Examining European Ancestry in African Americans to Assess the Implications of Functional Variants

A recent study found that European Americans carry significantly more nonsynonymous SNPs than do African Americans. [iii] This was followed by a subsequent study that found that extended regions of homozygosity (ROH) in European genomes are dense with nonsynonymous substitutions. [vii] Both of these findings are consistent with inbreeding that resulted from the migration of modern *Homo sapiens* out of Africa, some 50-75,000 years ago.[v] These results are important because amino acid changes are typically harmful to the function of proteins. For my graduate research, I will take an anthropological approach that uses European ancestry in admixed populations as a tool to identify the fitness consequences of nonsynonymous mutations associated with the out of Africa migration. Since fitness combines the effects of fertility and survival, my study will be informative about genetic aspects of health and disease.

Both of the aforementioned studies assumed that any nonsynonymous change was harmful. Researchers often make this assumption and it is embedded in the software PolyPhen[i], which both of the research groups used as well. However, amino acid substitutions and even loss of function can be either harmful or beneficial. There are many genetic diseases caused by loss of function, such as Duchenne Muscular Dystrophy and Phenylketonuria. There are also beneficial phenotypes caused by diminished or lost protein function. For example, α- and β-thalassemias protect individuals from malaria, and loss of function of C chemokine receptor type 5 protects against human immunodeficiency virus. One researcher has even proposed that, "loss of gene function may represent a common evolutionary response of populations undergoing a shift in environment and, consequently, a change in the pattern of selective pressures". [iv]

I propose that genomic ancestry provides a tool to identify whether the consequences of functional variants are beneficial or harmful. For my research, I will examine European ancestry in African American individuals. Both African and European ancestors have passed on chromosome segments to African Americans. European ancestry will have introduced into African Americans chromosomal segments that constitute ROH in Europeans. Admixture affects all loci of the autosome equally. If admixture were the only evolutionary process in operation, then we would expect every chromosomal segment in the African American population to hold the same fractions of European ancestry and African ancestry. However, if natural selection were operating in addition to admixture, natural selection will distort ancestry fractions across chromosomal segments. I expect that chromosomal segments that constitute ROH in Europeans containing favorable variants will be overrepresented in African Americans. Similarly, I expect that chromosomal segments that constitute ROH in Europeans containing unfavorable variants will be underrepresented in African Americans. In principle, I expect that the proportion of ROH carrying harmful variation will be greater than ROH carrying beneficial variation. Moreover, I should be able to rank unfavorable ROH, which could shed light on the most harmful variants. Additionally, I have the potential to identify novel variation or genes carrying beneficial variants.

I plan to use exome sequences from unrelated male and female African Americans in my research. I will also use combined data from Europeans, Africans, and African Americans from the HGDP-CEPH diversity panel and HapMap project. I will create my own computer program to convert these files to a common format. I would like to incorporate genomic data from African Americans participating in the Tallahassee Project into my sample. This ongoing project, led by Dr. Connie Mulligan, investigates whether genetic risk factors and sociocultural risk have association with blood pressure in African Americans. Thus far, Dr. Mulligan and Laurel

Pearson have assayed 4000 AIMs in close to 170 African Americans, which I can use to estimate genetic ancestry. I will be using the Yoruba as a proxy for the West African parental population. I will use the Tuscan Italians and CEPH populations as the European parental population proxy. I will also consider using the Yoruba, Mandenka, French, and Orcadians from the CEPH diversity panel. Before measuring ancestry, I will perform standard SNP quality controls.[iii] I will establish a threshold for inclusion based on the amount of missing SNP data. I will also calculate expected versus observed Hardy-Weinberg Equilibrium (HWE) values. I can write programs to filter out individuals based on the amount missing data and discrepancies in calculated HWE values.

I will download annotated data from dbSNP to classify SNPs in coding sequence as either synonymous or nonsynonymous. Next, I will find the global ancestry for all chromosomes in each individual using a maximum likelihood based program, such as ADMIXTURE[ii]. I will calculate global ancestry for each for both the autosomes and sex chromosomes. Next, I will reconstruct inferred ancestry at each SNP in the exome using the likelihood-based methods of LAMP[vi]. Finally, I will look at the distribution of European ancestry in African American chromosomal segments to determine whether the variants are beneficial or harmful. I expect that natural selection will cause favorable variants have high European ancestry and unfavorable variants to have low European ancestry. In order to decide high and low properly, one of my objectives will be to establish the proper criteria for an outlier.

I will perform a second analysis of European ancestry in African Americans to extend and confirm my results. I will create four categories of SNP variation. Group I will contain genic SNPs from areas that constitute ROH in Europeans. Group II will be composed of genic SNPs from non-ROH. Group III will be comprised of third codon SNPs from non-genic regions. Group IV will include random non-genic SNPs. I expect that the variability will decrease successively across each group. Natural selection will cause genic SNPs from ROH to have the most variability in European ancestry. Random non-genic SNPs would have the least variability, since they are not probable candidates for selection, but rather drift.

My project is important for both scientific and medical research. By using anthropological genetics in conjunction with computational genomics, my project will provide information about the functional consequences of genomic variation. My project will allow scientists to transition from the identification of nonsynonymous variation to its consequences in health and ecology. The mischaracterization on nonsynonymous variants can affect how scientists view and model evolutionary processes. The assumption that loss of function confers harmful phenotypes limits one's ability to see the evolutionary potential of loss of function mutations. Natural selection may in fact favor nonsynonymous SNPs that confer an advantage within populations.[iv]

My project will allow me to receive multi-disciplinary training in anthropological genetics, computational genetics, and bioinformatics. All of these skills are critical for my future research career. After graduate school, I plan to continue studying anthropological genetics and collaborating with both scientists and physicians. I believe that interdisciplinary collaboration is essential for investigating the relationship between genetics and health.

---

[i] Adzhubei IA, Schmidt S, Peshkin L, et al. Nat Methods. 2010;7(4):248-9.
[ii] Alexander DH, Novembre J, Lange K. Genome Res. 2009;19(9):1655-64.
[iii] Lohmueller KE, Indap AR, Schmidt S, et al. Nature. 2008;451(7181):994-7.
[iv] Olson MV. Am J Hum Genet. 1999;64(1):18-23.
[v] Ramachandran S, Deshpande O, Roseman CC, et al. Proc Natl Acad Sci U S A. 2005;102(44):15942-7.
[vi] Sankararaman S, Sridhar S, Kimmel G, Halperin E. 2008;82(2):290-303.
[vii] Szpiech ZA, Xu J, Pemberton TJ, et al. Am J Hum Genet. 2013;93(1):90-102.