Research Log

Your Name Here

June 19, 2019

Contents

1 Reference Genome Assembly

SUMMARY

Here is a summary of some results from genome assembly attempts.

NEXT TO-DO

1. Item 1

LOG

6 March 2018

An entry from March 6th, with some code on running qualimap.

ANOTHER SUBSECTION

Here are some details in another subsection of information.

2 Genotyping-by-sequencing

SUMMARY

Here is a summary of some results from my GBS analyses.

Table 1: Number of variants (SNPs) for the pilot genotyping-by-sequencing run for Lake Tanganyika *Lates* under different alignment and filtering conditions. Individuals were excluded from variant calling if they had fewer than 5,000 reads (5k) or 20,000 reads (20k). Raw SNPs are before filtering for missing data, minor allele frequency, and proximity to other SNPs (90bp threshold); these SNPs have already been filtered using bcftools based on QUAL $>$ 19 and GQ $>$ 9, and only SNPs (not indels or sites with $>$ 2 alleles) have been kept. The labels "with WGS" and "no nilo" refer to datasets containing whole genome re-sequencing data and omitting *Lates niloticus* individuals, respectively.

			Missing Data Allowed		
Reference Genome	Raw SNPs	MAF	≤ 0.5	≤ 0.3	≤ 0.2
Lates calcarifer	4,085,115	0.01	70,180	59,144	52,397
		0.05	33,854	28,991	25,641
Lates mariae	153,175	0.01	19,025	14,542	12,091
		0.05	10,541	7,721	6,205
Lates mariae, no nilo	139,881	0.01	17,422	13,334	10,832
		0.05	10,527	7,845	6,318
Lates mariae, no nilo, with WGS	3,076,112	0.01	22,247	14,933	12,252
		0.05	12,438	8,388	6,770
Lates mariae, with WGS (5k)	3,087,655	0.01	23,641	16,019	13,406
		0.05	12,124	8,228	6,619
Lates mariae, with WGS (20k)	3,086,364	0.01	24,907	17,275	14,619
		0.05	13,002	8,898	7,398

NEXT TO-DO

1. Item 1

LOG

24 Feb 2018

An entry from Feb 24th, with a figure (Fig. 1).

ANOTHER SUBSECTION

Here are some details in another subsection of information.

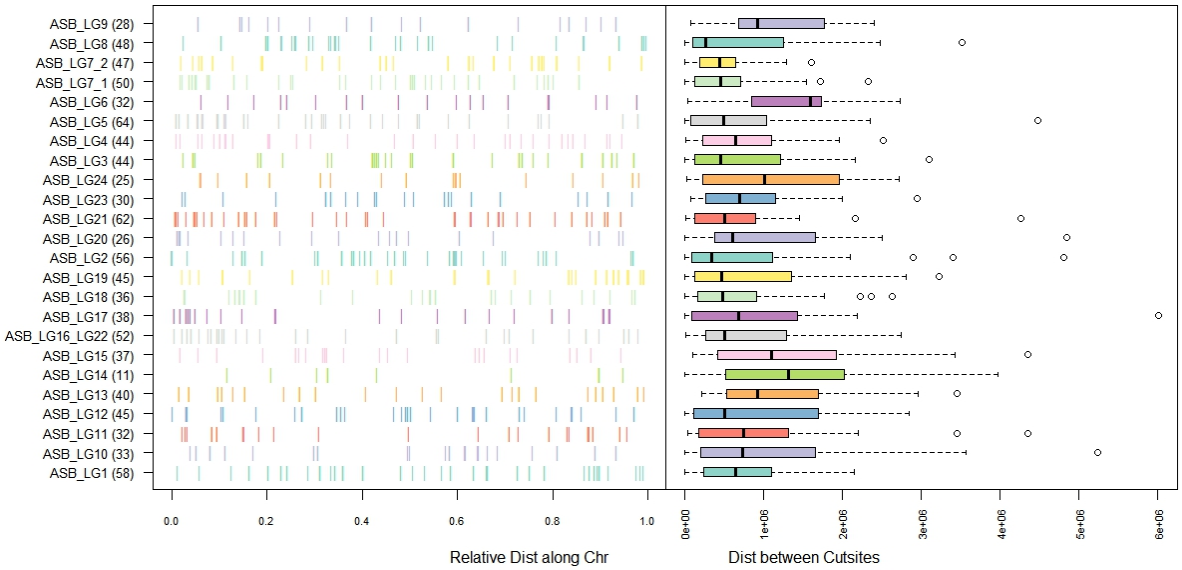


Figure 1: Plot of the location of cutsites shared by the GBS and RAD methods (where cutsites for SbfI occur within 150bp of a location where EcoRI and MseI cutsites fall within our size selection criteria of 200-350bp). Left-hand panel shows the spacing of potential shared SNPs along the chromosomes. Right-hand panel shows boxplots of the distances between these potential SNP locations, to give an idea of how useful these potential shared SNPs may be.

3 Phylogenetic Analyses

SUMMARY

Here is a summary of some results from phylogenetic analyses, and some pretty figures.

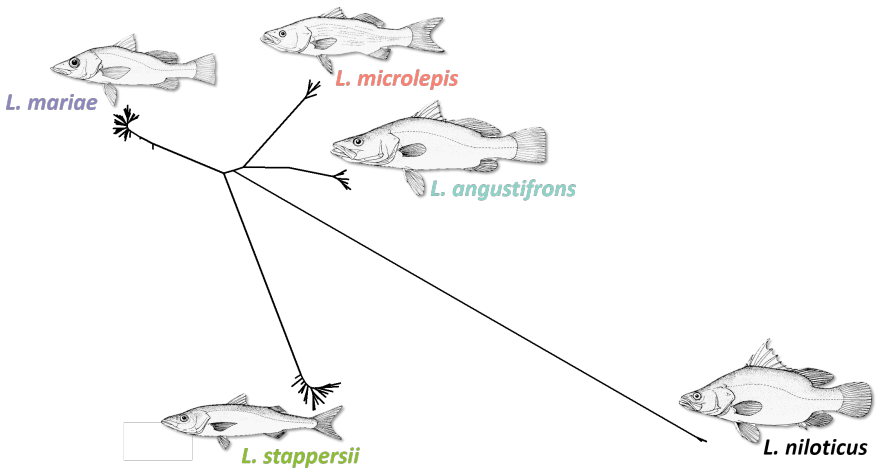


Figure 2: Preliminary RAxML phylogeny based off of *Lates* GBS data, aligned to the *L. calcarifer* genome, and following filtering for $\leq$ 20% missing data and MAF $\leq$ 0.01. These SNPs were not filtered for linkage disequilibrium.

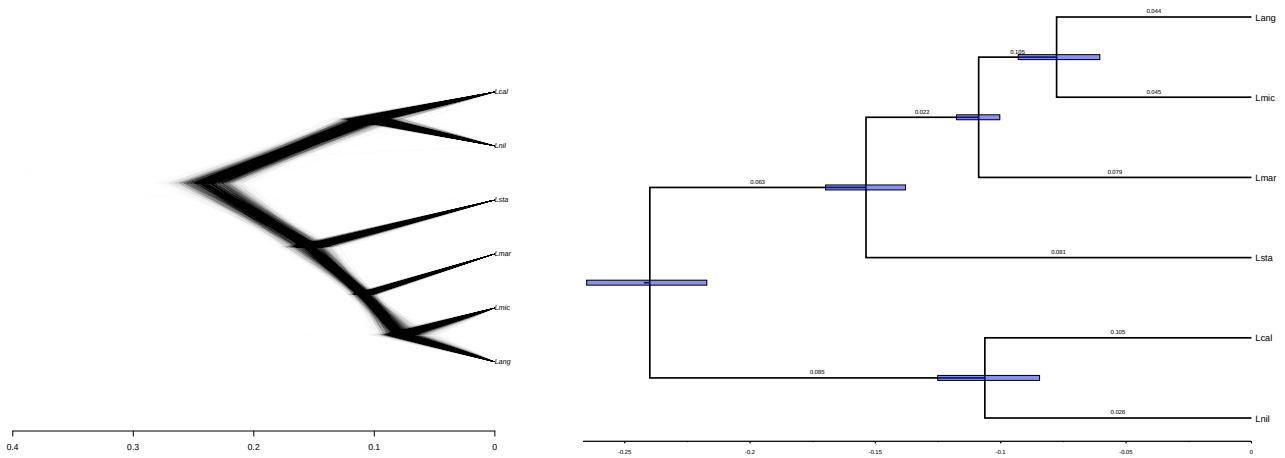
NEXT TO-DO

1. Item 1

LOG

5 March 2018

Here's an entry with a two-part figure.



(a) Densitrees from posterior

(b) Consensus tree

Figure 3: Trees from `SNAPP` runs, followed by summary in `TreeAnnotator`, run for 3 million MCMC iterations using parameter file `snapp_calcar_thin50k_small_030318.xml`. Data are SNPs called from GBS reads aligned to the *L. calcarifer* genome (v3, chromosome-level), thinned to a minimum distance of 50,000bp between SNPs and with invariant sites removed (4,696 SNPs total). Values on branches show the median theta value, and purple bars show 95% HPD for node depths.

Appendices

A SNAPP Workflow

example of a tutorial

This is an example of a tutorial made to share with others.

My working directory, including scripts: `/jrick/latesGBS_2018/snapp`

Starting from a VCF file of filtered SNPs (*note*: SNPs should be filtered for linkage disequilibrium in some way– I chose to just set `--thin 50000` in my vcftools specification), you can create a phylip file using Joana's `vcf2phylip.py` script. In the call for this script, `-r` indicates that you want the **r**eference included in the output; `-e` indicates that you want indels **e**xcluded.

This can either just be called, or can be run using Chad's shell script `slurm_convertVCF_CDB.sh`.

Next, we need to remove invariant sites from the phylip file and turn it into a nexus, which `BEAUti` requires as input. Chad has two different methods of doing this (you can ask him about that), but I couldn't get either to work so I just wrote an R script for it myself. The R package `phrynomics` seems to have some issues on Mt Moran (and even some on my own computer), so I have been running the R script `RscriptSNPs_removeInvar.R` on the phylip file on my computer. This both removes invariant sites and converts it to a nexus.

NOTE: You'll want to make sure you have a reasonable number of taxa/individuals in your nexus file that you're using for SNAPP, as it gets unwieldy with too many. I chose to keep a "random" subset (i.e. those with the highest coverage) of a maximum of 5 individuals/taxon for each of my 5 taxa.

Once you have a `NEXUS` file, you can load it into `BEAUti` to create the XML input for `SNAPP`. `BEAUti` is on Mt Moran (included in the `BEAST` module), but has a GUI, so ends up being easier to use locally instead. One thing that may make your life easier prior to loading the nexus file into `BEAUti` is if you edit your individual IDs to include "_species" at the end of the name (or somehow include taxa IDs in the names; e.g. I added "_Lmar", etc. at the end of my individual IDs). You can set the priors for all of the different parameters in `BEAUti` – I'd recommend chatting with Chad on how to figure out what to set these to, if you're unsure. I've been starting with 3 million steps for the MCMC chain, storing every 3000 steps.

Once I have the XML file from `BEAUti`, I copy it to Mt Moran, and then use `run_snapp.sh` to run `SNAPP` on the cluster. The general command is:

For some reason, the output folder (named `RUN_NAME`) has to be created before it will let you start the run. In this command, `threads` and `instances` need to both be equal to the number available in your SLURM request. In the script, I set the seed randomly by using the computer clock to make sure that I don't accidentally start multiple replicate runs with the same seed.

BONUS: if you want to check whether your data are overcoming your priors, you can start a run that samples only from the prior by adding `sampleFromPrior="true"` to your MCMC specification:

B appendix section b

subtitle for this appendix

This would be another appendix section