**ddRAD-seq phylogenetics based on nucleotide, indel, and presence-absence
polymorphisms: analyses of two avian genera with contrasting histories**

Jeffrey M. DaCosta^{a*} and Michael D. Sorenson^a

^aDepartment of Biology, Boston University, Boston, USA

*Corresponding author

Phone: 617-353-6989

E-mail: dacostaj@bu.edu

Postal: Department of Biology, Boston University, 5 Cummington Mall, Boston, MA 02215

Abstract

Genotype-by-sequencing (GBS) methods have revolutionized the field of molecular ecology, but their application in molecular phylogenetics remains somewhat limited. In addition, most phylogenetic studies based on large GBS data sets have relied on analyses of concatenated data rather than species tree methods that explicitly account for genealogical stochasticity among loci. We explored the utility of “double-digest” restriction site-associated DNA sequencing (ddRAD-seq) for phylogenetic analyses of the *Lagonosticta* firefinches (family Estrildidae) and the *Vidua* brood parasitic finches (family Viduidae). As expected, the number of homologous loci shared among samples was negatively correlated with genetic distance due to the accumulation of restriction site polymorphisms. Nonetheless, for each genus, we obtained data sets of ~3000 loci shared in common among all samples, including a more distantly related outgroup taxon. For all samples combined, we obtained > 1000 homologous loci despite ~20 my divergence between estrildid and parasitic finches. In addition to nucleotide polymorphisms, the ddRAD-seq data yielded large sets of indel and locus presence-absence polymorphisms, all of which had higher consistency indices than mtDNA sequence data in the context of concatenated parsimony analyses. Species tree methods, using individual gene trees or single nucleotide polymorphisms as input, generated results broadly consistent with analyses of concatenated data, particularly for *Lagonosticta*, which appears to have a well resolved, bifurcating history. Results for *Vidua* were also generally consistent across methods and data sets, although nodal support and results from different species tree methods were more variable. Lower gene tree congruence in *Vidua* is likely the result of its unique evolutionary history, which includes rapid speciation by host shift and occasional hybridization and introgression due to incomplete reproductive isolation. We

conclude that ddRAD-seq is a cost-effective method for generating robust phylogenetic data sets, particularly for analyses of closely related species and genera.

Keywords

ddRAD-seq, phylogenomics, species tree methods, *Lagonosticata*, *Vidua*, brood parasitism.

1. Introduction

Molecular phylogenetics has evolved during the past decade due to conceptual, technological, and analytical/computational developments. Among these has been a growing appreciation for the stochasticity of lineage sorting, resulting in variation among gene trees and potential incongruence with the species trees (Pamilo and Nei, 1988; Page and Charleston, 1997; Edwards and Beerli, 2000; Nichols, 2001; Rosenberg and Nordborg, 2002). This has led to an increasingly important interface between population genetics and phylogenetics (Edwards, 2009), and a shift toward datasets comprising multiple nuclear loci that more broadly sample the stochastic sorting of ancestral polymorphisms and the accumulation of informative variation along the internodes separating speciation events. Multi-locus datasets have in turn spurred the development of new phylogenetic methods that accommodate genealogical stochasticity in the estimation of phylogeny (e.g., Edwards et al., 2007; Liu, 2008; Heled and Drummond, 2010; reviewed in Knowles and Kubatko, 2010).

At the same time, advances in DNA sequencing technology have led to the development of methods that sample hundreds to thousands of genomic loci in a rapid and cost-effective manner. For most applications in population genetics and systematics, however, harnessing the power of this new technology requires methods that consistently recover a set of homologous loci across multiple samples (Davey et al., 2011; McCormack et al., 2013). The number of loci targeted should be large enough to take full advantage of the current throughput of DNA sequencers but still small enough to allow for adequate coverage for a large set of multiplexed samples and loci processed in a single run, thereby minimizing cost. Methods based on multiplexing a large number of locus-specific primer pairs (e.g., Binladen et al., 2007) or using hybridization probes

to isolate portions of the genome for shotgun sequencing (e.g., Mamanova et al., 2010) require prior knowledge of genomic sequences for the taxa of interest to design conserved primers/probes, an approach that has associated strengths and weaknesses (McCormack et al., 2013). An alternative approach is to use restriction-site associated DNA sequencing (RAD-seq) (Miller et al., 2007; Baird et al., 2008), which uses one or more restriction enzymes to target homologous loci among a set of samples and therefore requires no *a priori* genomic resources. The method provides a cost-effective means of broadly sampling the genome and, along with similar “genotype-by-sequencing” (GBS) methods, has emerged as a powerful and increasingly popular tool for an array of population level applications in molecular ecology (Davey and Blaxter, 2010; Narum et al., 2013).

Relatively few studies have explored the utility of RAD-seq data for phylogenetic analysis (e.g., Near and Benard, 2004; Rubin et al., 2012; Eaton and Ree, 2013; Jones et al., 2013; Keller et al., 2013; Nadeau et al., 2013; Wagner et al., 2013; Cruaud et al., 2014; Hipp et al., 2014). As with other methods based on restriction sites, an important complication of using RAD-seq in species level (or higher) phylogenetic analyses is the inevitable reduction in the number of homologous loci captured among samples as mutations in enzyme recognition sites accumulate with increasing genetic divergence among samples. Rubin *et al.* (2012) explored this issue by computationally extracting RAD loci from reference genomes of *Drosophila*, mammals, and fungi. While these simulated RAD-seq data produced accurate topologies for *Drosophila* and for shallow nodes in the mammalian and fungal trees, nodes representing speciation events > 60 Mya were not reliably recovered due in part to a dearth of homologous RAD loci recovered from highly divergent taxa. Patterns of locus recovery among taxa, reflecting the gain and loss of enzyme recognition sites, represent a potentially useful source of characters for phylogenetic

analysis. This approach, however, has not been tested; all RAD-seq phylogenetic studies to date have relied on nucleotide character matrices.

Applying methods that accommodate incomplete lineage sorting in the estimation of species trees to large RAD-seq datasets also presents a computational challenge. To date, empirical RAD-seq phylogenetic studies have largely side-stepped this issue by using concatenated data and assuming that phylogenetic signal in the aggregate overrides noise generated by loci that conflict with the underlying species tree (e.g., Nadeau et al., 2013; Wagner et al., 2013; Cruaud et al., 2014; but see Eaton and Ree, 2013). This approach may fail, however, if any relationships within the species tree fall within the “anomaly zone,” in which adding data leads to stronger confidence in a tree that is discordant with the species tree (Degnan and Rosenberg, 2006; Kubatko and Degnan, 2007; Rosenberg and Tao, 2008; but see Huang and Knowles, 2009). In addition to the computational load of processing a large number of loci, the application of more sophisticated methods may be hindered by the difficulty of estimating individual gene trees and/or model parameters from the typically short sequences generated by RAD-seq. For example, Bayesian methods (Liu, 2008; Heled and Drummond, 2010) that simultaneously estimate a large set of parameters from thousands of short loci are unlikely to converge on reliable parameter estimates even when using a simplified model (e.g., linking model and rate parameters over many loci). While each locus may provide relatively little phylogenetic information, the volume of loci generated by RAD-seq produces a large number of single nucleotide polymorphisms (SNPs), and new methods that use SNP data to infer species trees (Bryant et al., 2012; Kubatko and Chifman, 2014) may be more effective.

We explored the utility of RAD-seq data for phylogenetic analysis using a variety of character matrices and tree inference methods, and compared results for two avian genera with

contrasting ecology and evolutionary histories. The firefinches (genus *Lagonosticta*) comprise 10 species that are distributed in sub-Saharan Africa (Fry and Keith, 2004), and likely diversified over the course of several million years through a typical process of allopatric speciation (Sorenson et al., 2004). Firefinches are the primary hosts of the brood parasitic indigobirds, which together with whydahs comprise the genus *Vidua*, which has a total of 19 recognized species. Imprinting on hosts during development shapes the courtship and mating behavior of adult indigobirds (i.e., songs and mate choice preferences), providing a mechanism for rapid speciation when a novel host is colonized (Payne, 1973; Payne et al., 2000; Sorenson et al., 2003). The behaviors that drive speciation by host shift in indigobirds are shared by most other *Vidua* (i.e., the whydahs), appear to be ancestral in the clade, and may also lead to hybridization between established parasitic species (Payne and Sorenson, 2004). Thus, post-speciation gene flow in this clade may obscure species-level relationships. The contrasting evolutionary histories of these two genera are apparent in a comparison of mitochondrial DNA (mtDNA) phylogenies (Sorenson et al., 2004). Relationships among *Lagonosticta* species are generally well resolved, whereas the branching order of *Vidua* clades is uncertain. In this study, we present results of phylogenetic analyses for each genus based on thousands of “double-digest” RAD-seq (ddRAD-seq) loci (Peterson et al., 2012; DaCosta and Sorenson, 2014); we implement a variety of phylogenetic methods/approaches and test the utility of different categories of character information, including SNPs, indels, and the presence-absence of loci in the data set. We also evaluate the “decay” of homologous ddRAD-seq loci with increasing genetic divergence among samples and compare the extent of gene-tree/species-tree congruence between the two genera.

2. Materials and methods

2.1. Taxon sampling, DNA sequencing, and bioinformatics

Tissue samples were collected during fieldwork in Cameroon and Tanzania, or obtained from tissue collections at natural history museums (Table 1). The *Lagonosticta* dataset includes nine *Lagonosticta* samples representing seven species and a single brown twinspot (*Clytospiza monteiri*) sample as the outgroup; mtDNA sequence data support a sister group relationship between these two genera (Sorenson et al., 2004). The *Vidua* dataset includes 14 *Vidua* samples representing 12 species and a single cuckoo finch (*Anomalospiza imberbis*) sample as the outgroup (Sorenson and Payne, 2001). Genomic DNA was extracted from tissue samples using a DNeasy Tissue Kit (Qiagen Inc.). Mitochondrial DNA sequences from each taxon were collected following previously described methods (Sorenson and Payne, 2001; Sorenson et al., 2004). In some cases, different samples of a given species were used for mtDNA and ddRAD sequencing (see Table 1). We collected sequence data from the following mtDNA regions, comprising a total of 2,186 base pairs (bp): tRNA methionine, NADH dehydrogenase subunit 2 (ND2), tRNA tryptophan, NADH dehydrogenase subunit 6 (ND6), tRNA glutamic acid, and control region. Mitochondrial sequences for each taxon were aligned and reconciled in Sequencher v4 (Gene Codes Corporation); interspecific alignments were manually edited in Se-Al v2 (<http://tree.bio.ed.ac.uk/software/seal/>).

The protocol we used to generate and analyze ddRAD-seq fragment libraries yields data sets for which PCR and sequencing errors are inconsequential and genotypes that are concordant with neutral and Hardy-Weinberg expectations (see DaCosta and Sorenson, 2014). These methods are summarized below; additional details are provided in DaCosta and Sorenson (2014). Genomic

DNA (~1.0 µg) was simultaneously digested with 10 U of SbfI and 10 U of EcoRI. Following ligation of bar-coded adapters, fragments in the 300-450 bp size range (178-328 bp exclusive of adapters) were selected in an agarose gel, although smaller fragments were consistently captured in all samples (see DaCosta and Sorenson, 2014). Amplified fragment libraries were quantified using qPCR and then pooled in equimolar amounts for multiplexed sequencing (101 bp single-end reads) on an Illumina HiSeq 2000 machine. Data for this study were collected in two different sequencing runs, both of which included additional samples for population genetic analyses of indigobirds.

Sequence reads were assigned to individual samples based on barcode sequences and processed as described in DaCosta and Sorenson (2014), generally resulting in 97 bp reads beginning with the 8 bp SbfI recognition site. Sequences from some shorter loci (genomic SbfI-EcoRI fragments < 97 bp) were also recovered and included in our analyses after trimming of adapter sequences. Data for all samples were pooled and clustered into putative loci based on sequence similarity using the UCLUST function in USEARCH v5 (Edgar, 2010) with an *-id* setting of 0.85. The highest quality read from each cluster was mapped to the *Taeniopygia guttata* reference genome using BLASTN v2 (Altschul et al., 1990). Clusters that generated identical BLAST hits (± 50 bp) were combined, an approach that minimizes error associated with subjective similarity cutoff thresholds (see by Harvey et al., 2015) and substantially improves the clustering of loci with large indel differences between taxa. Clusters that did not generate a BLAST hit were carried through the analyses as anonymous loci. All sequence reads within each cluster were aligned using MUSCLE v3 (Edgar, 2004). Alignments with 5' or 3' gaps due to indels and/or restriction site polymorphisms were trimmed to the length of the shortest sequence during genotyping. In addition, alignments with ≥ 2 SNPs in the last five bp were flagged for

manual evaluation and editing based on comparison to the zebra finch reference genome; this allowed us to correct alignments for loci in which indels near the end of the alignment were not recognized by MUSCLE. Genotypes for each sample and locus were called using a custom python script (RADGenotypes.py, available at <https://github.com/BU-RAD-seq/ddRAD-seq-Pipeline>) based on the following categories: 1) “missing” – no reads recovered for that locus; 2) “good” – unambiguously genotyped as either a homozygote or heterozygote; 3) “low depth” – apparent homozygote with less than five reads; or 4) “flagged” – ambiguous genotype based on strong departure from the expected 50:50 ratio of sequence reads in a heterozygote and/or the recovery of three or more alleles in a single sample (see DaCosta and Sorenson 2014 for additional details). Putative loci with flagged genotypes for multiple samples result when two or more similar loci are clustered together; due to uncertain homology, these clusters are not appropriate for phylogenetic analyses and were excluded. In addition to genotyping SNPs, our script identifies and scores each unique indel, regardless of length, as a single 0/1 character.

We applied several filters to generate the largest possible data matrices with no missing data or ambiguous genotypes for each genus, plus their respective outgroups. First, we retained loci with at least one SNP and no missing or flagged genotypes, allowing up to two samples with low-depth genotypes, which in some cases will represent only one of the two alleles in heterozygous individuals. We then randomly selected one allele per locus for samples with “good” heterozygous genotypes. This approach allowed us to include low-depth genotypes, for which a second allele might be missing, without introducing missing data to the matrix. Rarely, this approach eliminated all polymorphism at a locus, so random sampling of alleles was repeated until at least one SNP was retained for each polymorphic locus. Unless otherwise noted, the *Lagonosticta* and *Vidua* datasets were analyzed separately.

2.2. Effect of phylogenetic divergence on the number of shared ddRAD loci

We quantified the “decay” of shared ddRAD loci with phylogenetic distance by relating pairwise genetic distance to the number of shared loci between samples. HKY (Hasegawa et al., 1985) genetic distances were calculated using a common set of loci ($n = 1,012$) recovered in all samples in both datasets. At a given ddRAD locus, one or more samples may have low sequencing depth due to a variety of reasons, including amplification biases, poor library quality, and/or nonspecific enzyme cutting (see DaCosta and Sorenson, 2014). Therefore, we calculated the number of loci that were shared between each pair of samples for a set of 14,942 loci for which every sample had a sequencing depth of either zero or ≥ 10 reads (i.e., all samples have coverage of either 0x or ≥ 10 x), thereby increasing the likelihood that patterns of locus presence and absence were primarily due to polymorphisms in restriction sites. As a less stringent metric, we also calculated pairwise locus sharing by counting the number of loci at which two focal samples both had ≥ 5 reads.

2.3. mtDNA phylogenetic analyses

Phylogenetic analyses of the concatenated mtDNA sequence data were conducted using unweighted maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI). Heuristic parsimony searches were performed in PAUP* v4.0b10 (Swofford, 2002) with 1000 random addition replicates, TBR branch swapping, and gaps treated as missing data. Nodal support was estimated with 1000 bootstrap replicates. ML and BI analyses were conducted with four data partitions: tRNAs, ND2, ND6, and control region. ML searches were conducted in RAxML v7.4.2 (Stamatakis, 2006) with the GTRGAMMA model, 20 searches for the best tree,

and 1000 standard bootstrap replicates. We used BEAST v1.8 (Drummond et al., 2012) for BI analyses, implementing a GTR+G substitution model, a relaxed lognormal clock, and a birth-death coalescent prior with incomplete sampling. Two independent searches were run, each with 50 million generations and sampling every 5000 generations. We discarded the first 10% of samples as burn-in, and calculated the maximum clade credibility tree from the combined posterior distribution of the two runs.

2.4. ddRAD phylogenetic analyses using concatenated data

Phylogenetic networks were computed using concatenated sequences and the Neighbor-Net algorithm in SPLITSTREE v4.12.3 (Huson and Bryant, 2006). Concatenated sequences, indel characters, and sequences+indels were used to reconstruct phylogenetic trees with MP, ML, and BI methods. Unweighted MP analyses were conducted in PAUP* with the same settings as in the mtDNA analyses. ML searches were conducted in RAxML with the GTRGAMMAI model for sequence data, the BINGAMMA model for indel data, 20 searches for the best tree, and 1000 standard bootstrap replicates. Four independent BI searches (50 million generations each with sampling every 5000 generations) were run in BEAST with a GTR+I+G substitution model for sequence data, the stochastic dollo model for indel data, a relaxed lognormal clock, and a birth-death coalescent prior with incomplete sampling. The number of generations required for convergence on a stable posterior distribution varied among runs (~5-30% of each run); thus, we summarized the two runs for each data set that converged earliest, discarding the first 40% of samples as burn-in, and calculating the maximum clade credibility tree from the combined posterior distribution of 12,000 samples.

2.5. ddRAD phylogenetic analyses using presence-absence data

Restriction site polymorphisms are expected to result in the gain and loss of ddRAD-seq loci among samples, such that the presence and absence of specific loci across taxa is a potential source of additional characters. Scoring the presence-absence of ddRAD loci for phylogenetic analyses is complicated by variable recovery of loci across samples due to insufficient sequencing depth, inconsistency in fragment library preparation, and star activity, which results in the haphazard appearance of loci that would otherwise be missing (DaCosta and Sorenson 2014). Thus, sufficient sequencing depth is needed to differentiate true presence-absence polymorphisms from biases associated with library preparation. Our experimental design produced relatively high per individual per locus sequencing depths (see Results), allowing us to generate a robust presence-absence character matrix for the *Lagonosticta* and *Vidua* datasets, respectively. We scored only those loci for which all samples had a sequencing depth of either zero or ≥ 10 . Each character in the data matrix represented a single ddRAD locus, with samples coded as either 0 (depth of zero) or 1 (depth of ≥ 10). Phylogenetic analyses of binary data matrices were conducted using MP, ML, and BI as described above for the concatenated indel data.

2.6. ddRAD phylogenetic analyses using individual gene trees

We used several methods to estimate species trees based on the individual gene trees for each ddRAD locus. Gene trees were estimated in PHYML v3.0 (Guindon et al., 2010) using an HKY substitution model, estimated base frequencies, and SPR branch swapping. Polytomies, which are common in the gene trees of short ddRAD-loci, are arbitrarily resolved by PHYML into ladderized bifurcations with internode branch lengths close to zero. Importantly, the order of

samples in the input file dictates the order of these bifurcations. We therefore randomized the order of taxa in each file before estimating gene trees in PHYML. We then used the Species Tree Analysis Web (STRAW) server (Shaw et al., 2013) to estimate species trees using the NJst (Liu and Yu, 2011), STAR (Liu et al., 2009), and MP-EST (Liu et al., 2010) methods. The NJst method takes unrooted gene trees as input, calculates a distance matrix among taxa based on the average number of internodes between all pairs of species, and then uses the neighbor-joining algorithm to estimate a species tree. Both the STAR and MP-EST methods require rooted gene trees; thus, individual gene trees were rooted using the respective outgroup for each clade and the “RerootTree” function on the STRAW server. The STAR method estimates a species tree using the average ranks of coalescences in gene trees, whereas the MP-EST method computes a pseudo-likelihood estimate of the multi-species coalescent. Species trees were also inferred with ASTRAL v4.4.4 (Mirarab et al., 2014) and STELLS v1.6 (Wu, 2012) using the gene trees estimated in PhyML. ASTRAL estimates a species tree that agrees with the largest number of quartet trees within a set of unrooted gene trees, whereas STELLS, taking rooted gene trees as input, uses maximum likelihood to search for the best species tree in a multispecies coalescent framework with incomplete lineage sorting. The following command line options recommended for large datasets were implemented in STELLS: -S, -d 1, -c 10, and -x 500. We randomly sampled gene trees from each dataset (with replacement) to create 100 non-parametric bootstrap replicates of the same size, which were run in NJst, STAR, MP-EST, ASTRAL, and STELLS to evaluate nodal support. Finally, species trees were estimated from rooted gene trees using the minimize deep coalescences (MDC) method in the program PHYLONET v3.4 (Than et al., 2008) with 100 bootstrap replicates.

We also used BUCKy v1.4.0 (Larget et al., 2010) for Bayesian concordance analysis, which estimates the proportion of loci that contain each clade (i.e., the clade concordance factor) while incorporating gene tree uncertainty by sampling from the posterior distribution of trees for each locus (Ané et al., 2007). BUCKy also implements a consensus method based on unrooted quartets to generate a “population tree” that consistently identifies the species tree when all discordance is due to incomplete lineage sorting (see BUCKy manual). For each ddRAD locus, posterior distributions of gene trees were generated in MrBAYES v3.1 (Ronquist and Huelsenbeck, 2003) with an HKY substitution model, 1 million generations, and sampling every 1000 generations; the first 500 samples (50%) were discarded as burnin. The α setting in BUCKy defines the *a priori* assumption of concordance among gene trees, with $\alpha = 0$ implying that all loci have the same topology and $\alpha = \text{infinity}$ implying that all loci have different topologies. For each data set, similar results were obtained with α settings of 1 and 100; results for $\alpha = 100$ are reported below. Each BUCKy analysis comprised two independent runs, with a burnin of 10,000 updates and 100,000 sampled updates per run.

2.7. ddRAD phylogenetic analyses using single nucleotide polymorphisms

Species trees were also estimated in a coalescent framework using SNP data with the SNAPP module (Bryant et al., 2012) built into BEAST2 v2.0.2 (Bouckaert et al., 2013) and the program SVDquartets v0.5 (Kubatko and Chifman, 2014). For SNAPP, which requires independent, biallelic SNPs coded as binary characters, we randomly sampled one biallelic SNP per locus. Two SNAPP runs were completed with default model parameters, six million generations, and sampling every 5000 generations. We assessed convergence in TRACER, and removed the first 10% of each sample as burn-in, leaving 2,160 samples from the combined runs to calculate the

maximum clade credibility tree. The SVDquartets analyses used the same sets of biallelic SNPs (coded as nucleotides) and sub-sampled 50,000 quartets. Sampled quartets were assembled into an unrooted species tree using the program QUARTET MAXCUT v2.1 (Snir and Rao, 2012). Nodal support of the SVDquartets tree was evaluated by generating 100 non-parametric bootstrap datasets using the SEQBOOT function in PHYLIP v3.69 (Felsenstein, 2005). Each replicate was analyzed in SVDquartets as described above, and nodal support values were calculated using the contree command in PAUP*.

2.8. Comparison of gene tree concordance between clades

In light of basic coalescent expectations, the contrasting evolutionary histories of *Lagonosticta* and *Vidua* suggest potential differences in patterns of gene tree concordance. As brood parasites, indigobirds (and other *Vidua* species) are expected to have smaller population sizes than their firefinch (*Lagonosticta* spp.) hosts (Payne and Payne, 1977), and thus intraspecific gene trees from the former should reflect higher rates of coalescence. Assuming a bifurcating history and all else equal, a greater proportion of *Vidua* gene trees should support a given node in the species tree (i.e., higher gene tree concordance in *Vidua*), after correcting for differences in the length of subtending internodes. However, imperfect host fidelity may result in hybridization among *Vidua* species, resulting in the migration of alleles between species and thus relatively greater gene tree discordance.

We evaluated these contrasting predictions by comparing estimates of population size, species tree branch lengths, and gene tree concordance for each genus. As a proxy for population size, we calculated nucleotide diversity for each taxon with DNAsp v5.10.1 (Librado and Rozas, 2009) using two alleles per sample for a set of 1,012 loci recovered in all taxa in both clades.

Mean nucleotide diversity for *Lagonosticta* and *Vidua* was compared with a student's t-test in R v2.15 (R Core Team, 2012). We examined gene tree concordance by plotting the relationship between subtending internode length and the proportion of gene trees recovering the adjacent node. Given that PHYML arbitrarily resolves polytomies (see above), we collapsed nodes of length ~0 before summarizing concordance of the ML trees for each locus and genus. Similarly, we used constrained searches in PAUP* to determine the number of loci providing unambiguous MP support for each node in the consensus species tree. To generate comparable estimates of species tree internode lengths, we use the 1,012 loci recovered in both datasets and one allele per sample per locus to estimate branch lengths for a constrained topology that matched the consensus species tree results for each genus (see Results). Internode lengths were estimated in a BEAST analysis, run as described above for other analyses of concatenated data. The linear regressions between subtending internode branch length and gene tree concordance for each genus were compared using an analysis of covariance in R.

3. Results

3.1 mtDNA phylogenies

Phylogenetic analyses of mtDNA yield a generally well-resolved phylogeny for *Lagonosticta*, with the exception of a weakly supported sister relationship between *L. rara* and *L. larvata* (Fig. 1). In contrast, basal relationships among several *Vidua* lineages are uncertain in the mtDNA tree (Fig. 1). Consistent with previous analyses (Sorenson et al., 2004), there is

essentially a four-way polytomy at the base of the *Vidua* tree, as well as uncertain relationships among members of a clade comprising *V. regia*, *V. fischeri*, and the indigobirds.

3.2. ddRAD data summary

We generated ~9.9 and ~13.3 million sequence reads for the *Lagonosticta* and *Vidua* ddRAD datasets, respectively. Demultiplexed, raw sequence reads are available in the National Center for Biotechnology Information Sequence Read Archive (see Table 1). A similar number of loci with ≥ 5 reads (~8,900) was recovered from all samples (Table 2), suggesting consistency in fragment library quality. Furthermore, variation in overall sequencing depth per individual did not have an effect on the number of loci recovered (linear regression, $F_{1,23} = 0.24$, $P = 0.63$; Fig. 2A) or median depth per locus in our final datasets (linear regression, $F_{1,23} = 2.95$, $P = 0.10$; Fig. 2B). This illustrates that our protocol resulted in consistent recovery of loci in the selected size range and sufficient depth to score both presence-absence of loci and SNPs within loci (i.e., individual genotypes). A comparable number of loci passed our selection criteria (i.e., ≥ 1 SNP, no missing or flagged genotypes, ≤ 2 low depth genotypes) in each dataset (2,860 for *Lagonosticta* and *Clytospiza*; 3,100 for *Vidua* and *Anomalospiza*; Table 2). Despite fewer samples and a more closely related outgroup, the *Lagonosticta* dataset had a slightly higher level of polymorphism (Table 2). For presence-absence analysis, there were 14,492 loci in which all 25 samples had either zero or ≥ 10 reads.

3.3. Genetic divergence and number of shared ddRAD loci

Using the 1,012 loci common to all samples in both datasets, intra-generic HKY genetic distances ranged from 0.004 to 0.021 in *Lagonosticta* and from 0.001 to 0.010 in *Vidua*. Genetic

distances between the two genera clustered at ~ 0.047 . As expected, the number of shared loci was negatively correlated with pairwise genetic divergence (Fig. 3). Using a conservative set of 14,492 loci in which all samples had either zero or ≥ 10 reads (Fig. 3A), intra-generic pairs shared $\sim 2,600$ loci on average, whereas inter-generic pairs shared $\sim 1,400$ loci. When using the less stringent criterion of ≥ 5 reads in both focal samples in a given pairwise comparison, the number of shared loci increased substantially (Fig. 3B), with the more recently evolved *Vidua* taxa sharing more loci on average ($7,137.8 \pm 412.8$) than *Lagonosticta* taxa ($6,096.5 \pm 706.4$).

3.4. ddRAD *Lagonosticta* phylogeny

The wide range of phylogenetic methods and data subsets described above (see Methods) converged on a single topology for *Lagonosticta*, with most analyses yielding 100% bootstrap support or posterior probabilities of 1.0 for all nodes (Fig. 4). This topology also matches the BEAST mitochondrial gene tree, including the sister relationship between *L. rara* and *L. larvata*. Exploration of gene trees for individual ddRAD loci also show overwhelming support for this sister relationship compared to alternative topologies (Fig. 5A). While most of our phylogenetic analyses defined *Clytospiza* as the outgroup *a priori*, Bayesian analyses that used the data to infer rooting (i.e., BEAST and SNAPP) recovered a monophyletic *Lagonosticta* (Fig. 4, node A). Support values for a few *Lagonosticta* nodes (D, E, and F) were somewhat lower in the NJst, MPest, STAR, and ASTRAL analyses, four of the six methods that build a phylogeny using individual gene trees as input. Analyses in BUCKy yield concordance factors, which indicate the proportion of gene trees including a given clade; these values are fundamentally different from the bootstrap and posterior probability values generated in other analyses. Relatively low

concordance factors in our analysis of short ddRAD-seq loci primarily reflect a lack of relevant phylogenetic signal in the gene trees for many loci rather than conflicting signals among loci.

3.4. ddRAD *Vidua* phylogeny

Despite a large data set of 3,100 loci (comprising > 276,000 bp per taxon), support values for several key relationships among *Vidua* lineages were variable across analyses (Fig. 4). Concatenated analyses, including nucleotides, indels, and locus presence-absence data, were generally consistent in resolving the basal relationships among four *Vidua* lineages as shown in Fig. 4, including a sister relationship between the paradise whydah clade and all other *Vidua* (node E), as well as a sister relationship between *macroura* and *hypochoerina* (node F). These basal relationships were also recovered in analyses based on SNP data (i.e., SNAPP and SVDquartets). In contrast, PhyloNet, which minimizes deep coalescences, and STELLS favored a sister relationship between paradise-whydahs and *hypochoerina*+*macroura*. Likewise, nodes E and F were not recovered in the BUCKy concordance tree, which placed *hypochoerina* as sister to all other *Vidua*. As in the firefinch analyses, four methods summarizing the results of individual gene trees (NJest, MPest, STAR, ASTRAL) yielded lower support values, but optimal trees that were generally consistent with other analyses (e.g., recovery of nodes E and F). Thus, the overall consensus phylogeny from the nuclear ddRAD-seq data (Fig. 4) suggests a different resolution of the poorly supported nodes in the mtDNA gene tree (Fig. 1B). In addition to resolving the placement of *hypochoerina* and *macroura*, the full range of analyses shows consistent though not overwhelming support for a sister relationship between *regia* and *fischeri*. A simple summary of the number of individual ddRAD gene trees supporting alternative placements of these taxa (Fig. 5B-C) reveals support for the consensus tree while also illustrating the variability among gene

trees in the *Vidua* data set. For example, 90 ddRAD loci recovered a *hypocherina-macrourea* sister relationship in ML analyses, whereas sets of 32 to 56 loci supported alternative placements of these taxa. Other relationships were more consistently supported, including monophyly of *Vidua* (node A) for analyses that inferred rooting of the tree, monophyly of the paradise whydahs (node B), and a clade (node H) comprising *regia*, *fischeri* and the indigobirds (*chalybeata*, *camerunensis*, *raricola*, and *purpurascens*). Finally, relationships among three closely related West African indigobirds (nodes K and L) were variable across analyses (Fig. 4).

3.5. Gene tree concordance in *Lagonosticta* and *Vidua* datasets

Qualitative differences in ddRAD gene tree concordance between the *Lagonosticta* and *Vidua* datasets can be visualized in phylogenetic networks based on concatenated data sets and cloudograms based on SNAPP results (Fig. 6). Uncertainty in the *Lagonosticta* network primarily involves possible alternative connections of the outgroup *Clytospiza*, corresponding to alternative rootings of the ingroup. The *Vidua* network in contrast has much shorter branch lengths (except for the highly divergent outgroup *Anomalospiza*), reflecting less genetic divergence overall and more recent common ancestry. In addition, the network illustrates the uncertain relationships discussed above, including a near polytomy involving four basal *Vidua* lineages and a trichotomy among *regia*, *fischeri* and indigobirds. Cloudograms based on SNAPP results also illustrate much greater uncertainty in the *Vidua* species tree as compared to *Lagonosticta*.

We used a set of 1,012 ddRAD loci shared in common between the *Lagonosticta* and *Vidua* data sets to estimate relative branch lengths in a single concatenated topology comprising both genera (Fig. 7A). The *Lagonosticta* clade has a deeper basal node and generally longer internode

branch lengths relative to the *Vidua* clade, reflecting both a longer history for *Lagonosticta* and a faster rate of nuclear sequence evolution in estrildid finches (Sorenson et al., unpublished data). Moreover, *Lagonosticta* nodes were recovered more consistently across individual loci as compared to *Vidua* nodes (Fig. 7A); nodes in the *Lagonosticta* species tree were recovered in 9-30% of ML gene trees, whereas recovery of *Vidua* nodes ranged from 1% to a maximum of only 18% of gene trees.

Using the length of the subtending internode as a proxy for time since the previous speciation event, we find a significant positive correlation with the proportion of ML gene trees recovering a given node in both *Lagonosticta* (linear regression, $F_{1,6} = 792.0$, $P < 0.001$) and *Vidua* (linear regression, $F_{1,11} = 768.2$, $P < 0.001$). The slopes of these two linear regressions, however, were significantly different (ANCOVA, branch*genus, $F_{1,15} = 8.32$, $P = 0.01$; Fig. 7B). Nearly identical results, including the significant difference in slope, were obtained for analyses based on the proportion of MP gene trees with a given node (results not shown). Estimates of nucleotide diversity were significantly lower in *Vidua* than in *Lagonosticta* (student's t -test, $t_{21} = 2.24$, $P = 0.04$; Fig. 7C), consistent with the expectation of smaller effective population sizes in *Vidua* (see above). All else equal, coalescent theory would predict that *Vidua* lineages should therefore have faster coalescence times and greater gene tree concordance for a given internode length as compared to *Lagonosticta*. We find evidence, however, for the opposite result; concordance increases more rapidly in *Lagonosticta* with increasing internode length (Fig. 7B).

4. DISCUSSION

4.1. Decay of shared ddRAD loci with increasing genetic divergence

As expected, the number of shared ddRAD loci between taxa was negatively correlated with the evolutionary divergence between samples (Fig. 3). Nonetheless, our ddRAD-seq protocol generated ample data for species-level phylogenetic analyses, as well as a large set of loci shared in common across all of the sampled parasitic finches and estrildid finches, families that likely diverged ~20 my (Sorenson and Payne, 2001). RAD-seq and other GBS methods that use restriction enzymes to subsample the genome are unlikely to produce enough shared loci to resolve relationships among taxa that diverged > 100 Ma (Rubin et al., 2012). For such higher-level phylogenetic questions, hybridization probes to capture ultraconserved elements or other conserved genomic sequences may be a more viable option (e.g., Carstens et al., 2012; Crawford et al., 2012; Lemmon et al., 2012; Faircloth et al., 2013).

Most of the analyses presented in this paper are based on a relatively conservative approach to data filtering (i.e., loci with data for all samples and ≤ 2 samples with low depth), but this resulted in data matrices of > 250 kbp for each of the clades, including 1,012 loci comprising ~90 kbp recovered in all samples in both data sets. Employing more permissive filtering (e.g., Fig 2B), or modifying our laboratory protocol (e.g., using restriction enzymes with higher cutting frequencies), would have generated an incomplete data matrix with thousands of additional loci. Other studies have suggested that the accuracy of some species tree methods is resilient to missing data (Hovmoller et al., 2013), and that excluding RAD loci with missing data results in a bias against portions of the genome with higher mutation rates (Huang and Knowles, In press). The effects of missing data on all of the various species tree methods used here, however, have not been rigorously analyzed. Further exploration of the effects of missing data is important because the set of loci recovered in all samples will inevitably shrink and the set of unique loci

will grow with an increasing number of samples. Even when adding closely related species, additional restriction site variants and variability in the quality of fragment libraries (see DaCosta & Sorenson, 2014) will result in some reduction in the number of loci common to all samples.

4.2. ddRAD phylogenetics using concatenated and presence-absence data

Previous phylogenetic studies using RAD-seq data have mainly relied on concatenated sequences to estimate phylogenies, an approach that is computationally expedient but ignores the potential incongruence among loci due to the stochasticity of lineage sorting and/or the effects of selection (Maddison, 1997; Nichols, 2001). Thus, concatenation may obscure patterns in the data that provide insight into the evolutionary history of a clade. Moreover, concatenation has the potential to be positively misleading, with additional data reinforcing an incorrect inference of the species tree (Kubatko and Degnan, 2007), although the combination of mutational and coalescent variance may make so-called “anomalous gene trees” relatively rare in practice (Huang and Knowles, 2009). Empirically, our analyses of *Lagonosticta* and *Vidua* based on concatenated sequence data produced topologies that were generally consistent with those generated using a range of species tree methods, but with higher nodal support values for some relationships (Fig. 4) due to fewer model parameters (Kelchner and Thomas, 2007; Liu et al., 2015). These large ddRAD-seq data sets, however, also provide the opportunity to examine topological variance across independent loci as a means of qualitatively assessing the likelihood of anomalous results (Fig. 5; see below).

Phylogenetic analyses of concatenated RAD-seq or GBS data typically rely on programs that treat gaps as missing data (e.g., PhyML, RAxML, default settings in BEAST, PAUP*), thereby ignoring two important sources of informative characters that can be derived from ddRAD-seq

data: indels within loci, and locus presence-absence polymorphisms generated by mutations in restriction sites. Indel characters may provide phylogenetic signal for nodes that are otherwise difficult to resolve (Rokas and Holland, 2000). We generated over a thousand indel characters in each of our data sets and analysis of just these indels produced results generally congruent with other analyses (Fig. 4), demonstrating their suitability for phylogenetic inference. As originally conceived, “RAD-tags” were designed to generate presence-absence polymorphisms using microarray technology (Miller et al., 2007), but we are not aware of other RAD-seq studies using presence-absence of loci for phylogenetic inference. We caution that scoring presence-absence characters from ddRAD-seq data requires consistent, high-quality fragment libraries that are sequenced at sufficient depth so that presence-absence is primarily a function of restriction site variation rather than the stochastic recovery of loci across samples. This contrasts with recent studies showing that very low coverage sequencing and Bayesian statistics may be effective for estimating population genetic parameters (e.g., Buerkle and Gompert 2013), and highlights the importance of matching experimental design with methods of data analysis. Scoring is also complicated by the infrequent, low-depth recovery of loci that should be “absent” due to restriction enzyme star activity (i.e., non-specific cutting at sites that differ at one position from the canonical recognition sequence) and other potential anomalies of ddRAD-seq fragment library preparation (DaCosta and Sorenson, 2014). We therefore used a conservative criterion for defining presence-absence characters (i.e., loci at which all samples had either zero or ≥ 10 reads). Nonetheless, this approach generated $> 10,000$ binary presence-absence characters for each dataset, and topologies generated from these data were largely consistent with those produced using other sources of data (Table 3, Fig. 4).

4.3. RAD phylogenetics based on gene trees or SNPs

Species tree methods based explicitly on coalescent theory and accommodating variation among gene trees are conceptually appealing, but potentially challenging in practice. One issue is the lack of resolution typically provided by individual nuclear loci, particularly the short loci generated by GBS approaches. Note, however, that recombination is less frequent in short loci and therefore the gene trees they produce are more likely to reflect a single genealogical history (Edwards, 2009). While a substantial fraction of the individual ddRAD loci in our data sets included relatively few SNPs (24.2% and 26.6% of loci had ≤ 3 SNPs in the *Lagonosticta* and *Vidua* datasets, respectively), and many SNPs simply represented divergence between the outgroup and ingroup, a variety of species tree methods generally converged on the same topology, or produced topologies that conflicted only at poorly supported nodes (Fig. 4). This is particularly impressive given that nodes in the most commonly recovered species tree of each genus were represented in only 9-30% (*Lagonosticta*) and 1-18% (*Vidua*) of individual ML gene trees (Fig. 7A). One exception is that analyses of *Vidua* using both minimize deep coalescences (PhyloNet) and a maximum likelihood estimate of the multispecies coalescent (STELLS) recovered a clade (*V. hypocherina*+paradise-whydahs) with high support that was not recovered in any of the other analyses. The STELLS settings used to decrease computation time restricted the search to near-MDC trees, which likely explains the similarity in results for these two programs. Also, the MDC method may fail to reconstruct the true species tree under some circumstances (Than and Nakhleh, 2009; Than and Rosenberg, 2011), and it is possible that the method is inconsistent given the short internodes at the base of the *Vidua* tree.

Species tree analyses based on SNPs (SNAPP and SVDquartets) produced results similar to those generated by other methods (Fig. 4), even when using just one randomly selected biallelic

SNP per locus. The utility of SNPs in estimating population-genetic parameters and reducing confidence intervals around these estimates has been recognized for some time (Brumfield et al., 2003). It is not until relatively recently, however, that cost-effective methods have become available to collect large numbers of independent SNPs from non-model organisms. As the generation of these data becomes increasingly feasible there should be increased interest in the utility of SNPs for estimating species trees based on the multi-species coalescent and further development of relevant computational tools (Brito and Edwards, 2009; Edwards, 2009).

4.4. Comparison of *Lagonosticta* and *Vidua* results

The contrasting evolutionary histories of firefinches (*Lagonosticta* spp.) and parasitic finches (*Vidua* spp.) were reflected in our analyses. Speciation in *Lagonosticta* likely followed a more gradual process of allopatric isolation, and all analyses of the ddRAD-seq data resulted in the same consensus topology, generally with high nodal support values (Fig. 4). The fact that a variety of character sets (i.e., sequences, indels, locus presence-absence, SNPs) and analytical approaches converged on this topology suggests that this is a robust estimate of their phylogeny. In contrast, the brood parasitic finches likely have a more complex evolutionary history, involving rapid speciation via host shift and post-divergence gene flow (Sorenson et al., 2003; Payne and Sorenson, 2004; Sorenson et al., 2004; Sefc et al., 2005). Thus, phylogenetic relationships within *Vidua* may be obscured by shared ancestral polymorphism in a group of recently evolved species with short intervals between speciation events as well as reticulation due to occasional hybridization and introgression. Estimates of the *Vidua* phylogeny were substantially less consistent than those derived for the *Lagonosticta* phylogeny, despite similarly sized datasets.

All else equal, greater gene tree concordance is expected in *Lagonosticta* given a longer history of more gradual diversification. By controlling for internode length, we sought to test whether smaller population sizes in *Vidua* as compared to *Lagonosticta* would result in faster coalescence and thus higher levels of gene tree concordance versus the alternative that incomplete reproductive isolation in *Vidua* would result in relatively less gene tree concordance (Fig. 7). A significant difference in the slope of the relationships between internode length and the proportion of supporting gene trees (Fig. 7B) supports the hypothesis that gene flow among *Vidua* species has increased the level of gene tree discordance in this genus, and that differences between the genera are not due solely to different timescales for speciation. Hybridization between *Vidua* species has been documented (Payne and Sorenson, 2004), and is likely facilitated by mate choice behaviors that are shaped by host imprinting (Payne et al., 2000), sympatry between ancestral and incipient species, and imperfect host fidelity by female parasites (Balakrishnan et al., 2009). An additional factor that may contribute to discordance among loci is the possibility that a given parasitic species derives from multiple independent host colonization events (in time and/or space) from different ancestral populations or species (Sefc et al., 2005), resulting in the admixture of alleles with disparate evolutionary histories.

It remains true, however, that rapid diversification at the base of the *Vidua* phylogeny undoubtedly contributes to generally lower gene tree concordance in this genus, raising the question of whether the history falls within the zone in which anomalous gene trees are expected. One approach to evaluating this is to assess the extent of conflict among gene trees. With long internodes, we expect most informative gene trees to be consistent with the species tree, whereas other trees are unresolved with respect to the node in question. In contrast, when there are true polytomies or a history of speciation events that falls near the edge of the “anomaly zone,” we

expect to see sets of gene trees supporting alternative relationships. This was clearly evident in the *Vidua* data set (Fig. 7), whereas sets of gene trees supporting possible alternative arrangements in *Lagonosticta* were small relative to the set of genes supporting the consensus tree (e.g., Fig. 5A).

Despite the complications discussed above and the likelihood of a more complex history of speciation in *Vidua*, our analyses suggest a single best hypothesis for relationships in *Vidua*, corresponding to the consensus of the generally concordant results among analyses (Fig. 4). This consensus phylogeny differs from the most likely mitochondrial gene tree (Fig. 1B), and provides a basis for evolutionary inferences. One interesting insight is provided by the recovery of a sister relationship between *V. hypocherina* and *V. macroura* (Fig. 4, node F), which are the only two *Vidua* species that are not known to mimic the vocalizations of their hosts (Payne, 1997; Fry and Keith, 2004). Together with the finding that paradise-whydahs (*V. interjecta*, *V. orientalis*, *V. paradisaea*, and *V. obtusa*) are sister to all other *Vidua* (Fig. 4, node E), this suggests a single origin followed by one loss of host song mimicry in the genus. Likewise, long tails appear to be the ancestral state for the genus, with a subsequent loss in indigobirds, which have the most elaborate vocal repertoires in the genus. The transition from long tails to more elaborate songs may be explained by the transference hypothesis, which posits a shift from more to less costly male display traits (Gilliard, 1956; Borgia, 1993; Kusmierski et al., 1993; Johnson, 1999).

4.5. Conclusions

ddRAD-seq and related GBS methods broadly sample the genome and produce datasets that are orders of magnitude larger than has been typical for molecular phylogenetic studies over the

past 20 years. In addition, the method requires no advance knowledge of genomic sequences for primer/probe design and so can be applied to almost any organism, although availability of a closely related reference genome aids in the clustering of homologous loci among taxa. In this study, we demonstrated that ddRAD-seq data provide various sources of informative characters for phylogenetic analyses (i.e., sequences, indels, presence-absence of loci, SNPs), thus allowing for the estimation of phylogenetic relationships using a variety of data sets, methods, and conceptual approaches. In general, phylogenetic analyses of these various character sets converged on similar results. Independent loci can be used to quantify support for alternative topologies, and discordance among loci and datasets may reveal neutral and/or non-neutral processes warranting further study. Although the feasibility of collecting and analyzing whole-genome sequence data in a phylogenetic context is improving (Jarvis et al., 2014), GBS methods offer a cost effective option for data generation in phylogenomic studies because a single lane of Illumina sequencing can now generate genotypic data for several thousand loci in hundreds of samples. The steadily increasing throughput and read length of next generation sequencing technology combined with the continued development of species tree methods, particularly those using SNPs, makes GBS methods a powerful and tractable option for data collection in modern phylogeographic and phylogenetic studies.

Acknowledgments

We thank Robert Payne (University of Michigan Museum of Zoology) for providing samples, and Christopher Schneider, Sean Mullen, Jeffrey Peters, and two anonymous reviewers for

comments on the manuscript. This study was funded by the National Science Foundation (grants DEB 0640759 and DEB 1011517).

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J Mol Biol* 215, 403-410.
- Ané, C., Larget, B., Baum, D.A., Smith, S.D., Rokas, A., 2007. Bayesian estimation of concordance among gene trees. *Mol Biol Evol* 24, 412-426.
- Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., . . . Johnson, E.A., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *Plos One* 3, e3376.
- Balakrishnan, C.N., Sefc, K.M., Sorenson, M.D., 2009. Incomplete reproductive isolation following host shift in brood parasitic indigobirds. *P R Soc B* 276, 219-228.
- Binladen, J., Gilbert, M.T.P., Bollback, J.P., Panitz, F., Bendixen, C., Nielsen, R., Willerslev, E., 2007. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *Plos One* 2.
- Borgia, G., 1993. The cost of display in the non-resource based mating system of the satin bowerbird. *Am Nat* 141, 729-743.
- Bouckaert, R., Heled, J., Kuhnert, D., Vaughan, T.G., Wu, C.H., Xie, D., . . . Drummond, A.J., 2013. BEAST2: a software platform for Bayesian evolutionary analyses. <http://beast2.cs.auckland.ac.nz>.
- Brito, P., Edwards, S.V., 2009. Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica* 135, 439-455.
- Brumfield, R.T., Beerli, P., Nickerson, D.A., Edwards, S.V., 2003. The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol Evol* 18, 249-256.
- Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N.A., RoyChoudhury, A., 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol Biol Evol* 29, 1917-1932.
- Buerkle, C.A., Gompert, Z., 2013. Population genomics based on low coverage sequencing: how low should we go? *Mol Ecol* 22, 3028-3035.
- Carstens, B., Lemmon, A.R., Lemmon, E.M., 2012. The promises and pitfalls of next-generation sequencing data in phylogeography. *Syst Biol* 61, 713-715.
- Crawford, N.G., Faircloth, B.C., McCormack, J.E., Brumfield, R.T., Winker, K., Glenn, T.C., 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biol Letters* 8, 783-786.
- Cruaud, A., Gautier, M., Galan, M., Foucaud, J., Saune, L., Genson, G., . . . Rasplus, J.Y., 2014. Empirical assessment of RAD sequencing for interspecific phylogeny. *Mol Biol Evol* 31, 1272-1274.
- DaCosta, J.M., Sorenson, M.D., 2014. Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *Plos One* 9, e106713.

- Davey, J.L., Blaxter, M.W., 2010. RADSeq: next-generation population genetics. *Brief Funct Genomics* 9, 416-423.
- Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., Blaxter, M.L., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* 12, 499-510.
- Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. *PLoS Genetics* 2, 762-768.
- Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29, 1969-1973.
- Eaton, D.A.R., Ree, R.H., 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants. *Syst Biol* 62, 689-706.
- Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32, 1792-1797.
- Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460-2461.
- Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63, 1-19.
- Edwards, S.V., Beerli, P., 2000. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* 54, 1839-1854.
- Edwards, S.V., Liu, L., Pearl, D.K., 2007. High-resolution species trees without concatenation. *Proc Natl Acad Sci U S A* 104, 5936-5941.
- Faircloth, B.C., Sorenson, L., Santini, F., Alfaro, M.E., 2013. A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). *Plos One* 8.
- Felsenstein, J., 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
- Fry, C.H., Keith, S., 2004. *The Birds of Africa, Volume VII: Sparrows to Buntings*. Princeton University Press, Princeton.
- Gilliard, E.T., 1956. Bower ornamentation versus plumage characters in bowerbirds. *Auk* 73, 450-451.
- Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59, 307-321.
- Harvey, M.G., Judy, C.D., Seeholzer, G.F., Maley, J.M., Graves, G.R., Brumfield, R.T., 2015. Similarity thresholds used in DNA sequence assembly from short reads can reduce the comparability of population histories across species. *PeerJ* 3, e895.
- Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22, 160-174.
- Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. *Mol Biol Evol* 27, 570-580.
- Hipp, A.L., Eaton, D.A.R., Cavender-Bares, J., Fitzek, E., Nipper, R., Manos, P.S., 2014. A framework phylogeny of the American oak clade based on sequenced RAD data. *Plos One* 9.
- Hovmoller, R., Knowles, L.L., Kubatko, L.S., 2013. Effects of missing data on species tree estimation under the coalescent. *Mol Phylogenet Evol* 69, 1057-1062.
- Huang, H., Knowles, L.L., In press. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. *Syst Biol*.

- Huang, H.T., Knowles, L.L., 2009. What is the danger of the anomaly zone for empirical phylogenetics? *Syst Biol* 58, 527-536.
- Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23, 254-267.
- Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., . . . Zhang, G.J., 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346, 1320-1331.
- Johnson, K.P., 1999. The evolution of bill coloration and plumage dimorphism supports the transference hypothesis in dabbling ducks. *Behav Ecol* 10, 63-67.
- Jones, J.C., Fan, S.H., Franchini, P., Scharl, M., Meyer, A., 2013. The evolutionary history of *Xiphophorus* fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. *Mol Ecol* 22, 2986-3001.
- Kelchner, S.A., Thomas, M.A., 2007. Model use in phylogenetics: nine key questions. *Trends Ecol Evol* 22, 87-94.
- Keller, I., Wagner, C.E., Greuter, L., Mwaiko, S., Selz, O.M., Sivasundar, A., . . . Seehausen, O., 2013. Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Mol Ecol* 22, 2848-2863.
- Knowles, L.L., Kubatko, L.S., 2010. *Estimating Species Trees: Practical and Theoretical Aspects*. Wiley-Blackwell, Hoboken, New Jersey.
- Kubatko, L.S., Chifman, J., 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30, 3317-3324.
- Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol* 56, 17-24.
- Kusmierski, R., Borgia, G., Crozier, R.H., Chan, B.H.Y., 1993. Molecular information on bowerbird phylogeny and the evolution of exaggerated male characteristics. *J Evol Biol* 6, 737-752.
- Larget, B.R., Kotha, S.K., Dewey, C.N., Ane, C., 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* 26, 2910-2911.
- Lemmon, A.R., Emme, S.A., Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst Biol* 61, 727-744.
- Librado, P., Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451-1452.
- Liu, L., 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* 24, 2542-2543.
- Liu, L., Xi, Z., Wu, S., Davis, C., Edwards, S.V., 2015. Estimating phylogenetic trees from genome-scale data. *arXiv[q-bio.PE]*, 1501.03578.
- Liu, L., Yu, L.L., 2011. Estimating species trees from unrooted gene trees. *Syst Biol* 60, 661-667.
- Liu, L., Yu, L.L., Pearl, D.K., Edwards, S.V., 2009. Estimating species phylogenies using coalescence times among sequences. *Syst Biol* 58, 468-477.
- Liu, L.A., Yu, L.L., Edwards, S.V., 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol Biol* 10.
- Maddison, W.P., 1997. Gene trees in species trees. *Syst Biol* 46, 523-536.
- Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., . . . Turner, D.J., 2010. Target-enrichment strategies for next-generation sequencing. *Nat Methods* 7, 111-118.

- McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C., Brumfield, R.T., 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenet Evol* 66, 526-538.
- Miller, M.R., Atwood, T.S., Eames, B.F., Eberhart, J.K., Yan, Y.L., Postlethwait, J.H., Johnson, E.A., 2007. RAD marker microarrays enable rapid mapping of zebrafish mutations. *Genome Biol* 8, R105.
- Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30, I541-I548.
- Nadeau, N.J., Martin, S.H., Kozak, K.M., Salazar, C., Dasmahapatra, K.K., Davey, J.W., . . . Jiggins, C.D., 2013. Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Mol Ecol* 22, 814-826.
- Narum, S.R., Buerkle, C.A., Davey, J.W., Miller, M.R., Hohenlohe, P.A., 2013. Genotyping-by-sequencing in ecological and conservation genomics. *Mol Ecol* 22, 2841-2847.
- Near, T.J., Benard, M.F., 2004. Rapid allopatric speciation in logperch darters (Percidae : Percina). *Evolution* 58, 2798-2808.
- Nichols, R., 2001. Gene trees and species trees are not the same. *Trends Ecol Evol* 16, 358-364.
- Page, R.D.M., Charleston, M.A., 1997. From gene to organismal phylogeny: reconciled trees and the gene tree species tree problem. *Mol Phylogenet Evol* 7, 231-240.
- Pamilo, P., Nei, M., 1988. Relationships between gene trees and species trees. *Mol Biol Evol* 5, 568-583.
- Payne, R.B., 1973. Behavior, mimetic songs and song dialects, and relationships of the parasitic indigobirds (*Vidua*) of Africa. *Ornithological Monographs* 11, 1-333.
- Payne, R.B., 1997. Field identification of the brood-parasitic whydahs *Vidua* and cuckoo finch *Anomalospiza imberbis*. *Bulletin of the African Bird Club* 4, 18-28.
- Payne, R.B., Payne, K., 1977. Social organization and mating success in local populations of village indigobirds, *Vidua chalybeata*. *Zeitschrift für Tierpsychologie* 45, 113-173.
- Payne, R.B., Payne, L.L., Woods, J.L., Sorenson, M.D., 2000. Imprinting and the origin of parasite-host species associations in brood-parasitic indigobirds, *Vidua chalybeata*. *Anim Behav* 59, 69-81.
- Payne, R.B., Sorenson, M.D., 2004. Behavioral and genetic identification of a hybrid *Vidua*: maternal origin and mate choice in a brood-parasitic finch. *Auk* 121, 156-161.
- Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., 2012. Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *Plos One* 7, e37135.
- R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rokas, A., Holland, P.W.H., 2000. Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol* 15, 454-459.
- Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572-1574.
- Rosenberg, N.A., Nordborg, M., 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet* 3, 380-390.
- Rosenberg, N.A., Tao, R., 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. *Syst Biol* 57, 131-140.

- Rubin, B.E.R., Ree, R.H., Moreau, C.S., 2012. Inferring phylogenies from RAD sequence data. *Plos One* 7.
- Sefc, K.M., Payne, R.B., Sorenson, M.D., 2005. Genetic continuity of brood-parasitic indigobird species. *Mol Ecol* 14, 1407-1419.
- Shaw, T.I., Ruan, Z., Glenn, T.C., Liu, L., 2013. STRAW: species tree analysis web server. *Nucleic Acids Res* 41, W238-W241.
- Snir, S., Rao, S., 2012. Quartet MaxCut: a fast algorithm for amalgamating quartet trees. *Mol Phylogenet Evol* 62, 1-8.
- Sorenson, M.D., Balakrishnan, C.N., Payne, R.B., 2004. Clade-limited colonization in brood parasitic finches (*Vidua* spp.). *Syst Biol* 53, 140-153.
- Sorenson, M.D., Payne, R.B., 2001. A single ancient origin of brood parasitism in African finches: implications for host-parasite coevolution. *Evolution* 55, 2550-2567.
- Sorenson, M.D., Sefc, K.M., Payne, R.B., 2003. Speciation by host switch in brood parasitic indigobirds. *Nature* 424, 928-931.
- Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688-2690.
- Swofford, D.L., 2002. *Phylogenetic Analysis Using Parsimony (*and Other methods)*. Sinauer Associates, Sunderland, MA.
- Than, C., Nakhleh, L., 2009. Species tree inference by minimizing deep coalescences. *Plos Comput Biol* 5.
- Than, C., Ruths, D., Nakhleh, L., 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9.
- Than, C.V., Rosenberg, N.A., 2011. Consistency properties of species tree inference by minimizing deep coalescences. *J Comput Biol* 18, 1-15.
- Wagner, C.E., Keller, I., Wittwer, S., Selz, O.M., Mwaiko, S., Greuter, L., . . . Seehausen, O., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol Ecol* 22, 787-798.
- Wu, Y.F., 2012. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution* 66, 763-775.

Table 1. Specimens used in the collection of ddRAD-seq data.

Family	Genus	species	English name	Locality (or captive)	Sample ID	Museum† accession no.	mtDNA accession no.	SRA accession no.
Estrildidae	<i>Clytospiza</i>	<i>monteiri</i>	Brown Twinspot	Cameroon	MDS054*	MCZ 364058	AY324242; AY323540	SRX1050691
	<i>Lagonosticta</i>	<i>senegala rendalli</i>	Red-billed Firefinch	Malawi	A056	UMMZ 231354	AY324238; AY323536	SRX1050675
		<i>s. rhodopsis</i>		Cameroon	A167	UMMZ 232449	AY324239; AY323537	SRX1050682
		<i>rufopicta</i>	Bar-breasted Firefinch	Cameroon	A204	UMMZ 232441	AY324240; AY323538	SRX1050685
		<i>larvata</i>	Black-faced Firefinch	Cameroon	A145	UMMZ 232458	AY324233; AY323531	SRX1050680
		<i>rara</i>	Black-bellied Firefinch	Cameroon	A248	UMMZ 232452	AY324232; AY323530	SRX1050686
		<i>rubricata rubricata</i>	African Firefinch	South Africa	SAR6894	UWBM 53053	AY324235; AY323533	SRX1050694
		<i>r. congica</i>		Cameroon	MDS065*	MCZ 364059	AY324234; AY323532	SRX1050692
		<i>rhodopareia</i>	Jameson's Firefinch	Malawi	A107	UMMZ 231405	AY324236; AY323534	SRX1050679
		<i>sanguinodorsalis</i>	Rock Firefinch	Nigeria	A328	UMMZ 233841	KT204257; KT204258	SRX1050687
Viduidae	<i>Anomalospiza</i>	<i>imberbis</i>	Cuckoo-Finch	Mali	A1794	UMMZ 236335	AY324265; AY323567	SRX1050684
	<i>Vidua</i>	<i>macroura macroura</i>	Pin-tailed Whydah	Cameroon	A147	UMMZ 232524	AY324261; AY323563	SRX1050681
		<i>m. arenosa</i>		Malawi	A089	UMMZ 231387	AF407109; AF407023	SRX1050678
		<i>orientalis</i>	Sahel Paradise-Whydah	captive	90.o53	UMMZ 231660	AY324264; AY323566	SRX1050674
		<i>interjecta</i>	Exclamatory Paradise-Whydah	captive	NKK511	UMMZ 228179	AY324262; AY323564	SRX1050693
		<i>obtusata</i>	Broad-tailed Paradise-Whydah	Malawi	A082	UMMZ 231380	AY324263; AY323565	SRX1050677
		<i>paradisaea</i>	Long-tailed Paradise-Whydah	Malawi	A081	UMMZ 231379	AF407108; AF407022	SRX1050676
		<i>hypochoerina</i>	Steel-blue Whydah	captive	90.o31	UMMZ 231669	AF407107; AF407021	SRX1050671
		<i>fischeri</i>	Straw-tailed Whydah	captive	90.o50	UMMZ 231655	AY322834; AY323561	SRX1050673
		<i>regia</i>	Shaft-tailed Whydah	captive	90.o38	UMMZ 231658	AY322836; AY323562	SRX1050672

<i>chalybeata neumanni</i>	Village Indigobird	Cameroon	A177*	UMMZ 232518	AF090341	SRX1050683
<i>c. amauropteryx</i>		Tanzania	JMD1285*	MCZ 364061	AY322747	SRX1050689
<i>raricola</i>	Jambandu Indigobird	Cameroon	MDS004*	UMMZ 236433	AY322796; AY323560	SRX1050690
<i>camerunensis</i>	Cameroon Indigobird	Cameroon	CNB018*	UMMZ 239680	AY322644	SRX484784
<i>purpurascens</i>	Purple Indigobird	Tanzania	JMD1249*	MCZ 364060	AY322693; AY323559	SRX1050688

*Due to the availability of existing data, a different conspecific sample was used in the analysis of mtDNA sequences.

† UMMZ: University of Michigan Museum of Zoology; MZC: Harvard University Museum of Comparative Zoology.

Table 2. Summary statistics for the *Lagonosticta* and *Vidua* ddRAD-seq datasets.

	<i>Lagonosticta</i>	<i>Vidua</i>
Number of samples	10	15
Reads per sample (avg $\pm$ sd)	994,064 $\pm$ 125,299	889,334 $\pm$ 104,578
Putative loci (≥ 5 reads) per sample (avg $\pm$ sd)	8,969 $\pm$ 443	8,901 $\pm$ 359
Passed Loci*	2,860	3,100
Loci with ≥ 1 biallelic SNP	2,856	3,098
Median depth per sample per locus (avg $\pm$ sd)	86.2 $\pm$ 95.7	80.9 $\pm$ 75.9
Alignment length (bp) (avg $\pm$ sd)	88.1 $\pm$ 16.5	89.2 $\pm$ 15.9
SNPs per locus (avg $\pm$ sd)	6.4 $\pm$ 3.6	5.9 $\pm$ 3.2
Haplotypes per locus (avg $\pm$ sd)	6.4 $\pm$ 2.6	5.4 $\pm$ 2.4
Indels per locus (avg $\pm$ sd)	0.6 $\pm$ 0.9	0.5 $\pm$ 0.8
Present-Absent Loci**	10,904	10,212

* All samples had either good or low depth genotypes, but not more than two low depth samples per locus.

** All samples had either zero or ≥ 10 reads.

Table 3. Summary statistics for *Lagonosticta* and *Vidua* character-based datasets. Consistency index (CI), homoplasy index (HI), and retention index (RI) statistics were calculated using consensus topologies for each genus (see Fig. 4).

Dataset	Data type	No. characters	No. variable characters	No. parsimony inf. char.	CI	HI	RI
<i>Lagonosticta</i>	mtDNA concatenated sequences	2186	515	299	0.66	0.34	0.54
	ddRAD concatenated sequences	251873	15848	6061	0.90	0.10	0.84
	concatenated indels	1599	1561	658	0.88	0.12	0.84
	concatenated seq+indel	253472	17409	6719	0.89	0.11	0.84
	SNPs	2856	2856	1033	0.90	0.10	0.85
	locus presence-absence	10904	8786	3342	0.83	0.17	0.73
<i>Vidua</i>	mtDNA concatenated sequences	2186	635	341	0.70	0.30	0.72
	ddRAD concatenated sequences	276476	16421	4385	0.87	0.13	0.80
	concatenated indels	1475	1408	387	0.86	0.15	0.80
	concatenated seq+indel	277951	17829	4772	0.87	0.13	0.80
	SNPs	3098	3098	772	0.88	0.12	0.81
	locus presence-absence	10212	7928	2457	0.82	0.18	0.75

Figure Captions

Fig. 1. Mitochondrial DNA phylogenies for (A) *Lagonosticta* and (B) *Vidua*. The topologies illustrated are from a Bayesian analysis in BEAST; topologies from maximum likelihood (ML) analyses were the same, and topologies from maximum parsimony (MP) differed only at poorly supported nodes. Pie charts at nodes summarize clade support in MP, ML, and Bayesian inference (BI) analyses.

Fig. 2. Relationship between the total number of sequence reads assigned to individual samples and (A) the number of loci recovered, and (B) the median per sample depth of loci in the final dataset. Neither relationship was significant (linear regression, $P > 0.05$).

Fig. 3. Correlation between the number of shared ddRAD loci and genome-wide genetic divergence. Genetic divergence was calculated using 1,012 ddRAD loci recovered in all samples in both data sets. Number of shared loci was calculated from (A) a pool of loci ($n = 14,492$) that were recovered with a depth of zero or ≥ 10 reads in all samples, or (B) all loci with ≥ 5 reads in the two focal samples for each pairwise comparison.

Fig. 4. Results of phylogenetic analyses for *Lagonosticta* (left) and *Vidua* (right) using ddRAD-seq data. The cladograms represent the consensus topology among all analyses for each genus. The tables below provide details on the input data for each method and support values for each node in the respective cladograms. Support values derived using non-parametric bootstrapping are on a scale of 0-100, and those from Bayesian and concordance methods are on a scale of 0-1.0. Nodes with maximum support values (i.e., 100 or 1.0) are shaded green. Empty cells denote nodes that were not recovered in a given analysis. In most analyses, the outgroup was defined *a priori*, and thus support values for the basal node of the genus (node A) cannot be estimated (n/a in table).

Fig. 5. Gene tree support for consensus and alternative relationships for: (A) *Lagonosticta rara* and *L. larvata*, (B) *Vidua hypocherina* and *V. macroura*, and (C) *V. regia*, and *V. fischeri*. Values next to each clade correspond to the number of loci with ML (above) or MP (below) support for the clade. Paradise-whydahs: *V. interjecta*, *V. orientalis*, *V. paradisaea*, *V. obtusa*; Indigobirds: *V. chalybeata*, *V. camerunensis*, *V. raricola*, *V. purpurascens*.

Fig. 6. Phylogenetic networks for (A) *Lagonosticta* and (B) *Vidua* taxa based on ddRAD-seq data. Both networks were constructed with the Net Neighbor algorithm, and are presented at the same scale. Inset boxes display portions of each network at a higher magnification. Lower panels display cloudograms representing the posterior distributions of trees recovered in SNAPP analyses. (C) Results from the *Lagonosticta* dataset ($n = 2856$ biallelic SNPs) show perfect topological concordance among trees (i.e., 1.0 posterior distribution for all nodes), whereas (D) results from the *Vidua* dataset ($n = 3098$ biallelic SNPs) illustrate conflicting topologies in the posterior.

Fig. 7. Estimates of gene tree concordance in the *Lagonosticta* and *Vidua* datasets. (A) Topology with branch lengths for both *Lagonosticta* and *Vidua* taxa based on a BEAST analysis of 1,012 ddRAD loci shared among all taxa in both datasets. Numbers at nodes represent the proportion of

loci in the *Lagonosticta* ($n = 2,860$ loci) or *Vidua* ($n = 3,100$ loci) datasets with ML support for the node. (B) For each genus, the correlation between length of the subtending internode and the proportion of ML gene trees with support for the node. Solid lines show separate linear regressions forced through the origin for each genus. Similar results were obtained when using maximum parsimony results. (C) Average ($\pm$ standard error) nucleotide diversity in *Lagonosticta* and *Vidua* samples.

Figure1

Figure2

A

B

Figure3

A

Number of shared ddRAD loci
(0 or 10+ reads in all samples)

B

Number of shared ddRAD loci
(5+ reads in focal samples)

Figure4

Lagonosticta *Vidua* *Lagonosticta**Vidua*

Data Type / Method	Program	A	B	C	D	E	F	G	H	A	B	C	D	E	F	G	H	I	J	K	L	M
Concatenated sequences		251873 characters								276476 characters												
Parsimony	PAUP*	n/a								n/a					97			87		67	73	
Likelihood	RAXML	n/a								n/a				77	99			95		81	68	
Bayesian	BEAST	0.92																				
Concatenated indels		1599 characters								1475 characters												
Parsimony	PAUP*	n/a								n/a		91	67	69	43			67			54	91
Likelihood	RAXML	n/a								n/a	99	91	84	40	57		96	77	97		60	87
Bayesian	BEAST													0.50								
Concatenated seq+indels		253472 characters								277644 characters												
Parsimony	PAUP*	n/a								n/a					97			93		68	81	
Likelihood	RAXML	n/a								n/a					98			95		79	73	
Bayesian	BEAST																					
Presence-Absence		10904 characters								10212 characters												
Parsimony	PAUP*	n/a								n/a				99				58		68		
Likelihood	RAXML	n/a								n/a				95	95			73		74		
Bayesian	BEAST																					
Gene trees		2860 gene trees								3100 gene trees												
Internode distances	PhyML+NJst	n/a			89	89	86			n/a		33	58	46	39	99		43	76		23	43
Maximum Pseudo-likelihood	PhyML+MPest	n/a		99	78	83	77			n/a		31	46	45	27	96	99		69	20	22	33
Coalescent ranks	PhyML+STAR	n/a			89	88	86			n/a		33	56	47	38	99		43	77		24	44
Quartet agreement	PhyML+ASTRAL	n/a			88	85	85			n/a		27	52	47	37	99		39	77		21	41
Minimize deep coalescences	PhyML+PhyloNet	n/a								n/a											72	88
Multispecies coalescent	PhyML+STELLS	n/a								n/a											73	86
SNPs		2856 biallelic SNPs								3098 biallelic SNPs												
SNP coalescence	SNAPP											0.89	0.93	0.99	0.59			0.70			0.27	0.32
Quartet agreement	SVDquartets	n/a								n/a		66	99	99	74			56				97
Gene tree distributions		2860 distributions								3100 distributions												
Concordance tree	MrBayes+BUCKy	n/a	0.52	0.37	0.36	0.34	0.35	0.51	0.70	n/a	0.18	0.11	0.12			0.20	0.11	0.07	0.05			0.09
Population tree	MrBayes+BUCKy	n/a	0.67	0.57	0.51	0.51	0.48	0.62	0.76	n/a	0.49	0.35	0.36	0.35	0.34	0.45	0.45	0.34	0.39	0.34		0.36

Figure5

Figure6

A

B

C

D

Figure7

A

B

C