Rationale

- A common goal of both EWAS and GWAS is to identify the genes and pathways associated with complex traits.
- If GWAS and EWAS signals of the same trait are predominantly found in the same regions or identify the same genes/pathways then using both study designs for this task is a little redundant. If the opposite is true then it suggests the study designs are capturing different facets of the trait of interest.

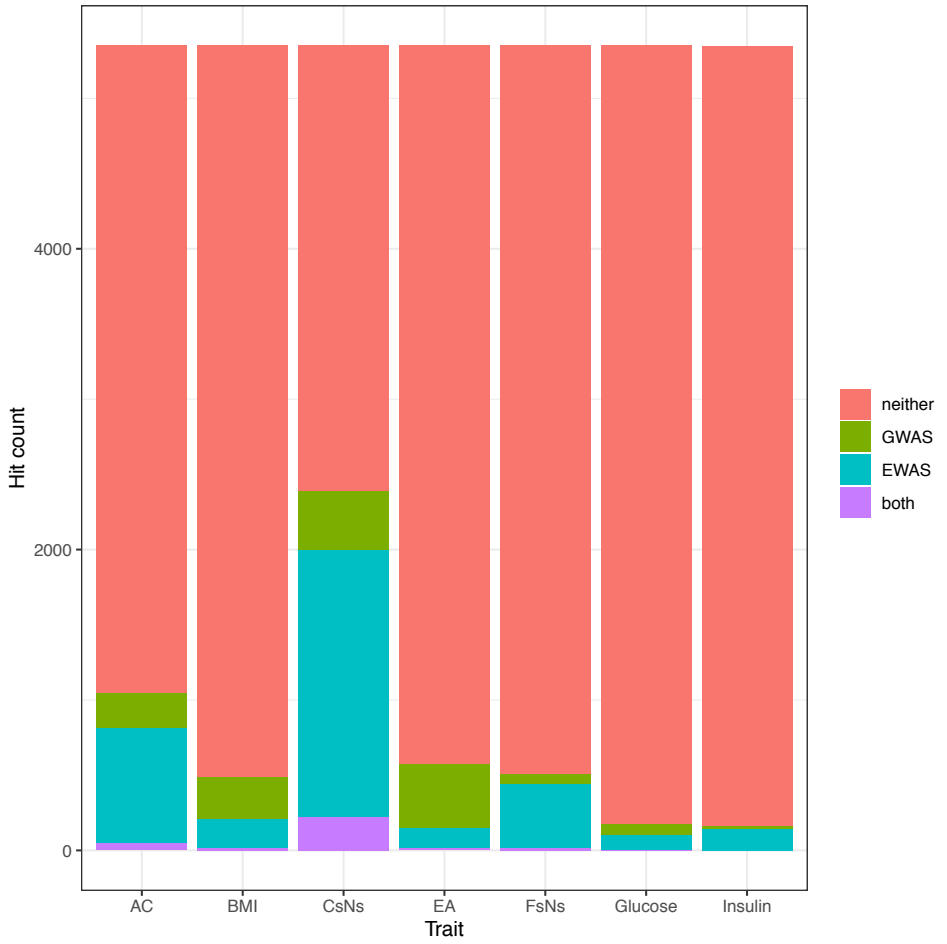
- Hypothesis: As DNAm variation is heritable and GWAS and EWAS are sampling from the same space, I'd expect a little more overlap than expected by chance. As EWAS results are more likely to be confounded and can be downstream of the trait, the overlap isn't expected to be too high.

Rough Methods

- Physical overlap between corresponding GWAS and EWAS (500kb regions)
- Testing best method and whether either method has enough power to detect overlap between pathways and genes
- Gene/pathway overlap between corresponding GWAS and EWAS
- Gene/pathway overlap between all GWAS (N>5000) and the other EWAS

Physical overlap

- Divide genome into 500kb regions and exclude regions that aren't tagged by 450k probes



Overlap of genes/pathways

- Map EWAS signal and GWAS signal to nearest genes
- Map these genes to GO terms, KEGG terms and linked proteins from Stringdb

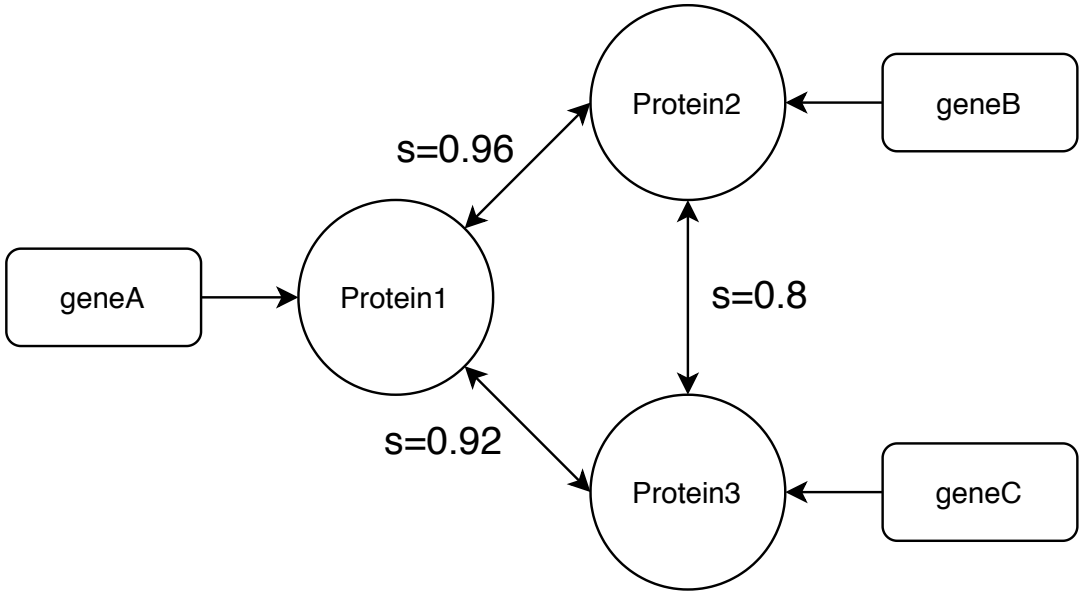
- Use some statistical test to assess if there is more overlap of the EWAS and GWAS genes/pathways than expected by chance

Quick note on stringdb

Stringdb is a database that uses various bits of evidence to assess whether proteins interact. The evidence includes evidence from other databases, text mining, co-expression experiments. Combines the evidence into a single score that they say defines how confident one can be that two proteins interact. The score ranges between 0 and 1 and from their website:

- low confidence - 0.15 (or better)
- medium confidence - 0.4
- high confidence - 0.7
- highest confidence - 0.9

If we define linked proteins as proteins with a score, s, greater than 0.9



Genes linked with protein1 = A,B,C

Genes linked with protein2 = A,B

Genes linked with protein3 = A,C

Testing best method (1)

- Aim: To test a couple of methods to see if either has power to detect gene/pathway overlap when it's present

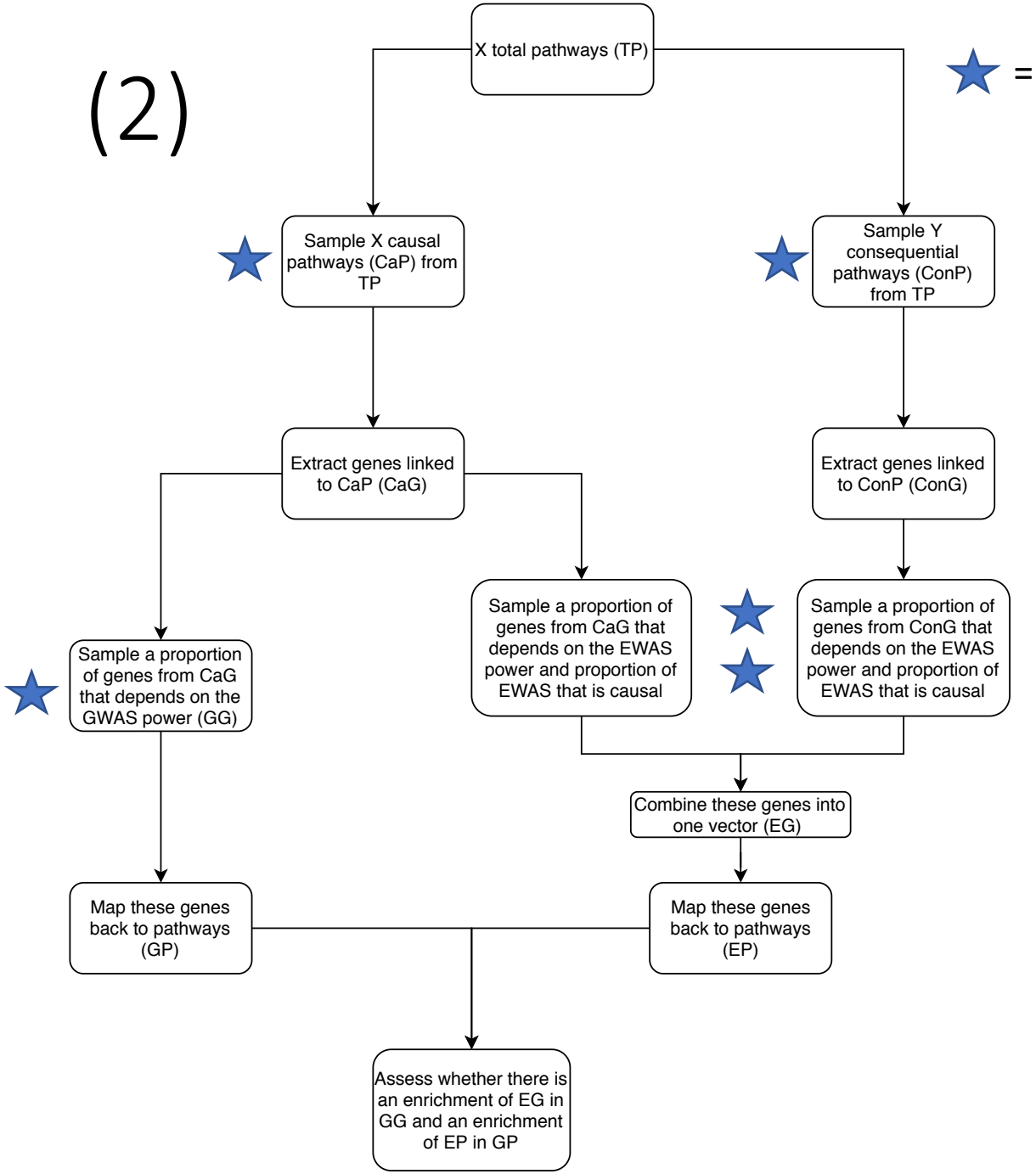
- Method 1: Use Fisher's exact test to assess if there is more overlap between all the genes and pathways

	In GWAS	Not in GWAS
In EWAS	3	50
Not in EWAS	30	200

- Method 2: Use Fisher's exact test to assess enrichment of pathways then assess correlation between "enrichment scores"

	In GWAS	Not in GWAS
In Pathway	3	50
Not in Pathway	30	200

(2)



★ = parameter that varies in simulations

Other simulation starts with genes – gene up approach

Expect overlap when:

- Proportion of EWAS hits that are causal is high
- EWAS and GWAS power is high

- Complete overlap when EWAS or GWAS power is 1 and EWAS hits are all causal

Assess overlap

- Use method that works best according to simulations
- Run permutations to generate a null distribution to compare the actual results too

Assess overlap 2

- Gene/pathway overlap between all GWAS (N>5000) and the other EWAS
- QC of GWAS data:
  - Sample size > 5000
  - N_{cases} > 10% of sample and $N_{controls}$ > 10% of sample
  - European population
  - No qtl data
- Duplicated traits used as positive control

Summary of results section

- Physical overlap between corresponding EWAS/GWAS
- Gene/pathway overlap between corresponding EWAS/GWAS
  - Simulations to test best method + power to detect overlap
  - Empirical EWAS/GWAS overlap
- Gene/pathway overlap between large group of selected GWAS and the EWAS used

Limitations

- Not a great number of traits and each trait will be unique → So can't generalise results
- Mapping signal to genes may not be done in the best way
- Databases used won't be "complete" so there a null result may just be a result of the fact that the true overlapping pathways have just not been annotated correctly
- Simulations may not be capturing the scenarios present in the empirical data
- May not have power to detect overlap (unclear what the actual scenarios are)