Genome-wide estimates of heritability

Ben Domingue
Institute of Behavioral Science
University of Colorado Boulder
ben.domingue@gmail.com

- Genes $\rightarrow$ behaviors & outcomes of interest.
- Genome-wide data: FHS, HRS, AddHealth, etc....
- Hard to get a handle on genotype/phenotype connection.
  - GWAS results help, but have limited availability.
  - Even when available, polygenic scores have limited predictive value.

What else can we do?

GCTA

Genome-wide Complex Trait Analysis (GCTA) tells us about heritability.

- GCTA estimates heritability without knowledge of causal variants.
- Instead uses "genetic similarity" (similar to logic of twin studies).

Method

1. Estimate genome-wide similarity:

$$A_{jk} = \frac{1}{N} \sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}$$

Method

1. Estimate genome-wide similarity:

$$A_{jk} = \frac{1}{N} \sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}$$

2. Then estimate mixed model:

$$y = X\beta + g + \epsilon$$

where $g \sim \text{MVN}[0, \sigma_g^2 A]$.

Method

1. Estimate genome-wide similarity:

$$A_{jk} = \frac{1}{N} \sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}$$

2. Then estimate mixed model:

$$y = X\beta + g + \epsilon$$

   where $g \sim \text{MVN}[0, \sigma_g^2 A]$.

3. Heritability: $\frac{\widehat{\sigma_g^2}}{\widehat{\sigma_g^2} + \widehat{\sigma_\epsilon^2}}$.

Method

1. Estimate genome-wide similarity:

$$A_{jk} = \frac{1}{N} \sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}$$

2. Then estimate mixed model:

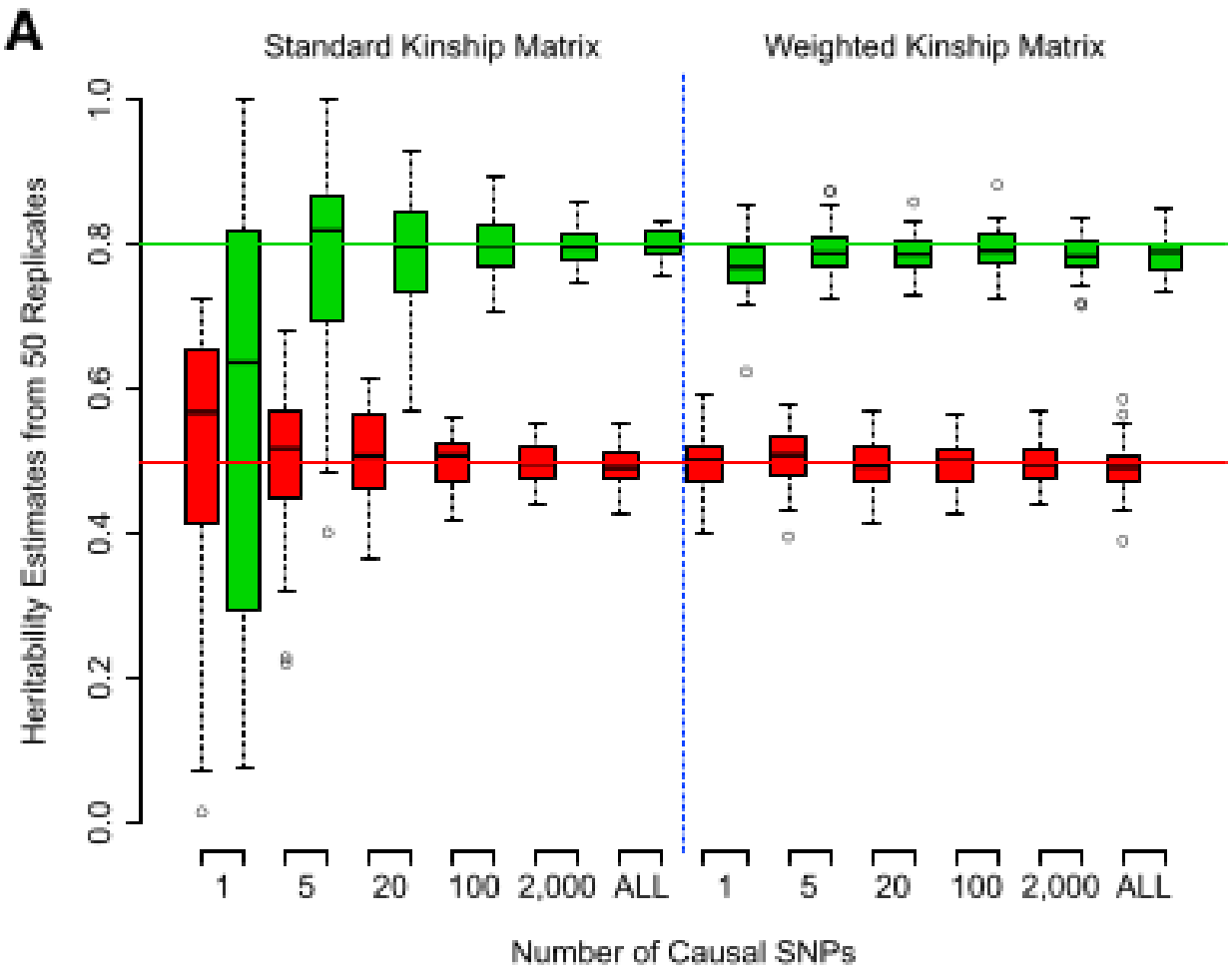
$$y = X\beta + g + \epsilon$$

   where $g \sim \text{MVN}[0, \sigma_g^2 A]$.

3. Heritability: $\frac{\widehat{\sigma_g^2}}{\widehat{\sigma_g^2} + \widehat{\sigma_\epsilon^2}}$.

Complicated model & not the DGP.

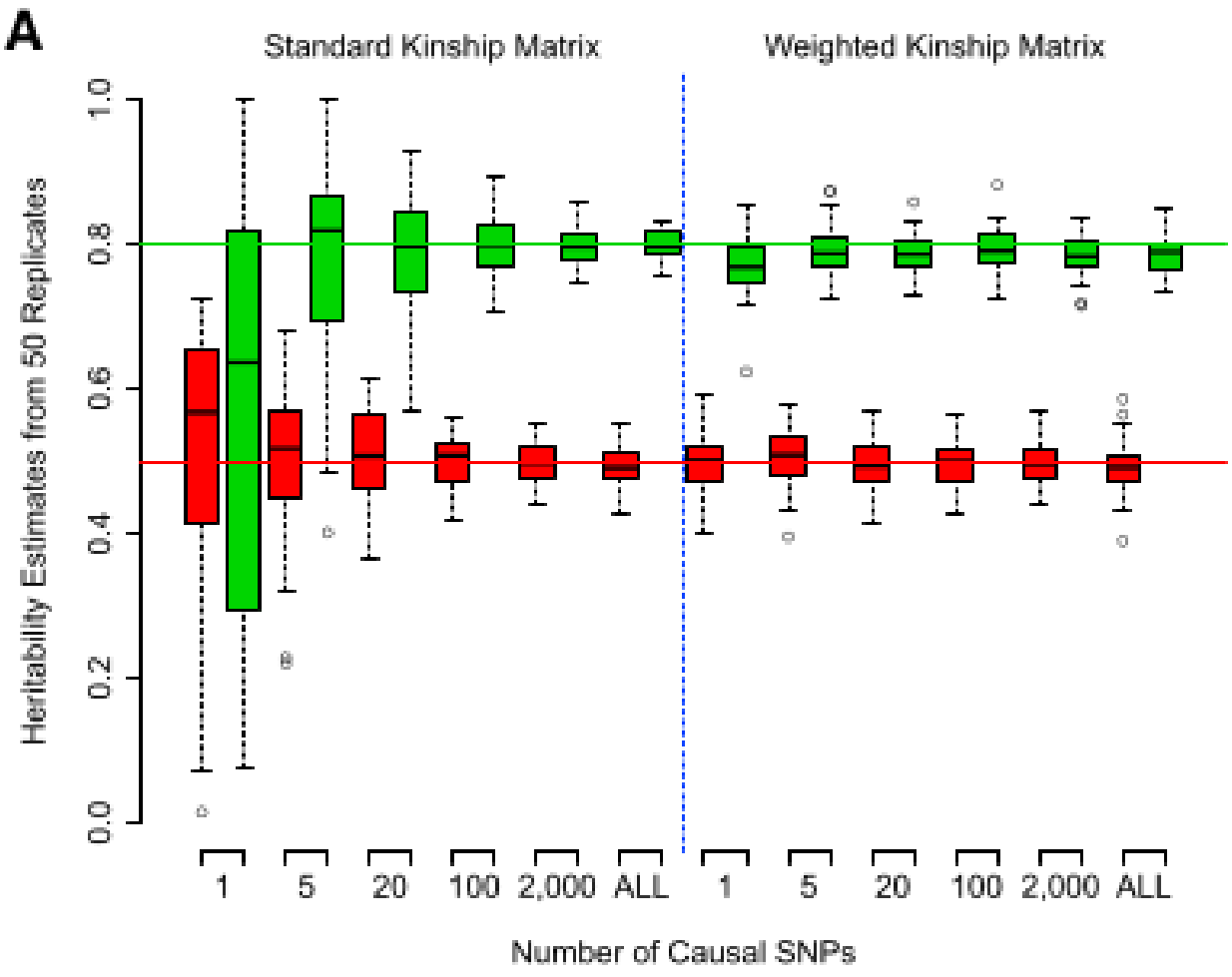
Sensitivity to genetic architecture?



- Robust to # of causal variants.

Speed et al.,2012, AJHG,

Sensitivity to genetic architecture?



Speed et al.,2012, AJHG,

- Robust to # of causal variants.
- Sensitive to LD.

Sensitivity to environment?

Could genetic similarity just be a proxy for environmental similarity?

Table 1 GREML heritability estimates for shared childhood urbanicity, height, BMI and education[a]

	h^2 No controls (2 PCs) A	h^2 Urban control (2 PCs) B	|Δ| A – B	h^2 No controls (10 PCs) C	h^2 Urban control (10 PCs) D	|Δ| C – D
Urban childhood N = 6439	0.29155 (0.0574)	NA	NA	0.14767 (0.0622)	NA	NA
Height N = 6379	0.32489 (0.0644)	0.32510 (0.0644)	0.00022 (0.0910)	0.30397 (0.0659)	0.30397 (0.0659)	0.02092 (0.0921)
BMI N = 6320	0.31300 (0.0674)	0.31323 (0.0675)	0.00023 (0.0953)	0.31300 (0.0674)	0.3190 (0.0678)	0.00596 (0.0956)
Education N = 6414	0.17493 (0.0650)	0.15217 (0.0652)	0.02276 (0.0921)	0.1749 (0.0650)	0.15939 (0.0656)	0.01554 (0.0923)

Conley et al., 2014, JHG

My goal: Offer intuition and basic guidance on when GCTA estimates may be reliable.

Data

HRS: 4950 non-Hispanic whites, $\approx$ 1.5M autosomal SNPs.

- Height: 0.40

Q1: Gen sim as function of SNPs

Correlation

Q1: Gen sim as function of SNPs

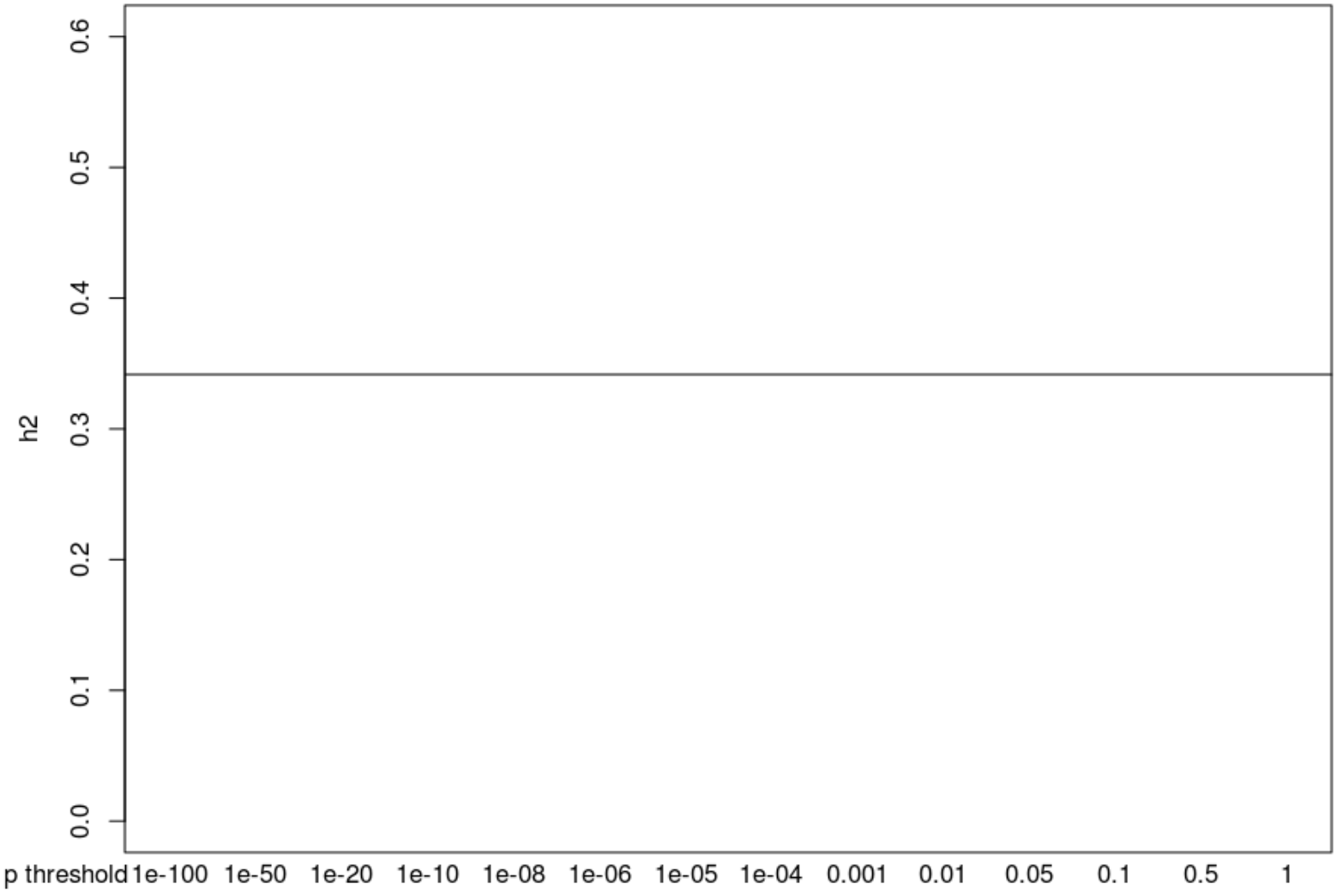
	Correlation
50% Sample	0.98
30% Sample	0.95
10% Sample	0.83

Q1: Gen sim as function of SNPs

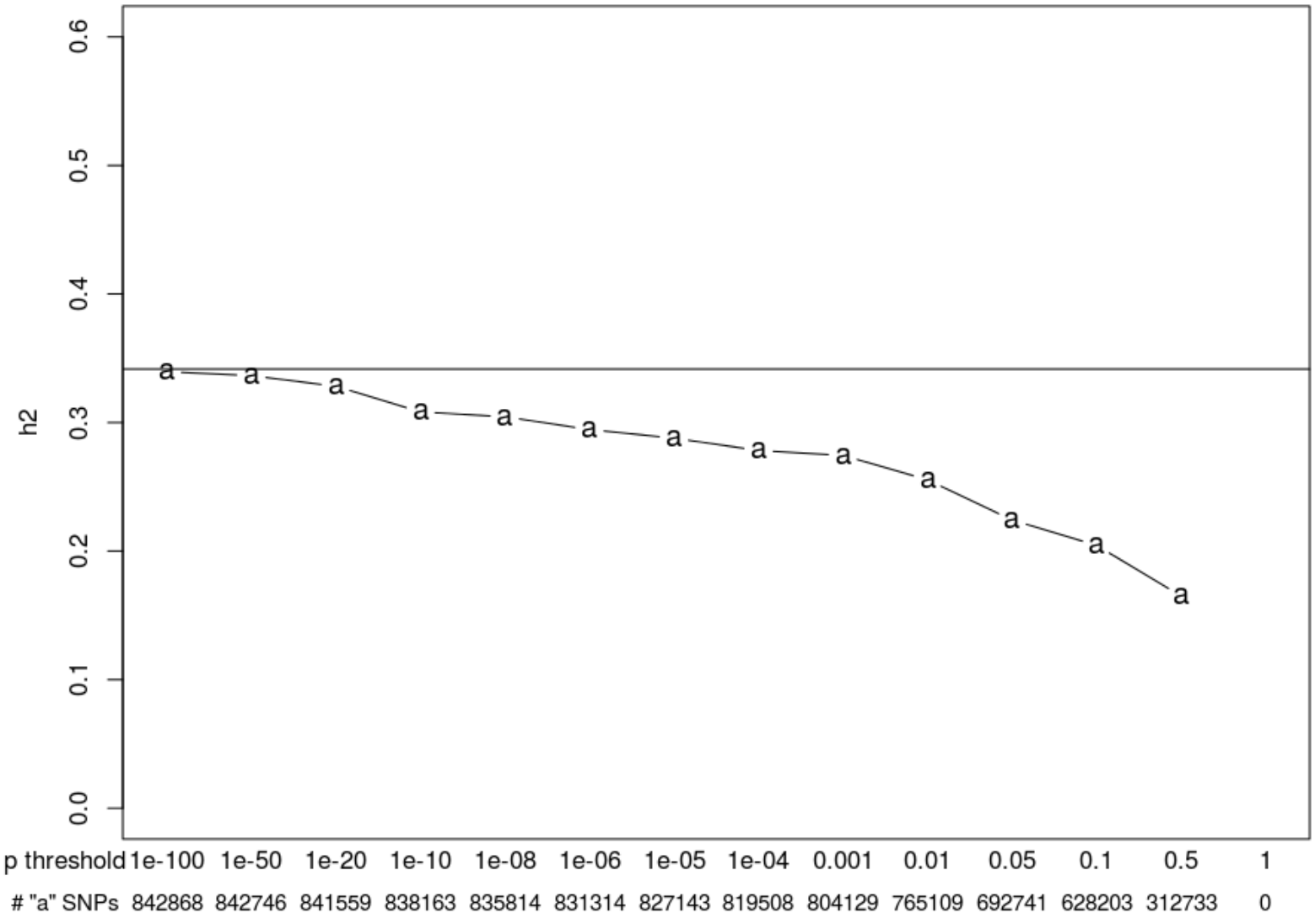
	Correlation
50% Sample	0.98
30% Sample	0.95
10% Sample	0.83
$r^2 = 0.01$	0.57
$r^2 = 0.2$	0.75
$r^2 = 0.5$	0.88

Q2: GWAS (height) variants

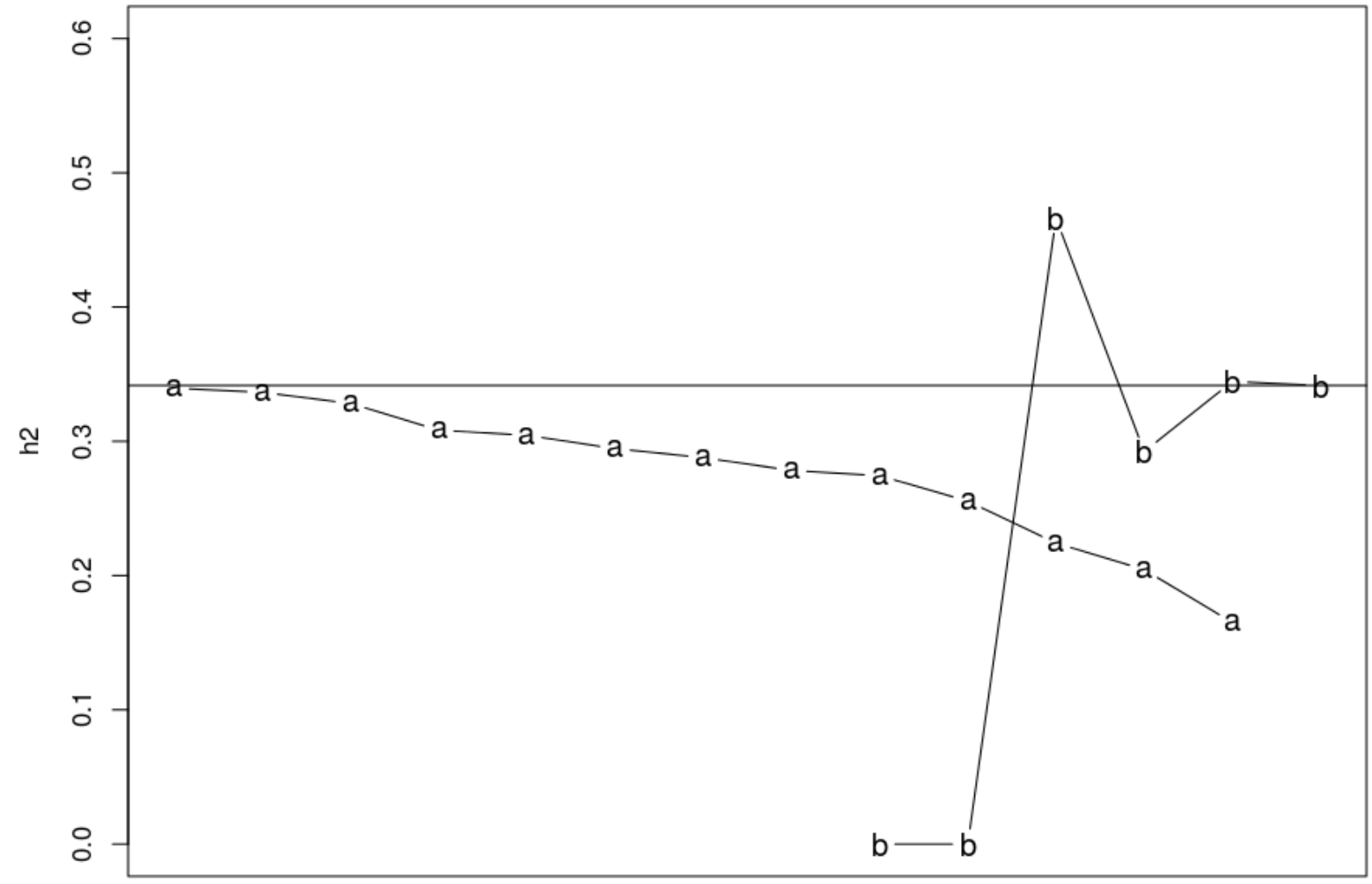
Q2: GWAS (height) variants



Q2: GWAS (height) variants



Q2: GWAS (height) variants



p threshold	1e-100	1e-50	1e-20	1e-10	1e-08	1e-06	1e-05	1e-04	0.001	0.01	0.05	0.1	0.5	1
# "a" SNPs	842868	842746	841559	838163	835814	831314	827143	819508	804129	765109	692741	628203	312733	0
# "b" SNPs	21	143	1330	4726	7075	11575	15746	23381	38760	77780	150148	214686	530156	842889

Q3: Heteroskedasticity

Heteroskedasticiy is common problem.

- weight on height.
- own education on paternal education.

Of concern here since we're estimating variance components.

Q3: Heteroskedasticity

Heteroskedasticiy is common problem.

- weight on height.
- own education on paternal education.

Of concern here since we're estimating variance components.

- Simulate outcome based on GCTA model.
- $y = 0.5 \cdot \text{height} + g + \epsilon$.
- ϵ_i has variance $\exp(\alpha \cdot \text{height} \cdot \sigma^2_\epsilon)$, where α controls level of heteroskedasticity and σ^2_ϵ controls heritability.

Q3: Heteroskedasticity

Heteroskedasticiy is common problem.

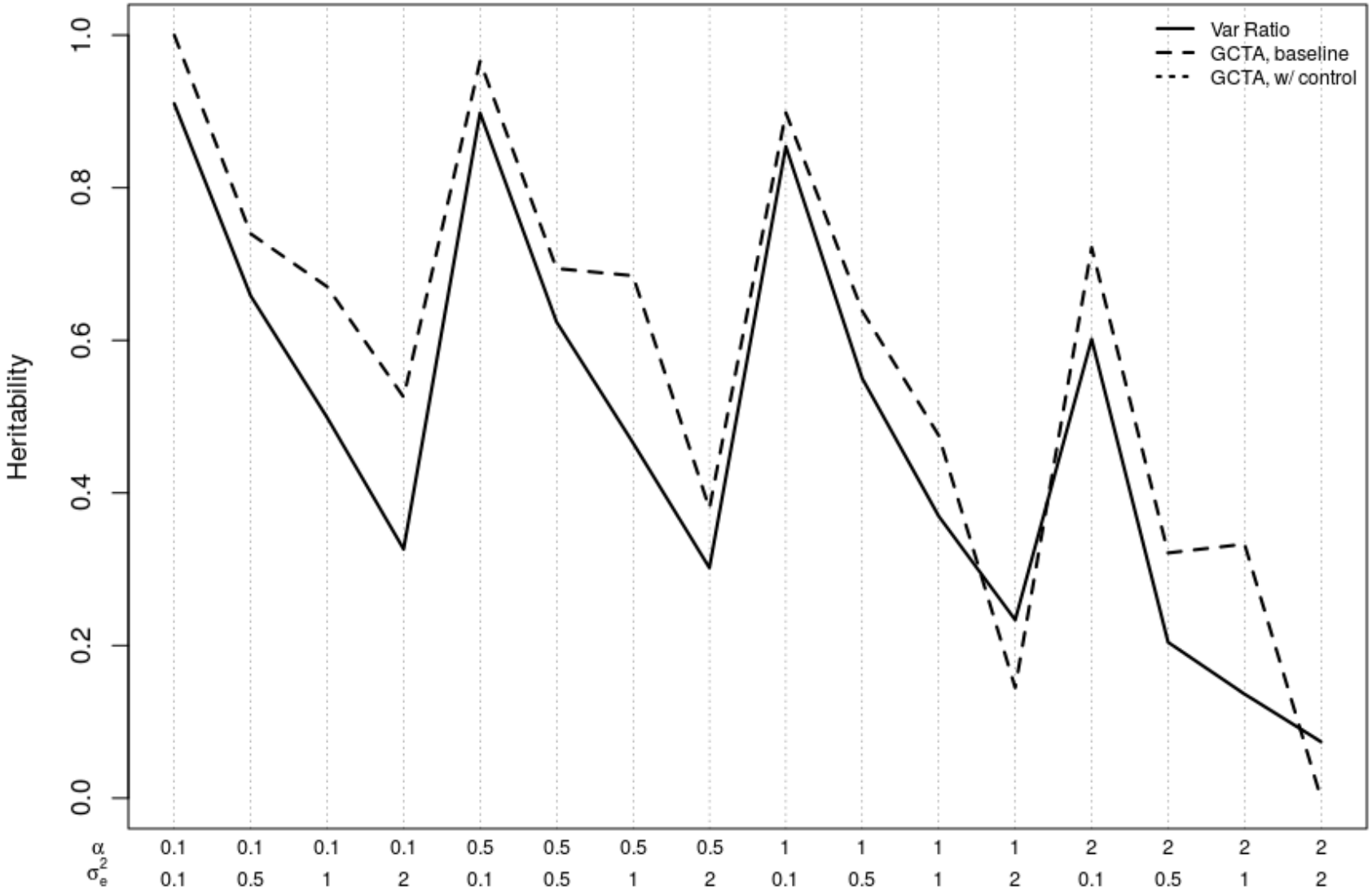
- weight on height.
- own education on paternal education.

Of concern here since we're estimating variance components.

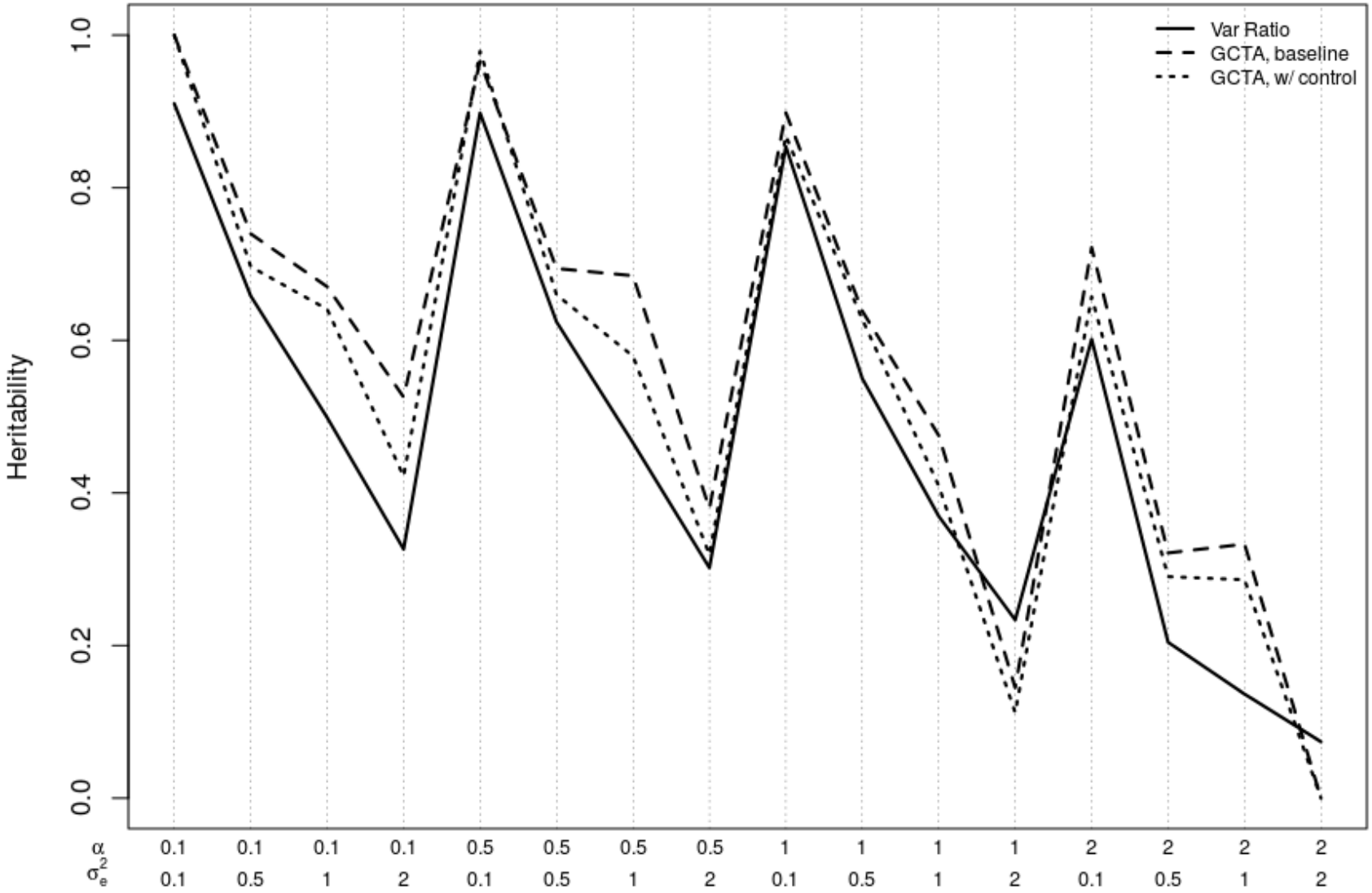
- Simulate outcome based on GCTA model.
- $y = 0.5 \cdot \text{height} + g + \epsilon$.
- ϵ_i has variance $\exp(\alpha \cdot \text{height} \cdot \sigma^2_\epsilon)$, where α controls level of heteroskedasticity and σ^2_ϵ controls heritability.

Examine recovery of heritability, but def'n no longer simple.

Q3: Heteroskedasticity



Q3: Heteroskedasticity



Q4: Environmental Moderation

Heritability not constant: What are implications for GCTA?

- Standard GCTA: $g \sim \text{MVN}[0, \sigma^2_g A]$.

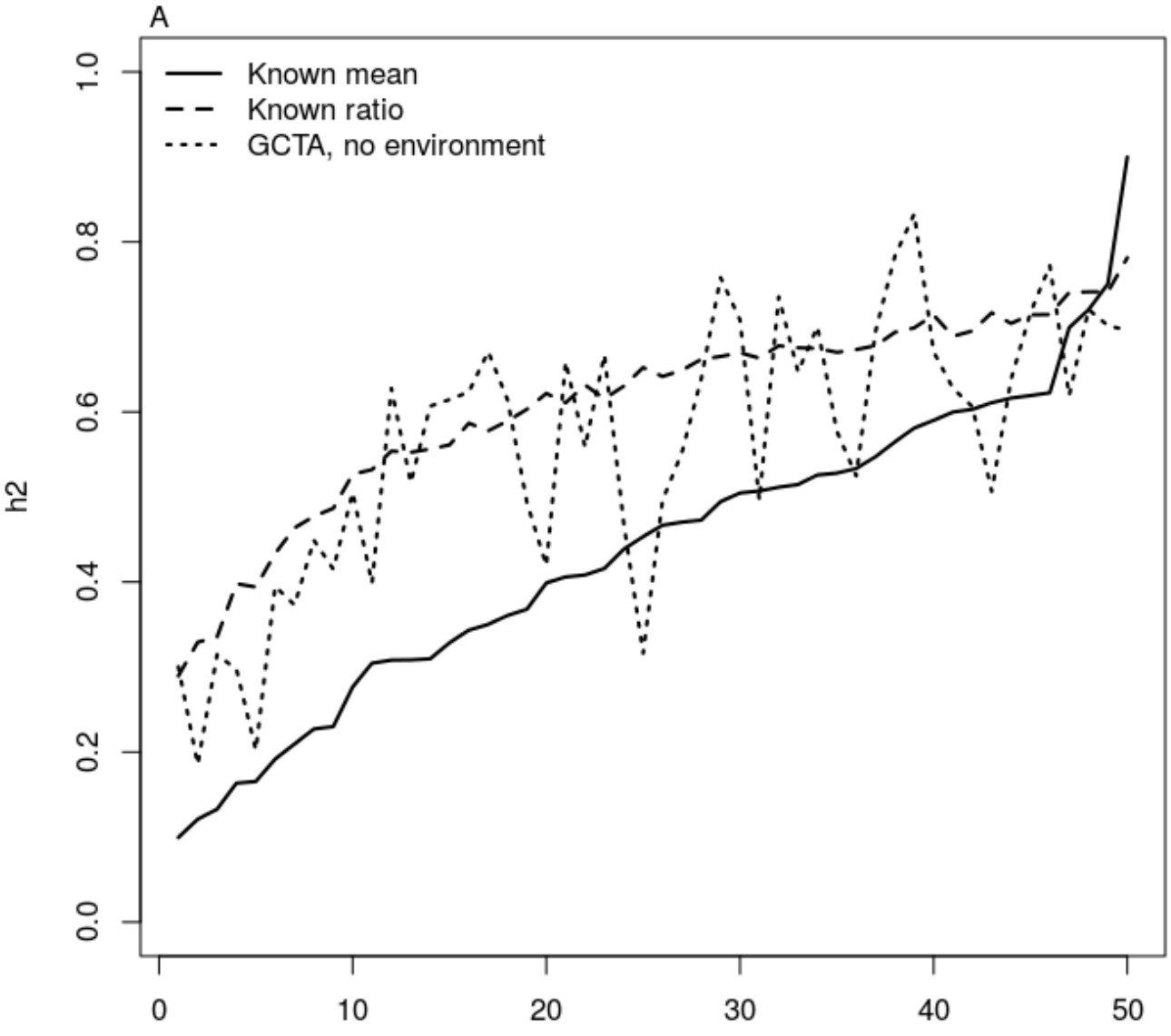
Q4: Environmental Moderation

Heritability not constant: What are implications for GCTA?

- Standard GCTA: $g \sim \text{MVN}[0, \sigma_g^2 A]$.
- We simulate data using $g \sim \text{MVN}[0, A']$ where (i, j)-th entry of A' is $h_i h_j A_{ij}$.

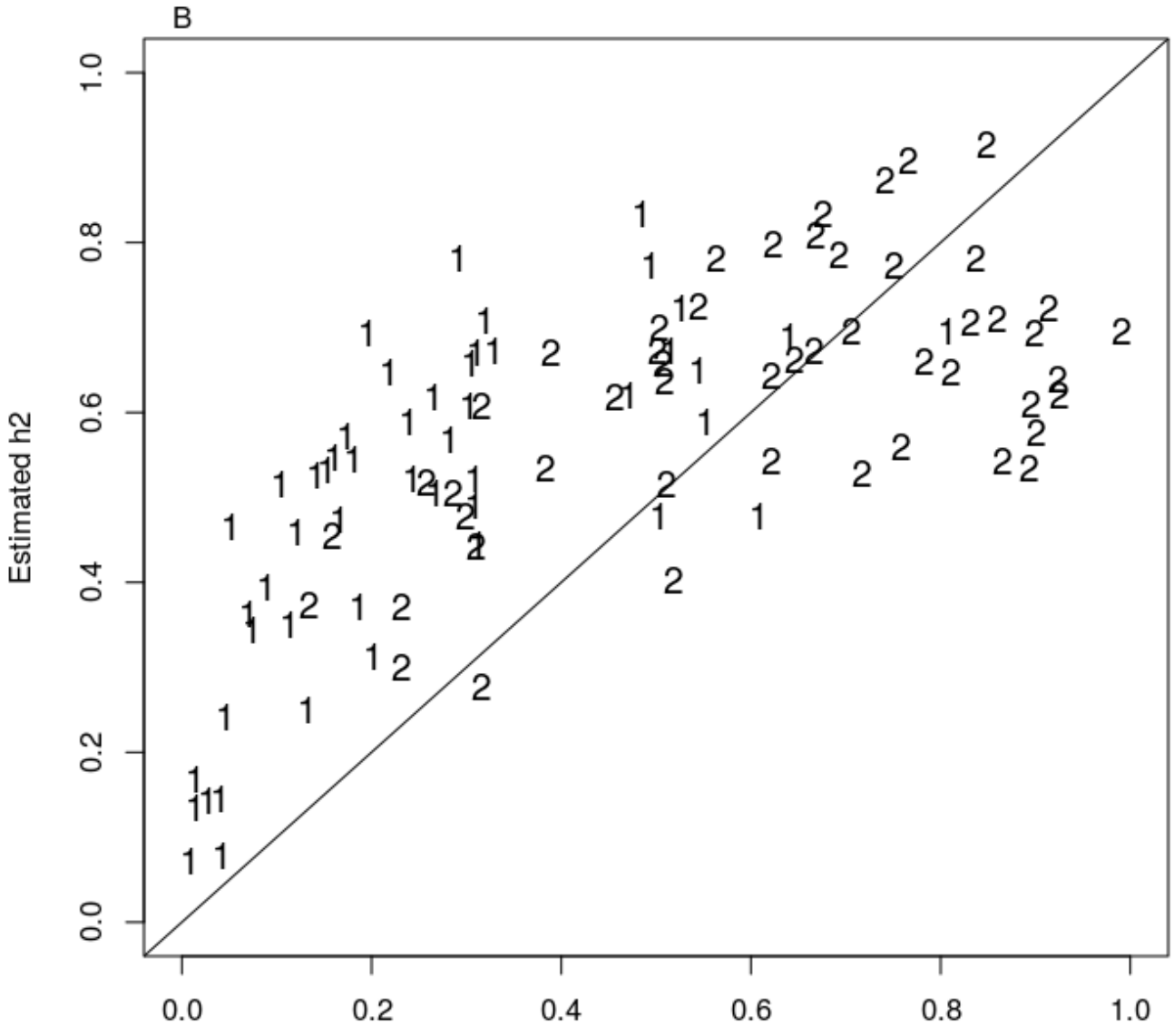
Q4: Environmental Moderation

What if we ignore environment?



Q4: Environmental Moderation

What if we allow for environmental variation?



- LD is important consideration (aside: I'm skeptical about using KING or REAP estimates).
- Heteroskedasticiy leads to inflation of h^2 estimates.
- Environmental differences are likely to be problematic (and yet may be rampant?).

- LD is important consideration (aside: I'm skeptical about using KING or REAP estimates).
- Heteroskedasticiy leads to inflation of h^2 estimates.
- Environmental differences are likely to be problematic (and yet may be rampant?).

In closing: GCTA is like a table saw.