Speciation proceeds in forward and reverse in real time in the Scrub-Jay species complex

**Abstract**: The Scrub-Jays are an iconic North American avian species complex with multiple recently diverged lineages experiencing variable levels of isolation and gene flow, making difficult both species delimitation and accurate reconstruction of the group’s evolutionary history. To tackle these questions, we performed restriction-enzyme associated DNA sequencing of thousands of genomic loci shared throughout the entire distribution of Scrub-Jay species complex. We used unlinked SNPs from these loci as input for multiple unsupervised machine learning and clustering approaches in search of a consensus number of independent evolutionary lineages within the Scrub-Jays. We then used multiple species tree reconstruction approaches to infer the evolutionary history of the identified lineages as accurately as possible, while highlighting tree-regions where branching patterns have been muddied by rapid and reticulate evolution. Additionally, we tested explicitly for gene flow between all identified lineages and searched the genome for signatures of introgressed haplotype blocks between species. Species delimitation combined with detailed reconstructions of evolutionary relationships and tests for gene flow lend support for recognition of an additional Scrub-Jay species, and highlight that the process of speciation appears to be proceeding both forwards and backwards in real time within the Scrub-Jay species complex.

**Introduction**

Species concepts have been contentious ground within evolutionary biology for decades (cite), but with the onset of the genomic revolution, our ability to determine the evolutionary history of previously difficult to resolve groups of organisms has never been greater (cite).

The taxonomy of the Scrub-Jay species complex has fluctuated historically, but currently the AOC recognizes four species, specifically California, Island, Woodhouse’s, and Florida Scrub-Jays.

Here, we follow the growing trend of sequencing thousands of genomic loci shared throughout the distribution of an entire species complex, in order to determine whether the current taxonomy accurately reflects the number of independently evolving metapopulations (De Queiroz, 2007) present within the species complex (Chan et al., 2017; Erickson et al., 2021; Georges et al., 2018; Hosegood et al., 2020; B. T. Martin et al., 2020; Newton et al., 2020; Tsai et al., 2019; Venkatraman et al., 2019). We first used multiple analytical methods for inferring species delimitation schemes from genomic data, in search of a biologically relevant consensus across methods (Carstens et al., 2013). We then used species tree reconstruction approaches to resolve the evolutionary history of the putatively delimited species, specifically using gene tree quartet frequencies to determine our level of confidence in resolving difficult nodes within the Scrub-Jay phylogeny. We also test for gene flow between putatively identified species, and search for signatures of introgressed haplotype blocks between species experiencing ongoing gene flow. These analyses facilitate a detailed understanding of the current diversity present within the Scrub-Jay species complex, and highlight the continuous and unending nature of the diversification process under natural conditions.

**Methods**

**DNA extraction, library preparation, and sequencing**

We extracted genomic DNA from 115 ethanol preserved tissue samples using a manual magnetic bead-based protocol based on Rohland & Reich (2012)

(full protocol available at: [github.com/DevonDeRaad/aph.rad/blob/master/dna.extraction.protocol.md](http://github.com/DevonDeRaad/aph.rad/blob/master/dna.extraction.protocol.md)). Library prep was performed on genomic DNA extracts at the University of Kansas Genomic Sequencing Core (KUGSC), following the protocol outlined in Manthey and Moyle (2015) (detailed protocol available at: [github.com/DevonDeRaad/aph.rad/blob/master/MSG.libraryprep.protocol.md](http://github.com/DevonDeRaad/aph.rad/blob/master/MSG.libraryprep.protocol.md)). Briefly, samples were digested with the enzyme NdeI, ligated with custom barcodes, size selected for fragments between 500-600 base pairs in length, PCR amplified, and bead purified. Libraries were then pooled across multiple projects and sequenced on a single flow cell using a high-output kit on a NextSeq550 Illumina sequencing machine to generate single-end 100bp raw reads.

**Reference genome synteny mapping and SNP calling**

For mapping and assembling our RAD loci, we utilized the publicly available Florida Scrub-Jay reference genome assembly, which was generated via whole genome shotgun sequencing on an Illumina HiSeq machine, with a mean sequencing depth of 221x. Specifically, we downloaded a highly contiguous (1,701 total scaffolds, scaffold N50 = 7.63 Mb) assembly created with ALLPATHS-LG v. 49184 (Gnerre et al., 2011) (available at: https://www.ncbi.nlm.nih.gov/assembly/GCA\_013398375.1#/st). A recent review of synteny mapping suggests a minimum scaffold N50 of 1Mb in order to make robust downstream analyses from synteny mapped genomes (Liu et al., 2018). Because of the strong conservation of genome architecture among birds, and the highly contiguous nature of our input reference genome, we were able to use Satsuma (Grabherr et al., 2010) to map this Florida Scrub-Jay assembly to a recent release of the chromosome assembled Zebra Finch reference genome (version: TaeGut3.2.4, available from: <http://ftp.ensembl.org/pub/release-97/fasta/taeniopygia_guttata/dna/>), giving us chromosomally assigned scaffolds onto which we aligned our RAD loci.

We demultiplexed our raw sequence files using the ‘process.radtags’ function from Stacks v2.41 (Rochette et al., 2019) in order to discard low-quality reads and reads with ambiguous barcodes. We then used BWA v0.7.17, to map the demultiplexed reads from each sample onto our reference genome using the ‘mem’ command with default settings (Li & Durbin, 2009), and used SAMtools v1.3.1 (Li et al., 2009) to generate sorted .bam files for downstream use. We used this sorted .bam file, containing aligned raw sequence reads, as input for the ‘gstacks’ module from Stacks, in order to build a catalog of individual RAD loci (Rochette et al., 2019). Finally, we used the ‘populations’ module from Stacks to output a completely unfiltered variant call format (vcf) file, which contained 210,336 SNPs each shared by multiple samples (Rochette et al., 2019).

**Quality filtering**

We then used the R (R Core Team, 2019) package SNPfiltR (cite) to filter the genotypes and samples within this unfiltered vcf file. Using functions from the package, we imposed a hard filter on each genotype requiring a minimum depth of 5, and a minimum genotype quality of 30. We then imposed an allele balance filter requiring heterozygous genotypes to have an allele balance >.25 and <.75. Additionally, we imposed a maximum mean depth filter of 100 reads per genotype, in order to remove likely paralogous loci. We visualized missing data per sample and chose to remove 20 samples which were missing data at > 91% of called SNPs. With our final set of samples for downstream analyses, we visualized an array of missing data cutoffs and visually determined that retaining SNPs with more than 85% non-missing genotypes maximized the number of SNPs retained while minimizing the overall proportion of missing data. Finally, we removed invariant sites generated during the filtering process, retaining all SNPs with a minimum minor allele count of 1 for downstream analyses. This filtering scheme resulted in a filtered SNP dataset containing 95 samples and 16,307 SNPs, with 5.7% overall missing data (referenced as ‘filtered SNP dataset’ going forward). We then filtered this dataset further, removing SNPs containing missing any genotypes, for input to downstream analyses that are intolerant of missing data (specifically our machine learning species delimitation approaches). This filtering scheme resulted in a dataset containing 95 samples, 1,779 SNPs, and 0% missing data (referenced as ‘complete SNP dataset’ going forward). Finally, we used VCFtools v0.1.15 (Danecek et al., 2011) to thin our filtered SNP dataset, retaining only SNPs separated by at least 1,000 base-pairs in our reference genome assembly. The subsequent dataset contains 95 samples and 2,725 SNPs, with 6.3% overall missing data (referenced as ‘filtered, unlinked SNP dataset’ going forward). R code detailing this filtering process with accompanying data visualizations, can be accessed at: devonderaad.github.io/aph.rad/filter.ref.aligned.radstackshelpr.html

**Species Delimitation**

We followed the clustering-based species delimitation protocol outlined by Derkarabetian et al. (2019), which involves both traditional clustering approaches and novel applications of unsupervised machine learning algorithms, in search of a biologically relevant consensus between methods. We began by executing a traditional population clustering approach using the R package ‘adegenet’ (Jombart, 2008) to execute discriminant analysis of principal components (DAPC) (Jombart et al., 2010). We began by performing successive K-means clustering to assign our 95 samples to populations and compare the goodness of fit of each clustering scheme based on Bayesian Information Criterion (BIC). Because the two most optimal clustering schemes (K=5 and K=6) were separated by less than a value of < 2 in BIC, these two schemes effectively cannot be differentiated, and both are presented here. Because DAPC is free of assumptions about linkage between SNPs, we used both our filtered SNP dataset (16,307 SNPs) and our complete SNP dataset (1,779 SNPs) as input for DAPC and found no discernible difference in clustering patterns (K=5/6 optimal in both cases). Here we present the results from our complete SNP dataset, which was also used as input for clustering analyses using unsupervised machine learning approaches. Following best practices outlined by Jombart and Collins (2015), for each scheme we retained enough PC axes to explain 75% of the overall variation in the data, and retained all possible discriminant factors.

We then performed dimensionality reduction using three distinct unsupervised machine learning algorithms, followed by two distinct clustering algorithms to determine the optimal number of species clusters present in the reduced-dimensionality space. Because some machine-learning algorithms (specifically random forest) cannot tolerate missing data, we used our complete SNP dataset (1,779 SNPs, 0% missing data) as input for all three unsupervised machine-learning approaches.

We began by using SNP genotypes as input for random forest (RF), implemented via the ‘randomForest’ R package (Liaw & Wiener, 2002). We ran an unsupervised RF with 5,000 individual decision trees, and the default proportion of variables sampled at each split (here sqrt(p) where p = the number of SNPs used as input variables), in order to generate a proximity matrix amongst our 95 samples. We ran repeated replicates varying the starting seed, and eventually settled on a seed value that reliably resulted in a representative spatial clustering pattern. We then used the function MDSplot() from the ‘randomForest’ R package (Liaw & Wiener, 2002) to perform classical multi-dimensional scaling of the proximity matrix output by random forest, with the statistically significant number of dimensions retained, as determined using the Broken Stick method implemented via the R package ‘PCDimension’ (Wang et al., 2018).

Next, we performed principal components analysis (PCA) via the R package ‘ade4’ (Dray et al., 2020) using our set of 1,779 complete SNP genotypes as input. We then used the first 10 principal component axes as input for t-distributed stochastic neighbor embedding (t-SNE) implemented via the R package ‘tsne’ (Donaldson, 2016), using 5,000 maximum iterations, a perplexity of five, five initial dimensions, and two dimensions for the resulting embedding (Derkarabetian et al., 2019). We repeated this procedure with multiple starting seeds before settling on a seed which reliably reproduced a commonly shared clustering pattern.

Additionally, we fit a variational auto-encoder (VAE) to our complete set of SNP genotypes using the python implementation ‘popVAE’ (Battey et al., 2021). We set the number epochs to run after the last improvement in validation loss to 500 and left the training proportion and network search sizes at their default values. Because popVAE generates stochastic results even with the specification of a constant starting seed, we performed 10 separate runs with identical parameters and starting seed and compared the resulting number of clusters delimited across these 10 runs. We subsequently chose to present one of the iterations that resulted in the most common clustering scheme shared across all 10 iterations.

Finally, we used two distinct analytical methods for determining the optimal number of clusters to partition the samples into in each of the resulting reduced dimensionality spaces (3 unsupervised machine-learning approaches x 2 clustering approaches = 6 distinct species delimitation schemes). We used partitioning around medoids (PAM) implemented via the R package ‘cluster’ (Maechler et al., 2018) to determine the optimal number of clusters (K) in this reduced dimensionality space by comparing the silhouette width across values of K=2-10, and choosing the highest value (Derkarabetian et al., 2019). Additionally, we used hierarchical agglomerative clustering (HAC) implemented via the R package ‘mclust’ (Scrucca et al., 2016) to determine the optimal number of clusters and assign each sample into a cluster. The entire workflow for executing this clustering-based species delimitation approach is depicted graphically in Figure 1, along with the optimal number of clusters identified by each approach. In addition, all species delimitation analyses described herein can be viewed and reproduced from the following link: https://devonderaad.github.io/aph.rad/ml.species.delimitation.html

**Reconstructing evolutionary relationships between samples**

We began by using another traditional clustering approach, STRUCTURE v2.3.4 (Pritchard et al., 2000), which uses genotype data to assign samples to population clusters, and can aid in identifying both population structure and admixed individuals. Because structure has been shown to be strongly affected by minor allele frequency (Linck & Battey, 2019), we used our filtered, unlinked set of 2,725 SNPs as input, and removed singletons, resulting in 1,828 total SNPs. We performed five replicate runs with unique starting seeds for each value of K (K = number of clusters) 1-8, each with a burn-in of 50,000 followed by 500,000 MCMC replicates. We used the R package ‘pophelper’ (Francis, 2017) to visually confirm consistency across replicate runs, indicating that 500,000 replicates was sufficient to achieve convergence. We also used ‘pophelper’ to determine the optimal K value based on the value of ΔK, as implemented by Evanno et al. (2005) (cite suppmat figure). Finally, we repeated this same procedure with identical parameters for a reduced dataset containing only the sampling localities 12-23, and 809 unlinked SNPs with singletons removed, in order to identify populations that may have gone unrecognized due to the hierarchical nature of the structure within the Scrub-Jay species complex. We visualized both of these sets of structure assignments as barplots using the ‘pophelper’ package, and annotated relevant species delimitation schemes onto each structure run.

Because concatenated phylogenetic reconstruction with RAD data has been shown to be most effective when using entire loci including invariant sites (Leaché et al., 2015), we generated a ‘whitelist’ of the 2,725 loci retained in our filtered, unlinked SNP dataset, and used the Stacks populations module to output the entire sequence for each of these loci in phylip format (Rochette et al., 2019). We used these concatenated loci as input for Raxml v8.2.11 (Stamatakis, 2014), and inferred a phylogenetic tree using the GTRGAMMA model of nucleotide substitution, and 1000 rapid bootstrap replicates to assess branch support. We used Figtree v1.4.4 (Rambaut, 2014) to visualize bootstrap support for internal branches, and mapped each putative species delimitation scheme onto the tips of the resulting tree.

Additionally, we sought to reconstruct relationships as a phylogenetic network, which is a graphical representation of relatedness that can clarify evolutionary reconstruction especially in cases of complex and reticulate evolutionary history (Huson & Bryant, 2006). For our 2,725 unlinked, filtered SNPs, we calculated Nei’s genetic distance (Nei’s D) (Nei, 1972) pairwise between all samples using the R package ‘StAMPP’ (Pembleton et al., 2013). We then used this pairwise divergence matrix between all samples as input for SplitsTree4 v4.15.1 (Huson & Bryant, 2006), to generate a phylogenetic network, specifically a Neighbor-Net (Bryant & Moulton, 2004), which can effectively present conflicting underlying data patterns. We subsequently annotated each sample tip with a color corresponding to its assignment under a species delimitation scheme of K=7, and the number corresponding to its sampling locality.

**Reconstructing species trees**

Because concatenation-based methods are known to perform poorly under conditions of high gene tree species tree discordance, we also sought to resolve evolutionary relationships using species tree methods which can account for sources of gene tree species tree discordance such as incomplete lineage sorting (ILS) (Kubatko & Degnan, 2007). One such method, SVDquartets (Chifman & Kubatko, 2014), infers bipartitions by generating singular value decomposition (SVD) scores for unrooted quartet topologies, and uses these inferred bipartitions to build a species tree under the multi-species coalescent model. We converted our phylip file containing 264,511 sites from our set 2,725 unlinked loci, into nexus format, and used the resulting nexus file as input for SVDquartets implemented via PAUP\* (Wilgenbusch & Swofford, 2003). We designated seven discrete species tips based on the clustering assignment from our K=7 species delimitation scheme, sampled 100,000 random quartets, performed 100 bootstrap replicates, and visualized the resulting tree with bootstrap support values using Figtree v1.4.4 (Rambaut, 2014). We performed multiple replicates and visually confirmed low variability between replicate reconstructions, indicating that 100,000 random quartets is an effective sample of all possible unrooted quartets for the given tree.

As an alternative approach to building a species tree, we used ASTRAL-III v5.7.7 (Zhang et al., 2017), which is a coalescent-based summary species tree approach requiring individual gene trees as input, rather than sequence data. We used the concatenated phylip file containing 2,725 unlinked loci output by Stacks (Rochette et al., 2019) to generate 2,725 separate locus alignments in fasta format using the R package APE v5.4.1 (Paradis & Schliep, 2019), and converted these fasta locus alignments to nexus format using a publicly available python script (<https://github.com/mmatschiner/tutorials/blob/master/ml_species_tree_inference/src/convert.py>). We then generated a separate gene tree for each nexus file using IQ-TREE v1.6.3 (Nguyen et al., 2015), and performed 1,000 bootstrap replicates for each gene tree. Although our short (100bp) RAD loci contain limited information for reconstructing species relationships, ASTRAL is capable of handling polytomies and missing data in input gene trees, and short loci free from intralocus recombination are ideal for estimating gene trees. Nevertheless, these loci are susceptible to generating poorly resolved gene trees, which may bias coalescent species tree reconstruction (Xi et al., 2015; Zhang et al., 2017). Therefore, we followed author recommendations (Zhang et al., 2017) and used Newick utilities (Junier & Zdobnov, 2010) to contract branches in the gene trees with less than 10% branch support, which subsequently increased the normalized quartet score of our overall species tree reconstruction. We used these resulting gene trees as input for ASTRAL and generated a species tree along with quartet frequencies for each of the internal branches by specifying the ‘-t 16’ option. For internal branches with posterior probability < .9, we annotated the branch with the observed frequencies of all three possible unrooted quartet topologies for the given branch. Finally, we used the ‘-t 10’ option in ASTRAL to perform polytomy tests (Sayyari & Mirarab, 2018). For a given internal branch in a species tree, a polytomy test tests the statistical power to reject the null hypothesis that the given internal branch has a true length of zero, meaning that the three species tips separated by this internal branch all branched from each other instantaneously (i.e. a hard polytomy) (Sayyari & Mirarab, 2018). We annotated each internal branch in our species tree reconstruction with the posterior probability calculated by ASTRAL, and the p-value for a polytomy test focused on the given branch.

We then used SNAPP to infer a species tree and visualize potential uncertainty in species tree reconstruction (Bouckaert et al., 2014). Because of computational constraints, we subsampled our filtered, unlinked SNP dataset to randomly include only 2 samples for each of the seven putative delimited species. To alleviate potential stochasticity associated with sample inclusion, we repeated this procedure five separate times. For each of these five randomly subsampled datasets, we implemented SNAPP via BEAST2 v2.6.4 (Bouckaert et al., 2014) running an MCMC chain for five million generations, with a burn-in of 500,000 generations, saving the output every 1,000 steps (4,500 total trees sampled from the posterior distribution for each run). For each replicate dataset, we used Tracer to visually verify chain convergence, and confirmed effective sample sizes (ess) > 200 for all estimated parameters. Four of the five replicates achieved convergence and ess > 200 for all estimated parameters after five million generations, so we used LogCombiner to combine the 18,000 sampled trees from these four replicate datasets, and used DensiTree to overlay all resulting species trees on a single plot in order to visualize uncertainty in the reconstruction of this species relationships (Bouckaert et al., 2014). Tracer, Logcombiner, and Densitree are all distributed as part of BEAST2 (Bouckaert et al., 2014).

To further investigate the source of uncertainty generating the conflicting species trees apparent in our SNAPP reconstruction, we used our filtered, unlinked SNP dataset as input for TreeMix (Pickrell & Pritchard, 2012), which uses allele frequency covariance to estimate a species tree in a maximum-likelihood framework, and then progressively adds ‘migration edges’ to explain variation in the data that cannot be fit to a tree-like model of evolution. We began by inferring a species tree between the seven putative species groups, and then added migration edges to the phylogeny until 99.8% of variance in allele frequencies could be explained by the tree (cite supplemental figure showing the curve). We then generated 100 bootstrap replicates of this species tree with the optimal amount of migration edges, resampling blocks of 100 SNPs, and used DendroPy (Sukumaran & Holder, 2010) to summarize branch support values on the species tree.

**Identifying gene flow**

We then sought to test explicitly for gene flow between the seven putative species identified under a K=7 species delimitation scheme. In order to explicitly test for introgression within a phylogenetic framework, we used program Dsuite (Malinsky et al., 2021). Dsuite uses allele frequency data from an input vcf file in order to calculate Patterson’s *D* (also known as the ABBA-BABA statistic) (Durand et al., 2011), and the *f*4-ratio statistic for all possible population trios, with a specified population serving as the outgroup for all comparisons. In addition, Dsuite can calculate the *f*-branch statistic, introduced in Malinsky et al. (2018), which explicitly tests for excess allele sharing (indicating gene flow) between all branches, including internal branches, on a specified guide tree which can be tested under an ((A,B)C,D) topology. We calculated the *f*-branch statistic for all possible comparisons using the species tree topology reconstructed from gene trees by ASTRAL. We also used Dsuite to perform a block-jackknife procedure to calculate Z-scores associated with each *f*-branch statistic, and converted these Z-scores into false-discovery rate (FDR) (Benjamini & Hochberg, 1995) adjusted p-values based on our 26 individual *f*-branch tests.

Additionally, we calculated and visualized the *D* Frequency Spectrum (*D*FS) using the pipeline developed by Martin & Amos (2021). This approach uses an outgroup to polarize SNPs, distinguishing between derived and ancestral alleles, before binning SNPs based on the frequency of the derived allele. Subsequently, Patterson’s *D* (Durand et al., 2011) is calculated separately for each allele frequency bin, under the specified topology of ((P1,P2),P3,OG), where a negative D value indicates excess allele sharing between P1 and P3 and a positive D value indicates excess allele sharing between P2 and P3. We performed this procedure for multiple variations of the ((P1,P2),P3,OG) topology where P2 and P3 were designated as the sampling localities 1-10 and 12-18, which were the two tips with the greatest genome-wide D value based on our previous comparisons implemented via Dsuite (Malinsky et al., 2021). *D*FS is a descriptive measurement, designed to generate intuitive insights into the demographic circumstances surrounding introgression, based on the general assumption that more recent introgression will result in D values skewed toward lower frequency (more recent) alleles. We used the accompanying *D*FS R shiny app to simulate *D*FS’s under a handful of simple conditions in order to build an intuitive understanding of the effects of demographic scenarios on the *D*FS, and present two plausible simulated *D*FS’s that generated similar distributions to those observed in our empirical data, in the supplemental material (S. H. Martin & Amos, 2021).

Then we searched explicitly for introgressed haplotype blocks between taxa with statistically significant signatures of gene flow, using the R package ‘introgress’ (Gompert & Buerkle, 2010). We began by performing a principal components analysis (PCA) using the R package ‘adegenet’ (Jombart, 2008), using our complete SNP dataset (1,779 SNPs with no missing data) to avoid the effect of missing data on clustering patterns, for samples from sampling localities 1-10, and 12-23 (i.e. all samples currently considered California and Woodhouse’s Scrub-Jays, excluding populations 24 and 25, which form a distinct population cluster in all previous analyses). We designated the three samples with the most extreme values for PC1 in each species as “parental” (i.e. least likely to contain introgressed loci) samples from which we could identify diagnostic differences between the two species, which we identified as SNPs with an allele frequency difference of > .8 between these parental populations. We then used this set of 128 diagnostic SNPs to calculate hybrid index and interspecific heterozygosity in each of the samples, and plot these values on a triangle plot using the ‘introgress’ package (Gompert & Buerkle, 2012). We also used ‘introgress’ to visualize the individual genotypes of each sample for all diagnostic SNPs, and visually searched the genome for runs of heterozygosity, which may indicate introgressed haplotype blocks between the two species. Additionally, we calculated FST between the two species at all 16,307 bi-allelic SNPs from our filtered dataset, using VCFtools v0.1.15 (Danecek et al., 2011). Because of missing data and uneven sample sizes between the two species, we generated some negative FST values which were manually converted to zero. We then used R (R Core Team, 2019) to identify 13 SNPs from our filtered SNP dataset which show a fixed ABBA pattern under a ((11,1-10)12-18,26) topology. These SNPs, fixed identical in the non-sister species California and Woodhouse’s Scrub-Jays, require the invocation of either ILS or introgression to explain their allele frequency patterns. Genome-wide patterns of FST, as well as SNPs showing this ABBA pattern, were visualized using the R package ‘ggplot2’ (Wickham et al., 2020), and all code for this investigation of introgression at the individual SNP level is available at: \_\_\_.

**Quantifying genetic diversity and divergence**

Finally, we calculated observed heterozygosity within each sample, and nucleotide diversity (Pi) within each of the seven putative species using the Stacks population module (Rochette et al., 2019). Both Pi and heterozygosity were estimated from all sites (including invariant sites) successfully mapped in the .bam file for each sample, before applying any SNP/locus filtering protocols. This prevents biasing our estimates by filtering for specific types of loci and SNPs and generates estimates that are robustly comparable across species and sequencing approaches. We then calculated pairwise FST between each of the seven putative species using the R package ‘StAMPP’ (Pembleton et al., 2013), and used R (R Core Team, 2019) to calculate the number of fixed differences for each pairwise species comparison. All of these species comparisons were visualized using the R package ‘ggplot2’ (Wickham et al., 2020), and analyses and accompanying visualizations can be viewed and reproduced at: githublink.

**Results**

Our species delimitation pipeline resulted in eight separate species delimitation iterations, which can be described by three unique schemes differentiated by the number of species supported, K=5, K=6, and K=7 (Fig. 1). Among these eight iterations, two support a species delimitation scheme of K=5, four support K=6, and two support K=7 (Fig. 2A-I). Clustering assignment from STRUCTURE identifies congruent sample assignments with a species delimitation scheme of K=5, with a single sample (FMNH 342048) identified as visibly admixed (Fig. 2J). Maximum-likelihood reconstruction using concatenated whole loci generates a well-supported tree with reciprocally monophyletic groups at species delimitation schemes of K=4 and K=5 (Fig. 2K). In contrast, the K=6 and K=7 species delimitation schemes generate paraphyletic groupings according to this reconstruction. Meanwhile neighbor net reconstruction identifies visually distinct sample clusters for each of the putative species delimited at K=7, with the longest branch leading to the first branching species in the complex, the Florida Scrub-Jay (Fig. 3A).

Species tree reconstruction using each sampling locality as a distinct tip recovers reciprocally monophyletic groupings for species delimitation schemes of K=4, 5, and 7, while the K=6 scheme is rendered paraphyletic with low branch support (Fig. 3B). Species tree reconstruction using gene trees as input, with the seven putatively identified species designated as separate tips, renders the K=6 species delimitation scheme paraphyletic with low posterior probability for the internal branch determining the branching order between populations 12:18, 19, and 20:24 (Fig 4A). Polytomy tests for the internal branches of this tree reveal that polytomies cannot be rejected



**Figure 1**. Diagram detailing the species delimitation approaches taken, resulting in 8 unique putative species delimitation schemes. Boxes depicting a dimensionality-reduction algorithm are highlighted in gray. Boxes depicting a clustering algorithm are highlighted in gold.

Chart, calendar

Description automatically generated

**Figure 2**. (A) Sampling scheme shown geographically with numbered localities, size corresponds to sample size, shape corresponds to current taxonomy, and color corresponds to the K=7 species delimitation scheme. Two dimensional plots visualizing sample relatedness following dimensionality reduction on complete SNP dataset and species delimitation clustering for (B) DAPC with K=5, (C) DAPC with K=6, (D) Random forest and PAM, (E) Random forest and HAC, (F) t-SNE and PAM, (G) t-SNE and HAC, (H) VAE and PAM, (I) VAE and HAC. For each scheme we show the two dimensions where population clusters are most visually differentiated. Samples are colored according to assigned cluster, and the color of each cluster matches its geography on the sampling map. Approaches which identified the most common delimitation scheme (K=6) are denoted with an asterisk. (J) Maximum-likelihood tree generated from 2,725 unlinked, concatenated loci input to Raxml, with each tip labeled according to its locality as shown on sampling map, and bootstrap support shown via node size. (K) Species tree generated from 2,725 unlinked SNPs input to SVDquartets, with each of the 26 sampling localities designated as a unique taxon, and nodes with less than 60% bootstrap support collapsed for visualization purposes. For both phylogenetic reconstructions, the current taxonomy (K=4) and the three identified genomic species delimitation schemes (K=5-7) are shown with colors corresponding to the sampling map. If a species delimitation scheme identified a single cluster containing more than one of the color-coded localities on the map, the color containing the most individual samples is used to depict the corresponding cluster on the phylogeny.

CHANGE raxml tree to branch support, add structure to K or just expand raxml tree

A picture containing text

Description automatically generated

**Figure 3**. (A) Neighbor net showing the relatedness of all samples in the format of an unrooted phylogenetic network. Each tip is color coded according to the K=7 species delimitation scheme, and labeled according to sampling locality. (B) Species tree reconstructed by SVDquartets, with each sampling locality designated as a separate tip, with bootstrap branch support values are shown on internal branches. Boxes colored according to the sampling map (Fig. 2A) are used to map each of the relevant species delimitation schemes onto this tree.

Diagram

Description automatically generated

**Figure 4**. (A) Summary species tree inferred from individual gene trees under the multi-species coalescent model, with the posterior probability and the p-value from a polytomy test shown for each internal branch (pp/p-value). A significant (p < .05) result from a polytomy test indicates a statistical rejection of the null hypothesis that the given branch represents a hard polytomy. For internal branches recovered with low confidence (posterior probability < .9) each of the three potential unrooted quartet topologies are shown, labeled by the proportion of gene trees supporting the given topology. (B) Simultaneous visualization of 18,000 trees sampled via Markov chain Monte Carlo from the posterior distribution of species trees inferred under the multispecies coalescent model using SNAPP. (C) Maximum likelihood tree inferred from allele frequency data for each of the seven population tips, with migration edges successively added until > 99.8% of the overall variance in allele frequencies could be explained by the tree. Branch support for internal branches shown based on 100 bootstrap replicates of the optimal tree.

Chart

Description automatically generated

**Figure 5**. (A) Heatmap corresponding to the value of the *f*-branch statistic tested under the topology supported by the greatest proportion of gene trees using an unrooted quartet summary species tree approach. Gray boxes indicate that the given tips/branches cannot be tested under an ((A,B),C,D) framework based on the specified species tree topology. Asterisk used to indicate significant *f*-branch values based on a FDR-adjusted p < .05 significance threshold. (B-C) *D* Frequency Spectrum plotted depicting the value of Patterson’s *D* for each allele frequency bin. Topology used to generate each *D*FS shown, with bracket identifying the focal comparison sharing excess alleles.

Scatter chart

Description automatically generated

**Figure 6**. (A) Principal components analysis (PCA) describing the two major axes of variation in the genotypes of samples from localities 1:10 and 12:18. Because PC1 distinguishes the two species, three samples each with the highest and lowest values on PC1 were used to identify diagnostic SNPs (allele frequency difference > .8) for panels B-C (B) Triangle plot depicting the estimated hybrid index, and heterozygosity at species diagnostic SNPs for each sample. An F1 hybrid between the two species would be located at the top of the triangle. (C) Plot showing the genotype for each sample at each of the 128 species diagnostic SNPs, where red indicates homozygous for the 1:10 allele, blue indicates homozygous for the 12:18 allele, gray indicates heterozygous, and white indicates a missing genotype. The longest identified run of heterozygosity is highlighted by a red bracket and labeled by the chromosomal positions of the first and last SNP. Sample with mixed ancestry according to STRUCTURE assignment (Fig. 2J) indicated with an arrow. (D) Manhattan plot showing Fst between the two species, ordered by genomic position. 13 SNPs highlighted in red indicate that the given SNP was fixed for an ABBA allele pattern across all samples using the rooted four species topology shown to the left.

Chart

Description automatically generated

**Figure 7**. (A) Dot plot showing per sample heterozygosity, with scale on the left axis, and nucleotide diversity (Pi) for each of the seven putative species depicted by an asterisk, with corresponding scale on the right axis. (B) Heatmap color-coded according to pairwise FST between each of the seven putative species. Cells above the diagonal contain the FST value, while cells below the diagonal show the number of fixed SNP differences for the given pairwise comparison.