Notes for Mary

Kristen Gruenthal

23 Aug 2017

From my draft of the cod paper: “Raw data were quality filtered and demultiplexed and sequence alignment, SNP discovery, catalog construction, and genotyping were performed in Stacks v1.21 (Catchen et al. 2011; Catchen et al. 2013) according to the methods of Gruenthal et al. (2014), with minor modification.”

1. Compressed data in \*.fastq.gz format were downloaded from the UOGCF website, and the raw reads were quality filtered and demultiplexed using the subprogram process-radtags (run\_process\_radtags\_b; flags set: -e sbfI –c –q –E phred33). Don’t forget to trim at least the last basepair of the reads using –t (like I did…oops). Confirm with Dan or Lorenz how much you guys are trimming now.



1. Demultiplexed sequence files were automatically renamed with the original IDs using a bash script (e.g. mv\_AD06\_samples).



1. I used the denovo\_map.pl wrapper program (denovomap\_RADrun2\_readfiltered) because I assumed no reference genome (i.e. I didn’t want to use Atl cod). Briefly, catalogs created in the cstacks subprogram were generated from the five most data-rich individuals from each sample. Flags (m = 3, M = 2, N = 4, n = 3, max\_locus\_stacks = 3) associated with increasing the number of loci, while reducing the SNP and allele calling error rates, were set according to Mastretta-Yanes et al. (2015).



1. A genotype file containing putative polymorphic SNPs present in ≥ 80% of fish per sample was filtered to include one SNP per RAD tag (flag: write\_random\_SNP) to minimize physical linkage, and a genotype file was output in Genepop format (flag: genepop). If you use write\_single\_snp, I believe it outputs the first SNP in the tag. If you want to filter for the most polymorphic SNP in the tag, you would have to filter outside of Stacks based on allele frequency (e.g. as part of step 5).



1. Final filtering removed loci with minor allele frequencies (MAFs) < 0.05 to minimize sequencing errors, as well as loci with uncorrected Hardy-Weinberg equilibrium (HWE) p-values ≤ 0.05. Allele frequencies and HW p-vals were generated in Genepop and parsed with perl. Then, I actually filtered by loading the Genepop file into Excel, but you should use whatever program with which you are comfortable. Obviously, you can filter by other metrics here, as well, like other measures of failed individuals or populations.





1. On the last step, I filtered out fish with really high relatedness using ML-Relate (Kalinowski et al. 2006), which can indicate accidental resampling or contamination. Lorenz has since filtered out more based on relatedness for concerns of contamination.