Data Matrix Assembly and Parameter Fitting

We examined a range of values for five key parameters in the STACKS 1.30 pipeline. These parameters can be seen in Supplemental Table 1. For each parameter, we tested 3 values, including the default value for values that have a default. The amount and distribution of missing data in each final phylogenetic matrix was then quantified using custom shell scripts (provided in the supplemental material) and by building distance matrices in PAUP 4b10 (cite Swofford). Distance matrices are useful in this situation, as they will highlight pairs of taxa for which there is no data upon which to make a comparison - i.e.; they have no sites in the matrix in common and there is no basis by which their relationship to each other can be evaluated. A value for a parameter was selected to be used in the final pipeline if that value both minimized the total missing data *and* the number of taxon pairs for which there is no data in common. The output of these two analyses can be found in the Supplemental Files X and Y.

The final step in the STACKS pipeline involves exporting the matrix of retained SNPs. At this step, it is possible to exclude incomplete SNPs - SNPs for which one or more individuals do not have data - from the final phylogenetic data matrix. We examined five completeness values. A small completeness value retains many highly incomplete SNPs, giving a large matrix with a large amount of missing data. A large completeness value gives a small amount of sites with little missing data.

Two individuals (Pecan Springs and Kickapoo Springs) were excluded due to low data yield. One further sample (Panther Canyon) was excluded due to specimen mislabeling.

Phylogenetic Analyses

We fit a model of DNA sequence evolution to each of the data matrices using jModelTest (cite Posada). The best-fit model that is implemented all three phylogenetic softwares that we utilized was the GTR model with Gamma-distributed rate variation. We performed phylogenetic estimation in three different software packages on each of the five final SNP matrices. The first two strategies were maximum likelihood methods. We used Garli to obtain a point estimate of topology for each data matrix. Because SNP datasets inherently exclude invariant sites, we also used RAxML (cite Stamatakis) to estimate a tree, as RAxML has recently implemented corrections for this type of ascertainment bias (cite Leache). To assess support for the conclusions, we performed 100 bootstrap replicates for each dataset, in each software package. Finally, we estimated a sample of trees in MrBayes, and summarized this sample with a majority-rule consensus tree. The MCMC chain was run for 10 million generations, and checked for convergence using Tracer (cite Drummond).

STRUCTURE

We also processed the 5 matrices for use with STRUCTURE. Because of the large number of sites in our matrices, STRUCTURE runs quite slowly. This made running many replicates to assess support for the number of populations present prohibitive. To make this more tractable, we used the software fastSTRUCTURE to do initial trials from K=2 to K=8. We

used the results of these estimations to target a smaller set of candidate values of K to use for analysis with STRUCTURE. The three most complete matrices supported K=2 in fastSTRUCTURE. The two least complete support K=4 and K=5. Therefore, in STRUCTURE, we sampled K values between 2 and 5 for four of the matrices. The least complete matrix is so large that we were only able to sample K=2 and K=3 for this matrix.

Results

Data Matrix Assembly

Our least-complete phylogenetic matrix contained 34518 sites, and the most complete contained 247 (Table XX) sites. When assembling a phylogenetic matrix, STACKS distills the dataset down to one individual in the matrix being equivalent to one population, not one organism. Therefore, some sites will be discarded will be discarded if they are not shared among many individuals in a population. Output for STRUCTURE does not perform this filtering, as each individual in the data matrix is equivalent to one individual organism. Therefore, the STRUCTURE matrices are much larger than the phylogenetic matrices (Table XX).

Phylogenetic Analysis

There is broad support for the backbone topology across all datasets and types of analysis we performed. All the trees support a deep North-South split in the group, and all show monophyly among the 14 proposed species.

Branch length and resolution on the trees is strongly correlated with the completeness parameter. As shown in Fig. XX, in trees estimated from matrices with a small completeness parameter (many sites with much missing data), generally yield a fully-resolved tree with non-zero branch lengths for each branch. Trees estimated from data with a high completeness parameter have many polytomies and tips with very short branches. This pattern is found in both the Bayesian and maximum likelihood analyses.

Garli and RAxML yield very similar trees in terms of topology. The correction for ascertainment bias appears to mostly affect branch lengths, leading to shorter branches being estimated across the tree (Supplemental Figure XX).

_____ BELOW HERE IS STUFF THAT SEEMS LIKE SUPPL. INFO. _____

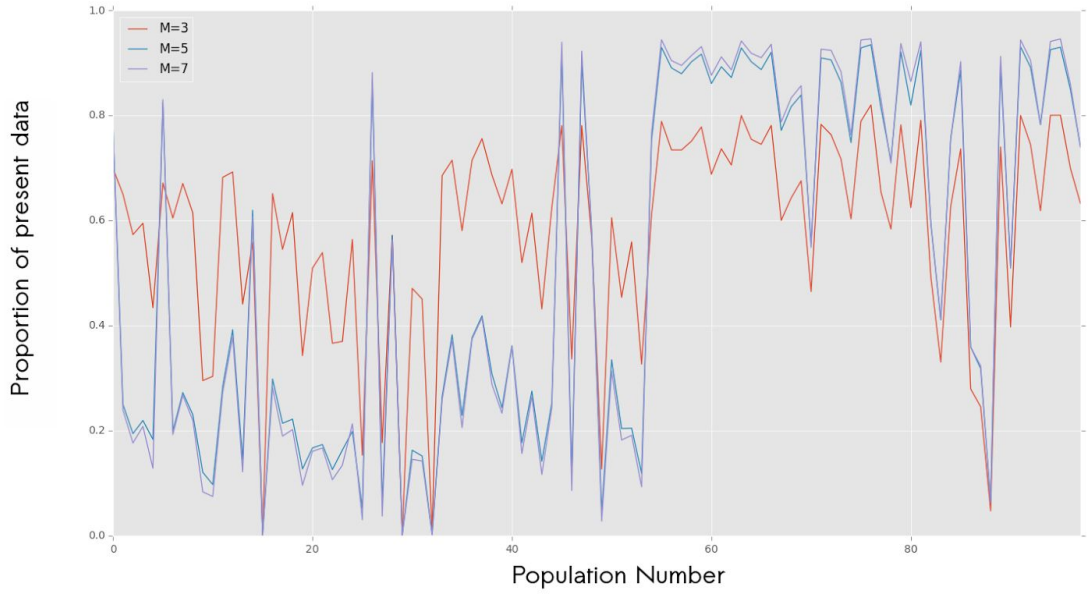
Supplemental Table 1: Choice of Analytical Parameters for STACKS 1.3 Pipeline. In the 'Values Tested' column, the default value is indicated in bold lettering.

Parameter	Pipeline Step	Parameter	Values Tested	Value Selected

		Function		
-M	ustacks	Minimum stack depth (in number of reads) to record a locus	**3**, 5, 7	3
-m	ustacks	Maximum mismatches allowed between reads in a stack	**2,** 3, 4	2
-n	cstacks	Maximum mismatches allowed between a proposed locus and a locus in the catalog	**0**, 1, 2, 3	2,3
-m	populations	Minimum stack depth required at a locus for individual to be added to population	3, 6, 9	3
-r	populations	Percent of individuals in a population that must possess a locus to include locus	20, 40, 60, 80, 90	

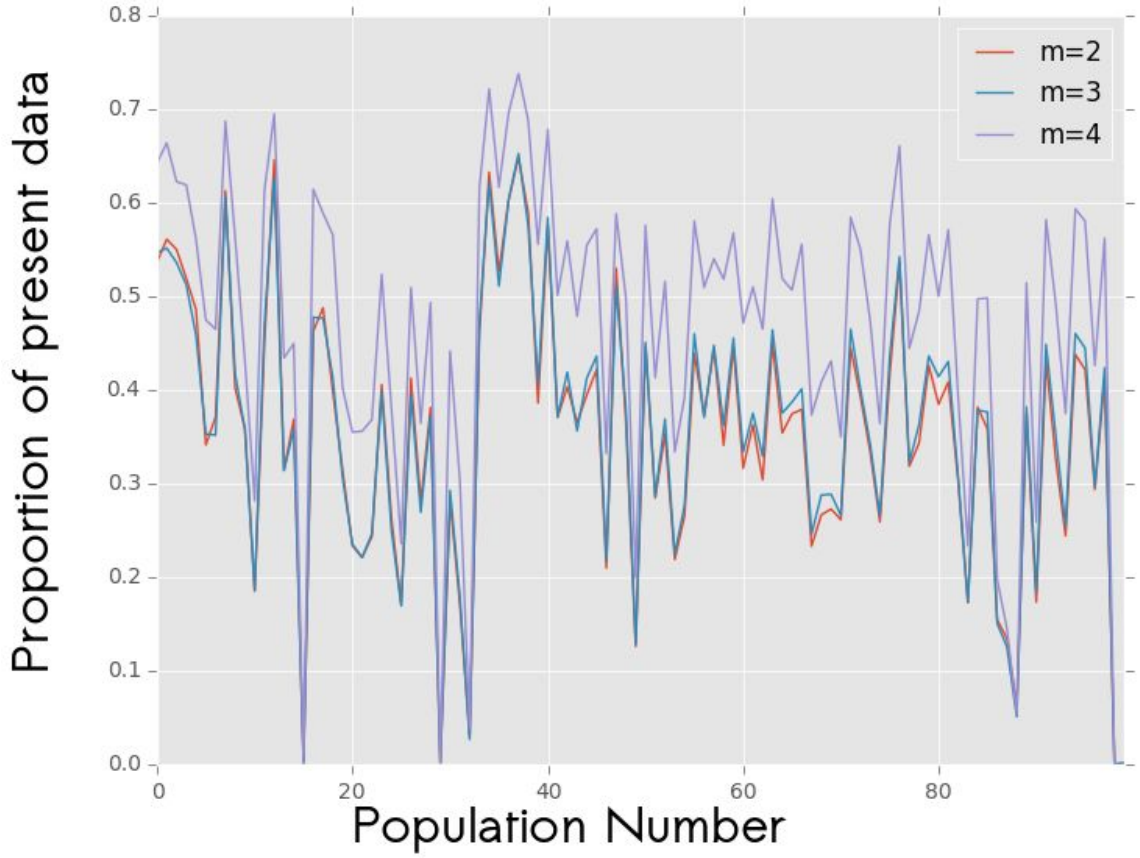
-M

We used 3 values for the minimum stack depth parameter, M=3, M=5 and M=7. A stack depth of M=3 produces a consistent amount of missing data across most individuals in the dataset. M=5 and M=7 produce missing data that are biased - populations from the second sequencing run have about twice as much missing data, in proportion to the total number of sites, as those from the first. These results are presented in Supplemental Figure XX.



-m

We tried 3 values for the maximum within-individual mismatch parameter, m=2, m=3 and m=4. M=2 is the default. m=2 maximizes the number of loci in the data matrix. m=2 and m=3 propose the same average amount of missing data per taxon (35%), which m=4 produces an average of 47%. In the distance matrix, m=2 minimizes the number of taxon pairs for which there is no data for a comparison (n=177). The proportion of data missing per taxon is visualized on Figure XX.



-n

We looked at four values of the maximum catalog mismatch parameter, n=0, n=1, n=2 and n=3. The default value of this parameter is n=0. The default value for our data severely limited the stacks depth values for the **-m** and **-r** parameters of the populations step: We did not retrieve any loci that had a stack depth of more than 2, or were present in more than 30% of populations. This suggests a value of n=0 is overly-conservative for our data. As can be seen on Figure 3, there is no clear value that always minimizes missing data. Both n=1 and n=3 maximize and minimize missing data for different individuals. n=2 tends to perform intermediately, never producing either the most or least missing data. However, n=3 produces the fewest taxon pairs for which there is no missing data. Therefore, both n=2 and n=3 were used in the fitting of the final two parameters. When fitting -m, n=3 retains more loci, but the distribution of missing data and number of taxon pairs for which there are no data in common is the same for n=2 and n=3. When the final parameter, r, is included, n=2 consistently produces more taxon pairs for which no data are available. We therefore selected n=3 to use in our final matrix.

-m (populations step)

We analysed the data using three values for the -m parameter, m=3, m=6 and m=9. The default is not specified in the manual. Much like the -n parameter, this parameter made little difference to the overall amount of missing data. However, increasing this value increases the amount of taxon pairs between which no data are shared (n=3 gives 174 taxon pairs for which there is no basis to compare, n=9 gives 354). Increasing this parameter also halves the number of loci retained in the phylogenetic matrix, from over 30,000 to 15,000. We, therefore, chose to use the fairly liberal value of m=3.

-r

The r parameter governs what percent of individuals must have a locus to use that locus in the final phylogenetic matrix. We looked at 5 values of this parameter: r=20, r=40, r=60, r=80, r=90. At r=90, we retain about 250 SNPs, twice the number to the number of tips on the tree, and we, therefore, did not sample more possible values of r. As the value of r increases, the number of taxon pairs for which there are no shared data points increases (Fig. 4). However, for taxa where data is present, the completeness of the data increases with the value of r (Fig. 5). We, therefore chose to evaluate all five matrices.

Table 2: Final matrix sizes for phylogenetic analyses and STRUCTURE analyses.

Matrix Name	Phylogenetic Matrix Size	STRUCTURE matrix size
323320	34518	75296
323340	10231	26401
323360	3416	8061
323380	1327	3661
323390	247	722