1. **Project PCs onto your Hispanic Sample using 6 reference populations from the 1000 Genomes including CEU, YRI, MXL, PUR, CLM, and CHB**

In your study sample

1. Exclude those SNPs with poor call rates (<90%), or individuals that fail other QC criteria (excess heterozygotes, gender discrepancy, etc).
2. Included common SNPs defined by MAC>10.
3. 1000 Genome population data is always on the + STRAND. So make sure the STRAND is all in the + direction within your data. If some are on the – STRAND, then drop any palindrome SNPs (AT and CG SNPs) and flip the rest of the SNPs to the + STRAND. You can flip them in plink as follows, where “SNPfliplist.txt” is a list of SNPS that need to be flipped:
   * 1. plink --file mydata --flip SNPfliplist.txt --recode --out mydataflipped

Download the plink files for the 1000 Genome reference populations (CEU, YRI, MXL, PUR, CLM, CHB) that have been generated based on the SNPs on the genotyping platform used to genotype your sample. These files are located at

<http://research.mssm.edu/kennylab/referencepops.html>

The reference populations are available for platforms:

Illumina\_660

Illumina\_2.5

Omni\_Express

Omni\_1S

Affy\_6.0

Affy\_5.0

Affy\_Axiom\_LAT

Exome\_Chip

NOTE: These files should include only SNPs common to the specific platform and 1000 Genomes, which do not fulfill any of the following exclusion criteria.

* SNPs on the X and Y chromosome
* Mitochondrial SNPs
* SNPs within the HLA region defined on Build 37 as chromosome 6 between base pair 27,000,000 to base pair 35,000,000
* SNPs within the Lactase Gene region defined on Build 37 as chromosome 2 between base pair 135,000.000 to base pair 137,000,000.
* SNPs on a common Chromosome 8 inversion defined on Build 37 between base pair 6,000,000 and 16,000,000.
* SNPs in a region of long LD in admixed population defined on Build 37 as chromosome 17 between base pair 40,000,000 to 450,000,000.

1. Find all SNPs that are in common between your sample and your platform-specific 1000 Genome reference population plink files and put this list of SNPs names into a text file.
2. Create new PLINK data files for your data and the 1000 Genome reference populations that includes SNPS that are in common in both samples as follows:

*Your sample:*

plink --file mydataflipped --extract commonsnps.txt --recode --out mydataflipped\_commonsnps

*1000 Genome reference populations:*

plink --file Refpops1000G --extract commonsnps.txt --recode --out Refpops1000G\_commonsnps

1. Check that the chromosome and base pairs are the same in the mapping files between your sample data and the 1000 Genome reference population (above called: “mydataflipped\_commonsnps” and “Refpops1000G\_commonsnps”). Change any that are not consistent.
2. Merge your sample data and the 1000 Genome reference population that include only the list of SNPs that are in common as follows:

plink --file mydataflipped\_commonsnps --merge Refpops1000G\_commonsnp.ped Refpops1000G\_commonsnp.map --recode --out mydata\_Refpops1000G

1. Prune the merged plink files containing your sample and the 1000 Genome reference population at r2<0.1 (or r2<0.5 if you are using a platform with more rare variants or less dense number of SNPs, i.e. Exome or Metabochip). Suggestions for pruning in PLINK:

We pruned at a window of 50 SNPs, then shifted 5 SNPs and looked at the next window of 50 SNPs considering R2<0.1 within each window as follows:

plink --file mydata\_Refpops1000G --indep-pairwise 50 5 0.1 --out prunedsnps

Then use the pruned SNP list to create a new set of PLINK data files that are pruned:

plink --file mydata\_Refpops1000G --extract prunedsnps.prune.in --recode –out mydata\_Refpops1000G\_pruned

1. Calculate PCs using analytic software and calculate the percent variance explained by each of the top 20 PCs.

Sample script for Eigenstrat:

genotypename: mydata\_Refpops1000G\_pruned.bed

snpname: mydata\_Refpops1000G\_pruned.bim

indivname: mydata\_Refpops1000G\_pruned.ind

evecoutname: Refpops1000G\_pruned.pca.evec

evaloutname: Refpops1000G\_pruned.eval

outlieroutname: Refpops1000G\_pruned.outlier

altnormstyle: NO

poplistname: Refpops1000G\_individuals.txt

logoutname: Refpops1000G\_pruned.log

numoutevec: 20

numoutlieriter: maxiter

numoutlierevec: topk

outliersigmathresh: sigma

qtmode: 0

* 1. The mydata\_Refpops1000G\_pruned.ind file needs to have three columns including: ID sex Population

(The ID is in the same order as the genotype file, i.e. .bed file).

* 1. The Refpops1000G\_individuals.txt should have the names of the Reference Populations, i.e.:

CEU

CHB

CLM

MXL

PUR

YRI

c. Run Eigenstrat using the smartpca program:

> smartpca –p <name of eigenstrat script>

1. The PCs will output by ID in the .evec file
2. The eigenvalues will output in the .eval file
3. Create a scree plot of the eigenvalues for the 20 PCs showing how much variance is explained by each PC.
   1. This is calculated by dividing each of the top 20 Eigenvalues by the total sum of all Eigenvalues (from .eval file)
4. Plot of PCs by self-identified Hispanic/Latino background with 1000 Genome reference populations
   1. Graph the first 5 PCs against each other by self-identified Hispanic/Latino background including the 1000 Genome reference populations (i.e. PC1 vs PC2, PC1 vs PC3, PC2 vs PC3 etc).

Note: We are aware that the fourth and fifth PCs may not explain much variance, however we would like to request that you look at these plots. If the resulting plots are generally uninformative, you do not need to include these in the plots you share.

* + 1. Identify any possible ancestry outliers.
    2. If outliers are present take note of these outliers for discussion.

(If interested in sample R code for these plots, please email migraff@email.unc.edu).

1. **Generate admixture estimates using the program ADMIXTURE.** These programs can estimate the percentage of Native or African admixture. The program is available at: http://www.genetics.ucla.edu/software/admixture/
2. We can use the same plink files we used for Eigenstrat. We’ll run the program on the full genome-wide data and use the differentiation between ancestries to estimate admixture proportions for each individual in each study.
3. **IMPORTANT:** remove eastern Asian samples from the plink binary mydata\_Refpops1000G\_pruned, unless your study clearly displays eastern Asian admixture (you can look at some of the lower PCs that may differentiate eastern Asians from Native Americans). Most will not. Here’s a sample plink command:

plink --noweb --bfile mydata\_Refpops1000G\_pruned --remove CHB\_indivs.txt --make-bed --out mydata\_Refpops1000G\_for\_admixture

1. Running the program is straightforward. We just need to specify K, the number of clusters:

./admixture mydata\_Refpops1000G\_for\_admixture.bed 3 > mydata\_Refpops1000G\_for\_admixture.k3.log

If you’re confident that there is significant eastern Asian admixture in your study, then use the original files:

./admixture mydata\_Refpops1000G\_pruned.bed 4 > mydata\_Refpops1000G\_pruned.k4.log

If you’re running this on a computer with multiple CPUs, you can add the –jX flag to run the program multithreaded using X threads. It’ll then run X-fold faster. X can be whatever number of jobs your computer can run in parallel.

1. The program will give two output files: a \*.Q file and a \*.F file. We’re interested in the \*.Q file, the one with the admixture proportions. The \*.F file will be much bigger as it includes all the ancestral allele frequencies, which can be used for other things, but we’ll ignore it here. Open the .Q file and see which column corresponds to which ancestry (these get assigned randomly by ADMIXTURE). You’ll need this to run the script. Also, note how many individuals are in your study versus the reference populations.
2. Compare the ancestry estimates for 1000 Genomes with those publically available at the following website. A sample comparison has been provided for your reference (Comparison admixture estimates-Admixture vs 1000 Genomes data.xlsx). Rerun your analysis if notable differences appear.

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis\_results/ancestry\_deconvolution/

1. Plot the output using R and the script plot\_admixture.R
2. **Please email Lindsay Fernández-Rhodes (**[**fernandez-rhodes@unc.edu**](mailto:fernandez-rhodes@unc.edu)**) with a finalized report of your findings.** She will collate these reports prior to the next call. For an example of how to report your findings to the larger consortium, please see summary file (PCA-Mexico City-Hispanic-project.pptx) provided by Esteban Parra for the Mexico City Study.