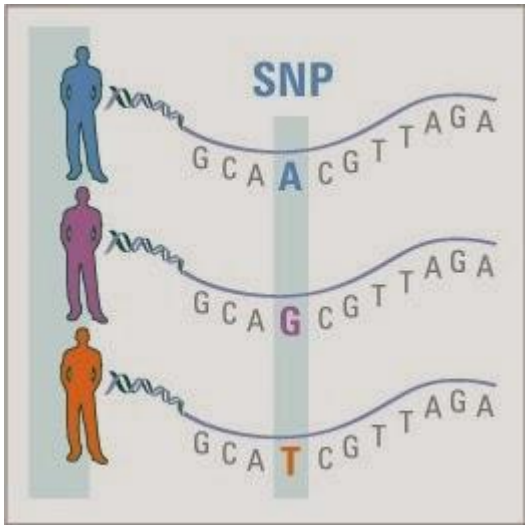


GENOMIC HERITABILITY AND LIKELIHOOD ESTIMABILITY USING THE G-BLUP

Yogasudha Veturi
II year PhD student
Department of Biostatistics
University of Alabama at Birmingham

Missing heritability

- Genome Wide Association Studies (GWAS) have reported **large numbers of variants** associated with important complex human traits and diseases (NHGRI, www.genome.gov/GWAStudies).
- A sizable proportion of **inter-individual differences** attributable to genetic factors remains largely **unaccounted for** (Manolio et al., 2009).
- GWAS **lack power** to detect associations of **small-effects variants**.
- With the **G-BLUP** we can estimate the proportion of variance that can be explained by **all-available markers** (Yang et al., 2010).

$$y_i = g_i + \varepsilon_i$$

Error

Genetic Value

Phenotype

$$var(\mathrm{y}_i) = var(\mathrm{g}_i) + var(\varepsilon_i) + 2cov(\mathrm{g}_i, \varepsilon_i)$$

$$\sigma_y^2 = \sigma_g^2 + \sigma_\varepsilon^2$$

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2}$$

Heritability: The proportion inter-individual differences of a trait (or in disease risk) that can be attributed to genetic factors.

Infinitesimal model

$$y_i = \mu + g_i + \varepsilon_i$$

$$\boldsymbol{g} \sim N(0, \mathbf{A}\sigma_a^2), \quad \boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma_\varepsilon^2)$$

Henderson, 1950

Subject	Y	SNP1	SNP2	SNP3	SNPj	SNPp
1	123	AA	gg	AA	…	GG
2	130	Ac	AA	AA	…	GG
3	146	AA	AA	AA	…	GG
i	…	…	…	…	…	…
N	101	Ac	Ag	AA	…	GG

X

G-BLUP

$$y_i = \mu + g_i + \varepsilon_i \text{ where } g_i = \sum x_{ij} u_j$$

$$\boldsymbol{u} \sim N(0, \sigma_u^2), \quad \boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma_\varepsilon^2),$$

$$\boldsymbol{g} \sim N\left(0, \sigma_g^2 = \mathbf{p}\sigma_u^2\right)$$

$$\boldsymbol{Var(y)} = G\sigma_g^2 + I\sigma_\epsilon^2$$

Computing genomic similarities

	I-1	I-2	II-1	II-2	II-3	II-4	III-1	III-2
I-1	1	0	0.5	0.5	0.5	0	0.25	0.25
I-2	0	1	0.5	0.5	0.5	0	0.25	0.25
II-1	0.5	0.5	1	0.5	0.5	0	0.25	0.25
II-2	0.5	0.5	0.5	1	0.5	0	0.25	0.25
II-3	0.5	0.5	0.5	0.5	1	0	0.5	0.5
II-4	0	0	0	0	0	1	0.5	0.5
III-1	0.25	0.25	0.25	0.25	0.5	0.5	1	0.5
III-2	0.25	0.25	0.25	0.25	0.5	0.5	0.5	1

$$\mathbf{G} = \frac{\frac{1}{p}\mathbf{X}\mathbf{X}'}{2\sum \theta_i(1-\theta_i)},$$

$$\mathbf{X} = \{x_i = 0,1,2\}$$

θ_i allele frequency

A: expected proportion of allele sharing.

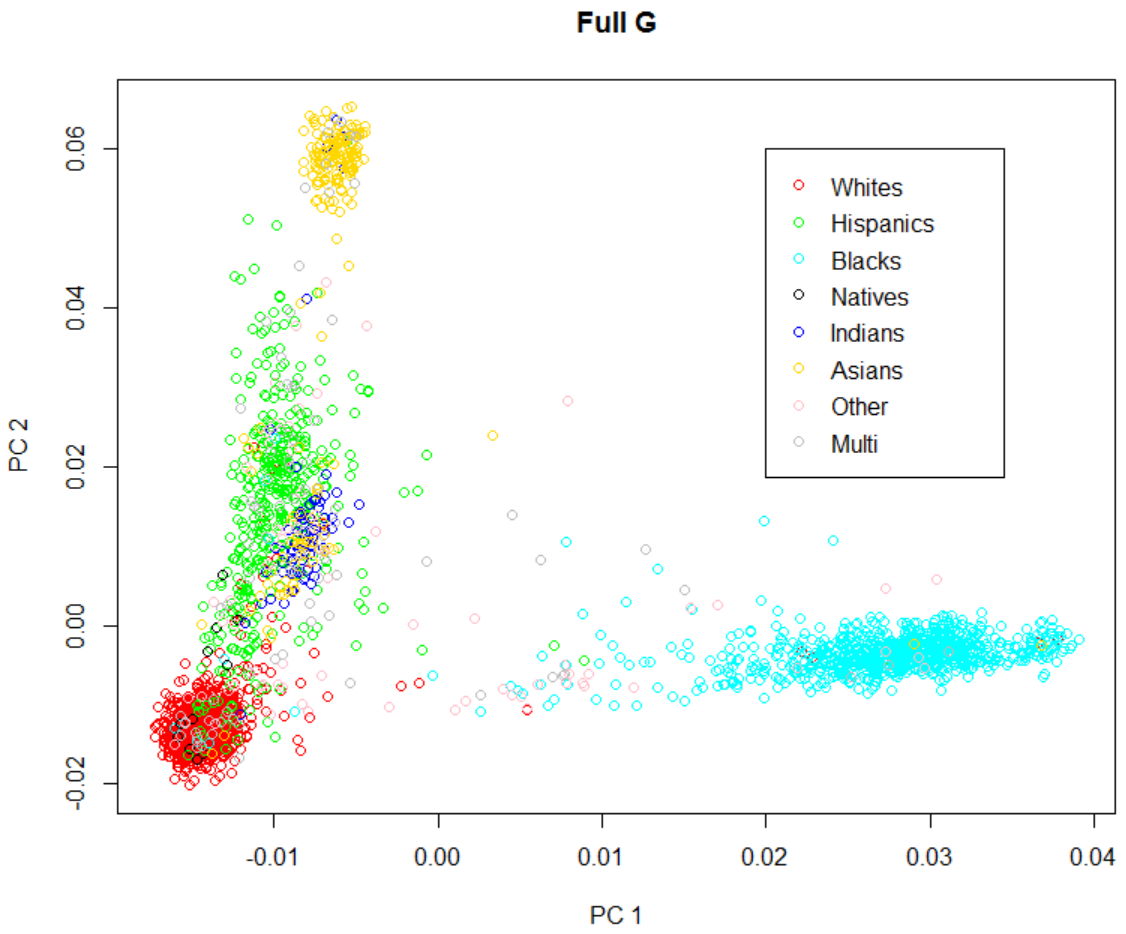
G: realized proportion of allele sharing at markers.

Materials and Methods

Training Interventions & Genetics of Exercise Response

Hypothesis: G-BLUP can explain a sizeable proportion of h^2_G for anthropomorphic traits in the TIGER study (Illumina metabochip)

RACE	SAMPLE SIZE	No. SNPs
Whites	1052	39438
Blacks	721	49679
Hispanics	321	75113
Asians	130	72057



Estimation of h^2_G using:

- **ML**
- **Bayesian methods**
- **Simulations**

ML/Bayesian Inference

$$p(\sigma_\varepsilon^2, \sigma_g^2 | \mathbf{y}, \boldsymbol{\theta}) \propto p(\mathbf{y} | \sigma_\varepsilon^2, \sigma_g^2)\, p(\sigma_\varepsilon^2, \sigma_g^2 | \boldsymbol{\theta})$$

$$\propto N(\mathbf{y} | \mathbf{0}, \mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_\varepsilon^2) \times \chi^{-2}(\sigma_g^2 | df_1, S_1) \times \chi^{-2}(\sigma_\varepsilon^2 | df_2, S_2)$$

Prior distribution

Likelihood function (used in ML)

Distribution of the unknowns given the data and hyper-parameters

---

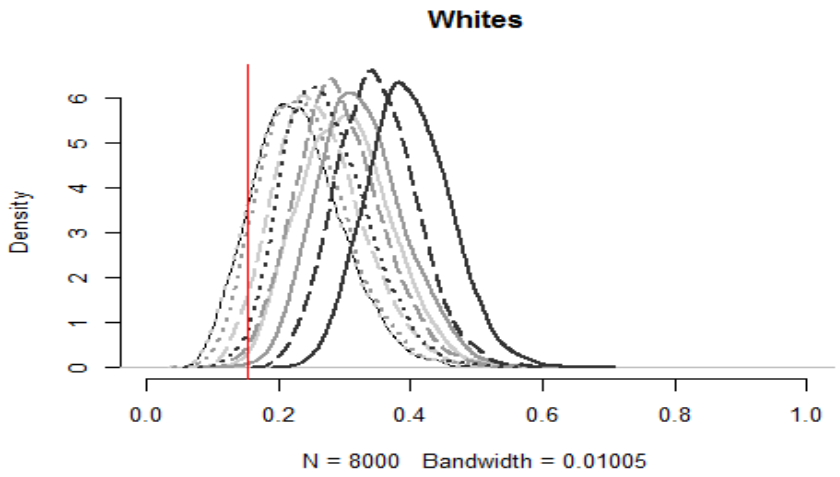
The Scale-Inverse Chi-Sq. Density

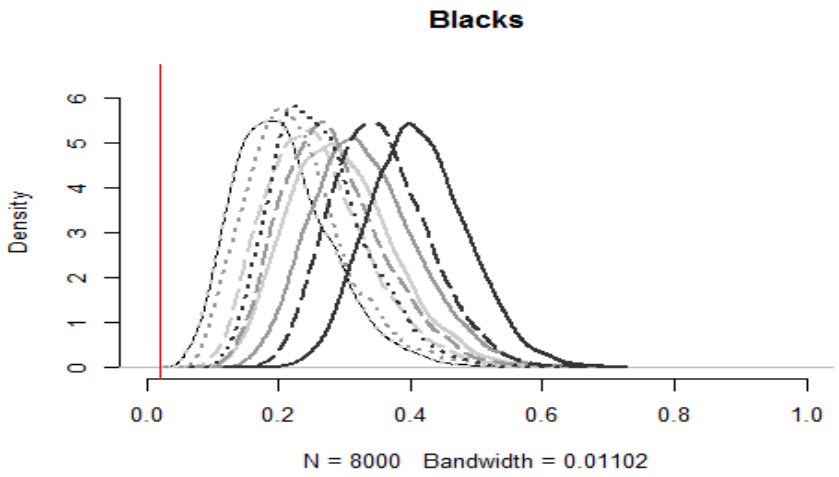
$$\frac{df_0\sigma_0^2 + \sum_{i=1}^{n}(y_i - \mu)^2}{df_0 + n} \longrightarrow \text{Posterior density}$$

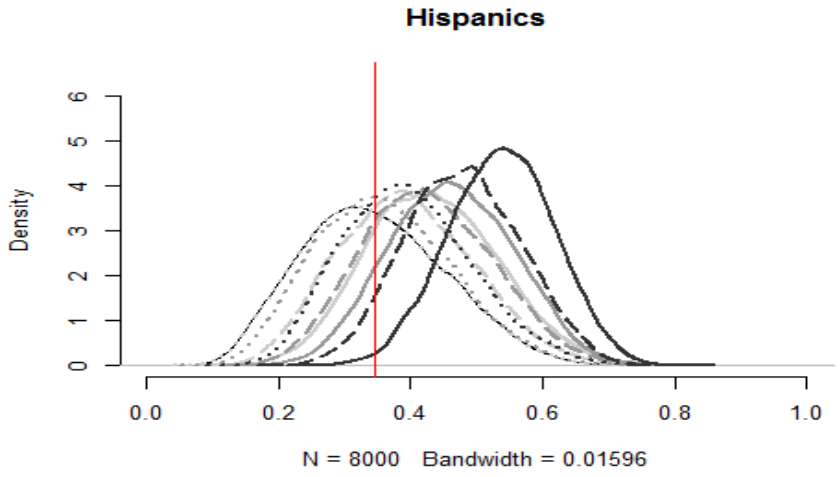
BGLR (R)

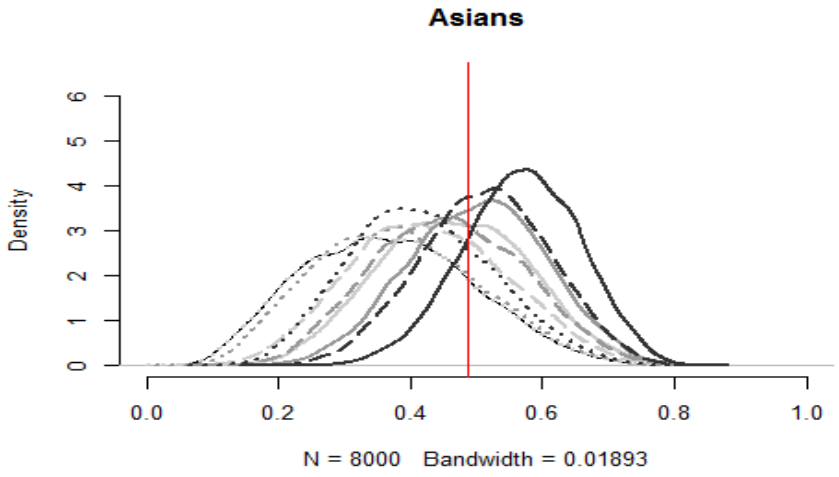
$df_0 = 1, 3, 5;\ \sigma_0^2 = 0.25, 0.5, 0.75$ → Hyperparameters

Distribution of h_G^2 for height using MCMC. Red line = ML estimate

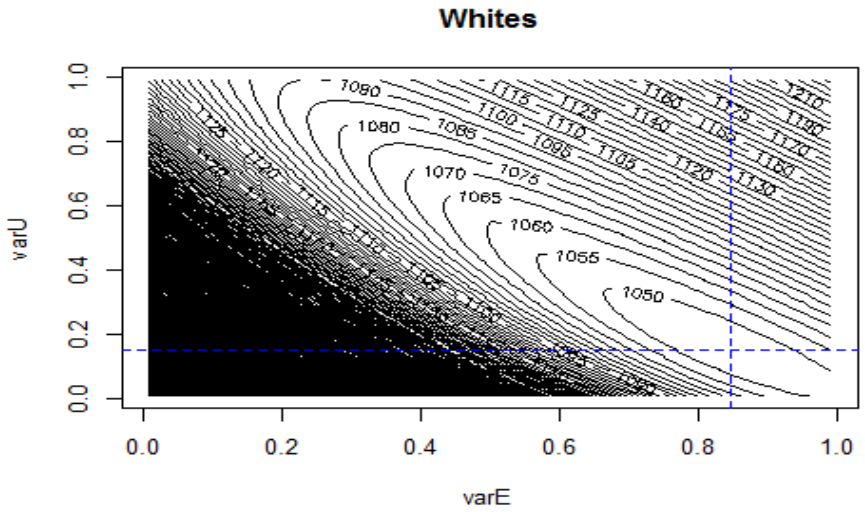


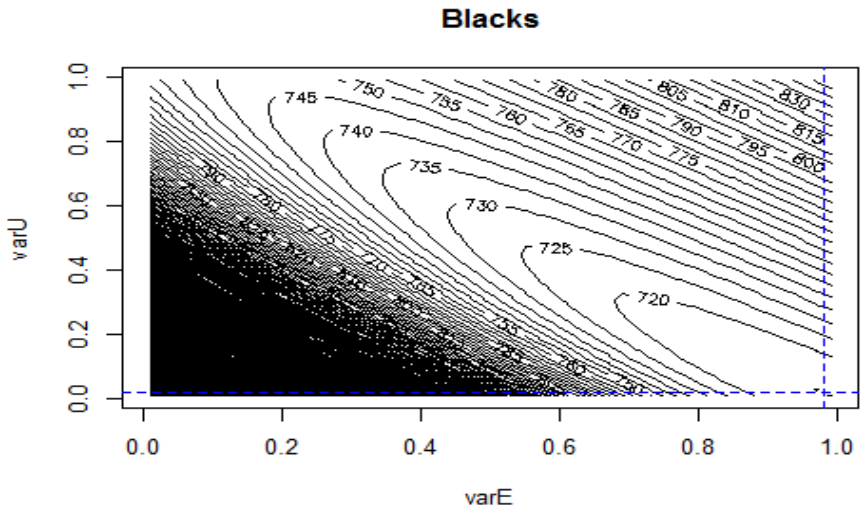


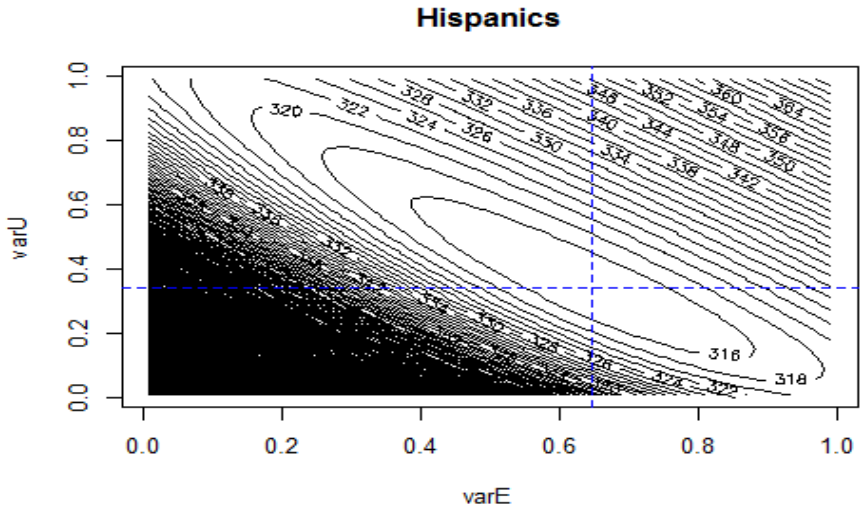


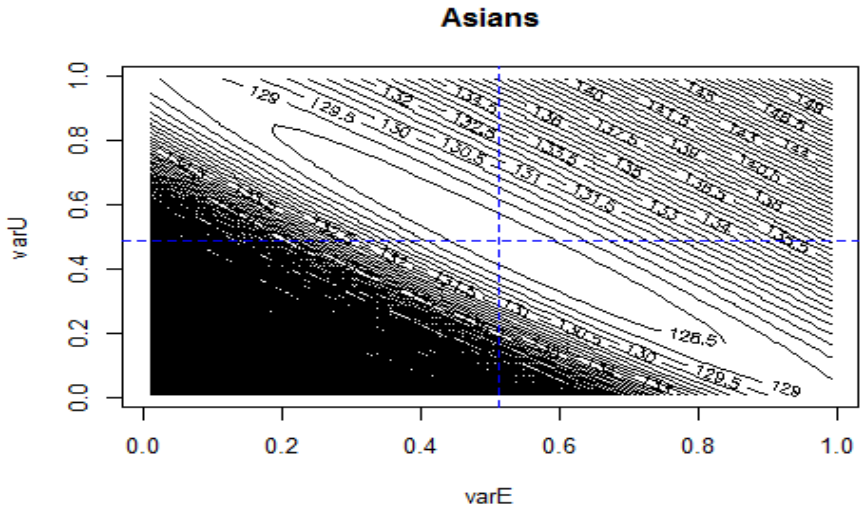


The Neg. Log-Likelihood Surface

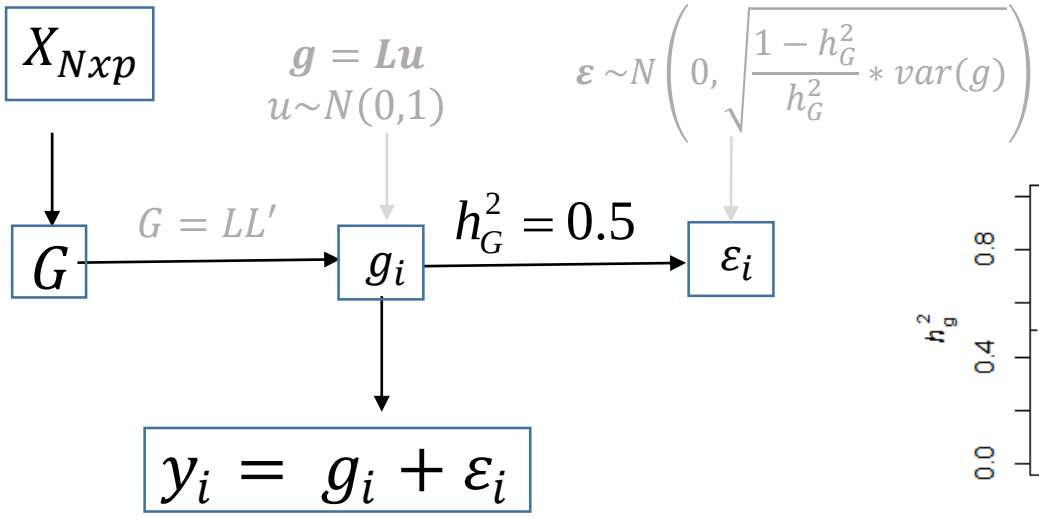




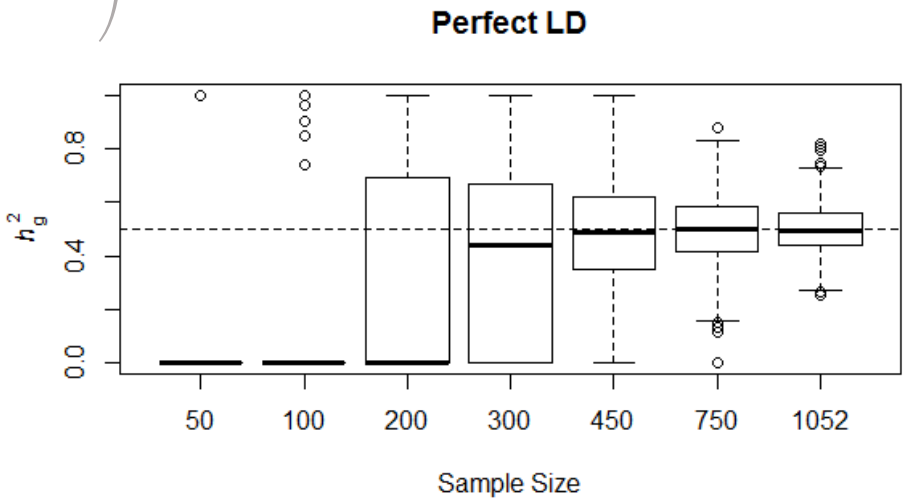


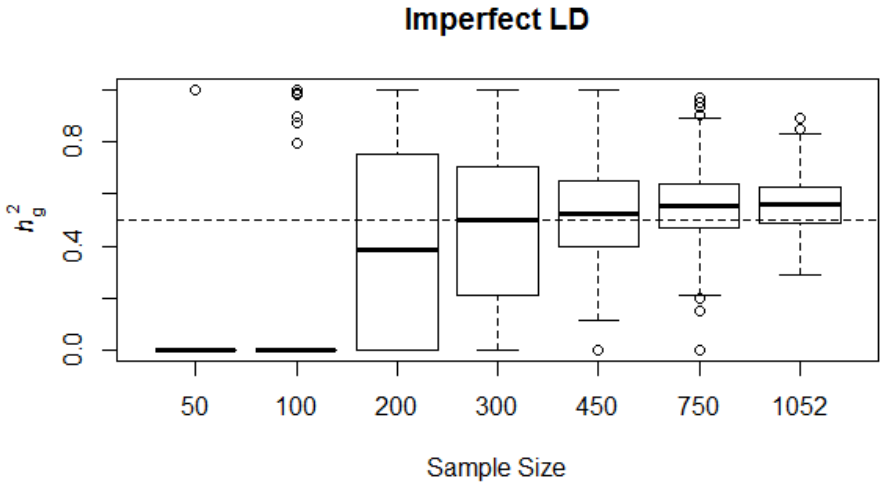


Simulations (ML estimates from 1000 reps)

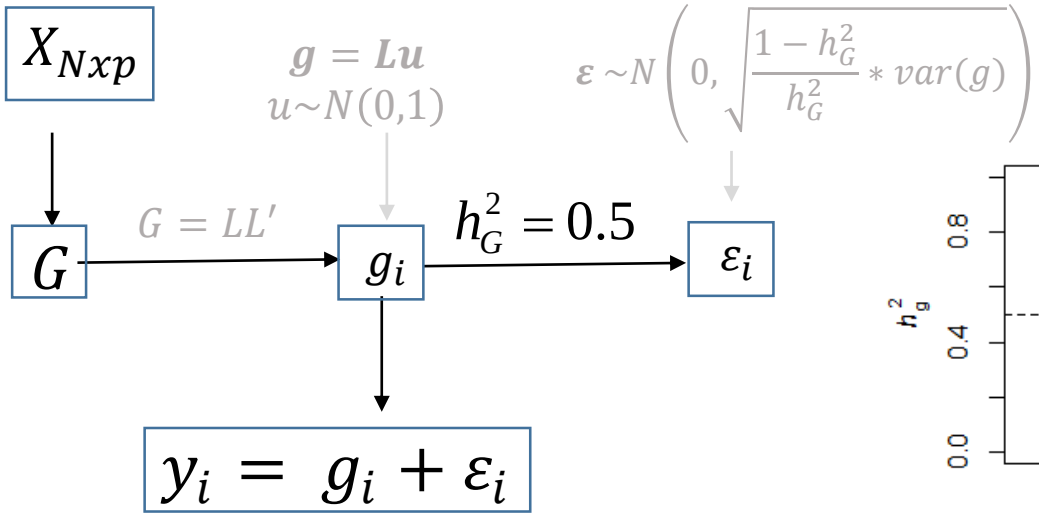


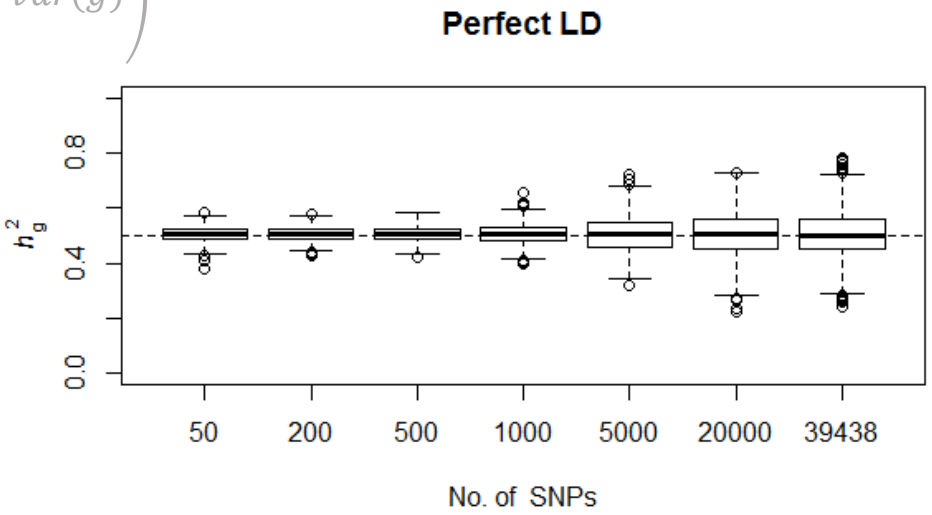
Z_{Nxp} Simulated dataset	W_{Nxp} Predictor dataset
$p = 39438$ $N = 50, 100 \ldots 1052$	$p = 39438$ $N = 50, 100 \ldots 1052$
$p = 250$ $N = 50, 100 \ldots 1052$	$p = 39188$ $N = 50, 100 \ldots 1052$



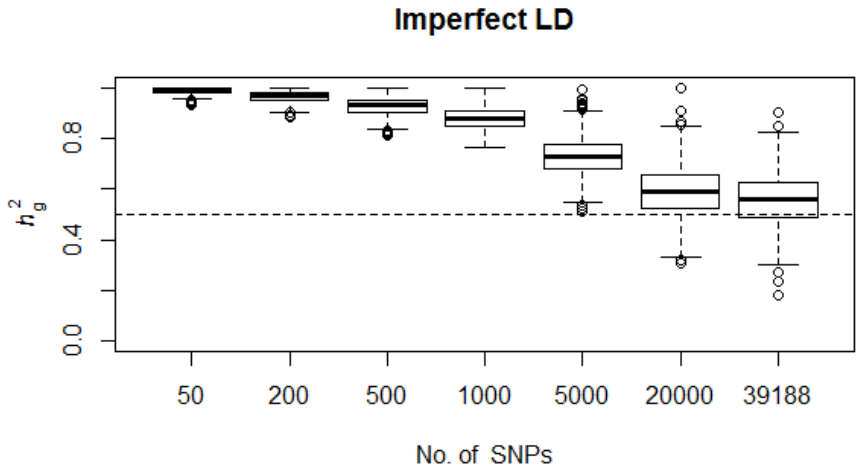


Simulations (ML estimates from 1000 reps)





Z_{Nxp} Simulated dataset	W_{Nxp} Predictor dataset
$p = 50,200, \ldots 39438$ $N = 1052$	$p = 50,200, \ldots 39438$ $N = 1052$
$p = 250$ $N = 1052$	$p = 50,200, \ldots 391888$ $N = 1052$



Conclusions

- Variance component and heritability estimation using ML resulted in some corner solutions.
- Bayesian estimates were very sensitive to choice of prior.
- This was the result of an estimability problem caused by a flat likelihood function; there was a large area corresponding to the same likelihood.
- Thus, variance component estimates should be reported after a careful study of likelihood profiles.
- Simulation studies suggest that the probability of corner solutions reduce by increasing sample size and decreasing SNP density, however lower SNP density might result in increased bias.

Acknowledgements

University of Alabama at Birmingham
Dr. A.I. Vazquez
Dr. G. de los Campos - Advisor

University of Texas at Austin.
M. Bray

Students, stuff and participants of the TIGER study.

Training Interventions & Genetics of Exercise Response

NIDDK National Institute of Diabetes and Digestive and Kidney Diseases

UAB

Nutrition Obesity Research Center

SSG