**Pleiotropy Simulation-Analysis Review and Methods Project**

**Background Readings:**

* MixFisher Paper:
  + Zhonghua Liu, Xihong Lin. Multiple phenotype association tests using summary statistics in genome-wide association studies. Biometrics. 2018 Mar;74(1):165-175. <https://pubmed.ncbi.nlm.nih.gov/28653391/>
  + <https://content.sph.harvard.edu/xlin/software.html#mpat>
* Intro-overview Paper:
  + Hanna Julienne, Pierre Lechat, Vincent Guillemot, Carla Lasry, Chunzi Yao, Robinson Araud, Vincent Laville, Bjarni Vilhjalmsson, Hervé Ménager, Hugues Aschard. JASS: command line and web interface for the joint analysis of GWAS results. NAR Genom Bioinform. 2020 Mar;2(1):lqaa003. <https://pubmed.ncbi.nlm.nih.gov/32002517/>
* LDSC Genetic Correlation Paper:
  + Brendan Bulik-Sullivan et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015 Nov;47(11):1236-41. <https://pubmed.ncbi.nlm.nih.gov/26414676/>
  + <http://ldsc.broadinstitute.org/>
* Paper of Potential Framework for combining PX methods:
  + Yaowu Liu, Sixing Chen, Zilin Li, Alanna C Morrison, Eric Boerwinkle, Xihong Lin. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. Am J Hum Genet. 2019 Mar 7;104(3):410-421. <https://pubmed.ncbi.nlm.nih.gov/30849328/>
  + <https://content.sph.harvard.edu/xlin/software.html#acat>
* Background Px Paper
  + Need to Add

**Key Questions:**

* Many PX methods –
  + What is the best method?
    - Goal is to select a universal best method/framework
    - Combining methods with different assumptions may improve modeling the of heterogeneity of pleiotropy genetic architecture between traits
  + Potential Methods: MPAT: WALD, mixFisher, VC, SUM, Others??
* Unified Test (Combining tests into a unified test) – ACAT Method (Xihong Lin)
* Meta-analysis (combining data from different sources) & Power Analyses Estimation
* Inflation due to sample overlap (e.g., Z% overlap in samples for GWAS for Trait A and GWAS for Trait B)
* Estimating and accounting for phenotype correlation: LDSC vs. correlation (Z-scores)
* How to deal with inflation in pleiotropy results (FDR?) –
  + Trait A (Beta=0.3, P=3E-4) and Trait B (Beta=0.1, P=8E-3) => PX Pval=4xE-8
    - Combined P-value improved vs. individual component trait p-value – would argue this is pleiotropy, rather than driven by Trait A or B specific signals
  + Trait A (Beta=0.3, P=2E-14) and Trait B (Beta=0.1, P=1E-1) => PX Pval=8E-14
    - Combined P-value is GWAS significant but does not improve on individual component trait p-value – would argue this is not pleiotropy, rather captures Trait A signal
* Digging into PX signals (Full vs. Partial PX)
  + Is the pleiotropy signal for all 5 traits (full) or a subset (partial)?
  + How to asses/identify and distinguish between partial vs. full pleiotropy
* Selecting phenotypes combinations for PX analyses – tough question
  + Pairwise Correlation that is problematic (R > 0.8 or different threshold)
  + Estimate the number of effective independent phenotypes

**Datasets:**

1. **Simulated Phenotypes (**Model on UKBB, large scale)
   1. User control of data/associations – use to test/develop, since we know the truth
   2. Sample Size: 400K, 200K, 100K, 50K, 25K
   3. Phenotypes & Pleiotropy Modeling
      1. Various levels of Heritability and pleiotropy between traits
      2. Pleiotropy with weakly correlated phenotypes
      3. Pleiotropy with strongly correlated phenotypes
      4. Heterogenous/homogenous effects, etc.
2. **UKBB**
   1. Real data – realistic situation to assess utility/performance of methods
   2. ~375K unrelated European Ancestry samples
   3. Split into sets of: 375K, 200K, 100K, 50K, 25K
   4. Phenotypes:
      1. Lipids
      2. Anthropometrics
      3. Metabolic Syndrome

**Scenario 1**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Scenario | Phenotype  (Pi – Pj) | Pi H2 | Pi N Loci | Pj H2 | Pj N Loci | N PXij Loci | Pi – Pj Correlation | Genetic  Correlation | Env.  Correlation |
| 1 | P1 – P2 | 0.5 | 90 | 0.5 | 90 | 10 | 0.05 | 0.05 | 0 |
|  |  |  |  |  |  |  |  |  |  |
| 2 | P1 – P2 | 0.5 | 90 | 0.5 | 90 | 10 | 0.10 | 0.10 | 0 |
|  |  |  |  |  |  |  |  |  |  |
| 3 | P1 – P2 | 0.5 | 90 | 0.5 | 90 | 5 | 0.25 | 0.25 | 0 |
|  |  |  |  |  |  |  |  |  |  |
| 4 | P1 – P2 | 0.5 | 90 | 0.5 | 90 | 5 | 0.20 | 0.10 | 0.10 |
|  |  |  |  |  |  |  |  |  |  |
| 5 | P1 – P2 | 0.5 | 90 | 0.5 | 90 | 5 | 0.10 | 0.00 | 0.10 |
|  |  |  |  |  |  |  |  |  |  |
| 6 | P1 – P2 | 0.5 | 90 | 0.5 | 90 | 5 | 0.30 | 0.05 | 0.25 |
|  |  |  |  |  |  |  |  |  |  |
| 7 | P1 – P2 | 0.5 | 90 | 0.5 | 90 | 10 | 0.10 | 0.10 | 0 |
| 7 | P1 – P3 | 0.5 | 90 | 0.5 | 90 | 5 | 0.05 | 0.05 | 0 |
| 7 | P2 – P3 | 0.5 | 90 | 0.5 | 90 | 5 | 0.05 | 0.05 | 0 |

Pi = GU + GS + EU + ES + err

**GU** = Unique Genetic

**GS** = Shared Genetics (Pleiotropy, and could be partitioned into sub-components to reflect shared genetics across multiple phenotypes)

**EU** = Unique Environment

**ES** = Shared Environment (this could be partitioned into sub-components to reflect shared environment across multiple phenotypes)

**err** = error/noise parameter

corr(Pi, Pj) = composed of GS + ES (need to update)

**Simulations Details-**

**Residual (e) simulation:**

Define the overall variance of the phenotype simulated as:

Model assumes that there is no G\*E interaction and therefore the and further assume the error term very small (>> than or )

where is the variance of the genetically explained portion of the variance of the phenotype, is the portion of phenotypic variance explained by the environment, and is the variance of the error term.

**Broad Sense Heritability** (includes additive and non-additive genetic components) of a phenotype:

**Narrow sense the heritability** (exclusively additive genetic component) of a phenotype:

[Need to Update]

Best software to generate simulated GWAS (vs doing in R)

Order of scales:

-work with ~5-20 pleiotropic variants out of ~100 for testing/modeling/development

- Then do GWAS Scale (~100K variants, rough estimate of number of genome wide haplotype blocks in genome) to assess/test for discovery of false positive

Also allows comprehensive assessment of impact of allele frequency on methods

-how to deal with Sex Chromosomes (XX and XY)

* Phenotypes
* P1 = Genetics1 (50 SNPs) + Env1 + Error
* P2 = Genetics2 (50 SNPs) + Env2 + Error
  + Select 10 SNPs that overlap G1 & G2 -> pleiotropy SNPs
* Variance(P1) = variance(G1) + variance(E1) + covariance (G1\*E1)
* Variance(P2) = variance(G2) + var(E2) + covariance (G2\*E2)
* Heritability (H2) = var(G2)/total var (P2)