High-Throughput Genomic Data in Systematics and Phylogenetics

Emily Moriarty Lemmon[1] and Alan R. Lemmon[2]

[1]Department of Biological Science, Florida State University, Biomedical Research Facility, Tallahassee, Florida 32306; email: chorusfrog@bio.fsu.edu

[2]Department of Scientific Computing, Florida State University, Dirac Science Library, Tallahassee, Florida 32306; email: alemmon@fsu.edu

Annu. Rev. Ecol. Evol. Syst. 2013. 44:99–121

First published online as a Review In Advance on October 9, 2013

The *Annual Review of Ecology, Evolution, and Systematics* is online at ecolsys.annualreviews.org

This article's doi: 10.1146/annurev-ecolsys-110512-135822

Keywords

high-throughput sequencing, next-generation sequencing, phylogenomics, genomic partitioning, target enrichment, hybrid enrichment, anchored phylogenomics, ultraconserved element enrichment, targeted amplicon sequencing, transcriptome sequencing, RAD sequencing, locus selection, model selection, phylogeny estimation

Abstract

High-throughput genomic sequencing is rapidly changing the field of phylogenetics by decreasing the cost and increasing the quantity and rate of data collection by several orders of magnitude. This deluge of data is exerting tremendous pressure on downstream data-analysis methods providing new opportunities for method development. In this review, we present (*a*) recent advances in laboratory methods for collection of high-throughput phylogenetic data and (*b*) challenges and constraints for phylogenetic analysis of these data. We compare the merits of multiple laboratory approaches, compare methods of data analysis, and offer recommendations for the most promising protocols and data-analysis workflows currently available for phylogenetics. We also discuss several strategies for increasing accuracy, with an emphasis on locus selection and proper model choice.

INTRODUCTION

Within the past decade there has been a dramatic increase in the range of methods available for genome-scale data collection. Beginning with the development of pyrosequencing (Ronaghi et al. 1998) and the release of the first commercial next-generation sequencing machine in 2005 (from 454 Life Sciences), researchers have seen the per-nucleotide cost of DNA sequence data drop precipitously by several orders of magnitude, without a clear lower limit yet in sight. These technological advances have generated great excitement in the systematics community as biologists have realized the tremendous potential for applying high-throughput sequencing to phylogenetic questions. Advances in DNA sequencing technologies have been accompanied by a proliferation of methods for genomic partitioning, which are techniques for enriching sequence libraries for specific regions of a genome (also termed target enrichment; Turner et al. 2009, Mamanova et al. 2010). Using these methods, researchers can focus sequencing effort on loci useful for their particular taxonomic scope, thereby reducing the burden of genome-scale data analysis, increasing the number of taxa that can be processed, increasing cost-effectiveness of projects, and improving the potential for phylogenetic resolution.

High-throughput sequencing: technologies that have dramatically decreased the per-base cost compared with conventional Sanger sequencing; also known as next-generation sequencing

Genomic partitioning: a series of methods for enriching a sequence library for specific regions of a genome

Sanger sequencing: the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication

WGS: whole-genome sequencing

Genomic partitioning strategies vary widely in several properties relevant to phylogenetics, including the taxonomic scope, the proportion of the genome targeted, and the ease of development and application. A subset of these strategies, such as those using hybrid enrichment (e.g., Albert et al. 2007, Gnirke et al. 2009), are rapidly emerging as the most useful for phylogenetics because they can be applied across broad taxonomic scales, thereby simultaneously reducing the burden of marker development and increasing the opportunity for coordination across diverse clades (Faircloth et al. 2012, Lemmon et al. 2012). These methods also provide an unprecedented opportunity for improving phylogenetic accuracy because they permit researchers to choose not only a sufficient number of genomic regions but also the specific genomic regions that are likely to be most useful for phylogenetic questions. One danger with these strategies, however, is the potential to bias the choice of markers in some manner, such as toward those under strong selection (e.g., Katzman et al. 2007). Future development of these methods, which are currently in their infancy, is expected to increase the quality of phylogenetic research across the Tree of Life.

As a result of these advances in genomic data collection, the systematics community has finally reached the frontier of locus selection, where more data can be obtained than are required to resolve the majority of phylogenetic questions. Researchers accustomed to utilizing every possible nucleotide obtained through painstaking and expensive efforts during the Sanger sequencing period bring to the high-throughput sequencing era the desire to include all available data in the hope that a large amount of phylogenetic information will overcome any phylogenetic error that may result from the increased data complexity. Recent phylogenomic studies, however, have dismissed the optimistic notion that large quantities of data will remove the need to ensure high data quality and adequate modeling (Rodriguez-Ezpeleta et al. 2007, Rannala & Yang 2008, Philippe et al. 2011). Researchers recognizing the potential dangers of analyzing large, complex data sets are beginning to wrestle with questions such as, Which genomic regions should be targeted? How many genomic regions should be targeted? How can phylogenetic accuracy be ensured given the high degree of complexity inherent in large-scale data sets? These questions become increasingly important as whole-genome sequencing (WGS) becomes tractable for large numbers of taxa and locus selection shifts to a postsequencing exercise. As the amount of available data begins to surpass the quantity needed to ensure phylogenetic accuracy, it becomes increasingly difficult to justify the inclusion of data that either is of low quality or cannot be modeled correctly. Although phylogenetic accuracy can be difficult to assess in empirical data sets, new methods for assessing

model adequacy (and thus for identifying potentially problematic taxa/loci) promise to ease this difficulty.

Here, we review recent advances in methods for targeted, large-scale data collection and the current workflows available for estimating phylogenies from these data; particular attention is given to ensuring phylogenetic accuracy through data subsampling to improve model fit and reduce systematic error. Given the size and scope of the field of phylogenetics, we restrict the focus of our discussion in several aspects. First, we focus on the generation and analysis of sequence data [as opposed to single-nucleotide polymorphisms (SNPs) or large-scale genomic information such as gene order], because the majority of recent method developments have been directed toward sequence analysis (but see Bryant et al. 2012). Second, we present the phylogenetic data analysis workflow, with particular focus on the post-assembly steps, because recent reviews of primary sequence analysis, such as read assembly, already exist (e.g., Godden et al. 2012, McCormack et al. 2013). Third, although we recognize the importance of taxon selection, we concentrate our survey on locus selection, because recent advances in data collection have had a greater impact on the number of loci that can be obtained as opposed to the number of taxa. After recommending methods for efficient data collection across broad taxonomic scales and proposing a phylogenetic workflow that avoids many of the dangers inherent in large-scale data analysis, we suggest fruitful areas for future research.

Systematic error: lack of accuracy in a phylogenetic estimate primarily as a consequence of model misspecification

SNP: single-nucleotide polymorphism

TAS: targeted amplicon sequencing

Uniplex PCR: when a single primer pair is used to amplify a single locus in a reaction

PCR: polymerase chain reaction

COLLECTION OF HIGH-THROUGHPUT PHYLOGENOMIC DATA

This section focuses on various genomic partitioning strategies that are necessary for phylogenetic studies when sequencing whole genomes is impractical. Here, we review laboratory approaches for targeting loci that are of interest to a particular research question. Some of the available genomic partitioning approaches, such as Cot-based cloning and sequencing (Peterson et al. 2002) and molecular inversion probes (Hardenbol et al. 2003), have not yet been applied to phylogenetics and thus are not included here.

Targeted Amplicon Sequencing or Parallel Tagged Sequencing

Targeted amplicon sequencing (TAS; Bybee et al. 2011), or parallel tagged sequencing (O'Neill et al. 2013), consists of uniplex PCR amplification of target genomic regions followed by multiplexed library preparation and sequencing of amplicon pools via a high-throughput sequencing platform (**Figure 1**). This approach was initially used for human biomedical applications (Craig et al. 2008; see Turner et al. 2009); its development was enabled by advances in library indexing protocols, which allow pooling of large numbers of polymerase chain reaction (PCR) amplicons and samples in a single sequencing lane. Bybee et al. (2011) first applied this approach to deep-level phylogenetics, developing both a wet-laboratory protocol and bioinformatics pipeline to analyze the data. Bybee et al. (2011) generated a phylogenetic hypothesis of Pancrustacea (crustaceans and hexapods) based on 6 loci and 44 individuals, obtaining many well-supported nodes and moderate to low levels of missing sequence data. O'Neill et al. (2013) developed a comparable approach, which they applied to the opposite end of the phylogenetic spectrum to address questions about species delimitation. A novel component of their bioinformatics pipeline included an allele-phasing step, which may be unnecessary at deep phylogenetic levels but essential for shallow-level studies. O'Neill et al. (2013) analyzed their data set in two ways: by converting sequence data to allelic data for population genetics analysis and by using the sequence data directly for species tree analysis. Using 95 loci and 93 individuals with minimal missing data, O'Neill et al. (2013) estimated the relationships among tiger salamander (*Ambystoma tigrinum*) lineages and also

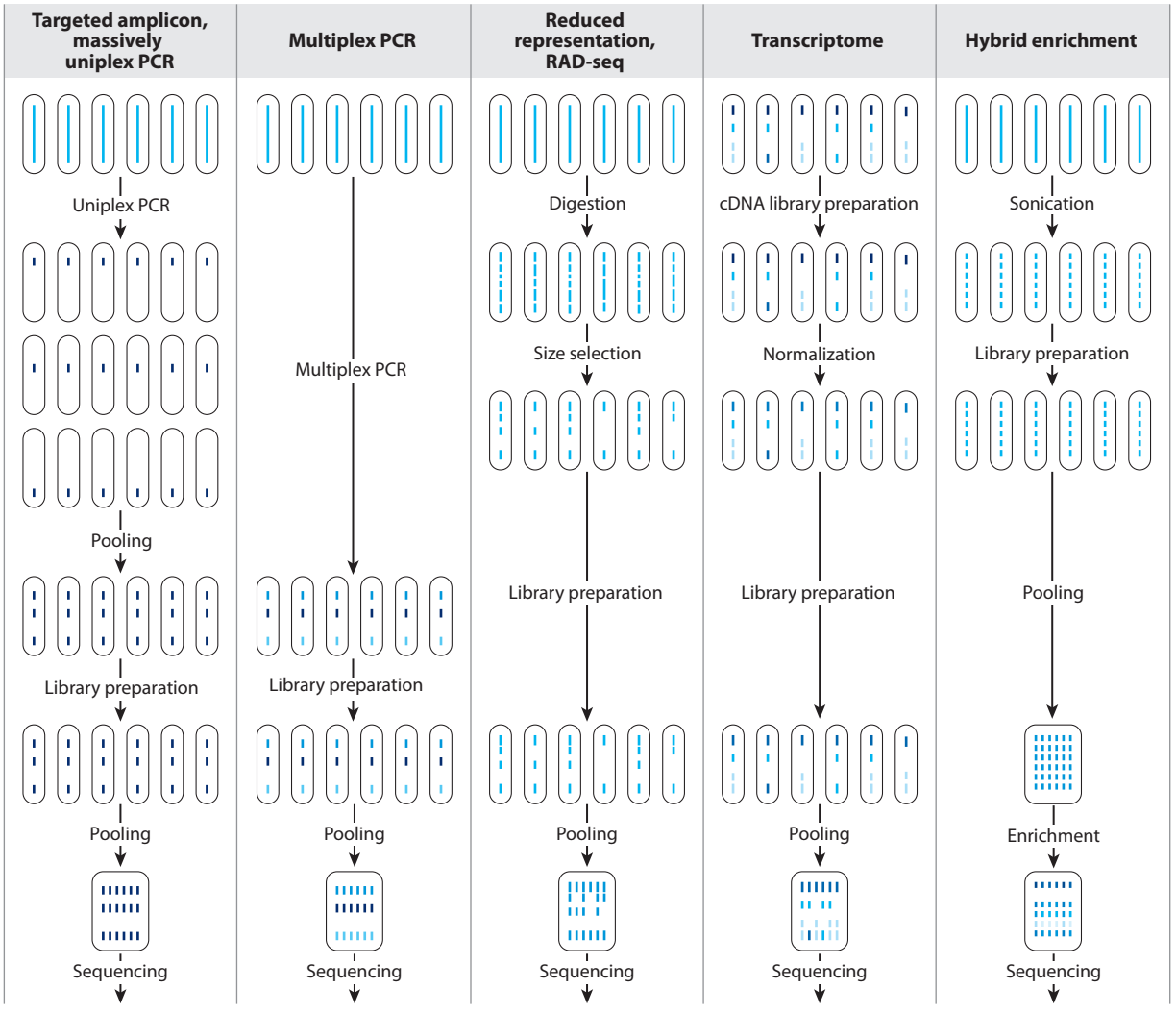


Figure 1

Genomic partitioning workflows for high-throughput phylogenetic data collection. Each oval represents a single sample or pool. Genomic DNA (or RNA, for transcriptome sequencing) is the starting material (*top row*), which undergoes polymerase chain reaction (PCR), enzymatic digestion and size selection, conversion from RNA to cDNA (transcriptome), or shearing (*upper middle rows*), followed by indexed library preparation (*lower middle rows*), pooling across samples (and enrichment in the *rightmost column*), and high-throughput sequencing (*bottom row*). Color intensity (*shades of blue*) indicates relative degree of enrichment of genomic regions during the different stages of each approach. Abbreviations: cDNA, complementary DNA; RAD-seq, restriction-site-associated DNA sequencing.

gained insight into their population structure. The primary disadvantages of these approaches are (*a*) the great amount of time required for marker development and (*b*) the labor-intensive nature of data collection (**Figure 2**).

Multiplex PCR: when primer pairs are combined to amplify multiple loci simultaneously in a single reaction

Multiplex Polymerase Chain Reaction

Multiplex PCR entails combining primers for different loci in the same reaction in order to amplify multiple target loci simultaneously (**Figures 1** and **2**; Chamberlain et al. 1988). This method

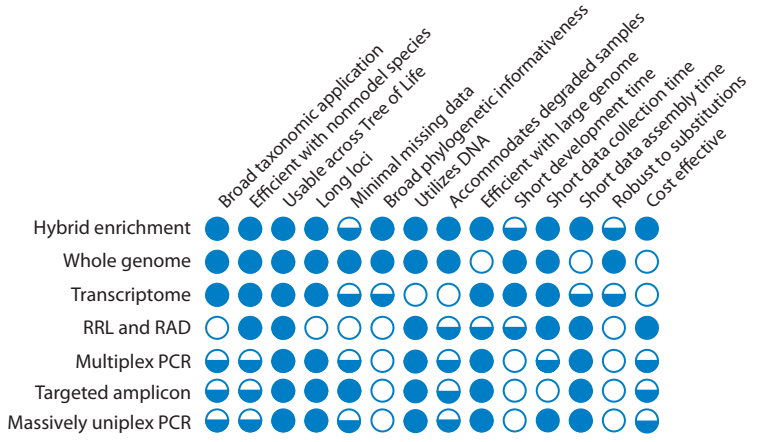


Figure 2

Comparison of attributes of genomic partitioning approaches for phylogenetic data collection. Whole-genome sequencing (WGS) is also included in order to illustrate the upper end of the data spectrum. Filled circles indicate the method performs well for a given criterion, whereas half-filled or hollow circles indicate moderate or poor performance, respectively. Note that assessments assume primers or probes are not currently available and must be developed. *Broad taxonomic application* indicates whether a single initial development effort [i.e., one round of marker design (PCR-based approaches), restriction digest design (RRL/RAD-seq), sequencing (transcriptome and WGS), or probe design (hybrid enrichment)] is sufficient for application of the method to a deep phylogenetic scale. *Efficient with nonmodel species* indicates whether, after initial development, the method can immediately be applied to nonmodel species without further optimization. *Usable across Tree of Life* indicates whether the method is applicable across all portions of the phylogeny of all life. *Long loci* indicates whether the method can be adjusted to obtain long (>1,000 bp) genomic regions. *Minimal missing data* indicates whether the method produces complete data matrices with few missing nucleotide sites or loci. *Broad phylogenetic informativeness* indicates whether, after initial development, the method can produce data that are informative across the spectrum of phylogenetic scales (i.e., from deep-level to intraspecific). *Utilizes DNA* indicates whether DNA (*filled circles*) or RNA (*hollow circle*) is used as starting material. *Accomodates degraded samples* indicates whether the method requires high quality DNA or RNA as starting material. *Efficient with large genome* indicates that the method works well in taxa with large genomes (i.e., up to 10 Gb). *Short development time* indicates the relative amount of time required for initial development of the method for a particular project. *Short data collection time* indicates the relative amount of time required to produce phylogenetic data for a given method. *Short data assembly time* indicates the relative amount of time required for bioinformatic analysis of raw sequence reads to produce alignments for phylogenetic analysis. *Robust to substitutions* indicates whether the method is sensitive to a small or to a moderate number of nucleotide substitutions in the target genomic regions (including primer/probe regions). *Cost effective* indicates whether the per individual cost for large-scale phylogenetic studies on the order of 96 loci × 96 individuals falls into one of three categories: <$200 = filled circle, $200–$500 = half-filled circle, and >$500 = hollow circle. Cost estimates are taken from multiple sources: hybrid enrichment (Lemmon et al. 2012; E. Moriarty Lemmon, unpublished data); whole genome sequencing (N. Chen, Beijing Genomics Institute, personal communication, April 22, 2013); transcriptome sequencing (N. Chen, Beijing Genomics Institute, personal communication, April 22, 2013); reduced-representation library (RRL) sequencing (Lemmon & Lemmon 2012, Lemmon et al. 2012); restriction-site-associated DNA sequencing (RAD-seq) (D. Bolnick, personal communication); multiplex polymerase chain reaction (PCR) (see Turner et al. 2009, Cronn et al. 2012, McCormack et al. 2013); targeted amplicon sequencing (Lemmon et al. 2012); massively uniplex PCR (D. Moon, RainDanceTechnologies, personal communication, February 21, 2013). Additional information on costs and other attributes of these methods are available from Turner et al. (2009), Good et al. (2011), Cronn et al. (2012), and McCormack et al. (2013).

requires much trial and error to optimize reaction conditions across all loci (Edwards & Gibbs 1994). The challenges of multiplex PCR include (*a*) off-target amplification as a consequence of combining many primers from different loci in the same reaction, (*b*) formation of an excess of primer dimers, (*c*) uneven or no amplification of some target loci, and (*d*) low amplification repeatability (Markoulatos et al. 2002). These challenges are magnified as phylogenetic depth increases, owing to problems such as target length variation across distantly related species and increased sequence variation in priming regions across taxa. As a result, although multiplex PCR has been applied to various shallow-level phylogenetic studies (e.g., Stiller et al. 2009), it has not been widely used in phylogenetics. Recent improvements (reviewed by Turner et al. 2009) may provide more opportunities for expansion of multiplex PCR.

RRL: reduced-representation library

RAD-seq: restriction-site-associated DNA sequencing

Massively Parallel Uniplex Polymerase Chain Reaction

Massive parallelization of uniplex PCR has become possible through microdroplet-based PCR enrichment, which uses microfluidics technology to combine reagents with samples in plate-level PCR reactions (Tewhey et al. 2009). Amplification of target regions occurs in millions of picoliter-sized microdroplets. After primer libraries are added to one set of microdroplets, they are merged with a second set containing fragmented genomic DNA template and PCR reagents in a microfluidics chip, where thermocycling takes place. Following amplification, the emulsion that keeps the microdroplets separate is broken, barcodes are ligated to the PCR products, and samples are pooled for high-throughput sequencing (**Figure 1**). This technology has been commercialized by RainDance Technologies (**http://www.raindancetech.com**) and Fluidigm Corporation (**http://www.fluidigm.com**) and has been widely applied to biomedical questions (Taly et al. 2012). Advantages include uniform numbers of target amplicons, high reproducibility, rapid processing, and low labor effort (Tewhey et al. 2009). One limitation is, however, cost. Amplifying 96 samples for 96 loci on a RainDance ThunderStorm system, for example, is ~$18,682 or $195 per sample (D. Moon, RainDance Technologies, personal communication, February 21, 2013), not including the cost of sequencing. A second limitation affects all PCR-based enrichment approaches: the substantial labor required to redesign and test primer pairs for application to each new clade (**Figure 2**). This approach is just beginning to be applied to phylogenetics (H. Mays, D. Weisrock, unpublished data).

Reduced Representation Library Sequencing, Restriction-Site-Associated DNA Sequencing, and Related Methods

A common workflow unites several data collection approaches: reduced-representation library (RRL) sequencing (Altshuler et al. 2000), restriction-site-associated DNA sequencing (RAD-seq; Miller et al. 2007, Baird et al. 2008), and related methods (e.g., CRoPs, Van Orsouw et al. 2007). These approaches involve digestion of genomic DNA samples with restriction enzymes, size selection of a subset of the restriction fragments, library preparation, and high-throughput sequencing of the size-selected fragments (**Figure 1**). The methods have been widely applied for SNP discovery and genotyping for genetic mapping and population genetic studies (Davey et al. 2011, Hohenlohe et al. 2013) and are now being applied to shallow-level phylogenetic and phylogeographic questions (e.g., Emerson et al. 2010, Reitzel et al. 2013).

McCormack et al. (2012) tested the RRL method in 20 individuals from each of four bird groups. Although they recovered 1,000–2,000 loci from each group, a low number of loci were obtained across individuals within each group (50–376 loci found in >7 of 20 individuals and 8–67 loci found in >15 of 20 individuals). Despite this limitation, McCormack et al. (2012) were

able to estimate the phylogenetic relationships among samples of two sister genera based on 30 shared loci and recover a tree that was largely consistent with a mitochondrial phylogeny. Lemmon & Lemmon (2012) utilized RRL sequencing to develop a high-throughput approach to marker development for shallow phylogenetic scales. Approximately 53,000–59,000 loci were recovered from three frog samples (two intraspecific samples and one interspecific sample), with 6,339 loci shared between intraspecific samples and >2,400 loci shared across all three congeneric samples. After Lemmon & Lemmon (2012) screened 187 of these loci, L. Barrow, H. Ralicki, S. Emme, E. Moriarty Lemmon (in review) used TAS to obtain data for 27 highly informative loci (~500–600 bp) and estimated a phylogeny for 44 taxa across the entire genus. Wagner et al. (2013) used RAD-seq to obtain sequence data for a 16-species flock of cichlid fish, including 156 individuals. Approximately ~90,000 loci were obtained, although many loci contained only one or a few individuals. Overall, only 23 loci contained data for all 156 individuals, indicating widespread missing data. Analysis of supermatrices of the short loci (65–84 bp each) with varying degrees of missing data resulted in greater species-level resolution and higher support on branches of the phylogeny than previously recovered.

Across all of these studies, the steep drop in the number of loci shared across progressively deeper phylogenetic scales indicates a fundamental limitation of RRL sequencing/RAD-seq to shallow-phylogenetic and phylogeographic questions. This problem results largely from evolution of restriction sites across taxa, particularly across deeper phylogenetic scales, which leads to inconsistent locus recovery during size selection and, consequently, missing data (**Figure 2**; Rubin et al. 2012). The issue of missing data may be partially mitigated through techniques such as double-digest RAD-seq (Peterson et al. 2012). Other limitations of RAD-seq are reviewed elsewhere (Davey et al. 2012, Arnold et al. 2013).

Transcriptome Sequencing

Transcriptome sequencing (also termed RNA-seq) consists of extracting whole RNA of an organism from a specific tissue or set of tissues, reverse transcribing complementary DNA (cDNA) from the RNA, and sequencing the cDNA on a high-throughput sequencing platform (**Figure 1**; Wang et al. 2009). This technique has been applied to phylogenetic questions because it produces a large number of exons and noncoding regions (derived from noncoding RNAs) flanking the exons that can be informative at various taxonomic scales. Several major initiatives are currently underway to collect data for thousands of transcriptomes, including plants (1KP Project, **http://www.onekp.com**; Medicinal Plant Transcriptome Project, **http://www.uic.edu/pharmacy/MedPlTranscriptome**), insects (1KITE, **http://www.1kite.org**), and microbial eukaryotes (Marine Microbial Eukaryote Transcriptome Sequencing Project, **http://www.marinemicroeukaryotes.org/**). Phylogenies derived from transcriptome data are being published for diverse taxa at an accelerating rate (e.g., arthropods, Oakley et al. 2012; land plants, Timme et al. 2012).

There are several challenges to transcriptome sequencing (Ozsolak & Milos 2011). First, fresh high-quality RNA is required as starting material, rendering many existing samples in tissue collections unusable, which is especially a problem for rare or extinct organisms. Second, because only a portion of the genome is expressed in a particular tissue at a given time, RNA must be sampled from the same types of tissue and from individuals at the same life-history stage to obtain as many orthologous loci across samples as possible. Third, because expression levels vary considerably across genes, high sequencing depth may be required to obtain sequences from weakly expressed genes. Finally, transcriptome assembly can be computationally challenging because alternative splicing of RNA following transcription results in isoforms derived from different

combinations of exons (**Figure 2**; Martin & Wang 2011, Ozsolak & Milos 2011, Cronn et al. 2012, Godden et al. 2012).

UCE: ultraconserved element

AE: anchored enrichment

Hybrid Enrichment

Hybrid enrichment (sequence capture) involves synthesis of ~60–120 bp oligonucleotide sequences (capture probes) that are complementary to target regions in the genome, hybridizing them to a DNA library and isolating the targets from the genome prior to high-throughput sequencing (**Figure 1**). Hybrid enrichment was originally performed in solid phase using microarrays (e.g., Albert et al. 2007) but is now more commonly performed in solution phase using biotinylated RNA probes or "baits" (Gnirke et al. 2009). The biotinylated probes, which form RNA-DNA complexes with target regions, are then captured with magnetic streptavidin-coated beads, unbound DNA is washed from the beads, and the captured DNA is then amplified via PCR using primers specific to adapters added during DNA library construction (Gnirke et al. 2009). Probe libraries can be custom designed by the researcher and purchased from companies such as Agilent Technologies (**http://www.agilent.com**) and MYcroarray (**http://www.mycroarray.com**) or designed by NimbleGen (**http://www.nimblegen.com**).

Hybrid enrichment approaches to genomic partitioning include several methods (**Figure 3**): anchored enrichment (AE; Lemmon et al. 2012), ultraconserved element (UCE) enrichment (Faircloth et al. 2012), exon enrichment (Burbano et al. 2010, George et al. 2011, Bi et al. 2012, Hancock-Hanser et al. 2013, Li et al. 2013), organelle enrichment (Briggs et al. 2009, Maricic et al. 2010, Mason et al. 2011, Cronn et al. 2012, Guschanski et al. 2013, Stull et al. 2013), PCR bait capture (Maricic et al. 2010; J. Penalba, L. Smith, M. Tonione, C. Sass, S. Hykin, P. Skipwith, J. McGuire, R. Bowie, C. Moritz, in review), filter-based hybridization capture (Herman et al. 2009), and primer-extension capture (Briggs et al. 2009). Hybrid enrichment has shown promising results for ancient DNA and other degraded samples (Briggs et al. 2009, Burbano et al. 2010). These methods share a similar workflow that includes a bioinformatic pipeline for designing capture probes and a laboratory phase, which includes library preparation, hybrid enrichment, and high-throughput sequencing. Drawbacks of these approaches include the high initial costs of investing in equipment and reagents for the laboratory protocol and the bioinformatic expertise required to develop probe sets and analyze raw sequence data (**Figure 2**). These challenges can be lessened by pooling resources across laboratories and/or by processing samples through a central facility.

The key to extending hybrid enrichment technology to broad-scale phylogenetics was determining how to design probes that not only capture successfully the model species from which the probes were designed but also capture effectively across a broad taxonomic group, including nonmodel species. Lemmon et al. (2012) and Faircloth et al. (2012) published alternative solutions to this problem. Lemmon et al. (2012) identified conserved DNA regions flanked by less conserved regions across five phylogenetically dispersed vertebrate genomes (model species). They designed a probe library targeting the conserved regions or anchors of 512 target loci, which contained a mix of probes derived from each of the five genomes, representing natural variation across lineages (AE). After applying the probe set to the five model species and five nonmodel species, they obtained hundreds of loci with high phylogenetic-information content from all species and estimated a fully resolved species tree of amniotes and vertebrates, respectively. Faircloth et al. (2012), alternatively, utilized UCEs identified in the genomes of two birds and one lizard for probe design in amniotes (UCE approach). As originally defined by Bejerano et al. (2004), UCEs are segments of mammalian genomes >200 bp that are perfectly conserved between human and rodents. Faircloth et al. (2012), however, refer to a UCE as any conserved DNA sequence with ≥80% identity over

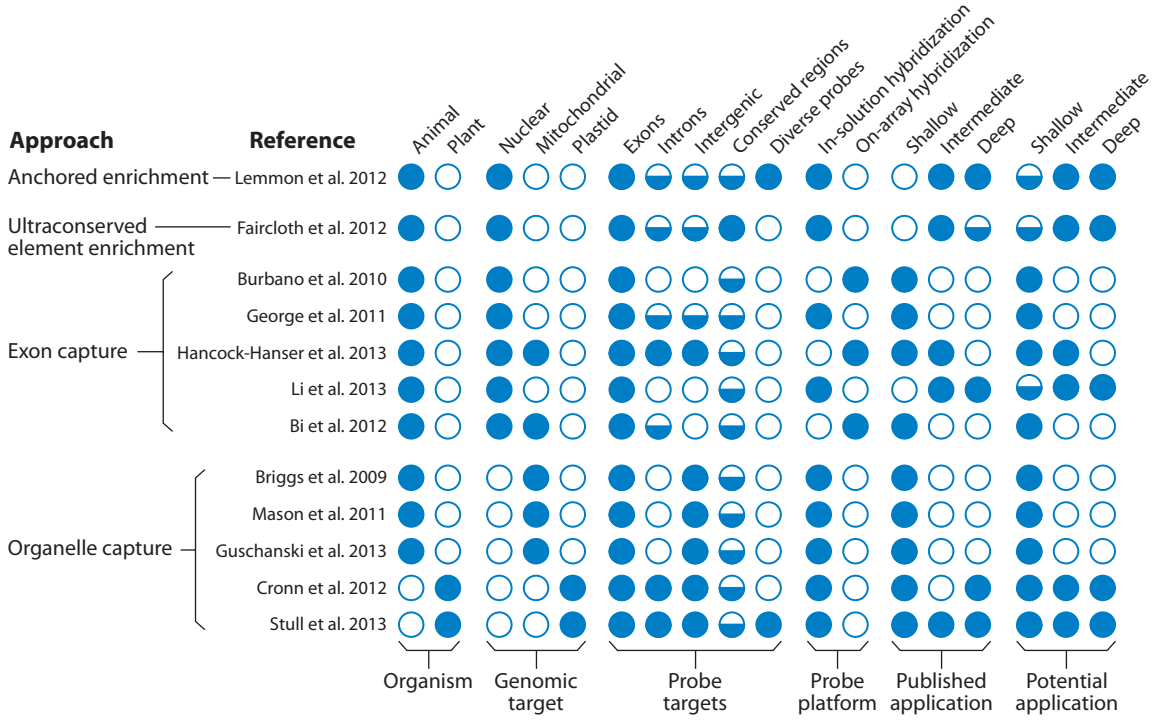


Figure 3

Comparison of the attributes of cross-species hybrid enrichment approaches for phylogenetic data collection. A filled circle indicates an approach has the particular attribute, whereas a half-filled or hollow circle indicates partial presence or absence of the attribute, respectively. Organism and genomic target columns indicate the target organisms and classes of loci for which a given method was designed. Probe targets and probe platform indicate the genomic regions targeted and hybridization method used. Published application columns demonstrate the phylogenetic scale to which the approach was applied in the publication. Potential application columns indicate the phylogenetic scale to which an approach could be applied in the future, given the constraints of the published probe design.

≥100 bp across taxa. UCEs are thought to be involved in gene regulation and development (e.g., Woolfe et al. 2005, Pennacchio et al. 2006) and to experience strong purifying selection (e.g., Katzman et al. 2007). Faircloth et al. (2012) designed a probe library targeting 2,386 UCEs (with probe sequences derived from the chicken genome) and performed hybrid capture in nine bird species, recovering >1,000 loci with moderate levels of phylogenetic information and obtaining a partially resolved species tree. From the research groups developing AE and UCE approaches, a burgeoning phylogenetic literature is emerging across a wide spectrum of taxa.

Comparison of Data Collection Approaches

Among the currently available genomic partitioning approaches, hybrid enrichment methods show the greatest utility for application across a wide phylogenetic spectrum (**Figure 2**). The combination of several critical factors including (*a*) high efficiency in nonmodel species; (*b*) flexibility in types, sizes, and numbers of target regions; (*c*) high phylogenetic-information content of enriched loci across shallow to deep taxonomic scales; (*d*) potentially low levels of missing data; (*e*) rapid rate of data collection; and (*f*) cost-effectiveness makes these methods useful for systematists working in nearly any taxon. Furthermore, because the same loci can be targeted across broad

taxonomic groups, these methods facilitate meta-analyses, thus accelerating resolution across the Tree of Life. For shallow-scale phylogenetics and phylogeography, RAD-seq is being successfully applied, improving resolution of species complexes and intraspecific relationships despite moderate to high levels of missing data in many studies. Other methods have various limitations, such as high cost (WGS, transcriptome sequencing), marker development challenges (multiplex PCR, massively uniplex PCR), narrow taxonomic applicability (massively uniplex PCR), or high labor investment in data collection (TAS), which lessen their impact on the field of phylogenetics.

Stochastic error: lack of precision in a phylogenetic estimate as a consequence of insufficient phylogenetic information

ANALYSIS OF HIGH-THROUGHPUT PHYLOGENOMIC DATA

Data Selection

Until recently, the prohibitively high cost of collecting genome-scale data has compelled systematists to utilize all available data. As the cost of data has decreased and the quantity has increased, however, we are quickly surpassing the point at which utilization of all available data is the optimal strategy. As the quantity of phylogenomic data increases, so does the quantity of potential sources of phylogenetic error because the set of all available data is necessarily increasingly complex (Philippe et al. 2005). The two primary types of phylogenetic error include stochastic error and systematic error (Swofford et al. 1996). Stochastic error (lack of precision) results from insufficient phylogenetic information and should decrease as the quantity of data increases. Systematic error (lack of accuracy) results primarily from model misspecification and may increase as the quantity of data increases because adequate modeling becomes increasingly difficult (Philippe et al. 2005, Kumar et al. 2012). Therefore, the possibility of obtaining a strongly supported but incorrect phylogenetic estimate may actually increase with an increasing amount of data (e.g., O'Neill et al. 2013) unless systematic error is minimized. For example, when estimating a gene tree from a single genomic region, increasing the number of sites decreases the stochastic error (increased support for a phylogeny). At some point, however, increasing the number of sites also begins to introduce systematic error if unmodeled processes, such as recombination, become increasingly important (Rannala & Yang 2008). Similarly, including a large number of loci in a species tree analysis tends to reduce stochastic error present in the species tree, but inclusion of genes subject to high stochastic error (due to insufficient phylogenetic information for gene tree resolution) may render convergence on a solution intractable. Because the magnitude of stochastic error, systematic error, and phylogenetic information are highly dependent on the properties of the data being analyzed, the most effective way to increase phylogenetic accuracy may be through data subsampling.

The primary aim in data subsampling is to increase the ratio of phylogenetic information to systematic error. Although systematic error may have the greatest effect when phylogenetic information is weak (e.g., Lemmon et al. 2009), some types of systematic error can overwhelm substantial phylogenetic information (Rodriguez-Ezpeleta et al. 2007, Philippe et al. 2011). The key then is to find the subset of data with sufficient information to resolve the phylogenetic history while minimizing systematic error. Data subsampling can occur at several stages in the phylogenomic pipeline (**Figure 4**) (**Supplemental Table**; follow the **Supplemental Material link** from the Annual Reviews home page at **http://www.annualreviews.org**), from selection of orthologous sequences to selection of individual sites within alignments.

Reducing error through selection of orthologous sequences. Most methods used in phylogenetics and phylogeography assume that the sequences in each alignment are orthologous. Under the Sanger sequencing paradigm, researchers invested substantial effort to utilize single-copy loci

in order to avoid obtaining misleading results or interpretations stemming from analysis of paralogous loci. By choosing single-copy loci, or by designing primers specific to one of multiple gene copies, researchers selected orthologous sequences for phylogenetic analysis a priori. Analyses of whole genomes have revealed that gene duplication and loss is widespread, suggesting that few single-copy loci are present in genomes (because the probability of observing a gene duplication and/or loss increases as the number of species considered in a study increases). Dehal & Boore (2005), for example, found that two genome-wide duplication events were followed by the loss of between zero and three gene copies in different lineages during the diversification of vertebrates, resulting in a "Swiss cheese" gene matrix across taxa. This example illustrates the difficulty of selecting loci a priori that have not undergone duplication or loss at some point during the evolutionary history of a clade. Although deep duplication events usually produce homologs for which orthology assessment is straightforward, shallow duplication events (i.e., within the time span of the phylogeny being estimated) can produce homologs for which orthology assessment is difficult.

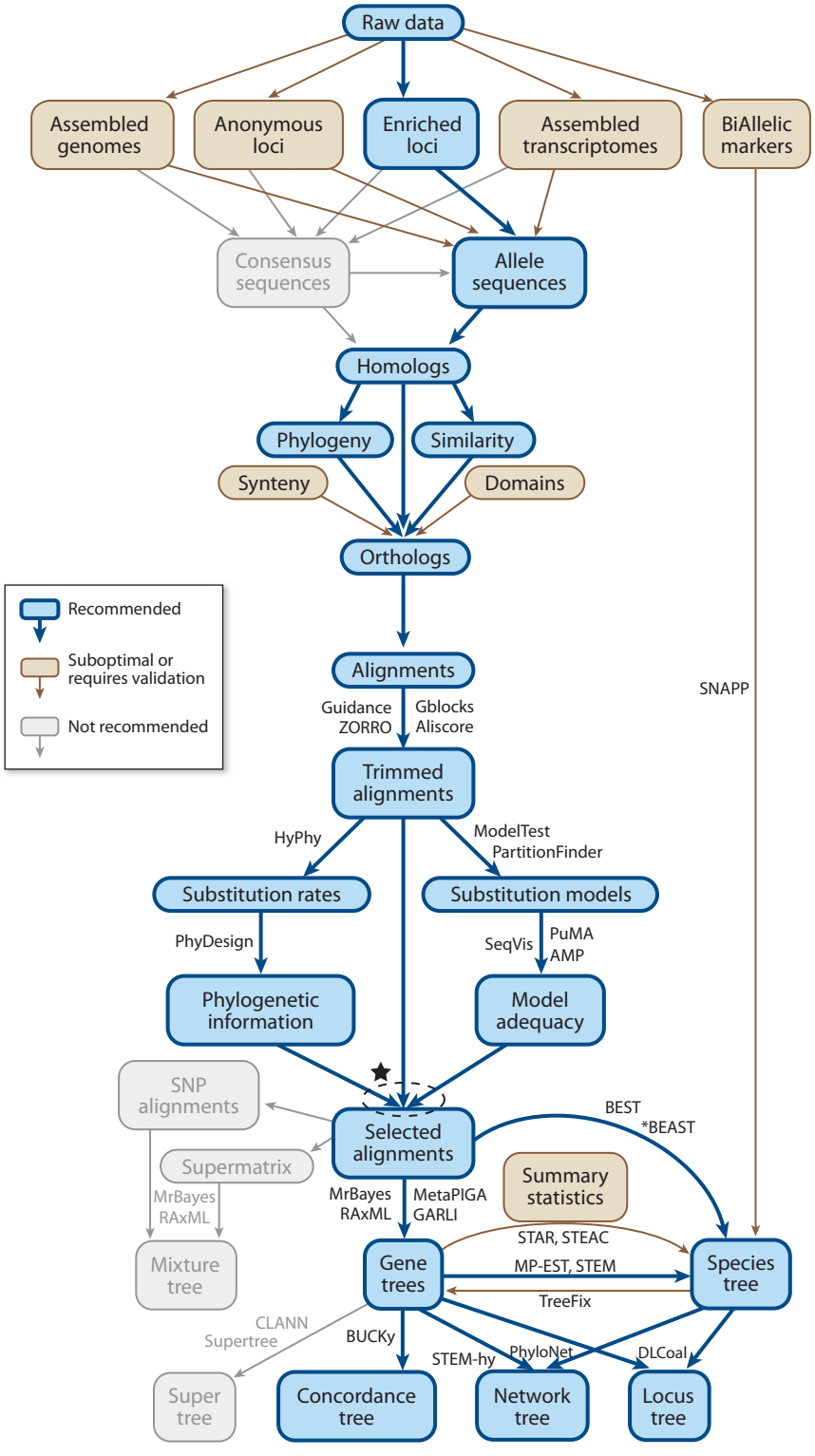
One solution to the difficulty of targeting single-copy loci is to sequence all homologs using high-throughput data collection techniques (i.e., via hybrid enrichment) and to establish orthology a posteriori using tree-, sequence similarity–, and/or synteny-based approaches (Gabaldón 2008, Kristensen et al. 2011). Tree-based methods explicitly model the evolutionary history of the genes with the goal of reconciling individual gene trees in the context of a species tree (e.g., Rasmussen & Kellis 2012, Boussau et al. 2013). Similarity-based methods, in contrast, do not explicitly model gene duplication and loss but instead rely on the assumption that orthologous genes are more similar to each other than paralogous genes. Typically, pairwise sequence similarity is measured for a set of homologs (e.g., using reciprocal BLAST scores), and orthologous sets are determined by clustering genes based on the similarities (Overbeek et al. 1999). Recent approaches increase the accuracy of orthology detection by focusing on protein domains in order to circumvent problems caused by recombination (Makarova et al. 2007). Accuracy may also be increased by considering synteny (similarity of gene order/position across individuals or species; Zheng et al. 2005), but this approach is potentially limited to shallow comparisons because genome evolution can quickly erode synteny through time.

Consensus sequence: the sequence that represents an individual and is obtained by combining two or more alleles for a locus

Coverage: the number of sequence reads covering a given nucleotide position

Reducing error through use of allele sequences. Coalescent-based species tree analyses (see below) implicitly assume that the sequences being analyzed represent individual gene copies (e.g., alleles). One common practice, however, is to collapse assembled reads into a single consensus sequence to represent an individual for a given locus. Using consensus sequences instead of separate alleles is an obvious case of model misspecification that could result in systematic error, although to our knowledge simulation studies quantifying this error have not yet been published. The use of alleles instead of consensus sequences may reduce stochastic error by allowing the population sizes of extant species to be estimated more accurately (see the section titled Phylogeny Estimation below). Although separation of alleles (haplotypes) previously required laborious cloning techniques or analysis of multiple samples from the same population (e.g., Stephens et al. 2001, Browning & Browning 2011), high-throughput sequencers now produce data that allow alleles to be phased bioinformatically (given adequate coverage and either long reads for single-end sequencing or a sufficiently broad fragment length distribution for paired-end sequencing) because each sequencing read is derived from a single chromosome (e.g., Bansal & Bafna 2008, O'Neill et al. 2013).

Reducing error through increasing alignment accuracy. Reducing alignment error is critical to reducing phylogenetic error (Felsenstein 2004, Susko et al. 2005). Alignment error is most



pronounced when high insertion-deletion rates produce substantial gaps and/or when high substitution rates produce saturated sites (Gatesy et al. 1993). Both sources of error can result in random alignment of nonhomologous regions (Misof & Misof 2009) that introduce model misspecification at a fundamental level. The most obvious way to reduce alignment error for highly divergent protein coding sequences is to construct amino acid alignments instead of nucleotide alignments. Iterative methods for improving alignment quality may also reduce error for highly divergent sequences (e.g., SATé; Liu et al. 2012). Regions that cannot be aligned with a high degree of confidence should be removed prior to downstream analysis. One growing challenge is ensuring high-quality alignments as the number of loci included in each study begins to exceed the number that can be manually inspected.

Reducing error through selection of informative loci. Recent simulation studies suggest that the number of loci needed to resolve a phylogeny can range from several to hundreds depending on several factors, including population size, time between speciation events, and the properties of the loci being considered (Knowles & Kubatko 2010, Leaché & Rannala 2011, Liu & Yu 2011). The exception to this general rule does exist: Thousands of loci may be required to resolve very short branches (Liu et al. 2010). Because an excess of data will soon be available for many systems, locus selection is expected to become one of the most effective means by which systematists may increase the ratio of phylogenetic information to error. One criterion by which loci can be distinguished is evolutionary rate, which determines the particular timescale for which a locus is most informative. Because loci evolve at different rates, only a subset of loci is likely to be well suited for phylogenetic questions pertaining to a given timescale. Although some work has been conducted in this area (Townsend 2007), development of a diversity of metrics for quantifying phylogenetic information will become increasingly valuable as the number of available loci begins to exceed the number that can be analyzed appropriately.

Model Selection

Proper model selection has long been recognized as a critical prerequisite to obtaining accurate estimates of phylogeny (Felsenstein 1978). There are numerous aspects of model selection that must be considered, two of which are mentioned here. First, the model of sequence evolution assumed should be appropriate (Posada & Crandall 1998, Lemmon & Moriarty 2004, Sullivan & Joyce 2005). Because phylogenetic error can increase as a result of model under- or overparameterization (Brown & Lemmon 2007), it is important to identify a partitioning strategy that captures model variation across sites (and/or branches) but does not introduce unnecessary parameters. More appropriate levels of model complexity can be obtained through mixture models (Pagel & Meade 2004) or data partitioning (Yang 1996). New computational tools offer powerful ways of

Figure 4

Workflows for phylogenomic data analysis. Types of data at different stages of analysis are presented in boxes, whereas methods of analysis are represented by arrows. Workflows we recommend for adoption and future development are indicated by thick blue arrows and thick blue lines around boxes. The star indicates methods especially needing development. For steps following alignment, examples of some commonly used and/or promising analysis programs are given next to arrows (see the **Supplemental Table** for full citations of methods; follow the **Supplemental Material link** from the Annual Reviews home page at **http://www.annualreviews.org**). For reviews of prealignment steps, see Notredame (2007), Li & Homer (2010), Kemena & Notredame (2009), Kristensen et al. (2011), Godden et al. (2012), and McCormack et al. (2013).

reducing systematic error for large data sets because they allow optimal partitioning strategies to be efficiently and accurately selected from predefined character sets (e.g., Lanfear et al. 2012).

A second important consideration is recombination, which reduces the correlation of coalescent histories across loci, thus contributing to gene tree discordance (Rannala & Yang 2008). Consequently, model misspecification resulting from concatenation of unlinked loci can result in increased phylogenetic error (Degnan & Rosenberg 2006, Edwards et al. 2007, Kubatko & Degnan 2007, Edwards 2009, Heled & Drummond 2010, Leaché & Rannala 2011, Weisrock et al. 2012). The effects of recombination in transcriptome data sets may be particularly underappreciated because sites that are adjacent in an alignment may actually be quite distant in the genome owing to the existence of unsequenced introns. Discordance across loci can be accommodated using species tree models that account for unlinked coalescent histories (Maddison 1997, Maddison & Knowles 2006, Edwards et al. 2007; and see below). Recombination within loci, if sufficiently infrequent, may introduce less phylogenetic error than other factors, such as the number of loci and individuals sampled in a study (Lanier & Knowles 2012).

Once a preliminary model has been chosen, tests should be performed to determine whether it adequately captures salient features of the evolutionary process. A battery of tools has been developed to assist in detecting model violations and for testing the fit of the model to the data. Some tools test for signatures of processes that violate specific model assumptions, such as recombination within or among loci (McGuire et al. 1997, Kosakovsky Pond et al. 2006), some types of selection (Delport et al. 2010), hybridization (Yu et al. 2011), and heterotachy (Pagel & Meade 2008). Other tools test for overall model adequacy (Bollback 2002, Brown & ElDabaje 2009, Reid et al. 2013), though the systematics community has yet to fully embrace these methods. If model violation is detected, a more appropriate model should be chosen. For example, if gene trees are discordant, loci should not be concatenated for phylogenetic analysis, but instead a model incorporating the appropriate processes (i.e., incomplete lineage sorting and/or hybridization) should be applied. If the model and/or data cannot be adjusted to remove the model misspecification, and phylogenetic error is known to result from that type of misspecification, then the offending subset of data should be removed. This process might involve removing loci under strong selection, sites within alignments that are saturated (Castresana 2000, Rodriguez-Ezpeleta et al. 2007, Goremykin et al. 2010, Philippe et al. 2011), or sites that contain ambiguous characters (Lemmon et al. 2009).

Phylogeny Estimation

The realization that many common biological processes cannot be adequately captured with traditional one-locus phylogenetic models has driven the rapid growth of more sophisticated models that accommodate complex biological processes. Development of these models has coincided with the generation of large multilocus data sets through WGS or genomic partitioning strategies. Whereas the set of methods that can be applied is defined by the type and scale of available data, the set of methods that should be applied is determined by both the biological process that gave rise to the data and the particular phylogenetic questions being addressed. The challenge lies in identifying the methods at the intersection of these two sets. Below, we describe the types of analyses that can be applied to most of the commonly collected data sets and outline the various types of phylogenetic trees that can be estimated to answer a variety of phylogenetic questions.

Modern phylogenetic methods typically rely on one of three types of data: sequence alignments, gene trees, and summary statistics. Methods relying on *sequence alignments* generally include a full statistical model of the evolution of sequences in the context of a gene region or a species. Maximum likelihood–based approaches, such as those implemented in RAxML (Stamatakis et al. 2005), are generally tractable for large data sets containing thousands of genes and/or thousands of taxa,

but typically rely on bootstrapping procedures to determine uncertainty in parameter estimates. Bayesian approaches, such as those implemented in BEST (Edwards et al. 2007), *BEAST (Heled & Drummond 2010), and MrBayes (Ronquist & Huelsenbeck 2003), provide direct estimation of parameter uncertainty but are typically only tractable for moderate or small data sets containing tens or hundreds of loci and/or individuals. Methods that utilize *gene tree distributions* to estimate other types of phylogenies (see below) tend to be much more computationally efficient than those relying on sequence alignments directly because the gene tree uncertainty is assumed to be negligible. Maximum likelihood–based methods of this type, such as MP-EST (Liu et al. 2010), are quite efficient and can be used to estimate phylogenies for hundreds of taxa and thousands of loci. Bayesian approaches of this type, such as BUCKy (Ané et al. 2007), are only tractable, however, when the data set contains a few dozen taxa and/or loci. The most computationally efficient methods (but not necessarily the most accurate) rely on *summary statistics*. Methods of this type, such as STEAC (Liu et al. 2009), can be applied to data sets containing thousands of loci and/or taxa. An important shortcoming of these methods, however, is that they estimate only topologies and not branch lengths, which is a major disadvantage for many downstream phylogenetic applications.

Answering phylogenetic questions typically requires the estimation of one or more types of phylogenetic trees. Proper interpretation of these estimates requires a clear understanding of the meaning of each type of tree that is estimated. Here we distinguish among seven commonly estimated types of trees: gene trees, mixture trees, species trees, supertrees, concordance trees, phylogenetic networks, and locus trees. The most familiar tree type is the *gene tree*, which reflects the pattern of ancestry of a set of homologous gene copies derived from a single genomic region within which recombination has not occurred along the lineages connecting the gene copies. We use the more inclusive term homologous instead of orthologous here to allow for gene duplication (but see *locus tree* below). Models used to estimate phylogenies of this type are those implemented in software such as PAUP (Swofford 2000), RAxML (Stamatakis et al. 2005), and MrBayes (Ronquist & Huelsenbeck 2003), which explicitly model nucleotide, amino acid, or other character evolution. It is important to emphasize that methods estimating gene trees assume recombination-free loci. Gene trees are typically estimated directly from sequence alignments.

We define a *mixture tree* as a phylogeny reflecting the similarity of a set of homologous gene copies derived from *more than one* genomic region among which gene tree discordance exists but is not explicitly modeled. Mixture trees, for example, are the result of supermatrix analyses in which alignments from multiple unlinked genomic regions are concatenated into a single matrix and analyzed using a model that assumes no recombination within the matrix (e.g., those employed in RAxML, MrBayes, etc.). For the purpose of this discussion, concatenated SNP alignments and transcriptome alignments of multiple exons are considered to be supermatrices. Mixture trees are difficult to interpret because they are not guaranteed to reflect the species tree (Degnan & Rosenberg 2006, Edwards et al. 2007, Kubatko & Degnan 2007, Heled & Drummond 2010, Leaché & Rannala 2010) or an equally-weighted average of gene trees (because different genes contain different amounts of phylogenetic information). In fact, mixture trees do not have a straightforward biological interpretation in contrast to gene trees or species trees. The problems associated with concatenation become especially apparent when separate alleles (instead of a single consensus sequence) are available for each individual, because there is no biologically justifiable way to concatenate alleles across unlinked loci. Attempting to do so can result in strong support for incorrect phylogenies (Weisrock et al. 2012).

Species trees reflect the relationships among species and account for processes such as incomplete lineage sorting that result in discordant gene trees across loci. Species trees can be estimated directly from sequence alignments (e.g., BEST: Edwards et al. 2007; *BEAST:

Heled & Drummond 2010), directly from biallelic markers (SNAPP: Bryant et al. 2012), indirectly from a set of gene trees (e.g., STEM: Kubatko et al. 2009; MP-EST: Liu et al. 2010), or indirectly using summary statistics (e.g., STAR, STEAC: Liu et al. 2009). Owing to the complex nature of joint gene tree and species tree estimation, analyses involving large numbers of loci and/or taxa can be computationally intensive, especially if a substantial number of the loci are uninformative. Two factors affecting analysis efficiency are often overlooked: the number of individuals sampled per species and the use of consensus sequences. Performance can be improved through the inclusion of multiple individuals (or alleles) for each species because population size estimates require at least two sequences per species. When only one sequence is provided for a particular species, no information pertaining to the population size of that terminal species is available, and thus the Markov chain samples from the broad prior distribution, increasing sampling error. Wherever possible, separate alleles (instead of consensus sequences) should be used because the model assumes that each sequence represents a single gene copy. For an in-depth discussion of species trees and their importance in phylogenetics, see Edwards (2009).

Supertrees are generated through heuristic algorithms designed to synthesize individual gene trees (or other types of phylogenies) containing incomplete but overlapping sets of taxa (Bininda-Emonds et al. 2002). Supertree methods are typically employed when a single analysis of the raw data is intractable owing to the taxonomic or temporal scope of the project (e.g., Pisani et al. 2007). One benefit to estimating supertrees is that phylogenies estimated from a diversity of data types and produced by different research groups can be combined into a single estimate of evolutionary history. Despite their utility for synthesizing disparate data, many supertree approaches have several substantial drawbacks, however, including lack of a statistical basis and/or failure to incorporate uncertainties in estimated subtrees (Rannala & Yang 2008).

A *concordance tree* is a phylogeny summarizing the clades supported by a sample of genes from a genome. The concordance factor for a branch is the proportion of sampled genes supporting that branch. In BUCKy (Ané et al. 2007), concordance trees are estimated from a set of gene tree posterior distributions estimated using MrBayes. Concordance trees are useful for identifying regions of the species tree in which processes such as incomplete lineage sorting and hybridization have resulted in discordant gene trees and thus may indicate situations in which concatenated analyses may be inappropriate. One important point to note is that BUCKy currently performs best when the posterior distributions of gene trees are centered on a small number of topologies for all loci. Consequently, inclusion of large numbers of taxa or loci (especially when gene tree support is weak for a large number of loci) can render the analysis intractable. Thus, large data sets may need to be reduced to include only loci with sufficient information for gene tree estimation. Removing loci with fewer informative sites, however, tends to bias the distribution of gene trees to those with longer branches (and thus deeper coalescent times), potentially resulting in inflated concordance factors.

A *phylogenetic network* represents a species history that includes one or more hybridization or reticulation events and therefore cannot be represented by a strictly bifurcating tree. Early phylogenetic network models of hybridization ignored incongruence due to incomplete lineage sorting (Nakhleh 2010 and references therein), whereas newer models account for both processes (Yu et al. 2011 and references therein). Likewise, coalescent-based species tree models typically have not accounted for hybridization, though some models now incorporate both processes (e.g., STEM-hy: Kubatko 2009). Two challenges associated with coestimating hybridization and incomplete lineage sorting should be mentioned. First, the difficulty of the problem increases very rapidly with the number of taxa and the number of potential hybridization events. Second, there is the potential for overestimating the number of hybridization events because any pattern of discordance can be explained given a sufficient number of hybridization events (Yu et al. 2011).

A *locus tree* (Rasmussen & Kellis 2012) represents the species history wherein each branching point represents either a speciation event (in which one species splits into two) or a gene duplication (in which one gene region gives rise to a second). A locus tree is analogous to a species tree but accounts for gene duplications and losses. In fact, the locus tree and species tree are identical in the absence of gene duplications/losses. When a gene tree is constructed from paralogous sequences, coalescent events can be mapped to a locus tree. Gene duplication/losses and speciation events can then be mapped to the species tree.

Recommendations for Data Analysis Workflow

The suite of analysis tools available to phylogeneticists has grown substantially in recent years, although further development is needed in order to take full advantage of the large quantities of data becoming available. **Figure 4** illustrates general workflows for phylogenetic data analysis, including the workflow we recommend, which involves phasing of alleles, a priori and a posteriori orthology assessment, alignment trimming, model estimation, model adequacy tests, and estimation of species trees using models that account for sources of gene tree discordance. Although acknowledging that alternative data analysis pathways may ultimately prove useful, we await the results of future simulation studies to test the utility of other data types (e.g., biallelic data) and the accuracy and precision of other analysis methods (e.g., summary statistics for species tree estimation). In sum, we encourage researchers to systematically assess model fit before proceeding with phylogeny estimation and, notwithstanding the temptation to include all possible genomic data in phylogenetic analyses, to carefully consider subsampling data sets to ensure selection of loci that can be properly modeled.

SUMMARY POINTS

1. A variety of genomic partitioning strategies for generating data in the laboratory have been developed to overcome the high cost of WGS and assembly. A sufficient amount of informative data (hundreds of unlinked loci) can now be efficiently collected for virtually any taxonomic group to resolve phylogenetic questions from shallow to deep timescales.
2. Hybrid enrichment approaches currently offer the greatest promise for high-throughput and cost-effective data collection across a broad phylogenetic spectrum.
3. Phylogenetic error cannot always be overcome by increasing the quantity of data. Accurate, well-supported phylogenies often require careful selection of subsets of the total available data that can be adequately modeled.
4. Data subsampling should involve consideration of orthology, alignment quality, missing data, phylogenetic information, and several measures of model adequacy/fit.
5. Concatenation has been shown in numerous studies to introduce phylogenetic error when gene tree discordance exists. Concatenation of unlinked loci should be avoided unless the data are also analyzed using methods that explicitly account for gene tree discordance.
6. Many species tree methods (especially those explicitly modeling coalescent processes) are quickly being saturated with the quantity of available data, emphasizing the need for careful subsampling of data and improved methods of analysis.

FUTURE ISSUES

1. To date, hybrid enrichment tools have been developed for a limited taxonomic spectrum, primarily vertebrates and plants. Extension of this approach to other organisms would greatly accelerate resolution of understudied portions of the Tree of Life.
2. One remaining bottleneck for data collection is the per-sample cost of library preparation. Development of low-cost library preparation methods would greatly facilitate large-scale phylogenetic, phylogeography, and population genetic studies.
3. Methods of obtaining separate alleles (as opposed to consensus sequences) from high-throughput sequence data should be further developed and adopted by the phylogenetic community.
4. Given the rapidly increasing quantity of available data, systematists would benefit from an integrated system for data subsampling that would allow selection of a sufficient number of loci using objective criteria for minimizing phylogenetic error (indicated by a *star* in **Figure 4**).
5. One increasingly important bottleneck in phylogenomic analysis is the restricted number of loci that can be utilized by some methods of species tree estimation. Further development of these methods to accommodate data sets containing hundreds or thousands of loci would benefit the phylogenetic community.
6. Future development of phylogenetic models should continue to integrate diverse processes such as incomplete lineage sorting, hybridization, and gene duplication/loss.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We are very grateful to Lisa Barrow, Jeremy Brown, Felipe Grazziotin, and David Weisrock for providing valuable comments on this manuscript. This review was supported by NSF DEB1120516 to E.M.L., NSF IIP1313554 to A.R.L. and E.M.L., and NSF DEB1145978 to D. Rokyta, A.R.L., and E.M.L.

LITERATURE CITED

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* 4:903–5

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, et al. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407:513–16

Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.* 24:412–26

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.* 22:3179–90

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3:e3376

Bansal V, Bafna V. 2008. HapCut: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* 24:i153–59

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. 2004. Ultraconserved elements in the human genome. *Science* 304:1321–24

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

Bininda-Emonds ORP, Gittleman JL, Steel MA. 2002. The (super)tree of life: procedures, problems, and prospects. *Annu. Rev. Ecol. Syst.* 33:265–89

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.* 19:1171–80

Boussau B, Szöllosi GJ, Duret L, Gouy M, Tannier E, Daubin V. 2013. Genome-scale coestimation of species and gene trees. *Genome Res.* 23:323–30

Briggs AW, Good JM, Green RE, Krause J, Maricic T, et al. 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* 325:318–21

Brown JM, ElDabaje R. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. *Bioinformatics* 25:537–38

Brown JM, Lemmon AR. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. *Syst. Biol.* 56:643–55

Browning SR, Browning BL. 2011. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* 12:703–14

Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* 29:1917–32

Burbano HA, Hodges E, Green RE, Briggs AW, Krause J, et al. 2010. Targeted investigation of the Neandertal genome by array-based sequence capture. *Science* 328:723–25

Bybee SM, Bracken-Grissom H, Haynes BD, Hermansen RA, Byers RL, et al. 2011. Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biol. Evol.* 3:1312–23

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17:540–52

Chamberlain JS, Gibbs RA, Rainer JE, Nguyen PN, Thomas C. 1988. Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. *Nucleic Acids Res.* 16:11141–56

Craig DW, Pearson JV, Szelinger S, Seker A, Redman M, et al. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* 5:887–93

Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, et al. 2012. Targeted enrichment strategies for next-generation plant biology. *Am. J. Bot.* 99:291–311

Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharb K, Blaxter ML. 2012. Special features of RAD Sequencing data: implications for genotyping. *Mol. Ecol.* 22:3151–64

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12:499–510

Reviews laboratory methods of genomic-scale data collection for population genetics and phylogeography.

Degnan JH, Rosenberg NA. 2006. Discordance of the species trees with their most likely gene trees. *PLoS Genet.* 2:e68

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* 3:1700–8

Delport W, Poon AFY, Frost SDW, Pond SLK. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 26:2455–57

Edwards MC, Gibbs RA. 1994. Multiplex PCR: advantages, development, and applications. *Genome Res.* 3:S65–75

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63:1–19

Edwards SV, Liu L, Pearl DK. 2007. High-resolution species trees without concatenation. *Proc. Natl. Acad. Sci. USA* 104:5936–41

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, et al. 2010. Resolving postglacial phylogeography using high-throughput sequencing. *Proc. Natl. Acad. Sci. USA* 107:16196–200

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* 61:717–26

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27:401–10

Felsenstein J. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer

Gabaldón T. 2008. Large-scale assignment of orthology: back to phylogenetics? *Genome Biol.* 9:235

Gatesy J, DeSalle R, Wheeler W. 1993. Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Mol. Phylogenet. Evol.* 2:152–57

George RD, McVicker G, Diederich R, Ng SB, MacKenzie AP, et al. 2011. *Trans* genomic capture and sequencing of primate exomes reveals new targets of positive selection. *Genome Res.* 21:1686–94

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, et al. 2009. Solution hybrid selection with ultralong oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* 27:182–89

Godden GT, Jordan-Thaden IE, Chamala S, Crowl AA, García N, et al. 2012. Making next-generation sequencing work for you: approaches and practical considerations for marker development and phylogenetics. *Plant Ecol. Divers.* 5:427–50

Good JM. 2011. Reduced representation methods for subgenomic enrichment and next-generation sequencing. In *Molecular Methods for Evolutionary Genetics*, *Methods in Molecular Biology*, Vol. 772, ed. V Orgogozo, MV Rockman. Totowa, NJ: Humana

Goremykin VV, Nikiforova SV, Bininda-Emonds OR. 2010. Automated removal of noisy data in phylogenomic analyses. *J. Mol. Evol.* 71:319–31

Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, et al. 2013. Next-generation museomics disentangles one of the largest primate radiations. *Syst. Biol.* 62:539–54

Hancock-Hanser BL, Frey A, Leslie MS, Dutton PH, Archer FI, Morin PA. 2013. Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Mol. Ecol. Res.* 13:254–68

Hardenbol P, Banér J, Jain M, Nilsson M, Namsaraev EA, et al. 2003. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* 21:673–78

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* 27:570–80

Herman DS, Hovingh GK, Iartchouk O, Rehm HL, Kucherlapati R, et al. 2009. Filter-based hybridization capture of subgenomes enable resequencing and copy-number detection. *Nat. Methods* 6:507–13

Hohenlohe PA, Day MD, Amish SJ, Miller MR, Kamps-Hughes N, et al. 2013. Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Mol. Ecol.* 22:3002–13

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, et al. 2007. Human genome ultraconserved elements are ultraselected. *Science* 317:915

Kemena C, Notredame C. 2009. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* 25:2455–65

Knowles LL, Kubatko LS. 2010. *Estimating Species Trees: Practical and Theoretical Aspects*. Hoboken, NJ: Wiley-Blackwell

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. GARD: a genetic algorithm for recombination detection. *Bioinformatics* 22:3096–98

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. 2011. Computational methods for gene orthology inference. *Brief. Bioinforma.* 12:379–91

Kubatko LS. 2009. Identifying hybridization events in the presence of coalescence via model selection. *Syst. Biol.* 58:478–88

Kubatko LS, Carstens BC, Knowles L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25:971–73

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56:17–24

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. *Mol. Biol. Evol.* 29:457–72

Describes the importance of identifying systematic and stochastic sampling error in phylogenomics.

Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* 29:1695–701

Introduces a computationally efficient method of reducing systematic error though optimal data partitioning.

Lanier HC, Knowles LL. 2012. Is recombination a problem for species-tree analyses? *Syst. Biol.* 61:691–701

Leaché AD, Rannala B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. *Syst. Biol.* 60:126–37

Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Syst. Biol.* 58:130–45

Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.* 61:727–44

Demonstrates the utility of hybrid enrichment for high-throughput phylogenetics of nonmodel organisms.

Lemmon AR, Lemmon EM. 2012. High throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. *Syst. Biol.* 61:745–61

Lemmon AR, Moriarty EC. 2004. The importance of proper model assumption in Bayesian phylogenetics. *Syst. Biol.* 53:265–77

Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJP. 2013. Capturing protein-coding genes across highly divergent species. *BioTechniques* 54:321–26

Li H, Homer N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinforma.* 2:473–83

Liu K, Warnow TJ, Holder MT, Nelesen SM, Yu J, et al. 2012. SATé-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst. Biol.* 61:90–106

Liu L, Yu L. 2011. Estimating species trees from unrooted gene trees. *Syst. Biol.* 60:661–67

Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.* 10:302

Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV. 2009. Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.* 53:320–28

Maddison WP. 1997. Gene trees in species trees. *Syst. Biol.* 46:523–36

Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55:21–30

Makarova KS, Sorokin AV, Novichkov PS, Koonin EV. 2007. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics and archaea. *Biol. Direct* 2:33

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, et al. 2010. Target-enrichment strategies for next-generation sequencing. *Nat. Methods* 7:111–18

Reviews in detail laboratory methods of genomic partitioning (target enrichment).

Maricic T, Whitten M, Pääbo S. 2010. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One* 5:e14004

Markoulatos P, Siafakas N, Moncany M. 2002. Multiplex polymerase chain reaction: a practical approach. *J. Clin. Lab. Anal.* 16:47–51

Martin JA, Wang Z. 2011. Next-generation transcriptome assembly. *Nat. Rev. Genet.* 12:671–82

Mason VC, Li G, Helgen KM, Murphy WJ. 2011. Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. *Genome Res.* 21:1695–704

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol. Phylogenet. Evol.* 66:526–38

McCormack JE, Maley JM, Hird SM, Derryberry EP, Graves GR, Brumfield RT. 2012. Next-generation sequencing reveals phylogenetic structure and a species tree for recent bird divergences. *Mol. Phylogenet. Evol.* 62:397–406

McGuire G, Wright F, Prentice MJ. 1997. A graphical method for detecting recombination in phylogenetic data sets. *Mol. Biol. Evol.* 14:1125–31

Miller MR, Atwood TS, Eames BF, Eberhart JK, Yan Y-L, et al. 2007. RAD marker microarrays enable rapid mapping of zebrafish mutations. *Genome Biol.* 8:R105

Misof B, Misof K. 2009. A Monte Carlo approach successfully identifies randomness sequence alignments: a more objective means of data exclusion. *Syst. Biol.* 58:21–34

Nakhleh L. 2010. Evolutionary phylogenetic networks: models and issues. In *The Problem Solving Handbook for Computational Biology and Bioinformatics*, ed. L Heath, N Ramakrishnan, pp. 125–58. New York: Springer

Notredame C. 2007. Recent evolutions of multiple sequence alignment algorithms. *PLoS Comp. Biol.* 3:e123

Oakley TH, Wolfe JM, Lindgren AR, Zaharoff AK. 2012. Phylotranscriptomics to bring the understudied into the fold: monophyletic ostracoda, fossil placement, and pancrustacean phylogeny. *Mol. Biol. Evol.* 30:215–33

O'Neill EM, Schwartz R, Bullock CT, Williams JS, Shaffer HB, et al. 2013. Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Mol. Ecol.* 22:111–29

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* 96:2896–901

Ozsolak F, Milos PM. 2011. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* 12:87–98

Pagel M, Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.* 53:571–81

Pagel M, Meade A. 2008. Modeling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philos. Trans. R. Soc. B* 363:3955–64

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444:499–502

Peterson DG, Wessler SR, Paterson AH. 2002. Efficient capture of unique sequences from eukaryotic genomes. *Trends Genet.* 18:547–50

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE* 7(5):e37135

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* 9:e1000602

Reviews strategies for study design and data analysis in phylogenomics.

Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. *Annu. Rev. Ecol. Evol. Syst.* 36:541–62

Pisani D, Cotton JA, McInerney JO. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol. Biol. Evol.* 24:1752–60

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–18

Reviews inferring phylogenies from genomic-scale data and provides justification for using proper species-tree methodologies.

Rannala B, Yang Z. 2008. Phylogenetic inference using whole genomes. *Annu. Rev. Genomics Hum. Genet.* 9:217–31

Rasmussen MD, Kellis M. 2012. Unified modeling of gene duplication, loss and coalescence using a locus tree. *Genome Res.* 22:755–65

Reid NM, Hird SM, Brown JM, Pelletier TA, McVay JD, et al. 2013. Poor fit to the multispecies coalescent is widely detectable in empirical data. *Syst. Biol.* In press. doi: 10.1093/sysbio/syt057

Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM. 2013. Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Mol. Ecol.* 22:2953–70

Studies several ways that stochastic error can render strongly supported but false phylogenies.

Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst. Biol.* 56:389–99

Ronaghi M, Uhlén M, Nyrén P. 1998. A sequencing method based on real-time pyrophosphate detection. *Science* 281:363–65

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–74

Rubin BER, Ree RH, Moreau CS. 2012. Inferring phylogenetics from RAD sequence data. *PLoS One* 7:e33394

Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21:456–63

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68:978–89

Stiller M, Knapp M, Stenzel U, Hofreiter M, Meyer M. 2009. Direct multiplex sequencing (DMPS)—a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. *Genome Res.* 19:1843–48

Stull GW, Moore MJ, Mandala VS, Douglas NA, Kates H-R, et al. 2013. A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Appl. Plant Sci.* 1:1200497

Sullivan J, Joyce P. 2005. Model selection in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 36:445–66

Susko E, Spencer M, Roger AJ. 2005. Biases in pylogenetic estimation can be caused by random sequence segments. *J. Mol. Evol.* 61:351–59

Swofford DL. 2000. *PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods*. Sunderland, MA: Sinauer

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In *Molecular Systematics*, ed. DM Hillis, C Moritz, BK Mable, pp. 407–514. Sunderland, MA: Sinauer

Taly V, Pekin D, Abed AE, Laurent-Puig P. 2012. Detecting biomarkers with microdroplet technology. *Trends Mol. Med.* 18:405–16

Tewhey R, Nakano M, Wang X, Pabón-Peña C, Novak B, et al. 2009. Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol.* 10:R116

Timme RE, Bachvaroff TR, Delwiche CF. 2012. Broad phylogenetic sampling and the sister lineage of land plants. *PLoS One* 7:e29696

Townsend JP. 2007. Profiling phylogenetic informativeness. *Syst. Biol.* 56:222–31

Turner EH, Ng SB, Nickerson DA, Shendure J. 2009. Methods for genomic partitioning. *Annu. Rev. Genomics Hum. Genet.* 10:264–84

Van Orsouw NJ, Hogers RCJ, Janssen A, Yalcin F, Snoeijers S, et al. 2007. Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One* 2:e1172

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, et al. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol. Ecol.* 22:787–98

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

Weisrock DW, Smith SD, Chan LM, Biebouw K, Kappeler PM, Yoder AD. 2012. Mouse lemur phylogeny: an empirical demonstration of the effect of allele sampling in phylogenetics. *Mol. Biol. Evol.* 29:1615–30

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* 3:e7

Yang Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* 42:587–96

Yu Y, Than C, Degnan JH, Nakhleh L. 2011. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Syst. Biol.* 60:138–49

Zheng XH, Lu F, Wang ZY, Zhong F, Hoover J. 2005. Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. *Bioinformatics* 21:703–10

Describes and compares in detail laboratory methods of genomic partitioning.

Reveals the problems resulting from concatenating alleles across unlinked loci.

Annual Review of Ecology, Evolution, and Systematics

Volume 44, 2013

Contents

Genomics in Ecology, Evolution, and Systematics Theme

Topical Reviews

Indexes

Errata

An online log of corrections to *Annual Review of Ecology, Evolution, and Systematics* articles may be found at http://ecolsys.annualreviews.org/errata.shtml