**Exome Wide Ancestry Informative Markers selection for Hispanics and Latino Americans**

Rohith Polishetty, Catherine W. Zhang, Yi-Dong Chen, PhD

**Problem Statement**

Race and ethnicity are widely used interchangeable in many populations based research. The ambiguity nature of Hispanic population, due to the genetic admixture with European, African and Native American ancestry, requires additional efforts in genetic studies where appropriate stratification may be considered in order to study the disease and human differentiation. In one of current studies, we performed the whole exome sequencing (WES) to profile genetic alterations in hepatocellular carcinoma (HCC) derived from local Hispanic population in San Antonio. All of them are self-reported as Hispanic. Our objective is to determine their race composition of European, African and Native American ancestry, such that we can confidently conclude the mutations detected in the HCC cohort will be attribute to the disease and uniquely to the population under investigation.

**Introduction**

America has been significantly shaped by the European Expansion which brought people from different origins into contact. Contemporary American genomes have been observed to derive membership from a variety of ancestral populations. This phenomenon is known as admixture where previously isolated parental populations interbreed and introduce new genetic lineages into a population. After a few generations, admixed individuals have a genetic mosaic of ancestry distinguishing them from their parental populations.

Amounting to 17% of the American population and the largest minority group in the United States (and in San Antonio specifically), Latino Americans are an admixture of European, Native American, and African ancestry. According to 23andMe’s database, their Latino American customers have about 70% European ancestry, 14% Native American ancestry and 6% African ancestry; however, these levels of admixture differ significantly in self-identified Latino Americans from state-to-state depending on the Native American density at the time of admixture and the European and African immigration in specific regions. By studying the specific admixture within Latino American subpopulations we can localize disease-linked genetic variants that show differential risk factors, eliminate confounding variables in genetic association studies, uncover disease characteristics as well as provide more accurate information for case-control association studies and precision medicine.

**Abstract**

Past approaches in population structure and admixture analysis were performed in the context of Genome Wide Association Studies (GWAS) and utilized ~100,000 randomly selected single nucleotide polymorphism (SNPs) each with moderate ancestry information to infer population structure within targeted populations. Analysis of several thousand markers is extremely computationally inefficient. This led to the development of ancestry informative marker (AIM) panels where 50 to a few hundred SNPs with high ancestry information were sufficient enough to compute ancestry proportions. However, multi-way admixture poses a problem in traditional AIM selection methods and only recently developed algorithms have produced well validated panels.

Studying admixture in Latino Americans with a three-way admixture has not been practical in past applications due to limited information on their Native American heritage. Consequently, there are few AIMs that have been well validated in the Native American population. In this study we chose to use the East Asian population as a surrogate for the Native American population since they diverged from Native Americans more recently than the divergence of Native Americans from Europeans. This allowed us to expand our AIM search and obtain a well-balanced AIM panel.

Additionally, utilizing whole genome sequencing as a database for SNPs can be extremely costly when only targeting a few hundred markers. Our panel utilized exonic SNPs as a database for our AIMs as Whole Exome Sequencing (WES) is a lot more cost-efficient comparing to Whole Genome Sequencing (WGS).

In this study we have ascertained a panel of 100 ancestry informative markers for determining ancestry proportions in admixed Latino Americans while addressing these limitations and have validated our panel against already published, well validated panels.

**Algorithm**

1. Developing a Primary database of SNPs

Since the application of our AIM panel would be targeted towards the Hispanics/Latinos in the San Antonio area with Hepatocellular Carcinoma (HCC), we started by pruning in-house Whole Exome sequencing data from 18 normal tissue samples in patients with HCC. Initial pruning was done by density of information for each given biallelic single nucleotide polymorphism (SNP). SNPs were eliminated if it had more than 10% missing information across 18 initial patients. We then selected SNPs that were bi-allelic only

**Notes**

*Exclude sites on the basis of the proportion of missing data (defined to be between 0 and 1) where 0 allows sites that are completely missing and 1 indicates no missing data allowed.*

1. Now, ancestry information is needed in order to select for AIMs.

1000 Genome Phase III Whole Genome Sequencing data was extracted by chromosome excluding Mitochondrial, chrY, and chrX[[1]](#footnote-1). 1000 Genome Phase III data is already aligned with hg19 human reference genome; however, they are currently working on aligning it to hg38[[2]](#footnote-2). 1000 Genome SNPs were then pruned by ancestral populations using VCFtools and the appropriate population alignment text files.

The following populations were pulled from the database:

|  |  |  |
| --- | --- | --- |
| **Super Population** | **Subpopulation** | **Count (n)** |
| **East Asian (EAS)** | Han Chinese (CHB), Japanese (JPT), Southern Han Chinese (CHS), Chinese Dai (CDX) | 405 |
| **Europe (EUR)** | Utah Residents with European ancestry (CEU), Toscani in Italia (TSI), Finnish in Finland (FIN), British in England and Scotland (GBR), Iberian in Spain (IBS) | 503 |
| **Africa (AFR)** | Yoruba in Nigeria (YRI), Luhya in Kenya (LWK), Gambian in Gambia (GWD), Mende in Sierra Leone (MSL), Esan in Nigeria (ESN) | 504 |

Individuals from the Caribbean and African Americans were excluded from the ancestral population of Africa due to high levels of admixture observed. The Vietnamese population were also excluded from the East Asian ancestral population.

1. Filtering the appropriate populations

Our original pruned SNPs from the 18 HCC patients were then pulled from the 1000 Genome vcf files after the appropriate populations were filtered out.

Alternatively, you can combine the last two steps into one.

1. **Ancestry Informative Markers selection**

All of the vcf files from the three ancestral populations were then converted to .ped format using VCFtools. Allele frequency is obtained using PLINK for every SNP in each individual ancestral VCF file. Maf is the minor allele frequency. This will output a file called plink.freq. Each ancestral frequency file will be utilized in the AIM selection algorithm. Additionally, we need to stitch all the individual VCF files from the ancestral populations into one large VCF file. Perl scripts require that the VCF files are compressed by bgzip and indexed by tabix.

If the format of the two vcf files being merged are not the same, the vcf files were hand stitched in Excel using the output file from vcf-merge.

Convert the merged .vcf file to PLINK format using the same commands described above. Then use PLINK to calculate genome wide pairwise linkage disequilibrium (r2) across all three ancestral populations. From this AIM selection algorithm and all of our previous filtering we were able to obtain a maximum of 2500 ancestral informative markers.

1. **Basic Ancestry/ethnicity analysis using STRUCTURE**

In this study we primarily used the 100 AIM panel. However, further investigation is being done on panels with larger number of AIMs.

Now that we have our AIM panel, we need to pull our AIMs from the ancestral, admixed[[3]](#footnote-3), and in-house patients. We then merge the ancestral VCF files with the admixed VCF files we want to run in STRUCTURE using the same command line described above.

**Results of running AIM panel through structure:**

****

**Github Link for the whole script:**

https://github.com/rpolishetty/ExomeCapAdmixture

**Software**

VCFtools was used for manipulation and stitching of VCF (<http://vcftools.sourceforge.net)>,

PLINK was utilized to calculate genome wide pairwise

linkage disequilibrium (<http://pngu.mgh.harvard.edu/~purcell/plink>), Galanter Python

Script was used as an AIM selection algorithm (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3869660>: File S1), STRUCTURE was used for inferencing population structure (<http://pritchardlab.stanford.edu/structure.html>),

PGDSpider was used to convert VCF files to STRUCTURE text format (<http://www.cmpg.unibe.ch/software/PGDSpider)>.

1. Reselection of 1000 Genomes data can be done to include chrX in the future. [↑](#footnote-ref-1)
2. Caution for future download of 1000 Genome data. Data must be compatible with our machines (hg19). [↑](#footnote-ref-2)
3. You may choose to leave out 1000 Genomes admixture for individual projects but for further investigation on the accuracy of our AIM panel we utilized admixed Americans from 1000 Genomes. [↑](#footnote-ref-3)