**BIOMI 609 Computational Genomics and Bioinformatics**

**Spring 2022**

**San Diego State University**

**Dr. Arun Sethuraman**

**Midterm Exam 2**

**Due via Canvas on 4/12/2022 (Tuesday) at 11:59 PM**

1. Tracing the evolutionary origins of HIV-1

Read this paper by Faria et al., 2014 (also on Canvas), where they utilize phylogenomics to understand the origins and spread of HIV-1. DOI: [10.1126/science.1256739](https://doi.org/10.1126/science.1256739). Now you will analyze data from this paper that I’ve uploaded as a multiple sequence alignment (466.fas), which is the envelope protein, aligned from across 466 strains from across the world (a full list of accessions, and their origins and times of sampling are also uploaded to Canvas). Now you will use RAxML that we have discussed in lab (refer to Lab 3 Phylogenomics) to construct a phylogeny using the GTRGAMMA model. Interpret these results now - particularly tracking the origins of viral transmission (25 points).

I just used this command, since my HIV sequences are already aligned.

raxmlHPC -s 466.fa -m GTRGAMMA -p 12345 -n hiv\

This should take some time to run and produce an optimal phylogeny which you should be able to view using some tool like FigTree. You’re welcome to “prettify” it as you see best. But the idea is to see the origins of populations of HIV, in comparison with the Kinshasa ones based on their phylogenetic resolution.

Extra Credit (10 points):

Build another multiple sequence alignment by adding sequences to this alignment from other countries other than the ones in this study (you can do this by just using BLASTN, then downloading more samples that match back). Then rebuild the phylogeny, and interpret those results.

This is subjective on the samples you’ve used - so there’s no one solution. I’ll grade these based on what you’ve done.

2. Write a piece of code to simulate genotype frequencies at a single bi-allelic locus (two alleles are A and G) in an inbreeding population. Assume that the population is inbreeding at the rate of F = 0.2. Assume that the starting population is of size 1000 diploid individuals, and is in HWE, with frequency of the A allele = 0.2. Plot the frequencies of AA, AG, and GG individuals in this population over 50 generations (25 points).

To do this, remember that there are two things at play - 1) genetic drift, since the population size is small (N = 1000), and 2) inbreeding, at the rate F = 0.2. So the strategy here is to sample allele frequencies based on drift, then obtain the expected genotype frequencies using the inbreeding coefficient equations. I’m just reusing code from the assignment to do this, with added functions to calculate and plot the genotype frequencies.

F=0.2 #inbreeding coefficient

N=1000 #Size of the diploid population

allfreqs<-c() #Empty array to store allele frequencies

allfreqs[1]=0.2 #Allele frequency in the first generation, set to 0.2 here = p

genfreqsAA<-c() #Empty array to store genotype frequency of AA

genfreqsAG<-c() #Empty array to store genotype frequency of AG

genfreqsAA[1]<-0.2\*0.2 #expected frequency under HWE is p^2 in the first generation

genfreqsAG[1]<-2\*0.2\*0.8 #expected frequency under HWE is 2pq in the first generation

plot(1,type="n",xlim=c(0,50),ylim=c(0,1),main="N=1000",xlab="Time in generations", ylab="Genotype Frequencies")#Creating an empty plot

for(sims in 1:10) { #Just repeating the below simulations 10 times

for(t in 1:50){ #Loop over 50 generations

gen<-sample(c(0,1),2\*N,replace=TRUE,prob=c(allfreqs[t],1-allfreqs[t]))

#Sampling with replacement (simulating drift)

allfreqs[t+1]<-length(which(gen==0))/(2\*N)

#Compute allele frequency in next generation

#Now compute genotype frequency expected in the next generation

genfreqsAA[t+1]<-allfreqs[t+1]^2+ allfreqs[t+1]^2\*(1-allfreqs[t+1]^2)\*F

genfreqsAG[t+1]<- 2\*allfreqs[t+1]^2\*(1-allfreqs[t+1]^2)\*F

}

points(genfreqsAA,col="red",type="l") #Plot AA freqs

points(genfreqsAG,col="blue",type="l") #Plot AG freqs

points(1-genfreqsAA-genfreqsAG,col="green",type="l") #Plot GG freqs

}



Et voila! You clearly see that heterozygosity has basically tanked to 0 almost instantly due to inbreeding. And the genotype frequencies are fluctuating, but they’re still entirely maintained as homozygotes in the population.

3. Population structure of a resident song sparrow - *Melospiza melodia*

Read this paper by Mikles et al., 2020 (also posted on Canvas): <https://doi.org/10.1111/mec.15647>

Thereon, you will analyze the VCF file provided for you from this paper to (1) estimate and plot genome-wide estimates of genetic diversity (pi) as a Manhattan Plot, (2) run ADMIXTURE analyses on the data, determine the most likely number of subpopulations and then plot the ancestry proportions. Compare your results with those of the Mikles et al., 2020 paper, and interpret your results (15 + 15 = 30 points).

I used vcftools to compute site-based diversity estimates:

vcftools --vcf sparrows.vcf --site-pi --out sparrowspi

Then I had to just replace all the “Contig” and “pilon” text in the first colum of the output file, to be able to plot this using the qqman package in R.

sparrowspi<-read.table("sparrowspi.sites.pi",header=TRUE)

manhattan(sparrowspi,chr="CHROM",bp="POS",p="PI",snp="POS",logp=FALSE,ylab="Pi")

Chart, scatter chart

Description automatically generated

From the looks of it, there don’t seem to be any visual outlier loci, but we’ll have to do some further tests to be able to detect anything.

I also used the code to use plink2 to convert the VCF file to a BED format that is given as input to ADMIXTURE. Looking at the cross validation errors, K = 2 seems to best fit the data. So I went ahead and plotted it:

Chart, bar chart, histogram

Description automatically generated

Tracking the ancestry of the red colored subpopulation, these individuals correspond to M. m. pusillula, while the other populations are relatively of greater blue ancestry, with some signatures of admixture.

4. What is going on at the genomic position marked by the arrow? Explain your answer with respect to Tajima’s D, LD, and Variability (differentiation, say measured as Fst - 10 points).



There are several things going on here - you see a lower value of Tajima’s D and correspondingly a lower degree of genetic diversity. This could potentially indicate a selective sweep at this position, resulting in fixation of an allele in a population. When this occurs, everything around this locus that are under linkage disequilibrium will also sweep to fixation, therefore resulting in a steep drop in diversity in and around this locus. You can also see this in the plot on the LD line (peak).

5. The MC1R allele is responsible for red hair color in humans, and is present across the global human population such that the observed homozygosity at this locus is 0.704. Assuming that this global population is randomly mating and that this locus is biallelic (MC1R, and not MC1R), what would be the expected frequency of individuals that are NOT red-haired? (10 points) HINT: You will have to solve a quadratic equation, and recall that the solution(s) to a quadratic equation of form ax2 + bx +c = 0 is given by (10 points)

If I assume that the allele is completely recessive, individuals that are homozygous for this allele will be red-haired, and everyone else will not be red-haired. So the expected frequency of non-red-haired folks in the population would be p2+2pq if this population is in HWE. It’s given that the homozygosity at this locus is 0.704 = p2+q2. Note that q = 1-p, so substituting this, we have:

p2+(1-p)2 = 0.704

p2+1-2p+p2 = 0.704

2p2-2p+1 = 0.704

2p2-2p+0.296 = 0

So here, a = 2, b = -2, c = 0.296; solving, we have p = (2)/4 = 0.82 or 0.18; hence the expected frequency of non-red-haired individuals in the population would be 0.9676.