Proceed with RADseq pipeline - speak with Dylan before doing so (try trimming - FASTQC (run before and after trimmomatic or Illumiprocessor or TrimGalore! (on raw data)

Ask Dylan if: barcodes on cavefish are the same as barcodes in document he gave me 1) if my data have adapters I need to account for 2) Is it one lane? Should I put in all samples from sheet you sent me? Why are those different from the sample sheet on Illumina? 3) show Dylan what I have so far, + look at log text file and interpret based on Rochette2017 4) how is coverage? 5) why do samples 8 and 12 have any retained reads? 6) can’t form primary stacks (too low coverage, why are so many reads being thrown out, why can’t find barcodes in so many – could it have to do with too much low quality sequence or uncalled nucleotides – for whatever reason, low percentage of sequence is being recovered - why?) why are sequences within sample different length?) 9) How long are the sequences? 10) Get a better text editor to view quality information from demultiplexing?

KBTO samples are from 25 through 39

6/01/2019 plate was what was used for library prep

Stacks couldn’t form any primary stacks with loci across samples to ultimately identify SNPs, so back to re-normalizing DNA. Next time isolating DNA, dry 240 nanograms of DNA and dilute in 12 microliters of water. Increasing concentration and more rigorously ensuring standardization help?

Neither my nor my advisors’ sequences had good enough quality after out-of-house Illumina sequencing.

Get in touch with Bayona-Vasquez and Catchen re: situation where can’t form primary stacks because only 50-100 reads per individual where barcode was found, 3/4 all sequences were thrown out because barcodes not found (due to high quantity of uncalled nucleotides). How do I allow for more mismatches at the process\_radtags phase than just 2 using **--barcode\_dist\_2** setting? Should I try **--disable\_rad\_check**?

Why do I have to allow for so many mismatches/why are there so many uncalled nucleotides in the barcodes? I know that’s normal and that quality increases as you get farther along in sequence. Shouldn’t do force\_diff\_len (not even sure why lengths vary within individuals, though understand across individuals due to varied length of barcodes). Is it overall sequencing quality? Maybe try Apparently, barcode quality/initial part of sequence will be low-quality but it will get better/correct as it gets farther into sequence. Should I re-normalize or try increasing mismatches (which I’m having trouble doing – tempted to email Bayona-Vasquez, Catchen, Pietro Mello for help re: this). How do I truncate paired-end reads of 300+/- 3 and 150+/- 3 lengths (vary based on barcode) to 290 and 150 to 140 nt, or would it work to force different lengths at **denovo\_map.pl** stage? How do I enter parameters to rescue sequences with an increased number of mismatches in the barcodes and trim lengths of sequences? Ask on google page for stacks.

Current script:

process\_radtags -1 /Users/jeffreylernercoleman/Desktop/KBTO/RadSEQ/stacks-2.5/jeff/Undetermined\_S0\_L001\_R1\_001.fastq.gz -2 /Users/jeffreylernercoleman/Desktop/KBTO/RadSEQ/stacks-2.5/jeff/Undetermined\_S0\_L001\_R2\_001.fastq.gz -b /Users/jeffreylernercoleman/Desktop/KBTO/RadSEQ/stacks-2.5/jeff/barcodes.lane1.tsv -o ./cleaned/ -c -q -r -t --barcode\_dist\_2 --inline\_inline --renz\_1 xbaI --renz\_2 ecoRI

*Data from each project were assembled independently using Stacks v1.40, v1.42, or v.1.44 (Catchen et al., 2013; Catchen et al., 2011). We used the process\_radtags program to demultiplex and/or clean and trim the sequence data. We removed reads with an uncalled bases (-c) and discarded reads with low quality scores (-q) with a default sliding window of 15% of the length of the read and raw Phred score of 10. We specified XbaI and EcoRI as restriction enzymes, and we rescued sequence tags (internal indexes) and RAD-tags (enzyme over-hang) within 2 mismatches of their expected sequence (-r); otherwise, reads were discarded. We truncated (-t) PE150 reads to 140 nt and PE75 reads to 64 nt to have equal length among all reads with different barcodes.*

*We parallel-merged the mates of paired-end reads using the ‘paste -d’ Unix command. We used the Stacks denovo\_map pipeline for each project with the following settings: the minimum number of identical raw reads required to create a stack (-m) = 3, the maximum distance between stacks (-M) = 3, and the number of mismatches allowed between samples when generating the catalog (-n) = 4. Coverage, number of loci, and number of SNPs recovered were scored for each species and compared to genome size and sequencing read length (PE75 or PE150). Next, we used the Stacks populations program with the following settings: the minimum populations per locus (-p) = 60–75% of total populations and the minimum individuals (within a population) per locus (-r) = 75%.*

Email to Taylor Hains:

My toucan project is on hold right now, as I have to go back to the lab to re-do the ddRAD library prep, as there was some issue with the sequencing quality such that by the time of de novo assembly in STACKS, very few reads were recovered. I worked on it a bit recently - I tried increasing the allowed mismatches in the barcodes as well as trimming all my reads to the same size (140bp) to see if I could capture more at that initial stage so as to see more downstream. What happened unfortunately is that whereas when the parameters are set to default (two inline barcodes with no trimming and recapturing of 2 mismatches), it’s able to demultiplex, but not when I specialize the commands - it pretends like my data are single- instead of paired-end). I even spoke with the STACKS creators as well as the developers of the 3RAD method I used (standard ddRAD but with a third restriction enzyme that binds to and gets rid of most adapter-dimers). Anyway, yes, I’ll be re-doing the benchwork - I know this is a TON of detail, but hopefully I’ll finish my project (one day?) But given that you’ve used enough methods within the realms of RADseq and microsats with relevant frameworks to make confident sibship and parentage reconstructions and strategic pairings, I’m sure your project is good to go (and even if not, while I’m sure you want it to be as rigorous as possible, you haven’t started your PhD yet... haha, but I see you with your high standards :) ) !

Follow up with Taylor Hains re: genome sequence, use trimmomatic to try trimming everything to 140bp