Reference genome choice and filtering pipeline jointly impact phylogenetic analyses

Jessica A. Rick[1,2], Chad D. Brock[1], Catherine E. Wagner[1,3]

[1]Botany Department, University of Wyoming, Laramie WY; [2]Program in Ecology, University of Wyoming; [3]Biodiversity Institute, University of Wyoming, Laramie, WY

Scan code to bring this poster with you

BACKGROUND

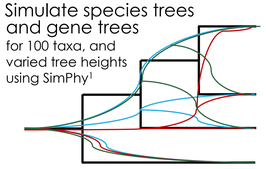
Although genomic datasets are increasingly used phylogenetics, we still have incomplete understanding of how decisions about the bioinformatic treatment of next-generation sequence data impact downstream phylogenetic results.

Analysis of genomic data requires researchers to make *a priori* decisions on **which data to retain and which to filter out**, which can introduce biases into the data used in analyses. Several recent studies have elucidated biases introduced by this data processing step, **which can substantially affect subsequent analyses** in both population genetics and phylogenetics.

Here, we investigated the effects of the average distance from the reference genome to the in-group taxa (i.e., whether the reference genome is an in- or outgroup individual) and the manner in which this interacts with bioinformatic filtering choices.

METHODS

1. Simulate species trees and gene trees
   for 100 taxa, and varied tree heights using SimPhy[1]

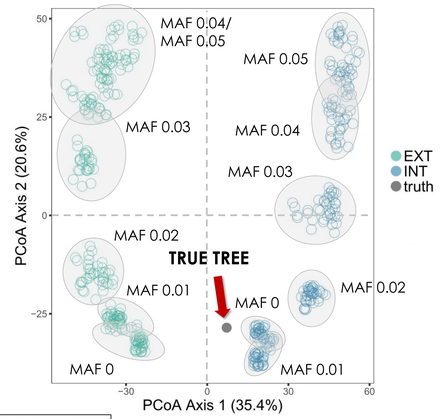


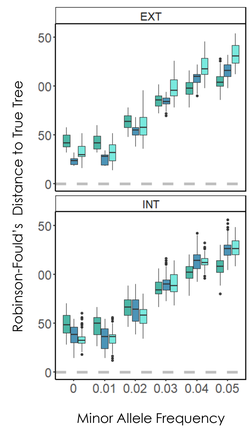
2. Simulate Illumina FASTQ reads for each taxon
   using TreeToReads[2]
3. Align reads to **outgroup** and one **random ingroup** reference, call and filter variants:
   - mapping quality
   - minor allele frequency
   - missing data per SNP
4. Build RAxML phylogeny, compare trees to simulated (known) species tree

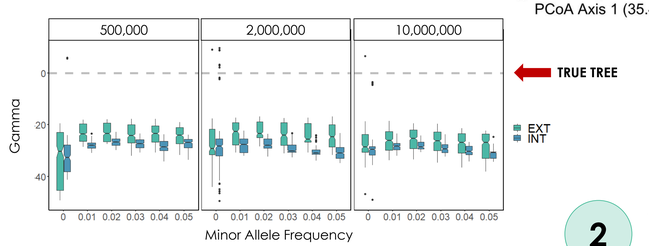
REFERENCES: [1] Mallo et al. 2015. Systematic Biology. [2] McTavish et al. 2017. BMC Bioinformatics.

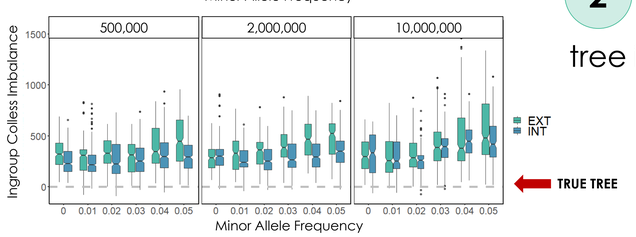
THE TAKEAWAYS

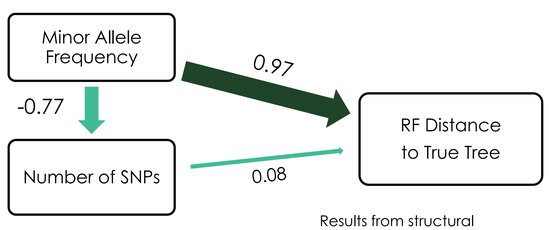
1 **Increasing** minor allele frequency cutoffs (MAF) leads to **more deviation** from the true tree, and this effect is greater in analyses using **ingroup reference genomes (INT)**.











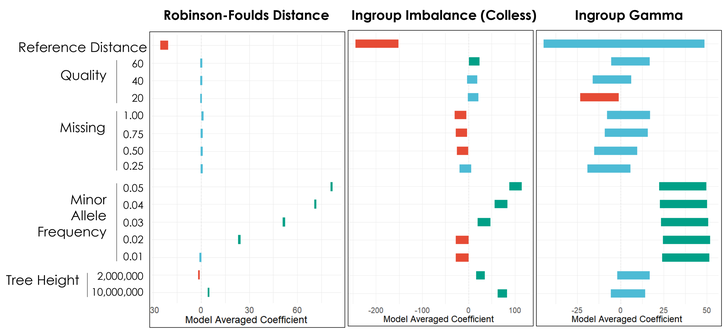
Results from structural equation model

2 This deviation from the true tree has a complex effect on tree **imbalance** and **gamma** statistics.

3 MAF remains a significant predictor of deviation from the true tree, even when **controlling for the number of SNPs retained** in the different data sets.

DETAILS AND DISCUSSION

Using linear mixed models with Robinson-Fould's distance (to the true tree), ingroup gamma, and ingroup Colless' imbalance statistic as the response variables, we found that **minor allele frequency** and **reference distance** impact all three of these measures, as shown below. In addition to these single variables, we found significant **pairwise interaction terms** between MAF, reference distance, and tree height for all three response variables. However, there remains much **unexplained variance** for the imbalance and gamma measures.



FUTURE DIRECTIONS for this project include using a **species tree method** (ASTRAL II) for inferring phylogenies to compare with the RAxML concatenation method. We also plan to **subsample the same number of SNPs** from all data sets, to further test whether minor allele frequency remains a significant consideration outside of its influence on data set size. We will also investigate how these filters alter the **mutation spectrum** of retained SNPs compared to the original set of SNPs.

ACKNOWLEDGEMENTS

Society of Systematic Biologists · ARCC · UWYO · Wyoming NASA Space Grant Consortium · Wyoming INBRE · PIE Program in Ecology