APPLICATIONS OF VARIABLE SELECTION AND SHRINKAGE IN STRUCTURED HUMAN POPULATIONS USING WHOLE GENOME REGRESSION MODELS

Yogasudha Veturi

Dissertation Proposal

Department of Biostatistics

June 5, 2015

THE UNIVERSITY OF ALABAMA AT BIRMINGHAM
Knowledge that will change your world

I) BACKGROUND

- Incorporating Genetic Heterogeneity
- Incorporating LD information into variable selection models
- Computationally efficient implementation of group variable selection and shrinkage

I) AIMS AND APPROACHES

- Data sets
- Chapter 1 - Whole Genome Regression with Data from Heterogeneous Populations
  - Preliminary results

    Aim 1.3 (1) Stratified and bi-cluster analyses on MESA
- Chapter 2 - Implementation of Group Variable Selection and Shrinkage in Whole Genome Regressions
- Chapter 3 - Computationally Efficient Implementation of Group Variable Selection using Variational Algorithms

I) BACKGROUND

- Incorporating Genetic Heterogeneity
- Incorporating LD information into variable selection models
- Computationally efficient implementation of group variable selection and shrinkage

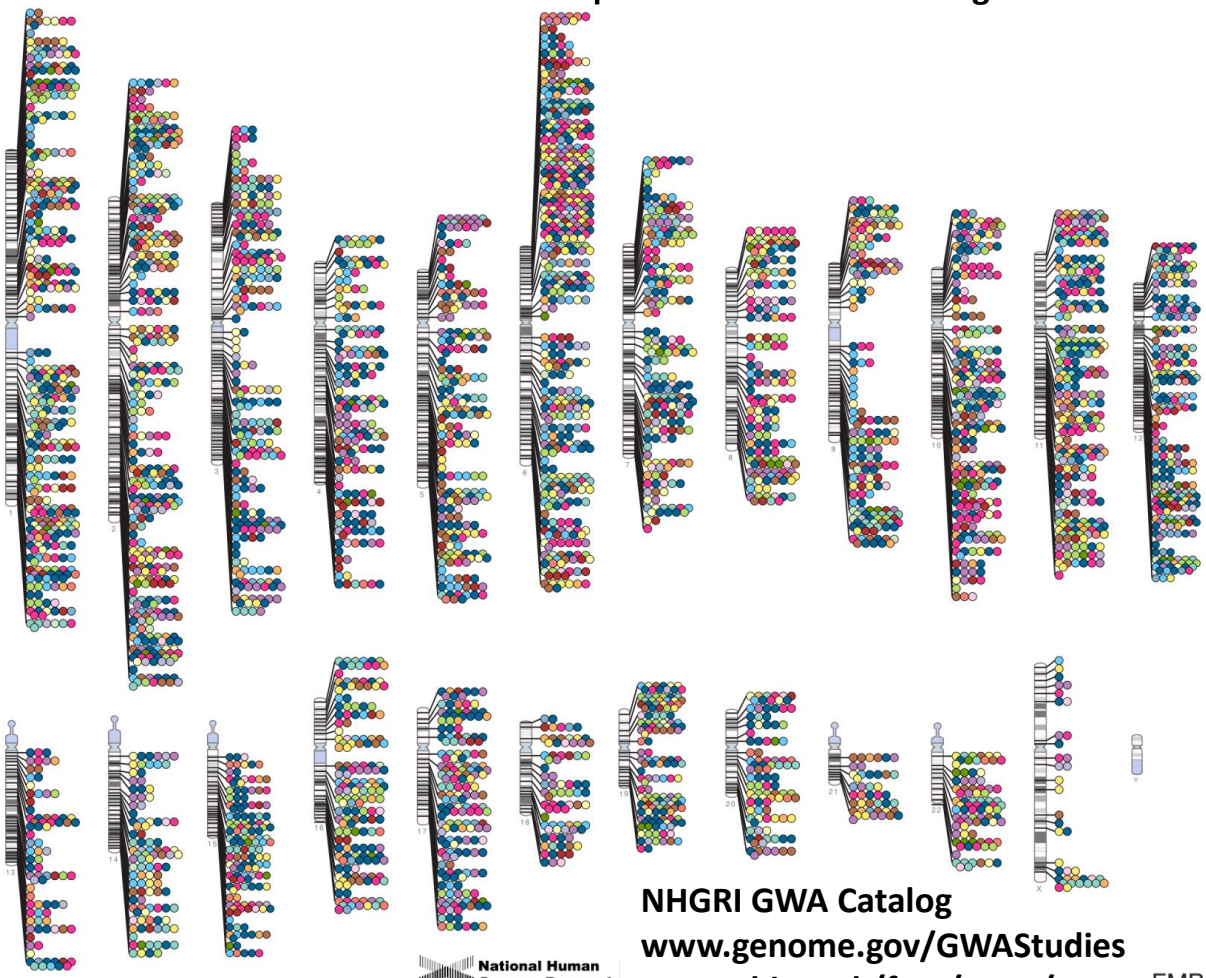
II) AIMS AND APPROACHES

- Data sets
- Chapter 1 - Whole Genome Regression with Data from Heterogeneous Populations
  - Preliminary results

    Aim 1.3 (1) Stratified and bi-cluster analyses on MESA
- Chapter 2 - Implementation of Group Variable Selection and Shrinkage in Whole Genome Regressions
- Chapter 3 - Computationally Efficient Implementation of Group Variable Selection using Variational Algorithms

GWAS

Published Genome-Wide Associations through 12/2013
Published GWA at $p \leq 5X10^{-8}$ 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

Missing heritability – Meta analyses

Consortium (acronym)	Phenotype (or phenotypes)	Publicly available genome-wide data?	Website
AMD	Age-related macular degeneration	Yes, accessible through the website	http://www.sph.umich.edu/csg/abecasis/public/amdgene2012
BCAC	Breast cancer	No	http://ccge.medschl.cam.ac.uk/consortia/bcac
CHARGE	Heart disease and ageing	No	http://web.chargeconsortium.com
GEFOS	Osteoporosis	Yes, accessible through the website	http://www.gefos.org
GIANT	Anthropometric traits	Yes, accessible through the website	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium
GLGC	TC, HDL-C, LDL-C, triglycerides	Yes, accessible through the website	http://www.sph.umich.edu/csg/abecasis/public/lipids2010
IIBDGC	Inflammatory bowel disease	Yes, accessible through the website	http://www.ibdgenetics.org
IMSGC	Multiple sclerosis	Yes, accessible through the website	https://www.imsgenetics.org/
ISC	Schizophrenia	No	http://pngu.mgh.harvard.edu/isc
MAGIC	Glycaemic traits	Yes, accessible through the website	http://www.magicinvestigators.org
NARAC-III	Rheumatoid arthritis	No	http://www.naracstudy.org/NaracStudy/narac.aspx
TREATOA	Osteoarthritis	Yes, accessible through the website	http://treatoa.eu
WTCCC	Various phenotypes	Yes, accessible through the website	http://www.wtccc.org.uk

HDL-C: high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; TC, total cholesterol.

Nature Reviews Genetics **14**, 379–389 2013)

Large consortia provide increased sample sizes

- Detected many more significant variants
- … *yet*, very small proportion of genetic variance explained by GWAS-significant variants

Missing heritability– Whole Genome Regression

ANALYSIS

nature genetics

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

Yang et al., Nature Genetics, July 2010

Human height: $h^2 = 80\%$

$h_g^2 \approx 45\%$

Uses all the available markers in the panel in a single modeling framework

Whole Genome Regression

Multi-locus marker-QTL LD

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

Sample size = n
Number of markers = p

- *p* >> *n*[1] ---- Need penalized or Bayesian regularized regression models
- These apply either:
  - Shrinkage (e.g. G-BLUP[2], Bayesian LASSO[3])
  - Variable selection and shrinkage (e.g. LASSO[4], elastic net[5])
- Popular in plant[6] and animal breeding[2] for genomic selection
- Recently adopted in complex human traits for:
  - Prediction[7,8]
  - Identification of marker-phenotype associations[9,10]
  - Estimation of the extent of missing heritability [11]

[1] Meuwissen. T.H.E. et al., 2001. Genetics **92**: 16–24
[2] VanRaden P. M et al., J. Dairy Sci. **92**: 16–24
[3] Park T., Casella G., 2008. J. Am. Stat. Assoc. **103**: 681–686.
[4] Tibshirani R., 1996 J. R. Stat. Soc. Ser. B **58**: 267 – 288.
[5] Zou H., Hastie T., 2005. J. R. Stat. Soc. Ser. B **67**: 301 – 320
[6] Heffner E. L., et al., 2009. Crop Sci. **49**: 1.
[7] Makowsky., et al, 2011. PLoS. Genet. **7(4)**
[8] de los Campos., et al, 2013. Genetics. **193(2)**: 327–345.
[9] Wu T., et al, 2009. Bioinformatics. **25(6)**: 714–721.
[10] Li J., et al, 2011. Bioinformatics. **27(4)**: 516–523.
[11] Yang J., et al, 2010. Nat. Genet. **42**: 565–9.

The three main topics of the proposal

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

$\Rightarrow$ Extend WGR to accommodate genetic heterogeneity

- Most of the available methods perform variable selection at the level of individual marker effects. *This approach does not incorporate LD patterns into the model.*

$\Rightarrow$ Develop Bayesian variable selection/shrinkage methods that incorporate LD information

- *The application of WGR methods (especially those that induce variable selection) is computationally intensive*

$\Rightarrow$ Develop Variational Bayes algorithm to implement the described models

I) BACKGROUND

- Incorporating Genetic Heterogeneity
- Incorporating LD information into variable selection models
- Computationally efficient implementation of group variable selection and shrinkage

II) AIMS AND APPROACHES

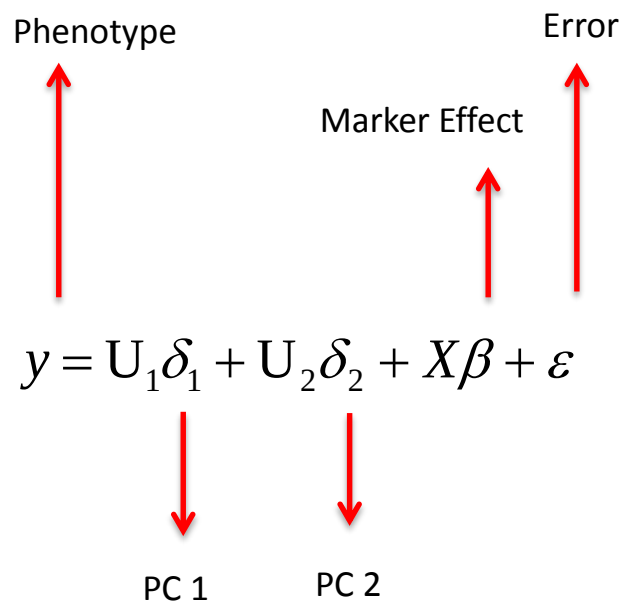
- Data sets
- Chapter 1 - Whole Genome Regression with Data from Heterogeneous Populations
  - Preliminary results

    Aim 1.3 (1) Stratified and bi-cluster analyses on MESA
- Chapter 2 - Implementation of Group Variable Selection and Shrinkage in Whole Genome Regressions
- Chapter 3 - Computationally Efficient Implementation of Group Variable Selection using Variational Algorithms

Population structure

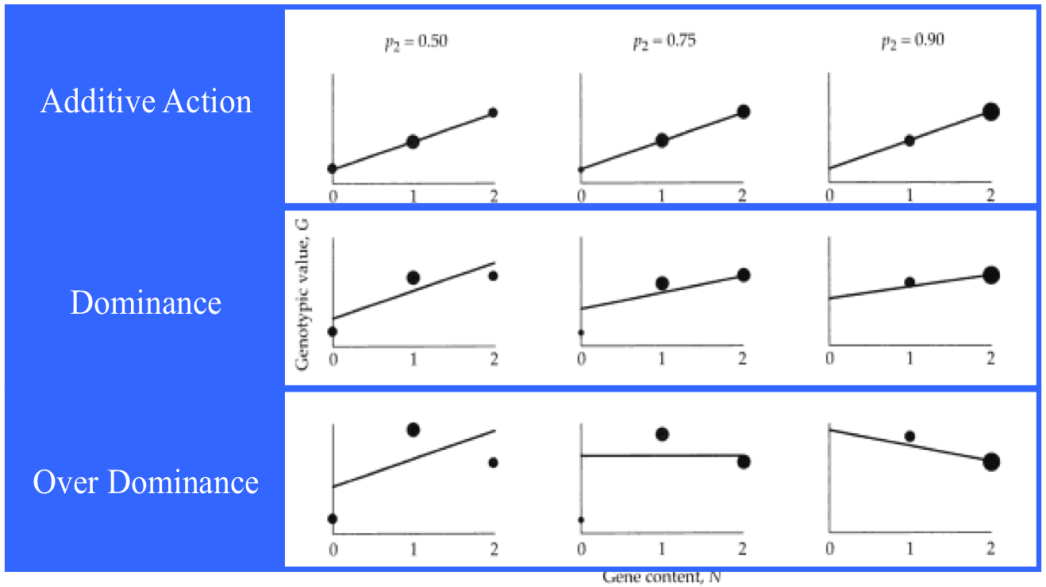
- Natural and artificially selected populations exhibit population structure
- Population differentiation occurred along geographic lines in humans
- Various evolutionary factors shape structure:
  - E.g. drift, selection, migration, population bottlenecks
- Heterogeneous subpopulations show differences 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 (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.

Influence of allele frequencies on additive effects



Lynch and Walsh (1998, p 68).

Hypothesis 1:

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

We propose the interaction model, an easy to apply model that gives:

(a) cluster-specific estimates of marker effects and genomic heritability
(b) between-cluster correlations of marker effects

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

I) BACKGROUND

- Incorporating Genetic Heterogeneity
- Incorporating LD information into variable selection models
- Computationally efficient implementation of group variable selection and shrinkage

II) AIMS AND APPROACHES

- Data sets
- Chapter 1 - Whole Genome Regression with Data from Heterogeneous Populations
  - Preliminary results

    Aim 1.3 (1) Stratified and bi-cluster analyses on MESA
- Chapter 2 - Implementation of Group Variable Selection and Shrinkage in Whole Genome Regressions
- Chapter 3 - Computationally Efficient Implementation of Group Variable Selection using Variational Algorithms

Variable selection Vs. Shrinkage

- Regularized regression methods use either shrinkage or variable selection or a combination of both.

- Shrinkage: marker effects are shrunk toward zero (this reduces the variance of the estimator)
  - E.g. Ridge regression, Bayesian Ridge Regression or G-BLUP

- Variable selection: a subset of predictors are included the rest are "zeroed-out"
  - E.g. subset selection

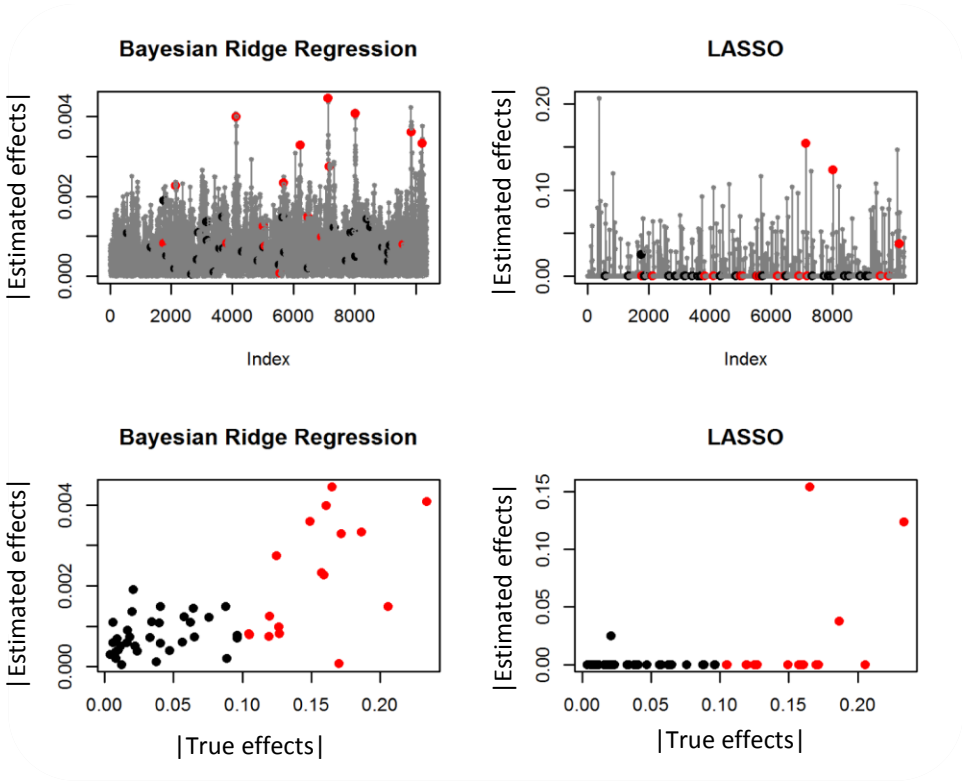
- Many methods combine both variable selection and shrinkage
  - E.g. LASSO, elastic net

- Shrinkage works better when all predictors have small effects and are strongly correlated.

- Variable selection is most effective when only a few predictors have effects and are weakly correlated.

- In human genomes, LD occurs in "blocks", yet variable selection has been applied at the <u>individual marker level</u>

Variable selection or Shrinkage?



Parameters

- *Mice data set*
- $nQTL = 50$
- $n = 487$ (simulated phenotypes)
- $h^2 = 0.4$
- $p = 10{,}346$
- LASSO: *grpreg*[1]
- BRR: *BGLR*[2]

Red : Large effect QTL |true effects| > 0.1
Black: Small effect QTL |true effects| < 0.1
Gray: Non-causal variants

Shrinkage

- BRR works better with correlated predictors
  - long stretches of LD in the mouse genome
- …*but* over-shrinks the effect estimates

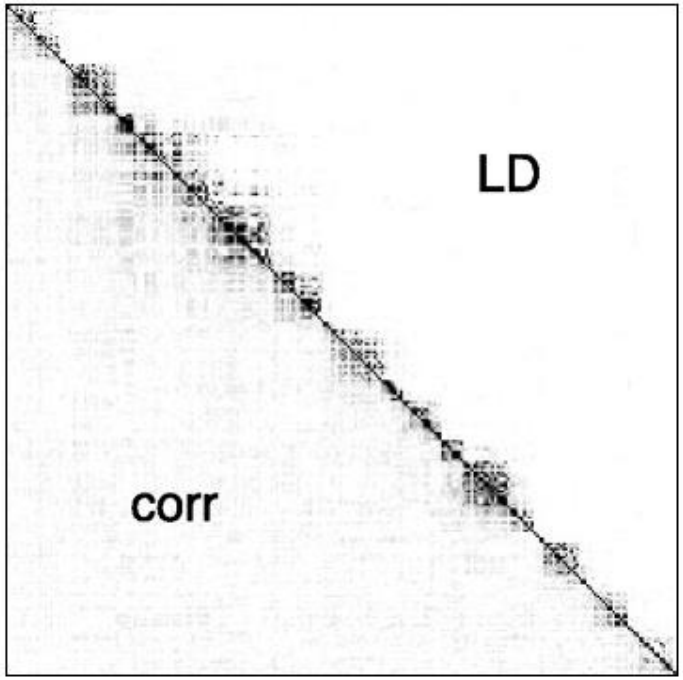
Variable selection and shrinkage

- LASSO gives bigger effect sizes
- …*but* misses over 90% of QTL because of correlated predictors

[1]Breheny P., Huang J., 2015 Stat. Comput. **25**: 173–187.
[2]de los Campos, G., Rodriguez P., 2014. BGLR: Bayesian Generalized Linear Regression.

Group variable selection and shrinkage

- LD occurs in blocks in the human genome; strong LD within blocks, weak LD between blocks



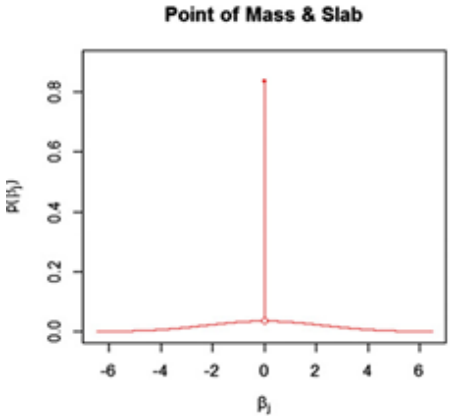
DEHMAN A., AMBROISE C., NEUVIAL P., 2015. BMC Bioinformatics **16**: 148

Hypothesis 2:

LD patterns in distantly related humans can be effectively incorporated in WGR using:

- variable selection on LD blocks (uncorrelated predictors)
- shrinkage within LD blocks (correlated predictors)

We propose a point-of-mass-at-zero on LD blocks and slab prior within LD blocks



I) BACKGROUND

- Incorporating Genetic Heterogeneity
- Incorporating LD information into variable selection models
- Computationally efficient implementation of group variable selection and shrinkage

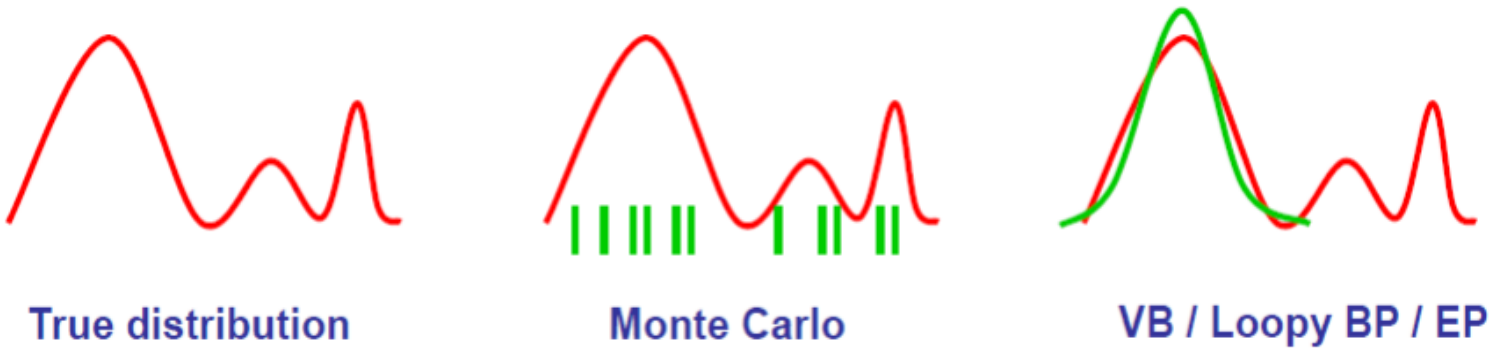
II) AIMS AND APPROACHES

- Data sets
- Chapter 1 - Whole Genome Regression with Data from Heterogeneous Populations
  - Preliminary results
    - Aim 1.3 (1) Stratified and bi-cluster analyses on MESA
- Chapter 2 - Implementation of Group Variable Selection and Shrinkage in Whole Genome Regressions
- Chapter 3 - Computationally Efficient Implementation of Group Variable Selection using Variational Algorithms

Bayesian *approximate* inference

- Limited application of variable selection methods in human data
- Markov Chain Monte Carlo (MCMC) is computationally expensive for high-dimensional data

- Variational Bayes uses **approximate inference**[1,2]



MCMC : approximate estimate of exact posterior
VB: exact estimate of approximate posterior

[1]Attias H., 2000 A Variational Bayesian Framework for Graphical Models
[2]Jaakkola T. S., 2001 Tutorial on Variational Approximation Methods

Variational Bayes

- Minimizes the Kullback-Leibler divergence between the true and approximating posteriors.
- The approximating distribution is obtained using an update and optimization algorithm similar to the EM algorithm
- Used when it is infeasible to compute the exact posterior
- Much faster than MCMC but can result in biased estimates
- Has been used to improve computational efficiency in genetics studies, for
  - GWAS[1-3]
  - Estimation of QTL effects with epistasis[4]
  - Heritability estimation with the G-BLUP[5]
  - Multi-trait analysis and prediction[6]
  - Efficient inference of population structure[7]

[1]Logsdon *et al.* 2010. BMC Bioinformatics **11**: 58
[2]Carbonetto and Stephens 2012. Bayesian Anal. **7**: 73–108
[3]Loh *et al.* 2015. Nat. Genet. **47**: 284–290
[4]Li and Sillanpää 2012. Genetics **190**: 231–49
[5]Arakawa 2014. *10th World Congress on Genetics Applied to Livestock Production*
[6]Hayashi and Iwata 2013. BMC Bioinformatics **14**: 34
[7]Raj *et al.* 2014. Genetics **197**: 573–89

Group variable selection with VB

- Variable selection methods implemented using VB result in poor quality of the approximation to the posterior mean when predictors are correlated[1]

- This can be addressed by applying:
  - variable selection at the level of uncorrelated LD blocks
  - shrinkage within blocks

Hypothesis 3:

VB implementation of **group variable selection and shrinkage on LD blocks** can lead to:

1) Better approximation to the posterior
2) Much improved computational efficiency relative to MCMC

[1]Carbonetto and Stephens 2012. Bayesian Anal. **7**: 73–108

Can we extend WGR models to analyze genetic heterogeneity in human populations?

- CHAPTER 1 - WHOLE GENOME REGRESSION WITH DATA FROM HETEROGENEOUS POPULATIONS

Can we develop WGR models that induce a combination of variable selection and shrinkage in LD blocks to more effectively incorporate LD patterns in human populations?

- CHAPTER 2 - IMPLEMENTATION OF GROUP VARIABLE SELECTION AND SHRINKAGE IN WHOLE GENOME REGRESSIONS

Can we apply algorithms alternative to MCMC to improve the computational efficiency of these methods?

- CHAPTER 3 - COMPUTATIONALLY EFFICIENT IMPLEMENTATION OF GROUP VARIABLE SELECTION USING VARIATIONAL ALGORITHMS

I) BACKGROUND

- Incorporating Genetic Heterogeneity
- Incorporating LD information into variable selection models
- Computationally efficient implementation of group variable selection and shrinkage

II) AIMS AND APPROACHES

- Data sets
- Chapter 1 - Whole Genome Regression with Data from Heterogeneous Populations
  - Preliminary results

    Aim 1.3 (1) Stratified and bi-cluster analyses on MESA
- Chapter 2 - Implementation of Group Variable Selection and Shrinkage in Whole Genome Regressions
- Chapter 3 - Computationally Efficient Implementation of Group Variable Selection using Variational Algorithms

Data

Traits: Height, HDL, LDL Cholesterol

Data sets	Sample size	Genotyping Platforms	# SNPs
1. Multi Ethnic Study for Atherosclerosis (MESA)	1. 8305 (Whites, Blacks, Asians, Hispanics)	1. Affymetrix 6.0 SNP array	1. ~900,000
2. British Cohort 1958 (BC 58)	2. ~3000 (Whites)	2. Illumina 1.2M	2. ~1 million
3. UNAM/INCMNSZ Diabetes Study (UIDS)	3. ~2067 (Mexicans)	3. IlluminaOMNI 2.5 array	3. 1.38 million

Data: MESA

Quality Control (QC)

- Removed monomorphic SNPs
- Removed SNPs with minor allele frequency < 0.05
- Removed genotypes > 5% missing values
- Retained only distantly related individuals.

No. of records / group, by trait after QC

Group	Height	HDL	LDL
Whites	2,032	1,923	1,904
Blacks	978	889	887
Chinese	470	499	492
Mexican	305	330	326
Total	3785	3641	3609

Phenotypes were adjusted for the effect of race, sex, and site

I) BACKGROUND

- Incorporating Genetic Heterogeneity
- Incorporating LD information into variable selection models
- Computationally efficient implementation of group variable selection and shrinkage

II) AIMS AND APPROACHES

- Data sets
- Chapter 1 - Whole Genome Regression with Data from Heterogeneous Populations
  - Preliminary results

    Aim 1.3 (1) Stratified and bi-cluster analyses on MESA
- Chapter 2 - Implementation of Group Variable Selection and Shrinkage in Whole Genome Regressions
- Chapter 3 - Computationally Efficient Implementation of Group Variable Selection using Variational Algorithms

Chapter Aims

Aim 1.1. To develop WGR models that can incorporate population structure as a potential effect-modifier

Aim 1.2. To design and conduct simulations to assess the finite sample statistical properties of estimates derived from the proposed model

Aim 1.3. To apply the proposed method to the analysis of real data

Aim 1.1 – Interaction model

Standard WGR model

$$\boldsymbol{b}_1 = \boldsymbol{b}_2 = \boldsymbol{0}$$

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

Stratified Model

$$\boldsymbol{b}_0 = \boldsymbol{0}$$

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

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 clusters

Cluster-specific deviations

Gaussian prior on marker effects:

Interaction G-BLUP

Aim 1.2 – Simulation study

No precedent for marker-effect correlations in the literature

(1) Data generation

i. Real genotypes from MESA (Blacks and Whites)
ii. Simulate phenotypes under different genetic architectures

(2) Model fitting

- Under different genetic architectures, compare bias and MSE in estimates of:
  - cluster-specific genomic heritabilities

$$h_g^2 = \frac{\sigma_g^2}{\sigma_y^2}$$

  - between-cluster marker effect correlations

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

Aim 1.3 – Real data analysis

(1) Analysis using a multi-ethnic cohort (MESA)

i. Stratified analysis
   - G-BLUP in each sub-population

ii. Across-cluster analysis
   - G-BLUP across all sub-populations

iii. Bi-cluster analysis
   - Interaction G-BLUP

cluster-specific genomic heritabilities

between-cluster marker effect correlations

(2) Joint analysis between BC58 and UIDS

i. Merge SNPs from Illumina 1.2M and IlluminaOMNI

ii. Stratified analysis

iii. Bi-cluster analysis

I) BACKGROUND

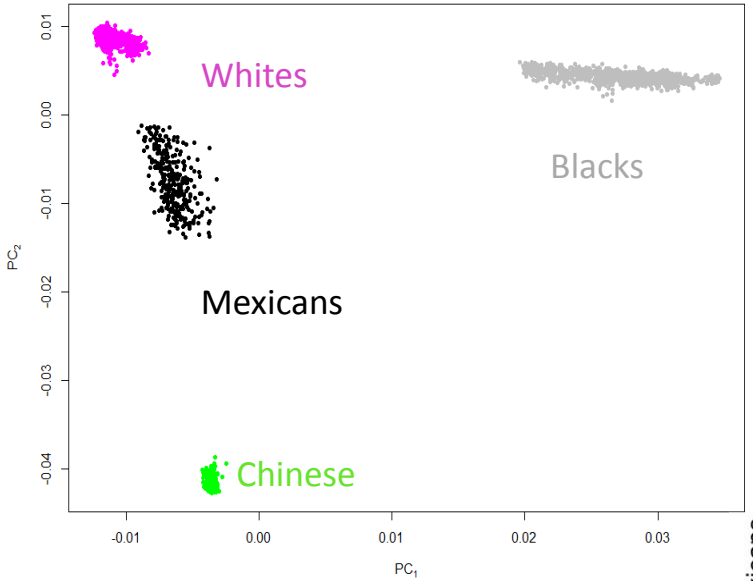
- Incorporating Genetic Heterogeneity
- Incorporating LD information into variable selection models
- Computationally efficient implementation of group variable selection and shrinkage

II) AIMS AND APPROACHES

- Data sets
- Chapter 1 - Whole Genome Regression with Data from Heterogeneous Populations
  - Preliminary results

    Aim 1.3 (1) Stratified and bi-cluster analyses on MESA
- Chapter 2 - Implementation of Group Variable Selection and Shrinkage in Whole Genome Regressions
- Chapter 3 - Computationally Efficient Implementation of Group Variable Selection using Variational Algorithms

Principal Components and Allele Frequencies

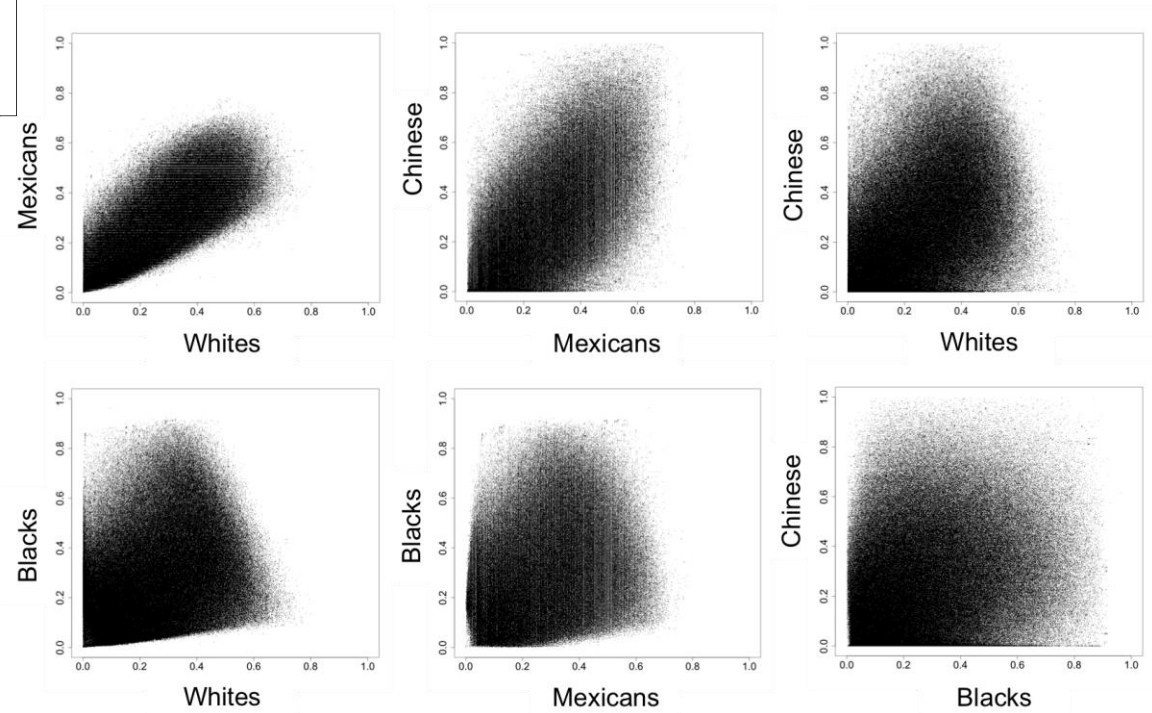


1st 2 Eigenvectors Post-QC

Clear stratification

Variation in allele frequencies by cluster->

Variable patterns of genetic correlation?



Allele frequencies by cluster

Estimates of Genomic Heritability (Stratified analysis)

Height $h^2 \approx 80\%$ $h_g^2 \approx 45\%$

	Mean	SD	Low-95	Up-95
White	0.63	0.10	0.44	0.82
Black	0.63	0.16	0.34	0.91
Mexican	0.47	0.20	0.13	0.82
Chinese	0.68	0.15	0.39	0.92

HDL-Cholesterol $h^2 \approx 40 - 60\%$

	Mean	SD	Low-95	Up-95
White	0.34	0.11	0.14	0.56
Black	0.39	0.16	0.11	0.70
Mexican	0.48	0.18	0.14	0.80
Chinese	0.41	0.18	0.09	0.75

LDL-Cholesterol $h^2 \approx 40\%$

	Mean	SD	Low-95	Up-95
White	0.17	0.07	0.05	0.30
Black	0.61	0.16	0.30	0.90
Mexican	0.53	0.18	0.19	0.87
Chinese	0.46	0.19	0.12	0.80

All analyses were performed using the R package BGLR[2]

Remarks

- Estimates lower than the trait heritability for all three traits
- In accordance with true heritability trends, estimates were:
  - Higher for height
  - Intermediate for HDL
  - Lower for LDL?
- Heritability for height higher than genomic heritability published in previous studies[1]
  - Much larger number of SNPs (~770,000) as opposed to previous studies[1] (~350,000)
- Wider credible intervals because of smaller sample sizes

[1]Yang J., et al. 2010a. Nat. Genet. **42**: 565–9
[2]de los Campos, G., Rodriguez P., 2014. BGLR: Bayesian Generalized Linear Regression.

Estimates of Genomic Correlation (Bi-cluster analysis)

Height

	Mexican	Chinese	Black
White	**0.49** [0.20;0.75]	**0.41** [0.15;0.67]	**0.31** [0.08;0.56]
Mexican		**0.44** [0.14;0.74]	**0.45** [0.16;0.75]
Chinese			**0.48** [0.18;0.77]

HDL-Cholesterol

	Mexican	Chinese	Black
White	**0.40** [0.16;0.65]	**0.42** [0.16;0.67]	**0.38** [0.15;0.65]
Mexican		**0.42** [0.13;0.71]	**0.44** [0.16;0.72]
Chinese			**0.46** [0.18;0.76]

LDL-Cholesterol

	Mexican	Chinese	Black
White	**0.39** [0.14;0.62]	**0.40** [0.16;0.63]	**0.39** [0.16;0.65]
Mexican		**0.47** [0.16;0.77]	**0.48** [0.16;0.78]
Chinese			**0.43** [0.14;0.73]

Remarks

- Intermediate correlations (0.30-0.50)
- Correlations are far away from:
  - zero -> genetic similarities
  - one -> genetic differences
- Wide credibility regions
- Some correlation patterns agree with patterns of allele frequencies (e.g. Whites and Mexicans for height)
- …while others go against the trend (e.g. Blacks and Chinese). This needs to be explored further

All analyses were performed using the R package BGLR[1]

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

I) BACKGROUND

- Incorporating Genetic Heterogeneity
- Incorporating LD information into variable selection models
- Computationally efficient implementation of group variable selection and shrinkage

II) AIMS AND APPROACHES

- Data sets
- Chapter 1 - Whole Genome Regression with Data from Heterogeneous Populations
  - Preliminary results

    Aim 1.3 (1) Stratified and bi-cluster analyses on MESA
- Chapter 2 - Implementation of Group Variable Selection and Shrinkage in Whole Genome Regressions
- Chapter 3 - Computationally Efficient Implementation of Group Variable Selection using Variational Algorithms

Chapter aims

- ***Aim 2.1***. To develop the group WGR model by applying group variable selection and shrinkage, i.e.
  - variable selection at the LD block level
  - shrinkage within LD blocks

- ***Aim 2.2.*** To develop the group interaction model by extending the interaction model (Aim 1.1) using group variable selection and shrinkage on LD blocks

- ***Aim 2.3.*** To apply the group interaction model to real human data
  - To identify genomic regions that have similar and variable effects across sub-populations

Aims 2.1 and 2.2 – WGR on LD blocks

Group WGR model

$b_1 = b_2 = 0$

$$\begin{bmatrix} \boldsymbol{y}_I \\ \boldsymbol{y}_{II} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{Ia} \\ \boldsymbol{X}_{IIa} \end{bmatrix} \boldsymbol{b}_0 + \begin{bmatrix} \boldsymbol{\varepsilon}_I \\ \boldsymbol{\varepsilon}_{II} \end{bmatrix}$$

Group Interaction model

$$\begin{bmatrix} \boldsymbol{y}_I \\ \boldsymbol{y}_{II} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{Ia} \\ \boldsymbol{X}_{IIa} \end{bmatrix} \boldsymbol{b}_0 + \begin{bmatrix} \boldsymbol{X}_{Ib} \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{b}_1 + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{X}_{IIc} \end{bmatrix} \boldsymbol{b}_2 + \begin{bmatrix} \boldsymbol{\varepsilon}_I \\ \boldsymbol{\varepsilon}_{II} \end{bmatrix}$$

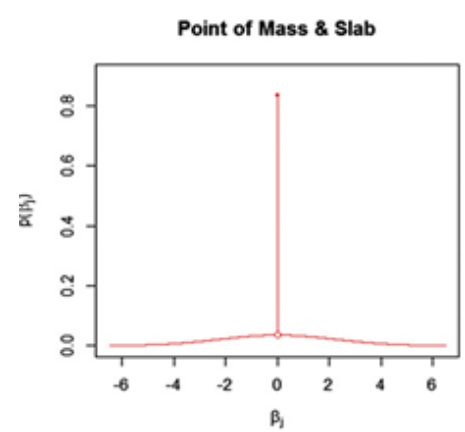
G groups or LD blocks

$$\boldsymbol{X}_{Ia} = [\boldsymbol{X}_{Ia,1}\delta_1 \quad \boldsymbol{X}_{Ia,2}\delta_2 \quad \cdots \quad \boldsymbol{X}_{Ia,G}\delta_G]$$

$$\delta_g = \begin{cases} 1 & g^{th}\text{group is included} \\ 0 & \text{0 otherwise} \end{cases}$$

Point-of-mass-at-zero and slab prior on marker effects

We will develop MCMC algorithms to implement these models in R



Aim 2.3 -- Data analysis

- Phenotypes and Genotypes
  - Adult height, LDL and HDL from MESA (after QC)
- Generating LD blocks
  - Divide markers into relevant LD blocks in each sub-population[1]
- Data Analysis
  - Bi-cluster analysis
    - group-interaction model
  - Compare model fit with that of the interaction model

BARRETT J. C., 2009 Haploview: Cold Spring Harb. Protoc. **2009**: pdb.ip71.

I) BACKGROUND

- Incorporating Genetic Heterogeneity
- Incorporating LD information into variable selection models
- Computationally efficient implementation of group variable selection and shrinkage

II) AIMS AND APPROACHES

- Data sets
- Chapter 1 - Whole Genome Regression with Data from Heterogeneous Populations
  - Preliminary results

    Aim 1.3 (1) Stratified and bi-cluster analyses on MESA
- Chapter 2 - Implementation of Group Variable Selection and Shrinkage in Whole Genome Regressions
- Chapter 3 - Computationally Efficient Implementation of Group Variable Selection using Variational Algorithms

Chapter aims

- ***Aim 3.1***. To develop VB algorithms for implementing the group WGR and group interaction models

- ***Aim 3.2***. To conduct simulations to compare from VB and MCMC implementations of the group-WGR model,
  - effect estimates
  - computational times

- ***Aim 3.3***. To apply the group interaction model implemented in VB framework to one real human data set and compare results with MCMC (Aim 2.3)

In summary …

The proposed work will implement novel WGR based methods in the analysis of genetic data from structured human populations.

These methods can provide a trait/disease-specific characterization of genetic heterogeneity and incorporate LD patterns using algorithms that can improve computationally efficiency.

Acknowledgements

University of Alabama at Birmingham

Dr. G. de los Campos - Advisor

Dr. A.I. Vazquez

National Institutes of Health

Turning Discovery Into Health

Friends and colleagues at UAB!

Thanks for your attention!

Questions?

Extra slides

Penalized vs. Bayesian methods

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \left\{ \sum_i (y_i - \sum_{j=1}^{p} x_{ij}\,\beta_j)^2 + \lambda J(\boldsymbol{\beta}) \right\}$$

Model unknown

Least squares function

Regularization parameter

Penalty function

$$p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \omega) \propto \prod_{i=1}^{n} N(y_i | \sum_{j=1}^{p} x_{ij}\,\beta_j, \sigma^2) \prod_{j=1}^{p} p(\beta_j | \omega)\, p(\sigma^2)$$

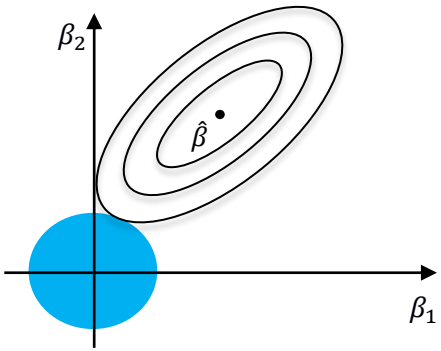
Likelihood function

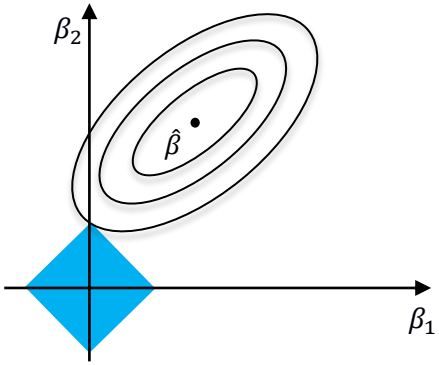
Prior distribution

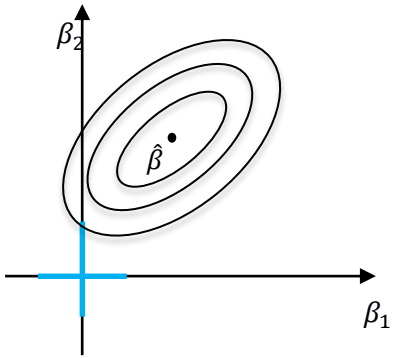
Posterior distribution of the unknowns given the data and hyper-parameters

- RLS = MAP
- Credible intervals
- A framework to deal with regularization parameters
- Computationally intensive

Penalized Estimators		$J(\beta)$
Shrinkage	Ridge Regression	$(\gamma = 2)\ \sum_{j=1}^{p} {\beta_j}^2$
	Bridge Regression	$(\gamma > 2)\ \sum_{j=1}^{p} \lvert\beta_j\rvert^{\gamma}$
Variable selection + Shrinkage	Elastic Net	$\alpha \sum_{j=1}^{p} \lvert\beta_j\rvert + (1-\alpha) \sum_{j=1}^{p} {\beta_j}^2$
	LASSO	$(\gamma = 1)\ \sum_{j=1}^{p} \lvert\beta_j\rvert$
Variable selection	Subset selection	$(\gamma \to 0)\ \sum_{j=1}^{p} I(\beta_j \neq 0)$

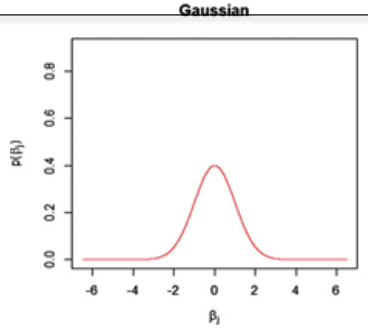


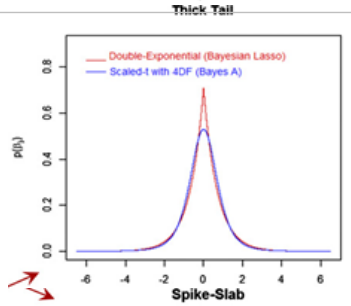


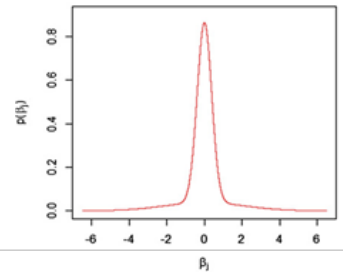


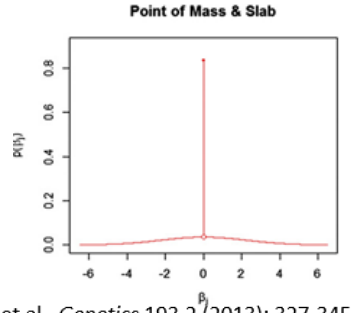
Bayesian Regularized Estimators

			$p(\beta_j \mid \omega)$
Shrinkage	Homogeneous	Bayesian Ridge Regression	$(\delta = 1, df_1 \to \infty)$ $N(\beta_j \mid 0, \sigma_\beta^2)$
Shrinkage	Differential	Bayes A	$(\delta = 1)\ t_{df,s}$
Shrinkage	Differential	Bayesian LASSO	$DE(\beta_j \mid \sigma^2, \lambda^2)$
Shrinkage	Differential	Spike Slab models	$\delta t_{df_1, s_1} + (1-\delta) t_{df_2, s_2}$
Variable selection + Shrinkage		Bayes B	$\delta t_{df,s} + (1-\delta) I(\beta_j = 0)$
Variable selection + Shrinkage		Bayes C	$(df_1 \to \infty)$ $\delta N(\beta_j \mid 0, \sigma_\beta^2) + (1-\delta) I(\beta_j = 0)$









de los Campos, Gustavo, et al. *Genetics* 193.2 (2013): 327-345.

Relationship between Penalized vs. Bayesian methods

We show that the Regularized Least Squares (RLS) estimate from Ridge Regression is the same as the posterior mode or maximum *a posteriori* (MAP) estimate obtained from Bayesian Ridge Regression

$$\widehat{\boldsymbol{\beta}}_{RR(RLS)} = \min_{\boldsymbol{\beta}} \left\{ \sum_i \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

$$= \max_{\boldsymbol{\beta}} \left\{ -\frac{\sum_i \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2}{2\sigma^2} - \frac{\lambda \sum_{j=1}^{p} \beta_j^2}{2\sigma_\beta^2} \left(\frac{\sigma_\beta^2}{\sigma^2} \right) \right\} \text{ for any } \sigma_\beta^2 \text{ and } \sigma^2 > 0$$

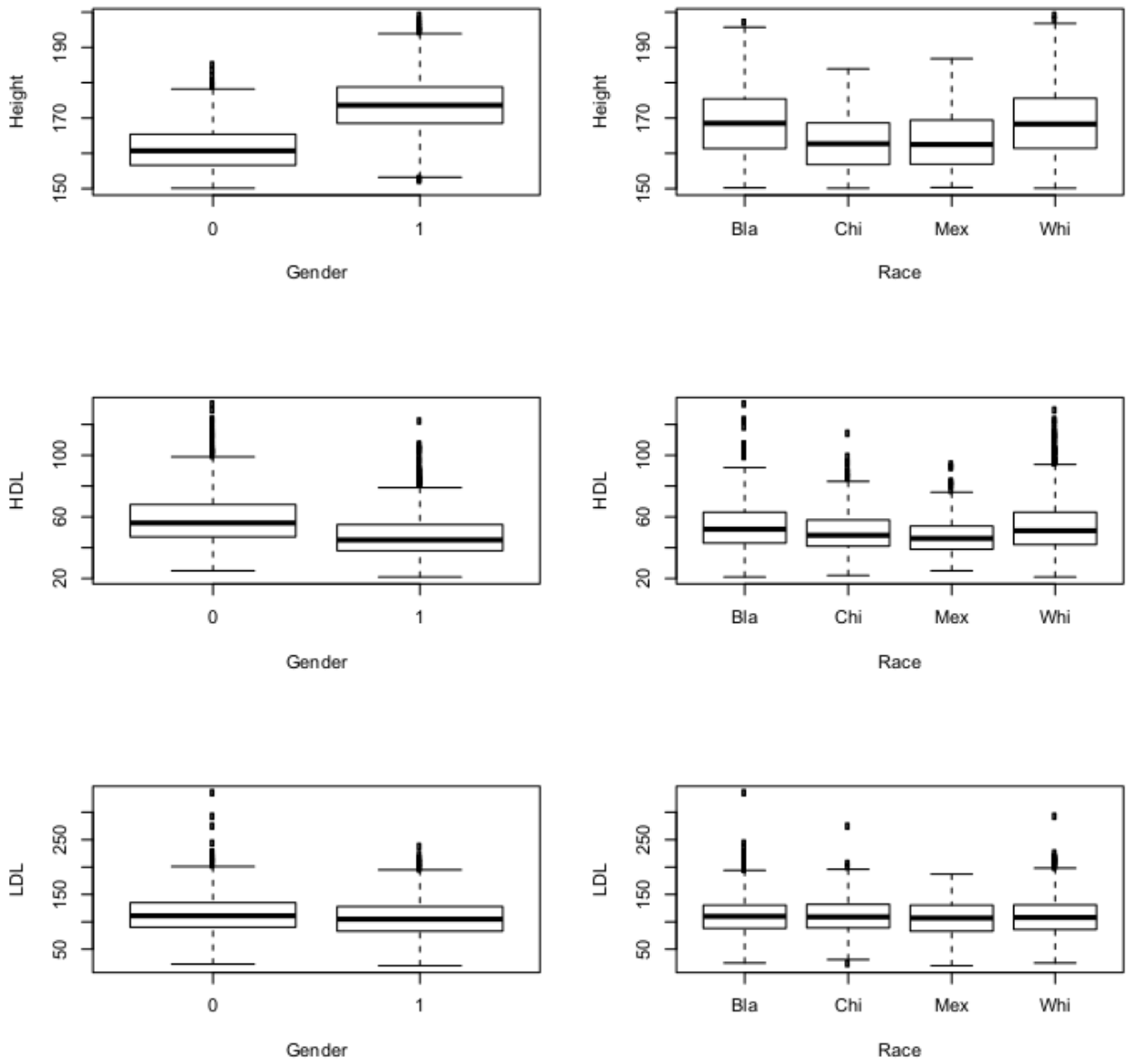
$$= \max_{\boldsymbol{\beta}} \left\{ exp\left\{ -\frac{\sum_i \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2}{2\sigma^2} \right\} exp\left\{ -\frac{\sum_{j=1}^{p} \beta_j^2}{2\sigma_\beta^2} \left(\frac{\lambda\sigma_\beta^2}{\sigma^2} \right) \right\} \right\}$$

$$\text{If } \lambda = \frac{\sigma^2}{\sigma_\beta^2},$$

$$\widehat{\boldsymbol{\beta}}_{RR(RLS)} = \max_{\boldsymbol{\beta}} \left\{ \prod_{i=1}^{n} N\left(y_i \middle| \sum_{j=1}^{p} x_{ij}\beta_j \ , \sigma^2\right) \prod_{j=1}^{p} N(\beta_j | 0, \sigma_\beta^2) \right\}$$

$$= \widehat{\boldsymbol{\beta}}_{BRR(MAP)}$$

Phenotypes in MESA



Aim 1.1

Standard WGR model

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

$$\mathrm{i} = 1,2 \ldots \mathrm{n}$$

Prior:

$$p(\boldsymbol{\beta}, \sigma^2, \omega) = p(\boldsymbol{\beta}, \omega) \times p(\sigma^2)$$

$$p(\sigma^2) \sim \chi^{-2}(S, d)$$

Interaction model

$$y_{ik} = \sum_{j=1}^{p} X_{kij}(b_{0j} + b_{kj}) + \varepsilon_{ik}$$

$$\mathrm{i} = 1,2 \ldots \mathrm{n};\ \mathrm{k} = 1,2$$

Prior:

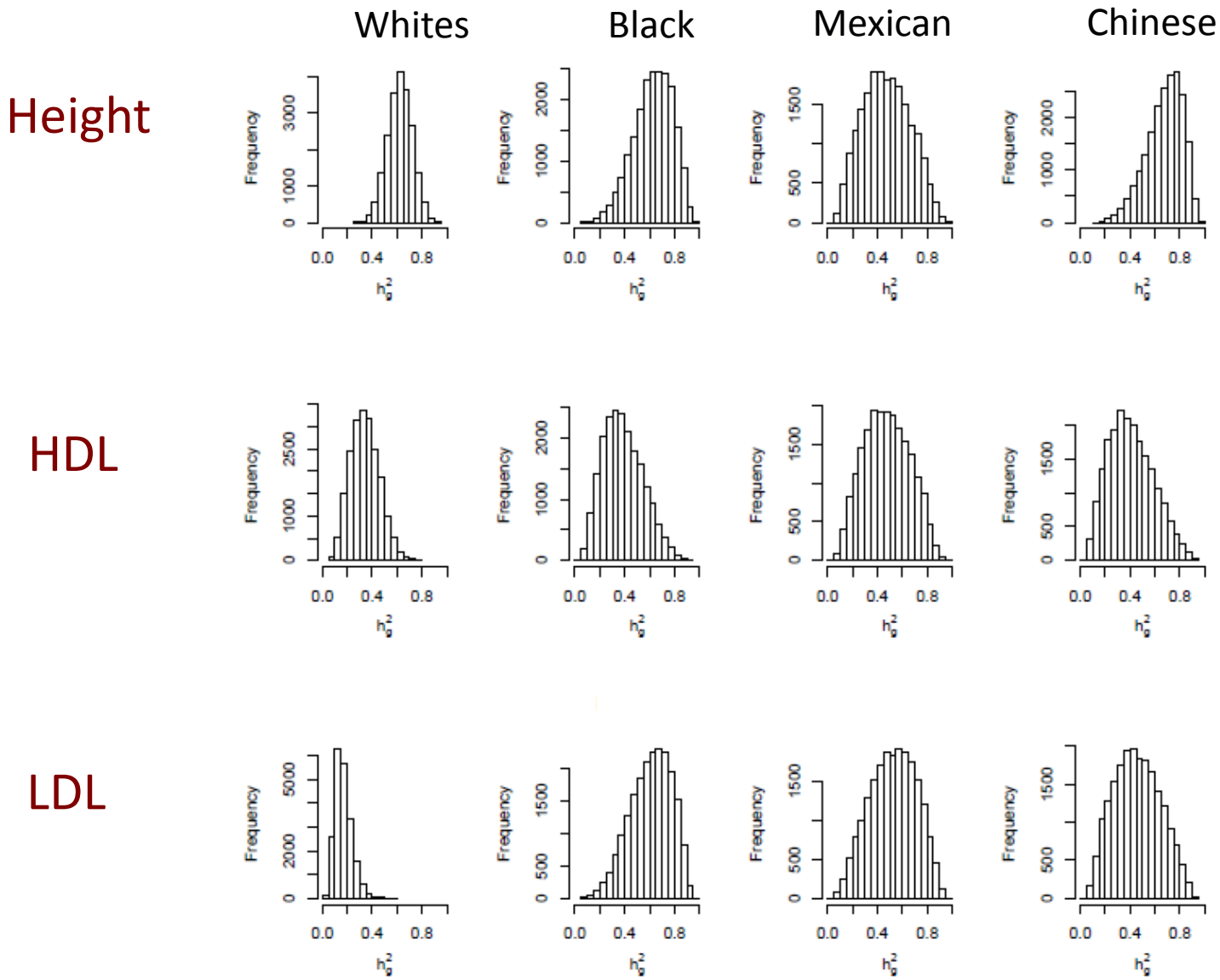
$$p(\boldsymbol{b_0}, \boldsymbol{b}_1, \boldsymbol{b}_2, \sigma_1^2, \sigma_2^2, \omega_0, \omega_1, \omega_2) =$$
$$= p(\boldsymbol{b}_0, \omega_0) \prod_{k=1}^{2} \{p(\boldsymbol{b}_k, \omega_k)\} \times p(\sigma_1^2) p(\sigma_2^2)$$

where

$$p(\boldsymbol{\beta}, \omega) \propto \prod_{j=1}^{p} p(\beta_j | \omega) \times \mathrm{p}(\omega)$$

For the G-BLUP $p(\beta_j|\omega)$ assigned a $N(0, \sigma_\beta^2)$ distribution

Aim 1.3 - Posterior Density (Genomic Heritability)



Aims 2.1 and 2.2

Group WGR model

$$y_i = \sum_{g=1}^{G} \delta_g \sum_{j=1}^{p_g} X_{ijg}\beta_{jg} + \varepsilon_i;$$
$$\mathrm{i} = 1,2 \ldots \mathrm{n}$$

Joint Prior:

$$p(\boldsymbol{\theta}, \sigma^2) = p(\boldsymbol{\theta}) \times p(\sigma^2)$$

$$\boldsymbol{\theta} = (\boldsymbol{\beta}, \omega, \boldsymbol{\delta}, \pi)$$

Group Interaction model

$$y_{ik} = \sum_{g=1}^{G_k} \sum_{j=1}^{p_g} X_{kijg}\left(\delta_{0g} b_{0jg} + \delta_{kg} b_{kjg}\right) + \varepsilon_{ik};$$
$$\mathrm{i} = 1,2 \ldots \mathrm{n};\ \mathrm{k} = 1,2$$

Joint Prior:

$$p(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma_1^2, \sigma_2^2) = p(\boldsymbol{\theta}_0) \prod_{k=1}^{2} \{p(\boldsymbol{\theta}_k)\} \times p(\sigma_1^2) p(\sigma_2^2)$$

$$\boldsymbol{\theta_0} = (\boldsymbol{b_0}, \delta_0, \omega_0, \pi_0) \quad \boldsymbol{\theta_k} = (\boldsymbol{b_k}, \delta_k, \omega_k, \pi_k)$$

where $p(\boldsymbol{\theta}) = p(\boldsymbol{\beta}, \omega | \delta_g) \times p(\delta_g | \pi) \times p(\pi) \times p(\omega)$

$p(\pi) \sim beta(a_1, a_2)$

$p(\sigma^2) \sim \chi^{-2}(S, d)$

Spike-slab prior: $p(\boldsymbol{\beta}, \omega | \delta_g) = \prod_{g=1}^{G} \prod_{j=1}^{p_g} \{p(\beta_{jg} | \omega)^{\delta_g} \times I(\beta_{jg} = 0)^{1-\delta_g}\}$

Bernoulli distribution: $p(\delta_g | \pi) = \prod_{g=1}^{G} \{\pi^{\delta_g} (1-\pi)^{1-\delta_g}\}$

We will develop MCMC algorithms to implement these models in R

Variational Bayes

Let's represent the approximate distribution by $q(\boldsymbol{\beta})$ of the true distribution $p(\boldsymbol{\beta}|\mathbf{y})$. The Kullback-Leibler divergence is then given by:

$$KL(q||p) = -\int \left(\log\left(\frac{p(\boldsymbol{\beta}|\boldsymbol{y})}{q(\boldsymbol{\beta})}\right)\right) q(\boldsymbol{\beta})d\boldsymbol{\beta} = -\int \left(\log\left(\frac{\frac{p(\boldsymbol{\beta},\boldsymbol{y})}{p(\boldsymbol{y})}}{q(\boldsymbol{\beta})}\right)\right) q(\boldsymbol{\beta})d\boldsymbol{\beta}$$

$$= -\int \left(\log\left(\frac{p(\boldsymbol{\beta},\boldsymbol{y})}{q(\boldsymbol{\beta})}\right)\right) q(\boldsymbol{\beta})d\boldsymbol{\beta} + \int \log\big(p(\boldsymbol{y})\big)\, q(\boldsymbol{\beta})d\boldsymbol{\beta}$$

$$= -E_g\left(\log\left(\frac{p(\boldsymbol{\beta},\boldsymbol{y})}{q(\boldsymbol{\beta})}\right)\right) + \log\big(p(\boldsymbol{y})\big)\int q(\boldsymbol{\beta})d\boldsymbol{\beta}$$

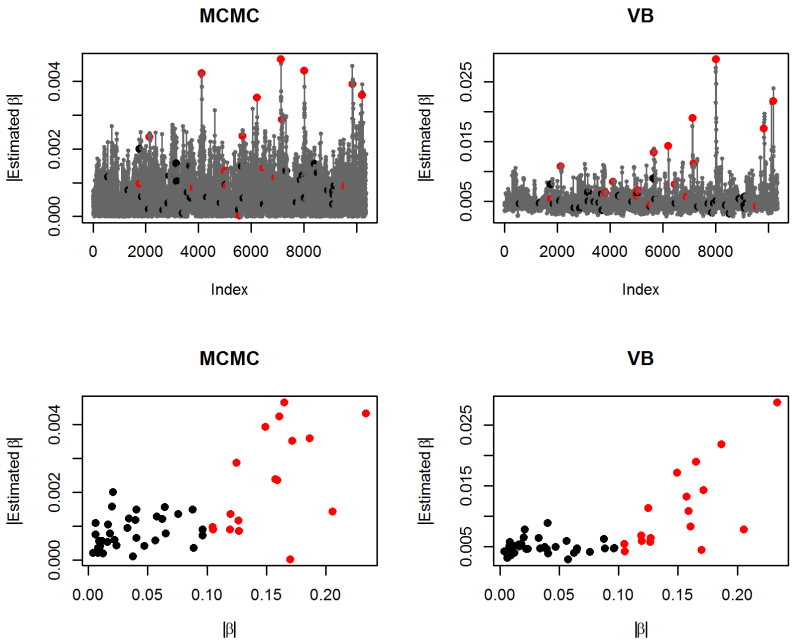
$$= -E_g\left(\log\left(\frac{p(\boldsymbol{\beta},\boldsymbol{y})}{q(\boldsymbol{\beta})}\right)\right) + \log\big(p(\boldsymbol{y})\big)$$

Variational lower bound

- VB maximizes the variational lower bound by **mean-field approximation**, given as

$$q(\boldsymbol{\beta}) = \sum_{j=1}^{p} q_j\,(\boldsymbol{\beta}_j)$$

VB vs MCMC?



Parameters

- *Mice data set*
- $nQTL = 50$
- $n = 487$
- h^2 = 0.4 (simulated phenotypes)
- $p = 10{,}346$
- BayesC-VB: *varbvs*[1]
- BayesC-MCMC: *BGLR*[2]

Red : Large effect QTL |true effects| > 0.1
Black: Small effect QTL |true effects| < 0.1
Gray: Non-causal variants

- Good correlation between true and estimated effects using both methods
- VB effect estimates affected by correlated predictors
- VB takes the same time as 1000 MCMC iterations

[1]Carbonetto and Stephens 2012. Bayesian Anal. **7**: 73–108
[2]de los Campos, G., Rodriguez P., 2014. BGLR: Bayesian Generalized Linear Regression.