Instruction for FastPop4

This document gives step-by-step instructions required to carry out the prediction of scores for intercontinental ancestry, make three plots using scores from discovery samples including hapmap samples, new samples, and from combined discovery samples plus hapmap samples and new samples, and calculate the distance metrics among three continental ethnicity definitions (CEU/European, CHB/Asian, YRI/African, and NA/Native American).

Step1:

Your genotype files must be properly cleaned. Please check the options --MIND in PLINK. The option --MIND 0.05 includes individuals with high genotyping at a least 95% complete. We have previously found that if --MIND is <.95, we observe an additional 'population' that reflects poor data quality.
You should keep the given 2318 markers because the SNP weights are generated from the specific 2318 markers.
Before running R-package, you have to create data with SNP genotypes recoded in terms of additive components which can be input file.
Please check strand flips and allele flips in your data. We use **"FORWARD" strand.** Also, it is critical to specify the reference allele with the attached file "**ref_2318SNPs.txt**" using the following command in PLINK;

plink --file your_file (with extension .map & .ped) --reference-allele ref_2318SNPs.txt --recodeA --out your_output(which will have extension .raw in your working directory)

In case of binary file, replace --file your_file with --bfile your_file(with extension .bed, .bim, &.fam)

If you would like to check **strand flip**, please use "**flip_top_updated.txt**".

From the uploaded plot, "**PCAplot.png**", three plots indicate that both discovery data and new data have no strand flip and the same reference alleles. In addition, when you see the scales on PC1 and PC2, they should be the same. Please check the log file from Plink before running R-codes.

More information on PLINK can be found here:
http://pngu.mgh.harvard.edu/~purcell/plink/
http://pngu.mgh.harvard.edu/~purcell/plink/dist/plink-doc-1.07.pdf

Step2:

Please first run the function “**Fastpop4_PredictioinPCAScoring.R**” to get the prediction of scores given by 2318 SNP weights. These SNPs were chosen from SNPs showing high FST values and differentiating well among continental populations. Also, it helps to visualize and compare new scores with discovery samples plus 601 Hapmap including European (red), Asian (green), African (blue), Native American (purple), and Mexican (dark green) samples chosen to indicate each centroid on four Continental definitions.

Step3:

The last function “**Fastpop4_Distance.R**” calculates the probability of each continental ethnicity definition from scores of new data.

If you have any question or encounter problems, please send an email to:
Jinyoung.Byun@bcm.edu , Younghun.Han@bcm.edu or Chris.Amos@bcm.edu