Erin Keller

Assignment #3

Ecological Genomics

April 5, 2017

In addition to information on gene expression, sequence data also detects thousands of single nucleotide polymorphisms (SNPs) that can be used for population genomic studies. Using multi-locus SNP data, it is possible to determine population genomic structure and diversity by looking at the frequency of SNPs among individuals. To determine the population genomic structure of *Pisaster ochraceus*, the program ADMIXTURE (Alexander et al. 2009) was used to determine the probability of an observed genotype belonging to a particular cluster; this program uses a likelihood-model to quickly establish ancestry coefficients that are then used as parameters for the model by simultaneously estimating allele frequencies and ancestry. The program generates cross-validation values for each value of K (potential number of populations) which are indicative of the most accurate model; this allows the user to identify the most likely number of populations the samples originate from.

SNP data from 24 individuals, comprised of 8 individuals from the subtidal site and 16 individuals from the intertidal site, was extracted by using the program read2snp and then filtered. Filtering was done using the program VCFTools which allows the user to exclude data that does not fit the desired criteria. All data was reduced to bi-allelic data only by setting the minimum and maximum number of alleles to 2; mutations at a single site are rare and, therefore, sites with more than 2 alleles may be the result of sequencing error and not be representative of actual biological phenomena. In addition, the data was then filtered with one of the following: the removal of data that deviates from Hardy-Weinberg equilibrium at a significance level of p < 0.01 or removal of data with rare alleles (allele frequencies < 0.03) as very rare alleles may be the result of sequencing error.

The VCF files generated from the filtering process were then converted to a .geno file needed for analysis in ADMIXTURE using the PGDSpider2 program. All analyses were done using K=3 with 10 iterations. K=3 was chosen as these sea stars are known to be somewhat panmictic due to their free-spawning strategy so it is unlikely that there would be many populations. Q plots (Figures 1a-b) were then generated from the results using R, the code for which can be found in Entry 20 of my online lab notebook.

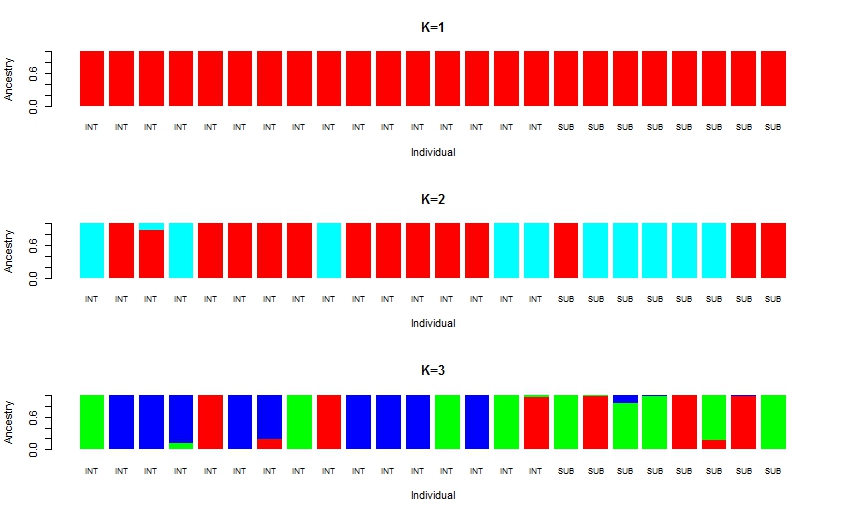
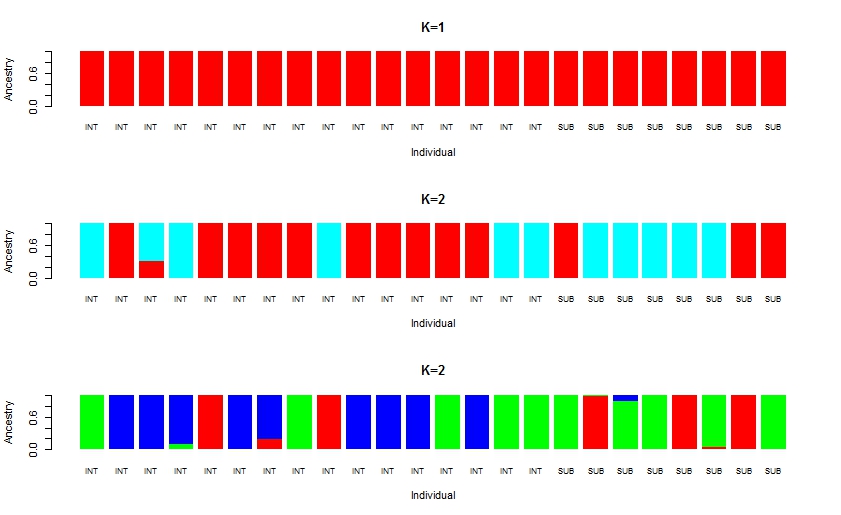
The results from the ADMIXTURE analysis indicate that the sea star samples all belong to a single population, as the cross-validation (CV) error is lowest when K=1. A low CV error indicates that the particular K being evaluated is the most accurately predictive model (Alexander et al. 2009). In addition to K=1 having the lowest CV error, analysis of the Q plots indicates that K=1 is the most likely number of clusters from which the sea star samples originate (Figures 1a-b). Q plots report the probability of ancestry (i.e. belonging to a particular population); therefore, the model of K for which samples have highest probability of ancestry is supported.

The results from the analysis indicating that the samples come from a single population, with no apparent substructure of subtidal and intertidal zones make biological sense, as sea stars are free-spawning organisms that release gametes into the water. Sea star populations can resemble panmictic populations as they have far spreading gametes and the random fertilization from free-spawning. In terms of filtering strategies, removing data that did not meet the filtering criteria being assessed produced a data set of similar number of SNPs (~5000) for both strategies. Removing data that deviate from Hardy-Weinberg equilibrium may lead to an underrepresentation of population divergence and diversity as loci with heterozygote excess or deficiency are excluded; however, removing only loci that deviate most significantly from Hardy-Weinberg equilibrium does not produce a different best-supported K than when Hardy-Weinberg filtering was not applied. Removing rare alleles may also reduce genetic diversity and mask population divergence but again the expected and supported value of K=1 was obtained.

In the future, it may be informative to isolate SNPs that significantly deviate from Hardy-Weinberg equilibrium and map them to a gene, as the gene may be under selection; if a particular SNP has a high frequency it may be due to selection acting on the gene and increasing the beneficial SNP. I would also reduce the minimum allele frequency to be excluded to 0.02, as that would indicate the frequency of an allele if only one individual in the sample had one copy of the allele. Sequencing error can create sporadic false-alleles but the probability of falsely generating the same allele due to sequencing is highly unlikely, thus, observing the same allele twice in the sample may represent true genetic diversity and not sequencing error.

Table 1: Cross-validation (CV) errors generated from ADMIXTURE for each value of K evaluated (K=1-3) for each model.

|  |  |  |  |
| --- | --- | --- | --- |
|  | CV Error | | |
| Filtering Strategy | K=1 | K=2 | K=3 |
| HWE | 1.16830 | 1.30136 | 1.20248 |
| MAF | 1.33780 | 1.49090 | 1.37265 |



a. b.

Figure 1: Q plots generated from ADMIXTRURE analysis for all K evaluated (K=1-3) for both filtering strategies: a. HWE- data significantly deviating from Hardy-Weinberg equilibrium (P<0.01) excluded b. loci with allele frequencies <0.03 excluded.

Alexander D.H., J. Novembre, and K. Lange. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19:1655–1664.