Final Project: Demographic history and patterns of molecular evolution from whole genome sequencing in the radiation of Galapagos giant tortoise

Tyce Schneider, ISG5312

Information/Background

- Dataset originated from *Demographic history and patterns of molecular evolution from whole genome sequencing in the radiation of Galapagos giant tortoise*, by Jensen et al (2021)
    - BioProject: PRJNA761229
- The aim of this paper was to use whole genome sequencing data from all extant species of Galapagos tortoise to elucidate population diversity among species and to reconstruct demographic history.
- Galapagos tortoises were chosen by the researchers due to the populations being closely related and the extensive existing background knowledge regarding these populations.

Sample Collection and Library Preparation

- Researchers selected samples from an archive of samples collected from previous studies
- 3 individuals were selected from each of the 12 extant lineages of Galapagos tortoise (the 11 named species, including both the PBL and PBR lineages of *C. becki*), and a closely related outgroup species (*Chelonoidis chilensis*)
- Once DNA was extracted, shotgun sequencing libraries were prepared and sequenced on an Illumina NovaSeq 6000 sequencer

Methods: Sequence Processing and Alignment

- The first step of my reanalysis involved trimming out the Illumina adapters using Trimmomatic
- Trimmed libraries were then aligned to the *Chelonoidis abingdonii* reference genome
  - The authors acknowledged that using an in-group reference genome can bias downstream analyses, but the very recent divergence (~6 mya) among all the species studied minimizes this bias, and there is no evidence that one species has better alignment than the others
- The genome was indexed and aligned using bwa-mem2 (ver 2.1)

Methods: Variant Calling

- Variant calling was carried out using bcftools (1.19) using the parameters outlined in the original publication
  - -q 25: ignore any bases with base quality <25
  - -Q 25: ignore any reads with mapping quality <25
- A 1000 kb window file was created to run call variants in parallel
- The resulting VCF file was filtered using vcftools (v0.1.16)
  - Indels were removed
  - Filtered for genotypes supported by a minimum depth of 6
  - Filtered for genotypes with a minimum genotype quality score of 18

Methods: Variant Calling Continued

- Vcftools (v0.1.16) was also used to calculate measure of heterozygosity on a per-individual basis for filtered file and the missingness on a per-individual basis
- The filtered file was further processed using plink (v1.9)
  - Pruned out loci in linkage disequilibrium using a sliding window size of 50 kb, step size of five loci, and r2 threshold of 0.5

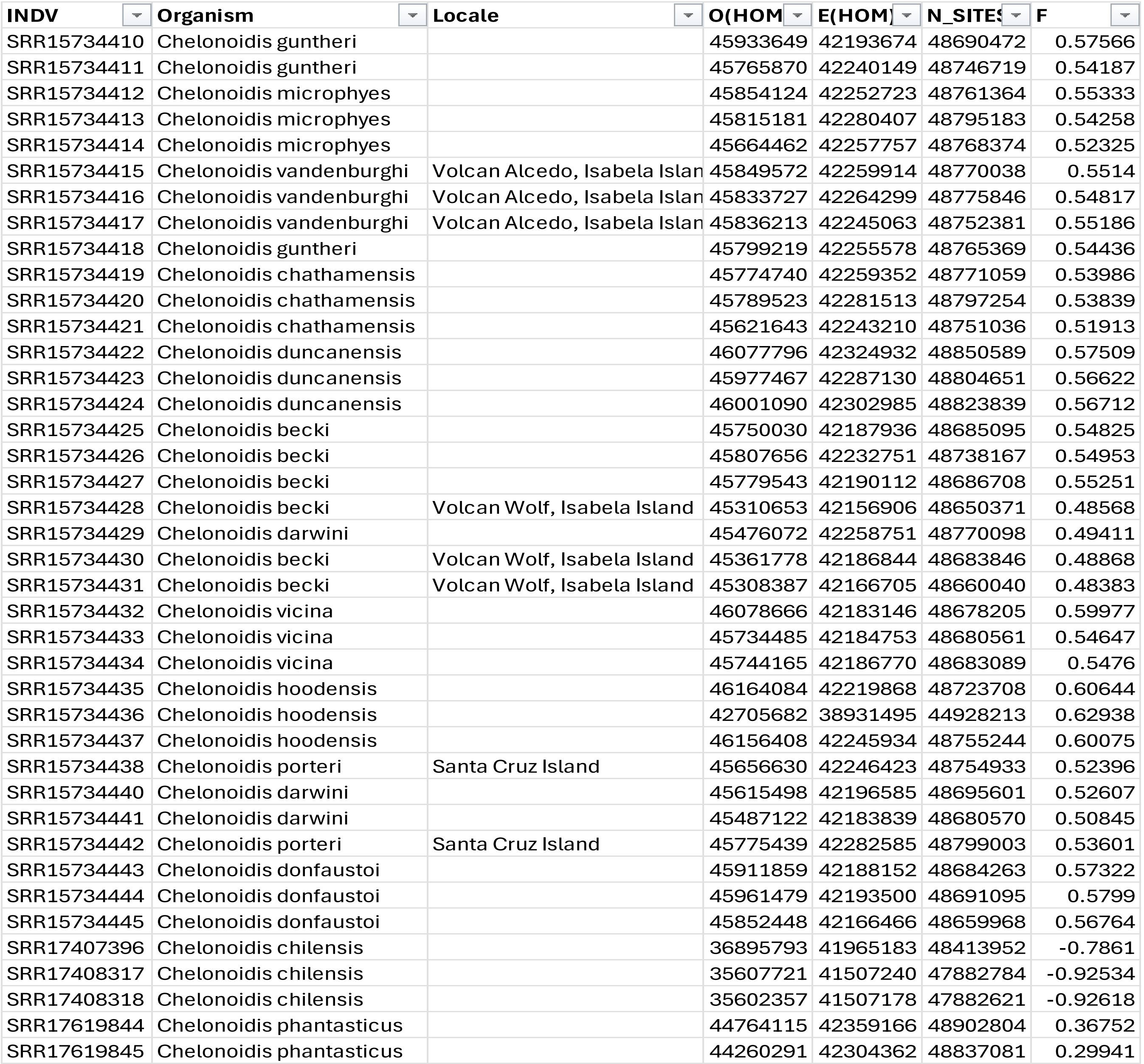
Methods: Data Analysis

- The distinctiveness of the 12 previously recognized Galapagos giant tortoise lineages was re-assessed using principal components analysis, also using plink (v1.9)

Results: Heterozygosity

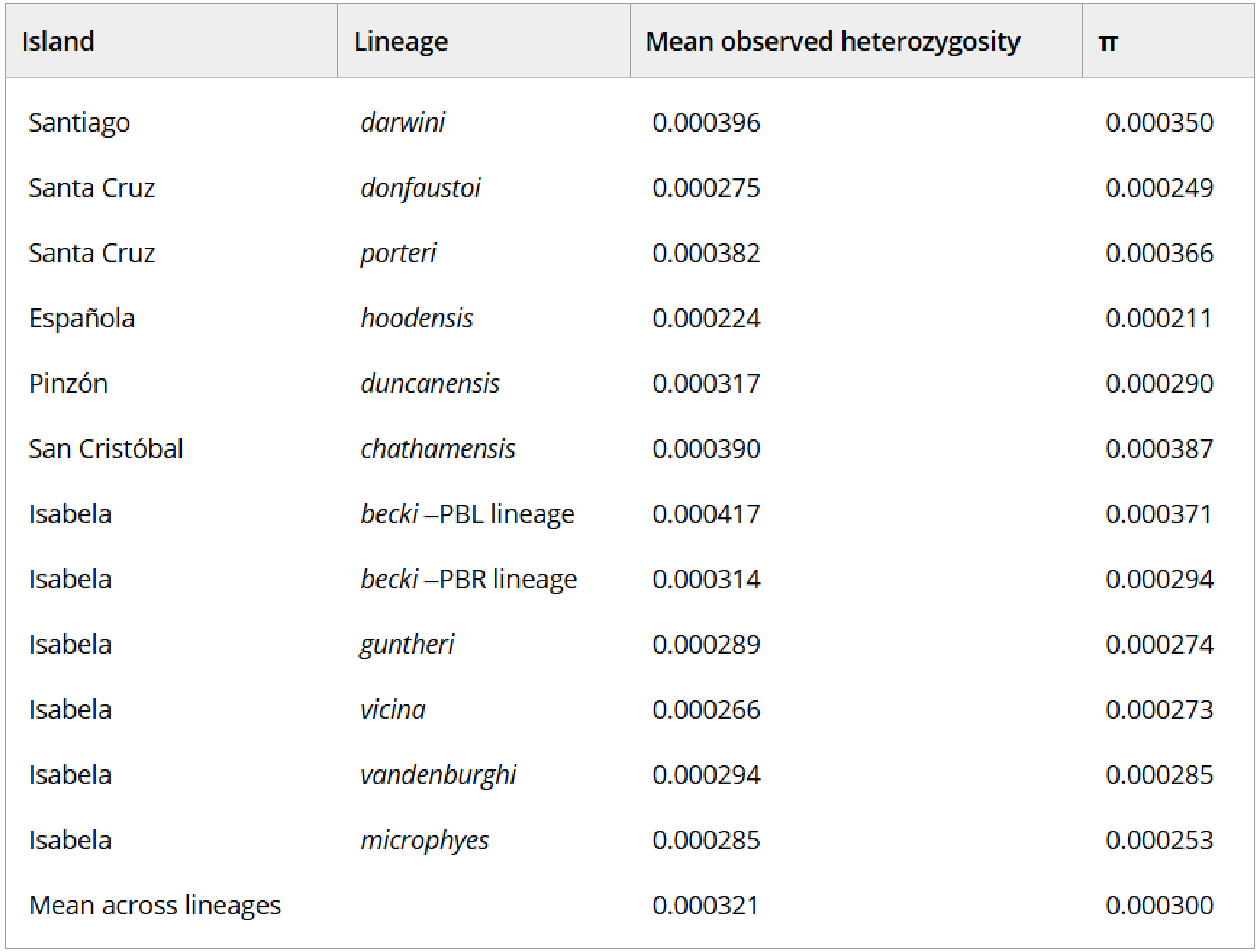
- Mean heterozygosity (reported as F) was similar across lineages, but there were some outliers
  - The out-group species (*Chelonoidis chilensis*) had a mean heterozygosity of -0.879, which makes sense given its distance from the reference genome
- *Chelonoidis phantasticus* exhibited the lowest heterozygosity of the in-group, with a mean heterozygosity of 0.333
- *Chelonoidis hoodensis* exhibited the highest heterozygosity of the in-group, with a mean heterozygosity of 0.612
- These findings are inconsistent with what was reported in the original publication

Table 1: Reanalysis Mean Heterozygosity



INDV	Organism	Locale	O(HOM	E(HOM)	N_SITES	F
SRR15734410	Chelonoidis guntheri		45933649	42193674	48690472	0.57566
SRR15734411	Chelonoidis guntheri		45765870	42240149	48746719	0.54187
SRR15734412	Chelonoidis microphyes		45854124	42252723	48761364	0.55333
SRR15734413	Chelonoidis microphyes		45815181	42280407	48795183	0.54258
SRR15734414	Chelonoidis microphyes		45664462	42257757	48768374	0.52325
SRR15734415	Chelonoidis vandenburghi	Volcan Alcedo, Isabela Islan	45849572	42259914	48770038	0.5514
SRR15734416	Chelonoidis vandenburghi	Volcan Alcedo, Isabela Islan	45833727	42264299	48775846	0.54817
SRR15734417	Chelonoidis vandenburghi	Volcan Alcedo, Isabela Islan	45836213	42245063	48752381	0.55186
SRR15734418	Chelonoidis guntheri		45799219	42255578	48765369	0.54436
SRR15734419	Chelonoidis chathamensis		45774740	42259352	48771059	0.53986
SRR15734420	Chelonoidis chathamensis		45789523	42281513	48797254	0.53839
SRR15734421	Chelonoidis chathamensis		45621643	42243210	48751036	0.51913
SRR15734422	Chelonoidis duncanensis		46077796	42324932	48850589	0.57509
SRR15734423	Chelonoidis duncanensis		45977467	42287130	48804651	0.56622
SRR15734424	Chelonoidis duncanensis		46001090	42302985	48823839	0.56712
SRR15734425	Chelonoidis becki		45750030	42187936	48685095	0.54825
SRR15734426	Chelonoidis becki		45807656	42232751	48738167	0.54953
SRR15734427	Chelonoidis becki		45779543	42190112	48686708	0.55251
SRR15734428	Chelonoidis becki	Volcan Wolf, Isabela Island	45310653	42156906	48650371	0.48568
SRR15734429	Chelonoidis darwini		45476072	42258751	48770098	0.49411
SRR15734430	Chelonoidis becki	Volcan Wolf, Isabela Island	45361778	42186844	48683846	0.48868
SRR15734431	Chelonoidis becki	Volcan Wolf, Isabela Island	45308387	42166705	48660040	0.48383
SRR15734432	Chelonoidis vicina		46078666	42183146	48678205	0.59977
SRR15734433	Chelonoidis vicina		45734485	42184753	48680561	0.54647
SRR15734434	Chelonoidis vicina		45744165	42186770	48683089	0.5476
SRR15734435	Chelonoidis hoodensis		46164084	42219868	48723708	0.60644
SRR15734436	Chelonoidis hoodensis		42705682	38931495	44928213	0.62938
SRR15734437	Chelonoidis hoodensis		46156408	42245934	48755244	0.60075
SRR15734438	Chelonoidis porteri	Santa Cruz Island	45656630	42246423	48754933	0.52396
SRR15734440	Chelonoidis darwini		45615498	42196585	48695601	0.52607
SRR15734441	Chelonoidis darwini		45487122	42183839	48680570	0.50845
SRR15734442	Chelonoidis porteri	Santa Cruz Island	45775439	42282585	48799003	0.53601
SRR15734443	Chelonoidis donfaustoi		45911859	42188152	48684263	0.57322
SRR15734444	Chelonoidis donfaustoi		45961479	42193500	48691095	0.5799
SRR15734445	Chelonoidis donfaustoi		45852448	42166466	48659968	0.56764
SRR17407396	Chelonoidis chilensis		36895793	41965183	48413952	-0.7861
SRR17408317	Chelonoidis chilensis		35607721	41507240	47882784	-0.92534
SRR17408318	Chelonoidis chilensis		35602357	41507178	47882621	-0.92618
SRR17619844	Chelonoidis phantasticus		44764115	42359166	48902804	0.36752
SRR17619845	Chelonoidis phantasticus		44260291	42304362	48837081	0.29941

Fig 2: Original Publication Mean Heterozygosity



Island	Lineage	Mean observed heterozygosity	π
Santiago	*darwini*	0.000396	0.000350
Santa Cruz	*donfaustoi*	0.000275	0.000249
Santa Cruz	*porteri*	0.000382	0.000366
Española	*hoodensis*	0.000224	0.000211
Pinzón	*duncanensis*	0.000317	0.000290
San Cristóbal	*chathamensis*	0.000390	0.000387
Isabela	*becki* –PBL lineage	0.000417	0.000371
Isabela	*becki* –PBR lineage	0.000314	0.000294
Isabela	*guntheri*	0.000289	0.000274
Isabela	*vicina*	0.000266	0.000273
Isabela	*vandenburghi*	0.000294	0.000285
Isabela	*microphyes*	0.000285	0.000253
Mean across lineages		0.000321	0.000300

Results: Principal Component Analysis

- The out-group species groups together and exhibits more variance than the ingroup lineages, as expected
- Clustering among thein-group lineages is similar to what was reported in the original publication and what has been observed in previous studies (according to authors)
  - Both *becki* lineages (SRR15734425, SRR15734426, SRR15734427 & SRR15734428, SRR15734430, SRR15734431) cluster closely together with the *darwini* lineages (SRR15734429, SRR15734440, SRR15734441), which is reflected in the original publication.
- While there is variance among lineages, it isn't a great deal of variation, according to the PC1 (~2%). This probably explains why they cluster so tightly and overlap.

Fig 1: PCA Plot, Jensen et al (2021)

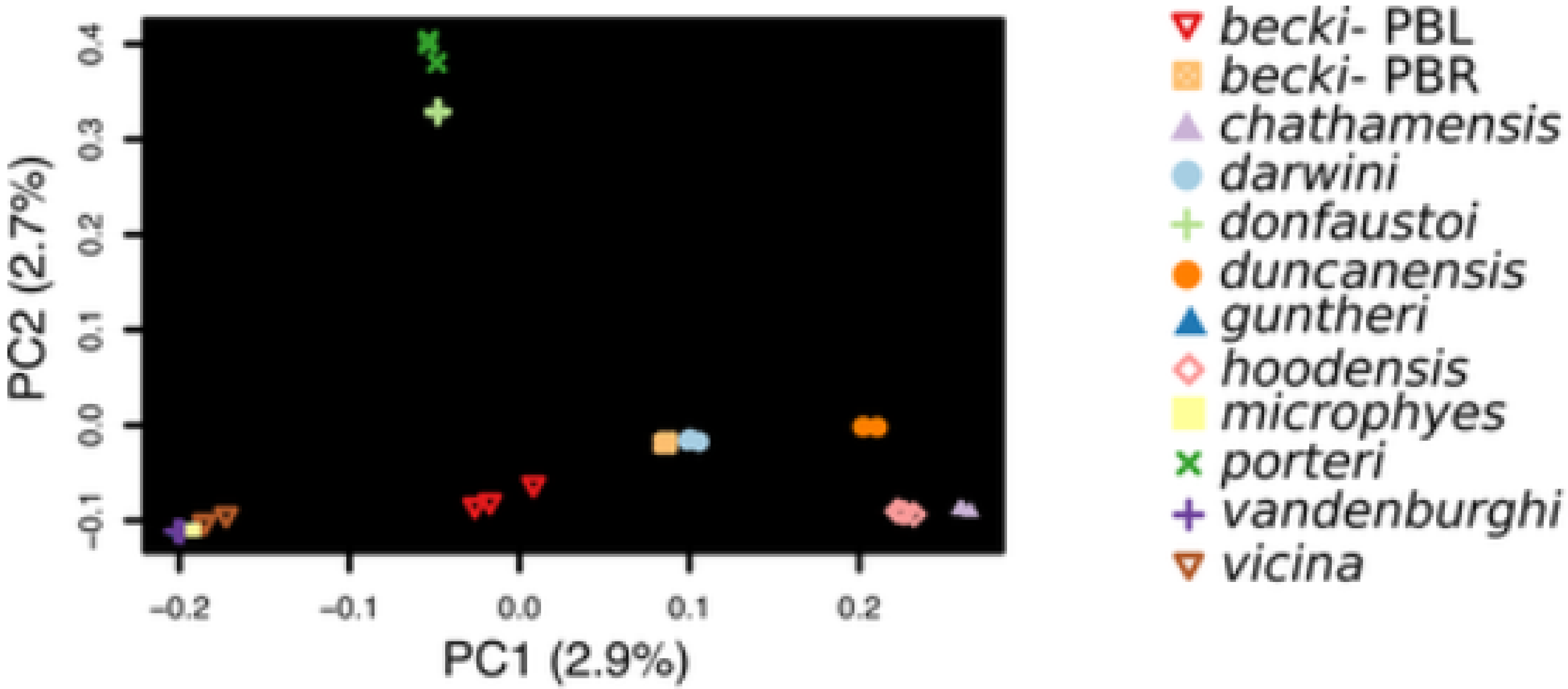


Fig 2: Reanalyis PCA Plot

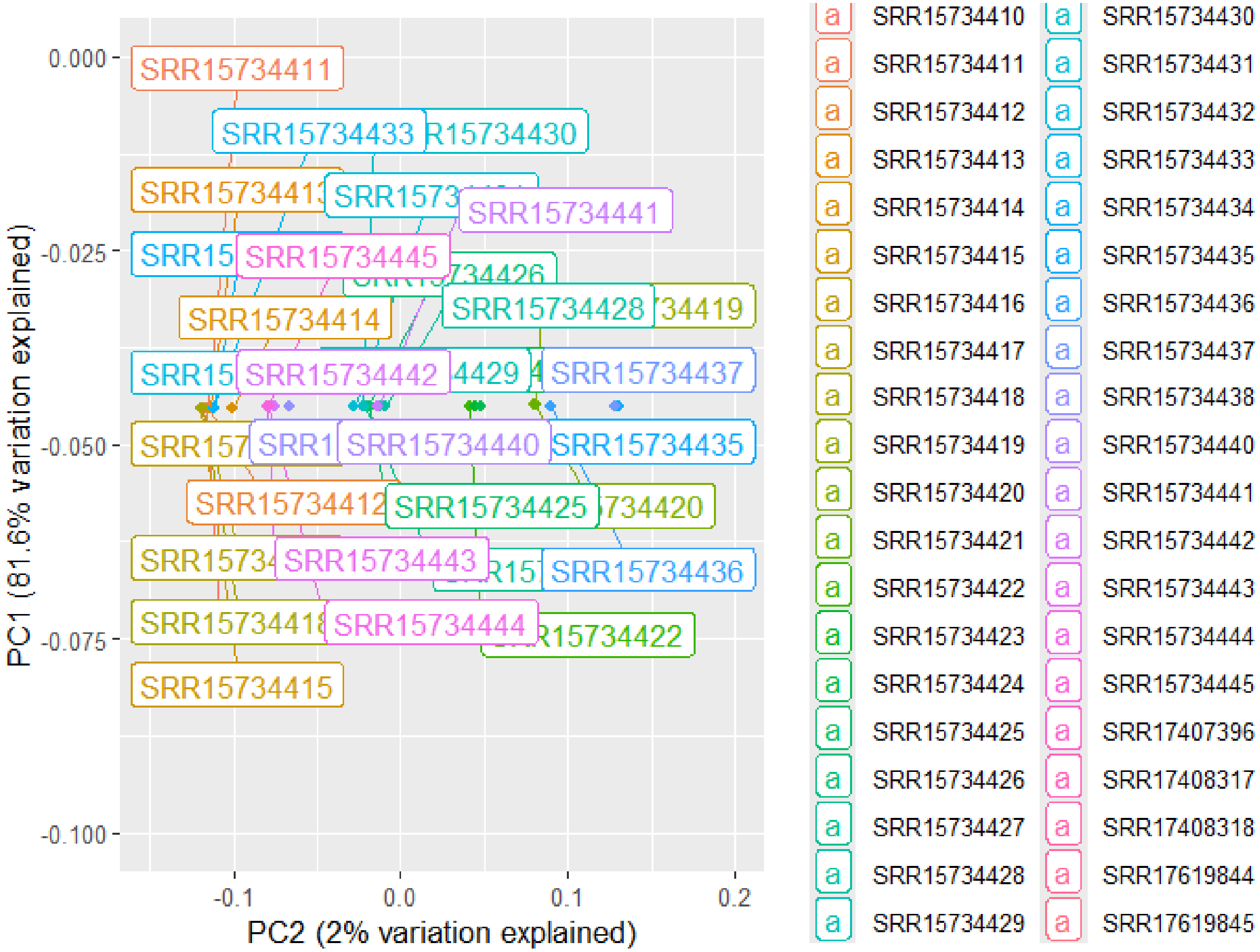
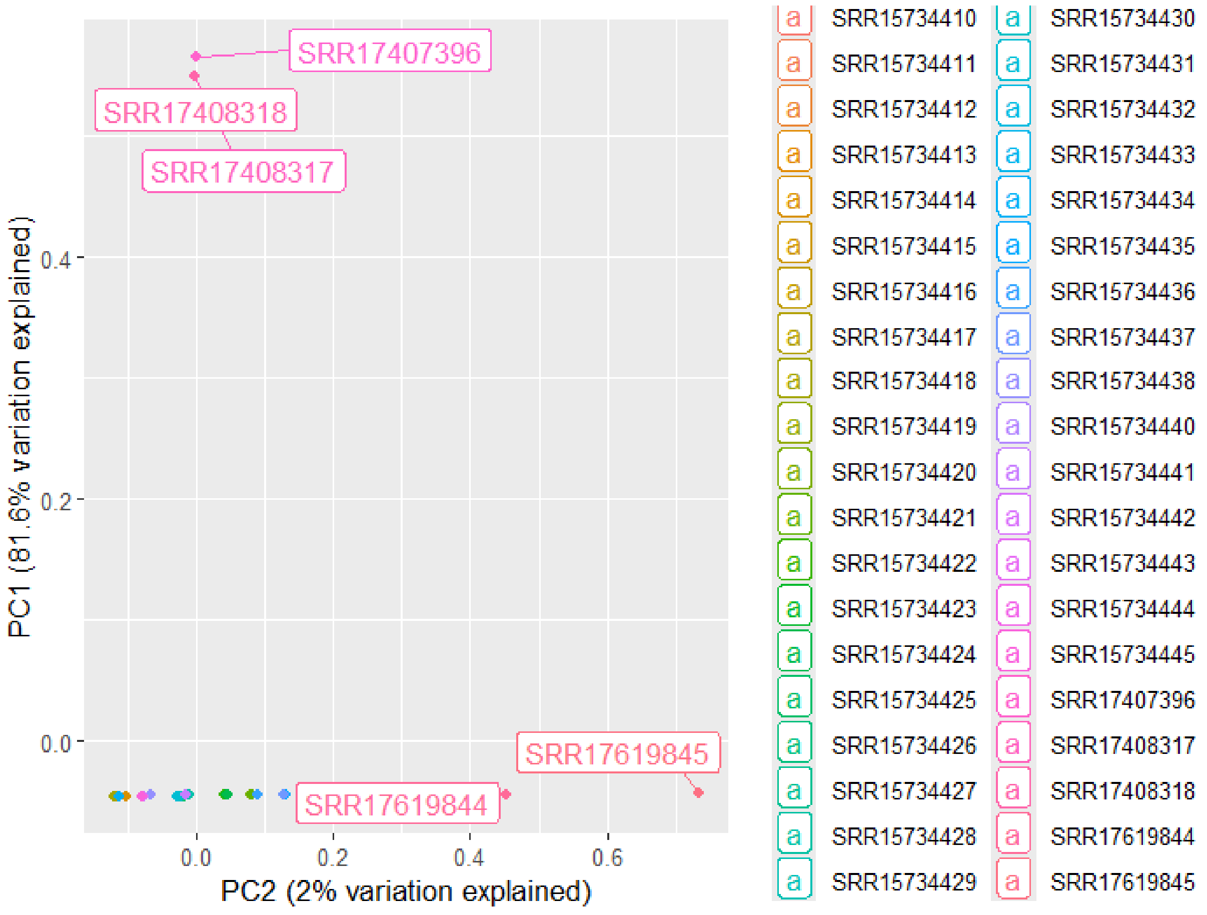


Fig 3: Full Reanalysis PCA Plot



Issues and Roadblocks to Reanalysis

- The main barrier encountered was getting bcftools to run on all the sequencing libraries in a timely fashion. Making alterations to the bcftools variant calling script consumed most of my time for this project.
- Since the reference genome was not chromosome level, a window of chromosomal windows could not be pointed to run bcftools in parallel. With the direct help of Dr. Noah Reid, a 1000 kb window file was created instead to aid in parallelization of bcftools.
- Replicating the full analysis was a very ambitious undertaking.
  - Unable to replicate mutation accumulation and coalescent rate due to unfamiliarity and time needed to learn those analysis workflows.

Issues and Roadblocks to Reanalysis

- While the original publication does give some of the important parameters used in variant calling or data analysis, some details are left out for the sake of keeping the manuscript manageable, hindering reproducibility.
- One sample (SRR15734439) for one of the lineages (*porteri)* was dropped, as fasterqdump brought back an empty sequence file. Apparently, fasterqdump can be inconsistent when pulling run data.
- I attempted to construct a phylogenetic tree based on the VCF file using VCF2PopTree in a browser. This attempt was unsuccessful. Since VCF2PopTree runs purely in the user’s browser, it is subject to the limitations of the local machine.

Citations

- Jensen, E. L., Gaughran, S. J., Garrick, R. C., Russello, M. A., & Caccone, A. (2021). Demographic history and patterns of molecular evolution from whole genome sequencing in the radiation of Galapagos giant tortoises. Molecular Ecology, 30, 6325–6339. https://doi.org/10.1111/mec.16176

GitHub Repository

- https://github.com/TSchneiderUCONN/ISG5312_final_project