC_Practical1_h2_SNP.R***

https://qimr.az1.qualtrics.com/jfe/form/SV_29tZDm8QmlN31Q2

Only TWO Primary Steps to Run LDSC

We are using **`{GenomicSEM}`** library in R to run LDSC

1. Munge the summary statistics: **`munge()`**

   munge = convert raw data from one form to another

2. Run LD-Score Regression: **`ldsc()`**

The summary statistics files input to **`munge()`** at a minimum need to contain five pieces of information:

1. The rsID of the SNP.
2. An A1 allele column, indicating the effect allele.
3. An A2 allele column, indicating the non-effect allele.
4. A signed (+/-) effect column.
5. The *p*-value associated with this effect.

The `munge()` function takes 6 arguments:

1. **files**: The name of the summary statistics files
2. **hm3**: The name of the reference file. Here we use Hapmap 3 SNPs.
3. **trait.names**: The trait names that will be used to name the saved files
4. **N**: The sample sizes associated with the traits.
5. **info.filter**: INFO filter. Package default is to retain SNPs with INFO > 0.9.
6. **maf.filter**: MAF filter. Package default is to retain SNPs with MAF > 0.01.

The `ldsc()` function takes 6 arguments:

1. **traits**: a vector of file names/paths to files which point to the munged sumstats.
2. **sample.prev**: A vector of sample prevalences of length equal to the number of traits. If the trait is continuous, the values should equal NA.
3. **population.prev**: A vector of population prevalences. If the trait is continuous the values should equal NA.

4. **ld**: A folder of LD scores used as the independent variable in LDSC

5. **wld**: A folder of LDSC weights (Typically the same folder as specified for the ld argument)

6. trait.names: The trait names.

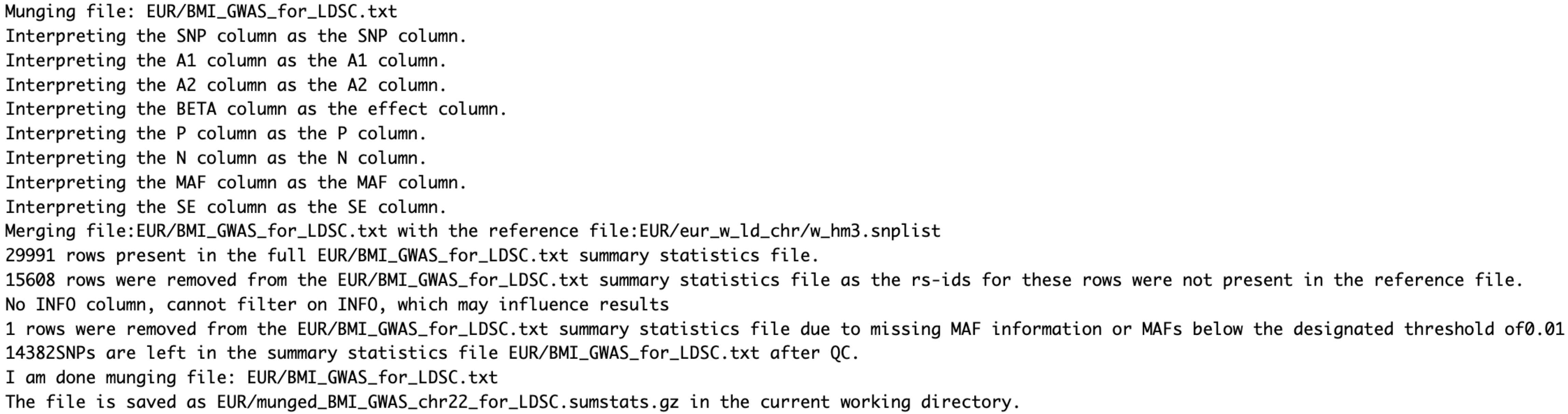
On To The Practical

Continuous phenotype – BMI (body mass index)

GWAS Sum Stats from the meta-analysis of GIANT Consortium (Locke et al., 2018) and UK Biobank

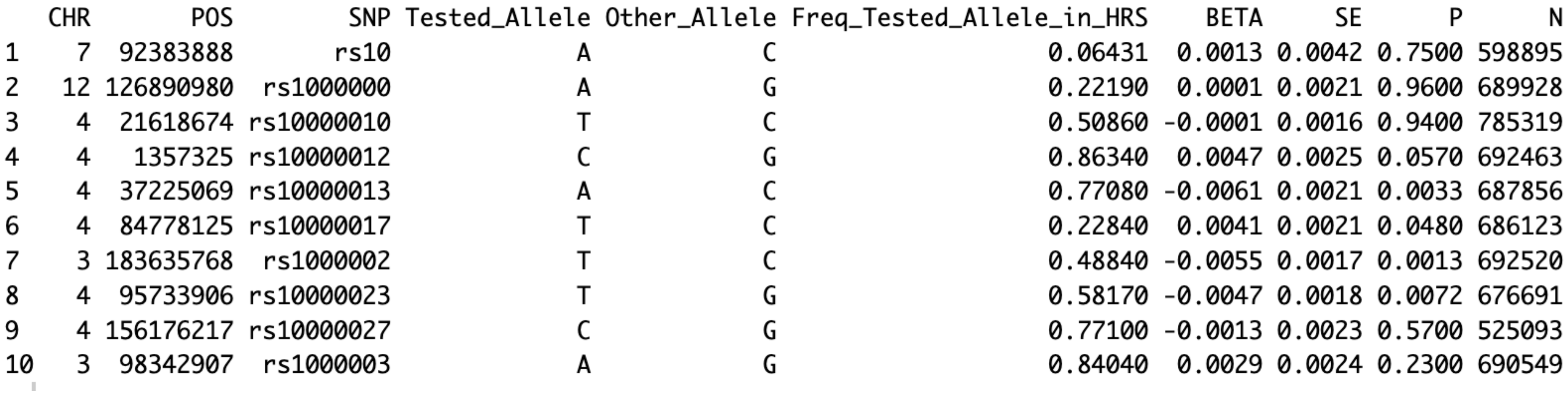
https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files

`munge` .log File



```
Munging file: EUR/BMI_GWAS_for_LDSC.txt
Interpreting the SNP column as the SNP column.
Interpreting the A1 column as the A1 column.
Interpreting the A2 column as the A2 column.
Interpreting the BETA column as the effect column.
Interpreting the P column as the P column.
Interpreting the N column as the N column.
Interpreting the MAF column as the MAF column.
Interpreting the SE column as the SE column.
Merging file:EUR/BMI_GWAS_for_LDSC.txt with the reference file:EUR/eur_w_ld_chr/w_hm3.snplist
29991 rows present in the full EUR/BMI_GWAS_for_LDSC.txt summary statistics file.
15608 rows were removed from the EUR/BMI_GWAS_for_LDSC.txt summary statistics file as the rs-ids for these rows were not present in the reference file.
No INFO column, cannot filter on INFO, which may influence results
1 rows were removed from the EUR/BMI_GWAS_for_LDSC.txt summary statistics file due to missing MAF information or MAFs below the designated threshold of0.01
14382SNPs are left in the summary statistics file EUR/BMI_GWAS_for_LDSC.txt after QC.
I am done munging file: EUR/BMI_GWAS_for_LDSC.txt
The file is saved as EUR/munged_BMI_GWAS_chr22_for_LDSC.sumstats.gz in the current working directory.
```

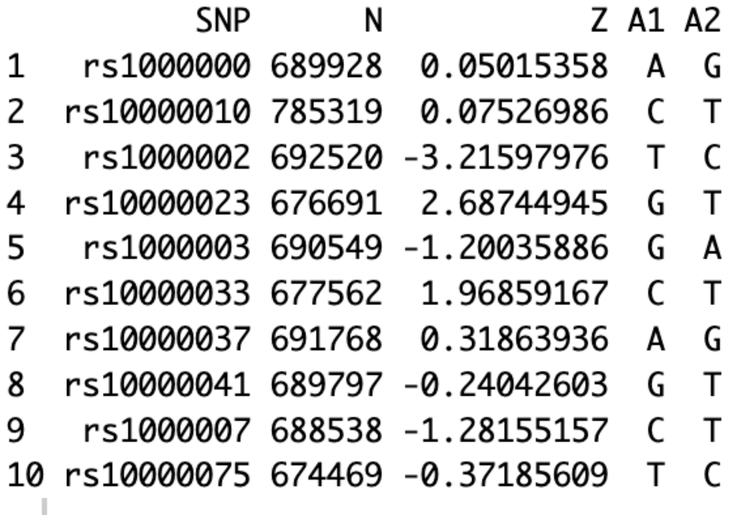
Raw GWAS Sum Stats



```
   CHR       POS        SNP Tested_Allele Other_Allele Freq_Tested_Allele_in_HRS    BETA     SE      P      N
1    7  92383888       rs10             A            C                   0.06431  0.0013 0.0042 0.7500 598895
2   12 126890980  rs1000000             A            G                   0.22190  0.0001 0.0021 0.9600 689928
3    4  21618674 rs10000010             T            C                   0.50860 -0.0001 0.0016 0.9400 785319
4    4   1357325 rs10000012             C            G                   0.86340  0.0047 0.0025 0.0570 692463
5    4  37225069 rs10000013             A            C                   0.77080 -0.0061 0.0021 0.0033 687856
6    4  84778125 rs10000017             T            C                   0.22840  0.0041 0.0021 0.0480 686123
7    3 183635768  rs1000002             T            C                   0.48840 -0.0055 0.0017 0.0013 692520
8    4  95733906 rs10000023             T            G                   0.58170 -0.0047 0.0018 0.0072 676691
9    4 156176217 rs10000027             C            G                   0.77100 -0.0013 0.0023 0.5700 525093
10   3  98342907  rs1000003             A            G                   0.84040  0.0029 0.0024 0.2300 690549
```

2,336,269 Genetic Variants

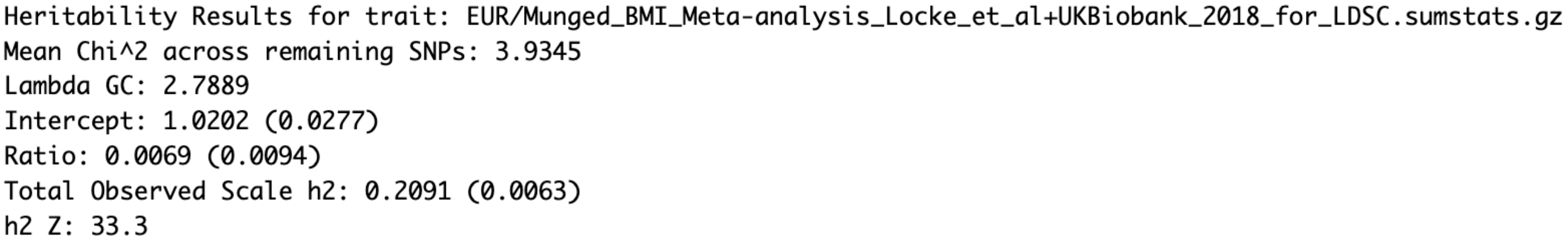
Munged GWAS Sum Stats



```
          SNP      N           Z A1 A2
1   rs1000000 689928  0.05015358  A  G
2  rs10000010 785319  0.07526986  C  T
3   rs1000002 692520 -3.21597976  T  C
4  rs10000023 676691  2.68744945  G  T
5   rs1000003 690549 -1.20035886  G  A
6  rs10000033 677562  1.96859167  C  T
7  rs10000037 691768  0.31863936  A  G
8  rs10000041 689797 -0.24042603  G  T
9   rs1000007 688538 -1.28155157  C  T
10 rs10000075 674469 -0.37185609  T  C
```

1,019,839 Genetic Variants

`ldsc` .log File



```
Heritability Results for trait: EUR/Munged_BMI_Meta-analysis_Locke_et_al+UKBiobank_2018_for_LDSC.sumstats.gz
Mean Chi^2 across remaining SNPs: 3.9345
Lambda GC: 2.7889
Intercept: 1.0202 (0.0277)
Ratio: 0.0069 (0.0094)
Total Observed Scale h2: 0.2091 (0.0063)
h2 Z: 33.3
```

Attenuation Ratio

$$\frac{LDSC\ Intercept\ -1}{Mean\ \chi^2\ -1}$$

Under no confounding,

LDSC Intercept = 1

Attenuation Ratio = 0

Take-home points

- LDSC allows us to estimate SNP heritability using only the GWAS sum stats.
- Importantly, LDSC helps us differentiate the true genetic signal (polygenicity) from confounding (population stratification) in GWAS.

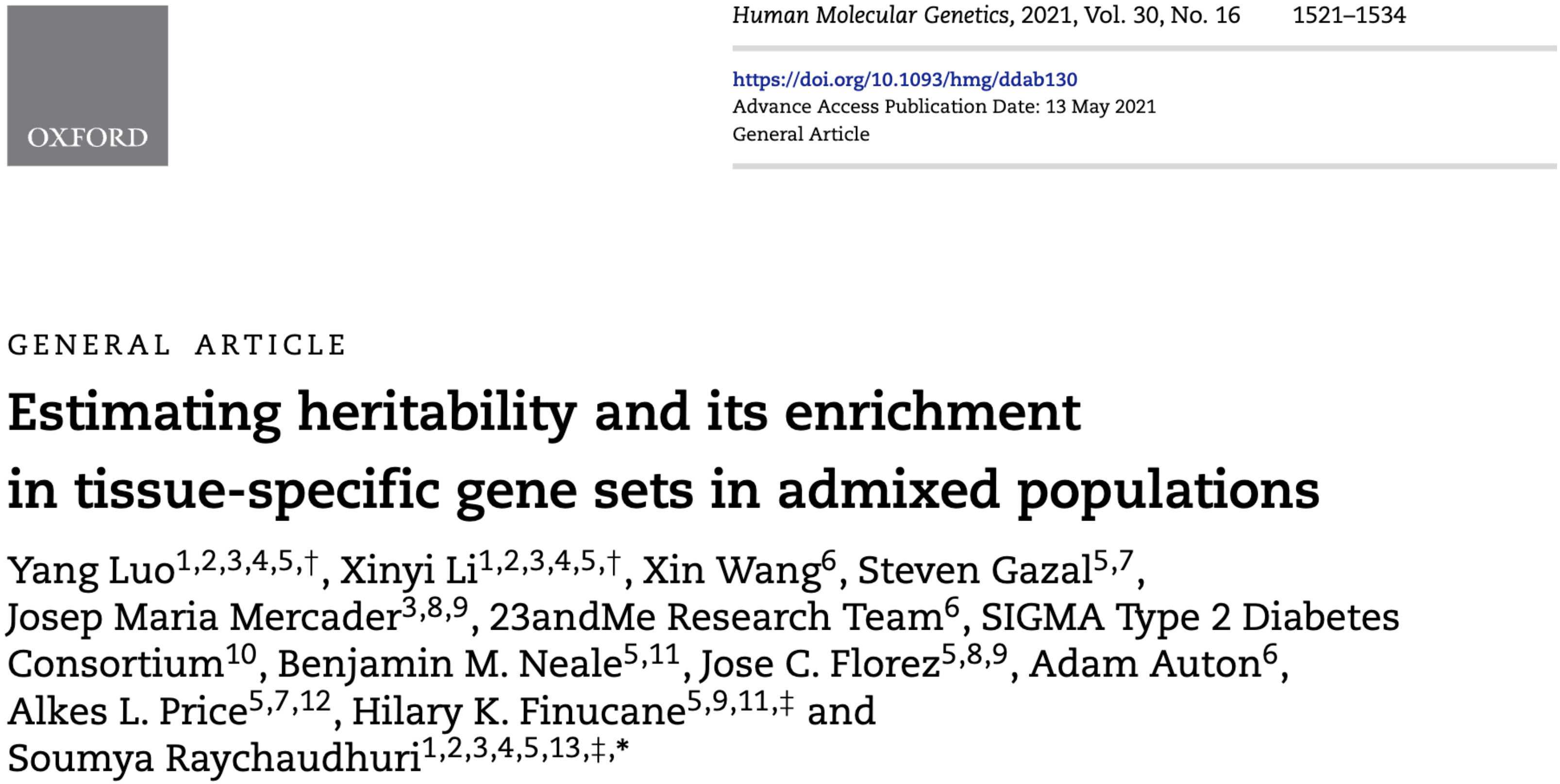
NOTE

- LDSC first estimates the heritability/variance per SNP. The program then computes the total variance across all common variants in the genome [i.e., the SNP heritability].
- **If the sum stats used to run LDSC analyses are NOT a random subset of all SNPs across the genome, the estimated per-SNP variance and, thus, the SNP-based heritability will be biased.**

Caveat 1: LDSC in Admixed Populations

- We use LD scores computed in an external, ancestrally matched dataset.
- We assume that the out-of-sample LD scores match the in-sample pairwise SNP correlations underlying the GWAS sum stats.
- This assumption would not hold in GWAS in a sample of individuals with admixed ancestry.
  - Due to long-range pairwise SNP correlations arising from admixture.

cov-LDSC



Human Molecular Genetics, 2021, Vol. 30, No. 16 1521–1534

https://doi.org/10.1093/hmg/ddab130
Advance Access Publication Date: 13 May 2021
General Article

GENERAL ARTICLE

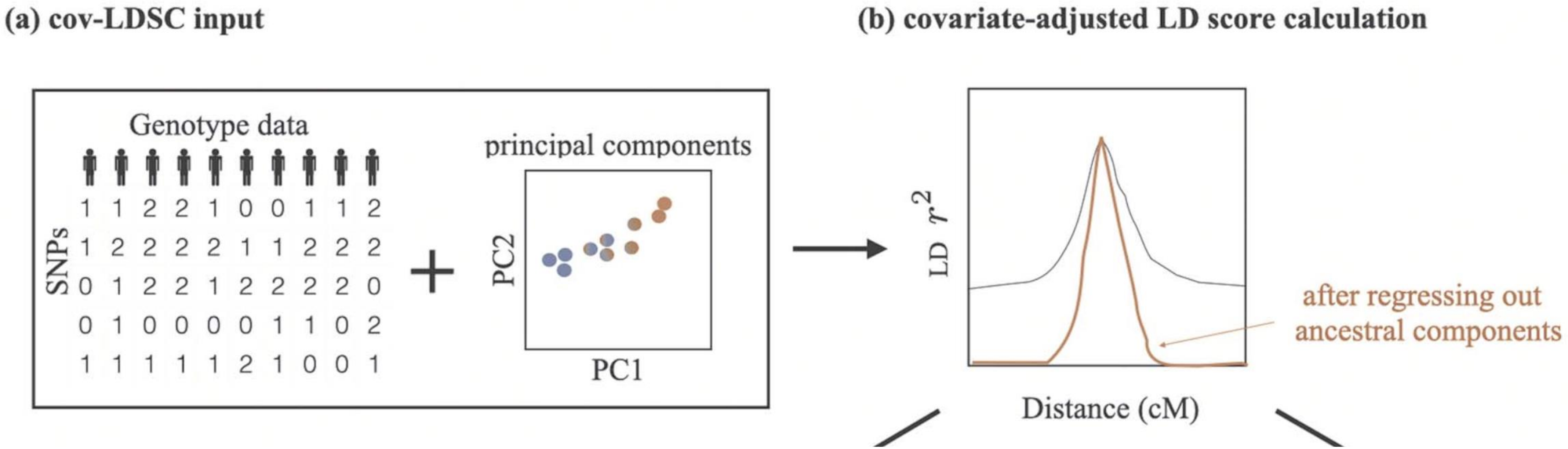
Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations

Yang Luo[1,2,3,4,5,†], Xinyi Li[1,2,3,4,5,†], Xin Wang[6], Steven Gazal[5,7], Josep Maria Mercader[3,8,9], 23andMe Research Team[6], SIGMA Type 2 Diabetes Consortium[10], Benjamin M. Neale[5,11], Jose C. Florez[5,8,9], Adam Auton[6], Alkes L. Price[5,7,12], Hilary K. Finucane[5,9,11,‡] and Soumya Raychaudhuri[1,2,3,4,5,13,‡,*]

cov-LDSC

- Using in-sample covariate-adjusted LD scores.
- Covariate-adjusted LD scores estimated
  - In (a random subset of) the GWAS sample
  - Conditional on the covariates (e.g., PCs) used in the GWAS

cov-LDSC



(a) cov-LDSC input

(b) covariate-adjusted LD score calculation

1	1	2	2	1	0	0	1	1	2
1	2	2	2	2	1	1	2	2	2
0	1	2	2	1	2	2	2	2	0
0	1	0	0	0	0	1	1	0	2
1	1	1	1	1	2	1	0	0	1

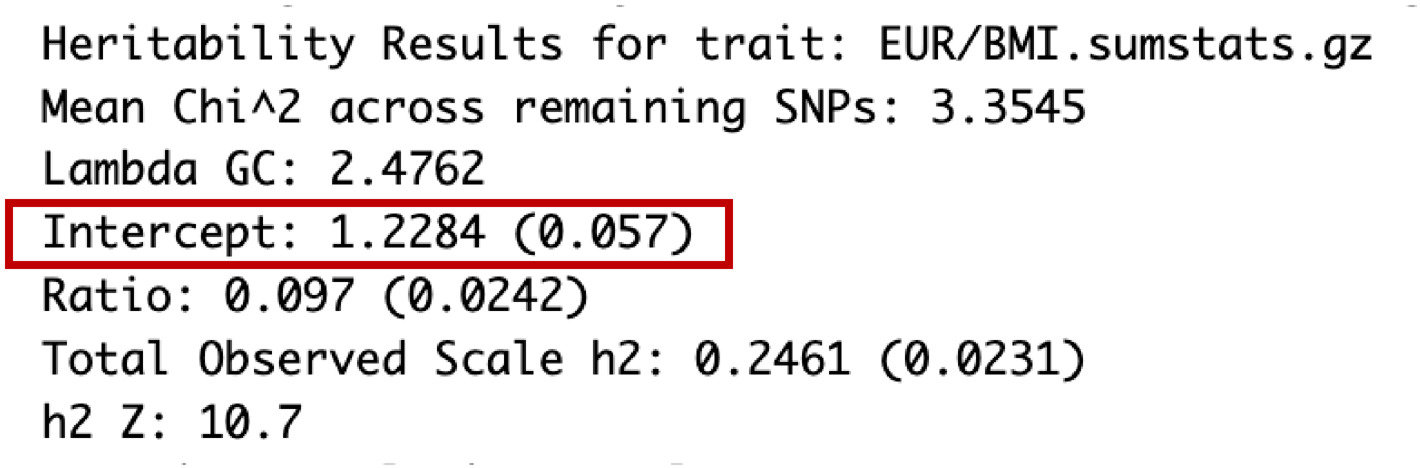
Caveat 2: LDSC with Sum Stats from Linear Mixed Models

Example: Sum Stats from Pan-UKBB project.

GWAS performed with SAIGE (linear mixed model)

Pan-UK Biobank

Pan-ancestry genetic analysis of the UK Biobank



```
Heritability Results for trait: EUR/BMI.sumstats.gz
Mean Chi^2 across remaining SNPs: 3.3545
Lambda GC: 2.4762
Intercept: 1.2284 (0.057)
Ratio: 0.097 (0.0242)
Total Observed Scale h2: 0.2461 (0.0231)
h2 Z: 10.7
```

Caveat 2: LDSC with Sum Stats from Linear Mixed Models

LDSC intercept *may* be >1

With large sample sizes (e.g., UK Biobank)

For traits with high SNP heritability

Ref: Loh, PR., Kichaev, G., Gazal, S. et al. Mixed-model association for biobank-scale datasets. *Nat Genet* 50, 906–908 (2018). https://doi.org/10.1038/s41588-018-0144-6

Points to consider

- Effective N (will be less than the raw N due to related individuals)
- Residual confounding?

References

Bulik-Sullivan, B., Loh, PR., Finucane, H. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 47, 291–295 (2015). https://doi.org/10.1038/ng.3211

Luo, Y., Li, X., Wang, X., et al. Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations, *Human Molecular Genetics*, 30, 1521–1534 (2021). https://doi.org/10.1093/hmg/ddab130

Loh, PR., Kichaev, G., Gazal, S. et al. Mixed-model association for biobank-scale datasets. *Nat Genet* 50, 906–908 (2018). https://doi.org/10.1038/s41588-018-0144-6