WHOLE GENOME REGRESSION WITH DATA FROM HETEROGENEOUS POPULATIONS – EXECUTIVE SUMMARY

Analysis of complex human traits and diseases has been revolutionized by the inundation of high-throughput genomic data using approaches such as Genome-Wide Association Studies (GWAS) and Whole Genome Regression (WGR). Standard GWAS is a one-marker-at-a-time approach to find the common variants associated with a trait of interest, whereas in WGR, phenotypes or disease risk are regressed concurrently on all available markers; this can explain a much greater proportion of genetic variability than GWAS.

Human populations are stratified in varying degrees (GAGGIOTTI et al. 2009; PFENNINGER et al. 2011; PUCKETT et al. 2014) and standard GWAS typically treat this structure as a “confounder”. They don’t consider the possibility that marker effects may be sub-population specific, which is a reasonable assumption considering the differences in allele frequencies and linkage disequilibrium patterns across heterogeneous sub-populations. We therefore propose a new approach that considers structure to be an “effect modifier” rather than a confounder. This approach models random-effect interaction between markers and inferred clusters and can accommodate cluster-specific effects. The proposed model yields cluster-specific genomic variances and between-cluster correlations. This provides a framework to test for the existence of genetic heterogeneity.

From the proposal, the three critiques for this chapter that we have addressed here are:

1. Use sample sizes larger than those from the Multi-Ethnic Study for Atherosclerosis or MESA to lower the standard errors in our estimates.

   > a) We used the Atherosclerosis Risk In Communities or ARIC dataset, which has much larger sample sizes than MESA for two racial clusters (Blacks and Whites).
   >
   > b) After standard quality control (QC) procedures, this dataset has ~6500 whites and ~1600 blacks that were nominally unrelated
   >
   > c) SNPs in this dataset were genotyped using the Affymetrix 6.0 array; 832766 SNPs remained after QC.

2. Use simulations to validate the accuracy and precision of estimates of genomic heritability and marker-effect correlations obtained from the interaction model.

   > a. We used real genotypes from the ARIC dataset
   >
   > b. Phenotypes were simulated by sampling at random different numbers of

> quantitative trait loci (QTL), at different levels of marker effect correlation and trait heritability for the two clusters.
>
> c. Interaction model was fit for each combination of simulation parameters using **only causal loci** for approximately 100 replicates to account for sampling variation. After this, comparisons were made between true and estimated variance components and marker effect correlations in terms of precision and accuracy.

3. Regress real phenotypes on *GWAS significant markers* instead of only using the genome-wide marker set.

> a) We used phenotypes whose true heritabilities range from high to low (standing height in cm, high-density lipoprotein or HDL, and low-density lipoprotein or LDL)
>
> b) We obtained GWAS significant markers from GWAS consortia such as Genome-wide Investigation of ANthropometric Measures or GIANT and The Global Lipids Genetic Consortium. These markers were chosen at different – log(p-value) cutoffs per phenotype.
>
> b)The interaction model for each phenotype was fit using SNPs obtained at each cutoff value. The estimates of cluster-specific genomic heritability and between-cluster marker-effect correlation were compared across the different cutoffs; i.e. when the number of markers used to fit the model ranged from small to large.

Methods

I) WGR Model for Structured Data

We model marker effects as the sum of a component that is common across clusters (b_{0j}) and interaction terms or deviations that are cluster-specific $(b_{kj}$ for $k = 1, 2)$. Therefore, marker effects for the k^{th} cluster are given by $\beta_{kj} = b_{0j} + b_{kj}$. Accordingly, the corresponding data equations for the two clusters are given as:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1\mu_1 \\ 1\mu_2 \end{bmatrix} + \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} b_0 + \begin{bmatrix} X_1 \\ 0 \end{bmatrix} b_1 + \begin{bmatrix} 0 \\ X_2 \end{bmatrix} b_2 + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

where $\mu_1, \mu_2, \beta_1 = b_0 + b_1, \beta_2 = b_0 + b_2, y_1 = \{y_{1i}\}_{i=1}^{n_1}, y_2 = \{y_{2i}\}_{i=1}^{n_1}$, $X_1 = \{X_{1ij}\}_{i=1,j=1}^{n_1,p}, X_2 = \{X_{2ij}\}_{i=1,j=1}^{n_2,p}, \varepsilon_1 = \{\varepsilon_{1i}\}_{i=1}^{n_1}$ and $\varepsilon_2 = \{\varepsilon_{2i}\}_{i=1}^{n_1}$ represent cluster-specific intercepts, effects, phenotypes, genotypes and model residuals, respectively. The total sample size $n = n_1 + n_2$. The residual vectors ε_1 and ε_2 are treated as independent because they pertain to measurements in different individuals, but the effect vectors b_1 and b_1 are correlated, where

$$Cor(\beta_{ij}, \beta_{2j}) = \frac{\sigma_{b0}^2}{\sqrt{(\sigma_{b0}^2 + \sigma_{b1}^2) \times (\sigma_{b0}^2 + \sigma_{b1}^2)}}$$

This model was fit using Bayesian methods with the R package BGLR

II) Quality Control (QC)

Standard quality control was applied. Monomorphic SNPs and those with minor allele frequency < 0.05 *in each subpopulation* were removed, along with markers and individuals with >5% missing values. We retained SNPs that had >5% allele frequency in at least one of the two sub-populations or clusters. Individuals that belonged to a pedigree were removed and only one individual per family was retained from the pedigree. Genomic relationship matrices were computed for each cluster (after centering the cluster-specific X matrices) and individuals that have a diagonal value outside the range [0.85, 1.15] and off-diagonal values > 0.075 were removed. Thus, the final data sets comprised only distantly related individuals.

III) Simulations

As explained above, we conducted simulations using real genotype data from the ARIC dataset in order to validate the estimates obtained from the interaction model. We simulated phenotypes for blacks and whites under different heritability levels (0.4 and 0.6), at two levels of marker effect correlation (0.25 and 0.75), and using two different sets of randomly sampled causal variants (500 and 5000). The interaction model was fit for each combination of levels for 100 replicates (to account for sampling variation) and the the cluster-specific genomic heritability and marker-effect correlation estimates were compared to the true values in terms of their precision and accuracy.

IV) Real data analysis

For real analysis we chose three phenotypes: Height, HDL and LDL cholesterol. We obtained GWAS-significant markers from GIANT consortium and The Global Lipids Consortium and sorted the p-values of the markers in ascending order. We chose 4 different $-log_{10}(p - value)$ cutoffs for each of the three chosen phenotypes. Each cutoff corresponded to different numbers of SNPs. The cutoffs (and corresponding number of SNPs) for each of the three phenotypes were:

> **Height**: 4 (3709), 5(2386), 8(917), 10(515)
> **HDL**: 3(2921), 4 (1668), 6(946), 9(515)
> **LDL**: 1(3651), 1.5(1784), 2(1019), 3(544)

We also added the case where analysis was performed with very large SNP coverage: i.e. 200,000 randomly chosen SNPs.

Preliminary Results

Simulations. From the results, we observed that the estimated correlations and error variances have very good precision and accuracy (Figs 1-3) under both values of underlying correlation (0.25 – *data not shown*; and 0.75 – Figs. 1-3) when the number of QTL is small (500 – Fig. 1) and large (5000 – Figs. 2-3) . The variance in estimates of marker-effect correlation were found to be marginally higher when the number of markers increased to 5000 – Fig. 2, because of the error associated with estimating additional parameters. **Overall, these simulations showed that on average, the interaction model results in accurate and precise estimates of cluster-specific variance components (Fig. 3) and between-cluster marker-effect correlation (Figs. 1-2) when models were fit using causal variants**.

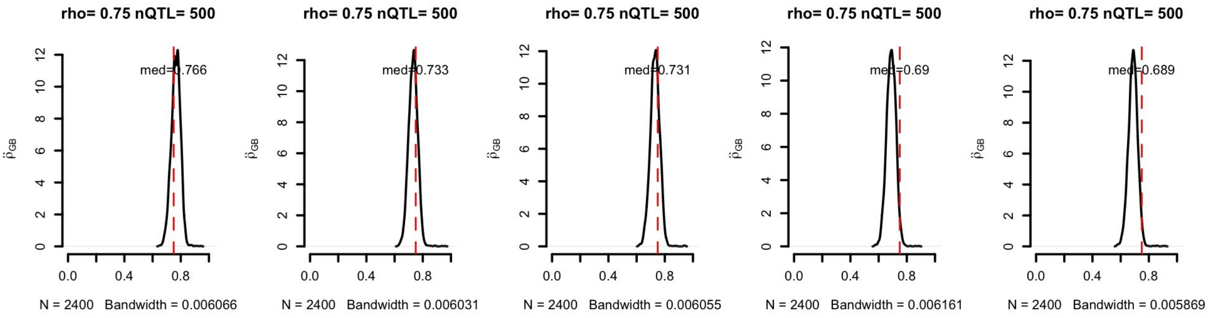

Figure 1. Density plots of MCMC samples of **estimated marker effect correlations** between blacks and whites for simulated data from five replicates. Phenotypes were simulated with **500 QTL** randomly sampled under heritability of 0.6 for whites and 0.4 for blacks and a between cluster correlation of 0.75. The red dotted line represents the true correlation value.

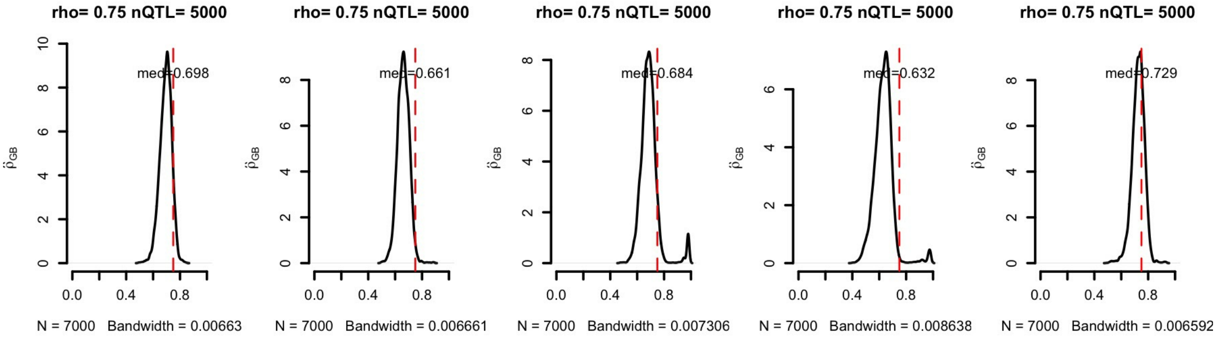

Figure 2. Density plots of MCMC samples of **estimated marker effect correlations** between blacks and whites for simulated data from five replicates. Phenotypes were simulated with **5000 QTL** randomly sampled under a heritability of 0.6 for whites and 0.4 for blacks and a between cluster correlation of 0.75. The red dotted line represents the true correlation value.

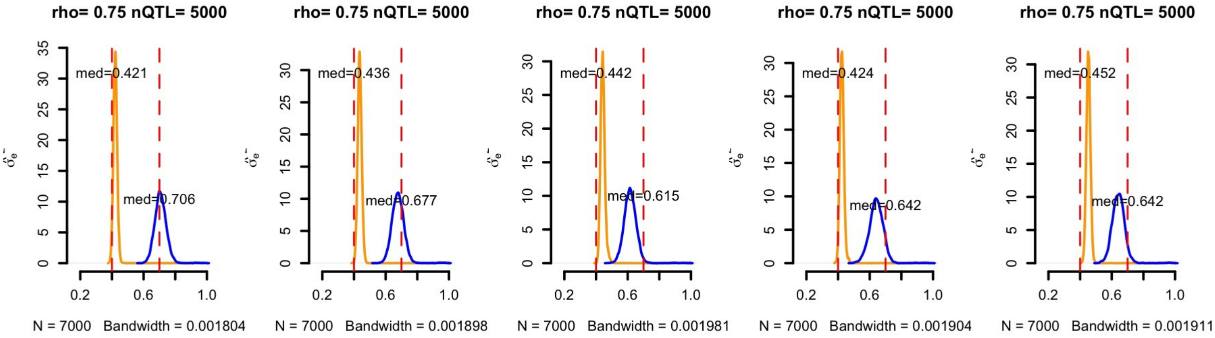

Figure 3. Density plots of MCMC samples of estimated **error variances** for blacks and whites with simulated data from five replicates. Phenotypes were simulated with **5000 QTL** randomly sampled under heritability of 0.6 for whites and 0.4 for blacks and a between cluster correlation of 0.75. The red dotted line represents the true error variance value for each cluster.

Real data analysis. As expected, models fit with GWAS-significant markers (in comparison to those fit with the 200K SNP set) resulted in lower standard errors for estimates of genomic heritability, albeit at the cost of an increase in bias (Fig. 4 – here, the interaction model was fit only using markers that passed the $-log_{10}(p-value)$ cutoffs for each of the three phenotypes). The estimated genomic heritability also increased with the number of markers; however, the trends for the three phenotypes were different. Height, being a quantitative trait, showed consistently significant increase in estimated genomic heritability as the number of markers increased. On the other hand, estimated genomic heritability for HDL and LDL did not vastly increase with the number of markers, especially for whites. This is because LDL has far fewer causal variants (GLOBAL GENETICS LIPIDS CONSORTIUM, 2013) than height (YANG et al., 2010). Importantly, these estimates were found to vary between clusters. For LDL, using a large number of markers, more of the genetic variation could be explained in blacks than in whites. Also, estimates of genomic heritability had much better precision among whites than among blacks, which could be attributed to the difference in sample sizes between the two clusters.

On the other hand, marker-effect correlation estimates between whites and blacks reduced as the number of markers increased. This is reasonable since we can expect considerable overlap in top causal variants between clusters. The correlation decreases as a greater percentage of the genome is included. However, the trends for the three phenotypes were again different; LDL had a mean correlation of 0.2 with 200K markers whereas height had a correlation of approximately 0.6 (Fig. 4 – top row). Also, the GWAS-significant markers for LDL had lower correlation than those for HDL and height (Fig 4 – top row). LDL appears to have many more causal variants for blacks that are not shared by whites, in contrast to height. **In summary, the real-data analysis showed that the extent of genetic heterogeneity among clusters (blacks and whites) varies by trait (Fig. 4)**.

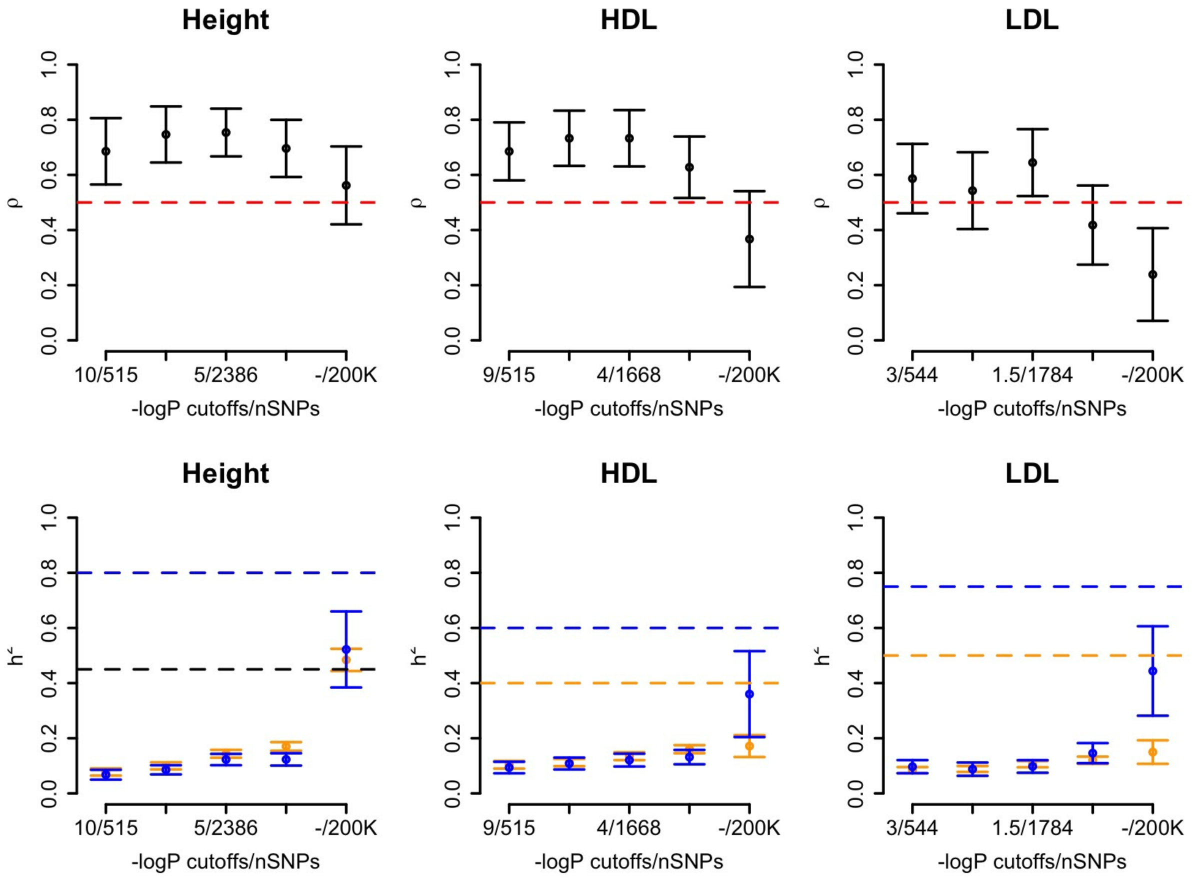

Figure 4. Point estimates (mean) and standard errors of MCMC samples of: marker effect correlations between blacks (1602) and whites (6569) (*top row*) and estimated genomic heritability for blacks (blue) and whites (orange) (*bottom row*) for the three phenotypes: standing height (in cm), HDL (mg/dL) and LDL (mg/dL). In each case, the first four axis points correspond to GWAS-significant markers chosen at four different cutoffs of $-log_{10}(p-value)$; the cutoffs (and corresponding number of SNPs) are: **Height**: 4(3709), 5(2386), 8(917), 10(515), **HDL**: 3(2921), 4(1668), 6(946), 9(515), **LDL**: 1(3651), 1.5(1784), 2(1019), 3(544) and the final axis point corresponds to the case with 200,000 randomly chosen SNPs.
Top row: The red dotted line represents the half-way mark
Bottom row: The blue dotted line represents the true heritability for blacks, the orange dotted line represents the true heritability for whites, the black dotted line represents the *genomic* heritability.

References

1) GAGGIOTTI O. E., BEKKEVOLD D., JØRGENSEN H. B. H., FOLL M., CARVALHO G. R., et al., 2009 Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. Evolution 63: 2939–51.
2) PFENNINGER M., SALINGER M., HAUN T., FELDMEYER B., 2011 Factors and processes shaping the population structure and distribution of genetic variation across the species range of the freshwater snail radix balthica (Pulmonata, Basommatophora). BMC Evol. Biol. 11: 135.
3) PUCKETT E. E., KRISTENSEN T. V, WILTON C. M., LYDA S. B., NOYCE K. V, et al., 2014 Influence of drift and admixture on population structure of American black bears (Ursus americanus) in the Central Interior Highlands, USA, 50 years after translocation. Mol. Ecol. 23: 2414–27
4) Global Lipids Genetics Consortium., 2013. Discovery and refinement of loci associated with lipid levels. Nature genetics, 45(11), 1274-1283.
5) YANG J., BENYAMIN B., MCEVOY B. P., GORDON S., HENDERS A. K., et al., 2010a Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565–9.