Molecular Phylogenetics and Evolution 66 (2013) 526–538

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Review

Applications of next-generation sequencing to phylogeography and phylogenetics

John E. McCormack [a,*], Sarah M. Hird [a,b], Amanda J. Zellmer [b], Bryan C. Carstens [b], Robb T. Brumfield [a,b]

[a] Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, United States
[b] Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, United States

ARTICLE INFO

Article history:
Available online 14 December 2011

Keywords:
Population genomics
Coalescence
Reduced representation library
Target enrichment
High-throughput sequencing

ABSTRACT

This is a time of unprecedented transition in DNA sequencing technologies. Next-generation sequencing (NGS) clearly holds promise for fast and cost-effective generation of multilocus sequence data for phylogeography and phylogenetics. However, the focus on non-model organisms, in addition to uncertainty about which sample preparation methods and analyses are appropriate for different research questions and evolutionary timescales, have contributed to a lag in the application of NGS to these fields. Here, we outline some of the major obstacles specific to the application of NGS to phylogeography and phylogenetics, including the focus on non-model organisms, the necessity of obtaining orthologous loci in a cost-effective manner, and the predominate use of gene trees in these fields. We describe the most promising methods of sample preparation that address these challenges. Methods that reduce the genome by restriction digest and manual size selection are most appropriate for studies at the intraspecific level, whereas methods that target specific genomic regions (i.e., target enrichment or sequence capture) have wider applicability from the population level to deep-level phylogenomics. Additionally, we give an overview of how to analyze NGS data to arrive at data sets applicable to the standard toolkit of phylogeography and phylogenetics, including initial data processing to alignment and genotype calling (both SNPs and loci involving many SNPs). Even though whole-genome sequencing is likely to become affordable rather soon, because phylogeography and phylogenetics rely on analysis of hundreds of individuals in many cases, methods that reduce the genome to a subset of loci should remain more cost-effective for some time to come.

Contents

* Corresponding author. Address: Moore Laboratory of Zoology, Occidental College, 1600 Campus Rd., Los Angeles, CA 90041, United States.
E-mail address: mccormack@oxy.edu (J.E. McCormack).

1055-7903/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2011.12.007

1. Introduction

1.1. Multilocus studies and the promise of next-generation sequencing

Using multiple loci to infer population and species histories has become the baseline in phylogeography and phylogenetics. Although these two fields came to the multilocus approach for different historical reasons (see Brito and Edwards, 2008), multilocus studies in both fields benefited from decreasing costs of DNA sequencing in the last three decades. More recently, statistical phylogeography (Knowles, 2009) and the emerging species-tree paradigm of phylogenetics (Edwards, 2009) provided convincing theoretical arguments for incorporating information from multiple loci into estimates of population and species history to account for random variation in patterns of gene inheritance (i.e., coalescent stochasticity). The practical outcome of these developments is that most practitioners of phylogeography and phylogenetics have spent a significant portion of their time in the last decade developing and screening molecular markers suitable to their study system and appropriate to their evolutionary timescale of interest.

Even with the increasing availability of molecular markers for non-model organisms (Edwards, 2008; Thomson et al., 2010), the process of data generation for a multilocus study is laborious. In the fast-moving field of molecular biology, it seems ever more unjustifiable to embark on the lengthy process of screening loci for variability (assuming primers already exist), amplifying and sequencing DNA for each sample at each locus, and phasing nuclear data via computation or cloning. Researchers of phylogeography and phylogenetics have understandably looked toward next-generation sequencing (NGS) with great interest as a potential means to condense the many steps of multilocus data generation for non-model organisms into a more time-efficient and cost-effective process (Ekblom and Galindo, 2010; Lerner and Fleischer, 2010).

Despite this obvious potential, NGS has been slow to take root in phylogeography and phylogenetics compared to other fields like metagenomics and disease genetics (Mardis, 2008). We suggest that this lag has been caused by four specific aspects of phylogeographic and phylogenetic research: the predominant focus on non-model organisms, the need for sequencing large numbers of samples per species, the lack of consensus regarding library preparation protocols for particular research questions, and the transitional state of the technology (whole-genome data are still neither cost-effective, nor even desirable for phylogeography and phylogenetics, but are paradoxically easier to collect). Consequently there are as yet relatively few published papers where NGS has been used to generate the kind of phylogeographic or phylogenetic data sets that appear in *Molecular Phylogenetics and Evolution*.

The purpose of this review is to address the specific challenges of applying NGS to phylogeography and phylogenetics of non-model organisms and to highlight emerging methods of sample preparation and data analysis that *MPE*'s readers may find useful. We will not focus on uses of NGS for ecological genomics or adaptive divergence because these topics have been reviewed in detail elsewhere (Stapley et al., 2010; Rice et al., 2011). Rather, we focus on uses of NGS to reconstruct evolutionary and demographic history from the level of populations to species. We briefly touch on the steady convergence of phylogeography with ecological genomics in the section on future directions.

1.2. Specific challenges to applying NGS to phylogeography and phylogenetics

There are two substantial challenges to empirical researchers wishing to collect sequence data using NGS: sample preparation and bioinformatics. Here and in Section 2, we address the former, while we devote Section 3 to the latter.

1.2.1. The need for homologous DNA regions from many individuals

Phylogeography and phylogenetics require homologous genomic regions from multiple individuals to infer gene genealogies and phylogenetic trees. Using traditional Sanger DNA sequencing methods, the process of generating overlapping genomic regions is highly targeted and straightforward, involving primer design followed by PCR for each individual at each locus. With NGS, DNA is not necessarily enriched for a single locus via PCR-based amplification, although this is one possible application, but for many loci through a variety of methods involving reduction of the size of the genome (Table 1; Fig. 1). Unlike with Sanger sequencing, with many NGS methods, there is less control over exactly what regions of the genome are sequenced. The trade-off is that the number of targets (and the number of reads associated with each target) is increased by orders of magnitude. Genomic DNA (gDNA) must be prepared in advance to contain orthologous regions among individuals, preferably without requiring hundreds to thousands of PCRs. The need to reduce the genome to overlapping subsets might become unnecessary as it becomes cost-effective to sequence whole genomes for hundreds of individuals and once the theory and analysis of full-genome data is better developed (see Section 4); however, for the time being the issue of how to reduce the genomes of many individuals to orthologous fragments remains a significant obstacle to incorporating NGS methods into phylogeography and phylogenetics. In Section 2 and Table 1, we discuss different wet lab methods for genomic reduction, including those that select expressed genes, random fragments from throughout the genome, and other targeted regions.

1.2.2. Cost-effective multiplexing and library preparation

The second challenge is that using NGS for phylogeography and phylogenetics is only cost-effective if many individuals can be combined (multiplexed) in the same sequencing run and the costs subdivided among many samples (Glenn, 2011). Multiplexing involves the application of short identifying DNA sequences (called "indexes", "barcodes", or the term we will use here – "tags") that are incorporated into the DNA fragments either by PCR (Binladen

Table 1
Methods of sample preparation for using NGS in phylogeography and phylogenetics.

Method	Other names or variants	Literature method	Literature examples	Benefits	Drawbacks	Best application
Amplicon sequencing	Multiplex PCR, parallel tagged sequencing	Binladen et al. (2007), Meyer et al. (2008), Tewhey et al. (2009b)	Chan et al. (2010), Griffin et al. (2011), Gunnarsdóttir et al. (2011); Morin et al. (2010), Parks et al. (2009)	Highly targeted. Results in nearly complete data matrices. Needed coverage easy to calculate. Circumvents individual sequencing reactions and phasing nuclear loci compared to Sanger sequencing	Requires PCR of each individual at each locus	Small- to medium-scale projects targeting a limited number of genes
Restriction-digest	Double-digest genome reduction, RAD sequencing (RAD-seq), complexity reduction of multilocus sequences (CRoPS), Genotyping by Sequencing (GBS)	Baird et al. (2008), Davey et al. (2011)	Andolfatto et al. (2011), Amaral et al. (2009), Bers et al. (2010), Emerson et al. (2010), Gompert et al. (2010), Hohenlohe et al. (2011), Hyten et al. (2010a,b), Kerstens et al. (2009), Ramos et al. (2009); Sánchez et al. (2009); Van Orsouw et al. (2007); Van Tassell et al. (2008), Wiedmann et al. (2008); Williams et al. (2010)	Broad, random genomic sampling of thousands of independent genomic regions. Requires no prior genomic resources whatsoever	Reproducibility and throughput may be limited by gel extraction step. Not targeted, thus coverage can be difficult to estimate. Null alleles could skew diversity estimates	Intraspecific studies of recent divergence
Target enrichment	Sequence capture, targeted resequencing, primer extension capture (PEC)	Albert et al. (2007), Gnirke et al. (2009), Hodges et al. (2007), Okou et al. (2007), Tewhey et al. (2009a), Maricic et al. (2010)	Briggs et al. (2009; Faircloth et al. in press)	Rapid collection of thousands of loci without individual PCR	Requires some prior genomic knowledge, but not necessarily a sequenced genome	Phylogenomics at taxonomic levels above at and above the species level
Transcriptome	RNA-seq	Morin et al. (2008); Marioni et al. (2008)	Barbazuk and Schnable (2011), Cánovas et al. (2010), Chepelev et al. (2009); Geraldes et al. (2011), Hittinger et al. (2010)	Can leverage data from expression studies	Skewed read distributions can outstrip coverage, making it difficult to find orthologous loci	Leveraging existing cDNA libraries

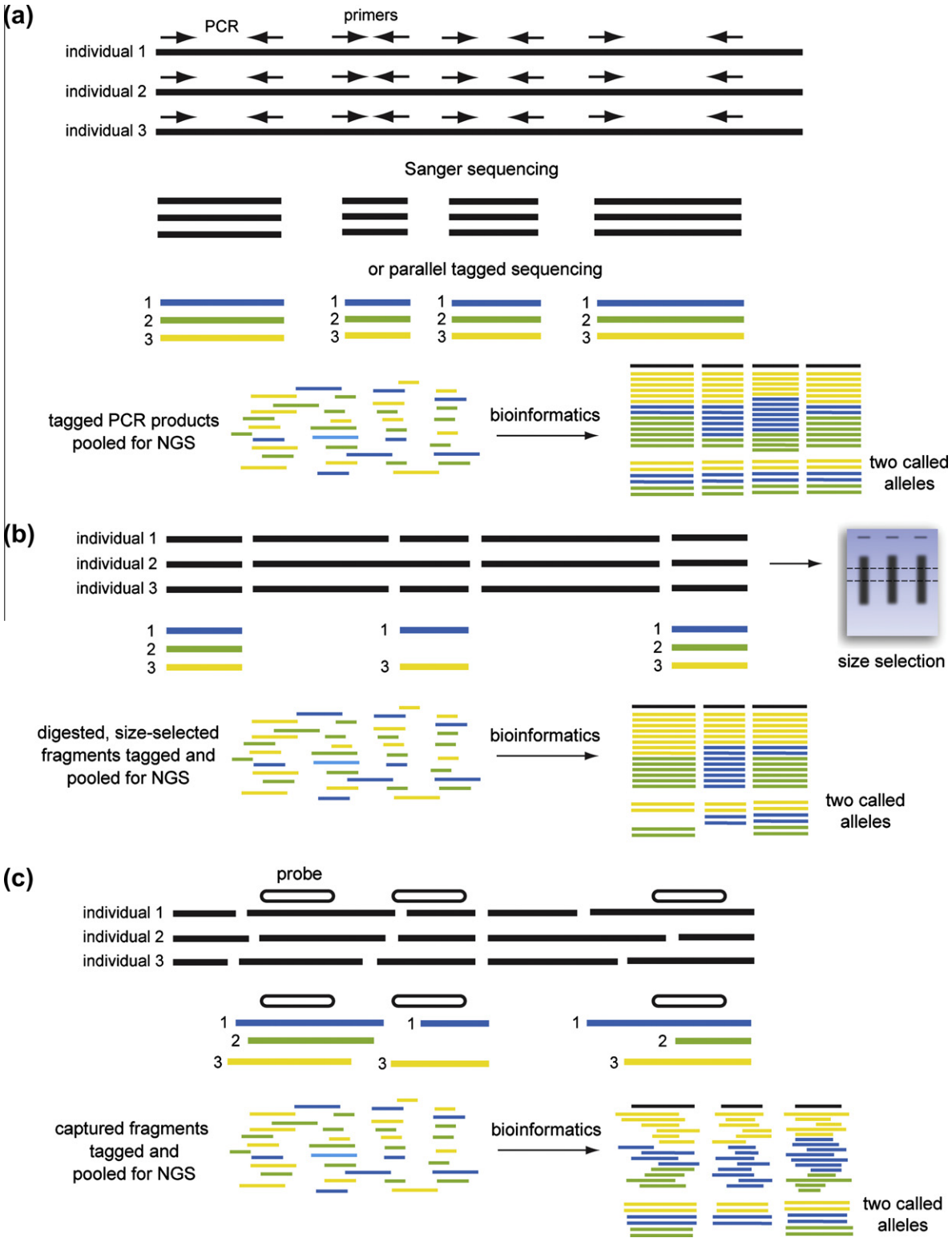

Fig. 1. Basic methods of sample preparation for NGS. (a) amplicon sequencing in which PCR products are tagged and pooled for NGS, resulting in a largely complete data matrix; (b) restriction-digest based methods, where genome reduction occurs by manual size selection. Note that mutations in the restriction site (as in individual 2) will result in a null allele; (3) target enrichment, in which probes "catch" gDNA, which is then pooled for NGS. The hybridization of probe to gDNA is robust to some variation, but mismatch or missing loci (show in individual 2) can result in missing data.

et al., 2007) or ligation (Meyer et al., 2008). These tags identify an individual prior to pooling with other tagged samples (Craig et al., 2008). Sequences are later sorted with bioinformatics. The most desirable tag sequences are those that are as short as possible while still being multiple base pairs away from other tags in the pool so that reads with sequencing errors in the tag sequence can still be allocated to the proper sample (Hamady et al., 2008). Simple methods for hierarchical tagging (Neiman et al., 2011) expand the possible number of pooled samples from hundreds to thousands.

NGS also requires platform-specific, often proprietary adaptor sequences to be incorporated into DNA fragments (this step is called library preparation and often occurs in conjunction with tagging). These adaptor sequences provide priming sites or hybridization targets for sequencing, depending on the NGS platform (e.g., linkers A and B for Roche 454 pyrosequencing or adaptors P1 and P2 for Illumina sequencing-by-synthesis). Incorporating both tags and adaptors to template DNA can be expensive if proprietary kits are used (e.g., Nextera™ kits list price at $2200 for 20 samples). A recent review calculated that library preparation was 10 times more expensive per sample than the sequencing itself (Glenn, 2011). In Section 2, we mention options for more cost-effective library preparation if they are available.

1.2.3. The long reign of the gene tree in phylogeography and phylogenetics

Another issue is the historical importance of utilizing gene trees in phylogeography and phylogenetics (Brito and Edwards, 2008), and the preeminence of coalescent-based analytical methods that either require, or are currently best utilized, in concert with gene trees (Kuhner, 2009; Liu et al., 2009a; Pinho and Hey, 2010). At present, gene trees are most robustly inferred from loci with high information content, for example, a non-recombining locus containing a series of linked SNPs. Individual SNPs, on the other hand, have low information content on a per-locus basis and have been used predominately with classification methods such as Structure (Pritchard et al., 2000) and principal components analysis (e.g., Novembre et al., 2008), although more versatile analyses are emerging (see below). While distance-based genealogies and phylogenies can be built from unlinked SNPs (e.g., Emerson et al., 2010), this ignores models of molecular substitution and probabilistic tree-searching algorithms that have led to more robust phylogenetic inference in the last several decades.

The practical problem is that most existing NGS technologies (e.g., Illumina) produce short reads and therefore are best suited for generating SNPs, not whole loci featuring many linked SNPs. This has limited the application of NGS data with respect to the standard analytical tool kit of phylogenetics and phylogeography. Certainly, this problem will resolve itself as NGS platforms converge on longer reads, and with the advent of third generation sequencing platforms (e.g., PacBio, Ion Torrent, Starlight). It is also possible that new analytical techniques will reduce our dependence on genes trees (Bryant et al., 2012; Naduvilezhath et al., 2011; Sirén et al., 2011). Until then, methods that can generate data amenable to gene-tree analysis will be preferred in phylogeography and phylogenetics. We point out which methods work well with gene trees versus those that generate unlinked SNPs.

1.3. Primary data collection or marker development?

A major consideration from the perspective of time investment is whether NGS data are amenable for use as primary data (e.g., Emerson et al., 2010) rather than being used to develop markers for later sequencing/genotyping (e.g., Williams et al., 2010). Downstream genotyping could still utilize high-throughput technologies, such as a SNP chip (Wang et al., 1998; Lipshutz et al., 1999; Buetow et al., 2001; Decker et al., 2009), or it could follow the more conventional route of primer design followed by individual PCR. The problem is that, unlike with Sanger sequencing, which is highly targeted to individual amplicons, NGS occurs on a pooled sample of amplicons. Although the researchers can make some decisions that influence which fragments of the initial pool are sequenced with sufficient coverage (e.g., the number of samples/individuals), pooled NGS sequencing is also subject to many stochastic factors, such as PCR bias and unequal sample pooling, which can result in patchy data matrices with large amounts of missing data. Methods that allow more even distribution of sequencing, such as PCR-less library preparation (Kozarewa et al., 2009), will ameliorate this problem to some degree, but stochasticity in coverage will likely still play a larger role in NGS methods than in Sanger sequencing. In Section 3, we discuss some factors relating to the use of NGS results as primary data, such as coverage and analytical methods that permit missing data.

2. Review of wet lab methods for sample preparation

There are several existing reviews of different NGS sequencing platforms and chemistries (Shendure and Ji, 2008; Glenn, 2011), so we discuss platform-specific details sparingly. The major difference for the purpose of this review is between platforms that produce fewer longer reads (~400–800 bp, 454 Titanium) versus those that produce orders of magnitude more short reads (~70–200 bp, Illumina HiSeq).

2.1. Multiplex PCR and amplicon sequencing

Perhaps the most straightforward application of NGS to phylogeography and phylogenetics is amplicon sequencing, or NGS sequencing of PCR products that have already been generated by Sanger sequencing (Fig. 1a). When multiple loci for an individual are tagged and pooled with tagged loci of other individuals, this method is called parallel tagged sequencing (Meyer et al., 2008). Benefits over typical Sanger sequencing include faster sequencing time and one-step phasing of nuclear DNA (because NGS results are single-stranded). Parallel tagged sequencing, however, does not remove the process of PCR for each individual at each locus, which can be the most onerous part of a multilocus phylogeographic study. An alternative to multiple PCRs is multiplex PCR methods with multiple primer pairs, but this approach has several limitations (described in Mamanova et al., 2009) including biased representation of some products as well as chimeric DNA sequences. Some promising alternatives have emerged, such as microdroplet PCR (Tewhey et al., 2009b), where millions of PCR reactions occur in picoliter-sized droplets before being pooled together, and the 96.96 Dynamic Array™ by Fluidigm (Seeb et al., 2009; Smith et al., 2011), which allows 96 primer combinations to be used on 96 samples (9216 total PCR reactions) using a plate-based nanofluid technology. These methods circumvent some of these challenges of multiplex PCR by ensuring that each primer combination is amplified separately (albeit in parallel), but thus far they have been little applied to phylogeography, so their ease-of-use and cost-effectiveness, though promising, is difficult to gauge.

Parallel tagged sequencing may be the most cost-effective method for small to medium-sized projects with few loci that amplify well across individuals (e.g., Griffin et al., 2011). For this reason, it has been used effectively in phylogenetic studies involving whole mitochondrial sequencing (Chan et al., 2010; Morin et al., 2010; Gunnarsdóttir et al., 2011) and whole chloroplast sequencing (Parks et al., 2009). It is also useful for uncovering rare sequences and has therefore been effectively employed in environmental sampling and metagenomics where a single locus is amplified from a sample containing DNA from many organisms (Fierer et al., 2008) as well as for characterizing the major histocompatibility complex where there are many alleles, some extremely rare (Babik et al., 2009; Kloch et al., 2010). Here, it is more cost-effective to build multiplexing tags (Faircloth and Glenn, 2012) and NGS platform-specific adaptor sequences into the amplification primers (Binladen et al., 2007) in order to circumvent costly kit-based library preparation.

2.2. Restriction digest-based methods

Several methods of genome reduction involve digestion of template DNA with restriction enzymes and manual excision of some fragment size range from an agarose gel to produce a reduced representation library (RRL) (Fig. 1b). RRLs pre-date the existence of NGS (Altshuler et al., 2000; Nicod and Largiadèr, 2003), and library preparation protocols were then adapted to NGS platforms (Van Tassell et al., 2008). No genomic resources are required in advance, so restriction digest-based methods are in many ways ideal for non-model organisms without a close relative with a sequenced genome.

There are several potential drawbacks of restriction digest-based methods. Selecting a fragment size range by manual excision is subject to some human error, potentially leading to fewer orthologous fragments across individuals. There are automated machines for size selection (PippinPrep™ from Sage Science and the LapChip® XT from Caliper LifeSciences), but they are somewhat expensive (currently $15,000–$20,000 for equipment). However, methods that add barcodes and adaptors prior to the cutting step allow gel cutting of one pooled sample.

Another issue that has not been well addressed is null alleles, where mutations in the restriction site result in the loss of a fragment in some individuals. Null alleles are easy to detect when mutations in restriction sites are fixed among populations. However, study populations that contain individuals heterozygous for null alleles present a more insidious problem, as it becomes more difficult to distinguish homozygotes from individuals with null alleles. Failure to detect null alleles in highly heterozygous populations could bias diversity statistics and phylogeographic inference.

Another potential drawback is that restriction-digest based methods are not suitable for deep-level phylogenetic studies, as mutations in the restriction sites quickly reduce the number of homologous fragments with increasing phylogenetic distance (McCormack et al., 2012; Althoff et al., 2007). Finally, most restriction-digest based methods are geared toward SNP generation (using short reads on the Illumina platform), which may not be ideal for researchers focusing on gene trees. However, most protocols could be adapted to platforms that produce longer reads (e.g., 454). We are also optimistic that this problem will likely be alleviated as all NGS platforms converge on longer reads.

2.2.1. RAD sequencing

Restriction-site Associated DNA (RAD) sequencing is the NGS method that has made the most impact on phylogeography and phylogenetics to date. As with other similar methods described below, DNA is digested with restriction enzymes and the resulting fragmented are size-selected from an agarose gel and sequenced via NGS. What sets RAD sequencing apart from other methods is that it combines tight control over the fragments resulting from the digest with ultra-deep sequencing across many individuals (Baird et al., 2008). For this reason, it is likely one of the most reproducible of the many restriction digest-based methods. Resulting NGS reads are mined across individuals for SNPs that occur immediately adjacent to common digest sites. This method has proven effective at generating data for marker development (Miller et al., 2007), genome scans (Hohenlohe et al., 2010), and building distance-based phylogenies for recent demographic events (Emerson et al., 2010). So far as we can tell, RAD sequencing has been carried out exclusively on the Illumina platform, which has limited resulting data to SNPs. However, paired-end Illumina sequencing should permit assembly of contigs up to 500 bp (Etter et al., 2011).

2.2.2. Other restriction digest-based methods

There are several alternatives to RAD sequencing that take advantage of restriction-digest based genome reduction (many are reviewed in Davey et al., 2011). The basic idea of these methods is akin to sequencing Amplified Fragment Length Polymorphism (AFLP) fragments (Vos et al., 1995), except instead of variation in the restriction site, the variation of interest lies between common restriction sites. Variations on this method have been used to generate thousands of SNPs for various organisms of agricultural importance (Van Orsouw et al., 2007; Van Tassell et al., 2008; Wiedmann et al., 2008; Amaral et al., 2009; Kerstens et al., 2009; Ramos et al., 2009; Sánchez et al., 2009; Hyten et al., 2010a,b), most involving many fewer individuals that would normally be assayed in a typical phylogenetic or phylogeographic study. Recent uses have involved wild populations (Bers et al., 2010) with application to phylogeography (Gompert et al., 2010; Williams et al., 2010).

In one example particularly relevant to phylogenetics, Decker et al. (2009) used a digest method to generate >40,000 SNPs and resolved the recent phylogenetic history of extinct and extant ruminants. In this case, Van Tassell et al. (2008) discovered the SNPs using a restriction-digest based method and then Decker et al. (2009) later genotyped a subset of the SNPs *en masse* for 61 species using a SNP chip (BeadChip, Illumina). Though producing a well-resolved tree based on a massive amount of data, this study also underscores the analytical limitations of SNP data, as the phylogenetic analysis was restricted to a parsimony method that required categorical coding of the data matrix (e.g., homozygotes coded as "0" or "1" and heterozygotes coded as polymorphic). In fact, most restriction-digest studies have targeted SNPs with short reads from the Illumina platform. However, two recent studies have used this method in conjunction with 454 sequencing to generate longer loci for use in coalescent-based phylogeographic analysis, in addition to SNP-based assignment tests (McCormack et al., 2012; Zellmer et al., 2012). Paired-end sequencing should also allow for the generation of whole loci for building gene trees with the Illumina platform. With large numbers of individuals, one economical approach is to add NGS adaptors and individual tags in a PCR step by incorporating them into the primer sequence (McCormack et al., 2012; Williams et al., 2010). Streamlined methods for adding adaptors and barcodes, while avoiding the use of proprietary kits, is a much needed area of methodological advancement.

2.3. Target enrichment

Target enrichment (also called "sequence capture" or "targeted resequencing") involves the selective capture of genomic regions prior to NGS (Mamanova et al., 2009). In target enrichment (Fig. 1c), fragmented gDNA is mixed with DNA or RNA probes, which hybridize to gDNA fragments either on an array (Albert et al., 2007; Hodges et al., 2007; Okou et al., 2007) or in solution (Gnirke et al., 2009; Maricic et al., 2010). Non-targeted DNA is then washed away, and targeted DNA is eluted and sequenced simultaneously via NGS.

Compared to restriction-digest methods, in which random fragments are sequenced, target-enrichment is a non-random method of genome reduction. Consequently, it requires some prior genomic resources to design probes, although they can potentially be from distantly related species (see below). One benefit to focusing on specific targets is that fragment size selection is easier and can make use of random genomic shearing (e.g., mechanical, acoustic, or enzymatic) as opposed to the laborious manual size selection step of restriction-based methods. With target enrichment, individual samples can be tagged and multiplexed either post-enrichment or prior to enrichment (Kenny et al., 2011). The latter technique is especially promising from the perspective of cost per sample given that the probes themselves can be quite expensive.

While the target enrichment technique is well described (Mamanova et al., 2009), most applications have been directed at human disease genes and exomes (Albert et al., 2007; Hodges et al., 2007; Ng et al., 2009; Tewhey et al., 2009a). Applications to phylogeography and phylogenetics will largely turn on the description of appropriate probe sets that are conserved enough that they bind genomic DNA across individuals (for phylogeography) and species (for phylogenetics) and yet show enough variation to be informative at the time depth of interest. We discuss different options for probe design below. Another promising, related technique is primer extension capture (PEC), which uses relatively short primer sequences as probes for sequence capture. Briggs et al. (2009) used PEC to capture and sequence the entire mtDNA genomes of five Neanderthals, allowing for their phylogenetic analysis with modern humans.

A final benefit of target enrichment is that probes can be tiled such that short reads from many tiled sections can later be assembled into larger contigs, obviating the problem of SNPs versus gene trees. This design maximizes the advantages of depth of coverage on a platform like Illumina, while minimizing the drawback of short read length. However, it should be noted that phasing loci generated from tiled probes can be problematic because the individual reads cannot be assigned to their respective chromosomal copies of origin within individuals. Thus, tiling might pose more of a problem for population-level studies, where coalescence among closely-related gene copies can bear strongly on the analysis, than it will for deep-level phylogenetics, where the coalescence times among species dwarf those between individual allele copies.

Once contigs are assembled from tiled reads, the thousands of independent loci that can be draw from target enrichment are ideal for use in species-tree analysis, an emerging systematic paradigm (Edwards, 2009) that has been little applied to phylogenomics. Here, the major problem is likely computational, as many species-tree programs rely on simultaneous optimization of individual gene trees and the species tree (e.g., BEST, Edwards et al., 2007; *BEAST, Heled and Drummond, 2010). Methods that use summary statistics to arrive at a fast, analytical solutions to the species tree (Liu et al., 2009b) offer one potential workaround to this problem, and are currently the only solution for phylogenomic data sets featuring hundreds to thousands of loci. Alternatively, algorithmic methods such as STEM (Kubatko et al., 2009) take gene trees as input, thus allowing gene trees from individual loci to be estimated in parallel prior to species tree estimation. However, there is a pressing need for further methodological development so that analytical robustness need not be sacrificed for time efficiency.

2.3.1. Probe sets designed from ultraconserved elements for phylogenomics

Much like universal priming sites for mitochondrial DNA (Kocher et al., 1989), conserved probes allow for maximal applicability across taxonomically diverse organisms. A conserved probe works much like a pair of primers located in conserved coding regions, except that variability would be sought in the regions flanking the probe instead of in between primers. What are the options for probe sets that are conserved enough to work across species and higher taxonomic groups for phylogenomics?

One promising development toward universal target enrichment for deep-level phylogenomics is the discovery of ultraconserved elements (UCEs) in mammals (Bejerano et al., 2004). The exact definition of a UCE differs among studies. Defined loosely, UCEs are genomic regions that show remarkable (in some cases 100%) conservation over a "long" stretch of DNA (generally 50–200 bps) among widely divergent organisms. The structure and function of UCEs is an active area of research. UCEs also possess properties that make them highly desirable as anchors for genetic markers. First, they are found in high numbers throughout the genome. Second, they appear to have little overlap with known paralogous genes (Derti et al., 2006), whose occurrence is difficult to detect and can confound phylogenetic inference (Philippe et al., 2011). Third, because variability increases moving toward the flanks, UCEs and flanking DNA might harbor phylogenetic signal useful for phylogenetic reconstruction at multiple evolutionary timescales (Faircloth et al., 2012).

A recent study showed that probes designed from UCEs conserved across amniotes (e.g., mammals, reptiles, and birds) have sufficient information content to resolve the primate tree of life and enrich over 800 loci in 9 bird species to resolve the phylogeny of three basal bird lineages (Faircloth et al., 2012). Another study of >2000 UCEs shared among the chicken, zebra finch, and *Anolis* lizard genomes found that nearly 1000 UCEs could also be located in the 27 mammals with sequenced genomes, elucidating their evolutionary history with as many as 917 loci (McCormack et al., 2012). Another study used UCE probes to capture a complete data matrix of over 1000 loci for six reptiles (Crawford et al., 2012). Combined with data obtained from existing genomes of birds and mammals, the resulting phylogeny revealed the evolutionary affinities of turtles with archosaurs (bird and crocodilians) with perfect support. The description of UCEs in diverse animal groups from tetrapods (Stephen et al., 2008) and reptiles (Janes et al., 2011) to invertebrates and yeast (Siepel et al., 2005) suggest that conserved probe sets for target enrichment may be applicable across a broad swath of the tree of life.

2.3.2. Probe sets designed from closely related genomes and transcriptome libraries

Given the amount of genomic resources now available, and the number of genome sequencing efforts currently being undertaken, it is also becoming feasible to design probe sets more targeted to individual species or groups using the genome of a closely related species. Due to their general conservation, yet high information content found at degenerate third codon positions, exons are a particularly appealing target for sequence capture. As there is currently more transcriptome data available than whole-genome data, probe sets could also be designed from transcriptome libraries, either undertaken especially from the species of interest or mined from existing transcriptomes of closely-related species. The potential drawback is that transcriptome libraries are highly variable depending on tissue type and many other variables. One attempt to align transcriptomes of 10 taxonomically divergent bird species did not find a large number of overlapping loci (Kuenster et al., 2010).

2.4. Transcriptome sequencing

Sequencing the transcriptome itself (RNA-seq) can be viewed as another method of genomic reduction where the remaining subset of gDNA is not random, as with restriction-digest, but contains the set of expressed genes (Marioni et al., 2008; Morin et al., 2008; Wang et al., 2009). While transcriptome sequencing is more prevalent in studies of adaptation and ecological genomics, a recent study by Hittinger et al. (2010) showed that transcriptome sequences could be used to recover the known phylogeny of a group of mosquitoes, demonstrating their potential utility to phylogeography and phylogenetics. Nabholz et al. (2011) sequenced the brain transcriptomes of nine bird species and used the resulting data to build a phylogeny. Transcriptomes have also been mined for SNPs in a variety of species (Chepelev et al., 2009; Cánovas et al., 2010; Barbazuk and Schnable, 2011; Geraldes et al., 2011).

3. Data analysis and bioinformatics

3.1. The difference between Sanger and NGS data and the importance of coverage

The most obvious difference between Sanger sequence data and NGS data is what is initially most daunting about NGS data: quantity. Simply storing unprocessed NGS data requires significant computer memory and often hardware upgrades or remote (i.e., online or "cloud") storage. Whereas a reasonably large Sanger dataset may contain 500 sequences, typical 454 and Illumina runs generate 1,000,000–2,000,000,000 sequences, and these numbers are increasing rapidly as the sequencing platforms are refined. Data set sizes now are measured in terabytes and file transfer is often conducted through normal postal mail due to the uploading/downloading time and cost of sending files through the internet or cloud. However, logistical difficulties aside, these numbers are in some ways deceptively large, because another major difference between NGS data and Sanger sequence data is the quality of the reads.

Chromatograms are an intuitive way to assess the quality of a given Sanger sequence because the colored peaks are a reflection of the strength of each nucleotide's signal. Frequently, with Sanger data, a human being has evaluated all or most of the bases called by the sequencer. This is not possible with even a small NGS dataset. NGS quality scores are an integral part of the sequence data itself, and come either as a series of integers or letters corresponding to every base called by the NGS platform. These are very different paradigms: with Sanger data you get, in essence, a more or less true snapshot of a pool of amplicons at every position. With NGS data, you get a slightly imperfect representation of some of the amplicons from a sample, along with their associated quality scores. This is why coverage (i.e., the number of reads that support a specific base call) is critically important with NGS data. Coverage affords confidence that every DNA fragment in a pool will be sequenced – and enough times to determine heterozygous positions. Coverage also ensures that NGS sequencing error can be detected and distinguished from heterozygosity.

3.2. Determining orthologous vs. paralogous loci

Although paralogous genes have long presented problems for researchers using Sanger sequence data, the detection of paralogs in Sanger data is relatively straightforward when using targeted primer pairs. A signal of more than two alleles (or more than one allele when mtDNA is studied) in the chromatogram is a clear sign that more than one gene copy has been sequenced. In addition, methods for detecting paralogs, such as SSCP (Sunnucks et al., 2000), are available for confirmation.

In contrast, NGS occurs on a single strand, so paralogy cannot be detected until after data are aligned. Evidence for paralogy from NGS alignments includes any biological signal that too many alleles have been grouped into a single, putatively homologous locus (e.g., three alleles for a diploid). These signals include (but are not limited to) more than two bases at a particular position (especially with greater than 1X coverage) or elevated values for observed heterozygosity at a given position in an alignment. Additionally, when viewing alignments of one's data, it is often obvious that too much variation exists for a single locus (Fig. 2); for this reason, it is a good idea to become familiar with some raw data, even if there is too much to check them all manually. One complication is that low coverage and uneven read distribution generated by stochasticity in PCR and sequencing can mask this evidence. To avoid paralogous loci, some researchers have eliminated the alignments (or contigs) with the highest coverage (Emerson et al., 2010) because copy number variants are often correlated with high coverage loci in NGS data (Alkan et al., 2009). Designating a maximum coverage cutoff is less than ideal, however, because one may end up throwing out good data. A method that incorporates error rate and evaluates all reads in an alignment to assign a measure of confidence in the number of supported alleles would be highly desirable and awaits development.

3.3. From raw NGS output to formats appropriate for phylogeography and phylogenetics

3.3.1. Filtering unprocessed NGS data and quality control

The first step in processing raw NGS data is eliminating low quality reads (quality control or QC), where "low quality" is generally determined by the user within some broad guidelines. When these values are reported, one frequently sees a minimum default value of $Q > 20$ (e.g., Oliver et al., 2010; Medinger et al., 2011), where Q is the quality assigned to a base. This translates to a 0.01 probability of a base call being inaccurate, or 99% accuracy. Many programs will QC raw NGS output (see Table 2). QC can also

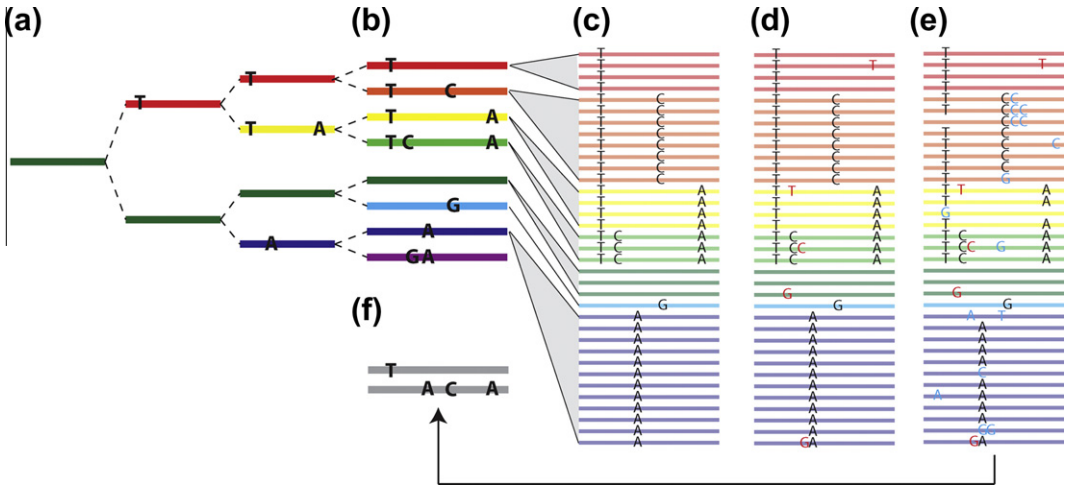

Fig. 2. The combined effect of PCR bias, PCR error, and sequencing error on calling alleles for paralogous loci. (a) Duplications (dotted lines) produce paralogs. (b) Paralogs accumulate different mutations to result in an underlying genetic signal for the eight paralogous loci. (c) Stochastic PCR bias during the amplification step of NGS library preparation results in some of the loci being amplified more than others, and one locus not being amplified at all. (d) PCR error during the amplification step results in another layer of stochastic noise. (e) Error during NGS adds an additional layer of noise. (f) The two alleles that would be called from the reads in (e) if they were processed in an allele-calling program like PRGmatic (Hird et al., 2011). Scanning the alignment by eye or by applying heterozygosity tests indicates that the alignment actually contains multiple paralogous loci.

Table 2
Programs for quality control, assembling, and analyzing NGS data for phylogeography and phylogenetics.

Program	QC	Alignment	Allele Calling	SNP Calling	Visualization	Open Source?	Computer Platforms	Other Functions	References
CLOTU	X	C				Y	Internet	Automated BLAST	Kumar et al. (2011)
Galaxy	X	R		X		Y	Internet		Goecks et al. (2010)
DNAstar SeqMan Ngen	X	R,D	X	X	X	N	Windows, MacOSX, Linux		http://www.dnastar.com
CLC Genomics Workbench	X	R,D		X	X	N	MacOSX, Linux, Windows		http://www.clcbio.com/
Geneious	X	R,D		X	X	N	WindowsVista, MacOSX		http://www.geneious.com/
GATK	X		X	X	X	Y	MacOSX, Linux		DePristo et al. (2011)
RDP Pyrosequencing Pipeline	X					Y	Internet	Microbial DNA Analyses	Cole et al. (2009)
Mothur	X					Y	Windows, MacOSX, Linux		http://www.mothur.org/
STACKS		C	X	X		Y	Unix	Genetic mapping	Catchen et al. (2011)
CAP3		C				Y	Windows,MacOSX, Linux, Solaris, Internet		Huang and Madan (1999)
PRGmatic		C,D	X	X		Y	MacOSX		Hird et al. (2011)
ABySS		D				Y	Any		Simpson et al. (2009)
SAMtools		R		X	X	Y	Unix		Li et al. (2009)
BWA		R				Y	Any (C++ source)		Li and Durbin (2009)
Bowtie		R				Y	Windows, MacOSX, Linux		Langmead et al. (2010)
Exonerate		R				Y	Unix		Slater and Birney (2005)
Novocraft		R				Y	MacOSX, Linux		Hercus, C. 2009. http://www.novocraft.com
Stampy		R				Y	MacOSX, Linux		Lunter and Goodson, 2011
SOAP		R,D		X		Y	Any (C++ source)		Li et al. (2008)
MIRA		R,D		X		Y	MacOSX, Linux	Automatic error removal	Chevreux et al. (1999)
Velvet		R*,D				Y	MacOSX, Linux, cygwin		Zerbino and Birney (2008)
Bambino				X	X	Y	Windows, MacOSX, Linux		Edmonson et al. (2011)
VarScan				X		Y	Any (JAVA source)		Koboldt et al. (2009)
Casava				X		N	Linux		Illumina propietary
Tablet					X	Y	Windows,MacOSX, Linux, Solaris		Milne et al. (2010)

QC = quality control.
R = reference.
D = *de novo*.
C = cluster generation.
* Velvet can use reference reads but it treats them as "just another" read, not a reference.

include discarding sequences shorter than some value, which is appropriate when there is a good idea for the target size, as with the restriction digest-based methods described above. One may also choose to discard sequence length outliers, as these are often associated with sequencing error (Oliver et al., 2010) or any reads containing an unidentified base ("N").

The second step of filtering is to sort the data by tag and remove primer and barcode sequence (if necessary) from the flanks of the reads. This is often performed by custom Perl or Python scripts, but there are some free online services such as the Ribosomal Database Project's Pyrosequencing Pipeline (Cole et al., 2009), the Galaxy website, or Mothur (Table 2). If one has selected a set of error-correcting tags (see above), one must make the choice whether to retain primer sequence and/or tags that contain errors (and how many errors are permitted). Although it may be tempting to relax quality standards to increase the number of reads retained, low-quality data will usually require more coverage, for instance for high-confidence heterozygote calls.

3.3.2. Alignments

Calling genotypes (SNPs or haplotypes) requires alignments, i.e. sets of homologous reads. Whereas processing and filtering NGS data is, at its most basic, a matter of text manipulation and editing, alignments require computational resources and efficient algorithms. There are two types of alignment methods, those that use a reference and those that do not (*de novo*). A reference does not necessarily imply a reference genome, such as that from a model organism, but rather some information on the output reads, including, for example, the probe sequences used for target enrichment.

A good, *de novo* assembler (e.g., Velvet, see Table 2) is the gold standard for analysis, but requires considerable computation time and resources. On the other hand, using a reference allows very quick alignment since alignment of high-quality reads can be restricted to the reference (instead of requiring pairwise comparison to each other). There are many NGS analytical tools that align a set of reads to a reference (see Table 2). If no reference is available, for example in most of the restriction-digest based methods, there are clustering programs (like CAP3), which, like *de novo* assemblers, collect reads into groups within a given percent similarity (in addition to other parameters) and generate alignments from these clusters. Several pipelines offer streamlined solutions for taking raw data (without a reference genome) to called genotypes (Table 2). Two examples of open-source software are PRGmatic (Hird et al., 2011), which builds high-confidence clusters into a provisional-reference genome, to which all reads are then aligned; and Stacks (Catchen et al., 2011), in which "stacks" of reads corresponding to loci are created, permitting genotype calling. Many commercially available multi-functional programs will do pre-processing and alignments (Table 2).

3.3.3. Genotype calling

The final step in taking raw NGS data to a format that is useable for phylogeography and phylogenetics is genotype calling, or calling two (diploid) alleles from all the reads for a given SNP or locus

(for a thorough review, see Nielsen et al., 2011). In its most straightforward application, genotype calling can be conducted based on threshold values. For instance, a base position would be declared a valid SNP if polymorphism were detected in a certain threshold percentage of total reads for an individual (say 20%). In some cases, thresholds can lead to bias toward rare alleles (Johnson and Slatkin, 2008). Thus, more statistically savvy genotype calling algorithms are based on probability theory, which permit incorporation of sequencing error (Johnson and Slatkin, 2006; Hellmann et al., 2008; Lynch, 2009; Hohenlohe et al., 2010; Andolfatto et al., 2011; Gompert and Buerkle, 2011).

Pooling individuals is another way of generating population-level allele frequency data (Cutler and Jensen, 2010). It saves money by limiting the number of barcode adaptors or library preparations needed and can be useful for marker discovery and describing some divergence statistics such as F_{ST} (Gompert et al., 2010; Kofler et al., 2011). However, it prevents simultaneous genotype calling at the level of the individual as well as paralog detection on the basis of observed heterozygosity. For the purposes of using NGS for phylogeography and phylogenetics, there seems to be much to gain by adding tags to individuals instead of population pools.

3.4. Data analysis

Because most existing NGS studies in phylogeography and phylogenetics are based on SNPs (Emerson et al., 2010; Gompert et al., 2010; Williams et al., 2010), most analytical approaches used to date are those amenable to SNP data, such as PCA and Structure (for phylogeography) and distance-based methods for inferring phylogenies. PCA has the limitation that it requires complete data matrices (or statistical imputation of missing data), and some methods (especially the restriction-digest methods) are prone to missing data. While a full treatment of NGS data analysis demands its own review, our observation is that NGS data as it is currently being produced is too computationally demanding for the popular suite of probabilistic coalescence-based methods that form the core of phylogenetic and phylogeographic analysis (e.g., BEAST, IMa, species-tree analysis), although using subsets of the data is always an option. We note two trends: (1) longer NGS reads are making analysis with gene trees more feasible and (2) SNP data are increasingly useful for testing demographic hypotheses, for example those involving gene flow (Durand et al., 2011). We address future prospects for data analysis in the next section.

4. Future directions

This is a time of incredible transition in sequencing technology. It is difficult to predict what the future holds, but it seems clear that at some point the important technological advances for phylogeographers and phylogeneticists will plateau with the emergence of affordable whole-genome sequencing. Although the technology currently exists for reasonably inexpensive genome sequencing on the Illumina Hi-Seq platform, the cost for the number of individuals typically employed in a phylogeographic or phylogenetic project is still beyond the reach of most labs. Perhaps more important than the technology is the pace of change to analytical resources. Phylogeography and phylogenetics is built on a firm foundation of resources for analyzing discrete loci. Meanwhile, whole genome analysis is still in its infancy (e.g., Sims et al., 2009; Vishnoi et al., 2010). We argue that, practically speaking, we are less limited by technology than we are by the ability of research labs – i.e., humans – to adapt to new technologies and effectively harness their information content. Thus we echo the sentiments of Davey et al. (2011) that sequencing methods based on reduced representations of the genome (however they are targeted) will remain useful for many years to come, until a strong foundation for analyzing whole genomes emerges. One beneficial contribution of whole-genome analysis will be to incorporate recombination into phylogenetic inference instead of ignoring it or mitigating its effects, as is the current mode when analyzing discrete loci.

An advance needed immediately is that current software for analyzing phylogenetic and population genetic data needs to be scaled-up to handle hundreds of loci in a reasonable timeframe. For example, although some analytical solutions to defining a species tree from many gene trees are nearly as accurate as probabilistic methods and return an answer almost instantaneously (Liu et al., 2009b), probabilistic methods are still more robust in most cases and preferred for difficult questions. The trade-off is whether we are willing to sacrifice some degree of accuracy or certitude in order to have an answer at all, or at least on a timeline suited to today's fast-paced speed of publishing. Alternatively, more thorough probabilistic methods will need to be streamlined and made more efficient.

Finally, as phylogeography becomes more genomic, it is only natural that it will increasingly merge with the now largely separate field of ecological genomics (or adaptation genomics). After all, both investigate the speciation process, but differ principally in their focus on adaptive versus neutral processes and their concomitant use of different subsets of markers. Ecological genomics has utilized candidate genes and, increasingly, the full subset of expressed genes interrogated with transcriptome sequencing. Phylogeography has traditionally used "neutrally evolving" markers; however, there are few phylogeographers that would not also be interested in the actual genes underlying speciation in their system. This joint interest is already flourishing in research utilizing genome scans to detect outlier loci potentially under selection from a large pool of other loci experiencing background (i.e., neutral) divergence. These studies have mostly employed AFLPs, thus it is no surprise that similar analytical techniques geared toward restriction-digest based NGS data (Section 2.2) are also now appearing (Hohenlohe et al., 2010; Gompert and Buerkle, 2011). We assume the integration of neutral and adaptive speciation processes will only accelerate with continuing analytical and technological advances at the level of the genome.

Acknowledgments

Funding was provided by the National Science Foundation (DEB-0956069 and DEB-0841729). We thank members of the Brumfield and Carstens labs, J. Good, T. Glenn, B. Faircloth, J.-M. Roulliard, and S. Herke for conversations regarding next-generation sequencing and comments on the manuscript.

References

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., 2007. Direct selection of human genomic loci by microarray hybridization. Nat. Methods 4, 903–905.

Alkan, C., Kidd, J.M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J.O., Baker, C., Malig, M., Mutlu, O., 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat. Genet. 41, 1061–1067.

Althoff, D.M., Gitzendanner, M.A., Segraves, K.A., 2007. The utility of amplified fragment length polymorphisms in phylogenetics: a comparison of homology within and between genomes. Syst. Biol. 56, 477–484.

Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L., Lander, E.S., 2000. A SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407, 513–516.

Amaral, A.J., Megens, H.J., Kerstens, H.H.D., Heuven, H., Dibbits, B., Crooijmans, R.P.M.A., Den Dunnen, J.T., Groenen, M.A.M., 2009. Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. BMC Genomics 10, 374.

Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T.T., Mast, J., Sunayama-Morita, T., Stern, D.L., 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res. 21, 610–617.

Babik, W., Taberlet, P., Ejsmond, M.J.A.N., Radwan, J., 2009. New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. Mol. Ecol. Resour. 9, 713–719.

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

Barbazuk, W.B., Schnable, P.S., 2011. SNP discovery by transcriptome pyrosequencing. Methods Mol. Biol. 729, 225–246.

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

Bers, N.E.M.V., Oers, K.V., Kerstens, H.H.D., Dibbits, B.W., Crooijmans, R.P.M.A., Visser, M.E., Groenen, M.A.M., 2010. Genome wide SNP detection in the great tit Parus major using high throughput sequencing. Mol. Ecol. 19, 89–99.

Binladen, J., Gilbert, M.T., Bollback, J.P., Panitz, F., Bendixen, C., Nielsen, R., Willerslev, E., 2007. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS One 2, e197.

Briggs, A.W., Good, J.M., Green, R.E., Krause, J., Maricic, T., Stenzel, U., Lalueza-Fox, C., Rudan, P., Brajkovi, D., Ku an, Ž., 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 325, 318–321.

Brito, P.H., Edwards, S.V., 2008. Multilocus phylogeography and phylogenetics using sequence-based markers. Genetica 135, 439–455.

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

Buetow, K.H., Edmonson, M., MacDonald, R., Clifford, R., Yip, P., Kelley, J., Little, D.P., Strausberg, R., Koester, H., Cantor, C.R., 2001. High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Proc. Natl Acad. Sci. USA 98, 581–584.

Cánovas, A., Rincon, G., Islas-Trejo, A., Wickramasinghe, S., Medrano, J.F., 2010. SNP discovery in the bovine milk transcriptome using RNA-Seq technology. Mamm. Genome 21, 592–598.

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

Chan, Y.C., Roos, C., Inoue-Murayama, M., Inoue, E., Shih, C.C., Pei, K.J.C., Vigilant, L., 2010. Mitochondrial genome sequences effectively reveal the phylogeny of Hylobates gibbons. PLoS One 5, e14419.

Chepelev, I., Wei, G., Tang, Q., Zhao, K., 2009. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic Acids Res. 37, e106.

Chevreux, B., Wetter, T., Suhai, S., 1999. Genome sequence assembly using trace signals and additional sequence information. In: Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB), vol. 99, pp. 45–56.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M., Tiedje, J.M., 2009. The ribosomal database project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37, D141–D145.

Craig, D.W., Pearson, J.V., Szelinger, S., Sekar, A., Redman, M., Corneveaux, J.J., Pawlowski, T.L., Laub, T., Nunn, G., Stephan, D.A., 2008. Identification of genetic variants using bar-coded multiplexed sequencing. Nat. Methods 5, 887–893.

Crawford, N.G., Faircloth, B.C., McCormack, J.E., Brumfield, R.T., Winker, K., Glenn, T.C., 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group to archosaurs. Biol. Lett. 8, 783–786.

Cutler, D.J., Jensen, J.D., 2010. To pool, or not to pool? Genetics 186, 41–43.

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

Decker, J.E., Pires, J.C., Conant, G.C., McKay, S.D., Heaton, M.P., Chen, K., Cooper, A., Vilkki, J., Seabury, C.M., Caetano, A.R., 2009. Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. Proc. Natl Acad. Sci. USA 106, 18644–18649.

DePristo, M.A., Banks, E., Poplin, R., et al., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498.

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

Durand, E.Y., Patterson, N., Reich, D., Slatkin, M., 2011. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28, 2239–2252.

Edmonson, M.N., Zhang, J., Yan, C., et al., 2011. Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. Bioinformatics 27, 865–866.

Edwards, S.V., 2008. PERSPECTIVE: a smörgåsbord of markers for avian ecology and evolution. Mol. Ecol. 17, 945–946.

Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging? Evolution 63, 1–19.

Edwards, S.V., Liu, L., Pearl, D.K., 2007. High-resolution species trees without concatenation. Proc. Natl Acad. Sci. USA 104, 5841–5936.

Ekblom, R., Galindo, J., 2010. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107, 1–15.

Emerson, K.J., Merz, C.R., Catchen, J.M., Hohenlohe, P.A., Cresko, W.A., Bradshaw, W.E., Holzapfel, C.M., 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proc. Natl Acad. Sci. USA 107, 16196–16200.

Etter, P.D., Preston, J.L., Bassham, S., Cresko, W.A., Johnson, E.A., 2011. Local de novo assembly of RAD paired-end contigs using short sequencing reads. PLoS One 6, e18561.

Faircloth, B.C., Glenn, T.C., 2012. Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. PLoS One 7, e42543.

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

Fierer, N., Hamady, M., Lauber, C.L., Knight, R., 2008. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. Proc. Natl Acad. Sci. USA 105, 17994–17999.

Geraldes, A., Pang, J., Thiessen, N., Cezard, T., Moore, R., Zhao, Y., Tam, A., Wang, S., Friedmann, M., Birol, I., 2011. SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. Mol. Ecol. Resour. 11 (Suppl. 1), 81–92.

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

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27, 182–189.

Goecks, J., Nekrutenko, A., Taylor, J., Team, T.G., 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11, R86.

Gompert, Z., Buerkle, C.A., 2011. A hierarchical Bayesian model for next-generation population genomics. Genetics 187, 903–917.

Gompert, Z., Forister, M.L., Fordyce, J.A., Nice, C.C., Williamson, R.J., Buerkle, C.A., 2010. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of Lycaeides butterflies. Mol. Ecol. 19, 1473–2455.

Griffin, P.C., Robin, C., Hoffmann, A.A., 2011. A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to Poa grasses. BMC Biol. 9, 19.

Gunnarsdóttir, E.D., Li, M., Bauchet, M., Finstermeier, K., Stoneking, M., 2011. High-throughput sequencing of complete human mtDNA genomes from the Philippines. Genome Res. 21, 1–11.

Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J., Knight, R., 2008. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat. Methods 5, 235–237.

Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27, 570–580.

Hellmann, I., Mang, Y., Gu, Z., Li, P., De La Vega, F.M., Clark, A.G., Nielsen, R., 2008. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome Res. 18, 1020–1029.

Hird, S.M., Brumfield, R.T., Carstens, B.C., 2011. PRGmatic: an efficient pipeline for collating genome enriched second generation sequencing data using a 'provisional reference genome'. Mol. Ecol. Res. 11, 743–748.

Hittinger, C.T., Johnston, M., Tossberg, J.T., Rokas, A., 2010. Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. Proc. Natl Acad. Sci. USA 107, 1476–1481.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., 2007. Genome-wide in situ exon capture for selective resequencing. Nat. Genet. 39, 1522–1527.

Hohenlohe, P.A., Amish, S., Catchen, J.M., Allendorf, F.W., Luikart, G., 2011. RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow trout and westslope cutthroat trout. Mol. Ecol. Res. 11 (Supp. 1), 117–122.

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

Huang, X., Madan, A., 1999. CAP3: A DNA sequence assembly program. Genome Res. 9, 868–877.

Hyten, D.L., Cannon, S.B., Song, Q., Weeks, N., Fickus, E.W., Shoemaker, R.C., Specht, J.E., Farmer, A.D., May, G.D., Cregan, P.B., 2010a. High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC Genomics 11, 38.

Hyten, D.L., Song, Q., Fickus, E.W., Quigley, C.V., Lim, J.S., Choi, I.Y., Hwang, E.Y., Pastor-Corrales, M., Cregan, P.B., 2010b. High-throughput SNP discovery and assay development in common bean. BMC Genomics 11, 475.

Janes, D.E., Chapus, C., Gondo, Y., Clayton, D.F., Sinha, S., Blatti, C.A., Organ, C.L., Fujita, M.K., Balakrishnan, C.N., Edwards, S.V., 2011. Reptiles and mammals have differentially retained long conserved noncoding sequences from the Amniote ancestor. Genome Biol. Evol. 3, 102–113.

Johnson, P.L.F., Slatkin, M., 2006. Inference of population genetic parameters in metagenomics: a clean look at messy data. Genome Res. 16, 1320–1327.

Johnson, P.L.F., Slatkin, M., 2008. Accounting for bias from sequencing error in population genetic estimates. Mol. Biol. Evol. 25, 199–206.

Kenny, E.M., Cormican, P., Gilks, W.P., Gates, A.S., O'Dushlaine, C.T., Pinto, C., Corvin, A.P., Gill, M., Morris, D.W., 2011. Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection. DNA Res. 18, 31–38.

Kerstens, H.H.D., Crooijmans, R.P.M.A., Veenendaal, A., Dibbits, B.W., Chin-A-Woeng, T.F.C., den Dunnen, J.T., Groenen, M.A.M., 2009. Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. BMC Genomics 10, 479.

Kloch, A., Babik, W., Bajer, A., Si Ski, E., Radwan, J., 2010. Effects of an MHC DRB genotype and allele number on the load of gut parasites in the bank vole *Myodes glareolus*. Mol. Ecol. 19, 255–265.

Knowles, L.L., 2009. Statistical phylogeography. Annu. Rev. Ecol. Evol. Syst. 40, 593–612.

Koboldt, D.C., Chen, K., Wylie, T., et al., 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25, 3.

Kocher, T.D., Thomas, W.K., Meyer, A., Edwards, S.V., Pääbo, S., Villablanca, F.X., Wilson, A.C., 1989. Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proc. Natl Acad. Sci. USA 86, 6196–6200.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R.V., Nolte, V., Futschik, A., Kosiol, C., Schlötterer, C., 2011. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One 6, e15925.

Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., Turner, D.J., 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat. Methods 6, 291–295.

Kubatko, L.S., Carstens, B.C., Knowles, L.L., 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25, 971–973.

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

Kuhner, M.K., 2009. Coalescent genealogy samplers: windows into population history. Trends Ecol. Evol. 24, 86–93.

Kumar, S., Carlsen, T., Mevik, B.H., et al., 2011. CLOTU: an online pipeline for processing and clustering of 454 amplicon reads into OTUs followed by taxonomic annotation. BMC Bioinformatics 12, 182.

Langmead, B., Trapnell, C., Pop, M., Salzburg, S.L., 2010. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25.

Lerner, H.R.L., Fleischer, R.C., 2010. Prospects for the use of next-generation sequencing methods in ornithology. Auk 127, 4–15.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 7.

Li, R., Li, Y., Kristiansen, K., Wang, J., 2008. SOAP: short oligonucleotide alignment program. Bioinformatics 24, 713–714.

Li, H., Handsaker, B., Wysoker, A., et al., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.

Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R., Lockhart, D.J., 1999. High density synthetic oligonucleotide arrays. Nat. Genet. 21 (Suppl. 1), 20–24.

Liu, L., Yu, L., Kubatko, L., Pearl, D.K., Edwards, S.V., 2009a. Coalescent methods for estimating phylogenetic trees. Mol. Phyl. Evol. 53, 320–328.

Liu, L., Yu, L., Pearl, D.K., Edwards, S.V., 2009b. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58, 468–477.

Lunter, G., Goodson, M., 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21, 936–939.

Lynch, M., 2009. Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics 182, 295–301.

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., Turner, D.J., 2009. Target-enrichment strategies for next-generation sequencing. Nat. Methods 7, 111–118.

Mardis, E.R., 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. 24, 133–141.

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

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y., 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 18, 1509–1517.

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

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

Medinger, R., Nolte, V., Pandey, R.V., Jost, S., Ottenwälder, B., Schlötterer, C., Boenigk, J., 2011. Diversity in a hidden world: potential and limitation of next generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. Mol. Ecol. 19 (Suppl. S1), 32–40.

Meyer, M., Stenzel, U., Hofreiter, M., 2008. Parallel tagged sequencing on the 454 platform. Nat. Protoc. 3, 267–278.

Milne, I., Bayer, M., Cardle, L., et al., 2010. Tablet – next generation sequence assembly visualization. Bioinformatics 26, 401–402.

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

Morin, R.D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T.J., McDonald, H., Varhol, R., Jones, S.J.M., Marra, M.A., 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 45, 81–94.

Morin, P.A., Archer, F.I., Foote, A.D., Vilstrup, J., Allen, E.E., Wade, P., Durban, J., Parsons, K., Pitman, R., Li, L., 2010. Complete mitochondrial genome phylogeographic analysis of killer whales (*Orcinus orca*) indicates multiple species. Genome Res. 20, 908–916.

Nabholz, B., Künstner, A., Wang, R., Jarvis, E., Ellegren, H., 2011. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. Mol. Biol. Evol. 28, 2197–2210.

Naduvilezhath, L., Rose, L.E., Metzler, D., 2011. Jaatha: a fast composite likelihood approach to estimate demographic parameters. Mol. Ecol. 20, 2709–2723.

Neiman, M., Lundin, S., Savolainen, P., Ahmadian, A., Andersen, M., 2011. Decoding a substantial set of samples in parallel by massive sequencing. PLoS One 6, e17785.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., 2009. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461, 272–276.

Nicod, J.C., Largiadèr, C.R., 2003. SNPs by AFLP (SBA): a rapid SNP isolation strategy for non model organisms. Nucleic Acids Res. 31, e19.

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

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., 2008. Genes mirror geography within Europe. Nature 456, 98–101.

Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J., Zwick, M.E., 2007. Microarray-based genomic selection for high-throughput resequencing. Nat. Methods 4, 907–909.

Oliver, T.A., Garfield, D.A., Manier, M.K., Haygood, R., Wray, G.A., Palumbi, S.R., 2010. Whole-genome positive selection and habitat-driven evolution in a shallow and a deep-sea urchin. Genome Biol. Evol. 2, 800–814.

Parks, M., Cronn, R., Liston, A., 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 7, 84.

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G., Baurain, D., 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9, e1000602.

Pinho, C., Hey, J., 2010. Divergence with gene flow: models and data. Annu. Rev. Ecol. Evol. Syst. 41, 215–230.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Ramos, A.M., Crooijmans, R., Affara, N.A., Amaral, A.J., Archibald, A.L., Beever, J.E., Bendixen, C., Churcher, C., Clark, R., Dehais, P., 2009. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS One 4, e6524.

Rice, A.M., Rudh, A., Ellegren, H., Qvarnström, A., 2011. A guide to the genomics of ecological speciation in natural animal populations. Ecol. Lett. 14, 9–18.

Sánchez, C.C., Smith, T.P.L., Wiedmann, R.T., Vallejo, R.L., Salem, M., Yao, J., Rexroad, C.E., 2009. Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. BMC Genomics 10, 559.

Seeb, J.E., Pascal, C.E., Ramakrishnan, R., Seeb, L.W., 2009. SNP genotyping by the 5-nuclease reaction: advances in high throughput genotyping with non-model organisms. In: Komar, A.A. (Ed.), Single Nucleotide Polymorphisms, Methods in Molecular Biology. Humana Press, Totowa, New Jersey, pp. 277–292.

Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135–1145.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.D.W., Richards, S., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050.

Simpson, J.T., Wong, K., Jackman, S.D., et al., 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123.

Sims, G.E., Jun, S.R., Wu, G.A., Kim, S.H., 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Natl Acad. Sci. USA 106, 2677–2682.

Sirén, J., Marttinen, P., Corander, J., 2011. Reconstructing population histories from single nucleotide polymorphism data. Mol. Biol. Evol. 28, 673–683.

Slater, G., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31.

Smith, M.J., Pascal, C.E., Grauvogel, Z., Habicht, C., Seeb, J.E., Seeb, L.W., 2011. Multiplex preamplification PCR and microsatellite validation enables accurate single nucleotide polymorphism genotyping of historical fish scales. Mol. Ecol. Res. 11, 268–277.

Stapley, J., Reger, J., Feulner, P.G.D., Smadja, C., Galindo, J., Ekblom, R., Bennison, C., Ball, A.D., Beckerman, A.P., Slate, J., 2010. Adaptation genomics: the next generation. Trends Ecol. Evol. 25, 705–712.

Stephen, S., Pheasant, M., Makunin, I.V., Mattick, J.S., 2008. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol. Biol. Evol. 25, 402–408.

Sunnucks, P., Wilson, A.C.C., Beheregaray, L.B., Zenger, K., French, J., Taylor, A.C., 2000. SSCP is not so difficult: the application and utility of single-stranded conformation polymorphism in evolutionary biology and molecular ecology. Mol. Ecol. 9, 1699–1710.

Tewhey, R., Nakano, M., Wang, X., Pabón-Peña, C., Novak, B., Giuffre, A., Lin, E., Happe, S., Roberts, D.N., LeProust, E.M., 2009a. Enrichment of sequencing

targets from the human genome by solution hybridization. Genome Biol. 10, R116.

Tewhey, R., Warner, J.B., Nakano, M., Libby, B., Medkova, M., David, P.H., Kotsopoulos, S.K., Samuels, M.L., Hutchison, J.B., Larson, J.W., 2009b. Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat. Biotechnol. 27, 1025–1031.

Thomson, R.C., Wang, I.A.N.J., Johnson, J.R., 2010. Genome enabled development of DNA markers for ecology, evolution and conservation. Mol. Ecol. 19, 2184–2195.

Van Orsouw, N.J., Hogers, R.C.J., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., Van Der Poel, H., Van Oeveren, J., Verstegen, H., 2007. Complexity reduction of polymorphic sequences (CRoPS™): a novel approach for large-scale polymorphism discovery in complex genomes. PLoS One 2, e1172.

Van Tassell, C.P., Smith, T.P.L., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., Sonstegard, T.S., 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5, 247–252.

Vishnoi, A., Roy, R., Prasad, H.K., Bhattacharya, A., Desalle, R., 2010. Anchor-based whole genome phylogeny (ABWGP): a tool for inferring evolutionary relationships among closely related microorganims. PLoS One 5, e14159.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 11, 4407–4414.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280, 1077.

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

Wiedmann, R.T., Smith, T.P.L., Nonneman, D.J., 2008. SNP discovery in swine by reduced representation and high throughput pyrosequencing. BMC Genet. 9, 81.

Williams, L.M., Ma, X., Boyko, A.R., Bustamante, C.D., Oleksiak, M.F., 2010. SNP identification, verification, and utility for population genetics in a non-model genus. BMC Genet. 11, 32.

Zellmer, A.J., Hanes, M.M., Hird, S.M., Carstens, B.C., 2012. Deep phylogeographic structure and environmental differentiation in the carnivorous plant *Sarracenia alata*. Syst. Biol. 61, 763–777.

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