Admixture Analysis of the Environmental Genome Project

The primary goal of this project is to set up gPLINK and FRAPPE and use them to estimate ancestry proportions from exome data.

Data

1. EGP exome data (egp.bed, egp.fam, egp.bim), N=80
2. Maya SNP array data (maya.bed, maya.fam, maya.bim), N=21

Software

1. Plink
2. gPlink
3. ADMIXTURE

Steps

1. Overall HINTs:
   1. “Ask questions, make mistakes, get messy” – Miss. Frizzle
   2. Never use the same name for your input and your output as it will override what you have done and if you make a mistake …
2. Download and install PLINK/gPLINK: <http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml>
   1. Double click on gPLINK which should open the interface
   2. Select a folder where you have put the data (*Project/Open*)
   3. Set the path to plink.exe (where it is on the computer). Use *Project/Configure*
3. Merge the EGP plink files with the Maya data. *PLINK/Data Management/Merge Filesets*
   1. This will throw an error. Don’t worry ;)
   2. Some SNPs will have the wrong strand and need to be flipped with a special command (--flip). Use *Advanced/Create PLINK command* on the maya data.
   3. Add the flags --flip filename\_from\_3a.missnp --make-bed in place of “{insert PLINK command here}”
   4. As a help, I have prefiltered the AT/GC SNPs.
   5. Try the merge again with the new maya files.
4. Filter out SNPs with a missingness of 10% or that are mono-allelic. *PLINK/Data Management/Generate Fileset*  under Threshold
   1. HINT: Use any non-zero minor allele frequency threshold, but make sure it is lower than 2/202. Kudos for those that tell me why ;)
5. Use *Advanced/Create PLINK command* with --indep-pairwise to perform LD pruning
   1. HINT: I usually use 50 SNP windows, with a 5 SNP step, and an r2 threshold of 0.5, e.g. --indep-pairwise 50 5 0.5
   2. This may take a bit of time.
6. “Extract” the correct SNPs for a combined and filtered plink file set. Use *PLINK/Data Management/Generate fileset* under Filter.
   1. Output as PLINK binary format (bed, bim, fam)
7. Download and Install ADMIXTURE: <https://www.genetics.ucla.edu/software/admixture/download.html>
   1. In a new folder copy the final egp/maya
8. Run the command
   1. ./admixture <finalEPGfile>.bed <number of K>
   2. This will output a finalEPGfile.Q and finalEPGfile.P.
      1. The Q file has the admixture proportions for each individual (same order as the fam file) for the various clusters (K)
      2. The P file has the estimates of the allele frequencies for the variants in the various K.
      3. Both are tab delimited
   3. Identify the clusters of individuals and select the cluster with individual “1800” in it.
      1. You can use paste to combine the fam and Q files.
   4. You should be able to import the results file into Excel and sort on the clusters.