2023-03-24

Comparing Multiple Sclerosis Age of Onset in American American and European Cohorts via Polygenic Risk Scores

Amy Arabel Hernandez
Brigham Young University - Provo

Honors Thesis

COMPARING MULTIPLE SCLEROSIS AGE OF ONSET IN AFRICAN AMERICAN AND EUROPEAN COHORTS VIA POLYGENIC RISK SCORES

by
Amy A. Hernandez Larrazábal

Submitted to Brigham Young University in partial fulfillment of graduation for University Honors

Microbiology and Molecular Biology Department
Brigham Young University
April 2023

Advisor: Mary Davis

Honors Coordinator: David Erickson

ABSTRACT

COMPARING MULTIPLE SCLEROSIS AGE OF ONSET IN AFRICAN AMERICAN AND EUROPEAN COHORTS VIA POLYGENIC RISK SCORES

Amy A. Hernandez Larrazábal

Microbiology and Molecular Biology Department

Bachelor of Science

Health disparities have been observed in autoimmune diseases such as multiple sclerosis (MS), which show that non–Hispanic black patients with MS die at an earlier age and have an increasing mortality trend compared to white MS patients.[3] New genetic information shows promising results to treat genetic diseases. For example, genome-wide association studies (GWAS) are being used to calculate polygenic risk scores (PRS) which are a method used to understand disease risk. However, current PRS have been developed using GWAS that overrepresent individuals of European ancestry,[4] showing that they do not equally apply to all individuals. We set out to examine if PRS could be calculated using 200 previously identified non-major histocompatibility (MHC) variants for increased MS risk in Europeans.[13] We also examined if PRS using those non-MHC variants could be used to generate a PRS for African Americans.

In order to evaluate the capability of PRS to predict MS age of onset for Europeans and African Americans using non-MHC variants, we acquired de-identified electronic health records (EHR) of MS patients that had been genotyped from Vanderbilt

University Medical Center BioVU. We replicated the 200 non-MHC MS risk variants in our European and African American cohorts. We then calculated PRS using the formula $PRS_j=\sum_i Si*G_{ij}$ for each cohort using the software program PRSice.[8] We analyzed the efficiency of the PRS through conducting a linear regression analysis. We also identified the variants that drove this association using linear regression analysis. Our results indicated that generating a PRS by only using non-MHC MS Risk SNPs for MS age of onset in Europeans and African Americans is not sufficient for a significant calculation.

ACKNOWLEDGMENTS

I would like to express my gratitude for the support and mentorship of Dr. Davis through the entirety of this project. Additionally, this endeavor would not have been possible without the undergraduate research grants offered by Brigham Young University Life Sciences College, who financed my research. I would also like to thank my family and friends for their unwavering support and encouragement that motivated me throughout this process. Lastly, I would like to give a special thanks to the Upson family for inspiring me to conduct this research.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

INTRODUCTION

Multiple sclerosis (MS) is a complex autoimmune disease where the myelin sheath that covers nerve fibers deteriorates and causes nerve damage. MS is most common in individuals of European descent; however, it still affects other populations.[2] Previous research studies have found that the average age of onset (AOO) for MS in non-European populations is 28.6 years, while for European populations the AOO is 32.8 years.[2] Furthermore, studies have shown that non–Hispanic black patients with MS died at an earlier age and have an increasing mortality trend in comparison to white patients with MS.[2,3] This suggests that MS takes a different toll on individuals depending on race; understanding why this difference in AOO occurs can aid us in enhancing our knowledge of MS onset, and how to better accommodate non-European individuals with MS.

Genome-wide association studies (GWAS) have increased the field's understanding of complex diseases and the prediction of clinical outcomes. However, the majority of GWAS analyses have been completed using data from individuals of European descent, while admixed individuals or individuals that possess ancestry from two or more genetically distinct sources, are often excluded due to a variety of historical and methodological reasons. Some methodological reasons for this exclusion come from the possibility of false positives (type 1 errors) or the possible reduction of power of the study and small sample sizes.[5] These methodological reasons result from differences in allele frequencies, differences in linkage disequilibrium, and a lack of methods and pipelines that account for ancestry.[22]

Despite awareness of lack of representation in genomic data, not much has changed throughout the years. As of 2018, the GWAS catalog reports that ~78% of all

GWAS participants were of European descent showing a large discrepancy in the representation in human genetic studies which further exacerbate health disparities.[4] The effects of the lack of diversity in human genetic studies can be seen as we compute polygenic risk scores (PRS). PRS have been used as a method to learn about genetic contributions to disease risk; however, the majority of PRS have been developed using GWAS that overrepresent individuals of European ancestry, making them heavily biased toward those individuals.[16]

Furthermore, the lack of diversity makes applying risk criteria across populations difficult when using different populations for the base and target data.[20] Previous research has found that variability of PRS estimates across multiple populations and ancestry groups exists and that scores transfer well across European populations but transfer poorly to individuals of African ancestry.[23] This shows that GWAS are currently less optimal for generating PRS for complex diseases when individuals of different ancestries since they may not be transferable across populations. Considering that GWAS data is likely biased towards Europeans, it is crucial to identify if such data can still be applied to non-European populations efficiently.

In the case of MS, disability accumulates overtime and early detection is crucial because it enables clinicians to initiate disease modifying treatments which have been seen to delay or prevent debilitating disability.[33] We believe that accurate PRS could be used clinically to encourage earlier screenings of MS thus leading to an earlier diagnosis and treatment options. In order to evaluate the capability of PRS to predict MS age of onset for Europeans and African Americans, we extracted previously identified non-MHC Single Nucleotide Polymorphism's (SNPs) associated with MS risk and used them

to calculate PRS. Using linear regression analysis, we evaluated the results and the SNPs that could potentially be driving the association.

MATERIALS AND METHODS

Data Acquisition

Vanderbilt University Medical Center BioVU is representative of patients that come from a variety of ethnicities, health statuses, and ages. BioVU is a resource of over 180,000 blood samples that have been collected from patients to extract DNA. After extraction, each sample is linked to the individual's clinical data in a Synthetic Derivative, a de-identified version of their electronic health record (EHR). The patients' DNA in this study was genotyped on the Illumina Expanded Multi-Ethnic Genotyping Array (MEGA^{EX}) platform. The MEGA^{EX} platform was used to provide genotyping coverage to European, African, East Asian, and South Asian populations.[12] We obtained access to these de-identified EHRs from Vanderbilt University Medical Center BioVU for our study.

Using previously published algorithms,[11] patients with MS were identified by the International Classification of Diseases 10th revision (ICD-10) billing codes associated with MS or demyelinating diseases, medications, and text keywords. We extracted EHRs of individuals of European and African American ancestry (Table 1). We limited our study to patients at or above the age of 18 due to the complexity of MS and lack of pediatric MS cases. The dates recorded for the first occurrence associated with MS or demyelinating diseases were extracted using ICD-10 billing codes and established as the

age at diagnosis (AAD) for each patient. AAD will be used as proxy for age at clinical onset (AOO) because it is the best estimation of initial MS onset that we have.

Table 1. Demographics

	European	**African American**
Females	1109	128
Total Sample Size	1438	171
Average AOO in years (SD)	41.37 (11.54)	38.94 (11.72)
Average Age in 2018 (SD)	54.99 (12.01)	51.56 (12.03)

Table 1. Demographics. This is the total sample size in our cohorts after the original quality control and the average age of onset in each cohort.

Quality Control

Standard quality control (QC) procedures for both the European and African American populations were completed in PLINK. PLINK is a software program commonly used as a toolset for whole-genome analysis; it is designed to perform a variety of functions such as data management, basic association testing, and result annotation.[27] These QC procedures include checking for sex and chromosomal anomalies, sample relatedness, sample genotyping efficiency, and minor allele frequencies. We filtered our dataset by an individual missingness > 0.01 and a minor allele frequency < 0.05. Related individuals were removed.

We checked for sex and chromosomal anomalies using the --*check-sex* command. A male call was made if the homozygosity estimate was more than 0.8, a female call was made if the homozygosity estimate was less than 0.2. No errors with sex checks were found in our dataset. We checked for sample relatedness using the --*genome* command.

We used denser marker data to compute pairwise kinship estimates between every individual.

We checked for sample genotyping efficiency using *--missing* command with a parameter of 0.01. We removed samples with a high proportion (proportion ≥ 0.01) of number of SNPs that failed to meet the missingness cut off and samples for which no genotype was called. We checked the minor allele frequencies (MAF) using the *--freq* command with a parameter of 0.05 for every SNP in our dataset. We removed rare SNPs (MAF ≤ 0.05) as they would reduce the power of our study if they remained in our analysis.

SNP extraction

We set out to identify the 200 previously identified non-major histocompatibility complex (MHC) MS Risk SNPs in our dataset.[17] Of these 200 MS risk SNPs, 78 SNPs were an exact match to those that were in our dataset. In order to find the remaining 122 SNPs, we used the software program SNiPA to obtain functional annotations and linkage disequilibrium information for these MS risk SNPs. SNiPA allowed us to identify SNPs in our dataset that were in linkage disequilibrium (LD) with the original 200 MS risk SNPs. LD patterns were used to identify the remaining MS risk SNPs in our dataset because these patterns allow for the identification of SNPs that co-occur with the actual causal variant.[19] We were able to identify 94 SNPs in our dataset that were in LD with the remaining 122 MS risk SNPs, therefore we were able to identify a total of 172 non-MHC SNPs associated with MS risk in our dataset. 12 of those SNPs were found to be ambiguous SNPs and were removed.

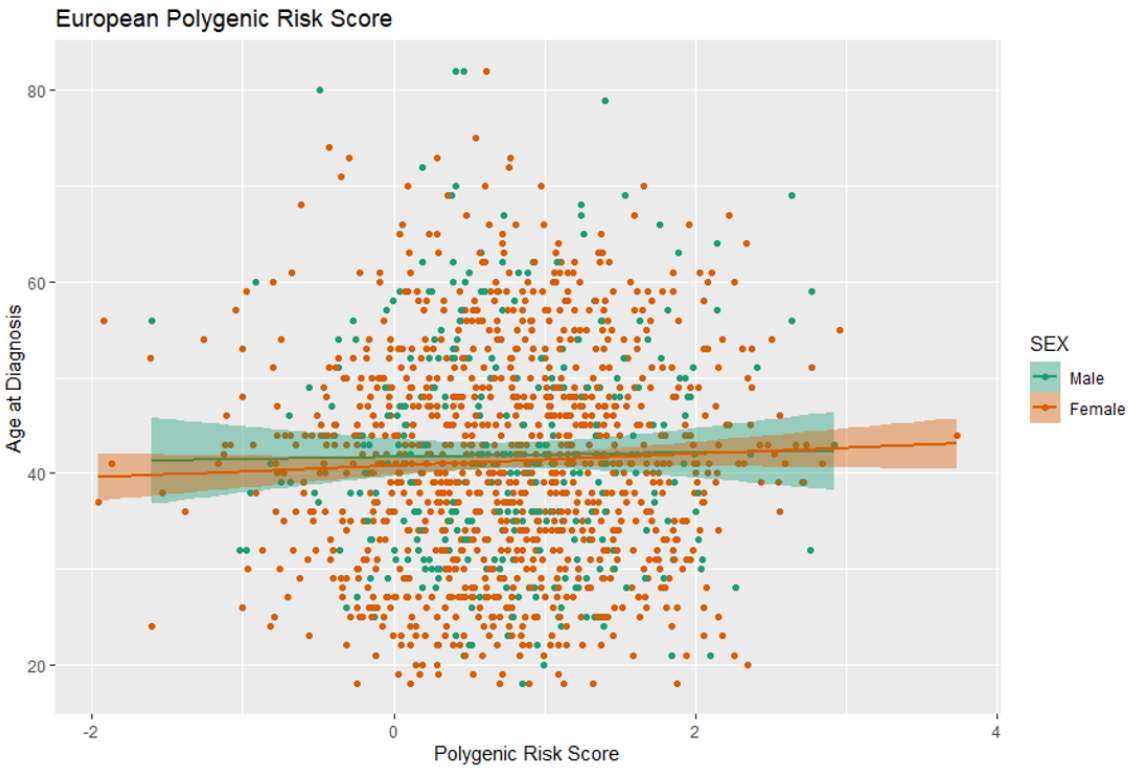

Figure 1. European Polygenic Risk Score
We used this PRS as our control because it was generated using non-MHC MS risk SNPs identified as significant in European populations. There is a weak positive correlation between PRS and age at diagnosis for females and no correlation for males in this cohort, n = 1400.

Polygenic Risk Scores

We calculated polygenic risk scores (PRS) for African American and European individuals via PRSice. PRSice is a software program that is dedicated to the calculation, application, and evaluation of PRS.[8] We used this software program in LINUX. It follows the standard use of the clumping and thresholding (C+T) approach in which SNPs are clumped and thinned according to linkage disequilibrium (LD) as well as P-value.[8] In order to account for differences in LD patterns, clumping is often used to remove SNPs in such a way that weakly correlated SNPs are retained by preferentially retaining the SNPs

most associated with the phenotype of interest. However, since the base data that we used was in the form of summary statistics that had already been pruned for LD patterns, we did not use clumping.

We calculated PRS for individuals in each cohort which contained the 160 of 172 SNPs that were identified by the International Multiple Sclerosis Genetics Consortium and were present in our target dataset.[17] We used the command --*score sum* in order to use the following formula: $PRS_j = \sum i\, S_i * G_{ij}$. In this formula the effect size of the SNP(*i)* is S_i. The number of effect alleles observed in individual *(j)* is G_{ij}.

Our phenotype of interest was age of MS onset. To best estimate this value for our PRS calculation we used AAD, which was determined from the SD for each participant. For our P-value threshold, we set the threshold to include all SNPs with P-values equal to or within the range of 0 to 0.05 for both cohorts. The test statistic that we used to generate effect size estimates was in the form of BETA. Additionally, we used the following covariates: sex, PC1, PC2, PC3, and PC4 to generate a summary file on the best fit PRS. Principal components (PC) are population structure and sample ancestry estimates that are used as covariates to account for differences in population structure and ancestry. To calculate PC, we used --*pca* command with a parameter of 4 in PLINK. We then extracted the first four PC values for each individual in our study.

In order to generate a file that contained a PRS output for every P-value threshold we used the command --*all-score*. We did this so that we could identify the P-value threshold that contained all of our base SNPs. For every PRS that we calculated, we used the following command in LINUX using an RScript: *RScript PRSice.R --prsice PRSice_linux --base baseSNPs --target target_QC_Data --no-clump --binary-target F --*

pheno MS_Phenotype --cov covariate_File --stat BETA --beta --score sum --all-score --out PRSicePRSFile. Additionally, while running the PRS command, 38 Europeans and 9 African Americans were excluded because there was no AAD available for them.

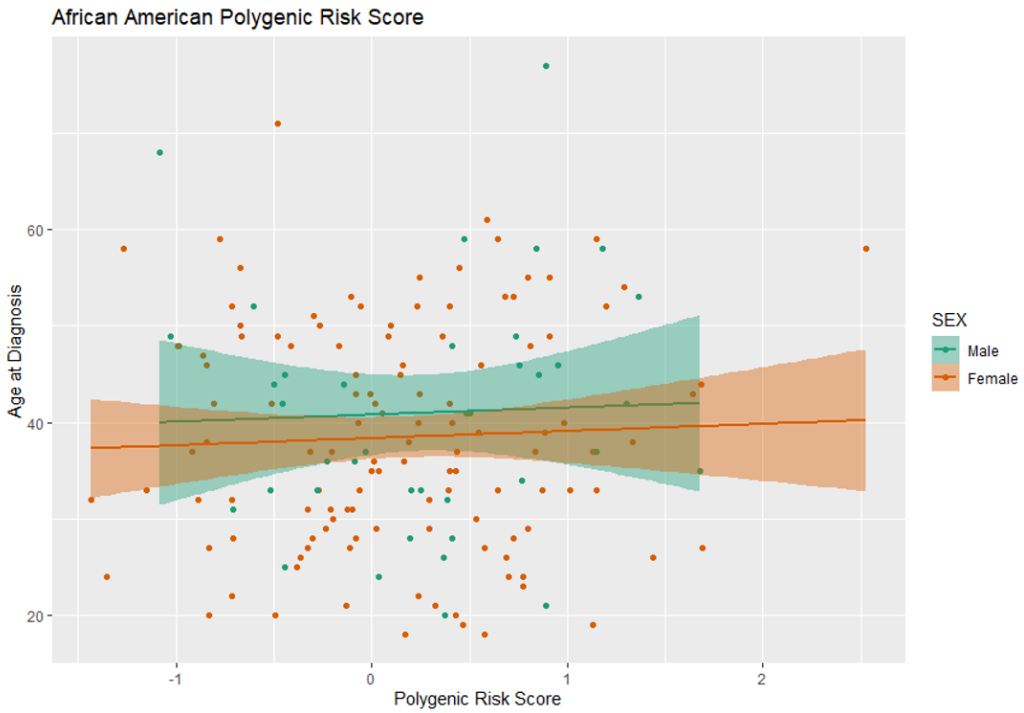

Figure 2. African American Polygenic Risk Score
PRS was generated using the non-MHC MS risk SNPs identified as significant in European populations. For females and males there is a weak positive correlation between PRS and age at diagnosis, n=162.

To identify the efficacy and accuracy of the PRS, we performed a linear regression analysis for AAD. For this linear regression we once again used the covariates sex and PC 1 – 4. This linear regression helped us identify the R^2 value of each PRS and its associated P-value. The following command *AAD ~ SCORE, data=Target_dataset +SEX + PC1 + PC2 + PC3 + PC4* was used in R for all of the PRS cohorts.

Table 2. PRS Results

Cohort	R^2	P-Value
European	0.009	**0.037***
African American	0.023	0.721

Table 2. PRS Results. European PRS was found to be significant (P-value ≤ 0.05) with a Multiple R^2 of 0.009. The African American PRS was not found to be significant and had a Multiple R^2 of 0.023. This means that less than 1% of the variance observed in the target variable is explained by the regression model for Europeans and that ~2% of the variance is explained for African Americans.

Linear Regression for Age of onset SNPs

Upon completion of the PRS, we completed a linear regression analysis via PLINK using the --*linear* command and assessed which SNPs could potentially be driving the PRS and AAD association. We ran this linear regression to find SNPs that had an association to MS AAD (P-value < 0.05). Six covariates were included in our linear regression analysis: age in 2018 (the year when the samples were genotyped), sex, PC1, PC2, PC3, and PC4. The basic linear regression command we used was *AAD ~ MS risk SNP, data=cohortFile** + *age in 2018 +SEX + PC1 + PC2 + PC3 + PC4* (*dataset changed depending on which cohort was examined). Analyses were performed in the African American dataset, as well as the European dataset.

From these analyses, we identified 16 SNPs associated with either an earlier or later MS age of onset. We identified the impact of an MS risk SNP on AAD by the BETA or regression coefficient value from the regression analysis. If the BETA value was negative the SNP was associated to an earlier age of onset. If the BETA value was positive the SNP was associated to a later age of onset.

Table 3. MS Risk SNPs Statistically Associated with Age of Onset in European Cohort

	SNP	BP	A1	BETA	P-value	R^2	PROXIMAL GENE(s)	Onset
1	exm-rs1335532	117100957	G	0.910	0.030	NA	CD58	Later
3	rs1962532	121751531	T	0.718	0.006	1	LEF1	Later
4	rs898518	109016824	C	-0.727	0.006	0.991	ILDR1(dist=24241), CD86(dist=8841)	Earlier
6	rs723054	16670901	C	0.638	0.018	0.992	ATXN1	Later
6	rs498549	137984935	G	-0.767	0.003	0.991	OLIG3(dist=143924), LOC102723649(dist=27328)	Earlier
6	rs1738074	159465977	T	0.533	0.047	NA	TAGAP(NM_152133:c.-4170A>G,NM_054114:c.-799A>G,NM_138810:c.-799A>G)	Later
7	rs1872881	3148092	C	-0.727	0.008	0.803	CARD11(dist=55838),LOC100129603(dist=41148)	Earlier
11	rs4755275	44734310	G	-1.11	0.031	NA	NCOA5(dist=15730),CD40(dist=12596)	Earlier
15	rs7183707	90883733	T	-0.567	0.046	NA	NGRN(dist=72141),GABARAPL3(dist=2179)	Earlier
17	exm2268006	40529835	A	-0.610	0.025	NA	STAT3	Earlier
22	rs2283792	22131125	T	0.724	0.005	NA	MAPK1	Later
22	rs4821544	37258503	C	0.794	0.005	NA	NCF4	Later

Table 3. MS Risk SNPs Statistically Associated with Age of Onset in European Cohort. These SNPs were found to be the most statistically significant (P-value ≤ 0.05) for Europeans with MS in our data set. If BETA is a negative value, we observed an influence of an earlier age of onset. If it is a positive value, we observed an influence of a later age of onset. CHR: Chromosome, SNP: Single Nucleotide Polymorphism, A1: Tested Allele (Minor Allele), BETA: Regression Coefficient. Proximal genes were pulled from Beecham AH, Amezcua L, Chinea A, et al. The genetic diversity of multiple sclerosis risk among Hispanic and African American populations living in the United States. *Mult Scler*. 2020;26(11):1329-1339. doi:10.1177/1352458519863764.

RESULTS

Polygenic Risk Scores

Following PRS QC, 1400 samples with European ancestry, 162 samples with African American ancestry, and 160 non-MHC MS risk SNPs were used to calculate PRS for each individual. Association of MS risk PRS with AAD in Europeans showed that there was a weak positive correlation for females and little to no correlation for males (Figure 1). Association testing of MS risk PRS with AAD in African American males and females found a weak positive correlation between PRS and AAD (Figure 2). However, the PRS results for African Americans was not statistically significant and the R^2 for both cohorts suggested that almost no variance was explained by the model (Table 2). These results suggest that non-MHC MS risk SNPs cannot solely be used to predict AAD or that larger samples sizes are necessary.

Table 4. MS Risk SNPs Statistically Associated with Age of Onset in African American Cohort

CHR	SNP	BP	A1	BETA	P-value	R^2	PROXIMAL GENE(s)	ONSET
1	rs6427518	160391398	T	1.613	0.043	0.919	CD48(dist=22324), SLAMF7(dist=4882)	Later
2	rs12373588	112466265	G	1.579	0.027	NA	MIR4435-1HG (dist=240294), ANAPC1(dist=32228)	Later
12	exm-rs1800693	6440009	C	-1.435	0.028	NA	TNFRSF1A	Earlier
12	rs701006	58106836	A	-1.48	0.028	NA	OS9	Earlier

Table 4. MS Risk SNPs Statistically Associated with Age of Onset in African American Cohort. These SNPs were found to be statistically significant (P-value ≤ 0.05) for African Americans with MS in our data set. If BETA is a negative value, we observed an influence of an earlier age of onset. If it is a positive value, we observed an influence of a later age of onset. CHR: Chromosome, SNP: Single Nucleotide Polymorphism, Proximal genes were pulled from Beecham AH, Amezcua L, Chinea A, et al. The genetic diversity of multiple sclerosis risk among Hispanic and African American populations living in the United States. *Mult Scler*. 2020;26(11):1329-1339. doi:10.1177/1352458519863764.

SNPs associated with MS Age of Onset

We found a total of 16 SNPs associated with AAD as statistically significant (P-value < 0.05). 12 of the 16 MS risk SNPs were found to be significant in our European dataset (Table 3), and the remaining 4 MS risk SNPs were found to be significant in the African American dataset (Table 4).

We identified a statistically significant association between individual non-MHC MS risk SNPs and AAD in African Americans and Europeans. In particular, SNP *rs12373588* was identified to have a later age of onset effect in African Americans with a risk allele having a frequency of 0.271 compared to a risk frequency of 0.473 in Europeans. SNPs *rs1335523* and *rs4821544* were identified to have a later age of onset effect in individuals of European descent. With SNP *rs1335523* having a risk allele frequency of 0.864 in Europeans and 0.490 frequency in African Americans (Table 5). SNP *rs4821544* has a risk allele frequency of 0.635 in Europeans and 0.417 in African Americans. This lower frequency MS risk SNPs associated with a later AAD in African

Americans could contribute to an earlier MS age of onset, which is consistent with previous studies.[2,3] Proximal genes and functions that could be interrupted by these variants can be seen in Table 5. Proximal genes are important to note because the SNPs could potentially be disrupting or influencing that gene's function.[15] However, the distance to that proximal gene should be considered and further studies on the expression of the Gene-SNP need to be conducted.[15]

Table 5. Risk Allele Frequencies and Gene Functions

SNP	CHR	POS	Risk Allele	Reference Allele	European Risk Allele Freq.	African American Risk Allele Freq.	Proximal Gene and function
rs1335532	1	117100957	A	G	0.863	0.490	CD58: Encodes a member of the immunoglobulin superfamily. The encoded protein is a ligand of the T lymphocyte CD2 protein, and functions in adhesion and activation of T lymphocytes
rs12373588	2	112466265	G	T	0.473	0.271	MIR4435-1HG(dist=240294): plays a critical role in the oncogenesis of renal cell carcinoma and may serve as a potential biomarker for renal cell carcinoma ANAPC1(dist=32228): Encodes a subunit of the anaphase-promoting complex. This complex is an E3 ubiquitin ligase that regulates progression through the metaphase to anaphase portion of the cell cycle by ubiquitinating proteins which targets them for degradation
rs4821544	22	37258503	T	C	0.635	0.417	NCF4: The protein encoded by this gene is a cytosolic regulatory component of the superoxide-producing phagocyte NADPH-oxidase, a multicomponent enzyme system important for host defense. This protein is preferentially expressed in cells of myeloid lineage.

Table 5: Risk Allele Frequencies and Gene Functions. These are SNPs associated with an effect on MS age of onset in African Americans and Europeans as well as their frequency for the risk allele in the respective populations. The proximal genes and functions of these genes, functions, and risk allele frequencies were pulled for reference from Beecham AH, Amezcua L, Chinea A, et al. The genetic diversity of multiple sclerosis risk among Hispanic and African American populations living in the United States. *Mult Scler*. 2020;26(11):1329-1339. doi:10.1177/1352458519863764.

DISCUSSION

Our results suggest that using the non-MHC MS risk SNPs would likely need to be coupled with other known or potential AAD risk variants in order to calculate a predictive PRS for European and African American individuals. These results are in line with another study that showed that carrying *HLA-DRB1*15:01,* a known risk factor for MS, and having a greater genetic burden from the 200 non-MHC risk variants is associated with an earlier MS AOO.[7] Considering that MS has a variety of risk factors that could contribute to AAD, it is important to recognize that PRS are not able to capture

the non-genomic variation that contributes to a disease (i.e., sociocultural factors, environmental factors, etc.). Contributions such as complex social disparities and systemic racism have also been suggested to contribute to the clinical heterogeneity in MS and should be considered.[27]

Interestingly, there were three SNPs in particular that we found to statistically correlate with AAD, that have also been found to biologically affect MS onset or disease progression. SNP *exm-rs133532* is proximal to *CD58* gene which was found to mediate both protection from onset of MS and inflammatory demyelination.[12] This is consistent with our association analysis that found *exm-rs133532* to confer a later effect on MS onset. SNP *exm-rs1800693,* which we found to be statically associated with an earlier AAD was also found to be primarily involved in the onset of MS by affecting the tumor necrosis factor pathway.[26] SNP *rs4821544* is proximal to *NCF4* gene which has been found to have a role in the NADPH-oxidase complex that produces reactive oxygen species and has been linked to autoimmune disorders.[25] Additionally, SNPs *rs12373588*, *rs1335532*, and *rs4821544* which we found to be associated with MS AAD in either Europeans or African Americans, were also found to have an effect on MS risk in Hispanic and African American populations.[6]

Our findings further urge the need to examine if current GWAS data can be used to conduct accurate PRS in different populations. Understanding the limitations of current data and the need for diversity in genetic samples should always be considered when conducting genetic studies.

Limitations

This study had several limitations. First, the sample size for African Americans was relatively small. The small sample size of African Americans is partially due to a smaller number of African Americans diagnosed with MS. Second, with multiple testing correction (Bonferroni Adjusted P-value $\leq 3.125 \times 10^{-4}$) the results of the study are not statistically significant; however, the results described in this study are the most significant. Lastly, the AAD for each patient with MS is an estimation for AOO, and variable of this proxy may decrease statistical power. Due to the nature of MS disease progression, most individuals that are affected by MS will not get assessed by a physician upon their first experience of symptoms. This will delay their diagnosis and it is likely that they will not be assessed by medical professionals until their symptoms worsen or become debilitating.[27] However, AAD based on EHR reporting is much more precise than recall biased AOO. Although, these limitations are commonly faced by the genetics and genomics fields when studying MS; they must be considered when examining the results that have been observed in this study.

Future Directions

To further account for ancestry differences, we recommend using software programs that generate ancestry specific genomic segments such as RFMIX or ADMIXTURE to incorporate them into the linear regression analysis as covariates. In addition, further analysis into the SNPs that we identified as significantly associated with AAD should be considered[10,24]. Further biological analysis of the non-MHC MS risk SNPs would also be beneficial. Chromatin immunoprecipitation (ChIP) can be used *in*

vitro to learn about the SNPs effects on gene regulation and effects on binding.[9, 21] Electrophoretic gel-mobility shift assay (EMSA) can also be used to identify *in vitro* DNA-protein interactions and can be used to distinguish sequence variation caused by a SNP.[9]

Animal models such as experimental autoimmune/allergic encephalomyelitis, Theiler's murine encephalomyelitis virus, and toxin-induced demyelination models should also be considered as a method to analyze the effects of the SNPs we found associated with AAD on MS onset.[28] Using animal models could potentially give an insight into biological mechanisms that theses SNPs could be altering or interfering with which could then be contributing to MS pathogenesis.[28]

In regard to the PRS, completing a PRS for AAD using the non-MHC MS risk SNPs previously identified as significant for African Americans would be beneficial and likely give a more accurate idea as to what SNPs, if any are driving an association between the PRS and AAD.[6] Additionally, using the SNPs that we found associated with AAD for African Americans and Europeans should be used to calculate PRS in another sample population and identify if a statistically significant association is observed.

Upon those modifications, it could be beneficial to evaluate the discriminatory effects and utility of the PRS generated. Area Under the receiver operating characteristic (AUROC), precision recall curve (AUPRC), or concordance statistic (C-index) are all analyses that can be used to evaluate the PRS performance and its ability to classify individuals between those who will develop the disorder and those who will not.[31] Additionally, the model's calibrations which gives insight into the predicted risk vs the observed risk could be calculated using Hosmer-Lemeshow test (X^2).[32]

Finally, increasing the sample size of the study would be ideal. Small sample sizes can reduce the statistical power of studies and increase the likelihood of type II errors.[14] Thus, increasing the sample size of the study could reduce the likelihood of such factors influencing the results of the study. Current efforts by the National Institutes of Health (NIH) *All of Us* research program could increase the sample size of this study. The NIH *All of Us* research program is a United States Department of Health and Human Services initiative to increase the diversity of genotyped samples available in a public dataset to researchers in the United States.[1] By using the *All of Us* dataset, increasing the sample size of African Americans is possible.

Implications

As science moves towards precision medicine, a special importance should be placed on increasing studies in non-European populations. One example of this is with the clinical use of PRS in non-European populations. With differences in the demographic relationships, allele frequency, and local linkage disequilibrium patterns among populations, using European-derived studies are not sufficient for precision medicine or adequate patient care. These observed differences lead to limited generalizability of PRS among ethnic/diverse populations and can exacerbate health disparities.

Using accurate genetic data in research studies for a variety of populations with MS could help to identify or promote an earlier age for screening for MS, reduce stigmatization around who is affected by MS, and help reduce MS mortality trends that are based on differences of race, ethnicity, and age.

References

1. All of Us Research Program Protocol. National Institutes of Health: All of Us Research Program. 2021. https://allofus.nih.gov/about/all-us-research-program-protocol

2. Alsaeed MO, Harding KE, Williams OH, et al. Multiple sclerosis: long-term outcomes in ethnic minorities. Analysis of a UK population-based registry. *Eur J Neurol*. 2018;25(4):701-704. doi:10.1111/ene.13571

3. Amezcua L, Rivas E, Joseph S, Zhang J, Liu L. Multiple Sclerosis Mortality by Race/Ethnicity, Age, Sex, and Time Period in the United States, 1999-2015. *Neuroepidemiology*. 2018;50(1-2):35-40. doi:10.1159/000484213

4. Ancestry Category Distribution in the GWAS Catalog. GWAS Catalog. 2017. https://www.ebi.ac.uk/gwas/docs/ancestry-data

5. Atkinson EG, Maihofer AX, Kanai M, et al. Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat Genet*. 2021;53(2):195-204. doi:10.1038/s41588-020-00766-y

6. Beecham AH, Amezcua L, Chinea A, et al. The genetic diversity of multiple sclerosis risk among Hispanic and African American populations living in the United States. *Mult Scler*. 2020;26(11):1329-1339. doi:10.1177/1352458519863764

7. Briggs FBS, Yu JC, Davis MF, et al. Multiple sclerosis risk factors contribute to onset heterogeneity. *Mult Scler Relat Disord*. 2019;28:11-16. doi:10.1016/j.msard.2018.12.007

8. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*. 2020;15(9):2759-2772. doi:10.1038/s41596-020-0353-1

9. Chorley BN, Wang X, Campbell MR, Pittman GS, Noureddine MA, Bell DA. Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. *Mutat Res*. 2008;659(1-2):147-157. doi:10.1016/j.mrrev.2008.05.001

10. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*. 2012;2:80-92. doi:10.4161/fly.19695

11. Davis MF, Sriram S, Bush WS, Denny JC, Haines JL. Automated extraction of clinical traits of multiple sclerosis in electronic medical records. *J Am Med Inform Assoc*. 2013;20(e2):e334-e340. doi:10.1136/amiajnl-2013-001999

12. De Jager PL, Baecher-Allan C, Maier LM, et al. The role of the CD58 locus in multiple sclerosis. *Proc Natl Acad Sci U S A*. 2009;106(13):5264-5269. doi:10.1073/pnas.0813310106

13. Explore genetic variation on a global scale: Research diverse human populations. *Illumina.* 2022. https://www.illumina.com/science/consortia/human-consortia/multi-ethnic-genotyping-consortium.html

14. Faber J, Fonseca LM. How sample size influences research outcomes. *Dental Press J Orthod*. 2014;19(4):27-29. doi:10.1590/2176-9451.19.4.027-029.ebo

15. Folkersen L, Ferdinand HV, Ekaterina C, et al. Association of Genetic Risk Variants With Expression of Proximal Genes Identifies Novel Susceptibility Genes for Cardiovascular Disease. *Circ*. 2010;3:365-373. doi:10.1161/CIRCGENETICS.110.948935

16. Genetics for all. *Nat Genet.* 2019;51:579. https://doi.org/10.1038/s41588-019-0394-y

17. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science*. 2019;365(6460):eaav7188. doi:10.1126/science.aav7188

18. Jafari M, Ansari-Pour N. Why, When and How to Adjust Your P Values?. *Cell J*. 2019;20(4):604-607. doi:10.22074/cellj.2019.5992

19. Joiret M, Mahachie John JM, Gusareva ES, Van Steen K. Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies [published correction appears in BioData Min. 2022 Apr 11;15(1):11]. *BioData Min*. 2019;12:11. doi:10.1186/s13040-019-0199-7

20. Lin M, Park DS, Zaitlen NA, Henn BM, Gignoux CR. Admixed Populations Improve Power for Variant Discovery and Portability in Genome-Wide Association Studies. *Front Genet*. 2021;12:673167. doi:10.3389/fgene.2021.673167

21. Liu X, Noll DM, Lieb JD, Clarke ND. DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res*. 2005;15(3):421-427. doi:10.1101/gr.3256505

22. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities [published correction appears in Nat Genet. 2021 May;53(5):763]. *Nat Genet*. 2019;51(4):584-591. doi:10.1038/s41588-019-0379-x

23. Mars N, Kerminen S, Feng YA, et al. Genome-wide risk prediction of common diseases across ancestries in one million people. *Cell Genom*. 2022;2(4):100118. doi:10.1016/j.xgen.2022.100118

24. McLaren, W, Gill L, Hunt SE, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(122). doi:10.1186/s13059-016-0974-4

25. Olsson LM, Lindqvist AK, Källberg H, et al. A case-control study of rheumatoid arthritis identifies an associated single nucleotide polymorphism in the NCF4 gene, supporting a role for the NADPH-oxidase complex in autoimmunity. *Arthritis Res Ther*. 2007;9(5):R98. doi:10.1186/ar2299

26. Ottoboni L, Frohlich IY, Lee M, et al. Clinical relevance and functional consequences of the TNFRSF1A multiple sclerosis locus. *Neurology*. 2013;81(22):1891-1899. doi:10.1212/01.wnl.0000436612.66328.8a

27. Petracca M, Palladino R, Droby A, et al. Disability outcomes in early-stage African American and White people with multiple sclerosis. *Mult Scler Relat Disord*. 2023;69:104413. doi:10.1016/j.msard.2022.104413

28. Procaccini C, De Rosa V, Pucino V, Formisano L, Matarese G. Animal models of Multiple Sclerosis. *Eur J Pharmacol*. 2015;759:182-191. doi:10.1016/j.ejphar.2015.03.042

29. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575. doi:10.1086/519795

30. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies [published correction appears in Cell. 2019 May 2;177(4):1080]. *Cell*. 2019;177(1):26-31. doi:10.1016/j.cell.2019.02.048

31. Wand H, Lambert SA, Tamburro C, et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature*. 2021;591(7849):211-219. doi:10.1038/s41586-021-03243-6

32. Wang, Yuzhuo, Zhu, Meng, Ma, Hongxia and Shen, Hongbing. "Polygenic risk scores: the future of cancer risk prediction, screening, and precision prevention" *Medical Review*, vol. 1, no. 2, 2021, pp. 129-149. https://doi.org/10.1515/mr-2021-0025

33. Waubant, E. Improving outcomes in multiple sclerosis through early diagnosis and effective management. *Prim Care companion CNS disorders*. 2019;14(5). doi:10.4088/PCC.11016co2cc