LDSC

# Environment

## Anaconda

Initialize conda:

source /home1/08005/cz5959/anaconda3/bin/activate

conda init

Activate conda and ldsc env:

conda activate /scratch1/08005/cz5959/ldsc/env

source activate ldsc

# Features

## LD Scores

* For European GWAS, no need to compute own LD scores and can instead use the pre-computed LD scores
  + This will create a eur\_w\_ld\_chr/ directory
  + Computed using 1000 Genomes European data

wget https://data.broadinstitute.org/alkesgroup/LDSCORE/eur\_w\_ld\_chr.tar.bz2

tar -jxvf eur\_w\_ld\_chr.tar.bz2

* Currently, bulik recommends using those same LD scores for both –w-ld-chr and –ref-ld-chr flags for non-partitioned ldsc

## Heritability and Genetic Correlation

<https://github.com/bulik/ldsc/wiki/Heritability-and-Genetic-Correlation>

#### Reformatting Summary Statistics

* Convert your results to .sumstats format using munge\_sumstats.py
* Requires following information: rsid; effect allele; non-effect allele; sample size; p-value; signed summary statistic (ex. beta)
* –merge-alleles makes sure that the alleles in your results match the alleles listed in the data used to estimate LD scores
* Check log file for errors with: grep ‘WARNING’ \*log



* Reformat Neale Lab to use munge\_sumstats.py with
  + Switch signs of betas is A1 != ALT; remove NA



#### Estimating Heritability and LDSC Intercept

* LDSC – get heritability and intercept



## Partitioned Heritability

Cahoy: isolates the cell types [astrocytes, neurons, oligodendrites] and extracts the mRNA and use Affymetric GeneChip Arrays to determine expression levels. They then use Significance Analysis of Microarrays (SAM) to find significant genes by cell type.

GTEx: isolates genotypes by tissue, annotate variants based on expression levels in specific cell type

Cell Type: annotations from 1000G, using Ensembl for annotation. Annotations based on histone marks

# Results

#### Heritability and Intercept of Single Phenotype

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| GWAS | | My LDSC | | | LDSC Neale | | Actual Neale | | |
| Phenotype | **Sex** | **Heritability** | **H2 SE** | **Intercept** | **Heritability** | **Intercept** | **Heritability** | **H2 SE** | **Intercept** |
| Height | Both | 0.419 | 0.0188 | 1.6768 | 0.423 | 0.485 | 0.485 | 0.0218 | 1.313 |
| Female | 0.4298 | 0.0196 | 1.3777 |  |  |  |  |  |
| Male | 0.434 | 0.0179 | 1.309 |  |  |  |  |  |
| Testosterone | Both | 0.0815 | 0.0102 | 1.0731 | 0.0889 | 1.0464 | 0.0771 | 0.0075 | 1.062 |
| Female | 0.0506 | 0.0053 | 1.0385 |  |  |  |  |  |
| Male | 0.166 | 0.0204 | 1.0725 |  |  |  |  |  |
| BMI | Both | 0.2275 | 0.0073 | 1.2145 | 0.2276 | 1.1172 | 0.248 | 0.0084 | 1.105 |
| Female | 0.2422 | 0.0081 | 1.0961 |  |  |  |  |  |
| Male | 0.2563 | 0.0099 | 1.0774 |  |  |  |  |  |
| IGF-1 | Both | 0.1987 | 0.0127 | 1.2771 | 0.201 | 1.1818 | 0.253 | 0.0185 | 1.125 |
| Female | 0.2032 | 0.0151 | 1.1672 |  |  |  |  |  |
| Male | 0.2382 | 0.0163 | 1.1193 |  |  |  |  |  |
| Total  Bilirubin | Both | 0.0786 | 0.0299 | 1.0891 | 0.0835 | 1.036 | 0.543 | 0.461 | 0.77 |
| Female | 0.0944 | 0.0405 | 1.056 |  |  |  |  |  |
| Male | 0.0942 | 0.0381 | 1.0458 |  |  |  |  |  |
| Creatinine | Both | 0.0954 | 0.0054 | 1.1313 | 0.1022 | 1.0757 | 0.211 | 0.0126 | 1.093 |
| Female | 0.1264 | 0.0074 | 1.0819 |  |  |  |  |  |
| Male | 0.0909 | 0.0066 | 1.0629 |  |  |  |  |  |
| RBC Count | Both | 0.1839 | 0.0129 | 1.3098 |  |  | 0.234 | 0.0233 | 1.156 |
| Female | 0.2026 | 0.0154 | 1.1803 |  |  |  |  |  |
| Male | 0.1831 | 0.0145 | 1.1162 |  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| GWAS | | My LDSC | | | Neale | | |
| Phenotype | **Sex** | **Heritability** | **H2 SE** | **Intercept** | **Heritability** | **H2 SE** | **Intercept** |
| Weight | Both | 0.2353 | 0.008 | 1.2583 | 0.266 | 0.0098 | 1.116 |
| Female | 0.2472 | 0.0089 | 1.1058 |  |  |  |
| Male | 0.2593 | 0.0103 | 1.1217 |  |  |  |
| Calcium | Both | 0.1031 | 0.0097 | 1.1499 | 0.135 | 0.0131 | 1.097 |
| Female | 0.1074 | 0.0137 | 1.10887 |  |  |  |
| Male | 0.1203 | 0.015 | 1.0655 |  |  |  |
| Total Protein | Both | 0.1281 | 0.0083 | 1.2189 | 0.167 | 0.0129 | 1.230 |
| Female | 0.1258 | 0.0081 | 1.124 |  |  |  |
| Male | 0.1355 | 0.0104 | 1.0958 |  |  |  |
| Urea | Both | 0.0930 | 0.0056 | 1.1072 | 0.119 | 0.0081 | 1.080 |
| Female | 0.1029 | 0.0063 | 1.0426 |  |  |  |
| Male | 0.0913 | 0.0075 | 1.0656 |  |  |  |
| SHBG | Both | 0.142 | 0.016 | 1.1572 | 0.230 | 0.0421 | 1.245 |
| Female | 0.1603 | 0.0188 | 1.1013 |  |  |  |
| Male | 0.2087 | 0.0289 | 1.0753 |  |  |  |
| Whole Body Fat Mass | Both | 0.2163 | 0.0068 | 1.2081 | 0.239 | 0.0080 | 1.096 |
| Female | 0.2351 | 0.008 | 1.0998 |  |  |  |
| Male | 0.2353 | 0.0087 | 1.0766 |  |  |  |
| FVC (best measure) | Both | 0.1935 | 0.0073 | 1.2027 | 0.238 | 0.0093 | 1.147 |
| Female | 0.2068 | 0.0088 | 1.1201 |  |  |  |
| Male | 0.2099 | 0.0103 | 1.0877 |  |  |  |
| HbA1c | Both | 0.1066 | 0.0069 | 1.1622 |  |  |  |
| Female | 0.112 | 0.0079 | 1.0833 |  |  |  |
| Male | 0.1159 | 0.0081 | 1.0766 |  |  |  |

#### Correlation Between Male and Female

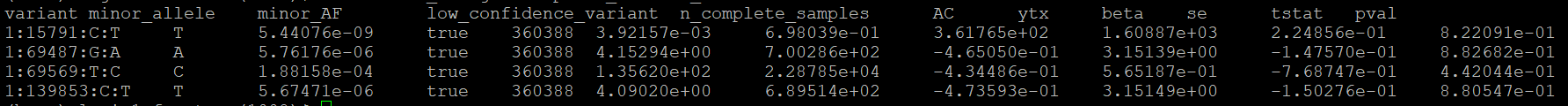
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Correlation | | | | | |
| Phenotype | **Genetic** | **Genetic SE** | **Baseline Enrichments** | **Cell Type Enrichments** | **Cahoy** | **GTEx Brain** |
| Height | 0.9723 | 0.0077 | 0.979 | 0.981 | 0.941 | 0.975 |
| Testosterone | 0.0156 | 0.0476 | 0.805 | 0.812 | 0.965 | 0.746 |
| BMI | 0.9272 | 0.0118 | 0.970 | 0.952 | 0.998 | 0.963 |
| IGF-1 | 0.8919 | 0.0178 | 0.942 | 0.982 | 0.968 | 0.973 |
| Total Bilirubin | 0.9815 | 0.033 | 0.848 | 0.989 | 0.977 | 0.942 |
| Creatinine | 0.897 | 0.0244 | 0.867 | 0.996 | 0.858 | 0.949 |
| RBC Count | 0.951 | 0.0166 | 0.977 | 0.997 | 0.953 | 0.981 |

|  |  |  |
| --- | --- | --- |
|  | Correlation | |
| Phenotype | **Genetic** | **Genetic SE** |
| Weight | 0.9383 | 0.0118 |
| Calcium | 0.9668 | 0.0238 |
| Total Protein | 0.954 | 0.023 |
| Urea | 0.9598 | 0.0266 |
| SHBG | 0.8354 | 0.0222 |
| Whole Body Fat Mass | 0.9191 | 0.0124 |
| FVC (best measure) | 0.9767 | 0.0175 |
| HbA1c | 0.9471 | 0.0242 |

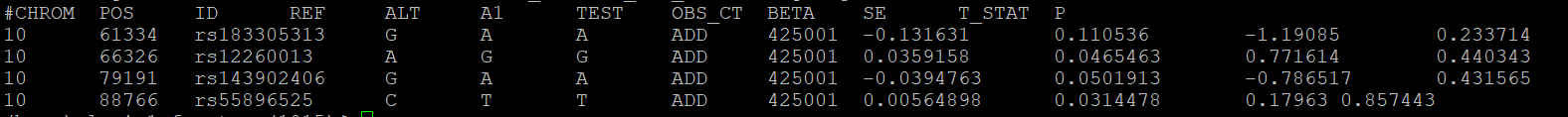
# Log

### 6/10/2021

* Set up environment by downloading anaconda for linux <https://github.com/bulik/ldsc#readme>
* Neale Lab Header



* + Neale doesn’t have rsid included; I can try to match with plink2 results
* Plink2 Header



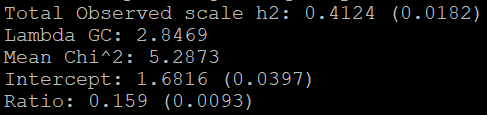
### 6/11/2021

* Download LD scores and alleles used to compute LD scores
* Create sumstats file format from results



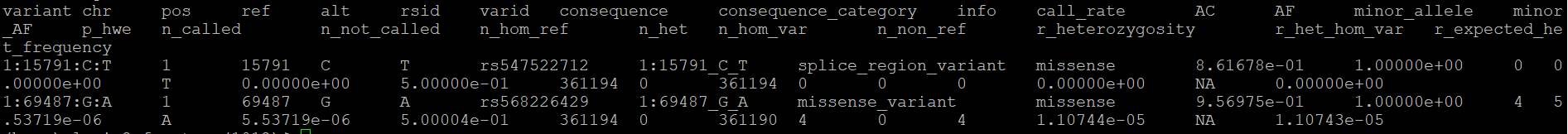
### 6/14/2021

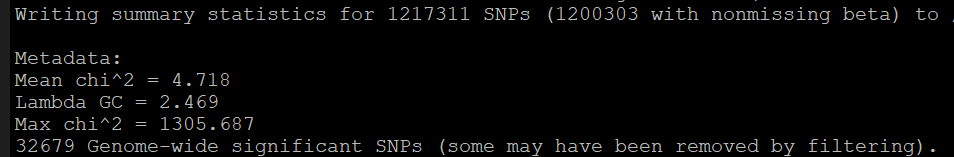
* Calculate heritability and ldsc intercept for height
* Error: IndexError while reading LD Scores
  + Forgot to add the “/” after eur\_w\_ld\_chr
* H2 and intercept for height
  + What is lamda and ratio
  + Lambda: percent variance in dependent variable (height) not explained by difference differences in levels of independent variable (genotype)
  + Ratio: Ratio is (intercept-1)/(mean(chi^2)-1), which measures the proportion of the inflation in the mean chi^2 that the LD Score regression intercept ascribes to causes other than polygenic heritability.



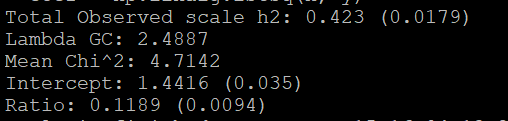
### 6/15/2021

* Format Neale lab to be suitable for munge\_py
  + Download variants file





* H2 and intercept for Height from Neale Results

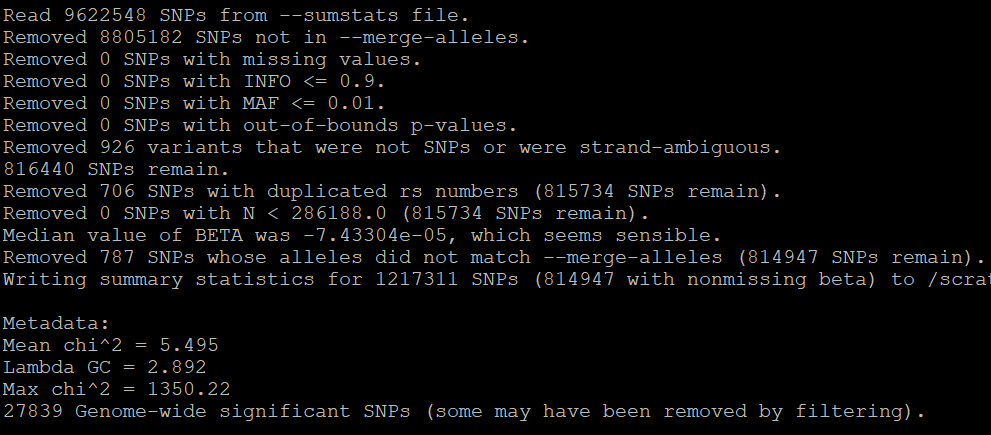


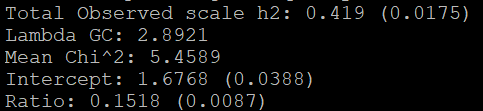
### 6/16/2021

* GC bias: genomic region of a higher GC content tends to have more (or less) Illumina reads covering that region
* Intercept, mean chi square when tagging 0 casual SNPs
* GC bias
  + Recombination: get errors one sis as GC; other sis has AT; should both have GC
    - Heteroduplex gets repaired, repairs more GC because stronger bond
  + Will see more GC in areas of higher recombination rates
* Neale Lab for height: heritability = 0.485 ; intercept = 1.313

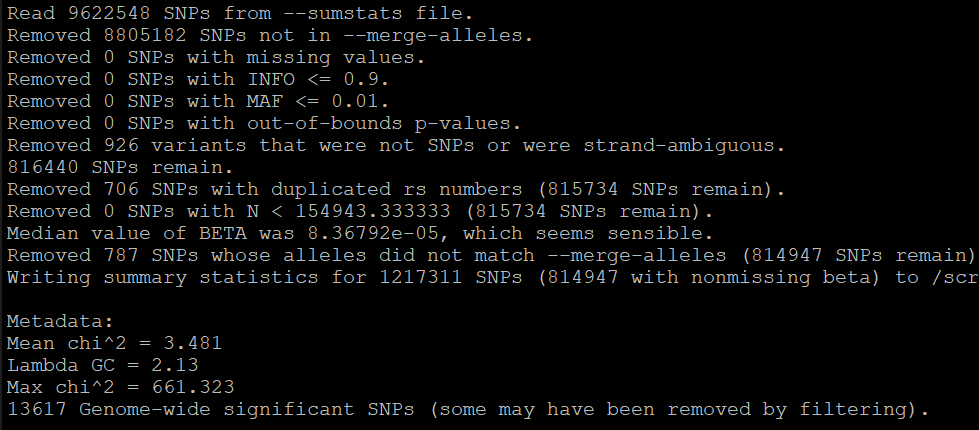
6/17/2021

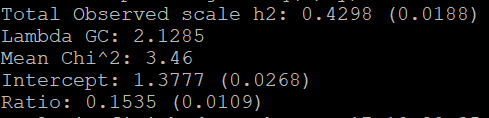
* Updates height (AX as non-effect), both sex



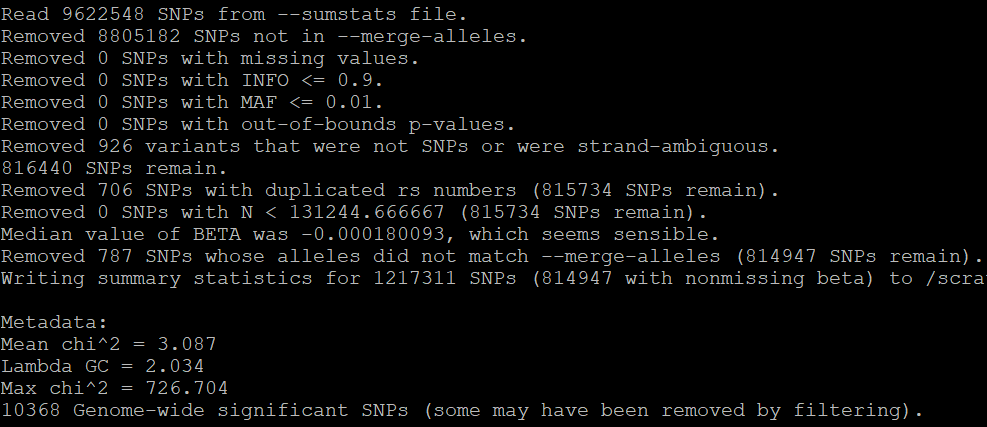


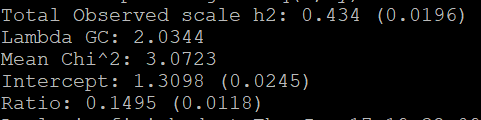
* Female





* Male



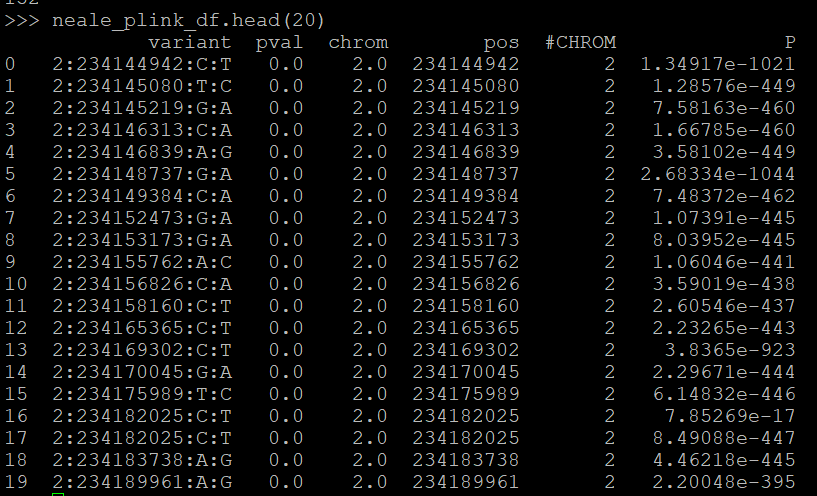


### 6/21/2021

* Get ldsc h2 and intercept for testosterone and BMI
* Write code for partitioned heritability

### 6/24/2021

* Couldn’t run code on total bilirubin since the results file contained p-values that were too small to be converted to numeric



* Create a format\_plink script to get rid of rows with p-values that can’t be converted
  + Reformat for bilirubin and RBC
* Get ldsc for Neale Lab IGF1 to examine my results and their results
  + Reformat script for Neale Lab Manhattan plot and format\_neale
  + My plink and my neale more similar than actual neale
  + They could be using a different code?
* Bar chart for partitioned heritability
  + Enrichment = proportion heritability / proportion SNPs

### 6/30/2021

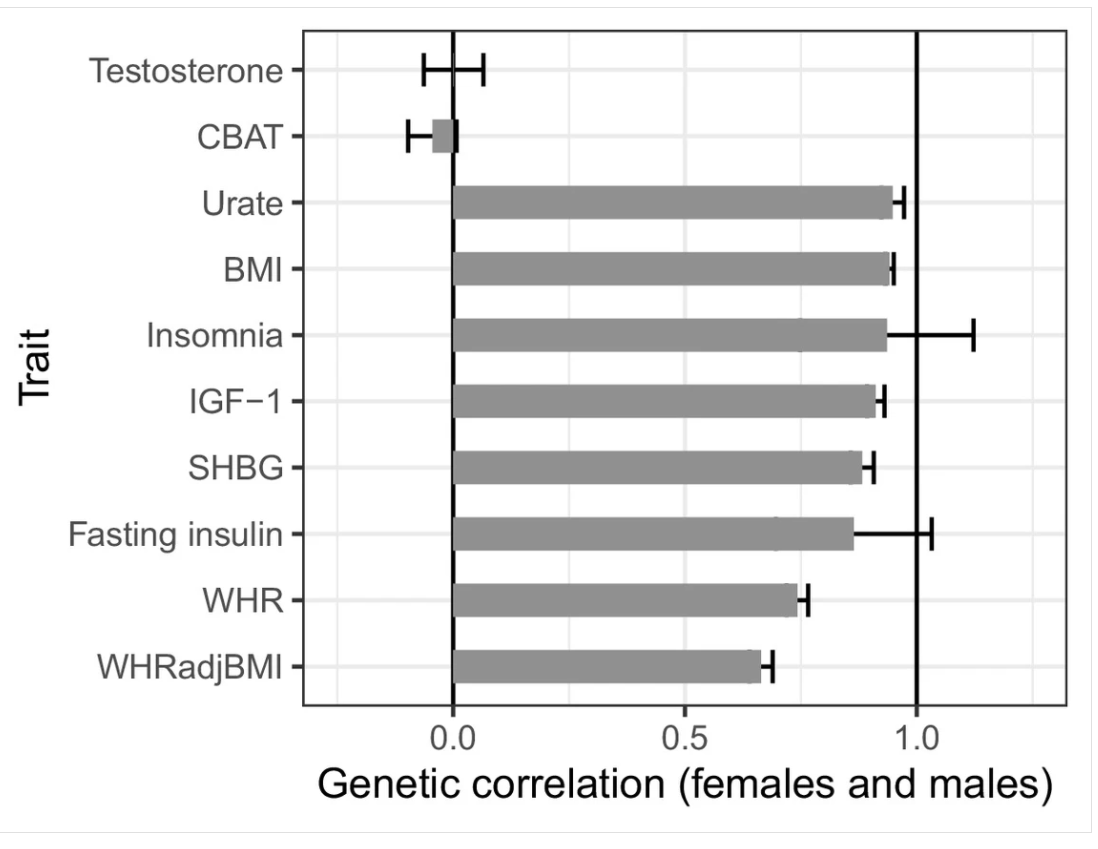
* Negative heritability~~: individuals with similar genotypes likely to have more divergent trait values than those with differing genotypes~~  estimated heritability is less than expected
* Enrichment: estimated share of heritability the category contributes divided by its expected share

Baseline Functional Categories

* Coding region: base sequence identical to mRNA transcript produced
* Conserved: identical/sim sequences across species
* CTCF: CCCTC-binding factor; highly conserved zinc finger protein, transcription factor
* DGF-1: dispersed gene family 1
* DHS: DNase I hypersensitive sites; regions of chromatin that are sensitive to cleavage by DNase I enzyme
* Enhancer: promote transcription
* FetalDHS
* H3k27ac: epigenetic modification to histone H3, acetylation, active enhancer mark
* H3k27me3: epigenetic modification to histone H3, methylation, downregulation
* TFBS: transcription factor binding sites
* TSS: transcription start site
* UTR 3’ and 5’: untranslated region on each side of mRNA

Testosterone Correlation

* Mine: 0.0156 sex-specific GWAS paper: 0.120
* GWAS of 3 traits:



### 7/1/2021

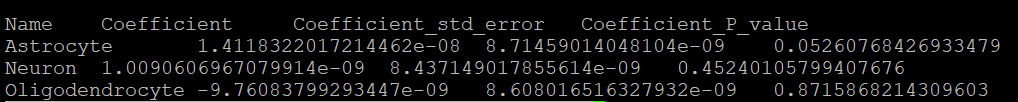
List of Cell Type Groups

* Adrenal\_Pancreas; Cardiovascular; CNS; Connective\_Bone; GI; Immune; Kidney; Liver; SkeletalMuscle; Other
* Code for cell type partitioned heritability

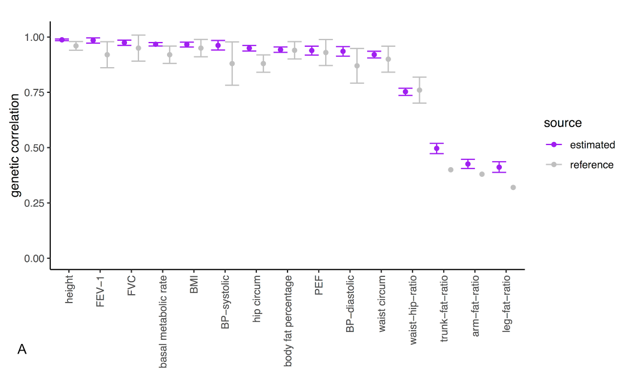
### 7/5/2021

* SEMM calculation of correlation seem to be higher than reference?

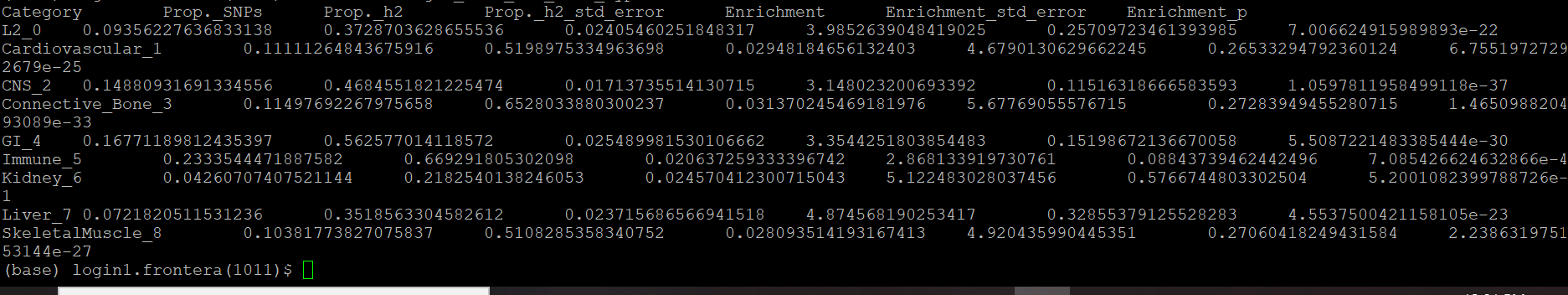
### 7/7/2021



* Cell type partitioned heritability
  + Move cell type LD scores to ldsc folder
  + Coefficient: estimate of first regression coefficient



* Cell type partition results



* Correlation among cell type and baseline annotations
* Create plots for cell-type with error bars

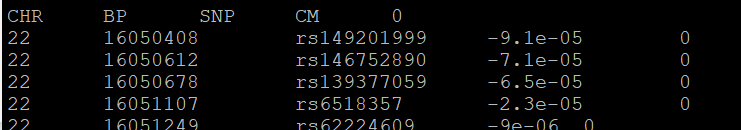
### 7/8/2021

Annotation Sources

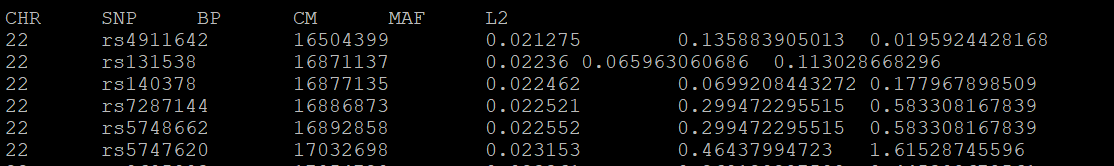
* Cahoy – transcriptome database
  + <https://www.jneurosci.org/content/28/1/264>
  + Isolate and purify astrocytes, neurons, and oligodendrocytes from mice forebrain
  + Extract mRNA and used Affymetrix GeneChip Arrays to determine expression levels and create transcription database
    - Filter by consistent expression by cell type: intensity level >200 in at least 2/3 samples
    - Significance Analysis of Microarrays used to determine genes significantly differentially expressed between cell types
  + Affymetrix GeneChip Array: measure expression for specific mRNA (microarray)
* GTEx Brain: Genotype Tissue Expression
  + <https://gtexportal.org/home/documentationPage>
  + <https://www.nature.com/articles/nature24277>
  + Genotype donors to assess genetic variation
  + Analyze Global RNA expression within individual tissues
  + eQTLs: variations in gene expression highly correlated with genetic variation
  + functional enrichment: annotated variants using chromatin state predictions
    - segment genome to determine state (enhancer, promoter, etc)
  + expression quantification based on GENECODE annotations

Annotation Files

* Cell Type – h2
  + .l2.M\_5\_50
    - Number for each annotation
    - Number of SNPs in annotation with MAF > 5%
  + .annot.gz
    - CM: centimorgans, genetic position
    - 0: annotations, binary

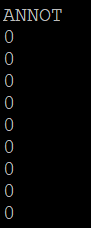


* + .l2.ldscore
    - L2: ld scores



Ch22 adrenal: 129365 annot, 17193 ldq:q!

* Cahoy – h2-cts
  + Annot



Chr 22 control: 141124 annot, 17490 ld

Ch22 3: 141124 annot, 17490 ld

* Issue: why I couldn’t use Cahoy and files with –h2: <https://github.com/bulik/ldsc/issues/103>
  + Mismatch between set of variants in LD score files and frqfile
  + Downloaded
    - 1000G\_Phase3\_frq/
    - 1000G\_Phase3\_frq/1000G.EUR.QC.9.frq

### 7/13/2021

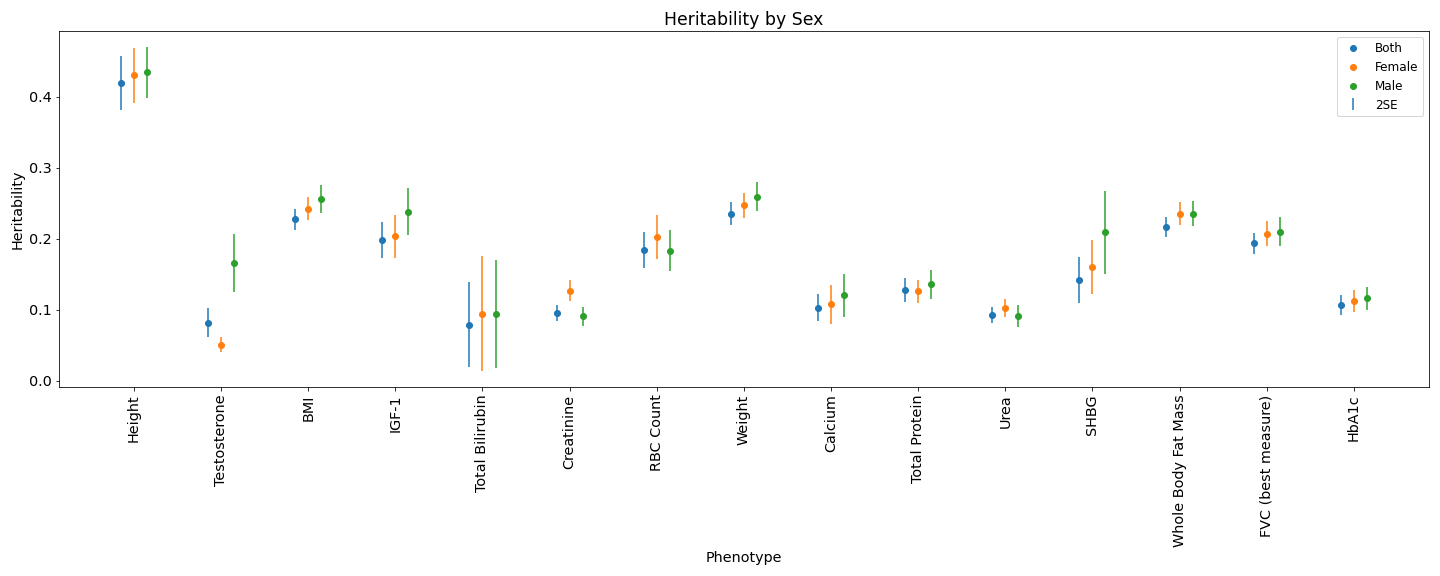
* Created partitioned bar plots for Cahoy and GTEx brain for first set of phenotypes

### 7/14/2021

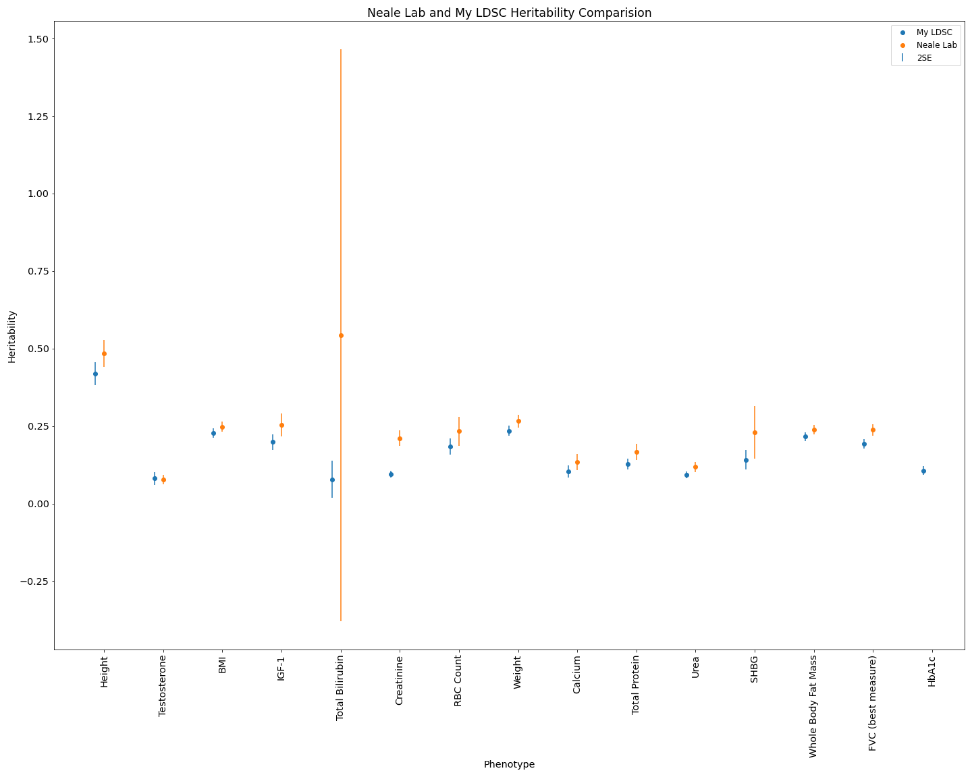
* Get ldsc results for second set of phenotypes

### 7/15/2021

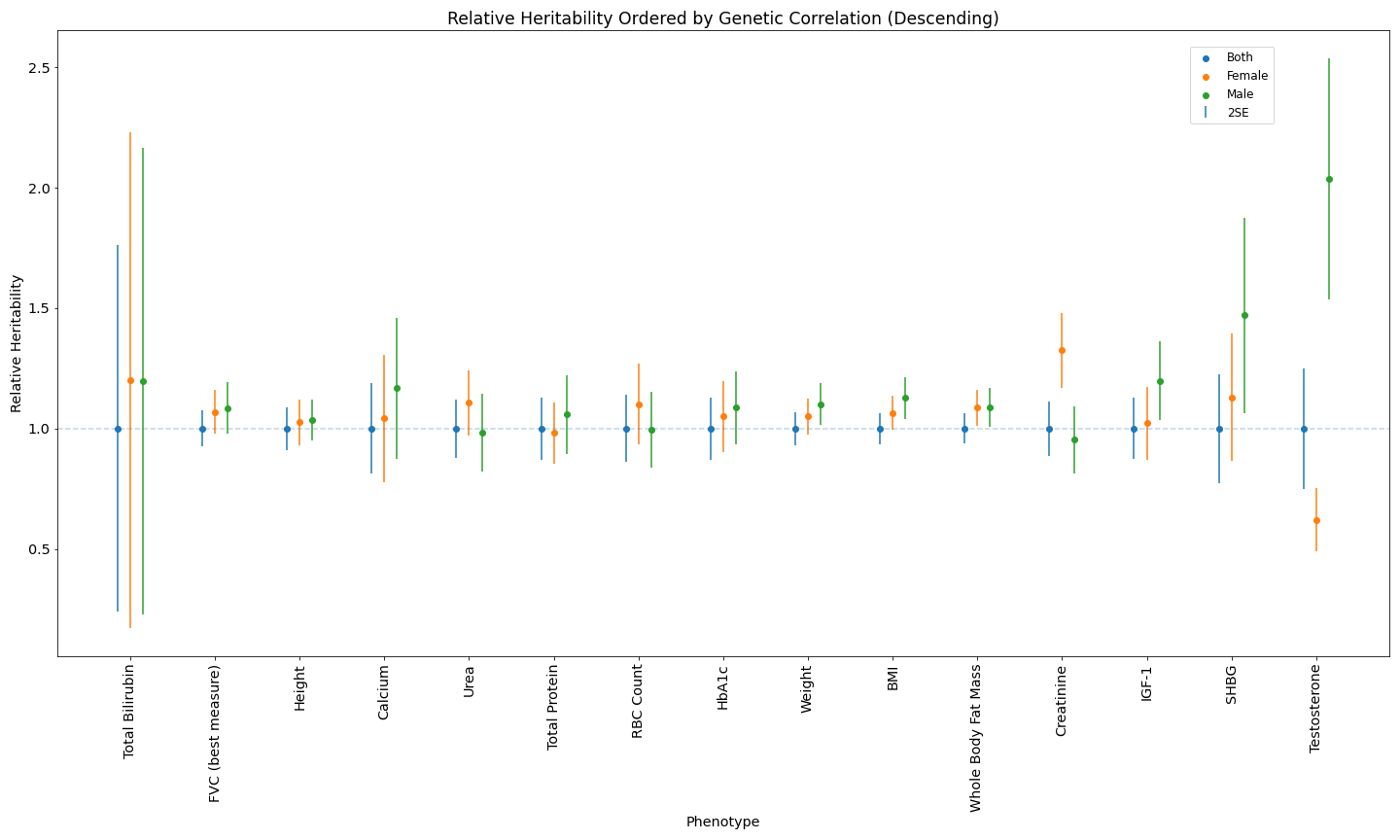
* Create scatterplot for heritabilities



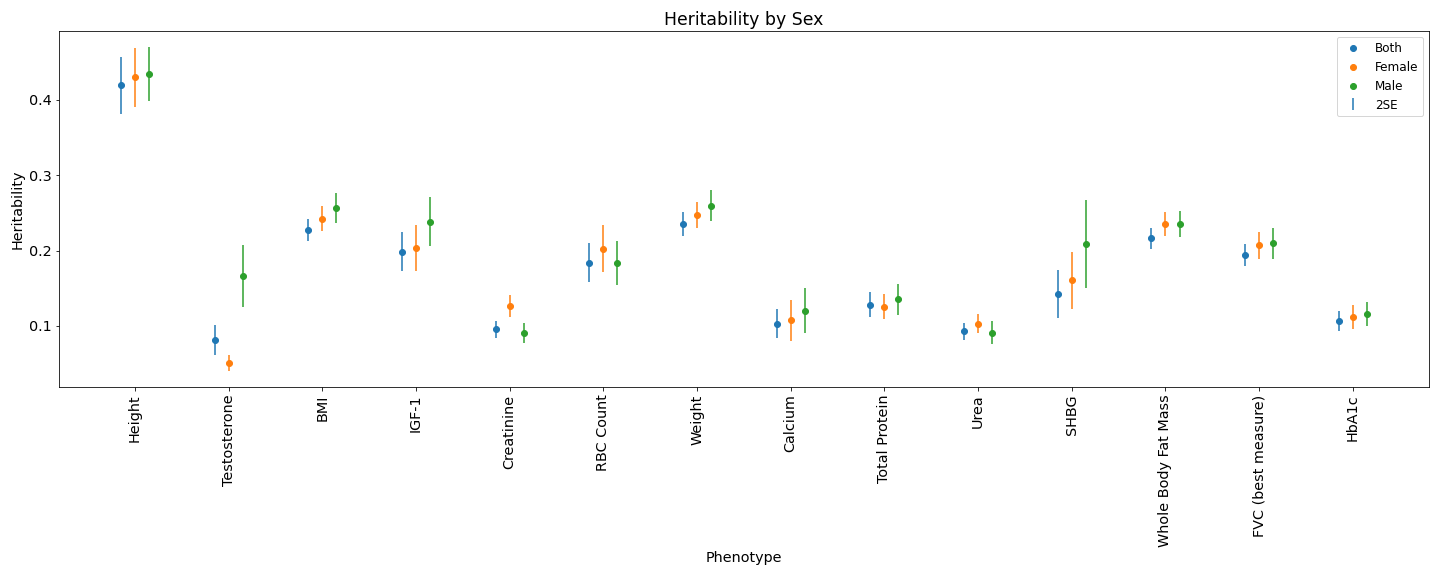
* Create scatterplot comparing with Neale heritabilities
  + Rank normalized probably contributing to why Neale heritability is consistently greater than ours; otherwise small problem since difference is small
    - Pushes estimates to normal distribution with mean 0
  + Get rid of total bilirubin since the confidence rating is so low

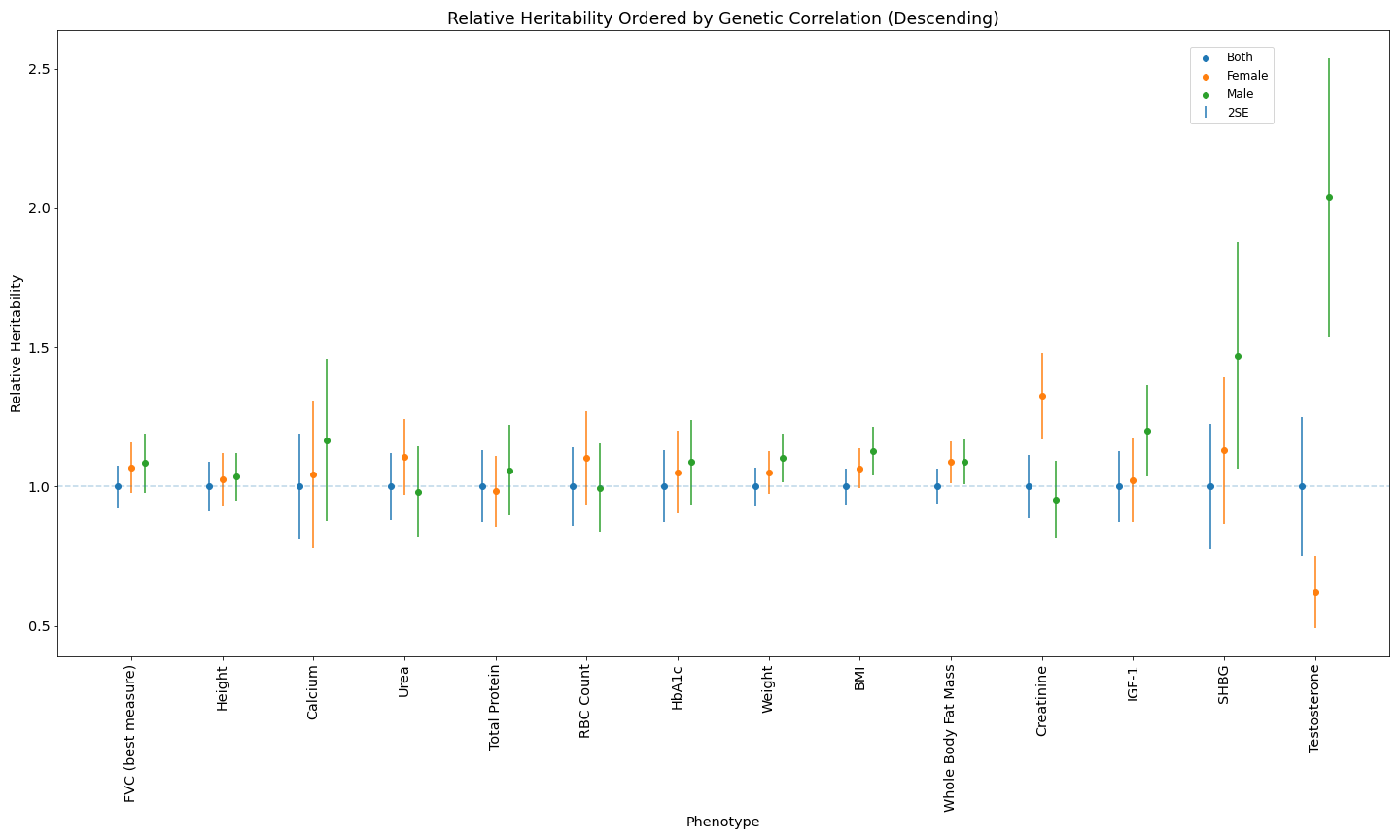


* Create scatterplot for relative heritabilities, ordered by genetic correlation (descending)
  + Relative heritability = h2 / {h2 of both}
  + SE = ( (h2+SE) / {h2 of both} ) – h2



* No bilirubin plots





7/27/2021

* ENTEX partition bar plot
  + Control: H3K27ac.union\_0 H3K27ac.sum\_0 H3K36me3.union\_0 H3K36me3.sum\_0 H3K4me1.union\_0 H3K4me1.sum\_0 H3K4me3.union\_0 H3K4me3.sum\_0

# Code

### LDSC and Partitioned LDSC





### Format Neale



### Format Plink



### Bar Chart – Partitioned Heritability



### Heritability Plots – Jupyter Notebook





