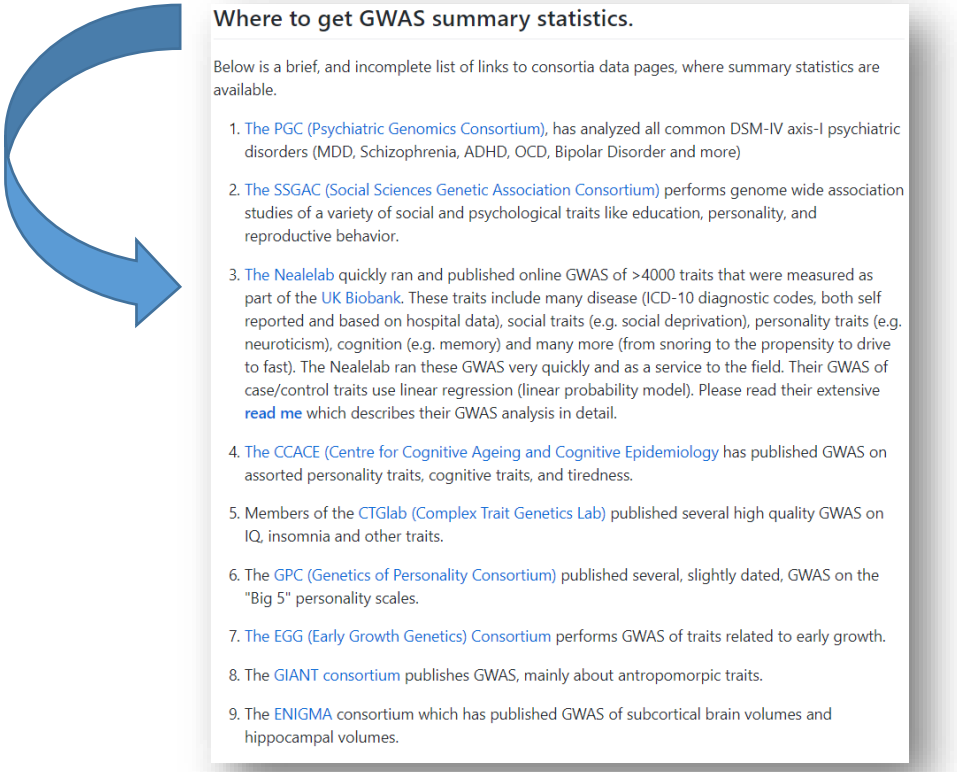
Step 1: The **munge** function

Where to get FREE, PUBLIC summary statistics

- List lots of resources on the Genomic SEM Wiki: https://github.com/GenomicSEM/GenomicSEM/wiki/2.-Important-resources-and-key-information



Where to get GWAS summary statistics.

Below is a brief, and incomplete list of links to consortia data pages, where summary statistics are available.

1. The PGC (Psychiatric Genomics Consortium), has analyzed all common DSM-IV axis-I psychiatric disorders (MDD, Schizophrenia, ADHD, OCD, Bipolar Disorder and more)
2. The SSGAC (Social Sciences Genetic Association Consortium) performs genome wide association studies of a variety of social and psychological traits like education, personality, and reproductive behavior.
3. The Nealelab quickly ran and published online GWAS of >4000 traits that were measured as part of the UK Biobank. These traits include many disease (ICD-10 diagnostic codes, both self reported and based on hospital data), social traits (e.g. social deprivation), personality traits (e.g. neuroticism), cognition (e.g. memory) and many more (from snoring to the propensity to drive to fast). The Nealelab ran these GWAS very quickly and as a service to the field. Their GWAS of case/control traits use linear regression (linear probability model). Please read their extensive **read me** which describes their GWAS analysis in detail.
4. The CCACE (Centre for Cognitive Ageing and Cognitive Epidemiology has published GWAS on assorted personality traits, cognitive traits, and tiredness.
5. Members of the CTGlab (Complex Trait Genetics Lab) published several high quality GWAS on IQ, insomnia and other traits.
6. The GPC (Genetics of Personality Consortium) published several, slightly dated, GWAS on the "Big 5" personality scales.
7. The EGG (Early Growth Genetics) Consortium performs GWAS of traits related to early growth.
8. The GIANT consortium publishes GWAS, mainly about antropomorpic traits.
9. The ENIGMA consortium which has published GWAS of subcortical brain volumes and hippocampal volumes.

https://www.ebi.ac.uk/gwas/

BioBank Japan PheWeb (PheWeb.jp)

https://pheweb.jp/

FINNGEN

https://www.finngen.fi/en/access_results

Pan-UK Biobank

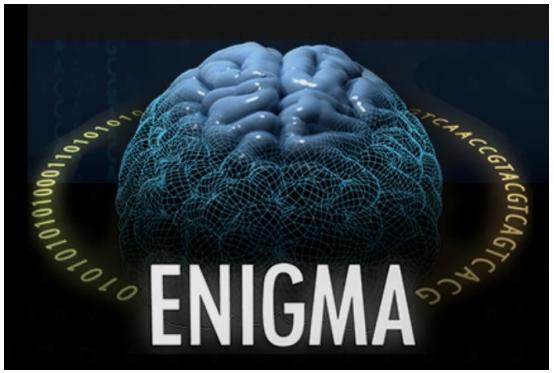
Pan-ancestry genetic analysis of the UK Biobank

https://docs.google.com/spreadsheets/d/1AeeADtT0U1AukliiNyiVzVRdLYPkTbruQSk38DeutU8/edit#gid=268241601

PGC

Psychiatric Genomics Consortium

https://pgc.unc.edu/for-researchers/download-results/



https://enigma.ini.usc.edu/research/download-enigma-gwas-results/

The summary statistics files input into the munge function 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** (+/-); logistic or continuous

5. The **p-value** associated with this effect.

What's in the data?

First rows of the PTSD GWAS data

	MarkerName	Allele1	Allele2	Effect	StdErr	P-value	Direction
1	chr8_125827954_I	d	i2	-0.3033	0.0574	1.259e-07	------+??
2	rs138517393	t	c	1.0194	0.1999	3.411e-07	++++++-??
3	rs1853871	t	c	0.4211	0.0840	5.381e-07	+++-+++++
4	chr5_154981038_I	d	i2	-1.2440	0.2537	9.391e-07	++-?---??
5	rs71247647	t	c	1.0956	0.2262	1.281e-06	+++-++-?+
6	rs9462413	a	t	-0.3971	0.0821	1.302e-06	---+-----

Using GWAS summary statistics for:

- Major Depressive Disorder (Cases = 170,756; Controls = 329,443; Howard et al., 2019)
- Anxiety Disorders (Cases = 31,977; Controls = 82,114; Purves et al., 2020)
- Alcohol use disorder (Cases = 8,485; Controls = 20,272; Walters et al., 2018)
- PTSD (Cases = 2,424; Controls = 7,113; Duncan et al., 2018)

*note that newer versions are available for all of these GWAS; these are being used to mirror what is on the Genomic SEM GitHub wiki pages

6 things to know before getting started

1. Be sure you are using summary statistics calculated within a single ancestry group

2. Be sure to use ancestry-specific LD scores that match the ancestry group of your included GWAS.

3. Typically advisable to only include summary statistics from a GWAS with N >= 10,000.

   The cut-off recommended by the LDSC developers for estimating genetic covariance is a SNP-based heritability Z-Statistic > 4

6 things to know before getting started

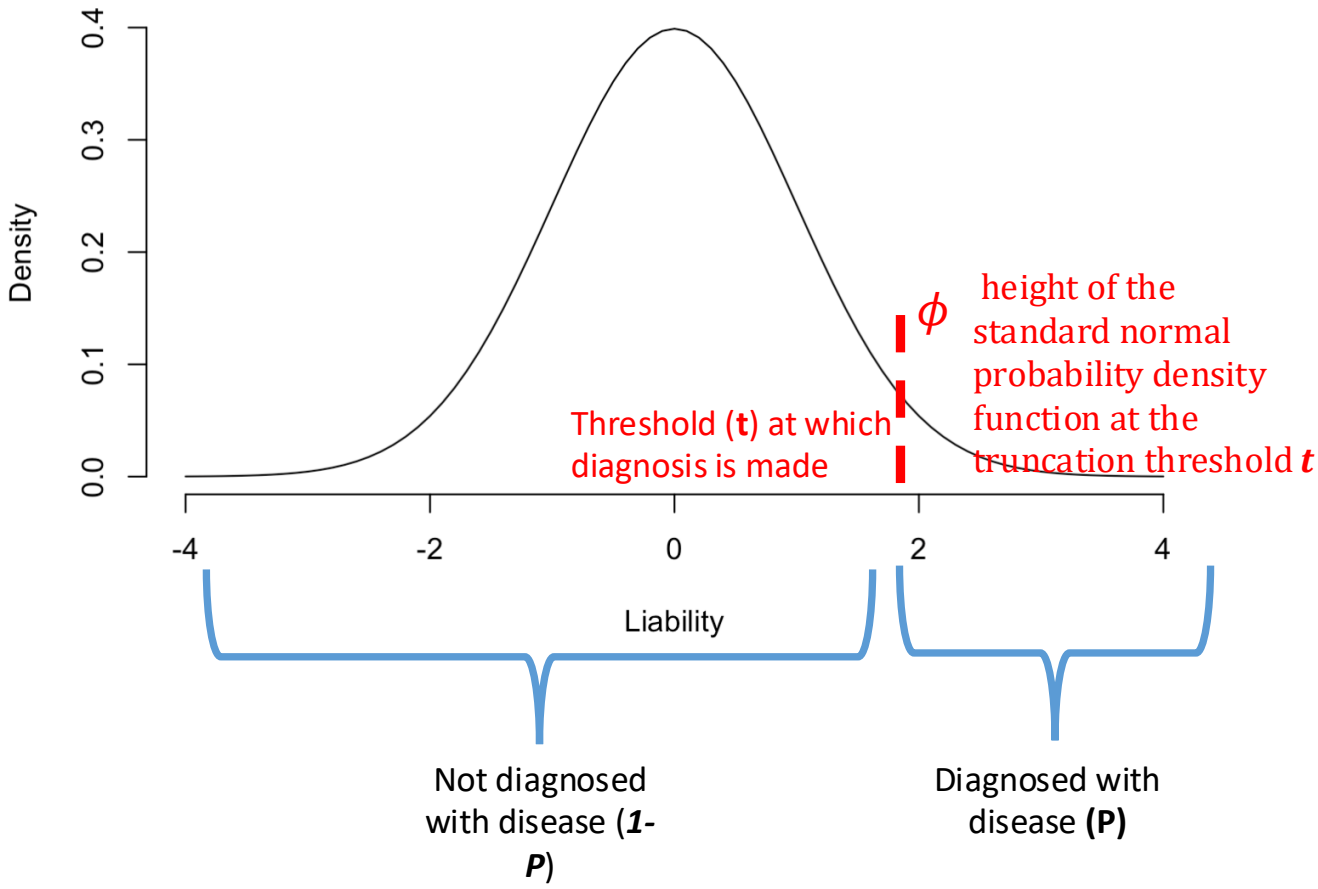
4. LDSC allows for varying and unknown degrees of sample overlap

- You do *not* need to know the specific levels of overlap and the LD-scores are often used from a separate, publicly available reference sample (e.g., 1000 Genomes)

5. Make sure you are not using a pruned list of summary statistics (e.g., the top 5,000 hits)

6. The munge function use sample size to perform necessary conversions. Sample size from summary statistics file or provided by the user. This one can be very tricky and requires different considerations for binary (case/control) traits)
    - Be wary of publicly available summary statistics that exclude certain cohorts (e.g., 23andMe).

Liability Threshold Model



Liability Scale Correction

$$h_l^2 = h_o^2 \frac{P(1-P)}{\phi^2} \frac{P(1-P)}{v(1-v)}$$

This first part of this equation backs out the expected heritability estimate on that continuous distribution of risk (liability)

P is the population prevalence of the disorder.

ϕ is the height of the continuous distribution of liability at the threshold *t* at which a diagnosis of the disorder is made

Liability Scale Correction

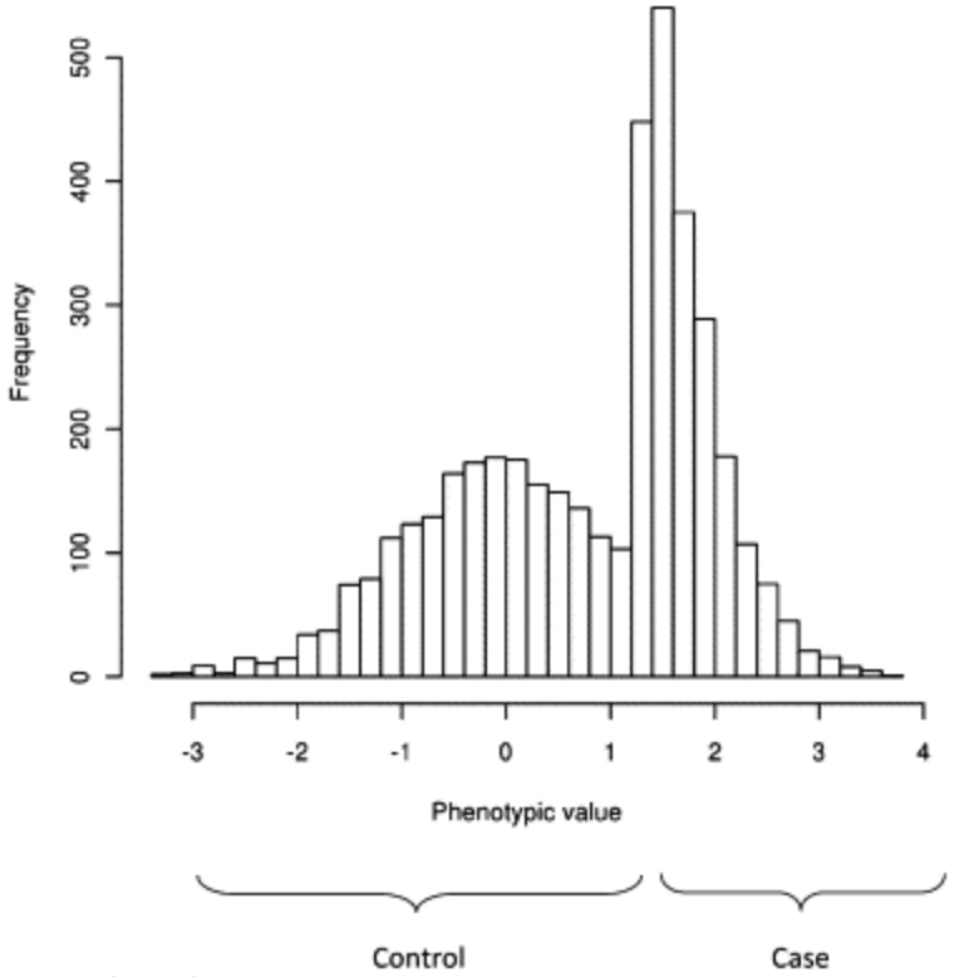
$$h_l^2 = h_o^2 \frac{P(1-P)}{\phi^2} \frac{P(1-P)}{v(1-v)}$$

This second part of this equation performs the correction for <u>participant ascertainment</u>

P is again the population prevalence of the disorder.

v is the prevalence of the disorder in our participant sample.

The ratio then reflects the degree of ascertainment



Lee et al. (2011)

"In case-control studies the proportion of cases is usually (much) larger than the prevalence in the population yet estimates of genetic variation are most interpretable if they are not biased by this ascertainment"

Cohort-specific ascertainment

- As GWAS have continued to grow in sample size they often reflect meta-analyses of a series of contributing cohorts.
- For pragmatic reasons, cohorts often share the GWAS summary stats to be meta-analyzed with other summary stats for the outcome of interest
- When this happens, a correction for ascertainment within each cohort is required.
  - The reason: the ascertainment calculated using total cases and controls is not the same as ascertainment calculated within cohort
    - Sum of parts != sum of totals

Cohort-specific ascertainment

$$h_l^2 = h_o^2 \frac{P(1-P)}{\phi^2} \frac{P(1-P)}{\sum_k v_k(1-v_k)}$$

In order to appropriately perform the ascertainment correction we need to calculate the sum of ascertainment across the contributing cohorts, *k*

Sum of Effective Sample Size Solution

$$EffN_k = 4v_k(1 - v_k)n_k$$

- In practice, we use what's call the effective sample size for this cohort-specific ascertainment correction.
  - The effective sample size is the sample size you would have had if the study design was balanced (50% cases and 50% controls)

- This corrects the sample size for ascertainment and allows for summing sample sizes across cohorts.

- We also use sum of effective N as many GWAS pipelines (e.g., Ricopili) automatically output this for the GWAS and it allows us to provide data in the format expected by LDSC

- This will either be:
  - In the GWAS data as a separate column
  - Something you can calculate from cohort-level information (often in supplementary tables)
  - Something you can back out directly from the data*

The *munge* function takes 4 key arguments:

1.files: The name of the summary statistics files

2.hm3: The name of the reference file. Here we use Hapmap 3 SNPs.

- Note that this means you will only end up with at most 1.1 million SNPs after running munge maximum. This is more than enough to get a good estimate from LDSC
- This HapMap3 file is available from the GenomicSEM GitHub

3.trait.names: The trait names that will be used to name the saved files

4.N: The sample sizes associated with the traits.