Syst. Biol. 66(3):399–412, 2017

DOI:10.1093/sysbio/syw092
Advance Access publication October 14, 2016

Misconceptions on Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants

DEREN A. R. EATON*, ELIZABETH L. SPRIGGS, BRIAN PARK, AND MICHAEL J. DONOGHUE

Department of Ecology and Evolutionary Biology, Yale University, PO Box 208106, New Haven, CT, 06520, USA;
**Correspondence to be sent to: Department of Ecology and Evolutionary Biology, Yale University, PO Box 208106, New Haven, CT, 06520, USA; E-mail: deren.eaton@yale.edu*

Received 29 March 2016; reviews returned 12 June 2016; accepted 10 October 2016
Associate Editor: Laura Kubtoko

Abstract.—Restriction-site associated DNA (RAD) sequencing and related methods rely on the conservation of enzyme recognition sites to isolate homologous DNA fragments for sequencing, with the consequence that mutations disrupting these sites lead to missing information. There is thus a clear expectation for how missing data should be distributed, with fewer loci recovered between more distantly related samples. This observation has led to a related expectation: that RAD-seq data are insufficiently informative for resolving deeper scale phylogenetic relationships. Here we investigate the relationship between missing information among samples at the tips of a tree and information at edges within it. We re-analyze and review the distribution of missing data across ten RAD-seq data sets and carry out simulations to determine expected patterns of missing information. We also present new empirical results for the angiosperm clade *Viburnum* (Adoxaceae, with a crown age >50 Ma) for which we examine phylogenetic information at different depths in the tree and with varied sequencing effort. The total number of loci, the proportion that are shared, and phylogenetic informativeness varied dramatically across the examined RAD-seq data sets. Insufficient or uneven sequencing coverage accounted for similar proportions of missing data as dropout from mutation-disruption. Simulations reveal that mutation-disruption, which results in phylogenetically distributed missing data, can be distinguished from the more stochastic patterns of missing data caused by low sequencing coverage. In *Viburnum*, doubling sequencing coverage nearly doubled the number of parsimony informative sites, and increased by >10X the number of loci with data shared across >40 taxa. Our analysis leads to a set of practical recommendations for maximizing phylogenetic information in RAD-seq studies. [hierarchical redundancy; phylogenetic informativeness; quartet informativeness; Restriction-site associated DNA (RAD) sequencing; sequencing coverage; *Viburnum*.]

Restriction-site associated DNA sequencing (RAD-seq) is a method for subsampling the genome to concentrate sequencing efforts as a way to efficiently attain high coverage data across many individuals for comparative genomics (Miller et al. 2007; Baird et al. 2008). Because it relies on the use of restriction enzymes to digest genomic DNA and isolate homologous fragments, the conservation of enzyme recognition sites is critical for recovering shared data among sampled individuals. For this reason, RAD-seq methods are expected to yield fewer shared data between more highly divergent taxa where the opportunity for mutations to disrupt shared restriction sites has been greater. This has led to a common conception that RAD-seq data are not applicable to deep-scale phylogenetic analyses.

A general motivation for combining multiple sequence markers (loci) into a single joint phylogenetic analysis is the expectation that a combination of markers contains more overall information than could be attained by examining each individually. Concatenation (de Queiroz and Gatesy 2007), binning (Bayzid et al. 2015), and supertree methods (Sanderson et al. 1998) represent just several of the many ways in which phylogenetic information can be combined. This logic extends similarly to the case of RAD-seq, in which there are often few loci that contain information for all taxa in a data set. When many loci with variable but overlapping taxon sets are combined there can be thousands of loci that inform any given split in a tree. A great concern, however, for applying RAD-seq data in this framework has to do with the distribution of missing data. If it is highly nonrandom, there may be too few markers with sufficiently overlapping taxon sets to infer certain splits in a tree — particularly those deeper in time.

The total proportion of missing data in supermatrices composed of RAD sequences has received considerable attention (Rubin et al. 2012; Cariou et al. 2013; Hipp et al. 2014; Huang and Knowles 2016; Leaché et al. 2015; Ree and Hipp 2015). However, there has been much less investigation of the underlying cause of missing data, and how different sources may contribute to its distribution. In many cases, it has been assumed that most missing information comes from a single source: allelic drop-out caused by mutations. However, there are multiple sources that can give rise to missing RAD-seq data: (i) variable recovery during library preparation due to technical issues related to digestion, size selection, and DNA quality (Escudero et al. 2014); (ii) bioinformatic errors in identifying homology (Eaton 2014); (iii) mutation-generation of new fragments that are not shared by common descent among all sampled lineages; (iv) mutation-disruption of ancestral fragments such that they are no longer shared by all descendant lineages (Rubin et al. 2012; Cariou et al. 2013); and (5) insufficient or uneven amplification and sequencing coverage (Huang and Knowles 2016).

The relationship between missing data at the tips of a tree and the loss of phylogenetic information across its edges (splits) depends on a number of factors, including topology and branch lengths, as well as the source of missing data, which determines its distribution.

For example, data loss from mutation-disruption will be hierarchical, with a phylogenetic pattern matching that of mutations; whereas low (or uneven) sequencing

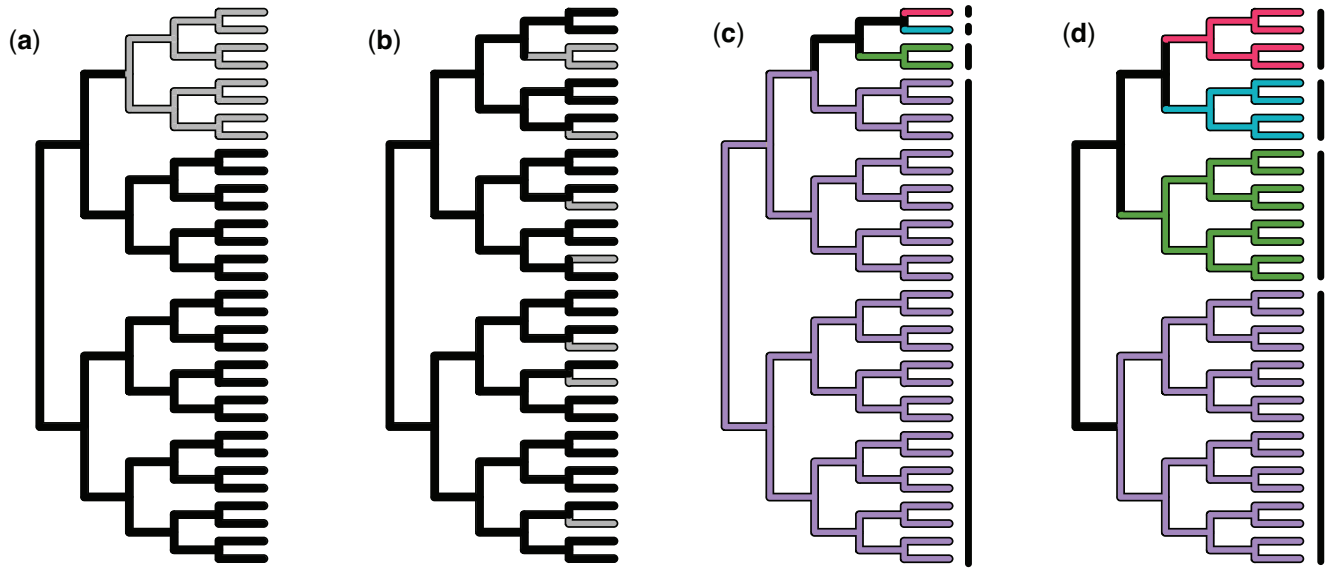

FIGURE 1. Expected distribution of missing information (grey) from a) mutation-disruption and b) low sequencing coverage. c) Quartet informativeness is a multi-locus measure of phylogenetic information content that measures the number of markers with sufficient taxon (tip) sampling to be potentially informative about a specific split/bipartition (black) in a tree. For a marker to be informative about a given split at least one taxon from each of the four connected edges (colored clades or vertical bars) must have data for the marker. d) Splits that have a greater number of descendant taxa that can be sampled from each connected edge have greater hierarchical redundancy.

coverage will generally produce a more stochastic distribution of missing sequences that is little if at all influenced by the relationships among samples (Fig. 1a-b). Examined phylogenetically, the first type of missing data perpetuates information loss deeper into the tree, because entire clades are likely to be missing information, whereas missing data resulting from low sequencing coverage will tend to be stochastically distributed toward the tips.

Here, we examine patterns of shared RAD-seq data observed in 10 empirical data sets to identify the most likely sources of missing data, and to investigate its effect on phylogenetic information content. Nine of these data sets are from published studies of a variety of organisms, and one is from a new study, presented here, of the angiosperm clade *Viburnum* (Adoxaceae). For the *Viburnum* data set, we examine the distribution of shared data across both shallow and deep splits within the tree, and investigate how sequencing coverage influences phylogenetic informativeness. These results are compared with simulations to explore expected patterns of data sharing when missing data arise from different sources. Together, our analyses yield a set of practical guidelines for improving the effectiveness of phylogenetic studies using RAD-seq data.

MATERIALS AND METHODS

Simulations

We developed a Python program, *simrrls* (http://github.com/dereneaton/simrrls), to simulate RADseq-like sequence data based on coalescent simulations performed with the Python package *egglib* (Mita and Siol 2012). To model realistic levels of mutation-disruption, haplotypes are dropped if (i) a mutation arises in the restriction recognition site (or sites if two cutters are used) relative to the ancestral sequence, or (ii) a mutation gives rise to a new restriction recognition site within an existing fragment and the resulting shortened fragment falls outside of a defined size selection window (Supplementary Figure S1 in Supplementary Material, available on Dryad at http://dx.doi.org/10.5061/dryad.g549v). To model variable sequencing coverage the number of reads sampled from each haplotype is drawn from a normal distribution. Mutations occur under the Jukes–Cantor model of sequence evolution with equal starting base frequencies.

We simulated 1000 diploid loci on three different topologies of equal tree lengths (6 coalescent units); all lineages had a constant population size of $1e^6$, and per-site mutation rate $1e^{-9}$ ($\theta = 0.04$). This yielded highly divergent loci (mean = 20 single nucleotide polymorphisms, SNPs), which were chosen in order to model the extreme at which mutation-disruption could give rise to missing data, but still fall within the range where homology could be identified in bioinformatic analyses. Data were simulated with a single cutter (e.g., RAD) or with two cutters (e.g., ddRAD; (Peterson et al. 2012)). The first used an 8-bp cutter, whereas the second used an 8-bp cutter in addition to a 4-bp cutter. When allowing for mutation-disruption, haplotypes were discarded if a RAD fragment was digested to less than 300 bp in length, or a mutation occurred within the cut site. When allowing for low sequencing depth, reads were sampled from each haplotype with mean = 2 and sd = 5, or mean = 5 and sd = 5.

To compare results with no missing sequences against those with only mutation-disruption, or only low sequencing coverage, data were simulated under three combinations of parameter settings, and on three different topologies (Fig. S2 available on Dryad). Each topology had an identical tree length but differed in shape. This included a completely balanced topology with 64 tips, a completely imbalanced topology with 64 tips, and an empirical topology inferred for the plant clade *Viburnum* (65 tips; empirical data set 1, described below).

We define a metric, *quartet informativeness*, as the number of loci in a multi-locus data set that have sufficient taxon sampling to be potentially informative about a given split (bipartition) in an unrooted tree (Fig. 1c and d). We used this metric to compare information loss at different edge depths for both simulated and empirical data sets that vary in their primary source of missing data, and thus the distribution of missing data across different edge depths in their trees.

Empirical Data Sets

We examined 10 RAD-seq data sets for which the raw data were accessible, spanning a range of crown ages, sizes, and library preparation methods (Table S1 available on Dryad). These included the following: 65 samples of *Viburnum* (generated for this study, described below); 56 samples of *Heliconius* (Nadeau et al. 2013); 36 samples of *Quercus* (Eaton et al. 2015); 32 samples of *Ohomopterus* (Takahashi et al. 2014); 31 samples of *Danio* (McCluskey and Postlethwait 2015); 13 samples of *Pedicularis* (Eaton and Ree 2013); 64 samples of *Orestias* (Takahashi and Moreno 2015); 74 samples of *Phrynosomatidae* (Leaché et al. 2015); 13 samples of Barnacles (Herrera et al. 2015); and 25 samples of Finches (DaCosta and Sorenson 2016). To minimize the influence of bioinformatic variables on our results, all data sets were assembled *de novo* under the same general parameter settings in the program *pyrad* v.3.0.66 (Eaton 2014); see supplemental notebooks (Table S1 available on Dryad). This software uses an alignment clustering algorithm (https://github.com/torognes/vsearch) that allows for insertion–deletion polymorphisms to minimize clustering bias between close versus distant relatives. Reads were clustered at 85% sequence similarity and we required a minimum depth of coverage of six for clusters to be retained in the assembly. Each data set was assembled into two final alignments: a "min2" data set that included all loci shared by at least two samples (non-singletons), and a "min4" data set that included all loci shared by at least four samples. Because data must be shared across four tips to attain phylogenetic information for an unrooted tree, the latter data set represents the maximum phylogenetic information in the data. The min4 assembly was used to infer a tree for each data set using *RAxML* v.8.1.16 (Stamatakis 2014) under the GTR+Γ nucleotide substitution model.

To investigate sources of missing data, we used phylogenetic generalized least squares (*PGLS;* Symonds and Blomberg 2014) to fit a regression between the observed number of loci shared among sets of four individuals (quartets), their amount of raw sequence data, and their phylogenetic distance. Phylogenetic distance was measured as the sum of GTR+Γ branch lengths joining quartet samples, and raw sequence data was measured as the log median number of reads among quartet samples. All values were mean standardized prior to analysis.

Because our data represent quartets of species, rather than individual species, we implemented a modified *PGLS* method. Many sets of quartets share taxa in common, or span the same internal edges of the tree, and thus are expected to have experienced the same mutation-disruption of RAD-seq loci. To model their nonindependence, we constructed a variance–covariance matrix measuring shared branch lengths among all sets of quartet samples. Models were fit using the *gls* function in the R package *nlme* (Pinheiro et al. 2016) with an imposed correlation structure derived from the quartet covariance matrix. Following (Symonds and Blomberg 2014) we estimated phylogenetic signal (λ) for each data set and used this value to transform off-diagonal elements of the covariance matrix such that if no phylogenetic signal is present the phylogenetic correction has no effect. For large data sets with many taxa the covariance matrices were too large to analyze in a single analysis, and so we used a subsampling approach to randomly sample 200 data points, and repeated this 100 times to attain a distribution of model estimates for each data set.

Viburnum Phylogeny

The flowering plant clade *Viburnum* (Adoxaceae) includes approximately 165 species of shrubs and small trees and has an estimated crown age of 50–60 Ma (Spriggs et al. 2015). Currently, the best estimate of *Viburnum* phylogeny is based on whole chloroplast genomes for 22 species representing all major clades, plus a 10-gene data set for 138 species including 9 cpDNA markers and nuclear ribosomal ITS sequences (Clement et al. 2014; Spriggs et al. 2015). Although many relationships are now well supported, several of the deepest splits in the tree are subtended by very short branches and receive low support, and several recent and rapid radiations remain unresolved due to insufficient variation.

A RAD-seq library was prepared for 95 individuals representing 65 species from all major clades within *Viburnum*. For the present study, we included only one representative per species by selecting the sample with the most data when replicates were present (Table S2). Genomic DNA was extracted from leaf tissues preserved in silica or from herbarium specimens, using Qiagen DNEasy kits (Valencia, CA). Multiple extractions from the same individual were sometimes required to obtain 1.0 μg of high molecular weight

DNA. RAD libraries were prepared by Floragenex inc. (http://floragenex.com) using a 6-bp restriction enzyme (PstI) for digestion followed by sonication. To minimize the influence of technical effects (e.g., stochasticity in amplification, sequencing, and PCR duplicates) on our results a second library was also prepared following the same protocol, and with a separate amplification. The two libraries were sequenced on separate lanes of an Illumina HiSeq 2000 at the University of Oregon GC3F facility (http://gc3f.uoregon.edu).

To evaluate the influence of sequencing effort on our results we first analyzed the data using only one lane of sequence data, and then pooled replicates from both lanes. Both data sets were assembled in *pyrad* with the same parameters described above. We refer to these as the "half" and "full" data sets. For each, the "min4" assembly was used to infer a maximum likelihood phylogeny in *RAxML* with 100 nonparametric bootstrap replicates. In addition, we inferred an unrooted species tree (i.e., consistent under the multi-species coalescent) with the program *tetrad* v.0.4.0 (http://github.com/dereneaton/ipyrad). This software implements the *SVDQuartets* algorithm (Chifman and Kubatko 2014) to infer quartet trees using the full SNP alignment for each sampled quartet of taxa. We inferred all 677,040 possible quartets for 65 taxa. The quartet trees are then joined into a supertree using the quartet-maxcut algorithm implemented by *wQMC* (Avni et al. 2015). This approach is well suited for RAD-seq data because it maximizes phylogenetic information available for each quartet of sampled taxa regardless of missing data among other taxa. We ran 100 nonparametric bootstrap replicates where each replicate resamples RAD-seq loci with replacement. For each sampled quartet, in each replicate, a single SNP is randomly sampled from the four-taxon alignment at each locus for which they share data. Heterozygous sites are randomly resolved in each replicate, since individually sampled SNPs are putatively unlinked. We report the tree as an extended majority-rule consensus with bootstrap supports. For visualization, all trees were rooted along the *V. clemensiae* branch based on previous analyses using outgroups (Clement et al. 2014).

A detailed description of an expanded *Viburnum* RAD phylogeny, its correspondence with cpDNA trees, and its implications for character evolution and biogeography will be presented elsewhere. Here we focus our discussion on how sequencing coverage impacted the distribution of phylogenetic information at different depths in the tree. Specifically, we focused on two regions of the phylogeny that have been especially difficult to resolve — one near the base of the tree and one near the tips. Near the base of the tree we used one representative of each of seven well-supported major clades to specifically test the relationship of *V. taiwanianum* (of the *Urceolata* clade) to *V. amplificatum* plus *V. lutescens* (as supported in cpDNA trees) versus to *V. lantanoides* (of the *Pseudotinus* clade, as suggested by morphological data). Near the tips we analyzed relationships among members of the *Oreinodentinus* clade, which here includes four species representing the rapid and previously weakly resolved radiation of *Viburnum* in neotropical cloud forests.

For both focused phylogenetic analyses we inferred a primary concordance tree with the program *BUCKy* (Larget et al. 2010) to examine the proportion of loci that recover each split while accounting for gene-tree estimation error and heterogeneity. Loci were only included in the analysis if they were sampled for all taxa chosen for the focused analysis. From these we sub-selected only loci that contained at least 1 parsimony informative site. A distribution of gene trees was estimated for each locus in *MrBayes* v.3.2.1 (Ronquist and Huelsenbeck 2003) under a GTR+Γ model from 4 MCMCMC chains run for 1M generations sampling every 1000 steps. These distributions were analyzed in *BUCKy*, combining 4 independent runs each with 4 MCMC chains run 1M generations and repeated at three different values for α (0.1, 1, 10), the prior expectation on the number of distinct trees.

Results

Simulations

Library types.—In simulations, as expected, allelic dropout caused by mutation-disruption of enzyme recognition sites occurred more frequently when the enzyme recognition site was longer. In contrast, allelic dropout caused by mutations giving rise to new cut sites within existing fragments occurred more frequently when enzyme recognition sites were shorter (Fig. S3a available on Dryad). Variation in the range of the size-selected window had little effect on these results. When both forms of mutation-disruption occur together the effect of cutter length is effectively nullified (Fig. S3b available on Dryad). A comparison of single-digest versus double-digest methods, however, shows that adding a second independent cutter approximately doubles the rate of data loss (Fig. S3b available on Dryad), consistent with predictions from *in silico* studies (Collins and Hrbek 2015) that have shown more rapid mutation-disruption in double digest data.

Quartet informativeness.—To investigate the impact of missing RAD-seq data on phylogenetic information loss we examined the distribution of RAD loci that were originally present in the common ancestor of a 64-tip tree and measured how many loci remain quartet informative about each edge after some proportion of data is removed. If a locus lacks data for any one of the four terminal edges of a quartet then it is not informative about that quartet. The two sources of data loss examined, mutation-disruption and low sequencing coverage, lead to different distributions of phylogenetic information, particularly with respect to quartet information.

Hierarchical redundancy.—On the balanced topology all edges are equal in length and thus mutation-disruption

is equally likely to occur along any edge. If mutation-disruption occurs for a given locus along a "tip" edge of the tree, then the locus no longer contains information to form any quartet that includes that tip taxon. However, the locus remains informative about all other quartets formed by all other sets of four taxa in the data set. The same is not true when mutation-disruption occurs along deeper edges. In that case, data are lost for all taxa subtending the edge (e.g., Fig. 1a), which disrupts all quartets that include a tip from the descendant clade. This means that mutation-disruption decreases quartet information the most for splits near the tips of the tree, since these edges can be affected by any mutations that occur between the tips and the root of the tree. The counter-intuitive result is that mutation-disruption actually has the least effect on information loss at deeper edges, since these are least affected by mutations occurring elsewhere in the tree. We refer to this property of the data as *hierarchical redundancy* (e.g., Fig. 1c and d), wherein the amount of information at a given split in the tree is affected by its hierarchical placement in the tree. A split with high hierarchical redundancy would have multiple individuals sampled from each of its edges, such that even if data were lost for some individuals from each edge, there is still likely to be at least one individual from each edge that retains data.

Mutation-disruption.—Of the original 1000 simulated loci, approximately 900 remained quartet informative across the deepest splits of the balanced tree following mutation-disruption, and this decreased to approximately 800 for edges near the tips (Fig. 2e). The rate of loss was approximately doubled for double-digest data (Fig. S2a available on Dryad). The effect of tree shape, and correspondingly edge length, is made clear when these results are compared with those from the imbalanced topology (Fig. 2f). On this tree, the shallowest edges have similar numbers of quartet informative loci as in the balanced tree, but the deepest edges have the least phylogenetic information. This is because the hierarchical redundancy at the shallowest edges is similar for both trees, but highly dissimilar at their deepest edges (Fig. 2a and d). For a locus to be quartet informative at the deepest split in the imbalanced tree would require no disrupting mutations to occur along any of the three longest edges of the tree, while only one edge of the quartet has high hierarchical redundancy (Fig. 2d). The structure of the balanced tree, in contrast, has redundancy of sampling, on average, across all edges (Fig. 2c).

For practical purposes, these results suggest that sampling a more balanced topology can increase phylogenetic information at deeper scales, and, in particular, that outgroup sampling would be more effective if multiple representatives from a sister clade were sampled as opposed to a grade of individual taxa representing multiple lineages. This would have the effect of breaking up longer branches on which mutation-disruption could occur, and increasing hierarchical redundancy, such that data lost for some members of a clade would more often still be represented by at least one other representative of that clade.

Sequencing coverage.—When data loss occurs more randomly, as in the case of low sequencing coverage, its effect on phylogenetic information loss is quite different. In Figure 2g-h, we show an example where approximately 50% of the data was randomly removed from each taxon. On the balanced topology, the quartet informativeness of these data is nearly completely recovered only 2–3 edges deeper into the tree (Fig. 2g). This is once again because of hierarchical redundancy; data lost for any one tip at a given locus can often be sampled from multiple other relatives as a substitute. Again, this effect is clear when compared against the imbalanced topology that lacks hierarchical redundancy (every split is connected to at least two tips). In that case, deeper nodes in the tree do not have more phylogenetic information because all internal edges depend on sampling a specific taxon — they do not increase in hierarchical redundancy (Fig. 2h). Thus, the expectation from low (or uneven) sequencing coverage is that large amounts of missing data will decrease quartet information near the tips in a balanced tree, or equally across all edges for an imbalanced tree, but will never lead to comparatively more information loss at edges that are deeper in time.

These results can be leveraged to maximize phylogenetic information at deeper depths of a tree. When taxa are sampled in a way that maximizes tree balance, deeper edges will have greater hierarchical redundancy, and low sequencing coverage will have less effect on phylogenetic information further back in time. As more taxa are added to a tree, sampling them in a balanced way can rapidly recover quartet information across deeper edges that was otherwise lost due to randomly missing data among tips of the tree (Fig. 2g).

Empirical Data Sets

The number of recovered RAD loci shared among two or four individuals varied by up to 15-30X across the 10 empirical data sets examined, with proportions of missing data ranging between 34% and 92% (Table 1). Size and completeness are affected by a range of factors that differ among the data sets, including genome size, sequencing effort, the restriction enzyme(s) used, the number of samples in the data set, and the amount of sequence divergence between them. From our small sample of studies, the influence of different restriction enzymes is perhaps most apparent. Despite many differences among the data sets, those that used similar restriction enzymes generally recovered a similar number of total shared loci (Table 1). Focusing on quartets as the minimum phylogenetic unit, we can crudely estimate the amount of potential phylogenetic information as the number of RAD loci with sequence data shared across at least four samples (Fig. 3). Data

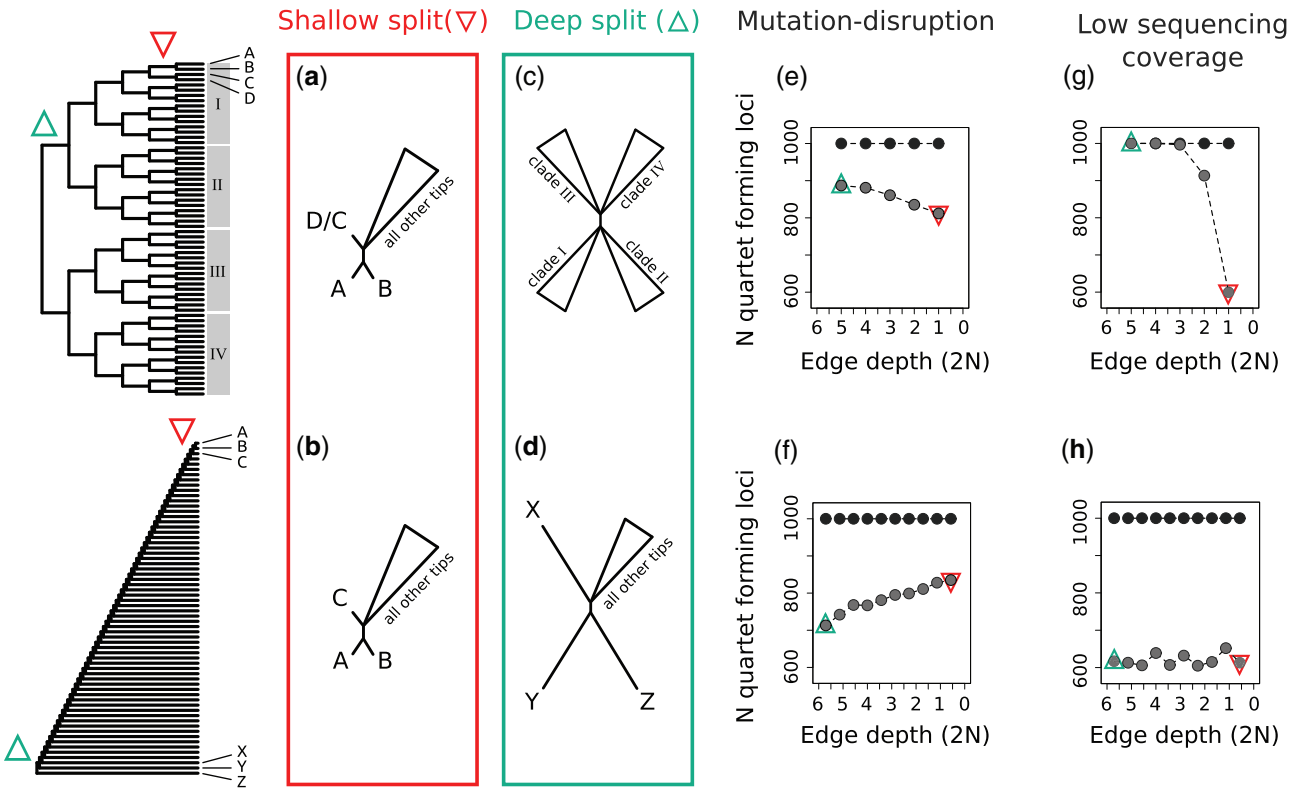

FIGURE 2. The impact of missing data on quartet informativeness when missing data arise from either mutation-disruption or low sequencing coverage. Data were simulated on two trees with contrasting balance and thus edge lengths. The shallowest splits on both topologies (marked throughout by a downward facing triangle) have similar topologies a) and b), and also similar numbers of loci that are quartet-informative (compare the shallowest edge depth in e) to f). In both topologies, the shallowest split contains three short terminal edges along which disrupting mutations are unlikely to occur, and a fourth edge at which many possible taxa could contain data for a locus to be informative a) and b). The two trees differ, however, in the information at their deepest splits c) and d) marked by a upward facing triangle). There is greatest redundancy in sampling across tips of the deepest split in the balanced topology c), but redundancy on only one edge of the deepest split of the imbalanced topology d), which also contains three long terminal edges on which mutation-disruption could occur. As a result, quartet information following mutation-disruption increases across deeper edges of the balanced topology e), but decreases across deeper deeper edges of the imbalanced topology f). When missing data arises from low sequencing coverage, the balanced topology quickly recovers quartet informativeness across deeper edges because of its hierarchical redundancy g). In contrast, the imbalanced topology does not increase in redundancy at deeper edges d), and therefore does not recover quartet informativeness h). Black circles in e–h show the number of quartet informative loci for each split without missing data, grey circles are with missing data, and triangles highlight the deepest and shallowest splits.

generated with a 6-bp restriction enzyme had an average of $9,837 \pm 2,496$ loci shared among four samples; similar to, but slightly more than in data sets using an 8-bp cutter (mean $= 7,026 \pm 6,983$). Both were substantially larger than double-digest data sets that used an 8-bp+4-bp combination (mean $= 1,292 \pm 1,011$) (Fig. 3). The large variance among the 8-bp data sets likely reflects the fact that it includes the *Orestias* data set, which is by far the youngest clade examined, and was also sequenced to relatively high coverage, making it an outlier.

The 10 empirical data sets examined vary with respect to size and tree balance, and thus also in the degree of hierarchical redundancy of their internal edges. In agreement with our simulation results, the number of quartet informative loci is not equally distributed across all edges in these trees, but instead is most concentrated on edges with greatest redundancy — those that have a greater number of descendants from each subtending edge. Figure 4 shows an example for *Viburnum*, *Quercus*, and *Orestias*, each of which accumulates more quartet informative loci on internal edges that have greater hierarchical redundancy (Spearman's rank correlations: $r_s > 0.6$, $P < 1e^{-7}$). This pattern is most prominent in data sets that have many taxa since many tips are required for some edges to have multiple descendant leaves from each subtending edge. It also is most conspicuous in data sets that have low or uneven sequencing coverage since hierarchical redundancy has a greater effect with randomly missing data. Because of the more balanced tree shape of the *Viburnum* data set hierarchical redundancy is greatest at many of its deepest edges, whereas, in contrast, the *Orestias* tree is less balanced across deeper edges and thus its information is more concentrated at shallow depths.

Sources of missing data.—Across data sets, phylogenetic distance was generally a poorer predictor of the number of loci shared among quartets of taxa than sequencing coverage (the number of input reads; Table 1; Fig. S4 available on Dryad). Phylogenetic distance was a better

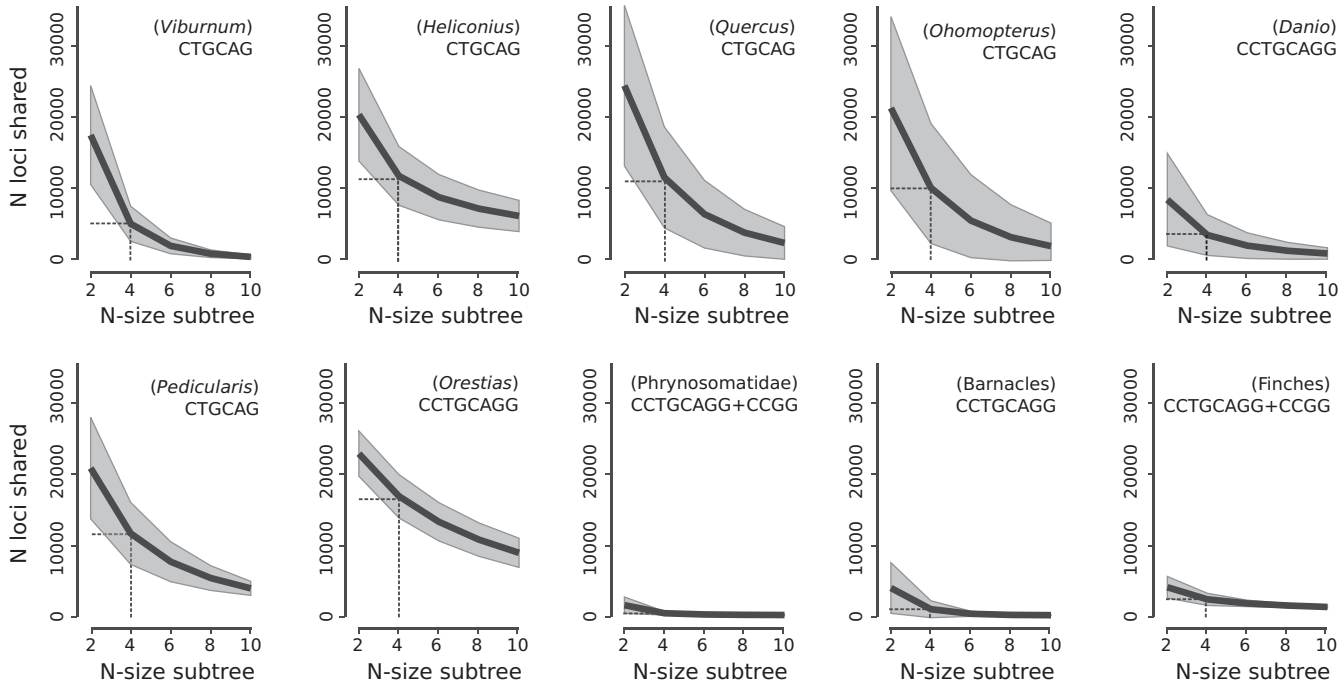

FIGURE 3. Number of loci (mean±SD) shared across subsampled trees of different size in 10 empirical RAD-seq data sets. Data sets are ordered by the total number of loci recovered in their min4 assemblies (Table 1). The restriction recognition sequence (cutter) used to prepare each library is shown. The mean number of loci shared across sets of four taxa (quartets) is marked with a dashed line. All axes are plotted on the same scale.

predictor when sequencing coverage was closer to saturation, either because the number of reads was very high, or the number of sequenced fragments was low (e.g., *Heliconius*, and Finches, respectively; Table 1). Sequencing coverage was uneven in most data sets, with some individuals receiving much higher input than others, and some loci receiving high coverage despite many other loci appearing as singletons. This was observed even in double digest data sets that select many fewer fragments, and thus are predicted to recover higher coverage data given equal sequencing effort (Fig. S5 available on Dryad). Differences across data sets in the proportion of low coverage clusters does not seem to be associated with single versus double digest preparations (Table 1), and may result from other factors such as differences in genome size, biases in fragment amplification, or sequencing off-target DNA fragments.

The 6-bp cutter data sets are predicted to contain a larger number of fragments than 8-bp cutter data sets, and correspondingly, the number of loci shared among sets of taxa in the 6-bp cutter data sets was generally better predicted by the number of input reads than by phylogenetic distance, suggesting these data sets are consistently under-sequenced. An exception was the *Heliconius* data set, which had a substantially larger number of input reads per sample. The 6-bp cutter data sets also generally contained the highest numbers of quartet informative loci (Table 1). *Heliconius* and *Orestias*, the two data sets with the greatest number of input reads, also had the most data shared across the largest number of individuals (Fig. 3). Interestingly, these two data set also show similar rates of data loss across increasing numbers of sampled taxa (Fig. 3b and g). These rates are similar despite the fact that *Orestias* is by far the youngest clade examined, representing essentially a population-level data set, whereas the *Heliconius* data includes many divergent species, with approximately 5X greater sequence divergence on average. This shows that in the near absence of missing data from under sequencing, phylogenetic scale RAD-seq data sets can recover similar amounts of shared data among individuals as has been observed in population-level studies.

Viburnum Phylogeny

Sequencing coverage.—The addition of a second lane of sequence data led to a substantial increase in the number of loci shared among multiple taxa. The increase was particularly prevalent for larger sets of taxa. For example, although the total number of loci shared across at least four samples approximately doubled (from 24,191 to 40,036) upon doubling our sequencing effort, the number of loci shared among 40 samples increased nearly 10X (from 184 to 1724). Because low sequencing coverage tends to cause random missing data, it makes sense that under-sequencing would rapidly reduce the number of loci shared among larger sets of taxa.

The total number of loci in the min4 data set increased from 142K to 199K with the additional lane of sequencing, and the average number of loci per sample from $28,605 \pm 8,731$ to $44,869 \pm 11,037$. The

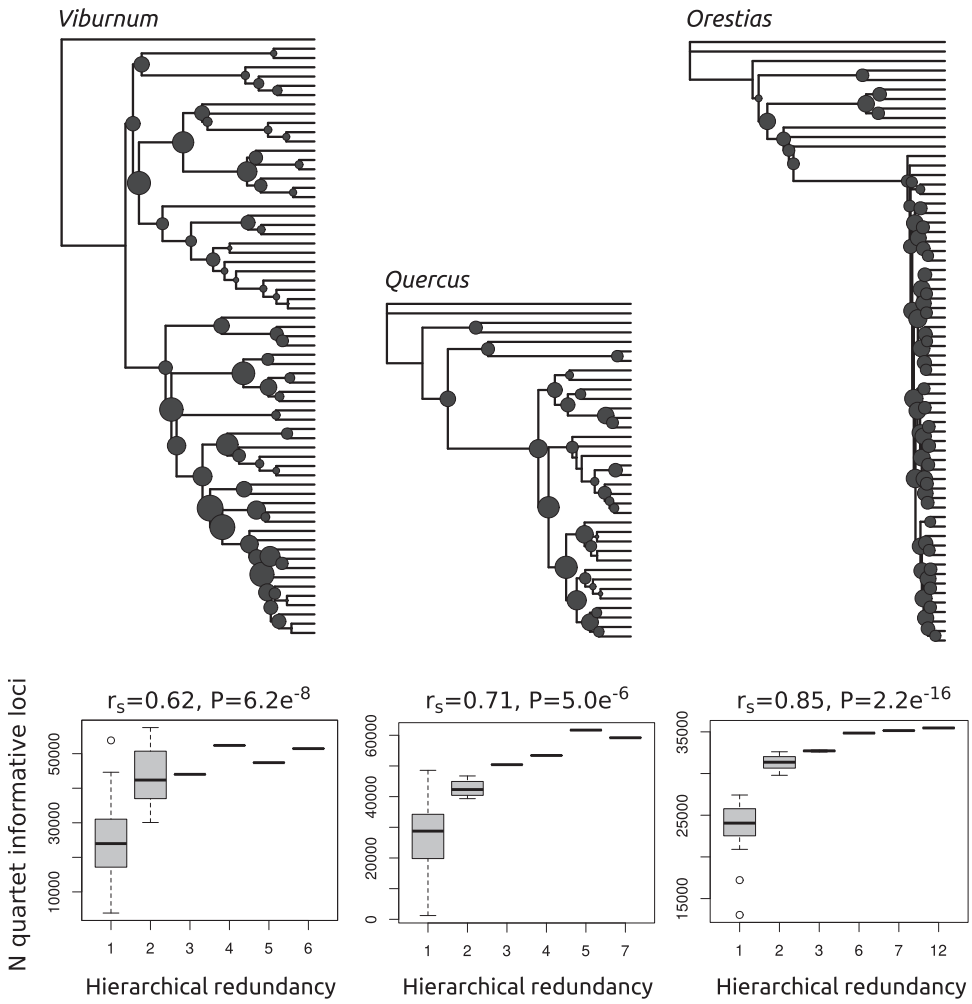

FIGURE 4. The number of loci that are quartet informative at each edge of three empirical RAD-seq data sets. The relative number of quartet informative loci is indicated by the size of the node below (to the right of) each edge. Node sizes are not on the same scale between trees. Boxplots show the actual number of quartet informative loci with respect to the level of hierarchical redundancy at each edge, measured as the minimum number of descendant leaves from any of the four tips of an edge. The *Viburnum* topology is the most balanced across deeper splits and also shows an increase in quartet informativeness across deeper (more internally nested) edges.

TABLE 1. Ten empirical RAD-seq data sets reassembled and reanalyzed for this study

Study	Cut	Ntips	phydist	input	lowcov	min2 (%miss)	min4 (%miss)	β_{phy}	β_{input}
1. *Viburnum*	6	65	0.061	2.7	0.25	496,509 (88.7)	199,094 (77.6)	−0.10	**0.31**
2. *Heliconius*	6	56	0.050	5.3	0.36	212,818 (81.0)	119,819 (70.0)	**−0.13**	0.04
3. *Quercus*	6	36	0.018	2.0	0.43	145,210 (69.2)	87,969 (53.7)	−0.12	**0.15**
4. *Ohomopterus*	6	32	0.040	3.9	0.18	119,441 (68.4)	76,778 (54.6)	−0.10	0.08
5. *Danio*	8	31	0.112	2.6	0.52	259,982 (88.9)	72,237 (74.3)	**−0.21**	−0.01
6. *Pedicularis*	6	13	0.027	1.3	0.73	80,639 (56.9)	44,875 (36.3)	−0.08	**0.24**
7. *Orestias*	8	64	0.010	4.0	0.05	45,146 (37.0)	43,000 (33.8)	−0.03	**0.26**
8. *Phrynosomatidae*	8, 4	74	0.063	1.5	0.72	87,325 (91.8)	41,011 (86.2)	**−0.43**	**0.17**
9. Barnacles	8	13	0.052	0.7	0.24	26,325 (64.4)	15,502 (51.9)	**−0.28**	**0.21**
10. Finches	8, 4	25	0.041	0.6	0.45	17,477 (60.9)	12,864 (50.5)	−0.07	0.07

Notes: Mean pairwise phylogenetic distances (GTR+Γ) between pairs of taxa (phydist), mean number of input sequence reads ($\times 10^6$; input) per taxon, and mean proportion of excluded low coverage clusters across samples (lowcov) is shown, in addition to the total number of RAD-seq loci recovered for pairs of taxa ("min2" data sets), and quartets ("min4" data sets) of taxa, along with the proportions of missing data in the assembled supermatrices (%miss). The number of sampled taxa (Ntips) and the length of restriction recognition sites for enzymes used to prepare libraries (Cut; detailed in Fig. 3) vary across data sets. We fit a phylogenetic least squares model to predict the number of loci shared among quartets of taxa in each data set. Regression coefficients for phylogenetic distance (β_{phy}) and log median number of input reads (β_{input}) are reported the mean value across 100 replicate subsamples of 200 quartets. Regression coefficients are bolded if the mean P-value across replicates was significant at $\alpha = 0.01$.

average number of samples with data for each locus increased from 12.9±9.1 to 14.6±12.7, and the number of parsimony informative SNPs from 793,827 to 1,227,424, corresponding to an increase from 5.5±4.4 parsimony informative SNPs per locus to 6.2±5.3. It is likely that sequencing coverage was still short of saturation, since the proportion of missing data in the half versus full min4 data sets remained similar (80.4% and 77.6%), meaning that many new loci were added that continued to be shared among a small proportion of samples.

Consistent with our simulation results, the number of quartet informative loci across splits of the *Viburnum* tree was greatest at edges with more hierarchical redundancy (Fig. 4). This relationship is very similar to the one predicted when data were simulated on the *Viburnum* topology with low sequencing coverage (Fig. S2f-g available on Dryad), and consistent with our PGLS estimates predicting that most missing data in this data set is due to low sequencing coverage rather than mutation-disruption.

Despite differences in size and completeness of the half and full data sets, both recovered highly similar ML topologies (Fig. 5a; Fig. S6 available on Dryad), with only a small difference being the placement of *V. erosum*. Nearly all splits in the tree have 100% bootstrap support (Fig. 5a). The quartet-based species tree inferred from SNPs recovered a similar topology as the ML analyses (Fig. 5b; Fig. S6 available on Dryad), with only slight differences that were also associated with lower bootstrap support. An important difference is in support for the placement of *Pseudotinus*, a clade that has historically proven difficult to place along the backbone of *Viburnum* (Clement et al. 2014; Spriggs et al. 2015). The placement of *Pseudotinus* in relation to the *Urceolata* clade is of special interest because, although these groups have not been united in cpDNA analyses, they share several distinctive morphological characterstics, including naked buds (lacking modified scales) and a unique sympodial branching architecture. Previous results from whole chloroplast genome data (Clement et al. 2014) placed *Urceolata* sister to *V. amplificatum* + *Crenotinus* (Fig. 5d). However, this clade was puzzling to the extent that it was named *Perplexitinus* to highlight the lack of morphological synapomorphies (Clement et al. 2014).

Our concatenated ML and quartet-based species tree analyses support a direct link between *Pseudotinus* and *Urceolata* (Fig. 5a and b), which makes the most sense from a morphological standpoint. However, while the concatenated analysis found perfect support for this relationship, the quartet-based analysis showed significant uncertainty (67% bootstrap support). The best supported alternative relationship in the quartet analysis includes the two clades in sequence (paraphyletically) along the backbone (Fig. 5c; Fig. S6 available on Dryad). As the placement of *Pseudotinus* and *Urceolata* in our analyses has mixed support, and conflicts with previous cpDNA analyses, we focused our Bayesian concordance factor analysis on exploring the distribution of RAD loci supporting each alternative hypothesis.

Bayesian concordance analysis.—Concordance factors are not estimates of support, but rather represent the proportion of sampled loci for which a clade is recovered (Baum 2007). We tested several values for the prior on the number of true underlying genealogies (a parameter which affects the correction of gene tree estimation errors) but present only $\alpha = 0.1$, since all results were qualitatively similar. The primary concordance tree for our deep-scale analysis, which represents the most frequently occurring nonconflicting clades across sampled loci, matches the topology from our concatenation and quartet-based analyses. This was true whether the concordance tree was inferred from the full data set, which contained 1,203 loci with at least one parsimony informative SNP and data for all 8 selected taxa in this test, or if we used the half data set, in which only 120 loci met the same criteria. The larger data set consistently yielded higher CFs and narrower confidence intervals.

On the primary concordance tree two clades (*amplificatum* + *Crenotinus*, and *Tinus* + *Imbricotinus*) were recovered with a 95% credibility interval on their CF that does not conflict with any other clade, while the remaining three clades show greater conflict (Fig. 5b). The placement of *Pseudotinus* in the alternative position that received moderate support in the quartet-based analysis (Fig. 5c) has a CF that is approximately half (95% CI = 0.08–0.17) that of the conflicting clade found in the primary concordance topology (95% CI = 0.15–0.23). The placement of *Urceolata* in the cpDNA topology, where it forms a clade with *amplificatum*+*Crenotinus*, received similarly low support (95% CI = 0.08–0.19); while the placement of *Urceolata* within a clade that includes *Tinus* + *Imbricotinus* to the exclusion of *Pseudotinus* (as supported by cpDNA) was very poorly supported (95% CI = 0.03–0.09) (Fig. 5d).

The shallow-scale relationships shown in Figure 5a and 5e for the *Oreinodentinus* clade are strongly supported here, in contrast to cpDNA results where the data were insufficient for confident resolution (Clement et al. 2014; Spriggs et al. 2015). Importantly, our RAD analyses are congruent with geography over morphological characters (leaf size, shape, and pubescence) in strongly supporting sister relationships between *V. sulcatum* and *V. acutifolium* from Mexico, and between *V. jamesonii* and *V. triphyllum* from Ecuador. It is noteworthy that we recovered 3,300 loci in the full data set that met our criteria for inclusion in the analysis, and only 364 loci in the half data set. Again, both data sets yielded the same topology, and the larger data set recovered higher concordance factors and narrower confidence intervals (Fig. 5e).

Discussion

Sources of Missing Data

A common intuition about RAD-seq is that it contains little phylogenetic information for resolving deep-scale relationships. Here, we examined both

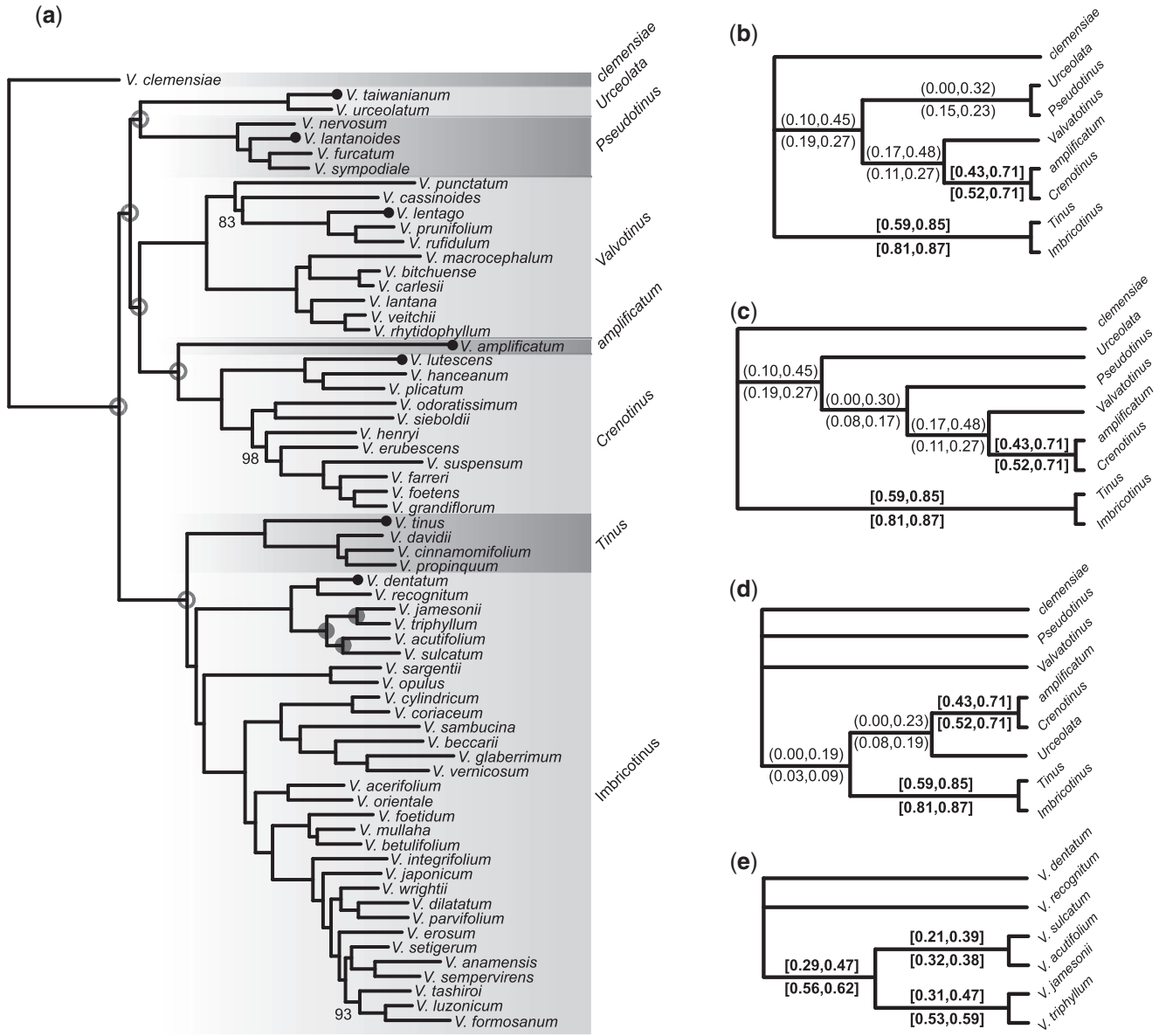

FIGURE 5. RAD-seq phylogeny for 65 species of *Viburnum*. a) Rooted maximum likelihood tree inferred from the largest phylogenetically informative (min4) concatenated data set. Bootstrap support is 100 except where indicated. Open and closed grey circles indicate deep and shallow splits, respectively, that were further analyzed with Bayesian concordance analysis b)–e). Closed black circles at the tips highlight taxa used as representatives of major clades in the deep-scale analysis (b-d). b) The primary concordance tree for deep splits among major clades of *Viburnum* matches the ML concatenated topology. The 95% CI on concordance factors (CFs) inferred from the "half" and "full" data sets are shown above and below each split, respectively. The two splits with CFs that do not overlap the 95% CI of any alternative split are shown in bold inside brackets, while the CFs for three splits that do conflict with at least one other split are shown in plain text inside parentheses. c) An alternative topology in which *Urceolata* and *Pseudotinus* do not form a clade, as supported by quartet species tree analysis, has lower support. d) The Clement et al. (2014) cpDNA topology has low CFs on splits that differ from the primary concordance topology shown in b. e) Splits within the recent and rapidly radiated *Oreinodentinus* clade are all strongly supported.

simulated and empirical data to show that although mutation-disruption causes more distant relatives to recover fewer shared RAD loci, this pattern does not necessarily lead to the corollary expectation of little (or even less) phylogenetic information across deeper splits in a tree. We identify several factors that influence how missing data among sampled tips of a tree translates into information that can be used to test support for splits within the tree using concatenation, quartet-based, and concordance phylogenetic approaches.

Our simulation study shows clearly an effect of tree shape and size (number of taxa) that together influence the information at edges of varying depths. Edges that are multiple nodes removed from a tip suffer less information loss from missing data among tips of the tree whether the data are missing from mutation-disruption or low sequencing coverage. In either case,

the hierarchical redundancy associated with these more nested edges makes it possible for more loci to have data for at least two individuals on each side of a split, such that the minimal quartet phylogenetic information is available. Sampling more individuals will necessarily shorten some branches of a tree and in doing so will increase the number of independent lineages along which data may be recovered. An interesting aspect of this property of RAD-seq data is that as larger data sets are assembled that include more sampled taxa, many loci that are initially found as singletons, or that are shared among few taxa, can become increasingly phylogenetically informative.

Although we did not include mutation-generation of new fragments in our simulations (as this would require assumptions about the size and composition of the genome), we expect that this type of missing data also varies greatly across data sets of different size and age, or prepared using different restriction digest methods. Loci that arise within a subclade from mutation-generation are expected not to be shared among all individuals in a larger clade. These loci can provide phylogenetic information for taxa within the subclade in which the locus is present, but are not informative about the placement of that subclade within the larger tree. Interestingly, of the 10 data sets examined, the double digest libraries, which tend to select fewer total fragments and should experience less mutation-generation of new fragments since they require two restriction enzymes to be present, did not tend to have less missing data than single digest libraries (Table 1). This suggests that mutation-generation of new fragments, at least over the time scales examined and considering the restriction enzymes used, is not a primary source of missing data in RAD-seq data sets.

How much Phylogenetic Information is Necessary?

In sequencing projects there is a persistent trade-off between cost, quality, and quantity, with decisions to be made about which protocols to use and how much sequencing effort to expend. One of the major benefits of "reduced complexity" methods like RAD-seq is that they simplify many cost-benefit decisions that must be made when preparing genomic libraries, because the restriction digestion more or less equalizes the number of expected markers that will be present across a group of closely related organisms. However, there remains the question of how many markers to sample (how common a cutter to use), and how many individuals to multiplex (how much sequencing effort per sample). For evolutionary questions at shallow scales, coverage as low as 1X is commonly used, and the benefit gained by multiplexing many individuals has been argued to outweigh the potential errors introduced from low coverage (Buerkle and Gompert 2013); but see (Harvey et al. 2013). A concern for using low coverage data is that sequencing errors, if not corrected, can significantly influence branch length estimates (Kuhner and McGill 2014), and perhaps even the topology. For phylogenetic-scale questions, individuals typically represent divergent populations or species, and thus it may be less appropriate to account for sequencing errors, or impute missing data (Li et al. 2009), using population genetic assumptions about the expected distribution of variation.

Across the 10 empirical data sets examined, those with greater sequencing effort per sample generally contained more shared loci among samples (Table 1; Fig. 3). In the case of *Viburnum*, which is a relatively old clade with a fairly large genome (~4 Gb; (Bennett and Leitch 2012)), a minimal increase in sequencing effort (2X), increased the number of markers with shared data across different subsets of the tree by 2–10X, and nearly doubled the number of parsimony informative sites. Consequently, the amount of information available for quartet-based inference using SNPs was approximately doubled, and the amount of information for our concordance analyses was increased by 10X, since this method requires fully sampled gene trees. Even in the case of species tree methods based on the multi-species coalescent that do not require fully sampled gene trees, such as *MP-EST* (Liu et al. 2010) and *ASTRAL* (Mirarab et al. 2014b), RAD loci that contain more SNPs shared across a greater number of taxa will provide more informative inputs, which has been shown to consistently improve performance of these methods (Mirarab et al. 2014a; Liu et al. 2015; Xi et al. 2015).

One can imagine two competing strategies for attaining large phylogenetic RAD-seq data sets given limited resources. The first is to sequence many loci to lower coverage since, as we have shown, randomly missing data have little effect on the amount of information at deeper edges of a tree (at least if it can be sampled in a balanced way). Examples of this are data sets 1–7 (Fig. 3), which recover relatively sparse matrices, but with enormous numbers of loci that are quartet informative, making them highly useful for concatenation and quartet-based phylogenetic approaches. A second strategy is to sequence fewer loci to higher coverage, which would be more useful when seeking to attain informative gene trees. Data sets 2, 4, 5, and 10 (Table 1) represent highly divergent data sets in which missing data are poorly explained by the amount of input data, suggesting that sequencing coverage was nearly saturated. However, it should be noted that these data sets are still composed of large amounts of missing data in their min4 assemblies (70, 54, 74, and 50%, respectively). Although sampling fewer loci to higher coverage can reduce the amount of data that is randomly missing, the use of additional cutters to reduce the number of loci comes with a trade-off of higher mutation disruption. Our simulation and empirical results both show that single digest preparations tend to recover much more phylogenetic information than double digest methods, and that when coupled with high sequencing input can yield orders of magnitude more data.

RAD-seq and related methods

A number of studies have used *in silico* approaches to investigate the expected distribution of missing data in RAD-seq assemblies, and these provided the earliest evidence that RAD could be informative for phylogenetic analyses (Rubin et al. 2012). However, many empirical RAD-seq data sets have proven less informative than these studies would predict, likely due to missing data from sources besides mutation-disruption, which *in silico* studies generally do not consider. As we have shown (Table 1), details of the RAD-seq preparation protocol substantially impact the amount of data that will be recovered.

An important point of comparison between RAD-seq and related reduced-representation genomic sequencing methods, such as anchored hybrid enrichment (Lemmon et al. 2012) and ultra conserved elements (UCEs; McCormack et al. 2012), is in the total amount of phylogenetic data that each can produce. Although RAD-seq can sample many more genomic regions in total than the other two methods, it is often thought that over deep time-scales the introduction of missing data will reduce its information content to a level below that of the other methods which contain very little missing data. (Collins and Hrbek 2015) recently overturned this expectation by comparing the *in silico* predicted phylogenetic informativeness of RAD, UCE, and anchored hybrid enrichment data that could be bioinformatically extracted from 33 primate genomes. In their comparison, even at the deepest time-scales examined (60–80 Ma), RAD provides far more phylogenetic information than alternative methods, despite the presence of substantial missing data in the RAD-seq data set.

How large are these differences? Consider the difficult phylogenetic problem of resolving the branching order of early diverging lineages of neoavian birds, which was recently investigated in two large-scale analyses using reduced-representation genomic methods. (Jarvis et al. 2014) combined full genome re-sequencing data with UCE loci, of which the latter provided a matrix of ~370 Kb for 48 taxa, whereas (Prum et al. 2015) sampled ~390 Kb from 198 taxa using anchored hybrid enrichment loci. Both data sets had negligible amounts of missing data. Because our *Viburnum* data set is of a similar, albeit slightly younger crown age (50–60 Myr) it provides an interesting comparison. The concatenated supermatrix of RAD loci in our min4 data set for 65 species of *Viburnum* was substantially larger than either reduced-representation bird data set, with 17.1 Mb. Despite its 78% missing data, this included 1.2 million parsimony informative SNPs (3.2M SNPs total), meaning that the *Viburnum* data set contains three times more informative sites than the two reduced-representation bird data sets contain characters.

As we have shown, the *Viburnum* RAD-seq data are not concentrated only among close relatives (Fig. 4), but rather, there is substantial information across all edges of the tree, and it actually increases across deeper edges (Fig. S2g available on Dryad). In fact, although there were only 1,203 loci (~100Kb) shared across the 8 taxa representing the most divergent splits in *Viburnum* (those selected in our *BUCKy* analyses), the total number of loci spanning these splits is far greater when considering the many additional combinations of 8 taxa that could be sampled from these 8 clades, since many contain multiple taxa. The counter-intuitive result that the sparcity of RAD-seq supermatrices does not translate into a sparcity of phylogenetic information is an important consideration when comparing reduced-representation genomic methods.

Missing Data and Phylogenetic Inference

Although it has been a topic of interest for many years, it remains unclear how missing data affects phylogenetic inference under a range of scenarios, including different inference methods, data types, and proportions of missing data. Many of the studies of these problems predate the availability of genomic sequence data. (Wiens 2003) showed that reduced accuracy from missing data is typically a function of having too few sampled characters shared between taxa rather than too many missing data cells. Similarly, the problem of terraces in phylogenetic tree space (Sanderson et al. 2011; 2015) was initially described for cases where many samples share little or no informative characters when partitioning data. This is not the problem we face in RAD-seq phylogenetics. Typically, any sample will share hundreds or thousands of loci with every other sample. As the problem for RAD-seq data is one of thousands of partially overlapping taxon sets, we focused our analyses on the distribution of the smallest informative sets, namely quartets (Berry and Gascuel 2000).

An important distinction of the quartet informativeness metric is that it is measured conditional on a topology. Thus, despite the fact that there may be a substantial amount of information to resolve a split on the correct topology, one could imagine that an uneven distribution of phylogenetic information could bias the ability to traverse tree space, and thus to find this topology. (Whidden and Matsen 2015) recently described a method for visualizing and measuring the efficiency with which MCMC chains traverse tree space which will likely be a useful direction for further research on how missing data affects the ability to reach optimal regions of tree space. Similarly, (Huang and Knowles 2016) showed that mutation-disruption can lead to a biased mutational spectrum, such that some taxa will share loci that evolve faster or slower than others. These topics deserve further consideration for RAD-seq as well as other data types for understanding how different distributions of missing data affect phylogenetic inference.

Conclusions

Our counter-intuitive results show that although RAD-seq data sets have high proportions of missing data, including more missing data between more

divergent taxa, there is potential to gather more phylogenetic information (measured in terms of quartet informativeness) about splits deeper in a tree than toward the tips. RAD-seq data can therefore be useful in resolving shallow (population-level) relationships, as generally assumed, but also in resolving much deeper phylogenetic relationships on the scale of the 50-60 million-year-old *Viburnum* phylogeny presented here. However, as we have also shown, high bootstrap values in concatenated analyses need to be thoroughly explored (e.g., using multi-locus coalescent and concordance approaches), and may reveal wide variation in the nature and strength of support (Salichos and Rokas 2013). By assembling a large RAD-seq data set containing millions of SNPs, as well as hundreds of well-sampled gene trees, we were able to compare multiple phylogenetic methods to test alternative phylogenetic hypotheses.

One of our primary findings is that increased sequencing coverage can greatly increase the phylogenetic utility of RAD-seq data sets. This is encouraging considering that sequencing costs are likely to continue to decrease for some time. Furthermore, we show that hierarchical redundancy reduces the impact that missing data among tips of a tree has on phylogenetic information at edges within it. Larger trees necessarily have more edges on which hierarchical redundancy can be increased. Sampling more taxa, and targeting more balanced tree shapes will further minimize the effects of missing data. Both of these findings suggest that early RAD-seq studies, which often sampled few taxa and were generated on sequencing platforms that yielded fewer reads, may not be representative of the potential that RAD-seq offers for phylogenetic resolution of clades from the scale of thousands of years to tens of millions of years. Our analysis of the angiosperm clade *Viburnum* makes clear that the amount of information that can be attained rapidly and efficiently using RAD-seq makes it a powerful approach for performing large-scale comparative genomic analyses.

Reproducibility

Code to download the 10 empirical data sets, assemble them, and reproduce our results are organized into Jupyter/IPython notebooks available at https://github.com/dereneaton/RADmissing. The demultiplexed *Viburnum* fastq data are archived in the NCBI SRA: SRP065788.

Supplementary Data

Supplementary tables and figures can be found in Dryad Repository: (http://dx.doi.org/10.5061/dryad.g549v).

Funding

This work was supported by National Science Foundation (DEB-1145606, IOS-1256706)

Acknowledgements

We thank Cecile Ané and Michael Landis for helpful discussions on implementing the regression analyses, and we also thank Laura Kubatko and two anonymous reviewers for suggestions that improved the article. We also thank the many researchers who made their data sets publicly available.

References

Avni E., Cohen R., Snir S. 2015. Weighted quartets phylogenetics. Syst. Biol. 64:233–242.

Baird N.A., Etter P.D., Atwood T.S., Currey M.C., Shiver A.L., Lewis Z.A., Selker E.U., Cresko W.A., Johnson E.A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376.

Baum D.A. 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. Taxon 56:417–426.

Bayzid M.S., Mirarab S., Boussau B., Warnow T. 2015. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. PLoS ONE 10:e0129183.

Bennett M., Leitch I. 2012. Plant DNA C-values database (release 6.0, Dec. 2012). http://www.kew.org/cvalues/. (Accessed: 2015-09-30).

Berry V., Gascuel O. 2000. Inferring evolutionary trees with strong combinatorial evidence. Theoret. Computer Sci. 240:271–298.

Buerkle A.C., Gompert Z. 2013. Population genomics based on low coverage sequencing: how low should we go? Mol. Ecol. 22:3028–3035.

Cariou M., Duret L., Charlat S. 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. Ecol. Evol. 3:846–852.

Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30:3317–3324.

Clement W.L., Arakaki M., Sweeney P.W., Edwards E.J., Dnoghue M.J. 2014. A chloroplast tree for *Viburnum* (Adoxaceae) and its implications for phylogenetic classification and character evolution. Am. J. Botany 101:1029–1049.

Collins R.A., Hrbek T. 2015. An in silico comparison of reduced-representation and sequence-capture protocols for phylogenomics. *bioRxiv*, p. 032565.

DaCosta J.M., Sorenson M.D. 2016. ddRAD-seq phylogenetics based on nucleotide, indel, and presence–absence polymorphisms: Analyses of two avian genera with contrasting histories. Mol. Phylogenet. Evol. 94:122–135.

Eaton D.A.R., Hipp A.L., González-Rodríguez A., Cavender-Bares J. 2015. Historical introgression among the American live oaks and the comparative nature of tests for introgression. Evolution 69:2587–2601.

Eaton D.A.R., Ree R.H. 2013. Inferring phylogeny and introgression using RADseq data: An example from flowering plants (Pedicularis: Orobanchaceae). Syst. Biol. 62:689–706.

Eaton D.A.R. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 30:1844–1849.

Escudero M., Eaton D.A., Hahn M., Hipp A.L. 2014. Genotyping-by-sequencing as a tool to infer phylogeny and ancestral hybridization: A case study in carex (cyperaceae). Mol. Phylogenet. Evol. 79:359–367.

Harvey M.G., Smith B.T., Glenn T.C., Faircloth B.C., Brumfield R.T. 2013. Sequence capture versus restriction site associated DNA sequencing for phylogeography. *arXiv:1312.6439 [q-bio]*. ArXiv: 1312.6439.

Herrera S., Watanabe H., Shank T.M. 2015. Evolutionary and biogeographical patterns of barnacles from deep-sea hydrothermal vents. Mol. Ecol. 24:673–689.

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:e93975.

Huang H., Knowles L.L. 2016. Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation study of RAD Sequences. Syst. Biol. 65:357–365.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., Fonseca R.R.d., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M.V., Alfaro-Núñez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jønsson K.A., Johnson W., Koepfli K.P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alström P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.

Kuhner M.K., McGill J. 2014. Correcting for sequencing error in maximum likelihood phylogeny inference. G3: Genes | Genomes | Genetics 4:2545–2552.

Larget B.R., Kotha S.K., Dewey C.N., Ané C. 2010. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26:2910–2911.

Leaché A.D., Chavez A.S., Jones L.N., Grummer J.A., Gottscho A.D., Linkem C.W. 2015. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biol. Evol. 7: 706–719.

Leaché A.D., Banbury B.L., Felsenstein J., nieto-Montes de Oca A., Stamatakis, A. 2015. Syst. Biol. 64:1032–1047.

Lemmon A.R., Emme S.A., Lemmon E.M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61:727–744.

Li Y., Willer C., Sanna S., Abecasis G. (2009). Genotype Imputation. Ann. Rev. Genomics Human Genet 10:387–406.

Liu L., Xi Z., Wu S., Davis C.C., Edwards S.V. 2015. Estimating phylogenetic trees from genome-scale data. Ann. N.Y. Acad. Sci. 1360:36–53.

Liu L., Yu L., Edwards S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10:1–18.

McCluskey B.M., Postlethwait J.H. (2015). Phylogeny of Zebrafish, a "Model Species," within Danio, a "Model Genus". Mol. Biol. Evol. 32:635–652.

McCormack J.E., Faircloth B.C., Crawford N.G., Gowaty P.A., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22: 746–754.

Miller M.R., Dunham J.P., Amores A., Cresko W.A., Johnson E.A. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17:240–248.

Mirarab S., Bayzid M.S., Boussau B., Warnow T. 2014a. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346(6215).

Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014b. Astral: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548.

Mita S.D., Siol M. 2012. EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genetics 13:27.

Nadeau N.J., Martin S.H., Kozak K.M., Salazar C., Dasmahapatra K.K., Davey J.W., Baxter S.W., Blaxter M.L., Mallet J., Jiggins C.D. 2013. Genome-wide patterns of divergence and gene flow across a butterfly radiation. Mol. Ecol. 22:814–826.

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.

Pinheiro J., Bates D., DebRoy S., Sarkar D., R Core Team 2016. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-128.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526:569–573.

de Queiroz A., Gatesy J. 2007. The supermatrix approach to systematics. Trend Ecol. Evol. 22:34–41.

Ree R.H., Hipp A.L. 2015. Inferring phylogenetic history from restriction site associated DNA (RADseq). In: (Hörandl E., Appelhans M., editors. *Next-generation sequencing in plant systematics*, International Association for Plant Taxonomy (IAPT), chap. 6.

Ronquist F., Huelsenbeck J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Rubin B.E.R., Ree R.H., Moreau C.S. 2012. Inferring phylogenies from RAD sequence data. PLoS ONE 7:e33394.

Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

Sanderson M.J., McMahon M.M., Stamatakis A., Zwickl D.J., Steel M. 2015. Impacts of terraces on phylogenetic inference. Syst. Biol. 64:709–726.

Sanderson M.J., McMahon M.M., Steel M. 2011. Terraces in phylogenetic tree space. Science 333:448–450.

Sanderson M.J., Purvis A., Henze C. 1998. Phylogenetic supertrees: Assembling the trees of life. Trend Ecol. Evol. 13:105–109.

Spriggs E.L., Clement W.L., Sweeney P.W., Madriñán S., Edwards E.J., Donoghue M.J. 2015. Temperate radiations and dying embers of a tropical past: the diversification of *Viburnum*. New Phytologist. 207:340–354.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Symonds M.R.E., Blomberg S.P. 2014. A primer on phylogenetic generalised least squares. In: Garamszegi L.Z., editor. *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*. Berlin Heidelberg: Springer, chap. 5, pp. 105–130.

Takahashi T., Moreno E. 2015. A RAD-based phylogenetics for Orestias fishes from Lake Titicaca. Mol. Phylogenet. Evol. 93:307–317.

Takahashi T., Nagata N., Sota T. 2014. Application of RAD-based phylogenetics to complex relationships among variously related taxa in a species flock. Mol. Phylogenet. Evol. 80:137–144.

Whidden C., Matsen F.A. 2015. Quantifying MCMC exploration of phylogenetic tree space. Syst. Biol. 64:472–491.

Wiens J.J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52:528–538.

Xi Z., Liu L., Davis C.C. 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. Mol. Phylogenet. Evol. 92:63–71.