**P/BIO 381 Spring 2017**

**Assignment #3: Population genomic diversity and structure**

Your assignment is to assess the sensitivity of our inferences of population genomic structure in our SSW data for two different SNP filtering strategies (not including the one we’ve used in the tutorials). You may want to consider some, though not necessarily all, of the following variables for filtering the SNP data:

* presence of multiple alleles/locus (--min-alleles; --max-alleles)
* read depth (--minDP)
* site missingness (--max-missing)
* minor allele frequency (--maf)
* deviation from Hardy Weinberg equilibrium (--hwe)
* removing individuals with large amounts of missing data (--remove-indv)

You should then analyze the 2 resulting datasets with one of the 3 population genomic techniques we’ve covered: PCA, DAPC, or ADMIXTURE.

Please use 2 pages maximum to demonstrate your understanding of the conceptual background and technical details for using SNPs derived from RNA sequencing to analyze population diversity and structure. You should include relevant tables or figures (within the two-page limit) with legends.

* Clear statement of objective (1 sentence).
* Conceptual background on what the analysis does (2-3 sentences).
* Verbal description of the mechanics of the pipeline (3-4 sentences).
* Present results (3-5 sentences).
* Tables and figures with legends.
* Interpretation (3-5 sentences). ***Critique the filtering strategies and their effects on your inference;*** ***place your interpretation of the structure of the SSW data into context based on what you know of their natural history and the sampling design.***
* Critical thinking (2-3 sentences). What would you do differently? What would you do next?
* Include a link to your code on github!

You may discuss the assignment with classmates, but the assignment should be prepared individually. Due **Wednesday, April 5th**.

Filters:

*Biallelic vs. multi-allelic SNPs:* Keep only sites with 2 alleles.

Rationale: When looking at diversity within species, it's very rare to have mutations occur at the same position. So, SNPs with >2 alleles probably reflect sequence or mapping errors. We also want to get rid of SNPs showing <2 alleles.

$ vcftools --vcf filename.vcf --min-alleles 2 --max-alleles 2

*Minor allele frequency (MAF):* Gets rid of very rare SNPs (based on a user-defined threshold).

Rationale: Sequencing errors are relatively common, but they tend to happen randomly and affect only 1 read at a time. Thus, if we have a SNP that is only seen very rarely, it may be a sequencing error, and should be discarded. For us, the most liberal MAF filters would be 1 allele copy out of the total 2N copies, or 1/48 = 0.02

vcftools --vcf filename.vcf --maf 0.02

*Missing data across individuals:* Get rid of sites where fewer than 80% of our samples have data.

Rationale: Missing data is a problem for any analysis, and population genetic statistics can behave oddly (i.e.. become biased) when a lot of individuals are missing data for a given SNP.

$ vcftools --vcf filename.vcf --max-missing 0.

Combining filters:

$ vcftools --vcf filename.vcf --min-alleles 2 --max-alleles 2 --maf 0.02 --max-missing 0.8 --recode --out ~/biallelic.MAF0.02.Miss0.8

VCFtools also can provide output in the form of many useful summary stats on a vcf file. Let's look at the observed and expected heterozygosity for each SNPs and test if any violate Hardy-Weinberg equilibrium expectations: **(1=p^2 + 2pq + q^2)**. Use the quality-filtered file we generated above as input.

$ vcftools --vcf filtered\_filename.vcf --hardy