AdjHE: An efficient way to estimate heritability

Christian Coffman

Advised by: Dr. Saonli Basu
University of Minnesota Twin Cities
Division of Biostatistics

January 9, 2023

1. **General trait influencers**

2. **Methods for detecting role of genetics**

3. **Simulations**

4. **Estimation on brain region volumes**

General influences on traits

Traits are determined by different contributions of genetics and environmental influencers.
Which traits are dictated by which set of influencers?

Image credit:

https://blogs.kcl.ac.uk/editlab/2019/05/07/if-something-is-genetic-it-can-still-be-influenced-by-the-environment/

GWAS

- Genome Wide Association studies (GWAS)

$$Y' = X_c\alpha + X_g\beta + \epsilon$$

- X_g: genotype
- X_c: other covariates
- Inference done on the β (sometimes millions)
- Pro: Great for highly influential SNP's
- Low: Low power for causality spread across multiple SNP's

Gene effects as random

- Consider $\beta \sim N(0, \sigma_g^2 I)$
- Then $X_g\beta \sim N(0, \sigma_g^2 X_g X_g')$
- Redefined as $N(0, \sigma_G^2 A)$
- Where A is called the **Genetic Relatedness Matrix**
- Model becomes

$$Y' = X_c\alpha + \epsilon, \epsilon \sim (0, \sigma_G^2 A + \sigma_e^2 I)$$

GRM based heritability estimation

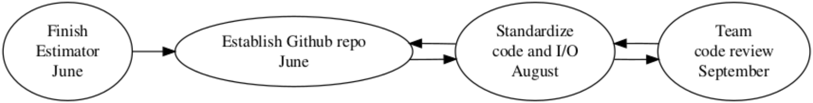
- Describe variation in phenotype as random effect (LMM)

$$Y' = X_c\alpha + \epsilon, \epsilon \sim N(0, \sigma_g^2 A + \sigma_e^2 I)$$

- Gain: power for dispersed genetic effects
- Loss: resolution on genome
- GCTA uses REML which can be slow with large studies (n x n matrix)
- Not efficient for exploration of mildly heritable traits (large sample sizes)
- Can solve via MOM, but what about when there is population substructure?

New tool: AdjHE

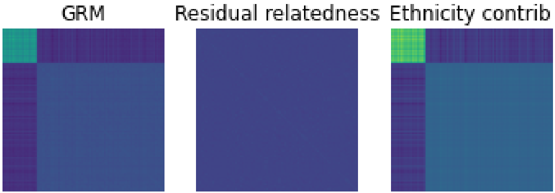
- Two-stage Method of Moments approach
- Accounts for ethnicity as PC's of GRM
- Key assumption: $X_c \perp PC's$
- Closed form $\therefore$ Much more efficient
- Benchmarked: 2x faster with 4000 subjects
- 10x faster with 45k subjects



Dealing with population substructure

- Relatedness has family relations A and ethnicity G

$$GRM = A + G$$



- PCA on GRM

$$GRM = \sum \lambda_i VV' = \sum \lambda_i A_i A_i' + \sum \lambda_i G_i G_i'$$

- Suppose G_i contribute more to variance
- First k PC's define G

Dealing with covariates

- Treat PC's as covariates ($X = [X_c, X_{pc}]$)
- Project away covariates ("Residualize")

$$Q = I - X(X'X)^{-1}X'$$

$$QY' = Y = QX_c + Q\epsilon = Q\epsilon$$

$$EY = 0$$

- 2nd moment

$$EYY' = Var(Y) = QVar(\epsilon)Q$$

$$= \sigma_G^2 A + \sum \delta_i G_i G_i' + \sigma_e^2 I$$

- Solve via OLS

Properties of OLS estimator

$$EYY' = \begin{bmatrix} A & G_1 G_1' & \vdots & G_k G_k' & I \end{bmatrix} \begin{bmatrix} \sigma_G^2 \\ \delta_1 \\ \cdots \\ \delta_k \\ \sigma_e^2 \end{bmatrix}$$

$$EYY' - G\delta = \begin{bmatrix} A & I \end{bmatrix} \begin{bmatrix} \sigma_G^2 \\ \sigma_e^2 \end{bmatrix}$$

$$\begin{bmatrix} \hat{\sigma}_G^2 \\ \hat{\sigma}_e^2 \end{bmatrix} = \begin{bmatrix} A & I \end{bmatrix} (\begin{bmatrix} trA^2 & trA \\ trA & n \end{bmatrix})^{-1} \begin{bmatrix} A \\ I \end{bmatrix} (YY' - G\hat{\delta})$$

Problem with multi-site estimation

- More studies using consortia to study smaller effects
- The Adolescent Brain Cognitive Development (ABCD) has $+10{,}000$ subjects > 20 sites
- Measures brain features
- Brain features sensitive to machine used (i.e. depends on site)
- Adding fixed effect blows up SE (coming up in a few slides)
- So treat it as random effect

$$Y \sim N(X_c\alpha, \sigma_G^2 A + \sum G_i\delta_i + S\sigma_s^2 + I\sigma_e^2)$$

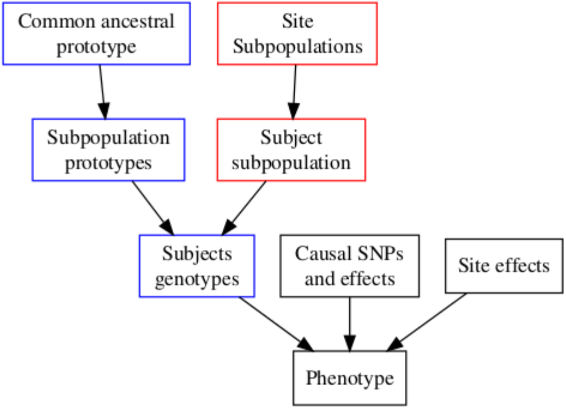
AdjHE site extension

- Assume $X_c \perp X_s, A, G$

$$EYY' - G\Delta G' = \begin{bmatrix} A & QSQ & I \end{bmatrix} \begin{bmatrix} \sigma_G^2 \\ \sigma_s^2 \\ \sigma_e^2 \end{bmatrix}$$

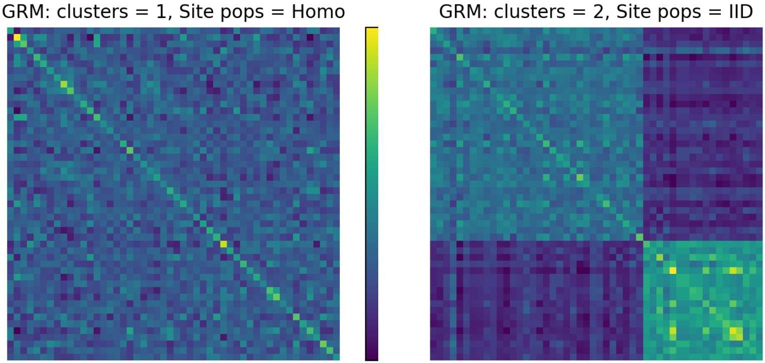
- X_s is site vector
- S is site similarity matrix $X_s X_s'$

Simulation tool



- Simulate realistically structured GRM's and phenotypes
- Determine what scenarios fit within AdjHE model

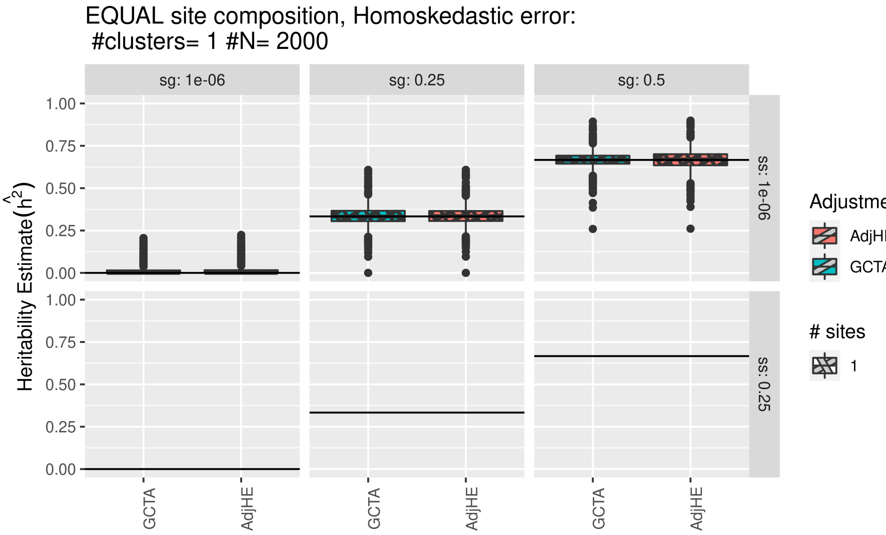
Simulating population structures



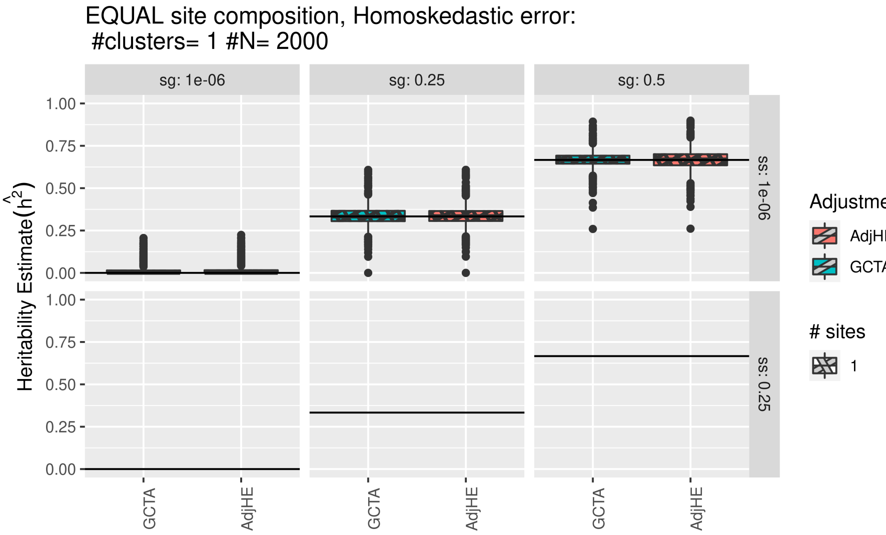
GRM: clusters = 1, Site pops = Homo

GRM: clusters = 2, Site pops = IID

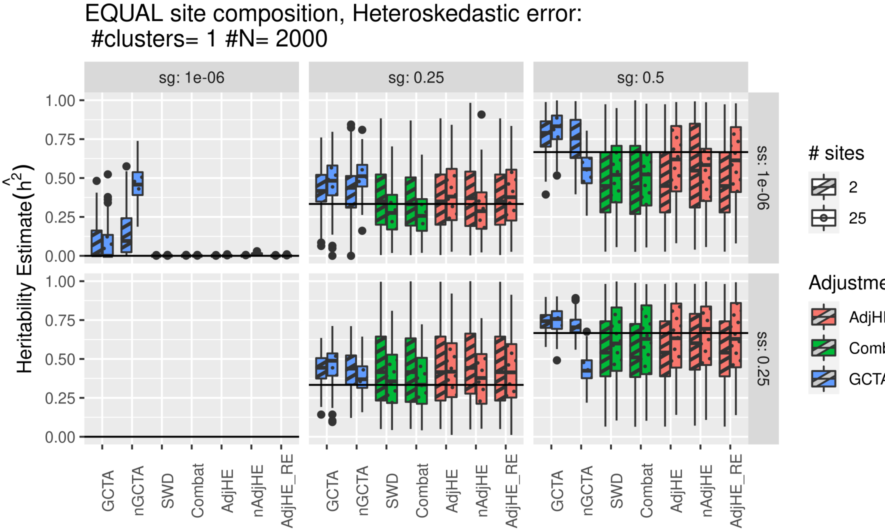
Estimation on Homogeneous



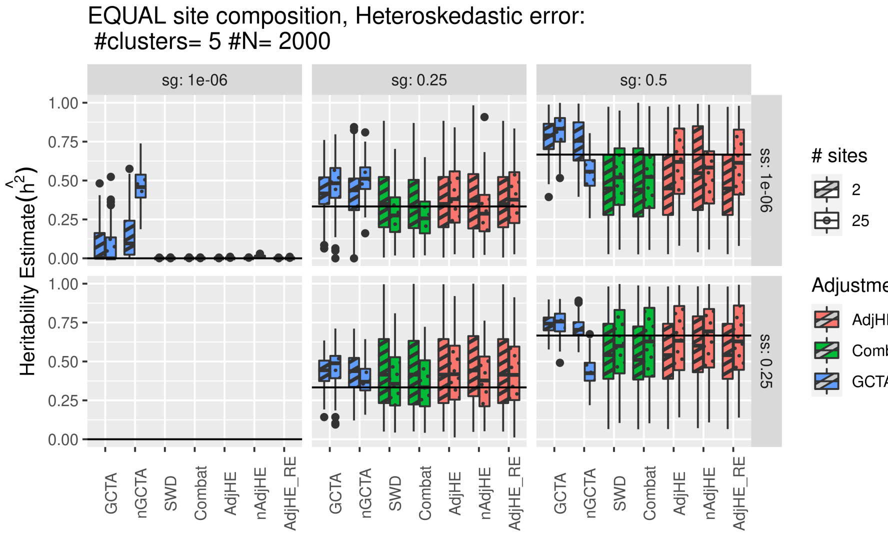
Estimation on Homogeneous



Estimation on Homogeneous



Multiple clusters heterogeneous

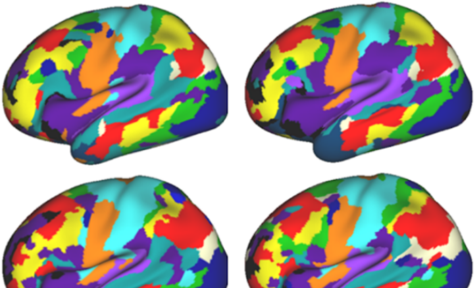


Naive estimates on Asegs data

Controlling for Site AdjHE

Conclusions and future aims

- AdjHE is efficient estimator and accounts for basic effect from site
- Early analysis suggests volumes in adolescent brains are heritable
- Estimates consistent with ADNI results
- Applications to functional topology
- Differing ethnicity distributions affect estimate?



Thank you for listening
Questions?

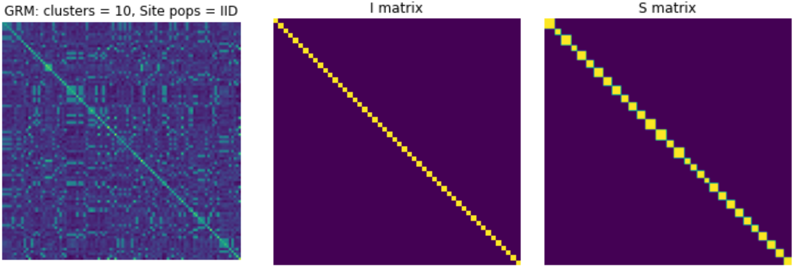
References

Lin, Seal, and Basu. “Estimating SNP Heritability in Presence of Population Substructure in Biobank-Scale Datasets.” Genetics 2022

Hermosillo et al. “A Precision Functional Atlas of Network Probabilities and Individual-Specific Network Topography.” 2022 bioRxiv

Zhao et al., 2019 “Heritability of Regional Brain Volumes in Large-Scale Neuroimaging and Genetic Studies.”

Source Identifiability problem

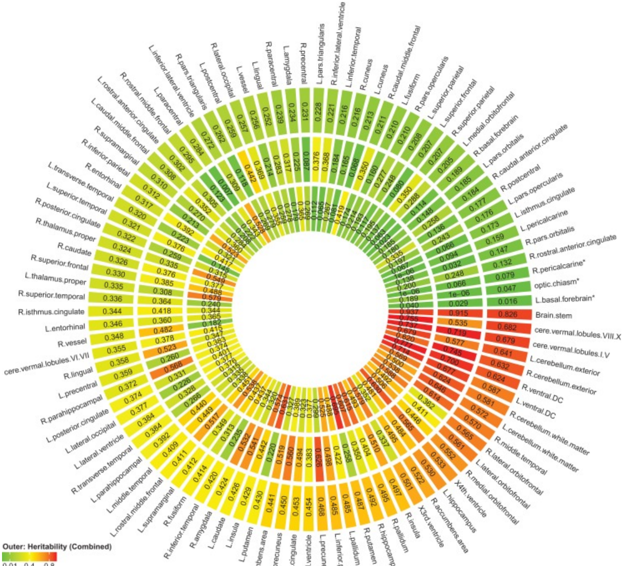


GRM: clusters = 10, Site pops = IID

I matrix

S matrix

Zhao paper results (ADNI)


L.pars.triangularis
R.inferior.lateral.ventricle
L.inferior.temporal
R.cuneus
L.cuneus
R.caudal.middle.frontal
L.fusiform
R.pars.opercularis
L.superior.parietal
L.superior.frontal
R.superior.parietal
L.medial.orbitofrontal
R.basal.forebrain
L.pars.orbitalis
R.caudal.anterior.cingulate
R.postcentral
L.pars.opercularis
L.isthmus.cingulate
L.pericalcarine
R.pars.orbitalis
R.rostral.anterior.cingulate
R.pericalcarine*
optic.chiasm*
L.basal.forebrain*
Brain.stem
cere.vermal.lobules.VIII.X
cere.vermal.lobules.I.V
L.cerebellum.exterior
R.cerebellum.exterior
R.ventral.DC
L.ventral.DC
R.cerebellum.white.matter
L.cerebellum.white.matter
R.middle.temporal
R.lateral.orbitofrontal
L.lateral.orbitofrontal
R.medial.orbitofrontal
X4th.ventricle
L.hippocampus
R.accumbens.area
X3rd.ventricle
R.insula
R.pallidum
R.hippocampus
R.putamen
L.pallidum
L.inferior.p
L.precune
ventricle
cingulate
recuneus
bens.area
L.putamen
L.insula
L.caudate
R.amygdala
R.inferior.temporal
R.fusiform
L.supramarginal
L.rostral.middle.frontal
L.middle.temporal
L.parahippocampal
R.transverse.temporal
L.lateral.ventricle
L.lateral.occipital
L.posterior.cingulate
R.parahippocampal
L.precentral
R.lingual
cere.vermal.lobules.VI.VII
R.vessel
L.entorhinal
R.isthmus.cingulate
R.superior.temporal
L.thalamus.proper
R.superior.frontal
R.caudate
R.thalamus.proper
R.posterior.cingulate
L.superior.temporal
L.transverse.temporal
R.entorhinal
R.inferior.parietal
R.supramarginal
L.caudal.middle.frontal
L.rostral.anterior.cingulate
R.rostral.middle.frontal
L.paracentral
L.inferior.lateral.ventricle
R.pars.triangularis
L.postcentral
R.lateral.occipital
L.vessel
L.lingual
R.paracentral
L.amygdala
R.precentral
Outer: Heritability (Combined)