Modeling genetic heterogeneity and sex differences using random effect interactions

SUDHA VETURI

Ph.D. candidate, Department of Biostatistics,
University of Alabama at Birmingham

sveturi@uab.edu

13 Apr 2016

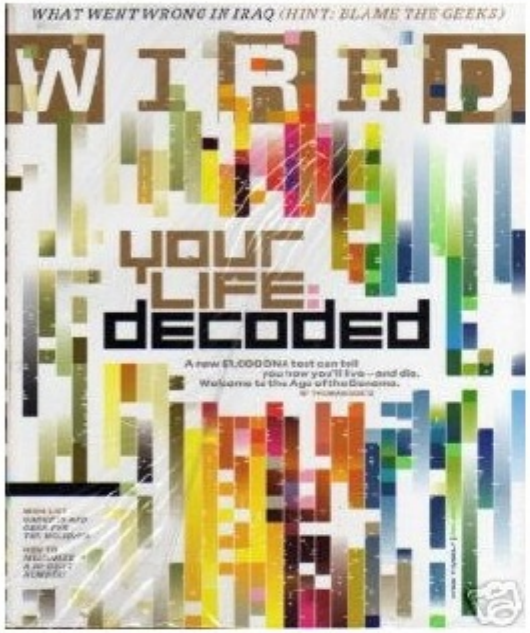
Outline

- Introduction
  - GWAS and Whole Genome Regression methods
  - Missing heritability
- Problem statements
  - Applications of WGR methods to study:
    - Population structure
    - Sex differences
- Interaction model
- Data
- Results and conclusions

Quantitative genetics

Genome Wide Association Studies

genome-wide association study

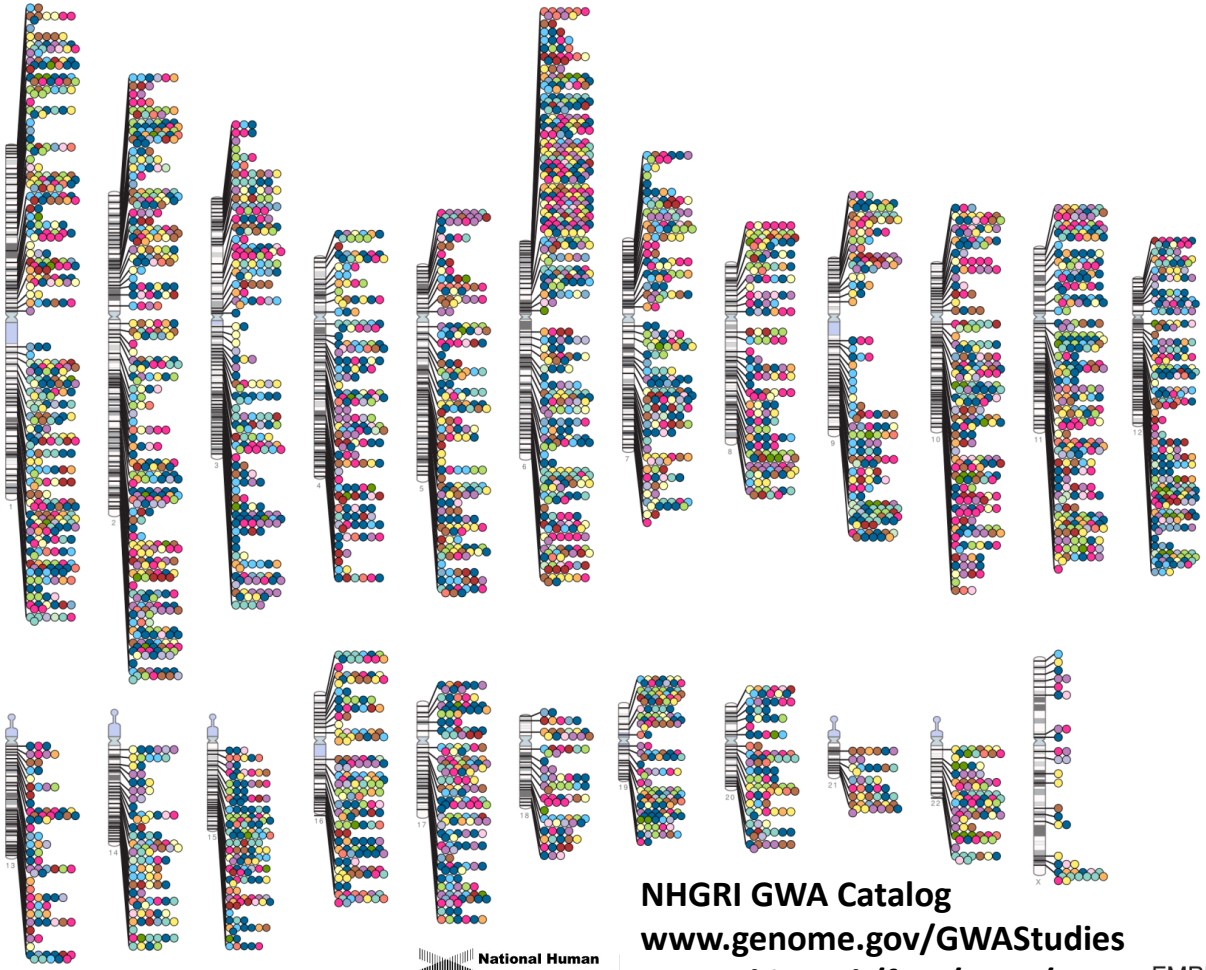


"... a study of common genetic variation across the entire human genome designed to identify genetic associations with observable traits."

-- National Institutes of Health, "Policy for sharing of data obtained in NIH-sponsored or conducted GWAS"

Genome Wide Association Studies

Published Genome-Wide Associations through 12/2013
Published GWA at p≤5X10^{-8} for 17 trait categories



- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body measurement
- Cardiovascular measurment
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease
- Other trait

NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI

Missing heritability

The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Proportion of genetic variance that is *unexplained*

Lack of power of GWAS to detect small-effect variants

MAHER B., 2008. Nature **456**: 18–21

Missing heritability

Medical Condition / Topic	Heritability Est.	References
Alcoholism	50 - 60%	[PMID 19785977]
Alzheimer's disease	58 - 79%	[PMID 16461860]
Anorexia nervosa	57 - 79%	[PMID 19828139]
Asthma	30%	[PMID 16117840]
Attention deficit hyperactivity disorder	70%	[PMID 22833045]
Autism	30 - 90%	[PMID 17033636]
Bipolar disorder	70%	[PMID 14601036]
Bladder cancer	7 - 31%	[PMID 21927616]
Blood pressure, diastolic	49%	[PMID 19858476]
Blood pressure, systolic	30%	[PMID 22479213]
Body mass index	23 - 51%	[PMID 25383972, PMID 18271028]
Bone mineral density	44 - 87%	[PMID 15750698, PMID 16025191]
Breast cancer	25 - 56%	[PMID 11979442, PMID 2491011]
Cervical cancer	22%	[PMID 11979442]
Colon cancer	13%	[PMID 11979442]
Coronary artery disease	49%	[PMID 15710764]

Source : Dr. Stephen Hsu Blog http://infoproc.blogspot.com/

Genetic architecture of complex traits

- Many complex traits are affected by large numbers of small effect genes

PLOS | GENETICS

Browse Publish About

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Ubiquitous Polygenicity of Human Complex Traits: Genome-Wide Analysis of 49 Traits in Koreans

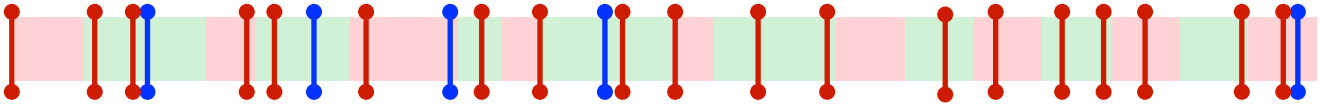
Jian Yang, Taeheon Lee, Jaemin Kim, Myeong-Chan Cho, Bok-Ghee Han, Jong-Young Lee, Hyun-Jeong Lee, Seoae Cho, Heebal Kim

Published: March 7, 2013 • DOI: 10.1371/journal.pgen.1003355

What statistical methodologies can we use to recover some of this missing heritability?

Whole Genome Regression

Multi-locus marker-QTL LD



$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$$

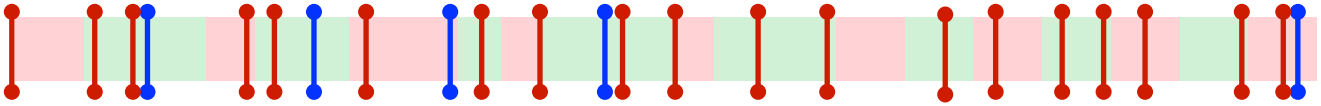
Sample size (i) = n
Number of markers (j) = p

$p >> n$

- Parametric/semi-parametric regression
  - Penalized regression (e.g. LASSO, elastic net)
  - Bayesian regularized regression (e.g. Bayesian SSVS, Bayesian LASSO)
- Non-parametric regression (e.g. neural networks, support vector regression)
  - Penalized regression
  - Bayesian regularized regression

Whole Genome Regression

Multi-locus marker-QTL LD



$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$$

Sample size (i) = n
Number of markers (j) = p

$p >> n$

- Parametric/semi-parametric regression
  - Penalized regression (e.g. LASSO, elastic net)
  - Bayesian regularized regression (e.g. Bayesian SSVS, Bayesian LASSO)
- Non-parametric regression (e.g. neural networks, support vector regression)
  - Penalized regression
  - Bayesian regularized regression

WGR in genetics

- In plant and animal breeding:
  - Genomic selection[1,2]
- In humans:
  - Estimation of genomic heritability[3,4] and genomic correlations
  - Detecting marker-phenotype associations[5,6]
  - Prediction of disease progression[7]
    - Integrate genetic data with other "omics" [8-10]

[1] Heffner E. L., et al., 2009. Crop Sci. **49**: 1.
[2] VanRaden P. M et al., J. Dairy Sci. **92**: 16–24
[3] Makowsky., et al, 2011. PLoS. Genet. **7(4)**
[4] Yang J., et al, 2010. Nat. Genet. **42**: 565–9.
[5] Wu T., et al, 2009. Bioinformatics. **25(6)**: 714–721.
[6] Li J., et al, 2011. Bioinformatics. **27(4)**: 516–523.
[7] de los Campos., et al, 2013. Genetics. **193(2)**: 327–345.
[8] Fridley B., et al, 2012. Genet Epidemiol **36(4)**: 352-359
[9] Kirk. P., et al, 2012. Systems Biology **28(24)**: 3290-3297
[10] Wheeler J., et al, 2014. Genet Epidemiol **38**: 402–415

Some unaddressed questions…

- WGR methods were developed and applied with reference to homogeneous populations. However,
  - Both breeding and human populations exhibit structure and admixture.

- Even among homogeneous populations, WGR methods have not been applied to study sex differences
  - Many studies have shown there is a genetic basis for sex differences in human (e.g. behavioral[1]) traits.

[1] Trabzuni, et al., 2013. Nature Comm. **4**: 2771.

Problem statements

⇒ Extension of WGR to accommodate genetic heterogeneity in structured populations

⇒ Extension of WGR to accommodate sex differences in homogeneous populations

Problem statements

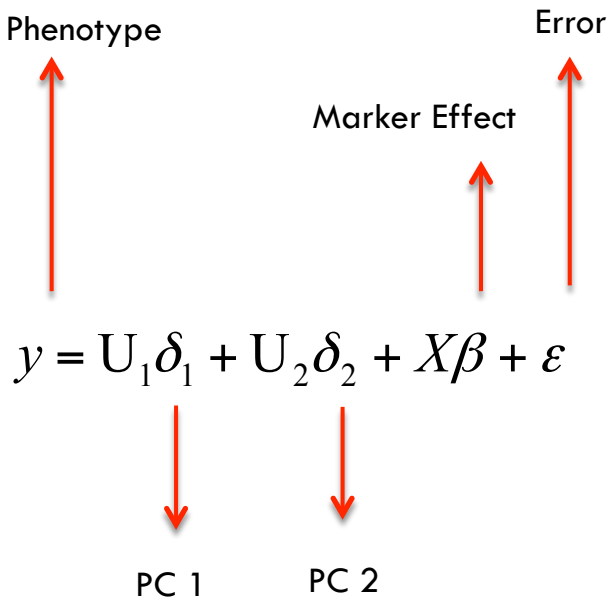
⇒ Extension of WGR to accommodate genetic heterogeneity in structured populations

⇒ Extension of WGR to accommodate sex differences in homogeneous populations

Population structure

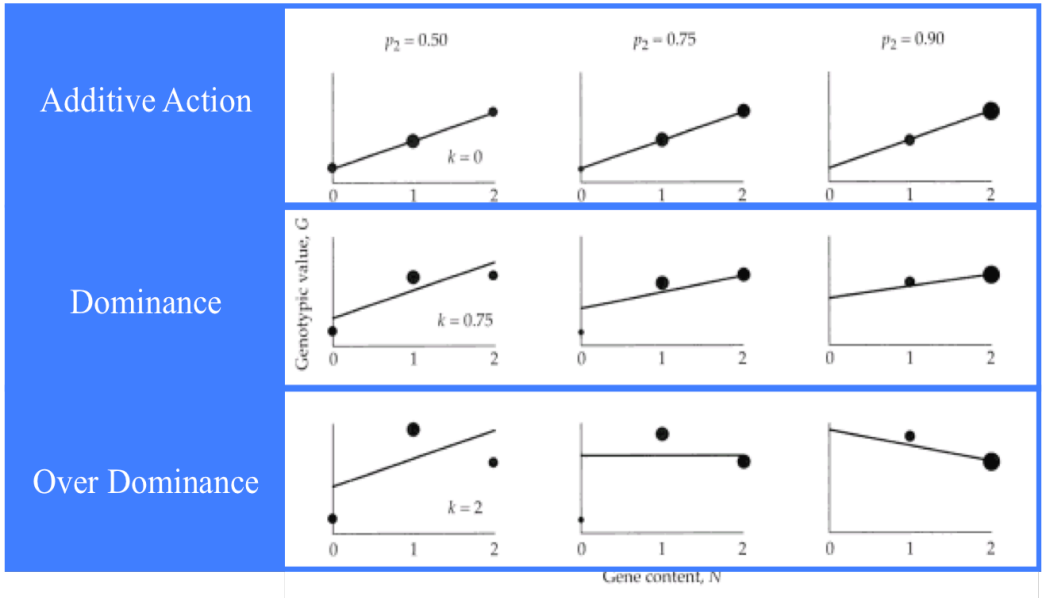
- Natural and artificially selected populations exhibit population structure
- Population differentiation occurred along geographic lines in humans
- Various evolutionary factors shape structure:
  - E.g. drift, selection, migration, population bottlenecks
- Heterogeneous subpopulations show differences in:
  - allele frequencies
  - linkage disequilibrium (LD) patterns
- However, most often,
  - Marker effects are assumed to be homogeneous (E.g. combined analysis[1,2])
  - population structure is treated as a confounder (E.g. PC correction[3])

Standard GWAS



[1]Daetwyler H. D et al., 2010. Anim. Prod. Sci. **50**: 1004–101
[2]Hayes B. J., et al., 2009. Genet. Sel. Evol. **41**: 51.
[3]Janss L., et al., 2012. Genetics **192**: 693–704.

Structure induces "group-specific" effects



Lynch and Walsh (1998, p 68).

Hypothesis:

Structure acts as an "effect modifier" rather than a confounder

We propose an interaction model that estimates average correlations of marker effects

[1]DE LOS CAMPOS G. DE, SORENSEN D., J. Anim. Breed. Genet. **131**: 163–4.

Interaction model

Standard WGR model

$\boldsymbol{b_1 = b_2 = 0}$

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \boldsymbol{b}_0 + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}$$

Interaction Model

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \boldsymbol{b}_0 + \begin{bmatrix} \boldsymbol{X}_1 \\ 0 \end{bmatrix} \boldsymbol{b}_1 + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{X}_2 \end{bmatrix} \boldsymbol{b}_2 + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}$$

$$\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{b}_0 + \boldsymbol{b}_1 \\ \boldsymbol{b}_0 + \boldsymbol{b}_2 \end{pmatrix}$$

Constant across groups

Group-specific deviations

Gaussian or other priors on marker effects

Average correlation of effects

$$Cor(\beta_{1j}, \beta_{2j}) = \frac{\sigma^2_{b_0}}{\sqrt{(\sigma^2_{b_0} + \sigma^2_{b_1}) \times (\sigma^2_{b_0} + \sigma^2_{b_2})}}$$

Data

Data	Source	Sample Size	#SNPs	#Groups	Traits/Environments
Wheat	CIMMYT	599	1,279	2	4 (E)
Pig	PIC	3,534	50,436	3	3 (T)
Humans	ARIC	8,228	800,000	2	4 (T)

Data

Data	Source	Sample Size	#SNPs	#Groups	Traits/Environments
Wheat	CIMMYT	599	1,279	2	4 (E)
Pig	PIC	3,534	50,436	3	3 (T)
Humans	ARIC	8,228	800,000	2	4 (T)

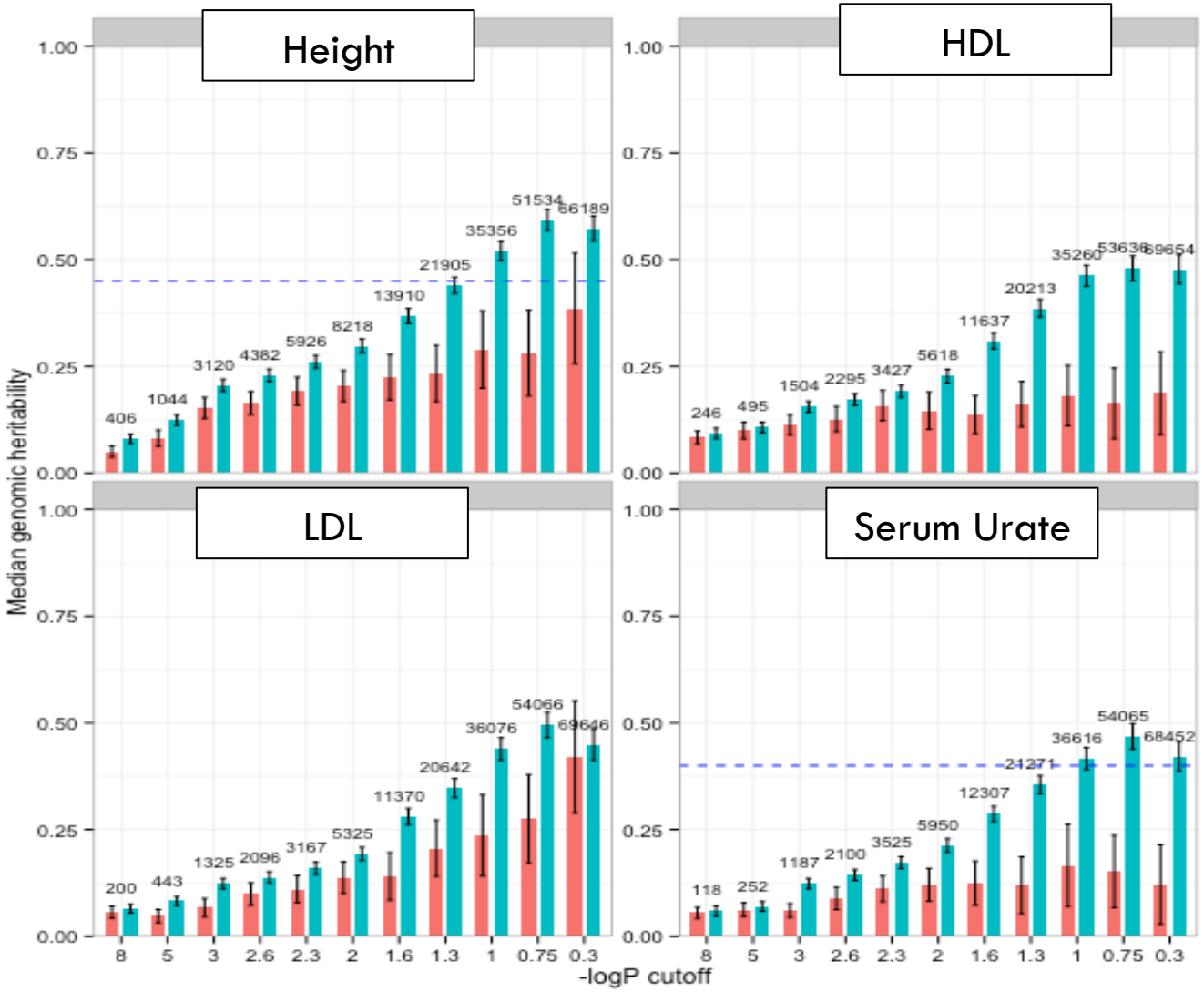
For Humans:

- 4 traits : Height (cm), High Density Lipoprotein (mg/dL), Low Density Lipoprotein (mg/dL), Serum Urate (mg/dL)
- 2 groups: 6627 Caucasians and 1601 African Americans
- ARIC is the Atherosclerosis Risk in Communities dataset
- Standard quality control procedures were applied

All models fit using a modified version of BGLR package[1] in R

[1]de los Campos, G., 2014. BGLR: Bayesian Generalized Linear Regression.

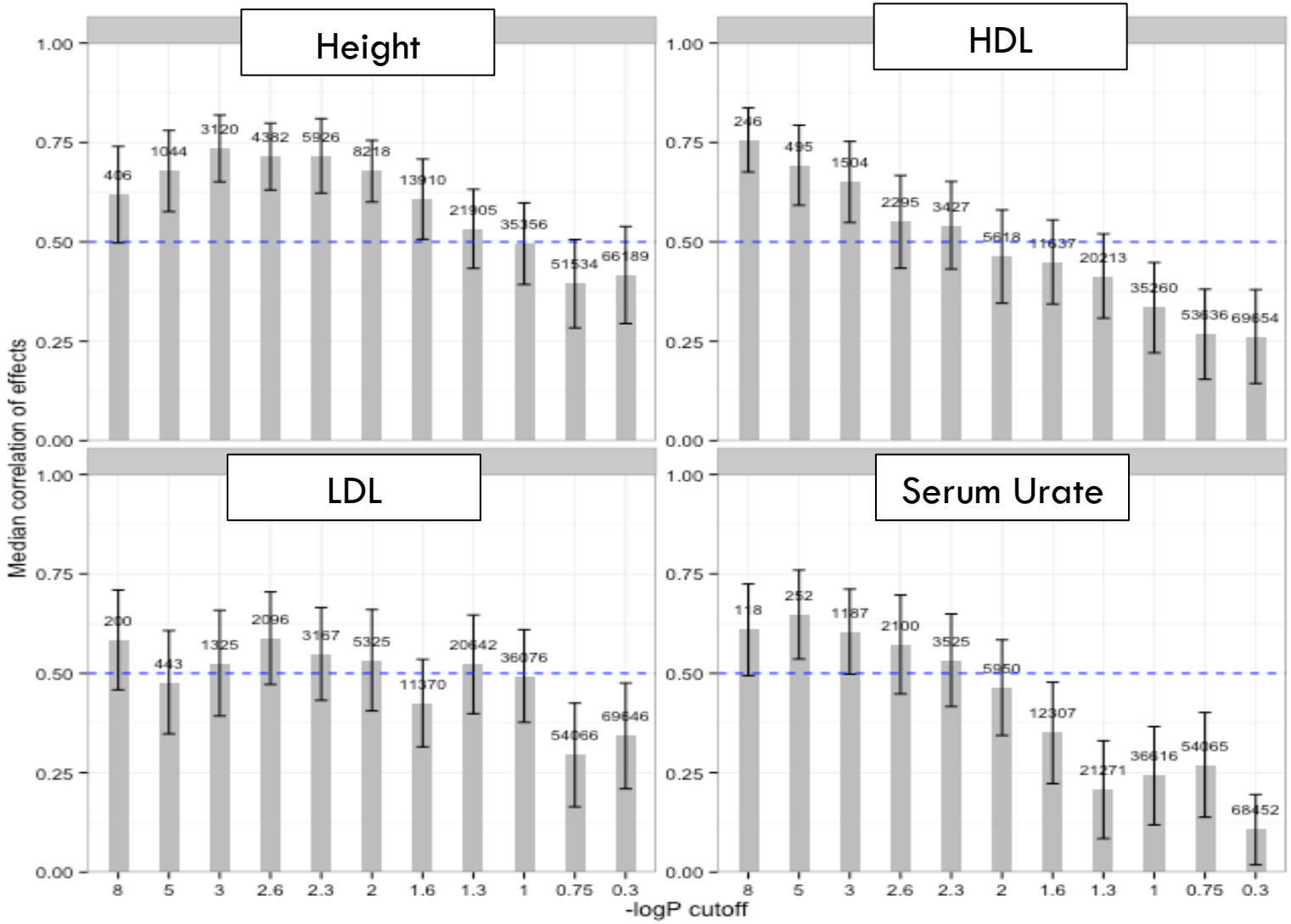
Estimated genomic heritability : ARIC



Markers chosen from GWAS consortia

- GIANT
- GLGC
- GUGC

Estimated effect correlation : ARIC



Markers chosen from GWAS consortia

Genetic heterogeneity varies by:

- Trait
- Effect size

What is driving these results?

- Genomic heritability higher for whites than blacks
  - GWAS conducted primarily on Caucasians
  - LD is stronger in whites
  - Lower trait heritability for blacks (e.g. serum lipids)

What is driving these results?

- Effect correlation varies with trait and effect size
- Large effect genes shared between groups
- Small effect genes different between groups
  - Epistasis responsible for small additive effects[1]
  - Thus effects will depend more on genetic background leading to low effect correlation

[1] MACKAY AND MOORE, Genome Medicine 2014. 6:42

Problem statements

⇒ Extension of WGR to accommodate genetic heterogeneity in structured populations

⇒ Extension of WGR to accommodate sex differences in homogeneous populations

Sex differences in humans

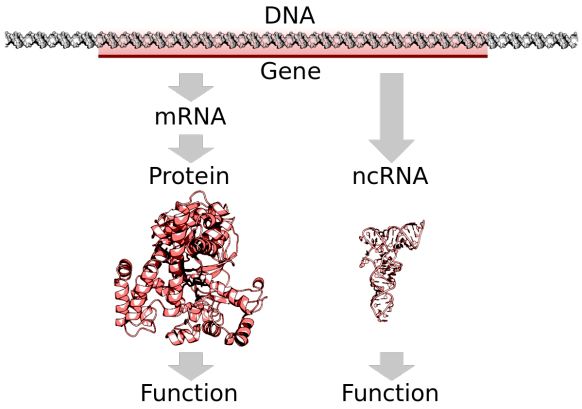
- Sex differences in:
  - General anatomy
  - Cognitive functions/brain anatomy
  - Pain sensitivity

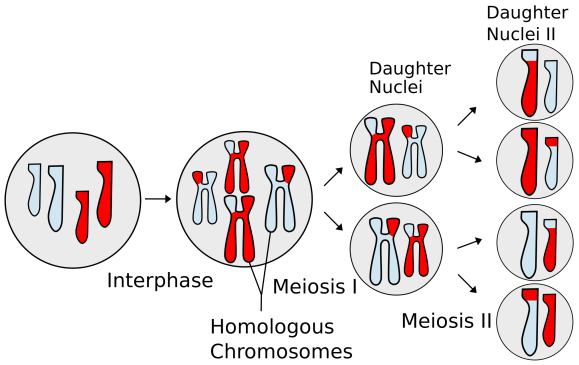
- Males die at a greater rate at every life stage than females (except at oldest ages)
- What protective factors exist in females?

http://www.ncbi.nlm.nih.gov/books/NBK53393/

Genetic basis for sex differences

- Mammals (all vertebrates) share 20,000 genes
- Base sequences vary slightly from person to person within a species
- All humans inherit two copies of each chromosome
  - Except the sex chromosomes!





Source: http://www.iflscience.com/health-and-medicine/

Genetic basis for sex differences

- Male is *heterogametic* (XY)
- Female is *homogametic* (XX)
- X > 1000 genes
- Y ~ 45 genes
  - *SRY* genes on Y chromosome determines gender [1]
  - 24 genes in Y are shared by X too[2]!
- X-chromosome inactivation equalizes X-gene dosage between men and women
  - Some genes escape inactivation[3]

Genetic differences between men and women lie in sex chromosomes!

- Male-specific genes : 18[2]
- Escapee X-genes : 150[3]

Approximately 160 genes on chr. X and Y different between men and women

[1]CRAIG I et al., 2010. Annals Human Genet. **68(3) 269: 213**
[2]WILLIAMS S., 2014. Science.
[3]BERLETCH J et al., 2010. Genome Biol. **11(6) : 213**
[4] http://www.iflscience.com/health-and-medicine/

Genetic basis for sex differences

- Not all of these will have the same (or any) effect
- *SRY* gene kick-starts a cascade of genes that are **not on** sex chromosomes[1,2]
- *Downstream effects of SRY are important!*
  - Androgens turn on hundreds or thousands of genes that determine several male functions[3]

Approximately 800 genes different between the sexes

Other genes might have similar cascading effects as well

[1]CRAIG I et al., 2010. Genome Biol. **11(6) : 213**
[2]WILHELM, D et al., 2007. Physiol Rev. **87(1) : 28**
[3]ARNOLD 2014. Trends in Genetics. **28(2) : 55-61**

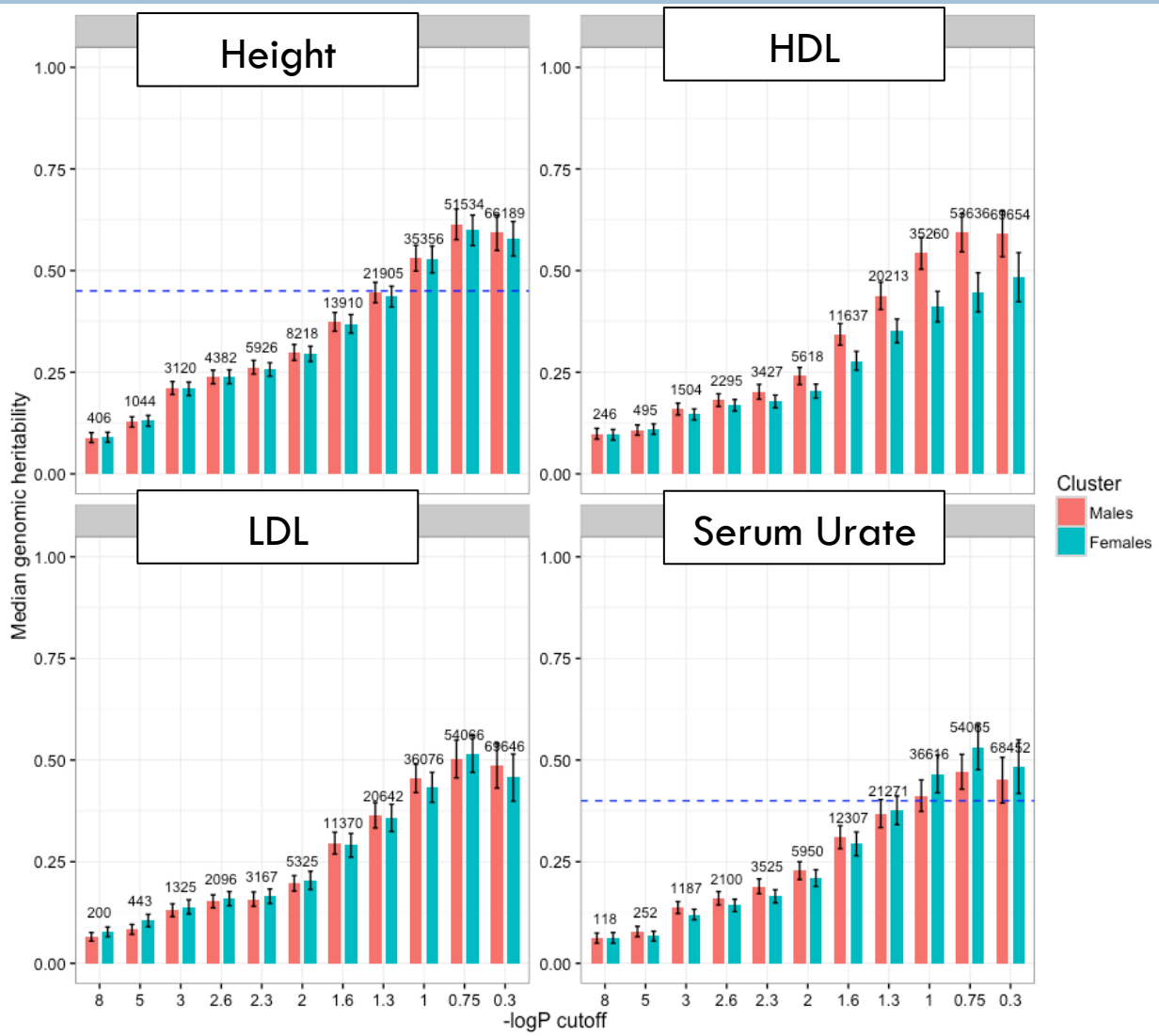
Which genes underlie sex differences?

Hypothesis:

Genes from across the genome will underlie sex differences for different human traits

We propose to estimate and test for effect correlation between sexes on per gene basis using score tests

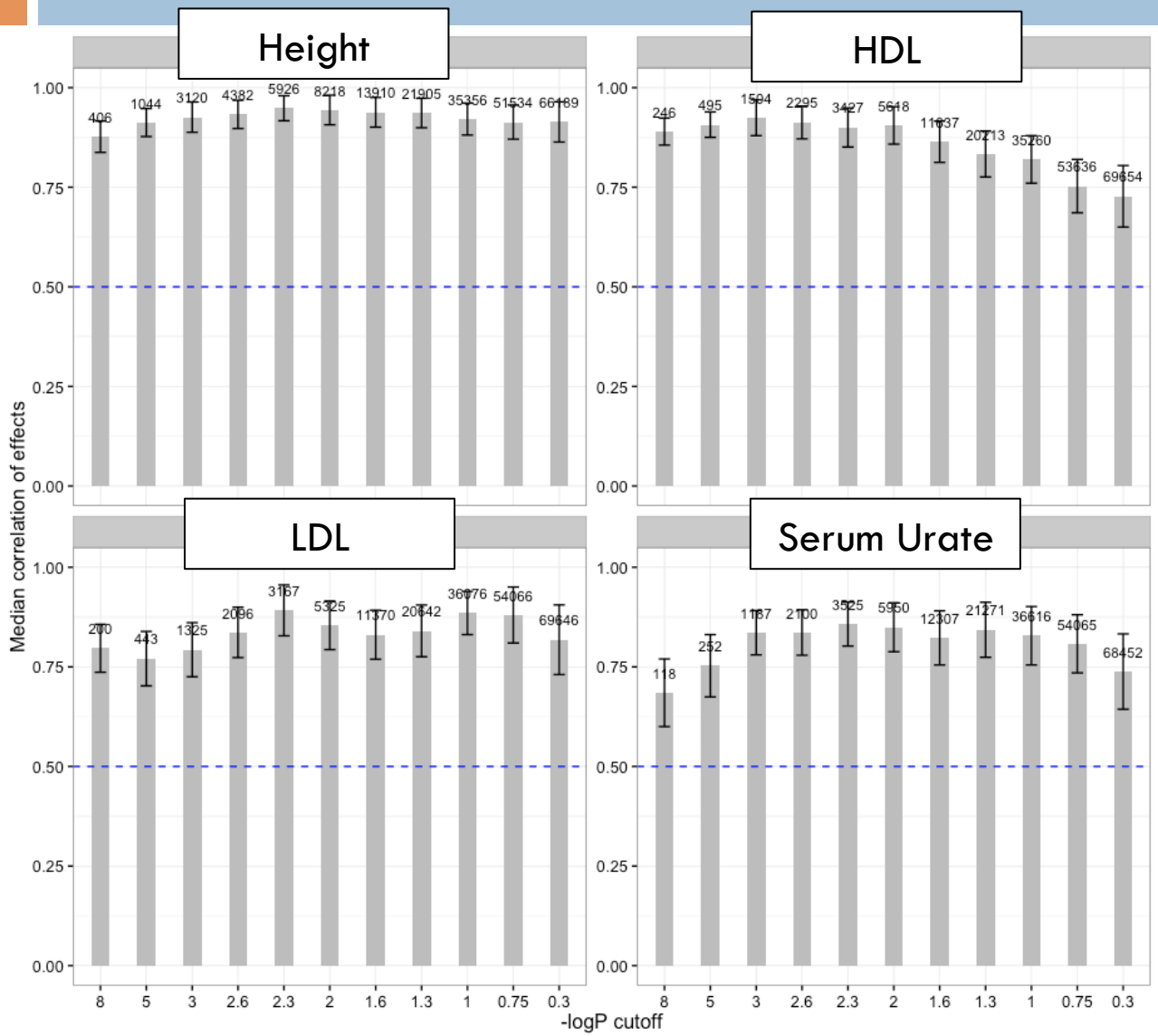
Estimated genomic heritability : ARIC



- 3114 males
- 3513 females

Heritability is constant between sexes

Estimated effect correlation : ARIC



Uniformly high correlations

Relatively low correlations for large effects : serum urate

Sex differences in Caucasians : ARIC

- 3114 males
- 3513 females
- 4915 genes

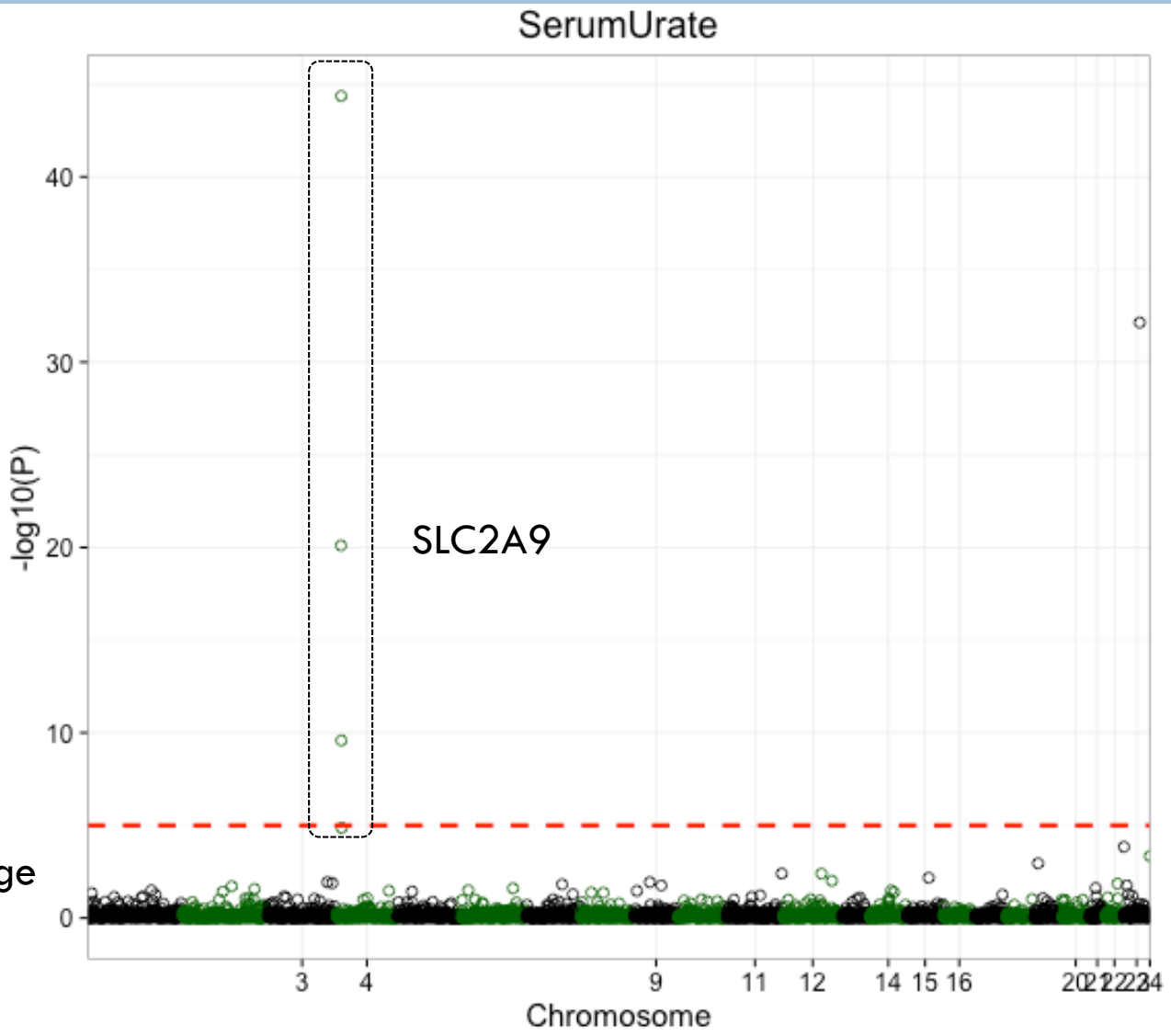
Models fit on a gene-by-gene basis using score tests

$$H_0 : \sigma^2_{b_1} = 0$$

$$H_1 : \sigma^2_{b_1} \neq 0$$

Models fit using SKAT2 package

https://github.com/lian0090/SKAT2



Summary

⇒ Model interactions between markers and genetic groups
⇒ Estimate/test average correlation of effects between groups

- ⇒ Incorporate genetic heterogeneity in structured populations
  - ⇒ Genetic heterogeneity between blacks and whites varies
    - ⇒ by trait and effect-size
  - ⇒ Don't need "whole genome" information to estimate genomic heritability

- ⇒ Incorporate sex differences in humans
  - ⇒ High effect correlations across all cutoffs
  - ⇒ Some genes show significant sex differences, which is purely driven by GxE (structure has no role here)

Acknowledgements

Michigan State University

- Dr. Gustavo de los Campos (Ph.D. advisor)

- Dr. Ana Ines Vazquez

University of Alabama at Birmingham

- Dr. Nengjun Yi
- Dr. Nianjun Liu
- Dr. Sadeep Shrestha
- Dr. Marguerite Irvin
- Dr. David Redden

Thank you!

Questions ?