Systematic Biology Advance Access published June 10, 2016

Sequence Capture Versus Restriction Site Associated DNA Sequencing for Shallow Systematics

Michael G. Harvey[1,2], Brian Tilston Smith[2,3], Travis C. Glenn[4], Brant C. Faircloth[1,2], and Robb T. Brumfield[1,2]

[1]*Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA*

[2]*Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA*

[3]*Department of Ornithology, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, USA*

[4]*Department of Environmental Health Science, University of Georgia, Athens, GA 30602, USA*

Corresponding author:

Michael G. Harvey

225-578-2855

mharve9@lsu.edu

ABSTRACT

Sequence capture and restriction site associated DNA sequencing (RAD-Seq) are two genomic enrichment strategies for applying next-generation sequencing technologies to systematics studies. At shallow timescales, such as within species, RAD-Seq has been widely adopted among researchers, although there has been little discussion of the potential limitations and benefits of RAD-Seq and sequence capture. We discuss a series of issues that may impact the utility of sequence capture and RAD-Seq data for shallow systematics in non-model species. We review prior studies that used both methods, and investigate differences between the methods by re-analyzing existing RAD-Seq and sequence capture datasets from a Neotropical bird (*Xenops minutus*). We suggest that the strengths of RAD-Seq datasets for shallow systematics are the wide dispersion of markers across the genome, the relative ease and cost of laboratory work, the deep coverage and read overlap at recovered loci, and the high overall information that results. Sequence capture's benefits include flexibility and repeatability in the genomic regions targeted, success using low-quality samples, more straightforward read orthology assessment, and higher per-locus information content. The utility of a method in systematics, however, rests not only on its performance within a study, but on the comparability of datasets and inferences with those of prior work. In RAD-Seq datasets, comparability is compromised by low overlap of orthologous markers across species and the sensitivity of genetic diversity in a dataset to an interaction between the level of natural heterozygosity in the samples examined and the parameters used for orthology assessment. In contrast, sequence capture of conserved genomic regions permits interrogation of the same loci across divergent species, which is preferable for maintaining comparability among datasets and studies for the purpose of drawing general

conclusions about the impact of historical processes across biotas. We argue that sequence capture should be given greater attention as a method of obtaining data for studies in shallow systematics and comparative phylogeography.

Keywords: next-generation sequencing, massively parallel sequencing, ultraconserved elements, allele frequency spectrum, coalescent methods, concordance analysis, birds

New sequencing technologies promise to provide increasingly detailed estimates of species and population histories by resolving rapid radiations (Wagner et al. 2013), improving demographic parameter estimates (Jakobsson et al. 2008), and identifying regions of the genome under selection (Wang et al. 2009). Researchers have recently adopted widely divergent strategies, however, in the approaches used to generate data for molecular systematics. Restriction site associated DNA sequencing (RAD-Seq) is the most widespread method for obtaining genomic datasets from non-model organisms, particularly for population genetic and phylogeographic studies (reviewed in Narum et al. 2013), and these data are also being increasingly used for phylogenetics (e.g., Eaton and Ree 2013; Wagner et al. 2013). In contrast, sequence capture approaches, typically targeting exons or other conserved portions of the genome, have been used primarily for reconstructing phylogenies (e.g., McCormack et al. 2013; Faircloth et al. 2013; Leaché et al. 2014). Sequence capture data are also useful for population genetic and phylogeographic studies (Carstens et al. 2013, Smith et al. 2014, McCormack et al. 2015), although few researchers have adopted this method for studies at shallow timescales. Other current genomic methods are less applicable to systematics, either because they require

high-quality samples for RNA extraction (transcriptomics; Morin et al. 2008, Künster et al. 2009), which are poorly represented in genetic resources collections, or because they remain prohibitively expensive when applied to many samples and species (whole genome sequencing; Ellegren 2014; but see Lamichhaney et al. 2015; Nater et al. 2015). Although RAD-Seq and sequence capture are both promising tools for population genetics, phylogeography, and shallow phylogenetic studies of non-model organisms, a more careful consideration of their potential weaknesses and strengths is warranted.

Differences in the potential utility of RAD-Seq and sequence capture stem from a set of issues that affect the resulting datasets. These issues are related to the function and distribution of the loci targeted, the cost of library preparation and sequencing, the assessment of sequence read orthology and locus assembly, the accuracy of variant calling and genotyping, and the information content within and across resulting loci. Each issue affects datasets in ways that may bias downstream analyses such as phylogeny reconstruction and demographic parameter estimation (Huang and Knowles 2014; Harvey et al. 2015; Mastretta-Yanes et al. 2014), and all issues may impact reproducibility and comparability of inferences across studies and species. Differences in the effects of each issue between RAD-Seq and sequence capture methods may determine which is preferable for particular applications in systematics, but there has been no thorough discussion of these issues that considers the relative merits of both RAD-Seq and sequence capture approaches.

Here, we review the major issues impacting the utility of next-generation sequencing datasets applied to systematics studies in non-model species, discuss differences in the importance of each issue relative to RAD-Seq and sequence capture datasets, and examine how each issue might affect down-stream systematics analyses. We focus on applications to “shallow

systematics", a term we use to encompass the diversity of population genetic, phylogeographic, and phylogenetic analyses currently employed by systematics researchers studying variation among populations or closely related species. We review existing studies and re-analyze previously published RAD-Seq and sequence capture datasets from the same population-level samples of a Neotropical bird (Plain Xenops, *Xenops minutus*) to provide an empirical example of the differences between methodological approaches. We argue that, although sequence capture and RAD-Seq are both useful for different applications in shallow systematics, sequence capture is better suited for making comparisons among datasets and studies and for drawing general conclusions about the processes responsible for similarities and differences in population history across species. Unlike RAD-Seq data, which are essentially one-off datasets, sequence capture data represent a lasting, amplifiable resource for comparative studies at multiple taxonomic scales.

Overview of RAD-Seq and Sequence Capture

Previous authors have described, in detail, the various strategies for conducting RAD-Seq (e.g., Davey et al. 2011; Elshire et al. 2011; Peterson et al. 2012; Wang et al. 2012; Stolle and Moritz 2013) and sequence capture studies (e.g., Mamanova et al. 2009; Gnirke et al. 2009; Faircloth et al. 2012; Lemmon et al. 2012; Bi et al. 2012; Hedtke et al. 2013; Li et al. 2013; Fortes and Paijmans 2015), so we only present a brief review of the methods. We use RAD-Seq to refer to the family of methods using restriction enzyme digests for genome reduction and high-throughput sequencing, including methods termed "genotyping by sequencing" (GBS). RAD-Seq involves digesting genomic DNA with one or more enzymes, adding platform-specific

adapters to the fragments, and selecting fragments for sequencing that fall within a particular size distribution (Fig. 1a). This digestion reduces the genome by sampling only those regions near cut sites or where cut sites occur within a certain distance of one another (Baird et al. 2008). Variations on this general method differ primarily in the number of enzymes used (one or two), the types of enzymes used and the frequency of their targeted cut sites, whether random shearing is used on one end, and the approaches used for size selection (Davey et al. 2011; Stolle and Moritz 2013). Sequence reads are distributed around cut sites depending on the method; for example, single reads bordering a single cut site or paired reads widely spaced around a single cut site in the original RAD-Seq (Baird et al. 2008), single reads adjacent to two nearby cut sites either for the same (Elshire et al. 2011; Stolle and Moritz 2013) or different (Peterson et al. 2012) enzymes, or a single read straddling a cut site (Wang et al. 2012; Fig. 1b). In most RAD-Seq methods, all fragments from a given locus have at least one static end (the cut site), meaning that sequence reads are not randomly distributed around a given cut site, which restricts the assembly of longer sequences from RAD-Seq reads. Although variations involving paired-end sequencing can produce longer alignments (Willing et al. 2011), most RAD-Seq techniques focus on collecting short sequences or single nucleotide polymorphism (SNP) data from groups of short sequences.

Sequence capture involves preparing DNA libraries from randomly fragmented DNA templates and hybridizing these libraries to sets of biotinylated synthetic oligonucleotide probes, also called baits (Gnirke et al. 2009). The probes typically have lengths of 60 to 120 bases and the sequence of each probe is complementary to one of hundreds or thousands of genomic regions of interest selected by the researcher from available sequence data (Fig. 1c). In the absence of existing genomic resources for a taxonomic group of interest, probes from genomic

regions that are conserved across divergent taxa (e.g., all amniotes, hymenoptera, or similar) can be used (Faircloth et al. 2012, 2015). Streptavidin-coated paramagnetic beads are used to attract the biotinylated probes and hybridized (target) DNA library fragments, unwanted (non-target) portions of the DNA library are washed away, and targeted fragments are then released from the beads for sequencing (Gnirke et al. 2009; Fisher et al. 2011). Because probes can be tiled across longer regions and enriched fragments are distributed in different positions across targeted loci, reads from sequence capture can be used for assembly of longer sequences (Fig. 1d). The length of contigs formed from sequence capture data is a function of the number and distribution of probes, library insert size and read depth.

RE-ANALYSIS OF EXISTING DATA

Although an increasing number of both sequence capture and RAD-Seq studies present results pertinent to the issues we describe below, drawing comparisons between studies is challenging because they often differ dramatically in sampling and, most importantly, in the methodological decisions made during the process of collecting and processing sequence data. We therefore supplement our review of existing studies with re-analysis of published RAD-Seq and sequence capture datasets, and we process these data using pipelines that maintain as much consistency as possible between each dataset. Specifically, we analyzed RAD-Seq (Harvey et al. 2015) and sequence capture (Smith et al. 2014) data collected from the same eight individuals (Table S1) of a non-model Neotropical bird, the Plain Xenops (*Xenops minutus*; family Furnariidae). Populations of *Xenops minutus*, which occurs in lowland Neotropical forests, began diverging roughly 5 Ma (Smith et al. 2014), and a deep divergence is present between

populations on either side of the Andes Mountains (Harvey and Brumfield 2015). We sampled four individuals from populations west of the Andes Mountains and four from populations east of the Andes. We collected RAD-Seq data from all samples using a GBS approach (Elshire et al. 2011), and we collected sequence capture data from ultraconserved elements as described in Faircloth et al. (2012) and Smith et al. (2014). Overall sequencing effort was higher for sequence capture (each sample was one of 44 on an Illumina Hi-Seq lane) than RAD-Seq (each sample was one of 96 on an Illumina Hi-Seq lane) resulting in an average of 4.96 times higher overall raw read counts in the sequence capture datasets (Table S2). We elected not to normalize read counts because there are diverse potential criteria for normalization (e.g. total number of reads, reads in the assembly, mean read depth at variable sites), none of which would necessarily yield equivalent datasets. Instead, we generally examine results that may be sensitive to variation in read depth across assemblies but not to absolute read depths. Although the fundamental attributes of RAD-Seq and sequence capture datasets necessitate the use of different methods for dataset assembly, thereby reducing comparability, we used approaches and parameter settings for processing that were as similar as possible between datasets (see Supplemental Information). For RAD-Seq data, we re-processed raw sequence reads and conducted *de novo* assembly using Stacks (Catchen et al. 2011, 2013), and for the sequence capture data we re-processed raw sequence reads using a custom pipeline for assembly of population-level sequence capture datasets (https://github.com/mgharvey/seqcap_pop; details in Supplemental Information) that uses some functions from the PHYLUCE package (Faircloth 2015). For both RAD-Seq and sequence capture data, we explored a series of sequence similarity thresholds for assembly and minimum read depths for calling alleles (see below), but we conducted all other analyses on datasets assembled using a 96% similarity threshold with the requirement of 7x minimum read

depth per allele (Table 1). We refer to these datasets throughout as the *Xenops minutus* RAD-Seq and sequence capture datasets.

ISSUES IN NEXT-GENERATION SEQUENCING DATASETS

The issues that determine the content of next-generation sequencing datasets are diverse and variable across methods, and we focus here on those issues that we think deserve the greatest weight when selecting RAD-Seq or sequence capture for a project in shallow systematics. We summarize differences in how each issue impacts sequence capture and RAD-Seq in Table 2.

1) Marker Distribution and Genomic Context

Restriction enzymes for RAD-Seq are often selected to cut at sites widely distributed across the genome while avoiding repetitive regions (Elshire et al. 2011). As a result, RAD-Seq sites may come from diverse coding and non-coding regions (Elshire et al. 2011; DaCosta and Sorenson 2014) having potentially heterogeneous genomic contexts and histories. RAD-Seq loci are not necessarily dispersed randomly throughout the genome, however, in part due to a preponderance of cut sites in regions with particular base compositions (DaCosta and Sorenson 2014).

Sequence capture in non-model species typically targets portions of the genome adjacent to highly conserved regions, such as ultraconserved elements (Faircloth et al. 2012) and conserved exons (Bi et al. 2012; Hedtke et al. 2013; Li et al. 2013). Conserved regions are generally selected such that they are distributed widely across available genomes (Faircloth et al.

2012). Ultraconserved elements may serve a structural or regulatory function and their conservation across deep evolutionary timescales may be indicative of strong purifying selection (Bejerano et al. 2004; Katzman et al. 2007), while exons may experience selection of various types, including purifying selection (Ward and Kellis 2012). Researchers using conserved sequences and exons have generally not selected loci based on their genomic distribution.

Because data directly comparing the distributions and genomic contexts of RAD-Seq and sequence capture are generally lacking, we explored these in our example taxon. We used Blastn (Altschul et al. 1997) to map both sets of loci to the closest genome assembly, a Golden-collared Manakin (*Manacus vitellinus*; Zhang et al. 2014) in a different family (Pipridae) but the same suborder (Tyranni) as *Xenops* (diverging 61-65 Ma; Barker et al. 2004). Using default Blastn settings, 99.4% of UCE loci successfully mapped to the *Manacus* genome compared to 17.7% of the RAD-Seq loci. The low proportion of RAD-Seq loci successfully mapping to the *Manacus* genome is consistent with the relatively low proportion of loci that are conserved across such deep evolutionary timescales in other studies of birds (e.g. McCormack et al. 2012). We used variance in the mean distance between loci across the 92,756 scaffolds in the *Manacus* genome as an index of the level of clustering (Fig. 2). Both UCEs and RAD-Seq loci recovered from *Xenops minutus* are more clustered than 1000 randomly positioned loci identified through simulations ($p < 0.001$; Figs. S1a,b, S2a,b), but the UCEs are more clustered than are all but one of 1000 random subsets of the RAD-Seq loci equal in size to the UCE dataset ($p = 0.001$; Figs. S1c). The RAD-Seq loci are closer both to predicted protein-coding genes (33.4 +/- 71.3 kbp) and repetitive elements (3.8 +/- 4.9 kbp) than are UCEs (55.0 +/- 84.1 kbp from genes, 4.3 +/- 4.3 kbp from REs). When mapped to a more distant genome (*Taeniopygia guttata*; Warren et al. 2010) with chromosome assemblies available, we found that the proportion of RAD-Seq and

sequence capture loci on each chromosome was similar ($R^2 = 0.85$, $p = 2.12 \times 10^{-12}$; Figs. S2, S3c,d).

In the *Xenops minutus* datasets, Tajima's *D* is lower (mean = -0.36, SD = 0.82) in sequence capture loci than in RAD-Seq loci (mean = 0.59, SD = 0.90), consistent with the expected effects of purifying selection (Hartl and Clark 2006). Recent evidence suggests few genomic regions are truly "neutral" (Andolfatto and Przeworski 2000; Schmid et al. 2005), and examinations of neutral population or species history may need to account for the action of selection, regardless of the loci under examination. Overall, RAD-Seq may better target heterogeneous genomic regions and be more applicable when used across species in taxonomic groups with very little genomic information. Sequence capture is flexible in that probe sets can be augmented or pruned as more genomic information becomes available for a group of interest or as loci are found to be more or less "neutral" or useful for a particular purpose.

2) Practical Considerations

Both RAD-Seq and sequence capture can be conducted with relatively small amounts of whole genomic DNA, such as those present in many museum samples. Sequence capture can often be achieved with templates of very low concentration or quality (Bi et al. 2012; Guschanski et al. 2013; McCormack et al. 2015). Many RAD-Seq methods require input DNA of higher molecular weight, but some protocols have been developed for samples of poor quality or concentration (e.g., Tin et al. 2014; Graham et al. 2015). In addition, sequence capture methods using RAD-Seq libraries as probes may allow RAD loci to be recovered from low-quality

samples (Suchan et al. 2015), and sequence capture of RAD loci may perform similarly (Ali et al. 2015).

Although next-generation sequencing platforms have dramatically reduced the cost and time involved in sequencing (Glenn 2011; Wetterstrand 2015), funding and time may still be limiting in large comparative studies due to expensive library preparations and limitations on the number of samples that can be multiplexed on a single sequencing lane (Harris et al. 2010). The cost of equipment purchase is negligible because both RAD-Seq and sequence capture can be conducted using equipment that is standard in most molecular labs (Gnirke et al. 2009; Elshire et al. 2011), although a sonicator is necessary for some library preparation methods used in sequence capture protocols. Sequence capture is generally more expensive than RAD-Seq due to the costs associated with more involved library preparation and purchasing enrichment probes. For our *Xenops minutus* datasets, sample preparation and sequencing for RAD-Seq datasets cost roughly $40 US per sample and sequence capture datasets cost roughly $60 per sample. Sequence capture may also require greater sequencing depth (to get sufficient coverage of more variable regions flanking conserved probe targets) and thus have a higher sequencing cost than RAD-Seq on a per locus basis. Depending on the success of an enrichment procedure, sequence capture may more efficiently target single-copy loci. However, in the *Xenops minutus* datasets, an average of only 5.0% of raw reads were in final assemblies in the sequence capture dataset, versus 40.1% in the RAD-Seq dataset (Table S2). This disparity may be due to poor enrichment success in the sequence capture samples we analyzed (Smith et al. 2014).

Similarly, time investment is modest for both methods (Gnirke et al. 2009; Elshire et al. 2011), although sequence capture is slower due to the additional hybridization and enrichment steps. For 96 samples, library preparation for RAD-Seq can be completed in about two days,

whereas an equivalent number of sequence capture libraries can be prepared in two to four days. Commercial library preparation and sequencing services, requiring only quantified whole genomic DNA, are available for both RAD-Seq and sequence capture. Commercial sequence capture services also require a list of target sequences from which to synthesize probes.

3) Assembly and Orthology Identification

In next-generation sequencing workflows, the process of dataset assembly is non-trivial, and its success depends on the attributes of the reads coming off the sequencer as well as the methodological decisions made during bioinformatics processing. Assembling reads into sequences and aligning them across individuals into loci is a critical component of processing next-generation sequencing datasets and has received the most attention, particularly in prior studies of the utility of RAD-Seq data for systematics (e.g. Rubin et al. 2012; Cariou et al. 2013). A primary initial concern in orthology assessment of next-generation sequence reads was whether, in divergent lineages separated by millions of years of evolutionary history, reads could be reliably recovered from sufficient loci for historical inference. It is now clear that, even in less conserved regions such as those potentially targeted by RAD-Seq protocols, at least some orthologous data can be recovered for population-level analyses and phylogenetic analyses involving species with divergences of up to 60 Ma or more (Rubin et al. 2012; Cariou et al. 2013).

A secondary issue, however, is whether the process of orthology assessment introduces biases in the resulting datasets that affect downstream analyses. Interactions between sequence divergence and the assembly parameters selected during data processing can have profound

effects on resulting datasets. Many assembly programs are available (e.g., Zerbino and Birney 2008; Simpson et al. 2009; Catchen et al. 2011) and all use sequence similarity, in some form, to assemble reads. Reads with high sequence similarity are expected to come from the same locus and are assembled, whereas those with low similarity are expected to come from different loci and are not (Pop and Salzberg 2008; Chaisson et al. 2009). A threshold is used to determine which reads belong to a single locus, but variation in genetic divergence across the genome and among study systems makes determination of an appropriate threshold challenging (Ilut et al. 2014; Harvey et al. 2015). If the similarity threshold applied is too low, reads from different loci will be assembled into a single locus and treated as orthologous ("undersplitting"), whereas if the threshold is too high, alleles belonging to a single locus may be split into separate alignments ("oversplitting").

The use of similarity thresholds for assembly is a concern for both RAD-Seq and sequence capture studies. Undersplitting may be frequent in RAD-Seq datasets if enzyme cut sites in different genomic regions fall within similar sequences, although previous results from simulated and empirical RAD-Seq data suggest that undersplitting is very infrequent (Ilut et al. 2014) and does not introduce enough signal to impact downstream analyses (Rubin et al. 2012). In many sequence capture approaches, loci are vetted to ensure they are single-copy in existing genome sequences (e.g., Faircloth et al. 2012), but the possibility of paralogous reads assembling to these loci in other taxa exists. That said, high sequence similarity within conserved regions may permit easier discrimination between orthologous and paralogous reads in sequence capture datasets, and the biology of ultraconserved elements suggests that paralogy is low (Derti et al. 2006).

We examined the relative frequency of undersplitting in RAD-Seq and sequence capture datasets from *Xenops minutus*. Examining raw assemblies, we used the number of alignments containing individuals with three or more alleles (birds are diploid) as an index of the frequency of putative paralogy (Ilut et al. 2014; Harvey et al. 2015). Given the use of a stringent depth setting per allele to remove errors (7x coverage), loci containing individuals with three or more alleles likely represent paralogous sequence rather than loci containing alleles resulting from sequencing errors. We found that undersplitting is of roughly equal and very low (<0.6% of loci) prevalence in both RAD-Seq and sequence capture datasets assembled under a range of similarity thresholds, although undersplitting increased slightly at more liberal thresholds (Fig. S4). The undersplit loci were identified and easily removed from both datasets. These results suggest undersplitting and paralogy are a relatively minor concern for both RAD-Seq and sequence capture datasets, at least in species without highly repetitive genomes and when examining relatively recently diverged samples that do not necessitate the use of liberal similarity thresholds.

Oversplitting may be frequent in short read datasets when high similarity among reads is required for assembly (Ilut et al. 2014). In *de novo* RAD-Seq assembly, oversplitting results in the separation of alternative alleles at a locus into separate alignments. Conversely, in sequence capture datasets, because reads are being aligned to a sequence determined *a priori*, oversplitting results in the loss of reads and therefore alleles that are highly divergent from the reference. High similarity thresholds for locus assembly, such as 98 or 99%, are often used with short read datasets (e.g., Catchen et al. 2011; Lu et al. 2013), potentially exacerbating the issue of oversplitting. The net result of oversplitting in both RAD-Seq and sequence capture datasets is a decrease in the average number of alleles detected within loci. We explored the frequency of

oversplitting in RAD-Seq and sequence capture using the datasets from *Xenops minutus*. We used the loss of alleles at a high similarity threshold (99%) relative to a lower similarity threshold (94%) as an index of the prevalence of oversplitting. We found that using a stringent similarity threshold resulted in an average loss of 19.4% of alleles in the RAD-Seq dataset, whereas the same similarity threshold resulted in a loss of 6.9% of alleles in the sequence capture dataset (Fig. 3a). Oversplitting appeared to level out around 96% similarity, hence the use of that threshold in the primary analyses. Oversplitting may be more severe in the RAD-Seq dataset both because of greater divergence among alleles within RAD-Seq loci relative to sequence-captured ultraconserved elements and because each oversplit locus results in two less variable alignments in RAD-Seq data. In sequence capture, conversely, oversplitting results in only one less variable locus because reads are aligned to a sequence that is determined *a priori*. Although using less stringent similarity thresholds for assembly can alleviate the impact of oversplitting (Ilut et al. 2014; Harvey et al. 2015), RAD-Seq datasets may be more sensitive to this key assembly parameter.

High conservation and low paralogy in sequence capture of conserved loci may improve discrimination of orthologous versus paralogous reads and be more amenable to assembly under low similarity thresholds. Correctly assessing orthology reduces bias in parameter estimates (Mastretta-Yanes et al. 2014), and improves the comparability of genetic diversity and inferences across studies (Harvey et al. 2015). The challenges associated with orthology assessment described above are also present, albeit less severe, in situations in which loci are assembled to a reference genome.

4) Variant Calling and Genotyping

Calling variants and genotyping individuals is the next important step after assembly when processing next-generation sequencing data, and this process is equally fraught with potential issues. PCR-related (Dunning et al. 1988, Eckert and Kunkel 1991) and short-read sequencing errors ("sequencing errors", hereafter) introduce spurious nucleotides or indels that may be identified as alleles if they are not correctly vetted (Dohm et al. 2008). Sequencing errors are problematic in both sequence capture and RAD-Seq datasets. The impact of sequencing errors on a dataset can potentially be reduced both by using filters and by calling alleles in a probabilistic framework (Nielsen et al. 2011).

Sequence read depth and evenness of sequence read depth across alleles are perhaps the most critical pieces of information researchers can use to distinguish true alleles from errors. Thus, differences in read depth or evenness across alleles between sequence capture and RAD-Seq may impact the relative success of variant calling between the two methods. Sequence capture and many RAD-Seq approaches require polymerase chain reaction (PCR) to obtain sufficient template for sequencing, and PCR can result in amplification bias and inconsistent coverage across alleles (Aird et al. 2011). Read depth in sequence capture datasets is often higher in the conserved regions targeted by the probe than in the more variable flanking regions (Fig. 1c), which are critical for calling variants. RAD-Seq datasets may also exhibit high variability in read depth across loci or amplification bias between alleles that decreases the evenness of coverage (DaCosta and Sorenson 2014). In both methods, the minimum of required PCR cycles should be used and DNA polymerases with reduced GC bias should be used to reduce amplification bias (Quail et al. 2012). PCR duplicate reads should be also removed during bioinformatics processing (Casbon et al. 2011). These are straightforward to remove from

sequence capture libraries (Sulonen et al. 2011), but PCR duplicate reads cannot be detected in RAD-Seq datasets without adapter modifications (Andrews et al. 2014; Schweyen et al. 2014; Tin et al. 2015) because even non-duplicate reads are often entirely overlapping.

We assessed the frequency of putative errors in RAD-Seq and sequence capture data from *Xenops minutus* by examining the relative read depth across rare (singleton) SNP alleles we identified in the alignments. As expected, we found that a low read depth filter (requiring 3x coverage per allele) resulted in larger datasets (Fig. S5), but a low read depth filter also resulted in more singleton alleles than assemblies requiring higher coverage (11x)(Figs. 3b, S6). The number of singletons recovered appeared to level out around 7x coverage in both datasets, which is the reason we selected this threshold for analysis of the primary datasets. The RAD-Seq dataset, however, was more impacted by the read depth filter we applied; we recovered 8.0 times as many singletons at 3x depth than we recovered at 11x depth, compared to only 4.6 times as many singletons at 3x versus 11x depth in the sequence capture dataset (Fig. 3b). This suggests that a high proportion of singleton alleles in our RAD-Seq dataset had low coverage and may represent spurious allele calls. Alternatively, the high frequency of singletons in the RAD-Seq dataset could be a result, not of lower absolute coverage, but of greater disparity in coverage across loci relative to the sequence capture dataset. It is possible protocols could be optimized to reduce coverage bias, for example by reducing the number of loci targeted in the RAD-Seq protocol or multiplexing fewer samples per sequencing lane.

Aside from sequencing errors, other artifacts can be observed in the allele frequency spectrum and can potentially be removed at the variant calling stage. Any lingering paralogous data present in an assembly (see above) can potentially be vetted during the variant calling process. High heterozygosity is typically attributed to paralogy because it may reflect the

inclusion of sequences from two divergent loci in a single alignment (Hohenlohe et al. 2011). Paralogs can be removed by filtering for heterozygosity (although this can also remove highly variable loci or loci under diversifying selection) or by filtering for loci with higher-than-expected read depth. Allele dropout due to restriction site polymorphisms is a different problem that may result in elevated homozygosity because individuals that would be heterozygous appear as homozygotes (but see Gautier et al. 2013), and it is unique to RAD-Seq datasets. Within recently diverged species and species with small effective population sizes, allele dropout may not be severe, but it is likely to increase in datasets including multiple species or deeply diverged populations (DaCosta and Sorenson 2014).

The spectrum of expected allele frequencies in a set of markers affects the ability to detect artifacts. Rare alleles representing errors may be more difficult to identify in conserved loci targeted by sequence capture because we expect a high proportion of rare alleles under purifying selection (Hartl and Clark 2006). Conversely, loci containing paralogous reads resulting in high heterozygosity may be easier to distinguish in conserved loci if there is lower overall heterozygosity in these regions.

Examining allele frequency spectra from the *Xenops minutus* datasets reveals patterns that may be due to the artifacts mentioned above and to real differences between RAD-Seq and sequence capture loci (Fig. 3c). The conserved loci recovered from sequence capture had higher overall frequencies of singleton alleles than the RAD-Seq loci (48% of alleles from sequence capture vs. 22% from RAD-Seq; Fig. 3c), consistent with the action of purifying selection. 77% of RAD-Seq genotypes were homozygous versus 56% of sequence capture genotypes in the *Xenops minutus* dataset, and the proportion of loci deficient in heterozygotes relative to Hardy-Weinberg expectations was slightly higher in the RAD-Seq (61%) than the sequence capture

dataset (55%). This discrepancy may be due to a greater effect of allele dropout, PCR bias, or low or uneven sequencing coverage across loci in the RAD-Seq dataset resulting in one allele call in heterozygotes, or it may be a result of real genotype frequency differences between the sets of markers. It is difficult to draw conclusions, however, from the differences in allele frequencies between RAD-Seq and sequence capture loci based on a single dataset due to the diversity of possible explanations for lower heterozygosity in the RAD-Seq dataset, including the possibility that they result from differences in sequencing effort.

Phasing alleles is a final important element in variant calling when researchers need to reconstruct haplotypes. In single-end RAD-Seq alignments, alleles are phased based on whether they occur on the same reads or not (read-backed phasing). In paired-end RAD-Seq and sequence capture, however, reads are not entirely overlapping and phasing of more distant alleles may require probabilistic models. These models use information from panels of reference individuals sampled previously or from other individuals in the dataset to impute the most probable combinations of alleles for heterozygous individuals. Model-based phasing introduces an extra step, and potentially additional estimation error, in datasets from paired-end RAD-Seq and sequence capture.

5) Information Content

RAD-Seq generally results in greater total aligned sequence and more potentially informative variable nucleotide sites (hereafter "information") than sequence capture. The information in RAD-Seq datasets, however, is partitioned into shorter loci. In *Xenops minutus*, for example, we assembled 158,329 RAD-Seq loci averaging 95.6 (SD = 0.62) bp in length,

whereas for sequence capture we obtained 1,358 loci averaging 590 (SD = 209) bp in length (Table 1). The total number of segregating sites for RAD-Seq (213,740) was much higher than for sequence capture (5,524), but the mean number of segregating sites per locus was higher for sequence capture: 4.07 (SD = 3.57) versus 1.35 (SD = 1.56). RAD-Seq may be preferable for estimating challenging parameters, at least in recently diverged samples, because the greater number of polymorphisms increases the chances of finding a shared allele on a very short phylogenetic branch or detecting a rare migration event. For approaches requiring more information per locus, sequence capture would be preferable.

Potential Effects of Biases on Inferences

The issues described above may shape datasets in ways that make them more or less appropriate or biased for downstream shallow systematics analyses. Sequence capture and RAD-Seq datasets yield broadly concordant results for phylogenetic analyses among species, depending on the steps used for dataset assembly (Leaché et al. 2015, Collins and Hrbek 2015, Manthey et al. 2016), but their relative utility for population genetic and phylogeographic analyses that are applied within species is largely unexplored. In this section, we discuss how these issues might impact a range of typical population genetic, phylogeographic, and phylogenetic analyses that are often applied at shallow timescales. The results of analyses of the empirical datasets presented here are not intended as a direct comparison of the applicability of RAD-Seq and sequence capture data, which in reality would probably not be examined with the same methods, but rather to demonstrate how the issues discussed above can result in divergent inferences between methods.

Genome-wide scans to identify signatures of selection or gene flow are often conducted in studies using RAD-Seq loci due to their dense distribution across the genome (Hohenlohe et al. 2010). Conserved regions targeted by sequence capture may be insufficiently dispersed across the genome for use in genome-wide scans. As discussed above, mapping RAD-Seq loci to divergent genomes is challenging, thus RAD-Seq may not be appropriate for identifying the genomic context of outlier loci in species without available genome assemblies. As with many markers, RAD-Seq loci may come from heterogeneous genomic regions impacted by diverse neutral and non-neutral processes, so scans will need to account for alternative explanations of outlier loci or migrant alleles.

Demographic inference is popular in population genetics and phylogeography, and may be affected by the distribution of allele frequencies in a dataset. Purifying selection on conserved regions may leave signatures, such as an excess of rare alleles, that complicate estimation of neutral demographic histories. Allele loss and heterozygote deficiencies in RAD-Seq datasets may also affect estimates of demographic parameters including theta ($\theta = 4N_e\mu$) and admixture. We estimated demographic parameters using gene trees in BP&P v.3.2 (Yang and Rannala 2010) and using SNP frequency spectra in ∂a∂i v.1.7.0 (Gutenkunst et al. 2009) with both RAD-Seq and sequence capture data from *Xenops minutus*. The demographic model used included two daughter populations comprising the four samples from west of the Andes Mountains and the four samples from east of the Andes Mountains, both of which diverged from a common ancestral population. We compared estimates of effective population size by normalizing the divergence time estimates from RAD-Seq and sequence capture datasets. We found that in both BP&P and ∂a∂i results, effective population sizes in the daughter populations were fairly similar between datasets (Tables S3, S4), but the estimate of ancestral effective population size was

lower from sequence capture than from RAD-Seq data (Fig. 4b, Fig. S7). The higher ancestral population size in the RAD-Seq data could be due either to the loss of shared variation among the daughter populations as a result of allele dropout in the RAD-Seq dataset, or to the high frequency of rare alleles restricted to a single population in the sequence capture alignments. In addition, heterozygote deficiencies in the RAD-Seq dataset may underlie the somewhat lower population sizes estimated in the daughter populations than those estimated in the sequence capture dataset.

Phylogenetic tree estimation to reconstruct the relationships between populations are commonly used in shallow systematics studies. Phylogeny estimation may be complicated if allele loss results in a downward bias in the mutational spectrum (Huang and Knowles 2014). This bias may produce shallower gene trees and lower genetic distances (Harvey et al. 2015), particularly between the most divergent individuals in a study. We examined branch lengths from *Xenops minutus* trees inferred using BUCKy v.1.4.3 (Larget et al. 2010), which are estimated in coalescent units based on quartet concordance factors for each branch. As observed in prior studies (Leaché et al. 2015), internal branch lengths from BUCKy trees estimated from RAD-Seq data were short relative to those estimated from sequence capture data in *Xenops minutus*, perhaps as a result of the loss of the most divergent alleles (Fig. 4c,d). Terminal branches in BUCKy trees for *Xenops minutus* are determined by the gene trees from loci in which individuals are homozygous for rare alleles. These branch lengths are longer in the RAD-Seq tree than in the sequence capture tree, consistent with the high levels of homozygosity we observed in the RAD-Seq dataset. The difference in relative branch lengths between RAD-Seq and sequence capture trees was not evident in trees estimated from SNPs using SNAPP (Bryant et al. 2012), likely because SNAPP removes sites with missing data, which would bias overall

tree depth rather than relative branch lengths (Fig. S8). Despite the differences in phylogenetic branch lengths, relative genetic distances corrected using a JC69 model (Jukes and Cantor 1969) among individuals were highly correlated between RAD-Seq and sequence capture *Xenops minutus* datasets (CADM test coefficient of concordance = 0.935, $p < 0.001$, Fig. 4a).

Both RAD-Seq and, to a lesser extent, sequence capture loci have low per-locus information relative to many of the genes traditionally targeted for Sanger sequencing in systematics. Low per-locus information content complicates analyses that depend on accurate parameter estimates from individual loci. It may be challenging to fit models of molecular evolution to loci due to their low information content, and poorly resolved gene trees may complicate analyses such as gene tree-species tree estimation (Lanier et al. 2014). Concordance analysis of gene trees from RAD-Seq and sequence capture in *Xenops minutus* using BUCKy (Larget et al. 2010) revealed that consensus relationships were supported by relatively few loci (Fig. 4c,d). Most gene trees contained polytomies as a result of low information content in alignments. Concordance was lower among RAD-Seq loci than among sequence capture loci, presumably due to the lower resolution of RAD-Seq gene trees. The consensus trees inferred across loci from both methods were topologically identical, however, both using BUCKy (Fig. 4c,d) and SNAPP (Fig. S8). Moreover, nearly all nodes had high support in the SNAPP trees from both RAD-Seq and sequence capture. Methods that successfully integrate across the small amounts of information present in many loci, including methods that examine independent SNPs, may be desirable for sequence capture and particularly RAD-Seq datasets.

The large datasets produced by RAD-Seq and sequence capture raise computational concerns. Although the sizes of both RAD-Seq and sequence capture datasets can be tailored according to researcher needs, RAD-Seq datasets are generally larger. Depending on the question

being addressed, very large datasets may not be needed and additional data may unnecessarily complicate analyses (Davey et al. 2011). Conversely, evolutionary events that are difficult to estimate may require large amounts of data to address, and larger datasets also offer the ability to subsample loci informing a research question *post-hoc*. To take advantage of the information in large datasets, computationally demanding methods may have to take a back seat to faster summary methods (e.g., Liu et al. 2009; Larget et al. 2010; Chaudhary et al. 2014).

COMPARING ACROSS DATASETS AND CALIBRATING PARAMETERS

The same RAD-Seq loci often cannot be recovered across divergent species due to mutations at restriction sites (Rubin et al. 2012) or variation in sequence coverage. Studies have successfully recovered some shared loci at moderately deep (~60 Ma) timescales in *Drosophila* (Rubin et al. 2012; Cariou et al. 2013), but sequence capture is substantially more effective for recovering the same loci, even at very deep timescales (up to about ~400 Ma; Faircloth et al. 2012; Faircloth et al. 2013). In birds, a comparison of population-level RAD-Seq datasets from across four species widely distributed across the avian tree of life found that only 0.3-0.8% of loci overlapped across all four (McCormack et al. 2012), whereas population-level sequence capture datasets between any two of 40 bird species from diverse families had an average of 92% overlap (Supplemental Information).

When identical loci cannot be recovered in different datasets, comparability among studies relies on the assumption that the set of loci in each species represents a random sample from the genome. Based on the discussion above, however, the diversity present in RAD-Seq datasets is not random with respect to the level of genetic variation and genome complexity in

the species being examined. Oversplitting is a major issue in RAD-Seq datasets and it disproportionately affects species with higher natural levels of divergence among alleles (Huang and Knowles 2014; Harvey et al. 2015). Species with higher divergence will lose more variation than those with lower divergence, resulting in a normalization of the variation present across datasets assembled with the same parameters. In addition, undersplitting may be a greater issue in species with repetitive genomes (Ilut et al. 2014; Harvey et al. 2015). Both oversplitting and undersplitting, therefore, could result in similarities or differences among species that are artefactual. Methods are available for informed selection of assembly parameters to reduce the effects of oversplitting and undersplitting (e.g. Ilut et al. 2014; Harvey et al. 2015), but they are not widely applied, and it is unclear whether they will be sufficient to permit comparisons across species. Any differences among datasets in the restriction enzymes or in the assembly strategies used among studies will further reduce comparability. Therefore, similar to studies of microsatellites, many analyses using RAD-Seq loci cannot easily be compared across species. Sequence capture of loci containing conserved regions is preferred for obtaining genomic data from a standard set of loci if there is to be hope that datasets or inferences could be directly compared across species.

Parameter calibration is also problematic when datasets are not comparable across species. In species or groups without fossil data or divergences tied to dated geological events, estimating absolute values for demographic and phylogenetic parameters requires calibration, typically by applying standardized substitution rates. Mutation rates, however, can only be adopted from other studies when the loci examined are the same or are expected to have similar rates of evolution. Because different loci are examined in RAD-Seq datasets, and the mutation rate in a dataset may be contingent on the impact of the assembly issues mentioned above, there

is little hope of developing standardized mutation rate estimates. Calibration across species should be possible in sequence capture datasets, however, if datasets are assembled and variants called in the same way, if the alignments are trimmed such that they contain the same sites across species, and if subsets of clock-like loci are identified (e.g., Doyle et al. 2015).

CONCLUSIONS

Although prior studies suggest RAD-Seq and sequence capture are both useful for shallow systematics and we observed broad concordance in RAD-Seq and sequence capture datasets and resulting inferences, the differences observed and discussed above suggest these approaches are not equally useful for different applications in shallow systematics. Sequence capture holds more promise for obtaining datasets that are comparable across species and for calibrating parameter estimates for demographic or phylogenetic studies. In addition, sequence capture is useful because marker sets can be tailored according to the needs of the researcher, because it is particularly effective with low-quality samples, because data from new samples can be easily added to an existing dataset, because orthology of sequence reads is relatively straightforward to assess, and because sequences could be useful in other studies at deeper evolutionary timescales. Completely or partially shared probe sets among studies will result in a growing, open source data matrix that can be used for comparative phylogeographic and phylogenetic analyses at multiple taxonomic scales. RAD-Seq will continue to be useful as a fast and inexpensive means to obtain large amounts of data, and its application to single-species population studies, genome scans, groups without genomic information, and species with very shallow histories is sure to continue. We suggest, however, that sequence capture should be

preferred, given sufficient resources, due to the higher comparability and extensibility of datasets.

We anticipate that the importance of the issues described in this paper on datasets from sequence capture and RAD-Seq will change as the methods for each evolve and improve. Moreover, new methods are sure to appear (e.g. Ali et al. 2015, Niedzicka et al. 2016) and existing methods such as whole-genome sequencing and re-sequencing will become more affordable in the near future. Many of the issues we have described transcend the genomic methods discussed here, however, and will continue to be relevant in discussions of the utility of new methods. Regardless of the method applied, a premium should be placed on maintaining comparability with other studies such that results and inferences can be properly incorporated into the body of systematics literature as a whole.

FUNDING

This work was funded in part by NSF grants DEB-1146265 and DEB-1210556 (a Doctoral Dissertation Improvement Grant for MGH's dissertation) to RTB and DEB-1242267 to BCF and TCG.

ACKNOWLEDGEMENTS

D. Willard (Field Museum), M. B. Robbins (University of Kansas Natural History Museum), and D. L. Dittmann (Louisiana State University Museum of Natural Science) provided genetic samples. J. M. Brown, B. C. Carstens, A. D. Leaché, and J. M. DaCosta discussed

experimental design. C. Locklear at Integrated DNA Technologies (IDT) provided adapters and sequencing. Bill Ludt, Prosanta Chakrabarty, Isaac Overcast, and the Systematics Discussion Group at LSU provided helpful feedback. Portions of this research were conducted with high performance computing resources provided by Louisiana State University (http://www.hpc.lsu.edu), and B. Thakur assisted in implementing analyses on computing clusters.

SUPPLEMENTARY MATERIAL

Supplementary material, including supplementary methods, figures, and tables and data files, can be found in the Dryad digital repository (http://dx.doi.org/10.5061/dryad.83fg0). Data files from the original sequence capture study are available from Dryad (http://dx.doi.org /10.5061/dryad.qm4j1) and from the original RAD-Seq study from the NCBI dbSNP (ss# 1536954775–1536958153) and Dryad (http://dx.doi.org/10.5061/dryad.3j0b1).

Aird, D., M. G. Ross, W.-S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, and A. Gnirke. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 12:R18.

Ali, O. A., S. M. O'Rourke, S. J. Amish, M. H. Meek, G. Luikart, C. Jeffres, and M. R. Miller. 2015. RAD Capture (Rapture): Flexible and efficient sequence-based genotyping. Genetics doi:10.1534/genetics.115.183665.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and J. D. Lipman. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucl. Acids Res. 25:3389-3402.

Andolfatto, P. and M. Przeworski. 2000. A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. Genetics 156:257-268.

Andrews, K. A., P. A. Hohenlohe, M. R. Miller, B. K. Hand, J. E. Seeb, and G. Luikart. 2014. Trade-offs and utility of alternative RADseq methods: Reply to Puritz et al. 2014. Mol. Ecol. 23: 5943-5946.

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

Barker, F. K., A. Cibois, P. Schikler, J. Feinstein, and J. Cracraft. 2004. Phylogeny and diversification of the largest avian radiation. Proc. Nat Acad. Sci. 101:11040-11045.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler. 2004. Ultraconserved elements in the human genome. Science 304:1321-1325.

Bi, K., D. Vanderpool, S. Singhal, T. Linderoth, C. Moritz, and J. Good. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics 13:403.

Bryant, D., R. Bouckaert, J. Felsenstein, N. A. Rosenberg, and A. RoyChoudhury. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917-1932.

Catchen J. M., A. Amores, P. A. Hohenlohe, W. A. Cresko, and J. H. Postlethwait. 2011. Stacks: building and genotyping loci de novo from short-read sequences. G3 Genes Genomes Genetics 1:171-182.

Catchen J. M., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology 22:3124-3140.

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

Carstens, B. C., R. S. Brennan, V. Chua, C. V. Duffie, M. G. Harvey, R. A. Koch, C. D. McMahan, B. J. Nelson, C. E. Newman, J. D. Satler, G. F. Seeholzer, K. Posbic, D. C. Tank, and J. Sullivan. 2013. Model selection as a tool for phylogeographic inference: An example from the willow *Salix melanopsis*. Mol. Ecol. 22:4014-4028.

Casbon, J. A., R. J. Osbourne, S. Brenner, and C. P. Lichtenstein. 2011. A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Res. 39:e81.

Chaisson, M. J., D. Brinza, and P. A. Pevzner. 2009. De novo fragment assembly with short mate-pair reads: Does the read length matter? Genome Res. 19:336-346.

Chaudhary, R., D. Fernández-Baca, and J. G. Burleigh. 2014. MulRF: A software package for phylogenetic anlaysis using multi-copy gene trees. Bioinformatics 31:432-433.

Collins, R. A. and T. Hrbek. 2015. An in silico comparison of reduced-representation and sequence-capture protocols for phylogenomics. BioRxiv doi:10.1101/032565.

DaCosta, J. M. and M. D. Sorenson. 2014. Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. PLoS One 9:e106713.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12:499-510.

Derti, A., F. P. Roth, G. M. Church, and C.-T. Wu. 2006. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. Nat. Genet. 38:1216-1220.

Dohm, J. C., C. Lottaz, T. Borodina, and H. Himmelbauer. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucl. Acids Res. 36:e105.

Doyle, V. P., R. E. Young, G. P. Naylor, and J. M. Brown. 2015. Can we identify genes with increased phylogenetic reliability? Syst. Biol. 64:824-837.

Dunning, A. M., P. Talmud, and S. E. Humphries. 1988. Errors in the polymerase chain reaction. Nucleic Acids Res. 16:10393.

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

Eckert, K. A. and T. A. Kunkel. 1991. DNA polymerase fidelity and the polymerase chain reaction. Genome Res. 1:17-24.

Ellegren, H. 2014. Genome sequencing and population genomics in non-model organisms. Trends Ecol. Evol. 29:51-63.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379.

Faircloth, B. C. 2015. PHYLUCE is a software package for the analysis of conserved genomic loci. BioRxiv doi:10.1101/027904.

Faircloth, B. C., M. G. Branstetter, N. D. White, and S. G. Brady. 2015. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol. Ecol. Res. 15:489-501.

Faircloth, B. C., J. E. McCormack, N. G. Crawford, M. G. Harvey, R. T. Brumfield, and T. C. Glenn. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61:717-726.

Faircloth, B. C., L. Sorenson, F. Santini, and M. E. Alfaro. 2013. A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). PLoS One 8:e65923.

Fisher, S., A. Barry, J. Abreu, B. Minie, J. Nolan, T. M. Delorey, G. Young, T. J. Fennell, A. Allen, and L. Ambrogio. 2011. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol. 12:R1.

Fortes, G. G., and L. A. Paijmans. 2015. Analyis of whole mitogenomes from ancient samples. arXiv 1503.05074.

Gautier, M., K. Gharbi, T. Cezard, J. Foucaud, C. Kerdelhué, P. Pudlo, J.-M. Cornuet, and A. Estoup. 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol. Ecol. 22:3165-3178.

Glenn, T. C. 2011. Field guide to next-generation DNA sequencers. Mol. Ecol. Res. 11:759-769.

Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ, S. Gabriel, D. B. Jaffe, E. S. Lander, and C. Nusbaum. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotech. 27:182-189.

Graham, C. F., T. C. Glenn, A. G. McArthur, D. R. Boreham, T. Kieran, S. Lance, R. G. Manzon, J. A. Martino, T. Pierson, S. M. Rogers, J. Y. Wilson, and C. M. Somers. Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). Mol. Ecol. Res. 15:1304-1315.

Guschanski, K., J. Krause, S. Sawyer, L. M. Valente, S. Bailey, K. Finstermeier, R. Sabin, E. Gilissen, G. Sonet, Z. T. Nagy, G. Lenglet, F. Mayer, and V. Savolainen. 2013. Syst. Biol. 62:539-554.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5:1-11.

Harris, J. K., J. W. Sahl, T. A. Castoe, B. D. Wagner, D. D. Pollock, and J. R. Spear. 2010. Comparison of normalization methods for construction of large, multiplex amplicon pools for next-generation sequencing. Appl. Environ. Microb. 76:3863-3868.

Hartl, D. L. and A. G. Clark. 2006. Principles of Population Genetics: Fourth Edition. Sunderland, Massachusetts: Sinauer.

Harvey, M. G. and R. T. Brumfield. 2015. Genomic variation in a widespread Neotropical bird (*Xenops minutus*) reveals divergence, population expansion, and gene flow. Mol. Phylo. Evol. 83:305-316.

Harvey, M. G., C. D. Judy, G. F. Seeholzer, J. M. Maley, G. R. Graves, and R. T. Brumfield. 2015. Similarity thresholds used in DNA sequence assembly from short reads can reduce the comparability of population histories across species. PeerJ 3:e895.

Hedtke, S. M., M. J. Morgan, D. C. Cannatella, and D. M. Hillis. 2013. Targeted enrichment: maximizing orthologous gene comparisons across deep evolutionary time. PLoS One 8:e67908.

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson, and W. A. Cresko. 2010. Population genomics of parallel adaptation in threespine stickleback using sequences RAD tags. PLoS Genet. 6:e1000862.

Hohenlohe, P. A., S. J. Amish, J. M. Catchen, F. W. Allendorf, and G. Luikart. 2011. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Mol. Ecol. Res. 11:117-122.

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

Ilut, D. C., M. L. Nydam, and M. P. Hare. 2014. Defining loci in restriction-based reduced representation genomic data from nonmodel species: Sources of bias and diagnostics for optimal clustering. BioMed Res. Int. 2014:675158.

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H.-C. Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, and R. Guerreiro. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998-1003.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21-123 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

Katzman, S., A. D. Kern, G. Bejerano, G. Fewell, L. Fulton, R. K. Wilson, S. R. Salama, and D. Haussler. 2007. Human genome ultraconserved elements are ultraselected. Science 317:915-915.

Künster, A, J. B. Wolf, N. Backström, O. Whitney, C. N. Balakrishnan, L. Day, S. V. Edwards, D. E. Janes, B. A. Schlinger, R. K. Wilson, and E. D. Jarvis. 2010. Comparative genomics based on massively parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. Mol. Ecol. 19:266-276.

Lamichhaney, S., J. Berglund, M. S. Allmén, K. Maqbool, M. Grabherr, A. Martinez-Barrio, M. Promerová, C.-J. Rubin, C. Wang, N. Zamani, B. R. Grant, P. R. Grant, M. T. Webster, and L. Andersson. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature 518:371-375.

Lanier, H. C., H. Huang, and L. L. Knowles. 2014. How low can you go? The effects of mutation rate on the accuracy of species-tree estimation. Mol. Phylo. Evol. 70:112-119.

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

Leaché, A. D., P. Wagner, C. W. Linkem, W. Böhme, T. J. Papenfuss, R. A. Chong, B. R. Lavin, A. M. Bauer, S. Nielsen, E. Greenbaum, M-O. Rödel, A. Schmitz, M. LeBreton, I. Ineich, L. Chirio, E. A. Eniang, S. Baha El Din, A. R. Lemmon, and F. T. Burbrink. 2014. A

hybrid phylogenetic-phylogenomic approach for species tree estimation in African Agama lizards with applications to biogeography, character evolution, and diversification. Mol. Phylo. Evol. 79:215-230.

Leaché, A. D., A. S. Chavez, L. N. Jones, J. A. Grummer, A. D. Gottscho, and C. W. Linkem. 2015. Phylogenomics of Phrynosomatid lizards: Conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biology and Evolution 7:706-719.

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

Li, C., M. Hofreiter, N. Straube, S. Corrigan, and G. J. Naylor. 2013. Capturing protein-coding genes across highly divergent species. Biotechniques 54:321-326.

Liu, L., L. Yu, D. K. Pearl, and S. V. Edwards. 2009. Estimating species phylogenies using coalescence times among sequences. Systematic Biology 58:468-477.

Lu, F., A. E. Lipka, J. Glaubitz, R. Elshire, J. H. Cherney, M. D. Casler, E. S. Buckler, and D. E. Costich. 2013. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. PLoS Genetics 9:e1003215.

Mamanova, L., A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, J. Shendure, and D. J. Turner. 2010. Target-enrichment strategies for next-generation sequencing. Nature Methods 7:111-118.

Manthey, J. D., L. C. Campillo, K. J. Burns, and R. G. Moyle. 2016. Comparison of target capture and restriction-site associated DNA sequencing for phylogenomics: a test in Cardinalid tanagers (Aves, Genus: *Piranga*). Systematic Biology doi:10.1093/sysbio/syw005.

Mastretta-Yanes, A., N. Arrigo, N. Alvarez, T. H. Jorgensen, D. Piñero, and B. C. Emerson. 2014. Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. Mol. Ecol. Res. 15:28-41.

McCormack, J. E., M. G. Harvey, B. C. Faircloth, N. G. Crawford, T. C. Glenn, and R. T. Brumfield. 2013. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS One 8:e54848.

McCormack, J. E., J. M. Maley, S. M. Hird, E. P. Derryberry, G. R. Graves, and R. T. Brumfield. 2012. Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. Mol. Phylo. Evol. 62:397-406.

McCormack, J. E., W. L. E. Tsai, and B. C. Faircloth. 2015. Sequence capture of ultraconserved elements from bird museum specimens. Mol. Ecol. Res. doi:10.111/1755-0998.12466.

Morin, R. D., M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. J. Pugh, H. McDonald, R. Varhol, S. J. Jones, and M. A. Marra. 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 45:81.

Nater, A., R. Burri, T. Kawakami, L. Smeds, and H. Ellegren. 2015. Resolving evolutionary relationships in closely related species with whole-genome sequencing data. Syst. Biol. doi:10.1093/sysbio/syv045.

Narum, S. R., C. A. Buerkle, J. W. Davey, M. R. Miller, and P. A. Hohenlohe. 2013. Genotyping-by-sequencing in ecological and conservation genomics. Mol. Ecol. 22:2841-2847.

Niedzicka, M., A. Fijarczyk, K. Dudek, M. Stuglik, and W. Babik. 2016. Molecular inversion probes for targeted resequencing in non-model organisms. Sci. Rep. 6:24051.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song. 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12:443-451.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra. 2012. Double digest RADseq: An inexpensive method for *de novo* genotyping in model and non-model species. PLoS One 7:e37135.

Pop, M., and S. L. Salzberg. 2008. Bioinformatics challenges of new sequencing technology. Trends Genet. 24:142-149.

Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. 2012. A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences, and Illumina MiSeq sequencers. BMC Genomics 13:341.

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

Schmid, K. J., S. Ramos-Onsins, H. Ringys-Beckstein, B. Weisshaar, and T. Mitchell-Olds. 2005. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics 169:1601-1615.

Schweyen, H., A. Rozenberg, and F. Leese. 2014. Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. Biol. Bull. 227:146-160.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol. 2009. ABySS: A parallel assembler for short read sequence data. Genome Research 19:1117-1123.

Smith, B. T., M. G. Harvey, B. C. Faircloth, T. C. Glenn, and R. T. Brumfield. 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Syst. Biol. 63:83-95.

Springer, M. S. and J. Gatesy. 2015. The gene tree delusion. Mol. Phylo. Evol. 94:1-33.

Stolle, E., and R. F. A. Moritz. 2013. RESTseq: Efficient benchtop population genomics with restriction fragment sequencing. PLoS One 8:e63960.

Suchan, T., C. Pitteloud, N. Gerasimova, A. Kostikova, N. Arrigo, M. Pajkovic, M. Ronikier, and N. Alvarez. 2015 Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on museum collection specimens. BioRxiv doi:10.1101/025551.

Sulonen, A.-M., P. Ellonen, H. Almusa, M. Lepisto, S. Eldorfs, S. Hannula, T. Miettinen, H. Tyynismaa, P. Salo, C. Heckman, H. Joensus, T. Raivio, N. Suomalainen, and J. Saarela. 2011. Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol. 12:R94.

Tin, M. M. Y., E. P. Economo, and A. S. Mikheyev. 2014. Sequencing degraded DNA from non-destructively sampled museum specimens for RAD-tagging and low-coverage shotgun phylogenetics. PLoS One 9:e96793.

Ward, L. D. and M. Kellis. 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science 337:1675-1678.

Wagner, C. E., I. Keller, S. Wittwer, O. M. Selz, S. Mwaiko, L. Greuter, A. Sivasundar, and O. Seehausen. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22:787-798.

Wang, S., E. Meyer, J. K. McKay, and M. V. Matz. 2012. 2b-RAD: A simple and flexible method for genome-wide genotyping. Nat. Methods 9:808-810.

Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10:57-63.

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Künstner, A., Searle, S., White, S., Vilella, A.J., Fairley, S., Heger, A., Kong, L., Ponting, C.P., Jarvis, E.D., Mello, C.V., Minx, P., Lovell, P., Velho, T.A.F., Ferris, M., Balakrishnan, C.N., Sinha, S., Blatti, C., London, S.E., Li, Y., Lin, Y.-C., George, J., Sweedler, J., Southey, B., Gunaratne, P., Watson, M., Nam, K., Backström, N., Smeds, L., Nabholz, B., Itoh, Y., Whitney, O., Pfenning, A.P., Howard, J., Völker, M., Skinner, B.M., Griffin, D.K., Ye, L., McLaren, W.M., Flicek, P., Quesada, V., Velasco, G., Lopez-Otin, C., Puente, X.S., Olender, T., Lancet, D., Smit, A.F.A., Hubley, R., Konkel, M.K., Walker, J.A., Batzer, M.A., Gu, W., Pollock, D.D., Chen, L., Cheng, Z., Eichler, E.E., Stapley, J., Slate, J., Ekblom, R., Birkhead, T., Burke, T., Burt, D., Scharff, C., Adam, I., Richard, H., Sultan, M., Soldatov, A., Lehrach, H., Edwards, S.V., Yang, S.-P., Li, X., Graves, T., Fulton, L., Nelson, J., Chinwalla, A., Hou, S., Mardis, E.R., Wilson, R.K.. 2010. The genome of a songbird. Nature 464:757–762.

Wetterstrand, M. S. 2015. DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP). Available from http://www.genome.gov/sequencingcosts (last accessed September 23, 2015).

Willing, E. M., M. Hoffmann, J. D. Klein, D. Weigel, and C. Dreyer. 2011. Paired-end RAD-seq for de novo assembly and marker design without available reference. Bioinformatics 27:2187-2193.

Yang, Z. and B. Rannala. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. U.S.A. 107:9264-9269.

Zerbino, D. R., and E. Birney. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821-829.

Zhang, G., B. Li, C. li, M. T. P. Gilbert, E. D. Jarvis, J. Wang, and The Avian Genome Consortium. 2014. Comparative genomic data of the Avian Phylogenomics Project. GigaScience 3:26.

a)

Sample 1 Sample 2 Sample 3

Enzyme digestion, size selection

Adapter/barcode ligation

Pooling, PCR, Illumina sequencing

b)

Original RAD-Seq

ddRAD

GBS and RESTseq

2b-RAD

Read Depth

Contig Length

c)

Sample 1 Sample 2 Sample 3

Random fragmentation

Adapter/barcode ligation, pooling, PCR

Hybridization to biotinylated probes

Enrichment with streptavidin beads

PCR, Illumina sequencing

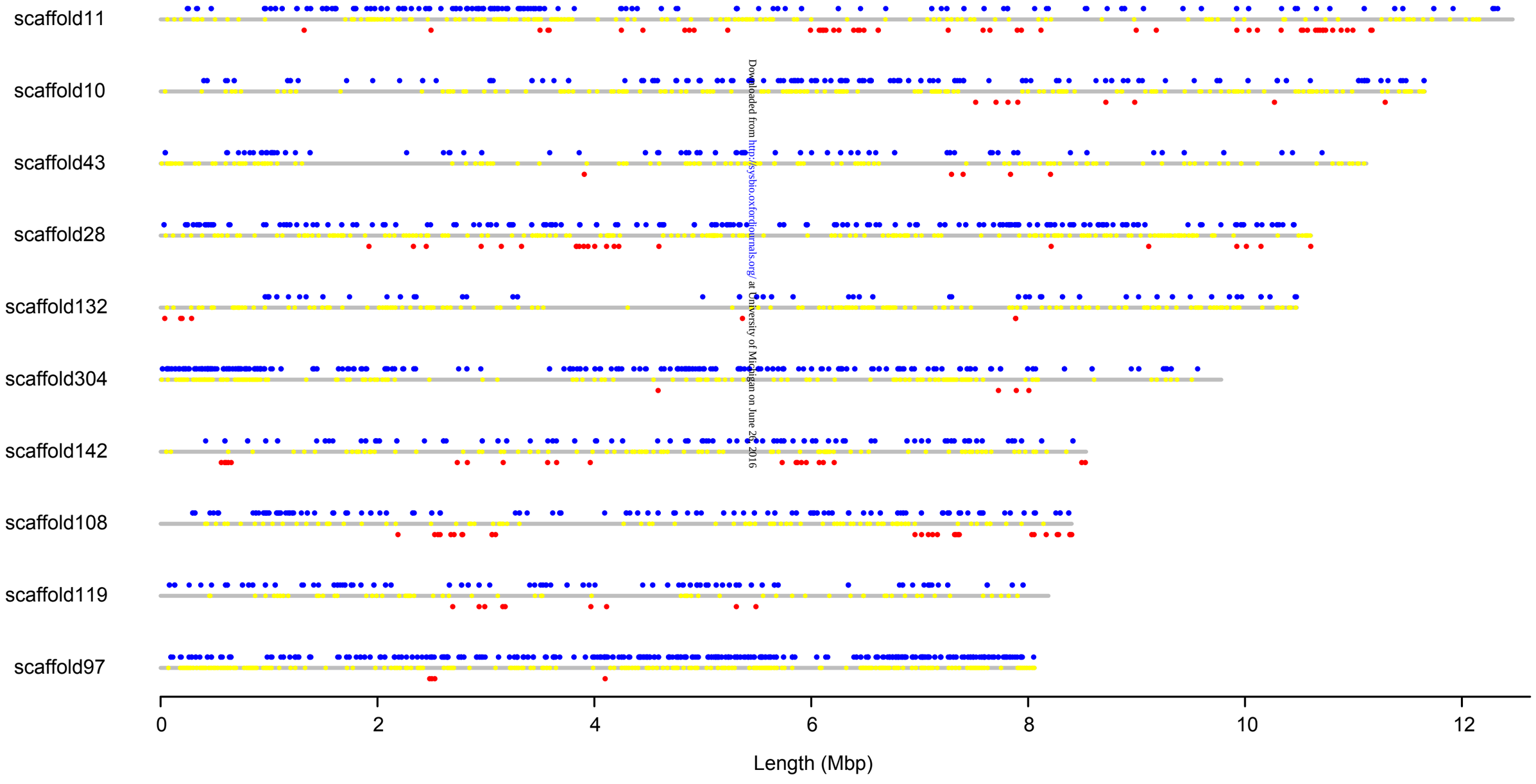
d)

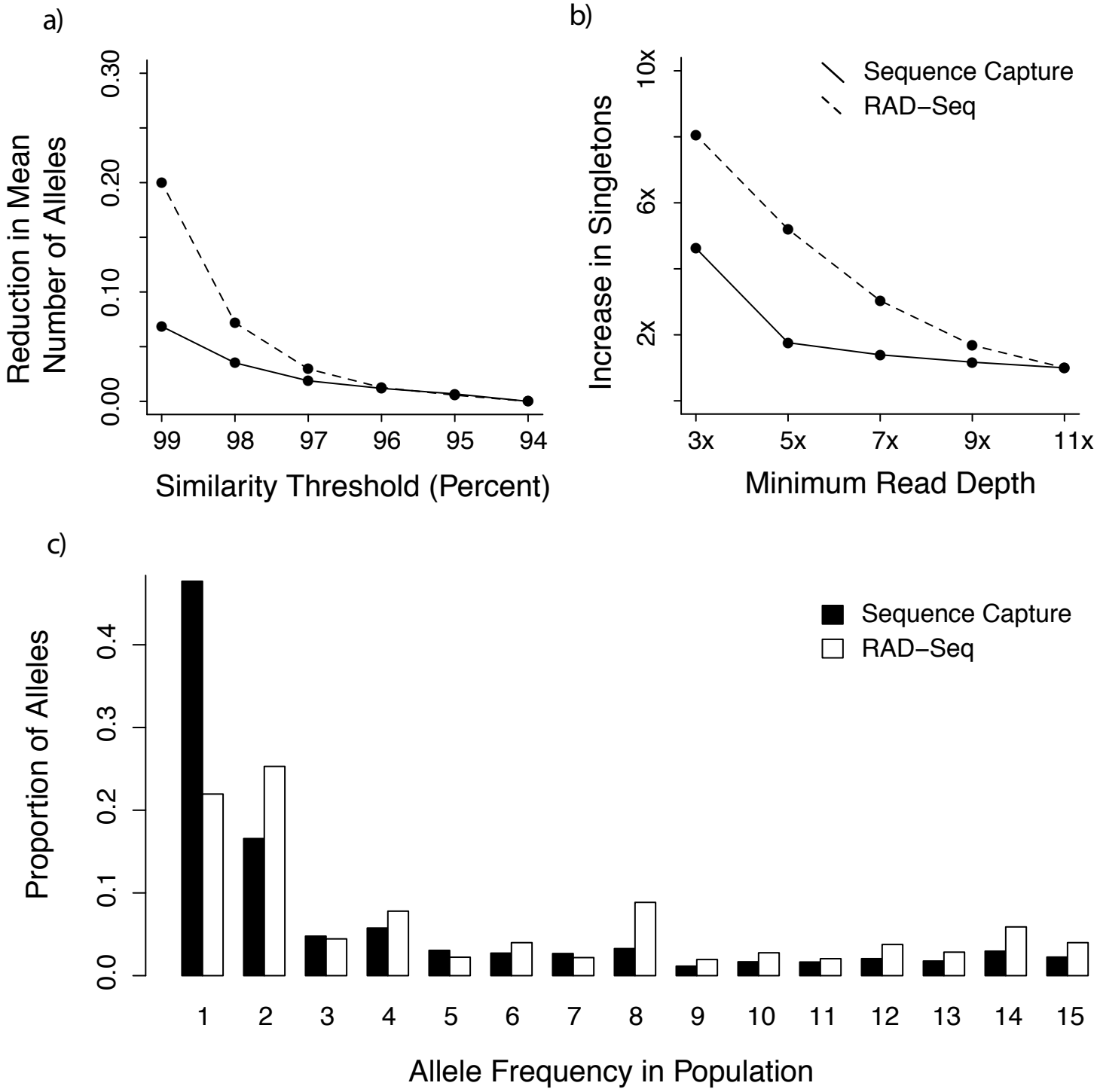
Single Probe

Tiled Probes

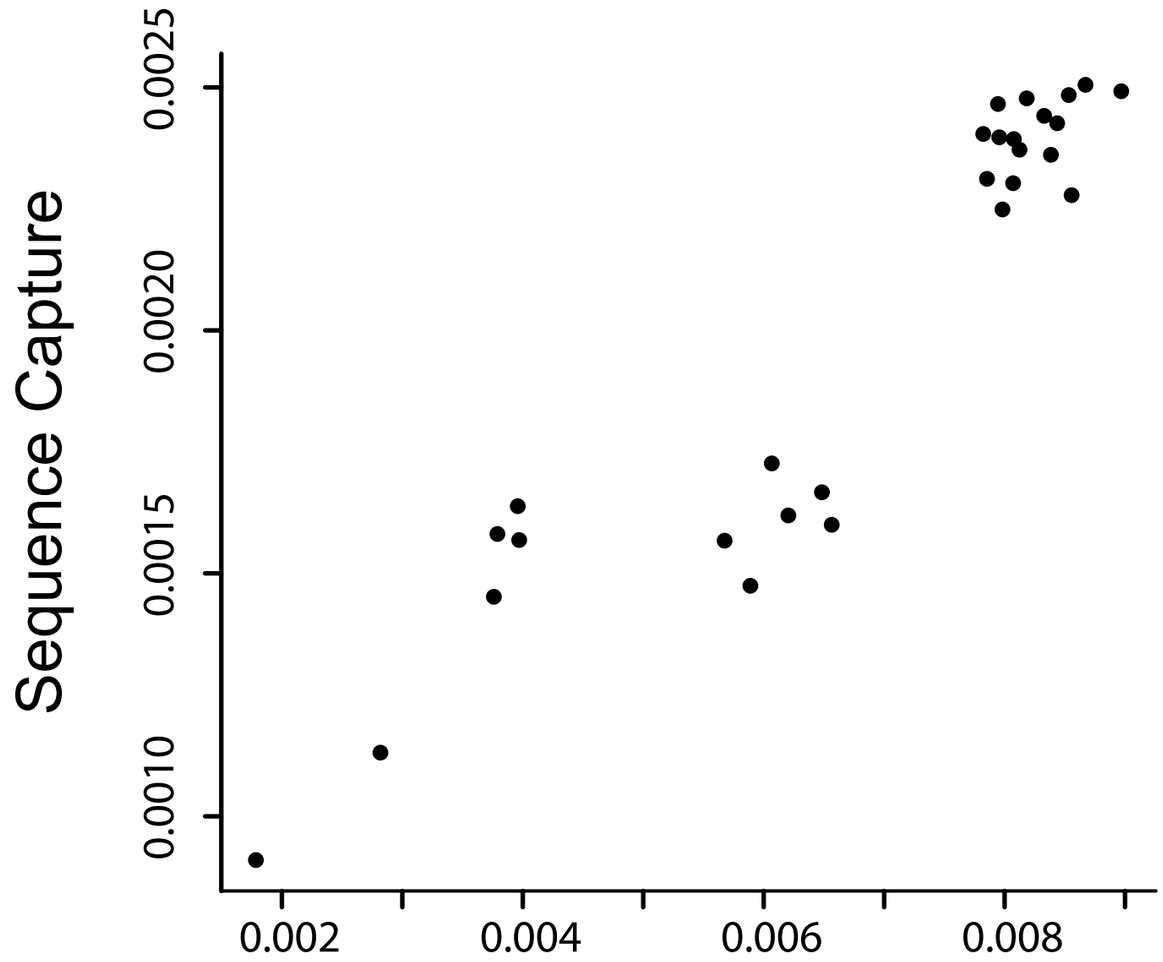
Read Depth

Contig Length

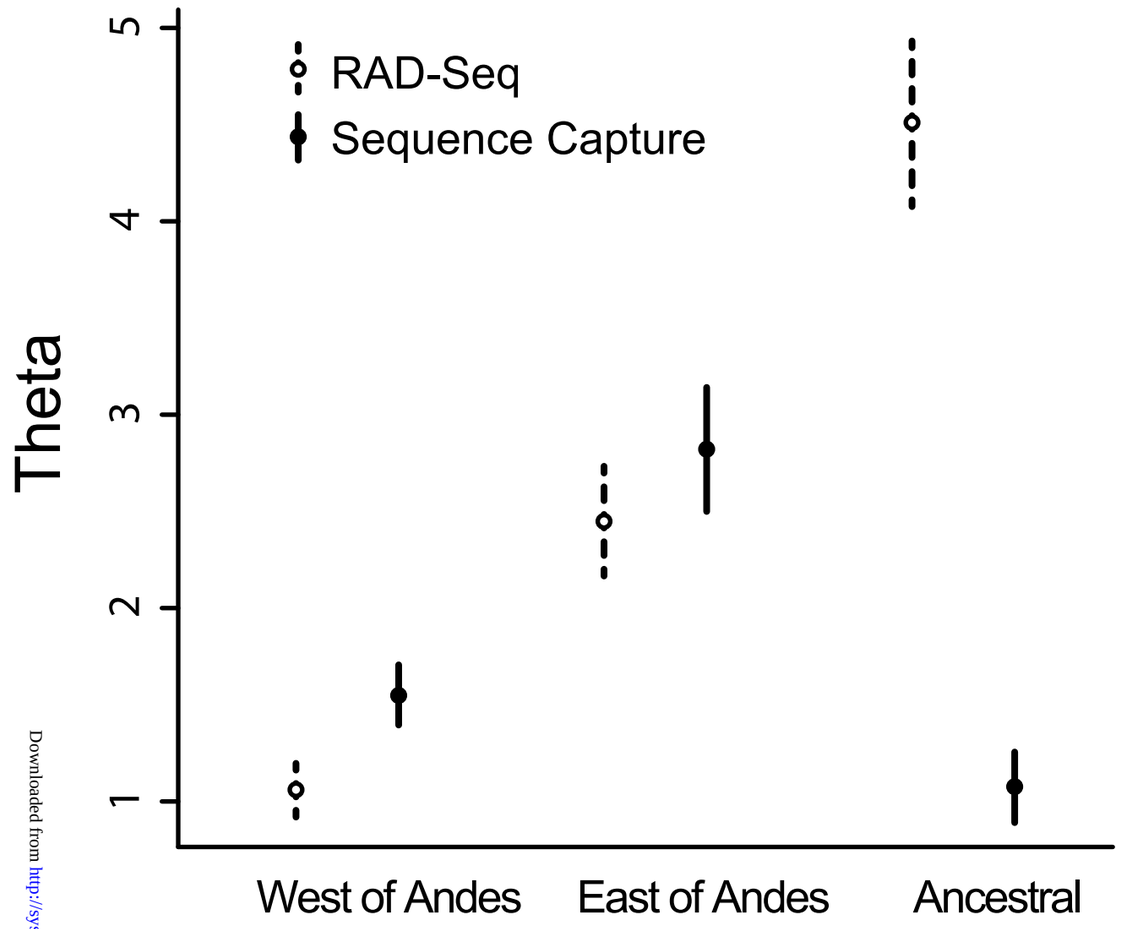




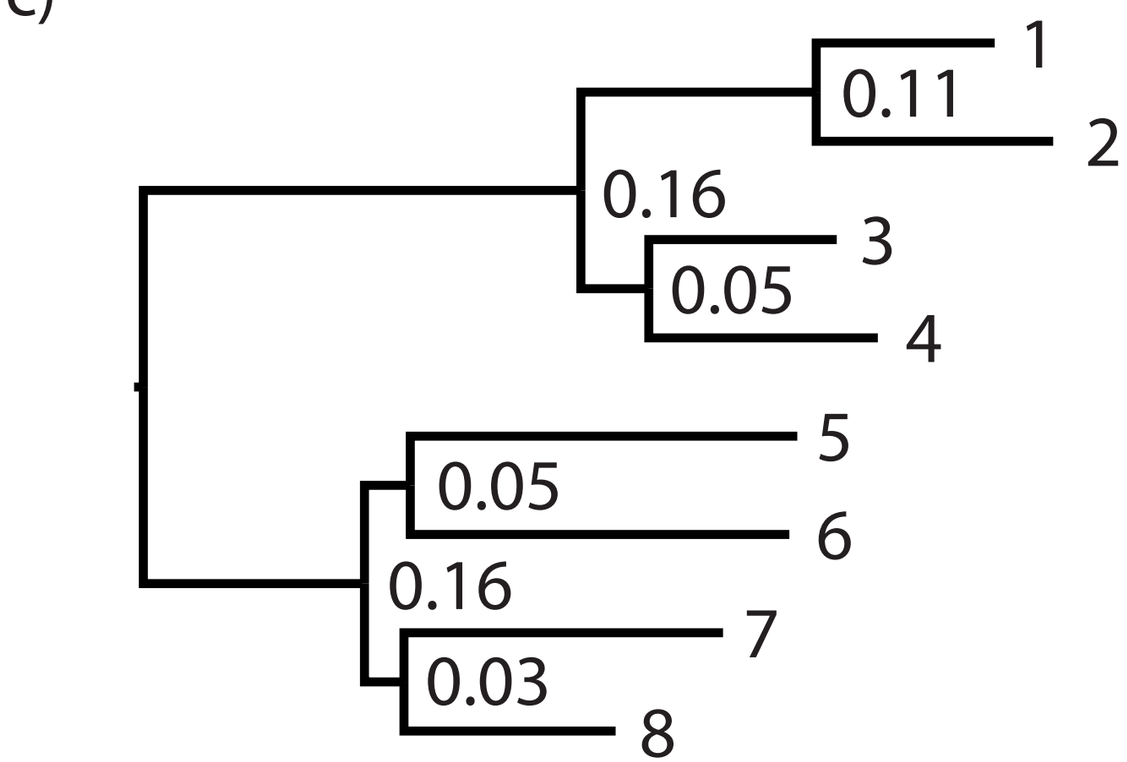
a)



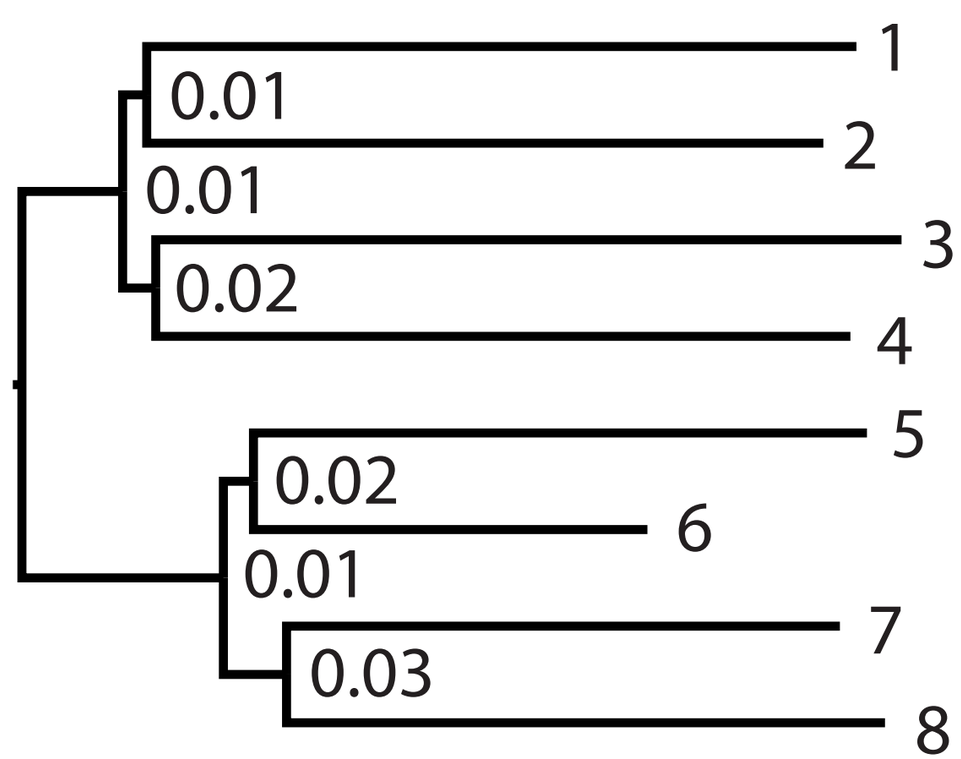
b)



c)



d)



	UCEs	RAD-Seq
Number of Loci	1,358	158,329
Mean Locus Length (sd)	590.36 (209.21)	95.55 (0.62)
Mean Number of Segregating Sites (sd)	4.07 (3.57)	1.35 (1.56)
Mean Number of Alleles (sd)	4.52 (2.88)	2.04 (1.14)
Mean Watterson's Theta (sd)	0.0021 (0.0017)	0.0057 (0.0065)
Mean Tajima's D (sd)	-0.36 (0.82)	0.59 (0.90)

Category	RAD-Seq
Marker distribution and genomic context	**Pro:** Widely dispersed across genome **Con:** Anonymous, evolutionary processes largely unknown
Practical considerations	**Pro:** Less expensive, faster
Assembly and orthology identification	**Pro:** Deep coverage, high read overlap
Variant-calling and genotyping	**Pro:** Fewer rare alleles may make errors easier to distinguish, phasing more straightforward
Information content	Pro: More overall information
Applications	Genome scans, rapid and inexpensive analyses, analyses using species in clades without genomic information, extremely shallow divergences and otherwise intractable relationships.

Sequence Capture
Pro: Can be tailored using new genomic information
Con: Purifying selection impacts allele frequencies
Pro: Works with low-quality and highly contaminated samples
Pro: Over-splitting less problematic
Pro: Fewer low-coverage rare alleles, no allele dropout
Pro: More information per locus
Comparisons across species, calibrating parameter estimates, targeting loci of known utility or interest, studies using poor-quality samples, studies requiring resolved gene trees, deeper phylogenetic studies.