# Chapter 6

## Inferring phylogenetic history from restriction site associated DNA (RADseq)

Richard H. Ree[1] & Andrew L. Hipp[1,2]

1 The Field Museum, 1400 South Lake Shore Drive, Chicago, Illinois 60605, U.S.A.
2 The Morton Arboretum, 4100 Illinois Route 53, Lisle, Illinois 60532, U.S.A.
Authors for correspondence: R.H. Ree, rree@fieldmuseum.org; A.L. Hipp, ahipp@mortonarb.org

**Abstract** Next-generation sequencing of restriction site associated DNA, or RADseq, was introduced in 2008 as a rapid genotyping method that does not require prior marker development. Developed for linkage mapping, genome-wide association, and population genetic studies, RADseq was initially viewed as ill-suited to interspecific phylogenetic questions. However, since 2012, approximately a dozen RADseq phylogenetic studies have been published. These studies utilize a variety of bioinformatic methods to identify loci, estimate orthology, and assemble phylogenetic matrices, and software pipelines customized for phylogenetic analyses are being rapidly developed. The resulting data matrices tend to be large (sometimes megabases in total aligned length) but relatively incomplete, presenting analytical challenges. Empirical and simulated RADseq studies have demonstrated that RADseq is suitable for phylogenetic inquiries at relatively deep scales, in some cases at least 60 million years. Yet its real strengths may show up at the boundary between within- and among-species inquiries. As sequence-based data, RADseq can be mapped to genomic resources for purposes of alignment, gene identification, and investigating the genomic architecture of introgression and differentiation. Given the ease with which RADseq data can be generated, particularly in comparison to targeted enrichment methods that require prior identification of candidate loci, in the coming years we anticipate greater use of RADseq for phylogenetic inference and predict that methods of species-tree estimation and genomic analysis will increasingly accommodate its characteristically large, incomplete, genome-scale data matrices.

**Keywords** concatenation; phylogeny reconstruction; reduced-representation genome sequencing; species tree inference

Hörandl, E. & Appelhans, M.S. (eds.), Next-Generation Sequencing in Plant Systematics.
Chapter 6, http://dx.doi.org/10.14630/000007

## Introduction

A primary endeavor in systematics is to infer, as accurately and comprehensively as possible, phylogenetic relationships among taxa. This poses a perennial and ultimately practical question for the researcher: How much character data can be gathered having the greatest signal, for the largest taxon sample, for the least cost and effort? In the current era of phylogenomics, approaches to this problem are being rapidly transformed by advances in next-generation sequencing (NGS), which present a panoply of new opportunities and challenges for molecular systematics (e.g., Lemmon & Lemmon, 2013; McCormack & al., 2013). In this paper, we focus on one particular NGS method, restriction-site-associated DNA sequencing (RADseq), and appraise its potential as a tool for phylogeny reconstruction.

To put RADseq in context, it is worthwhile to first consider how molecular systematics was previously transformed by technological advances in sequencing: namely, by the advent of PCR in the 1980s and automated Sanger sequencing in the 1990s. These enabled systematists to find and target loci having desirable properties for phylogenetic inference, the key properties being (1) easy amplification, (2) orthology, and (3) appropriate rates of nucleotide substitution across the clades of interest. Needless to say, great progress in reconstructing phylogeny ensued: automated Sanger sequencing unlocked floodgates, resulting in a torrent of new Sanger-sequence-based phylogenies.

Accompanying this progress was the growing realization that the relatively small number of commonly sequenced markers, which in plants emphasized regions of the chloroplast genome and ribosomal DNA, had insufficient signal to resolve many clades, particularly relationships among the most recently diverged species and, in deeper time, the backbones of rapid radiations. In addition, growing awareness of coalescent theory (e.g., Maddison, 1997; Edwards, 2009) shifted attention toward the need for multiple unlinked markers from the nuclear genome to reconstruct the "species tree", i.e., the branching demographic history of lineages as opposed to the genealogies of individual genes. In short, the Sanger sequencing revolution yielded great advances, particularly in taxon sampling, but has seemingly come up against practical limits to its resolving power from the standpoint of characters and phylogenetic signal. This leads naturally to the question: How can NGS be used to gather phylogenetically informative, genome-wide variation in a cost-effective manner for many taxa?

We use the terms RAD and RADseq here in a broad sense, to refer to a family of NGS methods based on the common premise of targeting a "reduced representation" of the genome associated with restriction sites. This includes

genotyping-by-sequencing (GBS; Elshire & al., 2011) and a host of other RADseq variants. With these methods, genomic DNA is digested with one or more restriction enzymes, and NGS sequencing adapters are ligated to the resulting fragments, yielding reads flanking the restriction sites. RADseq was originally designed as a genomic tool for the discovery of single-nucleotide polymorphisms (SNPs) among individuals of the same species, for purposes of population genetic inference, linkage mapping, association studies, and other intraspecific genomic analyses (Baird & al., 2008). This has led to a general perception that its utility as a tool for molecular systematics is similarly limited to the level of populations, or at most sibling species, not deeper levels of phylogenetic divergence (Lemmon & Lemmon, 2013; McCormack & al., 2013). However, some recent studies and our own research in plants has made us suspect that the extent of phylogenetic signal in RAD sequences has not yet been fully explored, and may be more useful at interspecific levels than generally expected.

To date, relatively few studies have used RADseq to infer interspecific phylogenetic relationships, but the number seems poised to increase rapidly. An early study used *in silico* methods on published genome assemblies of *Drosophila*, *Saccharomyces*, and a broad sample of mammals (Rubin & al., 2012), as did a concurrent but independent study of the same *Drosophila* data (Cariou & al., 2013). We are aware of published RADseq phylogenies in four genera in plants—*Pedicularis* (Eaton & Ree, 2013), *Quercus* (Hipp & al., 2013, 2014; Cavender-Bares & al., 2015; Eaton & al., 2015.; Deng & al. in prep.), *Carex* (Escudero & al., 2014), and *Valeriana* (Gonzalez, 2014)—and several in animals: two in ground beetles (*Carabus* spp.; Cruaud & al., 2014; Takahashi & al., 2014), zebrafish (*Danio* spp.; McCluskey & Postlethwait, 2014), swordtail fish (*Xiphophorus* spp.; Jones & al., 2013), cichlid fish (Wagner & al., 2013), octocorals (*Chrysogorgia* spp.; Pante & al., 2015), *Adelpha* butterflies (Ebel & al., 2015), and hydrothermal vent barnacles (Herrera & al., 2014). In addition, RADseq-based phylogenies have appeared in population-level studies of sibling species or species complexes, including irises (*Iris brevicaulis* and *I. fulva*; Hamlin & Arnold, 2014), butterflies (*Heliconius melpomene/cydno* complex; Nadeau & al., 2013), flycatchers (*Zimmerius viridiflavus* complex; Rheindt & al., 2014), barking frogs (*Craugastor augusti* and *C. tarahumaraensis*; Streicher & al., 2014), and geckos (*Hemidactylus fasciatus* complex; Leaché & al., 2014). Other published studies to date have focused on intraspecific RADseq variation.

Our objective in this chapter is to review the rapid progress in RADseq phylogenetics from 2012 through late 2014, with an eye toward the particular strengths and weaknesses of RADseq data, the bioinformatic challenges and

solutions associated with data analysis, and the potential limits of its utility. It is an area of research that is young and moving fast; trends in publication indicate that it is an area of exploding interest, especially for relatively young clades representing species complexes/flocks, adaptive radiations, and other recalcitrant problems of phylogenetic inference. More generally, we are interested in evaluating how RADseq fits into the landscape of molecular systematics methods and how it might evolve with technological trends.

## RAD sequence data: acquisition and analysis

The various protocols for RAD library preparation have been reviewed and compared in several previous articles (e.g., Davey & al., 2011—see especially fig. 1 in that article; Beissinger & al., 2013; Davey & al., 2013; Andrews & Luikart, 2014; Andrews & al., 2014; Puritz & al., 2014) and will not be rehashed in any detail here. Readers are advised to consult these, as well as the original literature on the methods (Miller & al., 2007a, b; Baird & al., 2008; Elshire & al., 2011; Etter & al., 2011; Peterson & al., 2012; Poland & al., 2012; Wang & al., 2012; Stolle & Moritz, 2013; Toonen & al., 2013), before embarking on a genotyping project using any of these methods. Briefly, the protocols vary in factors such as whether digested fragments are randomly sheared, in using one versus two restriction enzymes, in whether they use PCR enrichment, and in fragment generation using Type IIB versus typical Type II restriction enzymes. These translate into different types of bias in sequence acquisition and in the amount of data returned; thus, the decision on a particular method may carve a relatively deep channel in one's research program for years to come and should be considered carefully before one undertakes large-scale data-gathering. Here we use the terms RADseq and RAD to refer to all such methods.

All RADseq methods share a single salient characteristic: genome subsampling by limiting sequencing to the regions proximate to a specified restriction enzyme recognition sequence. By sequencing only the regions adjacent to or flanking the restriction sites cut during genomic DNA digestion at the outset of a library preparation, RADseq methods reduce the amount of sequencing needed to genotype each individual in the study. This is what gives RADseq methods their advantage over whole-genome or shotgun sequencing in terms of cost and time. In addition, RADseq methods as currently implemented using NGS typically share the following characteristics:

**Relatively short sequence reads. —** While not a necessary feature of RADseq, short sequence reads (50–100 bp) are characteristic of most RADseq projects,

the lion's share of which are sequenced on an Illumina platform (though see Etter & al., 2011 for use of paired-end sequencing to assemble much longer loci from short NGS sequencing reads). With continuing increases in sequencing read length and sequencing capacity (cf. Mardis, 2011; Liu & al., 2012), this characteristic of RADseq projects will presumably relax somewhat.

**Wide genomic distribution. —** RADseq data are widely distributed in the genome (cf. Miller & al., 2007a, b, Baird & al., 2008; Davey & al., 2011; McCluskey & Postlethwait, 2014) and thus are expected to provide a genome-wide view of phylogeny. As the data are sequence-based, RADseq loci can be mapped back to genetic or physical maps if appropriate sequence-based references are available, potentially allowing researchers to assess the genomic distribution of divergence and introgression in a phylogenetic context.

**Lack of distinction, at the outset of the project, between paralogous and orthologous sequences. —** RADseq differs in this respect from NGS methods based on targeted enrichment of predetermined loci (Faircloth & al., 2012; Weitemier & al., 2014). In the absence of reference genome or similar resource, orthology must be assessed informatically (e.g., Catchen & al., 2011; Lu & al., 2013; Eaton, 2014; see "Assembling datasets for phylogenetic analysis" below).

**Dropout of loci and alleles. —** RADseq datasets are expected to be incomplete for the samples sequenced in any given experiment for two primary reasons: sampling error (stochastic failure of a locus to be genotyped due to low read coverage) and the disruption of restriction sites by mutation. The latter results in a pattern of decline in locus-sharing with phylogenetic distance, documented in several studies (Rubin & al., 2012; Cariou & al., 2013; Hipp & al., 2014; Viricel & al., 2014).

Locus dropout is one of the most vexing RADseq problems, and the one expected to limit its utility at the deepest phylogenetic scales. If dropout is primarily driven by restriction site loss, then the pattern of missing data should be phylogenetically informative; however, no attempts have yet been made to incorporate this process into models or optimality criteria for tree inference. Locus dropout had no detectable effect on tree topology or branch lengths in Hipp & al. (2014), but Arnold & al. (2013) found that intermediate amounts of locus dropout are associated with biased estimates of higher genetic diversity (estimated using $\pi$ and $\theta$) and deeper coalescence times. This effect should be investigated further, both to understand how strong the effect is in ideal (simulated) cases and to understand how much it biases real-world estimates of divergence times. In cases in which locus dropout is not strongly systematic (cf. Simmons, 2012a, b; Roure & al., 2013), phylogenetic inference may not be as badly affected by locus and allele dropout as are the parameters estimated in population genetics and linkage mapping (cf. Gautier & al., 2013; Arnold &

al., 2013; Huang & Knowles, 2014). Where locus dropout is systematic, some of the recommendations in Simmons (2012a, b) may be helpful in evaluating the effects, if any, on phylogenetic inferences.

The typical RADseq dataset is, in summary, large, incomplete, and composed of relatively short anonymous loci that individually are expected to have little phylogenetic signal (i.e., insufficient variation for a fully resolved gene tree). Such datasets are rather foreign in character compared to those more typically favored in systematics, i.e., complete alignments of separate loci that each tells a resolvable story. Knowing what to do with RADseq data—including assessing its phylogenetic versus population genetic signal—is an interesting challenge.

## Assembling datasets for phylogenetic analysis

We consider here the general goal of producing a matrix containing a single orthologous sequence for each RAD locus from each organism sampled, with individuals being the terminal units of analysis. For most clades, a high-quality reference genome is not available as a bioinformatic resource, despite increasing availability of whole-genome assemblies across the tree of life (Lyons & al., 2015). Under these circumstances, the basic steps are as follows:

**1) Locus identification and genotyping. —** Within each individual, quality-filtered NGS reads must be clustered by similarity into groups representing putative loci for base-calling (genotyping), in which statistical corrections for sequencing errors are applied. Genotyping errors may also be caused by repetitive elements and polyploidy, requiring the use of tools such as UNEAK (Lu & al., 2013), a software pipeline that assembles single-mismatch networks to identify paralogous RAD sequences. Alternatively, genotyping errors can be estimated directly if individuals from a mapping population are analyzed alongside the broader phylogenetic sample (e.g., Henning & al., 2014), but this is often not the case for phylogenetic datasets. For phylogenetics, a consensus sequence for each locus is generally sufficient. If haplotype data are desired for population genetic analyses, randomly chosen alleles could be used in the phylogenetic matrix.

**2) Orthology estimation. —** Across individuals, locus genotypes must then be clustered into putatively orthologous groups and aligned. Clusters in which any single individual is represented by multiple sequences, which can arise if loci are duplicated or repeated in the genome without substantial sequence divergence, should be discarded (Eaton, 2014). There currently exists no objective method for optimizing the clustering threshold. If too relaxed, there

is greater risk of clustering paralogs (though paralogs may be filtered out by removing loci with more than two alleles or loci that are heterozygous for more than a threshold number of individuals at a given site; Eaton, 2014). If too stringent, a single-copy locus that is evolutionarily diverged across taxa is more likely to be erroneously split and appear as separate loci, similar in sequence, with each having incomplete and complementary taxon sampling (Rubin & al., 2012; Eaton, 2014).

**3) Locus filtering based on taxon coverage. —** A minimum of four leaves are needed for a phylogenetically informative unrooted tree, so loci having fewer taxa should generally be discarded. The "minimum taxa" parameter controls the amount of missing data in the final matrix, with a trade-off between the proportion of missing data and the number of loci included. Several studies (Rubin & al., 2012; Wagner & al., 2013; Hipp & al., 2014) found that larger amounts of missing data do not cause substantial problems for phylogenetic inference, and in fact may be preferable due to the larger number of loci included.

A variety of bioinformatic approaches have been used to implement these steps. In the first interspecific phylogeny inferred from RADseq data, an *in silico* study on existing genomes of *Drosophila*, mammals, and fungi (Rubin & al., 2012), sequences were assumed known without error, and allelic data were not available, eliminating the need for step 1 above. The study used UCLUST (Edgar, 2010) for clustering and MUSCLE (Edgar, 2004) for alignment in custom scripts to assemble concatenated RADseq matrices over a range of parameter values for sequence similarity and minimum taxonomic coverage per locus. A similar study of *Drosophila* only by Cariou & al. (2013) compared UCLUST, SiLiX (Miele & al., 2011), and BLASTN (Altschul & al., 1990) in their effectiveness for orthology estimation, finding that the latter two outperformed the first.

The first complete software pipeline for RADseq analysis, Stacks (Emerson & al., 2010; Hohenlohe & al., 2010; Catchen & al., 2011), has been used more than any other to generate RADseq data matrices for phylogenetic studies (e.g., Jones & al., 2013; Lexer & al., 2013; Reitzel & al., 2013; Wagner & al., 2013; Cruaud & al., 2014; Herrera & al., 2014; Leaché & al., 2014; Viricel & al., 2014; Pante & al., 2015). Originally designed for genetic mapping, Stacks has since been extended to include features for use in population genetics and phylogenetics. The software is self-sufficient and implements an off-by-*N* clustering strategy that ignores indels and may therefore be best suited to very fine-scale phylogenetic questions (see discussion in Eaton, 2014). Other off-by-*N* RADseq pipelines that have been less used are Rainbow (Chong & al., 2012),

RADtools (Baxter & al., 2011; used in Roda & al., 2013), and most recently, AftrRAD (Sovic & al., 2015), which differs from the others in allowing indels among alleles.

PyRAD (Eaton & Ree, 2013; Eaton, 2014) is the only RADseq pipeline developed to date that expressly targets phylogenetics as the intended application. Unlike Stacks, PyRAD utilizes external software for locus identification and orthology estimation (USEARCH or VSEARCH; https://github.com/torognes/vsearch) and multiple alignment (MUSCLE). In a simulation study, these clustering and alignment algorithms were less prone to locus splitting than Stacks (Eaton, 2014). Pante & al. (2015) compared PyRAD and Stacks in an empirical analysis of octocorals and found that PyRAD returned matrices with a greater number of loci that resolved more nodes of the phylogeny. PyRAD also implements hierarchical clustering (i.e., recursively clustering within and between clades) as a strategy for increasing the phylogenetic breadth of assembled matrices, and *D*-statistic genomic introgression tests (Green & al., 2010; Durand & al., 2011), including an extension developed to distinguish between current and ancestral introgression (Eaton & Ree, 2013; Eaton, 2014). PyRAD has been used in a relatively small but growing number of studies (Escudero & al., 2014; Herrera & al., 2014; Hipp & al., 2014; Pante & al., 2015; Takahashi & al., 2014; Ebel & al., 2015 used Stacks for preprocessing, PyRAD for final clustering). One other software pipeline, rtd (Peterson & al., 2012), utilizes graph clustering in lieu of the off-by-*N* approach taken by Stacks. That study finds similar increases in clustering sensitivity from the graph clustering approach relative to Stacks.

Several RADseq phylogenetic studies take a different approach to locus identification and orthology estimation: they map their sequence reads to genomic data in lieu of clustering (e.g., Hyma & Fay, 2013; Nadeau & al., 2013; Reitzel & al., 2013; McCluskey & Postlethwait, 2014). These studies begin with genomic scaffolds or a pseudogenome of assembled RADseq reads (e.g., Gompert & al., 2014; Rheindt & al., 2014), and map their sequence reads to this reference using alignment software such as Bowtie (Langmead & al., 2009), Stampy (Lunter & Goodson, 2011), or BWA (Li & Durbin, 2009), with post-mapping analysis and SNP-calling using genotyping software such as GATK (McKenna & al., 2010) or SAMtools (Li & al., 2009). This approach has the potential benefit of increasing confidence about orthology, but also risks discarding informative data or failing to detect paralogy if the genomic reference is incomplete. Several researchers have taken a hybrid approach, using available genomic data to prescreen sequence data, then applying the de novo approaches to locus identification described above. Wagner & al. (2013), for example, used Stacks initially, then mapped reads back to consensus sequences

to find those matching only one inferred locus, and finally performed SNP calling in GATK. Reitzel & al. (2013) took the opposite approach: they used Stacks *after* filtering their data by alignment against a reference genome using Bowtie. Note that this approach to generating a phylogenetic matrix is fundamentally different than post-clustering analysis of RADseq data in a genomic context, which also typically entails mapping RADseq loci to a genomic scaffold (see below, "Analyzing RADseq phylogenetic data in a genomic context").

What approach should one use? While mapping sequence reads back to a relatively complete genomic reference seems like an excellent way to identify loci and toss out potential paralogs, it is not immediately clear at what point having only partial coverage of the genome becomes more a hindrance than a help. If no reference genome is available, clustering methods such as PyRAD or rtd, which utilize global alignment and graph clustering respectively, are the most appropriate for phylogenetic datasets. Because of its ongoing development and focus on phylogenetics, at this point we consider PyRAD to be the pipeline of choice for RADseq phylogenetics. Additional studies comparing loci inferred using different methods (e.g., Pante & al., 2015) would help clarify their respective biases. In addition, the issues raised above regarding locus assembly and the effects of non-random missing data suggest the need for sensitivity analyses, e.g., in how the choice of clustering threshold simultaneously influences the dimensions, completeness, and distribution of evolutionary rates in the final matrix (Takahashi & al., 2014). Prudent researchers should generally investigate a range of data matrices assembled under different parameters.

## Inferring phylogeny from RADseq data

Among phylogenetic data types, RAD sequences have a unique combination of features: (1) they typically encompass a large number of loci (usually tens of thousands) relative to targeted enrichment approaches that sample fewer loci (hundreds to thousands) but yield longer sequences; (2) they are expected to sample very broadly from the genome; (3) missing data are inherent to the method and are distributed non-randomly with respect to phylogeny; (4) as sequences, they allow inferences from nucleotide substitution models, but the length of each individual locus is relatively short; and (5) they can also be processed to resolve alleles and heterozygosity, theoretically allowing both phylogenetic and population genetic inferences from the same dataset.

These characteristics place certain constraints on what methods of phylogenetic inference are applicable. In particular, having only short sequences

and missing data make it generally difficult to use multilocus methods that rely on resolved and complete gene trees, including some species-tree methods (e.g., Kubatko & al., 2009; Liu & al., 2009; Heled & Drummond, 2010) and concordance analysis (Baum, 2007; Ané & al., 2007). This is perhaps why most studies have opted for what is arguably the simplest approach, namely to infer a tree from a concatenated matrix of RADseq alignments. In all RADseq phylogeny studies we have seen to date, this has yielded highly resolved trees with strong clade support compared to comparable analyses of Sanger sequences (e.g., Wagner & al., 2013; Eaton & Ree, 2013; Escudero & al., 2014; Herrera & al., 2014; Hipp & al., 2014).

The large size of concatenated alignments, which can be megabases in length (e.g., Eaton & Ree, 2013; Wagner & al., 2013; Hipp & al., 2014; Takahashi & al., 2014), exacts computational demands that limit software choices for phylogeny reconstruction. At present RAxML (Stamatakis, 2014) seems the most capable at this scale, and is consequently the most frequently used. In all examples to date, none have attempted to partition the concatenated matrix: there seems to be no reasonable way to do it, given the number of loci and absent knowledge of codon reading frames, etc. In addition, statistical selection of a substitution model is generally dispensed with in favor of defaulting to the parameter-rich GTR+gamma model; in the only explicit test of which we are aware, McCluskey & Postlethwait (2014) found support for additionally including the invariant sites parameter, but this may be statistically superfluous (Yang, 2006: 113–114). And while some studies filter out invariant sites (e.g., Coghill & al., 2014), likelihood-based tree inference does not typically condition on all characters being variable, and thus may reconstruct biased branch lengths and topologies if invariant sites are excluded (see discussions in Felsenstein, 1992 and Lewis, 2001). Using RADseq loci in their entirety, including invariant sites, eliminates this ascertainment bias.

The concatenation approach is simple and convenient, but it obviously sweeps several sources of error under the rug, such as variation across loci in genealogy (due to stochastic coalescence, introgression, etc.), substitution parameters, and evolutionary rate; a clear summary of the main issues can be found in Wagner & al. (2013). One concern that has received particular attention is statistical inconsistency: namely, the potential in coalescent models for the most common gene tree to differ from the species tree, leading to convergence on an incorrect topology as loci are concatenated (e.g., Kubatko & Degnan, 2007; Degnan, 2013; Rosenberg, 2013; Roch & Steel, 2014). In this context, the simplifying assumptions of concatenation—in particular, failing to account for stochastic coalescence of unlinked loci—are seen as flying sufficiently in the face of reality as to render statistical interpretation of the

inferred tree moot (Rannala & Yang, 2008). However, while consistency is obviously the ideal in phylogenetic inference, in this case the theoretical basis for anomalous gene trees (prior probabilities associated with tree symmetry in the coalescent) and the empirical circumstances in which they arise (successive short internodes with large effective population size) are rather well known (Rosenberg, 2013), which suggests that potential errors could be identified post hoc with relative ease. Pending more sophisticated diagnostic tests, it seems prudent in the meantime to be at least skeptical of high support given to short internal branches in concatenation-derived trees.

The clearest reason to not concatenate may simply be that it does not estimate species-tree parameters relating to ancestral demography (effective population size, or branch "width") in addition to tree topology and branch lengths. That is to say, given that RADseq data can capture demographic parameters in the present, it seems natural to pursue phylogenetic inferences of the same parameters in the past. This raises a conceptual question: in ignoring demography and violating its assumptions, how does a concatenated-analysis RADseq tree relate to the species tree? Our view is that, in the absence of current or ancestral gene flow or hybridization, it reflects the genealogical central tendency of the sampled genomes and can be expected to accurately trace the species tree except in more or less predictable circumstances (e.g., Bayzid & Warnow, 2013). We stress, however, that this needs more attention, especially with respect to the effects of phylogenetically structured missing data. Concatenation is a valuable heuristic tool for extracting phylogenetic signal from interspecific RADseq data, providing at the very least an empirical framework on which to base further analyses, such as tests for introgression (e.g., using "ABBA/BABA" tests; see Eaton & Ree, 2013, where concatenation yielded an anomalous topology; also Escudero & al., 2014; Rheindt & al., 2014; Streicher & al., 2014).

It is worth noting that criticisms of concatenation have not yet been systematically investigated specifically in the context of RADseq data. Such studies are needed, but we are cautiously optimistic that they will not prove fatal to maximum likelihood on a concatenated RADseq matrix as a useful phylogenetic inference method. For example, the effects of heterogeneity in the evolutionary process within and between loci may be mitigated by the fact that RADseq loci, as previously mentioned, are many in number, short in sequence, and broad in genomic distribution, and may thus more closely represent a genome-wide average than datasets of fewer genes with longer sequences, as are typical of targeted enrichment methods. That is, the idiosyncratic effects of any single RAD locus are minimized. Moreover, the process of assembling RAD loci across samples based on sequence similarity

should automatically select against those evolving at high rates (e.g., having a preponderance of saturated sites), and should lead to reduced variation in the effective evolutionary rate across loci. This is an area ripe for investigation, particularly as rate heterogeneity, and high-rate sites in particular, have been implicated in topology estimation errors arising from concatenation of longer gene sequences (Xi & al., 2013, 2014).

## Inferring the species tree

The issues raised above lead naturally to the question: can RADseq data be used to infer species trees? The short answer is that while case studies are few at this point, in general it seems that current methods are not particularly well-suited for RADseq data; however, this is an active area of research in which improvements seem inevitable and imminent.

Eaton & Ree (2013) and McCluskey & Postlethwait (2014) used BUCKy to infer primary concordance trees and population trees, the latter being expected to converge on the species tree if gene tree discordance is entirely due to stochastic coalescence. BUCKy works with posterior samples of gene trees, with the expectation that each locus has enough signal to support a reasonably well-resolved gene tree, and in its current version requires all taxa to be sampled for all loci. Here again, RADseq data are hampered by locus dropout (missing taxa from most loci) and short sequence lengths, both factors that can drastically reduce the amount of data usable by this method. For example, out of 42,235 loci with at least four individuals in Eaton & Ree (2013), only 945 had both complete sampling of the ingroup (11 individuals) and at least two phylogenetically informative sites.

The only demography-oriented species-tree method used with RADseq data to date is SNAPP (Bryant & al., 2012), which uses a coalescent finite-sites model to calculate the likelihood of unlinked, biallelic markers given a species tree, allowing mutations but notably without explicitly considering individual gene trees. It has been applied to RADseq-derived SNPs in sister species of irises (Hamlin & Arnold, 2014), one species complex in birds (Rheindt & al., 2014) and two species complexes in frogs (Leaché & al., 2014; Streicher & al., 2014). In all cases, the empirical focus was on population-level sampling of very closely related taxa to test hypotheses of hybridization, admixture, introgression, and species limits, respectively.

Assuming that RAD loci are unlinked, using SNAPP requires selecting at most one binary-variable site from each locus alignment, which can mean discarding a potentially substantial amount of phylogenetically informative

data. For example, in a reassembly of the data from Eaton & Ree (2013), 45,668 RAD loci contain 183,980 variable homozygous sites; selecting one binary SNP from each locus yields 41,088 sites—still a large number, but the matrix is only 66% complete due to locus dropout. The current version of SNAPP allows missing data, but it is not yet clear how such a relatively large proportion impacts results or performance, particularly as the number of taxa increases. Moreover, its limits in terms of computational feasibility, and the maximum phylogenetic depth at which inferences of topology and demography are possible, particularly in comparison to concatenation, are also not well known at this point.

If a statistically consistent species-tree topology (without branch lengths or widths) is the primary goal, then a promising method for RADseq data is SVDquartets (Chifman & Kubatko, 2014), which calculates optimal quartet relationships for unlinked nucleotide site patterns (i.e., is not constrained to binary values) under the coalescent. The quartets can then be assembled into a species-tree topology, e.g., using Quartets MaxCut (Swenson & al., 2011; Snir & Rao, 2012). While the performance of SVDquartets has not yet been tested in empirical RADseq studies, simulations of longer gene sequences indicate it may be relatively robust to violations of linkage, implying that all variable sites in a RADseq matrix could be used; it is also apparently tolerant of missing data and computationally tractable for large numbers of loci and taxa (Chifman & Kubatko, 2014).

It is worth noting that demographic inference of treelike histories below what might be called the "species level" is also possible with RADseq data using a variety of methods that use allele frequencies and model genetic drift rather than mutation. These include TreeMix (Pickrell & Pritchard, 2012) and DADI (Gutenkunst & al., 2009). A notable example in this context is Gompert & al. (2014), who used TreeMix with RADseq-derived SNP data to derive an ancestral population graph of divergence and introgression events from an extensive sample (1536 individuals, 66 populations) of three sibling species of *Lycaeides* butterflies. At this level of analysis, an important consideration is the extent to which genetic covariances across populations reflect equilibrium processes of migration versus ancestral splits and mixtures (Felsenstein, 1992; Pickrell & Pritchard, 2012).

Species-tree inference using RADseq data is still a nascent field that can be seen as attempting to bridge, or at least straddle, the phylogenetic/population genetic divide. The economies of RADseq that increasingly allow studies to simultaneously sample across clades, species, and populations will fuel the impetus for unified inferences, particularly as researchers look to RADseq data with goals like species delimitation in mind (e.g., Leaché & al., 2014).

At present, in the "no-man's land" of species complexes and the like, where lineage sorting, introgression, and hybridization may each have significant effects, it is not always entirely clear what kind of inference method is most appropriate, as reflected in the variety of approaches taken in recent studies. Going forward, it seems likely that the most effective way to resolve the full complement of topology, divergence times, and ancestral demographics will continue to involve a combination of methods and a hierarchical approach.

## Analyzing RADseq phylogenetic data in a genomic context

RADseq loci have an advantage over other semi-anonymous markers, like AFLPs, in that they can be mapped back to a reference genetic resource such as an assembled genome, genetic (linkage) map, physical map, or EST library. Alternatively, the same RADseq marker system used to infer a phylogeny could be used to generate a linkage map (e.g., Baird & al., 2008; Baxter & al., 2011; Lu & al., 2013; Henning & al., 2014), without requiring that the RADseq loci be mapped back to external genomic resources. Integration of genomic and phylogenetic data is key to investigating the genomic architecture of diversification, and RADseq may allow researchers to undertake such integrative studies with fewer genomic resources than has previously been needed. The limiting factors in such studies are likely to be the level of detail and genomic distribution of genetic source data, and the coverage of the RADseq loci being investigated.

Investigating the genomic architecture of divergence, hybridization, and introgression depends on having a genetic resource of sufficient resolution and detail to address one's question, combined with a sufficient number of informative RADseq markers to map back to that resource (cf. The Heliconius Genome Consortium, 2012 for an example using full genomic data). Rheindt & al. (2014), for example, used the ABBA/BABA test (Green & al., 2010; Durand & al., 2011) to test alternative hypotheses about the origin of mosaic populations in flycatchers, then mapped introgressed flycatcher genome contigs onto the zebra finch genome to identify regions of the genome that are disproportionately introgressed. Nadeau & al. (2013) mapped raw sequence reads back to *Heliconius melpomene* genome scaffolds and used map positions to test a genomic-islands-of-divergence hypothesis and the hypothesis of introgression at wing-color loci. Even unmapped genomic scaffolds can be useful for such studies: Roda & al. (2013) mapped RADseq loci back to genomic scaffolds to identify neutral loci that were linked to outlier loci in a population comparison analysis, then compared topologies of the linked neutral loci vs.

unlinked neutral loci to investigate patterns of gene flow between ecologically divergent parapatric populations. It may turn out that this new power to detect introgression from large numbers of RADseq loci may come with a new challenge of teasing apart phylogeny from introgression: introgression may hide behind very high branch supports in RADseq datasets due to the large number of nucleotides used for inference, whereas in datasets composed of fewer loci, the conflicting phylogenetic signal due to introgression more often reduces branch support. We expect the integration of genomic and phylogenetic RADseq data to help make this separation more cleanly.

Without linkage-mapping data or sufficiently long genomic contigs to identify linkage relationships between RADseq loci, RADseq loci can still be BLASTed to EST libraries to identify putative orthologs, and gene-ontology (GO) analysis can be used to characterize distributions of gene functions. Rheindt & al. (2014) utilized GO-term analysis on mapped RADseq loci to determine whether introgressed genomic contigs draw non-randomly from gene functions, inferring tentatively that introgression may carry alleles influencing plumage coloration. Reitzel & al. (2013) used a combination of linkage mapping and GO analysis to characterize genes under balancing selection in a cnidarian phylogeographic study. Hipp & al. (2014) mapped RADseq loci back to EST libraries to identify a pool of loci that sample disproportionately from gene-rich regions of the genome: their analyses suggest that despite stronger constraints on and lower homoplasy of those loci, EST-linked RADseq loci are no more effective in their dataset at resolving deeper phylogenetic nodes. Such analyses rely on well-characterized databases of gene functions to which the RADseq loci can be mapped.

In our experience, the resolution and informativeness of the genetic resource, once it is intersected with the RADseq phylogenetic dataset, will often not be apparent until after analyses have been undertaken. It thus behooves researchers to characterize clearly what they expect they will need from non-RADseq genetic resources to address their questions, and then to scale their inquiry to avail themselves of what can be inferred from joint analysis of the resources at hand. One should anticipate adjusting one's expectations along the way in such analyses. Moreover, the completeness of the RADseq phylogenetic matrix should be expected to influence what questions can be addressed, and at what scale. Questions about the genes involved in introgressive gene flow, for example, can only be addressed using loci with sufficient sampling of the individuals hypothesized to be exchanging alleles; a seemingly enormous matrix of 30,000 or more RADseq loci may in fact be insufficient to address hypotheses that require mapped loci that are complete or nearly complete for a specific subset of the individuals sampled. While

RADseq datasets will certainly tend to be very large, perhaps unnecessarily large for some phylogenetic questions, they may be only barely large enough at times for studies linking RADseq phylogenetic markers to genomic data.

## The future of RADseq phylogenetics

RADseq is clearly emerging as a cost-effective and low-overhead method of easily surveying genomes for SNPs that segregate across the full spectrum of individuals, populations, and clades. For systematics, this represents an exciting new frontier, especially if the treasure trove of historical herbarium/museum collections can be opened wider to genome-scale phylogenetic inquiry with improvements in DNA sample preparation through whole-genome amplification (Blair & al., 2015). In our own labs, we have had reasonable success with RADseq using relatively young herbarium material, 10–15 years old in *Pedicularis* and up to 50 years old in *Carex* (unpub. data), supporting an experimental study demonstrating that RADseq can be successful even with moderately degraded DNA (Graham & al., 2015). However, for older specimens, tropical plants, or species in which dried material typically does not yield high amounts of high molecular weight DNA, the outlook may be less rosy.

Nevertheless, RADseq seems well positioned to provide empirical fuel for greater conceptual and methodological convergence, bringing phylogenetics, phylogeography, and population/evolutionary genetics into the same fold. It will provide new opportunities to infer branching descent and historical demographics from the same dataset, and new power to address questions relating to the interactions between adaptation, gene flow, and divergence. In addition, increasingly dense sampling within and between populations and across clades should spur activity relating to species delimitation and species concepts. We fully expect the results to challenge "simplistic notions concerning the organization of biological diversity into discrete, easily delineated and hierarchically structured entities" (Gompert & al., 2014: 4570).

As genome-scale phylogeny reconstruction rises to the mainstream, most methodological research to date has focused on the use of multiple gene trees to infer the species tree (e.g., Rannala & Yang, 2008; McCormack & al., 2013; Liu & al., 2015). By contrast, relatively little attention has been paid to the treatment of RADseq data and its widely distributed but incompletely sampled loci. As we have seen, much remains unknown about the performance and limits of species-tree inference with RADseq data, a deficit that demands attention if for no other reason than that interspecific RADseq datasets are

proliferating at a rapid pace. That is to say, the ease with which RADseq data can be acquired for non-model organisms—inexpensively, without prior identification of loci—ensures it an important role in systematics, at least in the near term, even if the data are less than ideal from a theoretical point of view.

We not only predict that RADseq will become more common in systematics, but that the data will continue to exhibit more phylogenetic signal than initially expected. It is clear that RADseq should not continue to be viewed primarily as a tool for population-level inferences (cf. Lemmon & Lemmon, 2013; McCormack & al., 2013). The growing number of case studies highlighted above convincingly illustrates that loci are conserved across surprisingly deep divergences. For example, RADseq analyses of *Drosophila* resolve nodes back to about 40 Ma (Rubin & al., 2012; Cariou & al., 2013), and in barnacles to about 68 Ma (Herrera & al., 2014). In our studies of oaks (*Quercus* spp.), we are able to resolve nodes dated to approximately 60 Ma (J. McVay, M. Deng, P. Manos, J. Cavender-Bares, and A. Hipp, unpub. data). While we do not expect these ranges to be universal across taxa, they argue strongly for the viable candidacy of RADseq as a phylogenetic data source for clades firmly embedded in the Cenozoic, though perhaps not any older.

An issue that has received little attention, but surely will rise in importance, is that of RADseq data reusability. It is common for new molecular phylogenies to be based partly (or in some cases entirely) on previously published Sanger sequences; it seems abundantly clear that the availability of well-characterized, commonly sequenced loci facilitates complementation of effort and taxonomic sampling, and has contributed substantially to progress and synthesis in systematics as a whole. Can RADseq data have the same lasting value? To what extent will RADseq data be re-used in new analyses of expanded or complementary taxonomic samples? Clearly, this would require employing the same restriction site(s), and likely the same laboratory protocol. One study (Hipp & al., 2014) investigated the combinability of RADseq data from different sequencing runs in different years, in which sequencing lengths varied from 60 bp in 2010 to 95 bp in 2012 and an average of $5.58 \times 10^5$ and $3.06 \times 10^6$ raw sequence reads per individual respectively. In this study, we found no effect on phylogenetic topology and almost no effect on branch lengths, despite the fact that pairs of technical replicates overlapped by only 43% to 64% of loci (compared against the union set of loci). Moreover, both authors of this chapter have combined RADseq data from numerous library preparations and sequencing runs from 2010 through 2014 and found no obvious effect on topology. However, in this case, the library preparations were all conducted in the same laboratory, using precisely the same protocol. Overlap in locus recovery using different protocols has not been studied in

any detail. Additionally, RADseq analysis has proceeded under a diversity of protocols, with no single emerging standard. For phylogenetics, standardization on a protocol might facilitate greater reuse of data, increasing its value over the longer term.

## Acknowledgments

The authors acknowledge collaborators in their RADseq phylogenetics work, particularly D.A.R. Eaton and J. McVay, for insights into the nature of RADseq phylogenetics. RR was supported in part by NSF Award DEB-1119098 and AH was supported in part by NSF Award DEB-1146488.

## Literature cited

**Altschul, S.F., Gish, W., Webb, M., Myers, E.W. & Lipman, D.J.** 1990. Basic Local Alignment Search Tool. *J. Molec. Biol.* 215: 403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2

**Andrews, K.R. & Luikart, G.** 2014. Recent novel approaches for population genomics data analysis. *Molec. Ecol.* 23: 1661–1667. http://dx.doi.org/10.1111/mec.12686

**Andrews, K.R., Hohenlohe, P.A., Miller, M.R., Hand, B., Seeb, J.E. & Luikart, G.** 2014. Trade-offs and utility of alternative RADseq methods. *Molec. Ecol.* 23: 5943–5946. http://dx.doi.org/10.1111/mec.12964

**Ané, C., Larget, B., Baum, D.A., Smith, S.D. & Rokas, A.** 2007. Bayesian estimation of concordance among gene trees. *Molec. Biol. Evol.* 24: 412–426. http://dx.doi.org/10.1093/molbev/msl170

**Arnold, B., Corbett-Detig, R.B., Hartl, D. & Bomblies, K.** 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molec. Ecol.* 22: 3179–3190. http://dx.doi.org/10.1111/mec.12276

**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. http://dx.doi.org/10.1371/journal.pone.0003376

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

**Baxter, S.W., Davey, J.W., Johnston, J.S., Shelton, A.M., Heckel, D.G., Jiggins, C.D. & Blaxter, M.L.** 2011. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLOS ONE* 6: e19315. http://dx.doi.org/10.1371/journal.pone.0019315

**Bayzid, M.S. & Warnow, T.** 2013. Naive binning improves phylogenomic analyses. *Bioinformatics* 29: 2277–2284. http://dx.doi.org/10.1093/bioinformatics/btt394

**Beissinger, T.M., Hirsch, C.N., Sekhon, R.S., Foerster, J.M., Johnson, J.M., Muttoni, G., Vaillancourt, B., Buell, C.R., Kaeppler, S.M. & De Leon, N.** 2013. Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics* 193: 1073–1081. http://dx.doi.org/10.1534/genetics.112.147710

**Blair, C., Campbell, C.R. & Yoder, A.D.** 2015. Assessing the utility of whole genome amplified DNA for next-generation molecular ecology. *Molec. Ecol. Resources*. http://dx.doi.org/10.1111/1755-0998.12376

**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. *Molec. Biol. Evol.* 29: 1917–1932. http://dx.doi.org/10.1093/molbev/mss086

**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. http://dx.doi.org/10.1002/ece3.512

**Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W. & Postlethwait, J.H.** 2011. Stacks: Building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genet.* 1: 171–182. http://dx.doi.org/10.1534/g3.111.000240

**Cavender-Bares, J., Gonzalez-Rodríguez, A., Eaton, D.A.R., Hipp, A.A.L., Beulke, A. & Manos, P.S.** 2015. Phylogeny and biogeography of the American live oaks (*Quercus* subsection *Virentes*): A genomic and population genetics approach. *Molec. Ecol.* 24: 3668–3687. http//dx.doi.org/10.1111/mec.13269

**Chifman, J. & Kubatko, L.** 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30: 3317–3324. http://dx.doi.org/10.1093/bioinformatics/btu530

**Chong, Z., Ruan, J. & Wu, C.-I.** 2012. Rainbow: An integrated tool for efficient clustering and assembling RAD-seq reads. *Bioinformatics* 28: 2732–2737. http://dx.doi.org/10.1093/bioinformatics/bts482

**Coghill, L.M., Hulsey, D.C., Chaves-Campos, J., García de Leon, F.J. & Johnson, S.G.** 2014. Next generation phylogeography of cave and surface *Astyanax mexicanus*. *Molec. Phylogen. Evol.* 79: 368–374. http://dx.doi.org/10.1016/j.ympev.2014.06.029

**Cruaud, A., Gautier, M., Galan, M., Foucaud, J., Sauné, L., Genson, G., Dubois, E., Nidelet, S., Deuve, T. & Rasplus, J.-Y.** 2014. Empirical assessment of RAD sequencing for interspecific phylogeny. *Molec. Biol. Evol.* 31: 1272–1274. http://dx.doi.org/10.1093/molbev/msu063

**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. http://dx.doi.org/10.1038/nrg3012

**Davey, J.W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K. & Blaxter, M.L.** 2013. Special features of RAD sequencing data: Implications for genotyping. *Molec. Ecol.* 22: 3151–3164. http://dx.doi.org/10.1111/mec.12084

**Degnan, J.H.** 2013. Anomalous unrooted gene trees. *Syst. Biol.* 62: 574–90. http://dx.doi.org/10.1093/sysbio/syt023

**Durand, E.Y., Patterson, N., Reich, D. & Slatkin, M.** 2011. Testing for ancient admixture between closely related populations. *Molec. Biol. Evol.* 28: 2239–2252. http://dx.doi.org/10.1093/molbev/msr048

**Eaton, D.A.R.** 2014. PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* 30: 1844–1849. http://dx.doi.org/10.1093/bioinformatics/btu121

**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. http://dx.doi.org/10.1093/sysbio/syt032

**Eaton, D.A.R., Hipp, A.L., Gonzalez-Rodriguez, A. & Cavender-Bares, J.** 2015. Introgression obscures and reveals historical relationships among the American live oaks. *Evolution* (preprint). http://dx.doi.org/10.1101/016238

**Ebel, E.R., DaCosta, J.M., Sorenson, M.D., Hill, R.I., Briscoe, A.D., Willmott, K.R. & Mullen, S.P.** 2015. Rapid diversification associated with ecological specialization in Neotropical *Adelpha* butterflies. *Molec. Ecol.* 24: 2392–2405. http://dx.doi.org/10.1111/mec.13168

**Edgar, R.C.** 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *B. M. C. Bioinf.* 5: 113. http://dx.doi.org/10.1186/1471-2105-5-113

**Edgar, R.C.** 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461. http://dx.doi.org/10.1093/bioinformatics/btq461

**Edwards, S.V.** 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63: 1–19. http://dx.doi.org/10.1111/j.1558-5646.2008.00549.x

**Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. & Mitchell, S.E.** 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLOS ONE* 6: e19379. http://dx.doi.org/10.1371/journal.pone.0019379

**Emerson, K.J., Merz, C.R., Catchen, J.M., Hohenlohe, P.A., Cresko, W.A., Bradshaw, W.E. & Holzapfel, C.M.** 2010. Resolving postglacial phylogeography using high-throughput sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 107: 16196–16200. http://dx.doi.org/10.1073/pnas.1006538107

**Escudero, M., Eaton, D.A.R., Hahn, M. & Hipp, A.L.** 2014. Genotyping-by-sequencing as a tool to infer phylogeny and ancestral hybridization: A case study in *Carex* (Cyperaceae). *Molec. Phylogen. Evol.* 79: 359–367. http://dx.doi.org/10.1016/j.ympev.2014.06.026

**Etter, P.D., Preston, J.L., Bassham, S., Cresko, W.A. & Johnson, E.A.** 2011. Local de novo assembly of RAD paired-end contigs using short sequencing reads. *PLOS ONE* 6: e18561. http://dx.doi.org/10.1371/journal.pone.0018561

**Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T. & Glenn, T.C.** 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* 61: 717–726. http://dx.doi.org/10.1093/sysbio/sys004

**Felsenstein, J.** 1992. Phylogenies from restriction sites: A maximum-likelihood approach. *Evolution* 46: 159–173. http://dx.doi.org/10.2307/2409811

**Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., Cornuet, J.-M. & Estoup, A.** 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molec. Ecol.* 22: 3165–3178. http://dx.doi.org/10.1111/mec.12089

**Gompert, Z., Lucas, L.K., Buerkle, C.A., Forister, M.L., Fordyce, J.A. & Nice, C.C.** 2014. Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molec. Ecol.* 23: 4555–4573. http://dx.doi.org/10.1111/mec.12811

**Gonzalez, L.A.** 2014. *Phylogenetics and mating system evolution in the southern South American Valeriana (Valerianaceae).* M.S. Thesis, Department of Biological Sciences, University of New Orleans, Louisiana, U.S.A.

**Graham, C.F., Glenn, T.C., McArthur, A.G., Boreham, D.R., Kieran, T., Lance, S., Manzon, R.G., Martino, J.A., Pierson, T., Rogers, S.M., Wilson, J.Y. & Somers, C.M.** 2015. Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). *Molec. Ecol. Resources*. http://dx.doi.org/10.1111/1755-0998.12404

**Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.-S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V.B.,**

**Golovanova, L.V., Lalueza-Fox, C., De la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. & Pääbo, S.** 2010. A draft sequence of the Neandertal genome. *Science* 328: 710–722. http://dx.doi.org/10.1126/science.1188021

**Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D.** 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLOS Genet.* 5: e1000695. http://dx.doi.org/10.1371/journal.pgen.1000695

**Hamlin, J.A.P. & Arnold, M.L.** 2014. Determining population structure and hybridization for two iris species. *Ecol. Evol.* 4: 743–755. http://dx.doi.org/10.1002/ece3.964

**Heled, J. & Drummond, A.J.** 2010. Bayesian inference of species trees from multilocus data. *Molec. Biol. Evol.* 27: 570–580. http://dx.doi.org/10.1093/molbev/msp274

**Henning, F., Lee, H.J., Franchini, P. & Meyer, A.** 2014. Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: Benefits and pitfalls of using RAD markers for dense linkage mapping. *Molec. Ecol.* 23: 5224–5440. http://dx.doi.org/10.1111/mec.12860

**Herrera, S., Watanabe, H. & Shank, T.M.** 2014. Evolutionary and biogeographical patterns of barnacles from deep-sea hydrothermal vents. *Molec. Ecol.* 24: 673–689. http://dx.doi.org/10.1111/mec.13054

**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. http://dx.doi.org/10.1371/journal.pone.0093975

**Hipp, A.L., Eaton, D.A.R., Cavender-Bares, J., Nipper, R. & Manos, P.S.** 2013. Using phylogenomics to infer the evolutionary history of oaks. *Int. Oaks* 24: 61–71.

**Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A. & Cresko, W.A.** 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD Tags. *PLOS Genet.* 6: e1000862. http://dx.doi.org/10.1371/journal.pgen.1000862

**Huang, H. & Knowles, L.L.** 2014. Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of RAD sequences. *Syst. Biol.* http://dx.doi.org/10.1093/sysbio/syu046

**Hyma, K.E. & Fay, J.C.** 2013. Mixing of vineyard and oak-tree ecotypes of *Saccharomyces cerevisiae* in North American vineyards. *Molec. Ecol.* 22: 2917–2930. http://dx.doi.org/10.1111/mec.12155

**Jones, J.C., Fan, S., Franchini, P., Schartl, 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. *Molec. Ecol.* 22: 2986–3001. http://dx.doi.org/10.1111/mec.12269

**Kubatko, L.S. & Degnan, J.H.** 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56: 17–24. http://dx.doi.org/10.1080/10635150601146041

**Kubatko, L.S., Carstens, B.C. & Knowles, L.L.** 2009. STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25: 971–973. http://dx.doi.org/10.1093/bioinformatics/btp079

**Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L.** 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10: R25. http://dx.doi.org/10.1186/gb-2009-10-3-r25

**Leaché, A.D., Fujita, M.K., Minin, V.N. & Bouckaert, R.R.** 2014. Species delimitation using genome-wide SNP data. *Syst. Biol.* 63: 534–542. http://dx.doi.org/10.1093/sysbio/syu018

**Lemmon, E.M. & Lemmon, A.R.** 2013. High-throughput genomic data in systematics and phylogenetics. *Annual Rev. Ecol. Evol. Syst.* 44: 99–121. http://dx.doi.org/10.1146/annurev-ecolsys-110512-135822

**Lewis, P.O.** 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* 50: 913–925. http://dx.doi.org/10.1080/106351501753462876

**Lexer, C., Mangili, S., Bossolini, E., Forest, F., Stölting, K.N., Pearman, P.B., Zimmermann, N.E. & Salamin, N.** 2013. 'Next generation' biogeography: Towards understanding the drivers of species diversification and persistence. *J. Biogeogr.* 40: 1013–1022. http://dx.doi.org/10.1111/jbi.12076

**Li, H. & Durbin, R.** 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760. http://dx.doi.org/10.1093/bioinformatics/btp324

**Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R.** 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352

**Liu, L., Yu, L., Pearl, D.K. & Edwards, S.V.** 2009. Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* 53: 320–328. http://dx.doi.org/10.1093/sysbio/syp031

**Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M.** 2012. Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* 2012: 251364. http://dx.doi.org/10.1155/2012/251364

**Liu, L., Xi, Z., Wu, S., Davis, C. & Edwards, S.V.** 2015. Estimating phylogenetic trees from genome-scale data. *arXiv:1501.03578 [q-bio.PE].*

**Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., Buckler, E.S. & Costich, D.E.** 2013. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. *PLOS Genet.* 9: e1003215. http://dx.doi.org/10.1371/journal.pgen.1003215

**Lunter, G. & Goodson, M.** 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* 21: 936–939. http://dx.doi.org/10.1101/gr.111120.110

**Lyons, E., Bomhoff, M., Tang, H. & Joyce, B.** 2015 [last modified: 7 Jan 2015]. CoGepedia: Sequenced plant genomes. https://genomevolution.org/wiki/index.php/Sequenced_plant_genomes

**Maddison, W.P.** 1997. Gene trees in species trees. *Syst. Biol.* 46: 523–536. http://dx.doi.org/10.1093/sysbio/46.3.523

**Mardis, E.R.** 2011. A decade's perspective on DNA sequencing technology. *Nature* 470: 198–203. http://dx.doi.org/10.1038/nature09796

**McCluskey, B.M. & Postlethwait, J.H.** 2014. Phylogeny of zebrafish, a 'model species', within *Danio*, a "model genus". *Molec. Biol. Evol.* http://dx.doi.org/10.1093/molbev/msu325

**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. *Molec. Phylogen. Evol.* 66: 526–538. http://dx.doi.org/10.1016/j.ympev.2011.12.007

**McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M.A.** 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297–1303. http://dx.doi.org/10.1101/gr.107524.110

**Miele, V., Penel, S. & Duret, L.** 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. *B. M. C. Bioinf.* 12: 116. http://dx.doi.org/10.1186/1471-2105-12-116

**Miller, M.R., Atwood, T.S., Eames, B.F., Eberhart, J.K., Yan, Y.-L., Postlethwait, J.H. & Johnson, E.A.** 2007a. RAD marker microarrays enable rapid mapping of zebrafish mutations. *Genome Biol.* 8: R105. http://dx.doi.org/10.1186/gb-2007-8-6-r105

**Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A. & Johnson, E.A.** 2007b. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* 17: 240–248. http://dx.doi.org/10.1101/gr.5681207

**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. *Molec. Ecol.* 22: 814–826. http://dx.doi.org/10.1111/j.1365-294X.2012.05730.x

**Pante, E., Abdelkrim, J., Viricel, A., Gey, D., France, S.C., Boisselier, M.C. & Samadi, S.** 2015. Use of RAD sequencing for delimiting species. *Heredity* 114: 450–459. http://dx.doi.org/10.1038/hdy.2014.105

**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. http://dx.doi.org/10.1371/journal.pone.0037135

**Pickrell, J.K. & Pritchard, J.K.** 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLOS Genet.* 8: e1002967. http://dx.doi.org/10.1371/journal.pgen.1002967

**Poland, J.A., Brown, P.J., Sorrells, M.E. & Jannink, J.-L.** 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLOS ONE* 7: e32253. http://dx.doi.org/10.1371/journal.pone.0032253

**Puritz, J.B., Matz, M.V., Toonen, R.J., Weber, J.N., Bolnick, D.I. & Bird, C.E.** 2014. Demystifying the RAD fad. *Molec. Ecol.* 23: 5937–5942. http://dx.doi.org/10.1111/mec.12965

**Rannala, B. & Yang, Z.** 2008. Phylogenetic inference using whole genomes. *Annual Rev. Genomics Human Genet.* 9: 217–231. http://dx.doi.org/10.1146/annurev.genom.9.081307.164407

**Reitzel, A.M., Herrera, S., Layden, M.J., Martindale, M.Q. & Shank, T.M.** 2013. Going where traditional markers have not gone before: Utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molec. Ecol.* 22: 2953–2970. http://dx.doi.org/10.1111/mec.12228

**Rheindt, F.E., Fujita, M.K., Wilton, P.R. & Edwards, S.V.** 2014. Introgression and phenotypic assimilation in *Zimmerius* flycatchers (Tyrannidae): Population genetic and phylogenetic inferences from genome-wide SNPs. *Syst. Biol.* 63: 134–152. http://dx.doi.org/10.1093/sysbio/syt070

**Roch, S. & Steel, M.** 2014. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor. Populat. Biol.* 100: 56–62. http://dx.doi.org/10.1016/j.tpb.2014.12.005

**Roda, F., Ambrose, L., Walter, G.M., Liu, H.L., Schaul, A., Lowe, A., Pelser, P.B., Prentis, P., Rieseberg, L.H. & Ortiz-Barrientos, D.** 2013. Genomic evidence for the parallel evolution of coastal forms in the *Senecio lautus* complex. *Molec. Ecol.* 22: 2941–2952. http://dx.doi.org/10.1111/mec.12311

**Rosenberg, N.A.** 2013. Discordance of species trees with their most likely gene trees: A unifying principle. *Molec. Biol. Evol.* 30: 2709–2713. http://dx.doi.org/10.1093/molbev/mst160

**Roure, B., Baurain, D. & Philippe, H.** 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Molec. Biol. Evol.* 30: 197–214. http://dx.doi.org/10.1093/molbev/mss208

**Rubin, B.E.R., Ree, R.H. & Moreau, C.S.** 2012. Inferring phylogenies from RAD sequence data. *PLOS ONE* 7: e33394. http://dx.doi.org/10.1371/journal.pone.0033394

**Simmons, M.P.** 2012a. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. *Cladistics* 28: 208–222. http://dx.doi.org/10.1111/j.1096-0031.2011.00375.x

**Simmons, M.P.** 2012b. Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data. *Molec. Phylogen. Evol.* 62: 472–484. http://dx.doi.org/10.1016/j.ympev.2011.10.017

**Snir, S. & Rao, S.** 2012. Quartet MaxCut: A fast algorithm for amalgamating quartet trees. *Molec. Phylogen. Evol.* 62: 1–8. http://dx.doi.org/10.1016/j.ympev.2011.06.021

**Sovic, M.G., Fries, A.C. & Gibs, H.L.** 2015. AftrRAD: A pipeline for accurate and efficient de novo assembly of RADseq data. *Molec. Ecol. Resources*. http://dx.doi.org/10.1111/1755-0998.12378

**Stamatakis, A.** 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313. http://dx.doi.org/10.1093/bioinformatics/btu033

**Stolle, E. & Moritz, R.F.A.** 2013. RESTseq – Efficient benchtop population genomics with RESTriction Fragment SEQuencing. *PLOS ONE* 8: e63960. http://dx.doi.org/10.1371/journal.pone.0063960

**Streicher, J.W., Devitt, T.J., Goldberg, C.S., Malone, J.H., Blackmon, H. & Fujita, M.K.** 2014. Diversification and asymmetrical gene flow across time and space: Lineage sorting and hybridization in polytypic barking frogs. *Molec. Ecol.* 23: 3273–3291. http://dx.doi.org/10.1111/mec.12814

**Swenson, M.S., Suri, R., Linder, C.R. & Warnow, T.** 2011. An experimental study of Quartets MaxCut and other supertree methods. *Algorithms Molec. Biol.* 6: 7. http://dx.doi.org/10.1186/1748-7188-6-7

**Takahashi, T., Nagata, N. & Sota, T.** 2014. Application of RAD-based phylogenetics to complex relationships among variously related taxa in a species flock. *Molec. Phylogen. Evol.* 80: 137–144. http://dx.doi.org/10.1016/j.ympev.2014.07.016

**The Heliconius Genome Consortium** 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487: 94–98. http://dx.doi.org/10.1038/nature11041

**Toonen, R.J., Puritz, J.B., Forsman, Z.H., Whitney, J.L., Fernandez-Silva, I., Andrews, K.R. & Bird, C.E.** 2013. ezRAD: A simplified method for genomic genotyping in non-model organisms. *PeerJ* 1: e203. http://dx.doi.org/10.7717/peerj.203

**Viricel, A., Pante, E., Dabin, W. & Simon-Bouhet, B.** 2014. Applicability of RAD-tag genotyping for interfamilial comparisons: Empirical data from two cetaceans. *Molec. Ecol. Resources* 14: 597–605. http://dx.doi.org/10.1111/1755-0998.12206

**Wagner, C.E., Keller, I., Wittwer, S., Selz, O.M., Mwaiko, S., Greuter, L., Sivasundar, A. & Seehausen, O.** 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molec. Ecol.* 22: 787–798. http://dx.doi.org/10.1111/mec.12023

**Wang, S., Meyer, E., McKay, J.K. & Matz, M.V.** 2012. 2b-RAD: A simple and flexible method for genome-wide genotyping. *Nature, Meth.* 9: 808–810. http://dx.doi.org/10.1038/nmeth.2023

**Weitemier, K., Straub, S.C.K, Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A. & Liston, A.** 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Appl. Pl. Sci.* 2(9): 1400042. http://dx.doi.org/10.3732/apps.1400042

**Xi, Z., Rest, J.S. & Davis, C.C.** 2013. Phylogenomics and coalescent analyses resolve extant seed plant relationships. *PLOS ONE* 8(11): e80870. http://dx.doi.org/10.1371/journal.pone.0080870

**Xi, Z., Liu, L., Rest, J.S. & Davis, C.C.** 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. *Syst. Biol.* 63: 919–32. http://dx.doi.org/10.1093/sysbio/syu055

**Yang, Z.** 2006. *Computational molecular evolution*. New York: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780198567028.001.0001