**1.a. Full Title:** A Mixed Linear Effects Model for Association Testing on the X Chromosome in Samples with Unknown Structure: Application to the Hispanic Community Health Study/Study of Latinos (HCHS/SOL)

**1.b. Abbreviated Title:** X Chromosome Association Testing with MLMs in GWAS with Unknown Structure

**1.c. Keywords:** ancestry, admixture, population structure, relatedness, kinship, principal components, variance components, association, mixed models, X chromosome, GWAS, Hispanic, Latino

**2. Proposer:** Caitlin McHugh

**3. Affiliation:** HCHS/SOL Genetic Analysis Center, Department of Biostatistics, University of Washington, Seattle

**4. Sponsoring PI and Ancillary PI:** Bruce Weir

**5. Co-Authors:** Conomos, MP; Jain, D; Laurie, CC; Thornton, TA; Weir, BS; Reiner, AP; North, K (\*add more blood cell coauthors?)

**6. First Author:** McHugh, Caitlin

**7. Will the DNA or biomarker data be used in this manuscript?** Yes

**8. Review manuscript proposals and no overlap?** Yes

**9. Where will the data analysis be performed?** HCHS/SOL Genetic Analysis Center, Department of Biostatistics, University of Washington, Seattle

**10. Is this manuscript proposal associated with any HCHS/SOL ancillary studies or use and ancillary study data?** No

**11. Rationale:**

As the number of genome wide association studies (GWAS) increases, discovery of genetic polymorphisms associated with disease continues to increase. These studies are beginning to focus on individuals from populations with multi-way admixture, such as Hispanic Americans. The Hispanic Community Health Study/Study of Latinos (HCHS/SOL) offers an opportunity to examine the genetic structure of complex disease in a sample of 13,000 individuals currently living in the United States.

Nearly every GWAS dataset includes genotypes from genetic markers across all autosomal chromosomes, as well as the sex chromosomes and the mitochondria. However, the vast majority of statistical association methods have focused on the autosomes, with less attention given to the analysis of X chromosome markers. Recent work has identified the lack of association testing on the X chromosome (Wise, et al. 2013). Although the size and number of genes on the X chromosome are similar to chromosome 7, the X chromosome is often ignored in association studies and, as a result, there are a disproportionate number of associations with X chromosome markers. Indeed, at the end of 2014, the GWAS catalog included only 135 of over 19,000 published GWAS results with associations on X chromosome SNPs (GWAS catalog). Association testing on the X chromosome presents a challenge, as females inherit one X chromosome from each parent, whereas a male only inherits a single X chromosome from his mother. The difference in sex is further complicated by X chromosome inactivation, or dosage compensation, which is still not entirely understood (Mugford, et al. 2014). As X chromosome association methods are developed, we must be sure we can account for confounders such as population structure and relatedness among our samples.

Proper association testing on the X chromosome allowing for relatedness among samples requires estimation of the pairwise kinship coefficient values on the X chromsome. As compared to autosomal relatedness, X chromosome relatedness must take into account the sex of the individuals. For more distant relatives, whether the individuals are related through a maternal or paternal lineage must also be considered, as this affects the amount of X chromosome genetic material shared between a pair of individuals. Relatedness estimates from the X chromosome can complement autosomal relatedness estimates in pedigree construction, and must be considered explicitly when performing association testing on the X chromosome in related samples.

Mixed linear models (MLMs) are a powerful and effective approach for analysis of GWAS with population and family structure. With MLMs, we can simultaneously account for both population structure and relatedness among our study samples by including a genetic relatedness matrix (GRM) as part of the covariance structure of the phenotype. In the past few years, the MLM approach has been made computationally feasible for analysis of GWAS data (Kang et al., 2010, Listgarten et al., 2012, Zhou et al., 2012). However, it still remains that existing MLM approaches are not directly applicable to analyzing X chromosome markers.

Here, we propose the MLM-X method for association testing on the X chromosome in samples with unknown structure. MLM-X includes a random effect for polygenic effects on the X chromosome and a random effect for the autosomes. Variance components of the random effects in MLM-X are calculated via restricted maximum likelihood (REML) with empirical GRMs for the autosomes and X chromosome. MLM-X improves over existing methods in terms of type I error and power for trait mapping with X chromosome variants.

**12. Main Hypothesis/Study Questions:**

Can our method demonstrate increased power compared to existing autosomal methods simply applied to X chromosome data?

Does our method yield improved type I error compared to existing methods?

Does our method provide accurately calibrated association testing in X chromosome markers in HCHS/SOL, compared to existing methods?

Does our method support known associations of phenotypes with X chromosome genetic markers previously discovered in Hispanic samples?

**13. Analysis Plan/Outline:**

We plan to utilize both simulated and HCHC/SOL observed genotype and phenotype data to assess the performance of our method. We compare results using MLM-X to existing MLM methods that do not account specifically for X chromosome relatedness or population structure, which is the typically employed autosomal model. Any HCHS/SOL data will be presented generally, in terms of group results and summary statistics. As a result, no individual identification will be possible. The population structure will be estimated using the PC-AiR method (Conomos and Thornton, 2015) and ancestry-adjusted GRMs will be calculated using the PC-Relate method (Conomos, Weir, and Thornton, manuscript in preparation).

Simulation studies will initially be used to demonstrate the amount of genetic relatedness we can detect using X chromosome markers. We will compare results using only the X chromosome to results using all autosomes together, along with using a subset of autosomal markers that is the same size as the set of X chromosome markers. The simulation will consist of samples from a homogeneous population that are related through a known pedigree with three generations. Varying numbers of independent SNPs from both the X chromosome and an autosomal chromosome will be generated for all founders in the pedigree at a specified allele frequency. Genotypes for the remaining members of the pedigree will be generated by dropping alleles down the pedigree. We then estimate the GRM using only X chromosome SNPs, only autosomal SNPs and a subset of autosomal SNPs. The estimated X chromosome kinship coefficient estimates tend towards the theoretical value as the number of markers increases. The distribution of GRM values will inform us of how variable the estimates are under varying numbers of genetic markers.

The second part of the simulation study assesses power and type I error rate for the MLM-X method. One million X chromosome SNPs will be generated for 8,000 samples that are related through 500 pedigrees for a resulting 2,500 unrelated samples and 5,500 that are related through known pedigrees. The alleles, generated at a frequency of 0.2, are passed through the pedigree as described above. To simulate a quantitative phenotype, first a causal SNP is chosen. The resulting phenotype is the sum of four terms: the effect size multiplied by the genotype at the causal SNP, the background polygenic effect due to the autosomes, the background polygenic effect due to the X chromosome, and an error term. The background autosomal and X chromosome polygenic effects are modeled as random effects. Thus, they will be simulated from multivariate normal distributions with mean zero and the GRM estimated from the autosomes and the X chromosome, respectively, as the covariance matrices. A constant variance term of 0.3, 0.5 or 0.8 will be multiplied by the covariance matrices. The error term is sampled from a normal distribution with mean zero and variance one, where all error terms are independent between samples. The effect size will vary from 0.01 to 1.3. We will simulate 1,000,000 phenotypes independently and perform MLM-X association tests with the causal SNP, as well as with 1,000,000 SNPs that are not associated with the phenotype. The proportion of false positives that are detected with MLM-X will be presented in a table. The proportion of true positives will be compared to the proportion of false positives, for a given nominal significance level. These values will be presented in a figure. We will compare these results to the performance of fitting a model that excludes correction for background polygenic effects due to the X chromosome, as well as a model that excludes correction for background polygenic effects due to the autosomes. The comparison will allow us to evaluate the performance of our method in terms of power and type I error against the commonly used autosomal method and will allow for a recommendation of which model to fit using GWAS data on X chromosome genetic markers.

We will apply the MLM-X method to the HCHS/SOL data. The GRM will be estimated from three different sets of genetic markers: a set of independent autosomal markers and a set of independent X chromosome markers. We will compare the autosomal GRM to the X chromosome GRM for particular pairs of individuals, proving the X chromosome relationships mirror the relationships detected using autosomal markers. These results will be presented in a scatterplot format. To examine population structure on the X chromosome beyond what is detected on the autosomes, we will perform an eigendecomposition of the X chromosome GRM, adjusted for autosomal ancestry. The top two eigenvectors from this analysis will be presented. We will look at a scree plot describing the proportion of variance explained by the first 10 eigenvectors in each analysis. In order to understand what we expect when adjusting for the correct population structure, we will look at the eigendecomposition of the GRM estimated using X chromosome SNPs, after adjusting for X chromosome population structure. The results will be plotted, allowing us to conclude there remains structure on the X chromosome even when correcting for autosomal population structure.

Finally, we will present results from applying the MLM-X model to the red blood cell count (RBC) trait. X chromosome hits have previously been published for this phenotype (Chen, et al. 2013) and were replicated in the SOL data. We will fit five models of interest, including the usual autosomal model. The remaining four models all include X chromosome principal components (PCs) as fixed effects and the X chromosome GRM as a random effect, but vary in whether they include or exclude autosomal principal components (PCs) as fixed effects and whether they include or exclude the autosomal GRM as a random effect. Variance components will be estimated in each of these models, and we will present the results of the estimated proportion in a table. P-values from association tests for all five models will be presented as scatterplots of the –log10(p-values) and effect size estimates will be reported for known hits. Furthermore, the genomic inflation factor will be presented for each model, genome-wide and specifically for the X chromosome, in a table.

**Sample Tables and Figures**

**(Note: none of the data presented is real)**

Table 1. Type I Error Rate for Various MLM Approaches

|  |  |  |  |
| --- | --- | --- | --- |
|  | Correction for X + Autosomes | Correction for X | Correction for Autosomes |
| 0.05 |  |  |  |
| 0.01 |  |  |  |
| 0.005 |  |  |  |
| 0.001 |  |  |  |

Table 2. Estimated Proportion Variance for Each of the Considered Components

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Autosomal GRM | X Chromosome GRM | Household | Block Group | Environment |
| Autosomal |  |  |  |  |  |
| Model 1 |  |  |  |  |  |
| Model 2 |  |  |  |  |  |
| Model 3 |  |  |  |  |  |
| Model 4 |  |  |  |  |  |

Table 3. Genomic Inflation Factor for Models of Interest

|  |  |  |
| --- | --- | --- |
|  | X Chromosome | Genome-wide |
| Autosomal |  |  |
| Model 1 |  |  |
| Model 2 |  |  |
| Model 3 |  |  |
| Model 4 |  |  |

Figure 1. Statistical Power for Comparison of MLM Approaches

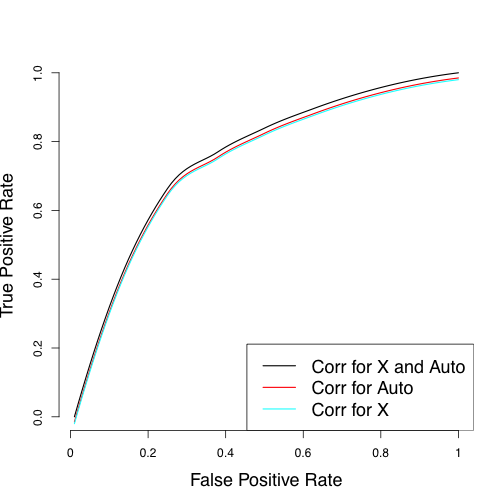


Figure 2. X Chromosome GRM Estimates for Known PO Pairs

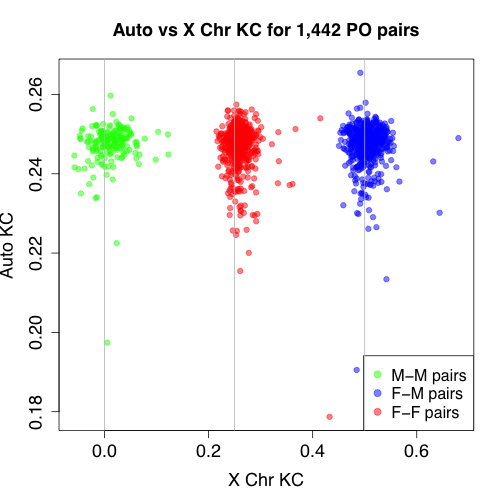


Figure 3. Top 2 Eigenvectors from X Chromosome GRM Adjusted for Autosomal Structure

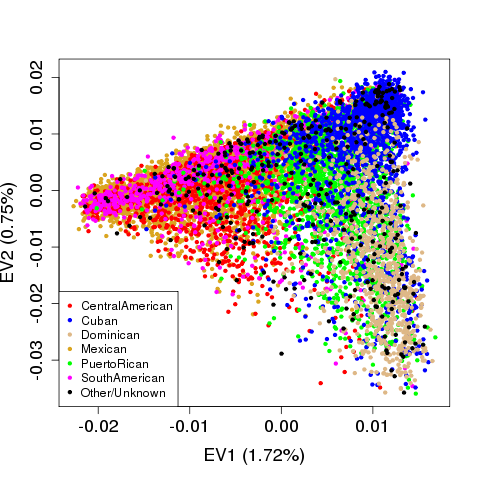


Figure 4. Top 2 Eigenvectors from X Chromosome GRM Adjusted for X Chromosome Structure

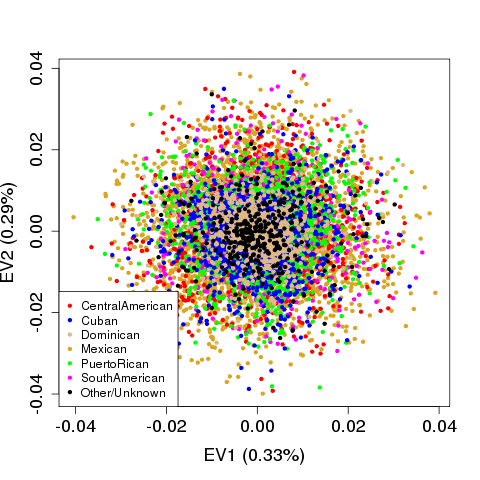
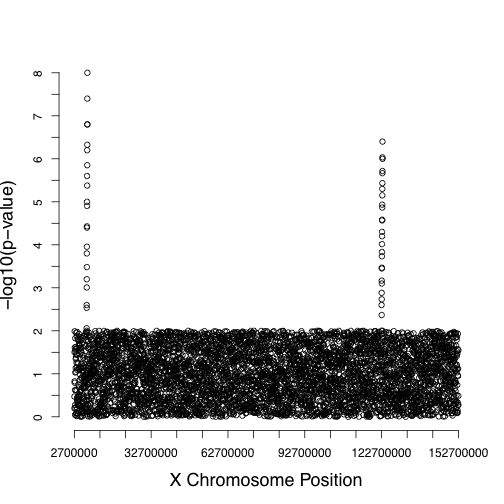


Figure 5. Manhattan Plot of X Chromosome Association Results



**14. Relevant References:**

Chen, Z., Tang, H., Qayyum, R., Schick, U.M., Nalls, M.A., Handsaker, R., et al. Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. *Hum Mol Genet*, 22(12): 2529-2538.

Conomos, M.P., Thornton, T.A., 2015. Robust population structure inference in the presence of known or cryptic relatedness. *Genetic Epidemiology*, in press.

Hindorff, L.A., MacArthur, J., Morales J., Junkins, H.A., Hall, P.N., Klemm, A.K., Manolio, T.A. A Catalog of Published Genome-wide Association Studies. Available at: [www.genome.gov/gwastudies](http://www.genome.gov/gwastudies).

Kang, H.M., Sul, J.H., Zaitlen, N.A., Kong, S., Freimer, N.B., Sabatti, C., Eskin, E., et al., 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4): 348-354.

Listgarten, J., Lippert, C., Kadie, C.M., Davidson, R.I., Eskin, E., Heckerman, D., 2012. Improved linear mixed models for genome-wide association studies. *Nature Methods,* 9(6): 525-526.

Mugford, J.W., Starmer, J., Williams Jr., R.L., Calabrese, M., Mieczkowski, R., Yee, D., Magnuson, T., 2014. Evidence for Local Regulatory Control of Escape from Imprinted X Chromosome Inactivation. *Genetics,* 197: 715-723.

Wise, A.L., Gyi, L., Manolio, T.A., 2013. eXclusion: Toward Integrating the X Chromosome in Genome-wide Association Analyses. *American Journal of Human Genetics,* 92: 643-647.

Zhou, X., Stephens, M., 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics,* 44(7): 821-824.