**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)

*Biallelic vs. multi-allelic SNPs:* Keep only sites with 2 alleles. $ vcftools --vcf filename.vcf --maf 0.02

* 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.
  + Keep only sites with 1 allele??
* site missingness (--max-missing) : removes SNPs with a high amount of missing data

$ vcftools --vcf filename.vcf --max-missing 0.8 #used in class for 80%

* 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.
  + Me: Get rid of sites where fewer than 70% of our samples have data
* 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**.

**From Laurent et al. 2016:**

“Besides true variation, these initial variant calls contain false positives due to systematic sequencing artefacts, mismapped reads and misaligned indels. Such false-positive calls often (i) exhibit excessive depth of read coverage, (ii) show an allelic imbalance, (iii) occur preferentially on a single strand, (iv) appear in regions of poor read alignment and (v) arise in unusual close proximity to multiple

other variants.” (i) Three or more variants were found within 10 bp (clusterWindowSize = 10). (ii) The depth of coverage at the given position (summed across individuals) was <500 or >3000 (DP<500; DP>3000). (iii) There was evidence of a strand bias as estimated by Fisher’s exact test (FS > 60.0) or the Symmetric Odds Ratio test (SOR > 1.0). (iv) The read mapping quality was low (MQ<80). (v) At least one of the samples was not called (NCC > 0). After applying the initial filter criteria, the variant data set was limited to biallelic sites using VCFtools (Danecek et al. 2011). Genomic positions that fell within repeat regions of the reference assembly were excluded because erroneous alignment of reads to these regions often leads to an increased frequency of heterozygous genotype calls. In particular, five different classes of repeats (i.e. LINE, LTR, DNA, simple repeats and lowcomplexity regions) were annotated using REPEATMASKER (Smit et al. 2013–2015) and variants within these regions were excluded from further analyses. To minimize genotyping errors, all variants with either missing data for any individual or genotype quality of less than 20 (corresponding to P[error] = 0.01) for any individual were excluded using VCFtools.”

**From Gompert et al. 2014:**

When identifying variants, we used the full prior with h set to 0.001, and we ignored aligned sequences with phred-scaled mapping quality <20 and bases with phred-scaled base quality <13. We only designated genetic variants at nucleotide sites where we had sequence data for at least 80% of the sampled butterflies and where the posterior probability of the sequence data under a null model that the nucleotide was invariant was <0.01. We estimated global variant MAFs directly from genotype likelihoods using the maximum-likelihood approach described by Li (2011) and used these estimates to assign loci to MAF classes. We identified 801 218 SNVs based on these criteria, which included 28 701 common variants (MAF ≥ 5%), 180 043 low-frequency variants (0.5% ≤ MAF < 5%) and 349 497 rare variants (0.1% ≤ MAF < 0.5%). We ignored very rare variants (i.e. MAF < 0.1%) as these occurred in only one or two individuals. The median sequencing depth per individual per variable site was 7.6 reads (mean = 8.5 reads, standard

deviation = 4.5 reads).

**From McCoy et al. 2014:**

To therefore reduce potential false positives, we modified our pipeline to allow only one mismatch per

aligned read. We then used a hard filter to extract potential false SNPs with at least one sample sequenced

to ≥109 coverage with reads supporting both alleles and >75% of reads supporting the reference allele. We

likewise extracted putative true SNPs for which all samples were sequenced to ≥109 coverage, and any individual with nonzero counts of each allele had an allele balance between 30 and 70%.

**From Rominguier et al. 2014:**

Then polymorphic positions were filtered for possible hidden paralogues (duplicated genes), using a likelihood ratio test based on explicit modelling of paralogy16. The across-species average percentage of SNPs discarded for suspicion of paralogy was 7.65%. Positions at which a genotype could not be called in a sufficient number of individualswere discarded.Calling k the number of sampled individuals for a given species (2#k#11), the minimum number of genotyped individuals required to retain a position was set to k/2 when k.5, to k21 when k54 or k55, and to k when k,4. Control for contamination. Each locus of each species was translated to protein and

1. For Homework #3 that you're working on, I wanted to clarify a couple of things. First, when you go to generate your new filtered vcf files based on your chosen filtering strategy, you should use the original unfiltered vcf file (7.4 million SNPs) as your starting point with vcftools. The path to this file on the server is:

/data/project\_data/snps/reads2snps/SSW\_by24inds.txt.vcf.gz

2. I had suggested using depth (number sequence reads per SNP) as one of your possible filters. However, I discovered that during the original SNP calling process I did using the 'reads2snps' algorithm, all SNPs were called based on a minimum read depth of 10, but after that the depth info was not retained in the vcf file. This means you won't be able to filter on depth, and instead should choose one of the other variables to filter. That still leaves many interesting possibilities to try (e.g., # alleles, MAF, removing SNPs with high amounts of missing data, removing individuals with high amounts of missing data).

Keep vs remove part: look at vcftools page