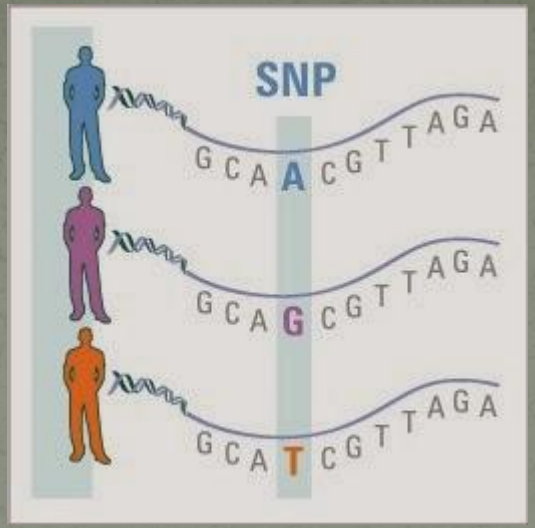
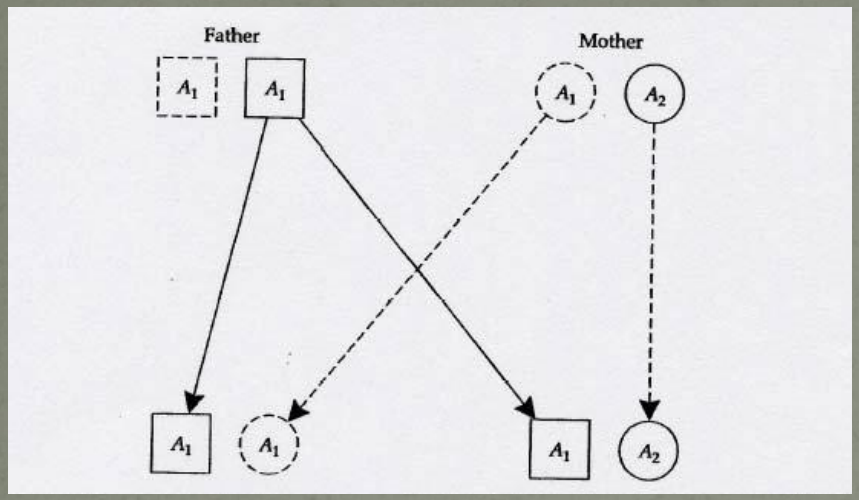


Estimation of genetic relatedness and heritability

Yogasudha Veturi
BST 775, Fall 2013
University of Alabama at Birmingham

Genetic Relatedness. Why study it?



Lynch M, Walsh B(1998) Genetics and analysis of quantitative traits.

Utility:

- Forensics
- Agriculture and Animal breeding
- Ecology
- Human Genetics

P(IBD): Probabilities that sets of genes have descended from a single ancestral gene

- Mapping human genes
- Predicting genotype frequencies
- Estimating genetic variances

Genetic relatedness over the years..

- Traditionally, estimates of genetic relatedness (probabilities of IBD) were calculated from known pedigrees (Malecot 1969, Wright; 1943)
- Later, there were methods to estimate cryptic relatedness or recent ancestry

- **E.g.: Thompson's MLE:** The joint probability of genotypes G_1 and G_2 of individuals 1 and 2, conditional on their degree of pairwise relatedness parameterized by $\mathbf{k} = (k_0, k_1, k_2)$ and conditional on the allele frequencies in the population is:

$$P(G_1, G_2 | k) = \mathbf{k_0} P(G_1) P(G_2) + \mathbf{k_1} P(G_1) P(G_2|G_1) + \mathbf{k_2} P(G_1) I(G_2=G_1)$$

For L multiple linked loci: $$P(\mathbf{G_1}, \mathbf{G_2} | \mathbf{k}) = \prod_{j=1}^{L} P(G^{(j)}_1, G^{(j)}_2 | \mathbf{k})$$

MLE is the value of $\mathbf{k}$ that maximizes this joint probability; subject to constraints $0 \leq \mathbf{k_0}, \mathbf{k_1}, \mathbf{k_2} \leq 1$ and $\mathbf{k_1}^2 \geq 4\, \mathbf{k_0}\, \mathbf{k_2}$

- Queller and Goodnight (1989), Ritland (1996), Lynch and Ritland (1999), and Wang (2002) are some other non-likelihood based methods for estimating genetic relatedness

Genomic relationship matrix (G)

- Originated in animal breeding
- Used to estimate the proportion of chromosome segments shared by individuals
- Genes that are IBS (identical by state) can be shared through common ancestors not recorded on pedigree
- Greatly useful for genomic prediction in quantitative genetics

Why am I concerned about G?

- *To estimate genetic variances and heritability for human traits!*
- h^2 :Proportion of variation in phenotype that is attributable to the genotype; the additive genetic component is called **narrow-sense** heritability

$$h^2 = \frac{\sigma^2{}_G}{\sigma^2{}_G + \sigma^2{}_E}$$

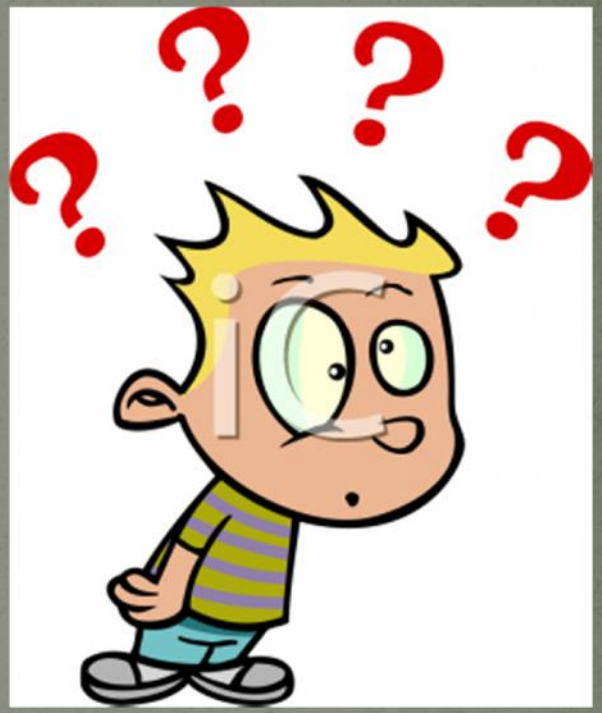
- Dense genotype data can explain large amount of genetic variation when using whole genome statistical models
- LD is generated by the short genomic regions passed by remote common ancestors
- h^2 =causal variant heritability that is tagged by the genotyped SNPs

…however, despite dense genotypic data….

- There is missing heritability!!

Why is heritability missing?

- Rare variants and undetected CNVs
- Insufficient sample sizes
- Causal variants are not in complete LD with the genotyped SNPs
- Mismatch between genetic architecture and statistical modeling



Yang Study (2010)

- Using Whole Genome Prediction (WGP) method on human height:

Common SNP variation explained 45% of the phenotypic variance, accounting for more than 50% of the expected heritability of height (approx. 80%)

- Results suggest infinitesimal model for height

How to make use of G?

Basic Model

- $y_j = \mu + g_j + e_j$ where $g_j = \sum z_{ij} u_i$
- w_{ij} is the genotype of individual *i* at the j^{th} of *m* diallelic loci with additive coding of genotypes, $E(w_{ij}) = 0$ and $Var(w_{ij}) = 1$; $e_i \sim iid\ N(0, \sigma^2_e)$ $u_i \sim iid\ N(0, \sigma^2_g/m)$
- $Var(Y) = \sigma^2_g \frac{WW'}{m} + \sigma^2_e I = \sigma^2_g G + \sigma^2_e I$
- In reality G is unknown so a G matrix is estimated using genome-wide sample of SNPs

Methods for estimating G

- If P is the matrix of allele frequencies, *n* is the number of individuals, *m* is the number of markers and X is the allele sharing matrix (0,1,2) and Z is the allele sharing matrix centered at its mean, i.e. Z =X-P:

- Van Raden et al. 2008 (VAN)
- $G = \frac{ZZ'}{2 \sum p_i(1-p_i)}$

- Leutenegger et al. 2009(DEF)
- G = ZDZ' where D is diagonal with $D_{ii} = \frac{1}{m[2p_i(1-p_i)]}$ (weights markers by reciprocals of expected variance

- Legarra et al. 2009 (LEG)
- $G = \frac{ZZ'}{tr(ZZ')/n}$

Methods for estimating G

- Gianola et al. 2010 (GIA)
- $$G = \frac{ZZ'}{\left((p_0 - q_0)^2 + \left(\frac{2\sum_{i=i}^{m} p_i(1-p_i)}{m}\right)\left(\frac{\alpha+\beta+2}{\alpha+\beta}\right)\right)m}$$

p_0 and q_0 are expectations of allele frequencies from a Beta distribution with hyper-parameters α and β

- Unified Additive Relationship - Yang et al. 2010 (UAR)

$$G = ZDZ'$$

$$\text{diag}(G) = G_{ii} = 1 + \frac{1}{m}\sum_k \frac{x_{ik}^2 - (1+2p_k)x_{ik} + 2p_k^2}{2p_k(1-p_k)}$$

Provides an unbiased estimate of inbreeding coefficient

- Adjusted Unified Additive Relationship - Yang et al. 2010 (aUAR)

$$G^*_{ij} = 1 - \frac{1}{m * Var(G)} * G \quad i \neq j$$

$$= 1 + \left(1 - \frac{1}{m * Var(G)}\right) * (G - 1) \quad i \neq j$$

Corrects for sampling error of UAR

- So.. the diagonal of the G is important..

Simulation – Toy example

- 1500 individuals simulated with 1000 SNPs at 50% heritability
- Allele frequencies drawn from a beta distribution
- Effects assigned to 100 QTL (drawn from a normal distribution)and genetic signal = X*b_0
- Error calculated from a N$\left(0, \sqrt{\frac{1-h^2}{h^2} * var(b_0)}\right)$
- Phenotypes simulated as y = signal + error

Simulation – Toy example

- After model-fitting, almost the entire h^2 was recovered using any of the Gs when all QTL were included (except GIA which over-estimated h^2)

$$h^2 = \frac{\sigma^2{}_G}{\sigma^2{}_G + \sigma^2{}_E}$$

Upon adding a small constant to the diagonal up to 0.5 ↓

$$h^2 = \frac{\sigma^2{}_G}{\sigma^2{}_G + \sigma^2{}_E}$$

Dataset

The TIGER Study

- …"how variation in DNA sequence may influence levels of body fatness and fitness both prior to and following a 30-week exercise program".
- …"how genes may alter response to exercise and diet interventions is not known."
- 3,200 men and women (18-30 yrs) drawn from the student population at UAB
- Genotyped using the Illumina Metabochip; ~200,000 SNPs of interest for metabolic and atherosclerotic / cardiovascular disease traits

Source: www.tigerstudy.org

Hypothesis

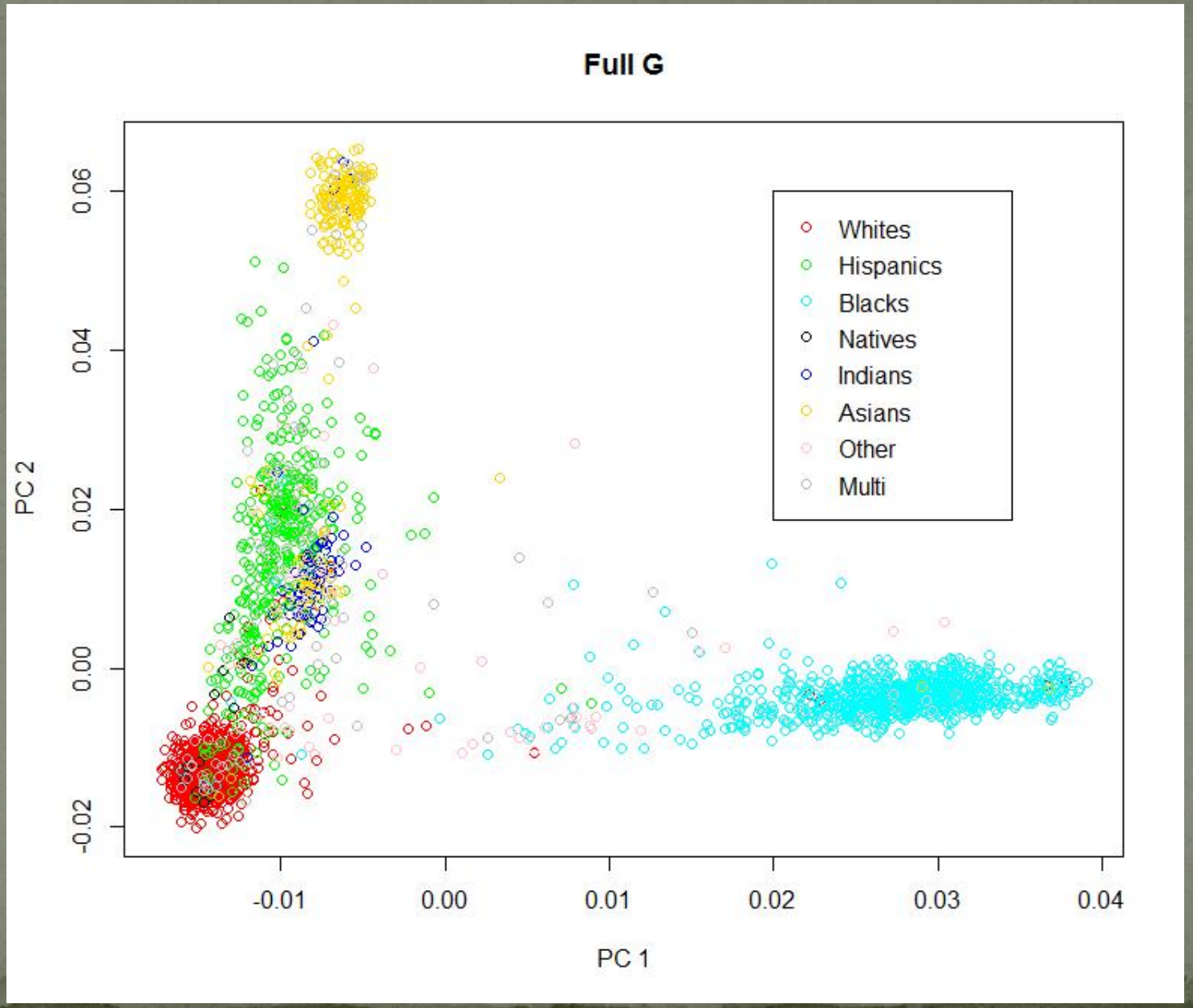
Genetic variance in body composition, obesity and bone composition can be explained using a whole-genome genetic model e.g., (Yang et al., 2010) with the high-density-genotyping from the Illumina MetaboChip platform for each trait.

Quality control:

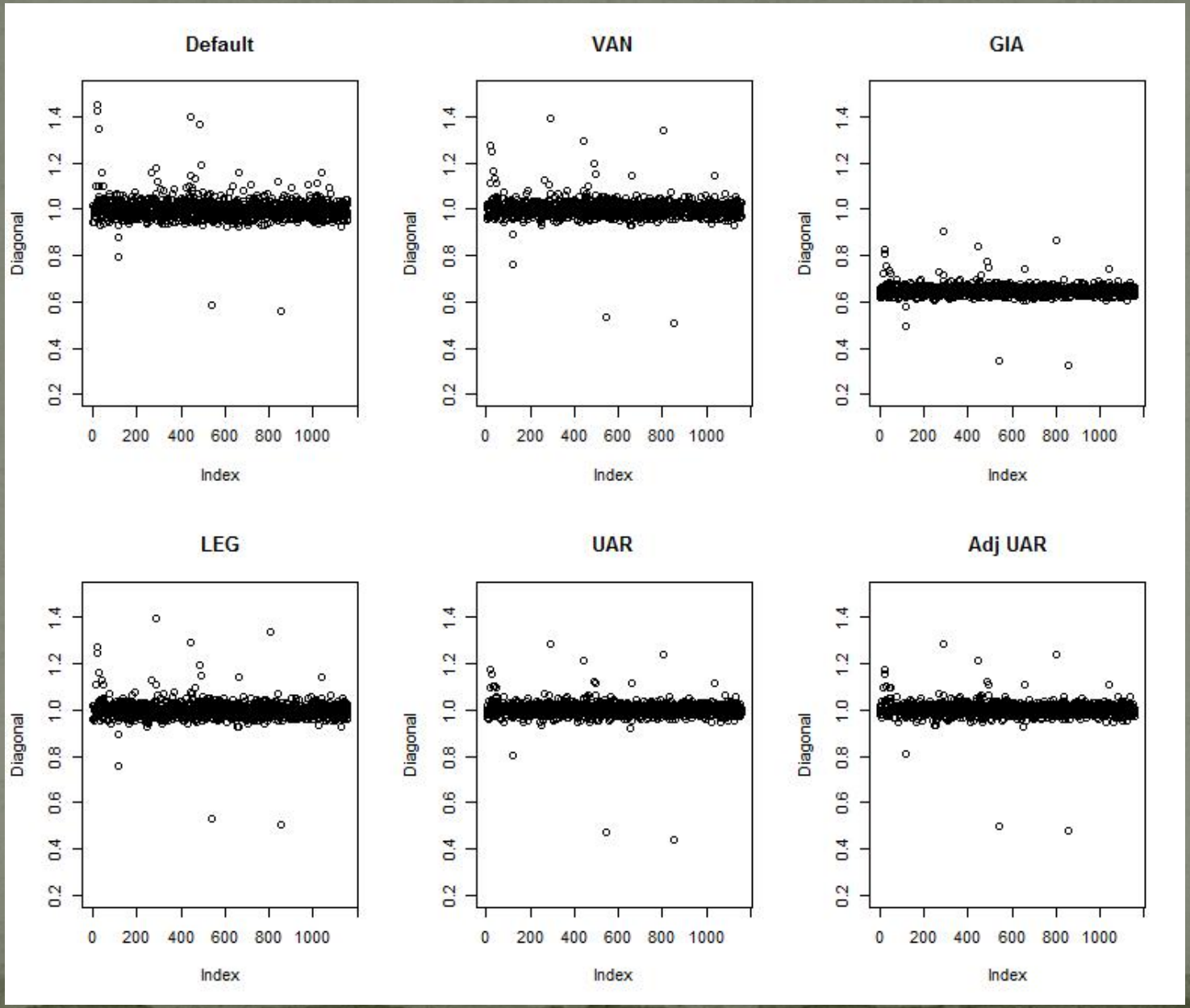
- SNPs with minor allele frequency lesser than 5% were removed
- Individuals with missing values greater than 5% were removed

Results

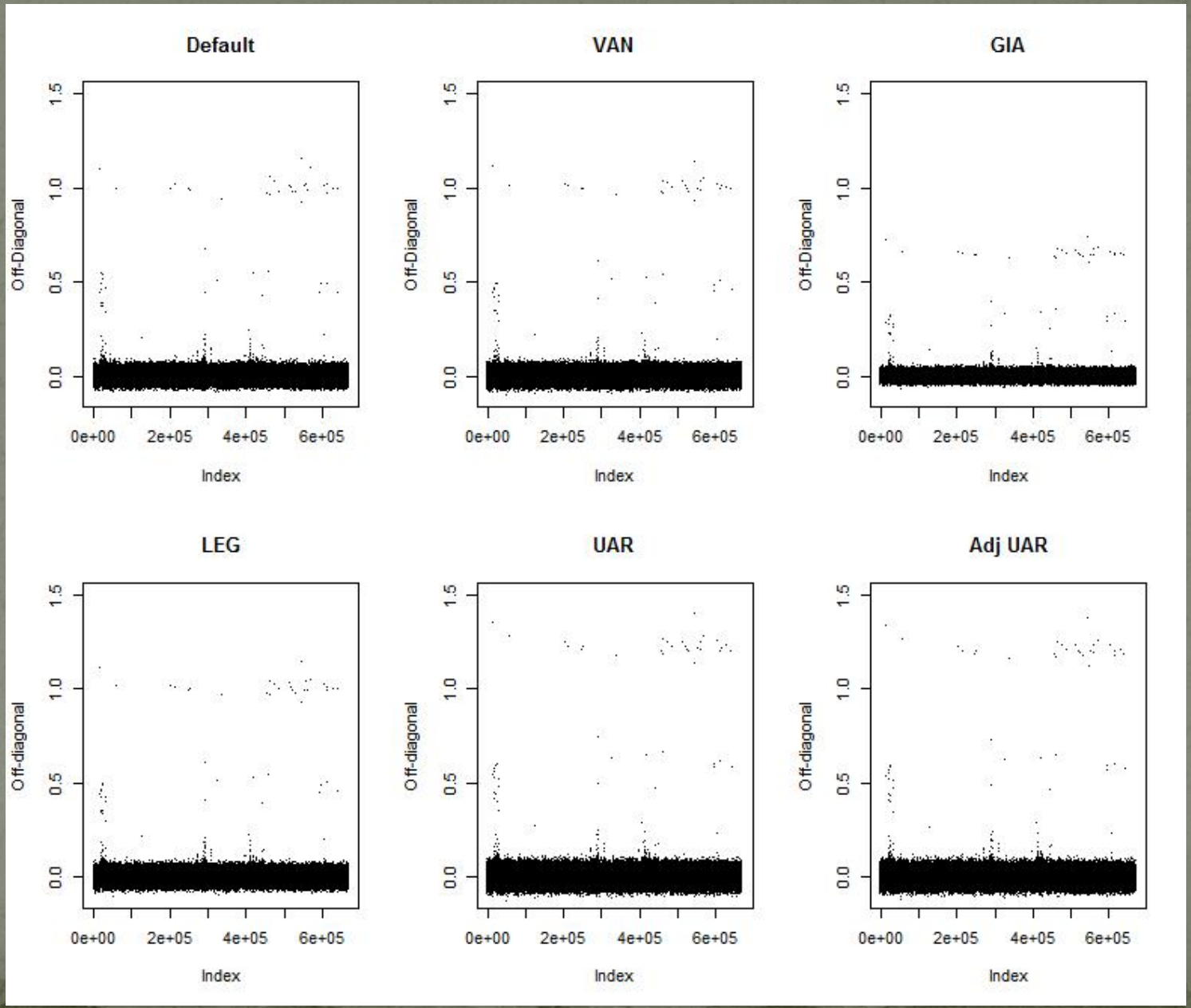
Principal Component Analysis



G matrix analysis (Whites)



G matrix analysis (Whites)

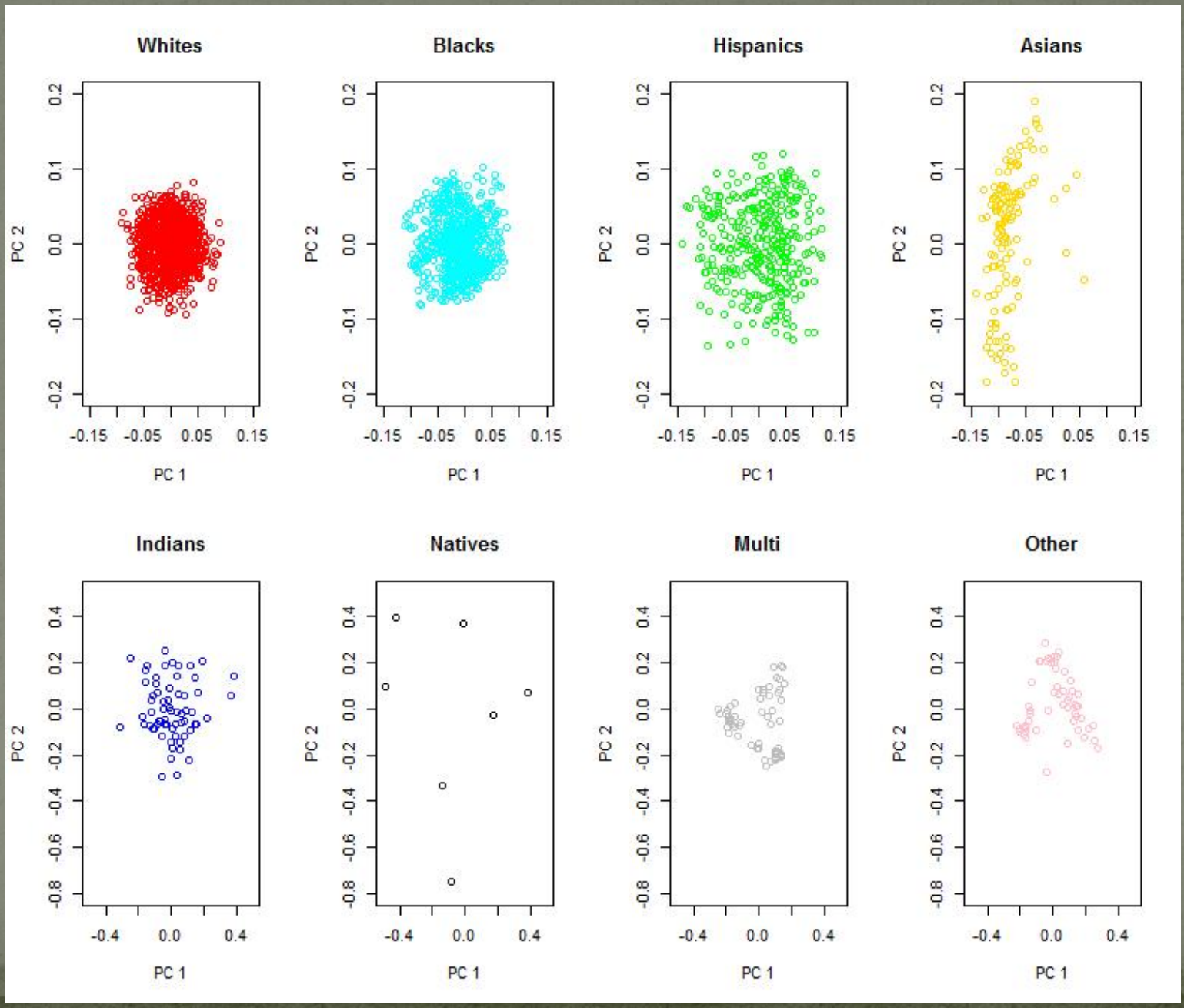


Further QC

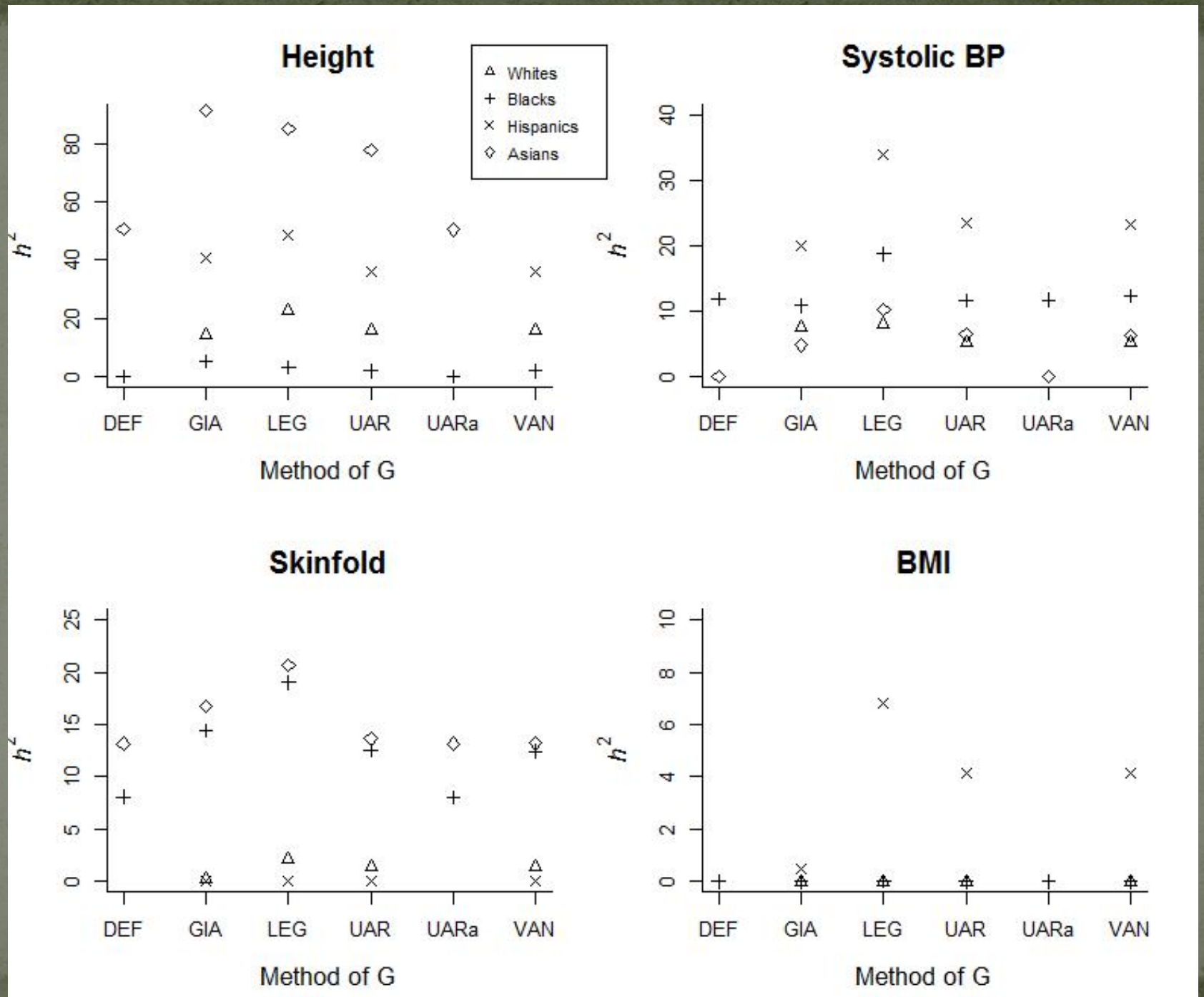
- QC was performed *by race*, as before
- Individuals with relatedness > 10% (per race) were removed

RACE	SAMPLE SIZE	No. SNPs
Whites	1054	39438
Blacks	721	49679
Hispanics	321	75113
Asians	130	72057

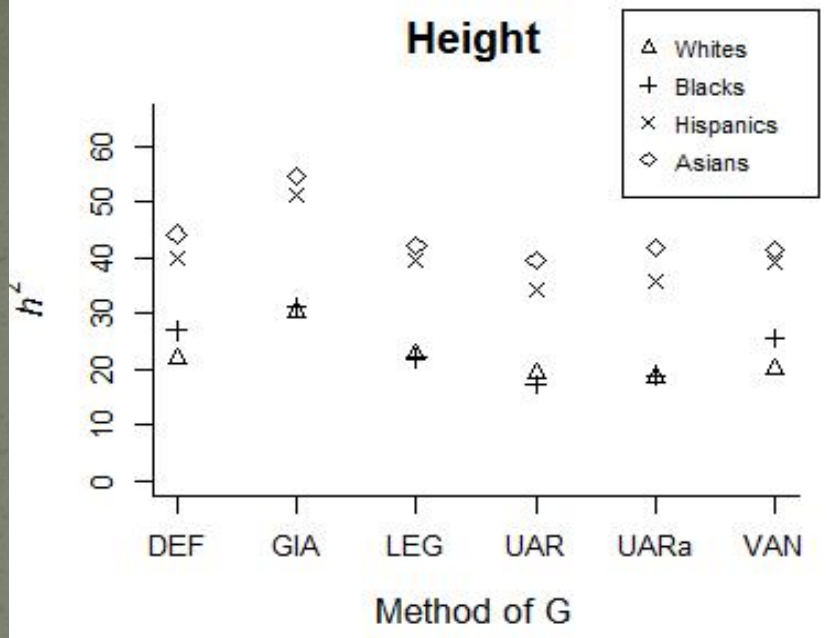
Principal Component Analysis – by race

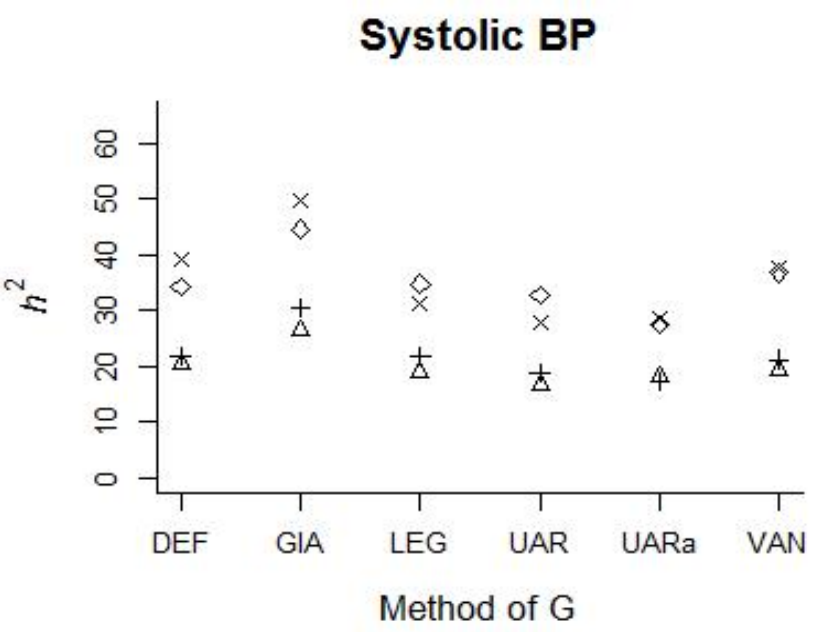


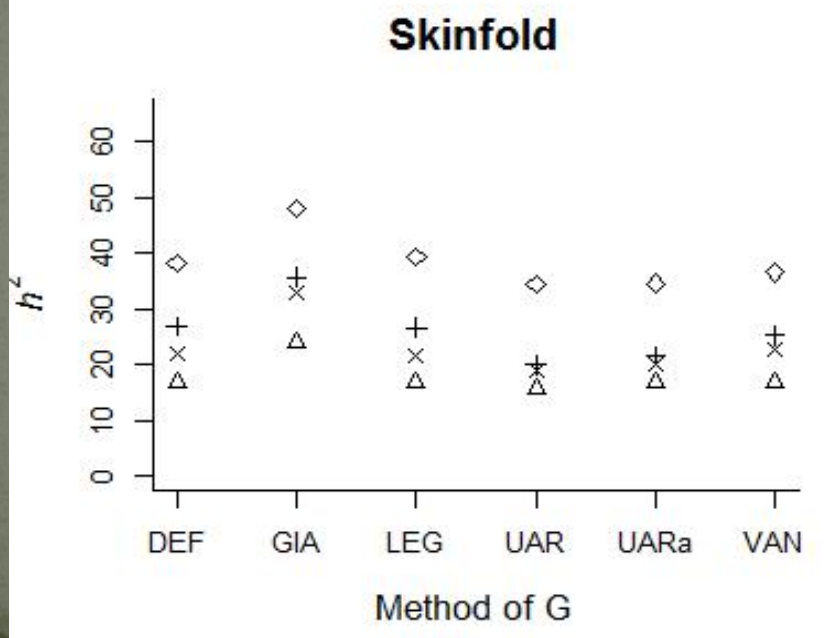
Heritability estimates using REML

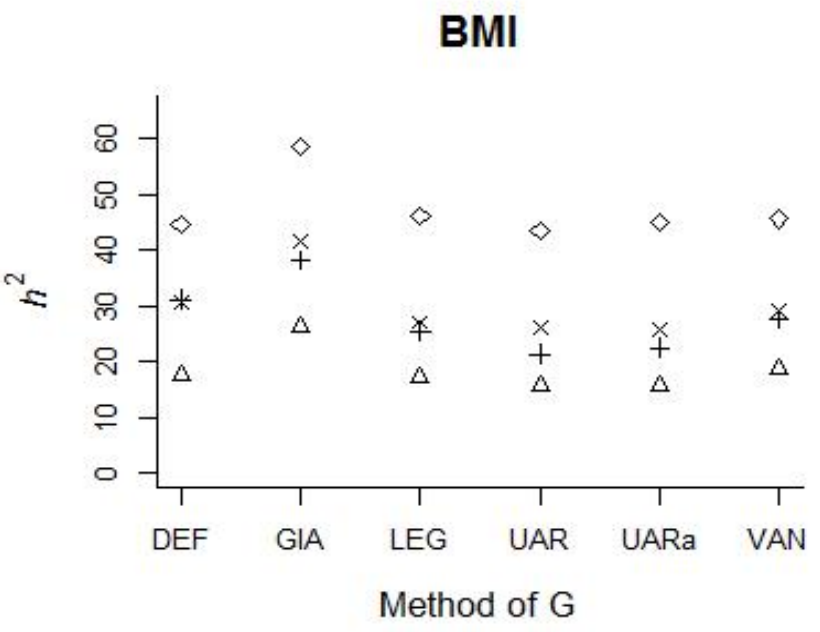

Height
Whites
Blacks
Hispanics
Asians
h^2
0
20
40
60
80
DEF
GIA
LEG
UAR
UARa
VAN
Method of G
Systolic BP
h^2
0
10
20
30
40
DEF
GIA
LEG
UAR
UARa
VAN
Method of G
Skinfold
h^2
0
5
10
15
20
25
DEF
GIA
LEG
UAR
UARa
VAN
Method of G
BMI
h^2
0
2
4
6
8
10
DEF
GIA
LEG
UAR
UARa
VAN
Method of G

Heritability estimates using MCMC


Height
Whites
Blacks
Hispanics
Asians
h^2
0
10
20
30
40
50
60
DEF
GIA
LEG
UAR
UARa
VAN
Method of G


Systolic BP
h^2
0
10
20
30
40
50
60
DEF
GIA
LEG
UAR
UARa
VAN
Method of G


Skinfold
h^2
0
10
20
30
40
50
60
DEF
GIA
LEG
UAR
UARa
VAN
Method of G


BMI
h^2
0
10
20
30
40
50
60
DEF
GIA
LEG
UAR
UARa
VAN
Method of G

Conclusions

- Differences in h^2 ranged from 0% – 5% between the different methods to estimate G (excluding GIA) for MCMC.
- UAR and aUAR were sometimes not positive definite, preventing model fitting using REML.
- Sample size and SNP density could affect h^2 estimates.
- REML could not partition the total variability between genetic and residual for traits like BMI whereas MCMC was able to.

Thanks!