Chapter 3
Targeted DNA Region Re-sequencing

Karolina Heyduk, Jessica D. Stephens, Brant C. Faircloth, and Travis C. Glenn

3.1 Different Types of Re-sequencing Methodologies

Multiple re-sequencing approaches have been developed and reviewed (McCormack et al. 2013a; Lemmon and Lemmon 2013). Below, we briefly summarize the major re-sequencing methods, indicating their advantages and disadvantages (Table 3.1) and the scale at which they are most appropriate (Fig. 3.1). For all methods, we assume that sequencing coverage will be reasonably deep to achieve high accuracy (Table 3.2), especially at heterozygous sites. All methods are usually paired with DNA sequence tags (also known as barcodes, indexes, or molecular identifiers, MID tags; see Faircloth and Glenn 2012) to identify individual samples from a pool of samples. We assume that lower costs will increase how widely the techniques will be adopted, and that total costs of ≤$100 US/sample, including personnel costs, are highly desirable.

Karolina Heyduk and Jessica D. Stephens are contributed equally with all other contributors.

K. Heyduk • J.D. Stephens
Department of Plant Biology, University of Georgia,
2502 Miller Plant Sciences, Athens, GA 30602, USA
e-mail: heyduk@uga.edu; jdstephe@uga.edu

B.C. Faircloth
Department of Biological Sciences and Museum of Natural Science,
Louisiana State University, 202 Life Sciences Bldg., Baton Rouge, LA 70803, USA
e-mail: brant@faircloth-lab.org

T.C. Glenn (✉)
Department of Environmental Health Science, University of Georgia,
150 East Green St, Athens, GA 30602, USA
e-mail: travisg@uga.edu

A.M. Aransay, J.L. Lavín Trueba (eds.), *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing*,
DOI 10.1007/978-3-319-31350-4_3

Table 3.1 Advantages and disadvantages of DNA re-sequencing methods

Re-sequencing approaches	Advantages	Disadvantages
Whole genome re-sequencing	– Easy to implement in lab	– Must have a reference genome from same or closely related species
	– Most complete data	– Large genomes require a lot of sequencing
	– Many robust software options	– Large genomes require more computational effort
Transcriptome sequencing (RNA-seq)	– Already reduced complexity of genome	– Differences in expression across tissue types, developmental stage, etc.
	– Template for future marker design	– Data may violate population genomics assumptions
	– Good platform for comparison across species/individuals	– Expensive
		– Computationally intensive
PCR amplicon sequencing	– Works well with limited starting material	– Issues with sequence diversity on Illumina platforms
	– Cost efficient with few loci	– Assay development time and costs increase with number of loci
Restriction-site-associated DNA markers (RADseq)	– Discovery, development, and screening of markers are time and cost efficient	– Loci are dominant
	– Established bioinformatics pipelines	– Significant variation in reproducibility
	– Cost-efficient method	– Can result in large amounts of missing data
		– Issues with paralog determination and coverage due to untargeted loci
Target enrichment	– Less likely to have allelic loss	– Need prior genetic resources to design baits
	– Baits can target areas of interest	– Bait design can be challenging
	– Useful for large, complex genomes	– Higher up-front costs (e.g., library prep, bait design) relative to RADseq
	– Upstream methods help avoid paralogs, complexity, and repetitive regions	

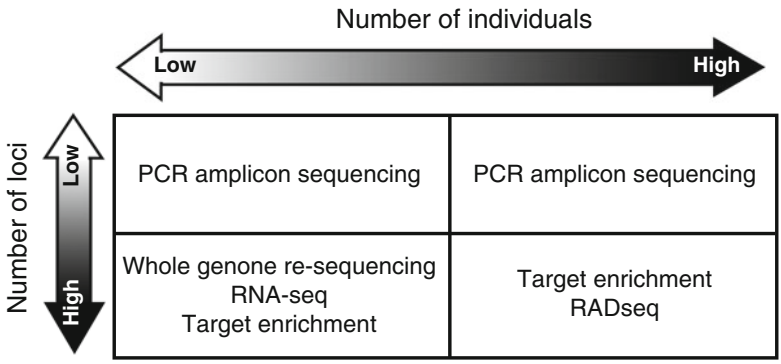

Fig. 3.1 Methods for re-sequencing based on number of individuals and loci for analyses

3.1.1 Whole Genome Re-sequencing

Whole genome re-sequencing (WGRS) is the easiest method to implement in the lab, offers the most complete data, and has excellent software support due to its widespread use in human genomics (for a review of software, see Bao et al. 2011). While WGRS studies are being published in nonhuman systems, these are mostly limited to agriculturally important crops (rice, Xu et al. 2010; soybean, Li et al. 2013) or model organisms (*Arabidopsis*, www.1001genomes.org; *Mus*, Keane et al. 2011; *Drosophila*, Zhu et al. 2012). The lack of WGRS studies are due to the inherent problems associated with WGRS; these include (1) a required reference genome from the same or a closely related species, (2) the amount of sequencing is directly proportional to genome size (i.e., big genomes require a lot of sequencing), and (3) computational efforts increase as a power function of genome size (i.e., large genomes require much more computational effort than small ones)—all of which increase costs. As of 2015, it is possible to re-sequence a human genome at 30× coverage for ~$1000 on Illumina HiSeq 4000s (www.illumina.com). Thus, it is possible to sequence *Drosophila*-sized genomes for a cost approaching $100/sample, but most other non-model and large-genome organisms remain uneconomical for WGRS efforts.

3.1.2 Transcriptome Sequencing

Transcriptome sequencing (RNA-seq) has the advantage of using the cellular transcriptional machinery to naturally reduce the complexity of genomes and enrich for functional elements. There are multiple advantages of focusing on genome reduction through transcriptomics. For example, transcript profiles for polymorphism comparisons are predicted to be similar if using the same tissue across

Table 3.2 Recommended amount of starting nucleic acids, major constraints of methods, minimum average recommended sequencing depth, and recommended read lengths (Illumina) for DNA re-sequencing methods

Technique	Recommended starting material (ng)	Constraints	Average Sequencing depth	Recommended sequencing run	Reference(s)
Whole genome re-sequencing	500	Cost of sequencing	6–30×	HiSeq or NextSeq PE75–PE150	Sims et al. (2014), Ekblom and Wolf (2014)
Transcriptome sequencing	1000	Sample material, cost of libraries, cost of sequencing	≥10 million reads per sample	HiSeq or NextSeq PE100–PE150	Ozsolak and Milos (2011), Wang et al. (2009), Wang et al. (2011)
PCR amplicon sequencing	20	Combinatorial tags, pool with diverse libraries	20×	MiSeq PE250–PE300	Feng et al. (2016), Mamanova et al. (2010)
Restriction-site-associated DNA markers (RADseq)	100	Consistency	10×	HiSeq or NextSeq SE75–PE150	Davey et al. (2011), Davey et al. (2013)
Target enrichment by sequence capture	500	Probes (information to design and cost)	30×	HiSeq or NextSeq PE100–PE150	Mamanova et al. (2010), Mertes et al. (2011)

individuals or species. There are large-scale initiatives attempting just that through a consortium of universities (plants, 1KP project, http://www.onekp.com; insects, 1KITE, http://www.1kite.org; eukaryote microbes, Marine Microbial Eukaryote Transcriptome Sequencing Project, http://www.marinemicroeukaroytes.org). Another benefit of transcriptome sequencing is that the assembled template can be used to develop markers for future studies (Ekblom and Galindo 2011).

RNA-seq has several disadvantages. First, differences in gene expression will vary depending on which tissues are collected, developmental stage of tissue, time of day, and nutritional status of individuals; this can limit comparison of orthologous loci across samples. Variation between libraries can be mitigated, however, by pooling several life stages, tissues, etc. during cDNA library preparation (Hahn et al. 2009). Second, RNA-seq requires significant sequencing depth to account for loci that are weakly expressed. Third, models relating to demographic history and population structure generally assume neutral evolutionary processes, which may be violated by transcribed genes and thus may cause problems with downstream analyses for these types of studies. Finally, RNA-seq currently costs one to a few hundred dollars per sample; thus, sampling a large number of individuals and species can be costly for reagents and sequencing and can increase computational time requirements for transcriptome assembly and subsequent analysis (Wang et al. 2009; Ozsolak and Milos 2011).

3.1.3 PCR Amplicon Sequencing

PCR can be used to produce amplicons that are sequenced using MPS. This has most frequently been done for 16S metagenomics (Wang and Qian 2009; Haas et al. 2011) and specific disease panels (Easton et al. 2015), but many other applications of this technique have been developed (Faircloth and Glenn 2012). Amplicon sequencing has the advantage of working from very limited amounts of starting material, building on well-known techniques, and can be done for well under $100 US per sample if the number of target loci is limited. The major disadvantages of amplicon sequencing are that (1) costs increase significantly as the number of target loci increases, (2) amplicons generally need to be combined with other samples to increase sequence diversity on Illumina platforms and to take advantage of capacity, and (3) assay development time and costs increase significantly as the number of target loci increases; thus, amplicon sequencing is generally limited to surveying only a very small portion of the genome.

3.1.4 Restriction-Site-Associated DNA Makers (RADseq)

RADseq uses restriction enzymes to reduce genome complexity and isolate a smaller, repeatable fraction of the genome and is combined with MPS to genotype thousands of genetic markers without having prior genetic information for the

organism(s) under study. Multiple flavors of RADseq have been developed, making use of one, two, three, or more restriction enzymes (Davey et al. 2011; Puritz et al. 2014). The method used is often selected based on the genome size of the organism and the predicted amount of coverage resulting from the enzyme combination selected. RADseq was developed for and has been extensively utilized for questions pertaining to genetic mapping and population genomics (Davey et al. 2011; Puritz et al. 2014). RADseq data have also been used for phylogenetic assessments (Rubin et al. 2012; Cariou et al. 2013; Wagner et al. 2013), but these are often in small, species-level phylogenies. A major advantage of RADseq is that discovery, development, and screening of markers generally happens in only one or two rounds of MPS, making RADseq time efficient and cost-effective (Davey and Blaxter 2010). In addition, there are well-developed downstream bioinformatics pipelines to handle these data (e.g., Stacks—Catchen et al. 2013; PyRAD—Eaton 2014). Although RADseq is inefficient in its use of MPS data (i.e., most data are discarded), because MPS data are cheap, most RADseq projects still achieve costs well below $100 US/sample. Thus, RADseq represents a generally reasonable approach for acquiring genotype information dispersed across large genomes.

Unfortunately, RADseq also suffers from several disadvantages. First, RADseq loci are untargeted (i.e., any fragment of DNA with the restriction site(s) will be obtained). Thus, the loci may be less evenly spread across a genome than desired and may miss important portions simply due to chance or bias (Davey et al. 2013). Second, RADseq loci are dominant—substitutions that cause the loss of restriction sites create null alleles (Gautier et al. 2012; McCormack et al. 2013a). Thus, RADseq is not recommended for deeper-level phylogenetics because variation in restriction sites that occurs across divergent taxa yields large amounts of missing data across a given taxonomic sample (McCormack et al. 2012). Third, most RADseq users experience significant variance in reproducibility among taxa or projects, which can cause many samples to fail quality control, increasing the number of samples that must be repeated. Fourth, the variance inherent in RADseq (Davey et al. 2013) frequently results in sparse data matrixes. Finally, RADseq also presents challenges post-sequencing when trying to determine whether fragments are paralogs and have appropriate coverage, because they were not targeted (McCormack et al. 2013a).

3.1.5 Target Enrichment

Target enrichment approaches (also known as sequence capture and gene capture) use baits (also known as probes) to specifically pull out fragments of interest from a genomic library, keeping the fragments of interest while fragments that do not hybridize to the baits are washed away (Mamanova et al. 2010). In contrast to RADseq, target enrichment has higher up-front costs, both for library preparation and the cost of baits and capture, but is more efficient than RADseq because specific targeted areas make up large portions of the data (Grover et al. 2011). Target

enrichment is less likely than RADseq to suffer from allelic loss (null alleles) because alleles with one to several substitutions are recovered at a higher rate across individuals and species. In addition, target enrichment baits can be designed to target a variety of genomic locations including intergenic regions assumed to evolve under neutral processes, making this method ideal for population-level questions. Target enrichment is also useful for organisms with large, complex genomes (such as plants or amphibians) because targeting specific regions can avoid repetitive elements. These strengths of target enrichment result from *a priori* upstream methods to eliminate potentially paralogous sequences, regions of low complexity, and repetitive regions while focusing on those targeted regions of interest and returning data having high coverage across these regions. Moreover, baits can be designed to target regions of varying size depending on different treatments of the data during library preparation and the MPS platform used (McCormack et al. 2013a).

Disadvantages of target enrichment include: (1) prior genetic resources are needed to design baits (e.g., genomes, genomic regions, or transcriptomes of related species); (2) bait design can sometimes be challenging when targeting genomic regions that are highly variable within and among species (e.g., introns, immune-coding loci); and (3) most target enrichment studies to date have focused on using genomic libraries of randomly sheared DNA, which are more expensive to create than RADseq libraries and result in less coverage of targeted bases per sequence. Below, we discuss study design and bioinformatic methods to ameliorate many of these disadvantages, with a focus on target enrichment for population genetic and phylogenetic studies.

3.2 Experimental Design Considerations

As with any study, understanding the biology of the organism(s) of interest is critically important to study design and downstream analyses. For instance, knowing whether the organism under study has undergone recent gene/genome duplications, whether the organism is polyploid, and/or whether the lineages being studied frequently hybridize can have a dramatic influence on data collection and subsequent inference. Paralogs, hybridization, and horizontal gene transfer can influence gene tree discordance for phylogenetic analyses. In addition, many programs have a long list of assumptions or may not properly model aspects of the study system if the proper number of samples has not been sequenced. As an example, *BEAST is an excellent program for coestimating gene trees and their underlying species tree using a Bayesian MCMC procedure; however, the authors of *BEAST recommend the use of at least two individuals per species to properly estimate population parameters (Heled and Drummond 2010). Knowing this prior to sequencing can help better inform experimental design and simplify downstream analyses.

When considering the correct number of individuals per species to sample, in a phylogenetic context, it is mostly based on preference, study system, sample availability, and downstream analyses. If the study system has frequent hybrids or

taxonomic designations below the species level, then one may consider including multiple exemplar individuals for a given species to examine reciprocal monophyly within species. In this case, a phylogenetic program that assigns individuals to species and then infers the phylogeny of the species may be more appropriate than having a phylogeny where every individual represents a lineage. Moreover, some phylogenetic programs require that every gene has a representative sequence from an out-group (Table 3.3). Therefore, it may be advantageous to include multiple exemplar individuals of the out-group species to increase the likelihood of capturing a high number of targets in the out-group. This is especially important to consider if the out-group was not used in the bait design and is distantly related to the in-group species, which would result in more sequence variability in regions targeted by the hybrid enrichment baits between out-group and in-group members. Whenever possible, it is recommended that multiple individuals per species are sequenced, as it not only helps analyses but safeguards against species or population dropout due to unexpected low sequence coverage or low enrichment efficiency of any particular sample. While multiple exemplars per species or populations are beneficial to both phylogenetic and population genomic inferences, if the taxonomic sample is large, then it may not be cost-effective or computationally efficient to include multiple individuals per species.

In contrast to phylogenomic studies, the number of individuals used for population genomic studies is more contingent on capturing rare alleles within a population. Having prior knowledge of the system (i.e., population size, generation time, etc.) can better inform this decision. Ideally sampling a larger number of individuals per population is better, but sample size is dependent upon sample availability, number of populations, number of sequence tags needed for pooling samples, and overall sequencing costs, including the benchwork costs and amount of sequencing required to obtain adequate coverage. Obtaining samples for population-level work can also be more difficult. However, for both phylogenetic and population-level sequencing, DNA from preserved samples (i.e., herbaria, zoological collections, etc.) have been successfully sequenced using target enrichment methods (e.g., Carpenter et al. 2013; Enk et al. 2014; Comer et al. 2015; McCormack et al. 2015). The ability to use fragmented DNA for target enrichment greatly facilitates the sequencing of larger sets of individuals.

When deciding on the number of loci to target, it is best to plan on some modest proportion of the loci being dropped from analysis due to low coverage or poor enrichment across taxa. Thus, designing baits for a large amount of target loci will help to keep the final number of loci analyzed at the desired level, even after filtering poorly covered targets. The number of targeted loci that may actually be used for analysis varies among studies, ranging from 35 % to close to 100 % (Heyduk et al. 2016; McCormack et al. 2013b; Stephens et al. 2015a). These numbers can vary depending on biology and evolutionary history of the focal organisms, the phylogenetic scope or population divergence among the samples, and the number of samples that will be included (e.g., if a locus needs to be present in at least 50% of individuals to be analyzed, then increasing the number of samples makes this threshold harder to reach).

Table 3.3 Some examples of phylogenetic programs that handle gene tree discordance due to incomplete lineage sorting

Program[1]	Phylogenetic method	Multiple accessions	Missing data[2]	Requirements	Command line vs. GUI	Computational time[3]
BEST (Liu 2008)	Sequence alignment	Can assign accessions to species	Accepts missing data	None	Command line	Very slow (months)
*BEAST (Heled and Drummond 2010)	Sequence alignment	Multiple accessions recommended	Does not allow missing data for a given species	Priors	GUI	Very slow (months)
SVDquartets (Chifman and Kubatko 2014)	Sequence alignment	Cannot assign accessions to species	Accepts missing data	None	Command line	Very fast (hours to days)
STEM (Kubatko et al. 2009)	Summary method	Cannot assign accessions to species	Accepts missing data	Must have an estimate of individual gene evolution relative to each other	Command line	Fast (days)
MP-EST (Liu et al. 2010)	Summary method	Can assign accessions to species	Accepts missing data	All gene trees must be rooted	Both	Fast (days)
ASTRAL (Mirarab et al. 2014)	Summary method	Cannot assign accessions to species	Accepts missing data	Gene trees need to be fully resolved (can be unrooted)	Command line	Very fast (hours)

[1]There are many widely used programs not mentioned here that should be considered as well (e.g., NJst, STAR). Program parameters (e.g., multiple accessions, computational time, requirements) should be considered prior to re-sequencing depending on purpose of the study

[2]While certain programs can handle missing data, it should be noted that how missing data influences species tree estimation is not well known across programs. Thus, caution is warranted when including missing data in any phylogenetic analysis, although the use of complete matrixes can also introduce bias (Huang and Knowles 2014)

[3]Computation time is based on a phylogenetic analysis of ~70 individuals with ~200 loci on a Linux cluster with up to 75 CPUs using 8 or 12 core nodes

Determining the number of targeted loci may also be dependent on the system of interest and the study question. Questions pertaining to population genomics would benefit from sampling as many loci as the cost of sequencing allows to ensure detection of outlier loci which can improve parameter estimates such as effective population size and relatedness (Luikart et al. 2003). For studies that are examining population differentiation in phenotypic space, a larger number of loci are important to be able to accurately pinpoint genomic regions responsible for any local adaptation. On the other hand, genomic studies assessing population structure at a fine scale would benefit from highly informative loci. When selecting the number of loci to target for phylogenomic studies, the decision is equally situational. For example, if the study system has been historically difficult to resolve due to rapid or recent radiation and/or high levels of gene tree discordance, then including more genes or more informative genes in the analyses should improve resolution of species relationships. Although one would always prefer highly informative loci, it is difficult to predict which loci will be informative *a priori*. Lastly, computational time should be taken into account when adding more loci to any study, as many statistically robust methods (e.g., *BEAST, see "Post-sequencing") are unable to handle large datasets, and analysis time increases with each locus.

The types of genomic regions (e.g., exons, introns, etc.) collected using target enrichment can vary within or across studies. General approaches range from collecting single loci with single baits to using multiple baits to collect loci spread throughout the genome to collecting data from a single long region of interest with overlapping (tiled) baits (see bait design below) Exons are common targets, including collection of all the exons (i.e., the exome) of model organisms, but any region of the genome may be targeted by baits.

The use of ultraconserved elements (UCEs) for target enrichment is becoming popular given their applicability across extremely divergent taxa (Bejerano et al. 2004; Faircloth et al. 2012; McCormack et al. 2012). UCEs are highly conserved genomic regions that are ≥60 bp and found among widely divergent taxa (Bejerano et al. 2004; Dermitzakis et al. 2005). UCEs are appealing as targets because they are abundant, extremely conserved, straightforward to identify, and found within many groups of organisms (Stephen et al. 2008). In addition, UCEs tend to be orthologous (Derti et al. 2006) with few retroelement insertions. Finally, their utility for phylogenomic approaches is that while UCEs themselves show reduced variation, making them easy to capture, the flanking regions show much higher counts of informative sites (Faircloth et al. 2012). Several research groups have targeted conserved elements for target enrichment approaches, and much work remains to test and optimize the methods of identifying and using such loci. Here, we have focused on those methods that are open-access, because they are amenable to continued optimization and improvement by the research community.

3.2.1 Cost Reductions

The method used for re-sequencing can vary based on the number of individuals and number of loci required to address the questions of interest (Fig. 3.1). For questions that require sampling a limited number of individuals (<50) at very few loci (1–3), traditional PCR and Sanger sequencing may be the most cost- and time-effective methods. On the other end of the spectrum, a one-time study requiring many loci for few individuals might be best served by transcriptome sequencing. For studies requiring the collection of large numbers of loci from large numbers of individuals, then RADseq and/or target enrichment could be warranted. RADseq produces libraries at the lowest cost per sample, but more funds are spent on sequences that ultimately will not be used. Target enrichment significantly reduces both cost and time spent on sequencing, but methods to reduce costs prior to sequencing are important. Below we focus on ways to reduce costs for target enrichment.

Although a variety of home-brew methods are possible, commercial synthesis of target enrichment baits is the most convenient and cost-effective method for most researchers to conduct target enrichment (Fig. 3.2). Most companies that provide baits offer both premade kits and custom bait designs. A wide spectrum of baits can be accommodated, ranging from single biotinylated oligos from traditional oligonucleotide manufacturers (e.g., IDT, Life Technologies, Sigma, etc.) to companies that use high-density microarray technologies (e.g., Agilent, MYcroarray, NimbleGen, etc.) to construct massive numbers of unique baits. If <100 baits are needed, traditional biotinylated oligonucleotides are generally most economical. For example, if a study requires few loci for a large number of individuals, one might consider homemade baits complementary to the sequences of interest (e.g., for studies focusing on one pathway or known genes of interest). This methodology typically requires the bait sequence of interest to be PCR amplified, then subsequently size selected and biotinylated (see Peñalba et al. 2014 for methodological descriptions). If >1000 baits are needed, then high-density approaches for bait construction are most economical. Whole-exome capture kits for humans and model species can include hundreds of thousands of baits.

Although custom, commercial, high-bait number kits have list costs of hundreds of dollars per sample, many methods are available for reducing the costs of target enrichment when using such kits. First, it has long been appreciated that pooling sample libraries prior to conducting enrichment hybridization is an efficient way to reduce costs (Fig. 3.2; Cummings et al. 2010; Shearer et al. 2012). In this strategy, individual samples are tagged during library construction and pooled prior to target enrichment. This allows the costs of target enrichment to be divided among multiple samples. Pooling generally ranges from 2 to 96 samples per pool, with trade-offs between better coverage (i.e., less variance in capture efficiency and read depth with fewer samples per pool) and better cost savings (more samples per pool). In practice, we generally pool 4 to 12 samples prior to enrichment (Faircloth et al. 2012; Heyduk et al. 2016; Stephens et al. 2015a; http://ultraconserved.org). When pooling, samples should have similar: molarity (i.e., accounting for insert size and concentration), copy number (i.e., accounting for genome size and ploidy), and sequence

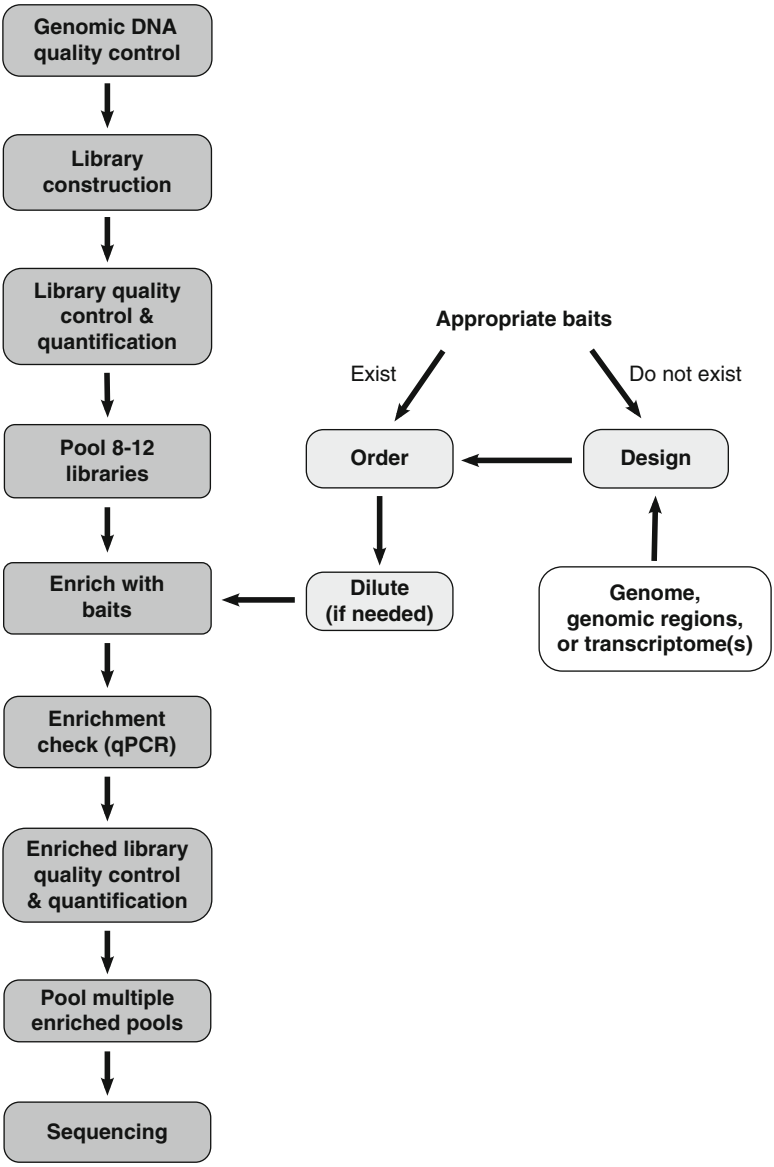

Fig. 3.2 Overview of the wet-lab workflow for target enrichment

divergence from the baits (or phylogenetic distance from the taxon used for bait design). Any of these three factors can lead to preferential capture of loci in higher number from some of the taxa in the pool (i.e., those with more targets or those with targets more similar to the baits than other individuals in the pool).

In addition to pooling prior to hybridization reactions, the quantity of baits per reaction may also be decreased if the targeted number of base pairs is significantly smaller than the protocol assumes (Faircloth et al. 2012; Heyduk et al. 2016; http://ultraconserved.org). Indeed, flooding the reaction with an overwhelming excess of baits relative to genomic targets can reduce capture efficiency rather than increase it. As a simple example, consider a project in which a researcher wishes to survey 1000 loci from 960 individuals. That research might design 2 baits per locus × 1000 loci = 2000 baits. A single custom bait kit that normally allows 12 captures, each with a 20,000 bait pool, is all that is necessary to conduct this experiment because the researcher can dilute the baits tenfold (20,000/2000 = 10; yielding enough baits for 120 captures instead of 12) and pool 8 samples per capture (120 × 8 = 960). Additional hybridization reagents will be necessary, but these can be purchased commercially or made from common reagents (Blumenstiel et al. 2010; http://ultraconserved.org).

Library preparation costs are another significant expense for target enrichment. Library costs can be reduced by decreasing reaction sizes and/or using home-brew protocols (e.g., Meyer and Kircher 2010; Fisher et al. 2011; Glenn et al. 2016; http://ultraconserved.org) rather than commercial kits. Strategically choosing a sequence tagging scheme can reduce costs as well. Illumina sequencing was once limited to a single 6 nt index. Newer methods allow two indices per fragment, employing a combinatorial approach that increases the versatility of indexing. With the dual-indexing method, n unique barcodes for each side of the fragment can be used on n^2 libraries to reduce the number, complexity, and cost of barcode oligos.

Finally, in addition to the on-target sequences captured, target enrichment methods also yield off-target bonus sequences (i.e., DNA sequence lagniappe). Off-target sequences are unavoidable because no target enrichment process is perfectly efficient. Thus, sequences that have partial similarity to the baits or were simply present in the pre-enrichment library, especially in high-copy numbers, will be present post-enrichment. As a result, high-copy DNA from chloroplasts, mitochondria, and ribosomes are commonly sequenced as off-target reads. These sequences are often informative however, and studies in both plants and animals have used these bonus sequences to assemble complete or mostly complete chloroplast and mitochondrial genomes (Weitmeier et al. 2014; Stephens et al. 2015a,b; Meiklejohn et al. 2014; Raposo do Amaral et al. 2015).

3.2.2 Workflow Bottlenecks

Sequence capture is highly effective at generating a large number of sequences for many individuals rapidly and consistently. While sequencing methods continue to improve, a number of bottlenecks exist in current workflows for sequence capture. The speed at which hundreds of libraries can be generated is limited by human labor, although protocols exist for robotic library preparation (e.g., Fisher et al. 2011; Rohland and Reich 2012). Quantification of hundreds of libraries

pre-hybridization is expensive in both time and cost, depending on the method used. Most hybridization methods currently require ≥12 h for libraries to hybridize to baits. Shorter hybridization times are possible but generally require shorter baits, which require trade-offs in specificity and ability to capture library fragments with small sequence differences. Post-sequencing bioinformatic analysis is often not limited by human labor but by computational power; the same hundreds of libraries that take human hours to create may take many days and gigabytes of memory to analyze. For both pre- and post-sequencing, the number of individuals is the most influential limitation to sequence capture projects. As library protocols become more efficient and analysis programs are written to accommodate large numbers of individuals sequenced at many loci, sequence capture bottlenecks will decrease, and multi-species phylogenies and robust population genomics studies will become the norm.

3.3 Bioinformatics

3.3.1 Pre-sequencing

Initial bioinformatics work will depend on whether capture baits are being designed in-house or are available from a prior study (e.g., ultraconserved elements (UCEs), Faircloth et al. 2012). Bait design *de novo* requires genomic resources and can be conducted using genome sequences, transcriptomes, or even EST databases (Fig. 3.2). Comparative analyses of genomic data from divergent taxa can be used to design baits that will work across study systems including divergent taxa; for example, using regions that are conserved across a family will result in baits more likely to anneal to targeted regions and thus give more representative sequences per species. If the study requires examination of intra- and interspecific variation, then baits must be designed so they capture fragments with informative intraspecific sequence differences while still being able to capture targets across species (Stephens et al. 2015b), or sufficient amounts of sequence polymorphisms must accumulate in the regions immediately flanking the conserved sequences used for baits (Faircloth et al. 2012; Smith et al. 2014). This technique could also be applied to bait design for population-level questions. In particular, having genomic resources for multiple populations across the range of interest will help ensure baits are designed that maximize differences between and among populations.

Avoiding duplicated sequences is paramount to both phylogenomic and population genomic analyses, and care should be taken to exclude regions of the genome present in more than one copy (Faircloth et al. 2012, 2015). Prior to bait design, all repeat-like regions across the source data should be masked, and bait design protocols should avoid these regions. It is also recommended that potential areas for targeting should be aligned within and among species to ensure that targets are orthologous and only present in a single copy, especially in systems where polyploidy is abundant (e.g., low-copy genes across angiosperms described in Duarte

et al. 2010, as done in Heyduk et al. 2016). Once targets have been determined, baits can be designed in-house (cf. http://ultraconserved.org), or target sequence information can be sent to commercial companies for bait design and synthesis. Bait sets may be designed having one bait per target or including multiple baits that are overlapped (tiled) across longer regions. Whether or how much to overlap baits depends primarily on the size of the targets, the number of baits, and research budget. Additionally, the sequence similarity of the taxa of interest will influence not only the optimal amount of overlap but also if multiple baits per locus (i.e., baits designed from multiple taxa) are necessary or desirable. Light (2×) tiling (i.e., each target nucleotide has two baits) can increase capture success even when targets are small and the target species are similar, thus decreasing sequencing costs but increasing bait costs relative to no tiling.

3.3.2 Post-sequencing for Phylogenomics Designs

Bioinformatics analysis post-sequencing can be quite daunting, but more pipelines and programs are being designed to handle these data. For example, those targeting UCEs can use phyluce (Faircloth 2016; https://github.com/faircloth-lab/phyluce) to go from raw reads to final alignments for phylogenetic analyses, with an added bonus of flexibility regarding how baits were designed. Throughout this process, phyluce will output relevant summary information that can be reported in a table as a supplement to the manuscript (see reporting section below). An alternative method from Heyduk et al. 2016 (https://github.com/kheyduk/reads2trees) is less streamlined than phyluce but allows for more customizable parameters throughout the bioinformatic pipeline. Together these programs and pipelines are achieving the same goal with very similar methodological steps (Fig. 3.3). First, all raw reads must be cleaned by removing Illumina adapters and trimming reads with poor quality scores. These clean reads are then used for assembly, which can either be reference based or *de novo*. Users can assemble reads through both routes and then merge similar sequences or opt to use one type of assembly program. The resulting assembled contigs can then be matched via local alignment searches (e.g., BLAST or LASTZ) against the initial targets and retained for further analyses. Contigs that match the target areas should be sorted into loci (e.g., by merging exons from the same gene), aligned, and trimmed prior to downstream analyses. A second round of duplicate removal may be necessary, depending on the target loci, because paralogous sequences may be captured or make it through as nontarget data that were not in the initial reference used for bait design.

We have seen a dramatic increase in the amount of data that can be collected using recent genomic techniques, and this trend is likely to increase as sequencing costs continue to decrease. The bottleneck with handling high-throughput data generally arises from the computational time required for their analysis and from our current understanding of phylogenomics and population dynamics. Historically, phylogeneticists would concatenate genes to estimate the species tree, but both

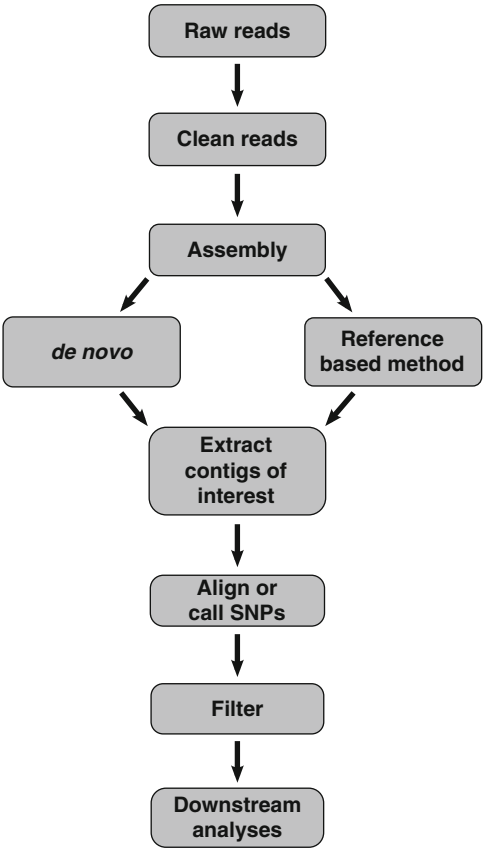

Fig. 3.3 Overview of a bioinformatic pipeline for re-sequencing data. Programs for each step should be determined based on assumptions regarding data and downstream analyses. Assembly can be conducted using multiple programs, or a single optimal assembly method can be implemented

empirical and theoretical data suggest that this is not always a robust method. Specifically, it has been known for some time that gene trees can have different histories from each other and from the species tree. Gene tree discordance can impact phylogenetic analyses, and modeling the processes that lead to discordance (i.e., incomplete lineage sorting [ILS], recombination, hybridization, etc.) has been challenging. To date the majority of phylogenetic programs can only estimate species trees when accounting for ILS. Programs are emerging to model the process of hybridization (STEM-hy—Kubatko 2009; PhyloNet—Yu et al. 2011; Yu and Nakhleh 2015), and, in general, the analysis of multilocus data is rapidly developing, making it hard for newcomers to find appropriate programs for analyzing their data. Care should also be taken to consider the biology of your taxa of interest. Therefore, we recommend that researchers consider the programs and the underlying models they are most likely going to be implementing given their system. For example, understanding the phylogenetic relationships of a recent or rapid radiation will most likely involve high levels of ILS and possibly hybridization. In this example, it may be worthwhile to sequence multiple individuals per species to increase

the accuracy of parameter estimation for the coalescent models (Heled and Drummond 2010), but not all programs are capable of taking into account multiple individuals per species (Table 3.3). In addition, some programs may take an exceedingly long time (or fail) to run depending on the number of loci and number of taxa input (Table 3.3). Computational biologists are developing new ways to reduce the size and complexity of datasets for phylogenetic analyses (e.g., Bayzid and Warnow 2013), though these methods should be carefully evaluated on individual projects to assess their suitability.

3.3.3 Post-sequencing for Population Genomic Designs

Many of the difficulties described above for phylogenetic analyses hold true for population genomic analyses, as well. Pipelines for analyzing target enrichment data collected at the population level are generally lacking (but see Faircloth 2016; https://github.com/mgharvey/seqcap_pop). With a bit of legwork, one can identify genomic features of interest, including SNP and indel calls and use these data to estimate heterozygosity, FST, Tajima's D, and others, using the bcftools (https://github.com/samtools/bcftools) command line program (among others). The program requires reads to be mapped to some sort of assembly or reference genome, and it extracts and analyzes relevant information from those mappings. Note, however, that the estimates of population genomic statistics through bcftools are only as good as the reads and reference contigs that are used in mapping; duplicated loci of any kind could allow for a read to map to multiple locations and create false allele calls and erroneous estimates. Low-coverage contigs are particularly problematic because they may contain erroneous homozygous SNP calls.

3.3.4 Computational Resource Requirements

Although it is possible to run most of the individual programs on desktop computers, parallel compute clusters are highly recommended or necessary to process the data in a timely and efficient manner. For projects that have an especially large number of individuals that need to have sequence data assembled *de novo*, parallelization will greatly increase the speed at which assemblies can be completed. Similarly, for many loci, performing many calculations across all loci will be untenable without the help of parallel computing. In addition to large clusters housed at universities and research centers, researchers interested in attempting large-scale analyses can use third-party computing such as CyVerse (http://www.cyverse.org/), Amazon (www.amazon.com/hpc), and XSEDE (https://www.xsede.org/home). While parallelization greatly reduces time spent on the bioinformatics side of target enrichment, researchers should note the memory requirements for a number of programs. For example, Trinity (Grabherr et al. 2011) recommends 1 Gb of RAM per

every 1 M reads; RAxML requires ~2.8 Gb for a 100 kb alignment of 50 taxa (http://www.exelixis-lab.org/software.html). Perhaps most important for consideration is the sheer size of storage space required to store raw reads, cleaned reads, assemblies, and various intermediate files that are produced during analysis. Projects with many individuals and loci can quickly use a terabyte of hard-drive space.

3.4 Results Reporting and Community Resources

3.4.1 Standards of Reporting

Sequence capture methods, no matter how baits are designed, are fundamentally similar in their attempt to reduce genomic representation in the sequenced reads. As a result, similar statistics are important for assessing the quality and efficiency of sequence capture. For example, the number of on-target contigs assembled per library, relative to how many were targeted, gives a general impression of how well hybridization worked, although this metric is slightly confounded by sequencing depth, which alone can increase the number of assembled contigs. Coverage statistics—both for assembled contigs from targeted regions and off-target regions and perhaps for exon and intron sequences separately (see Heyduk et al. 2016)—indicate whether the depth of sequencing was adequate to call polymorphisms and whether hybridization of certain baits was more efficient than others, perhaps due to sequence similarity or genomic copy variation. For studies that attempt to capture loci from taxa across broader phylogenetic distance, assessing hybridization variation in baits across taxa helps to define the phylogenetic boundary of effective capture using a particular bait set. In addition, it is often important to know how efficient capture was across the entire library—in other words, researchers might be interested in how many reads were on target or how many reads map to contigs used in the final analyses. Consistent reporting of such metrics enables comparisons of various methods and techniques across different sampling schemes and bait designs, leading to informed decision-making by researchers looking to implement sequence capture methods.

While numerical information about a given sequence capture project is useful for those looking to replicate methodology, the raw and cleaned data generated can be used by the larger scientific community as a whole. For this reason, researchers should take special care to deposit raw reads, alignments, and downstream analyses into common repositories (e.g., NCBI's Short Read Archive (http://www.ncbi.nlm.nih.gov/sra) and Dryad (http://datadryad.org/)). The bait sequences should be shared after publication as well. The time and effort put into designing effective and informative baits should be stretched beyond a single project. Indeed, some bait sets have sufficient utility that commercial companies may synthesize them in bulk, making them available to the research community at far lower cost than custom kits (http://www.mycroarray.com/mybaits/mybaits-UCEs.html).

Annex: Quick Reference Guide

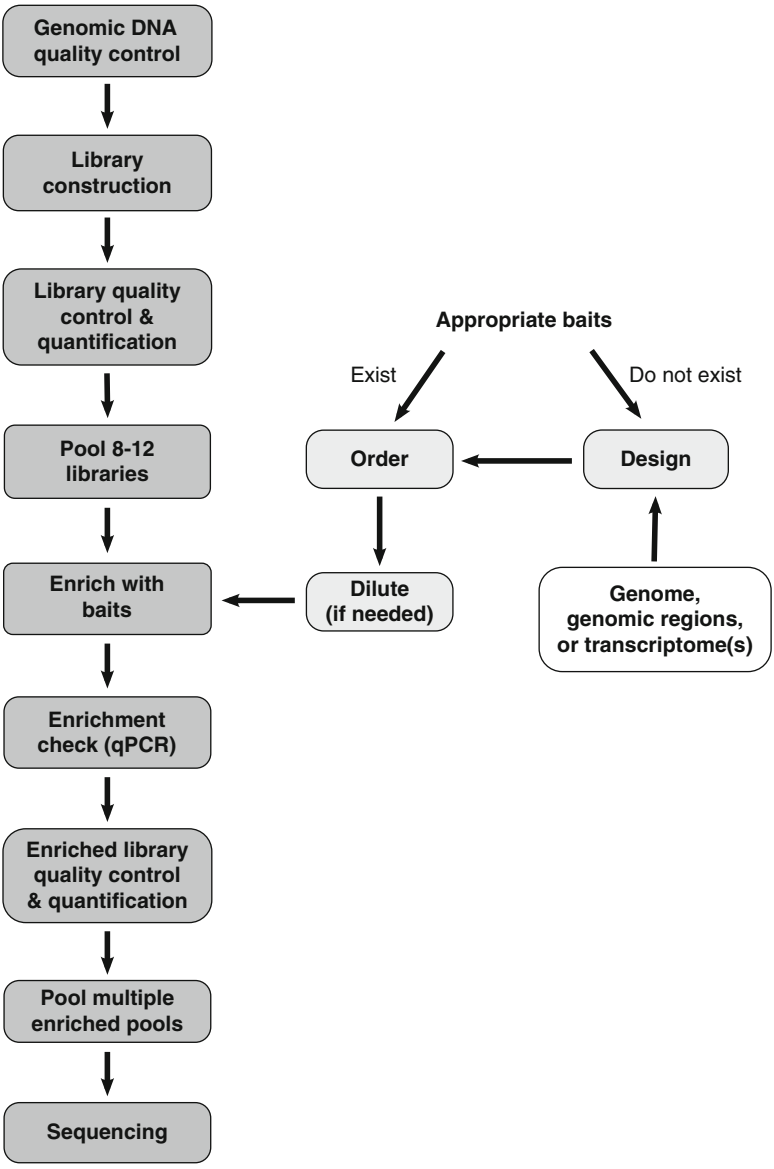

Fig. QG3.1 Representation of the wet-lab procedure workflow

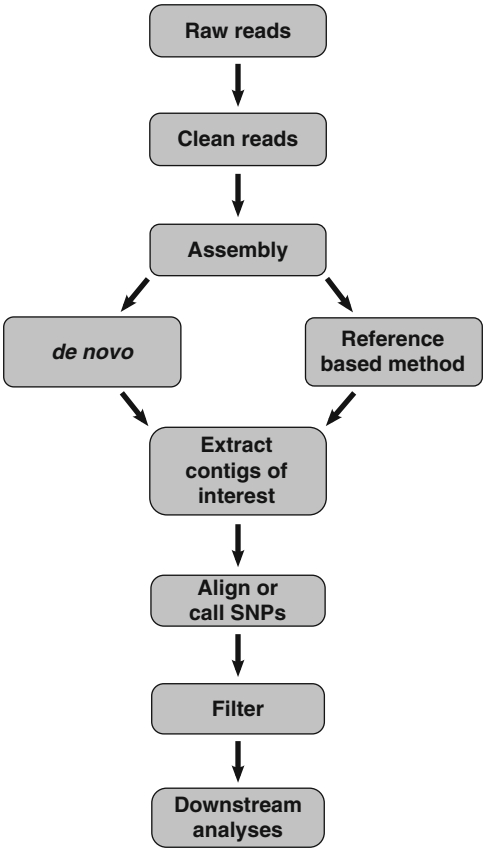

Fig. QG3.2 Main steps of the computational analysis pipeline

Table QG3.1 Experimental design considerations

Technique	Recommended starting material (ng)	Constraints	Average sequencing depth	Recommended sequencing run	Reference(s)
Whole genome re-sequencing	500	Cost of sequencing	6–30×	HiSeq or NextSeq PE75–PE150	Sims et al. (2014), Ekblom and Wolf (2014)
Transcriptome sequencing	1000	Sample material, cost of libraries, cost of sequencing	≥10 million reads per sample	HiSeq or NextSeq PE100–PE150	Ozsolak and Milos (2011), Wang et al. (2009), Wang et al. (2011)
PCR amplicon sequencing	20	Combinatorial tags, pool with diverse libraries	20×	MiSeq PE250–PE300	Feng et al. (2016), Mamanova et al. (2010)
Restriction-site-associated DNA markers (RADseq)	100	Consistency	10×	HiSeq or NextSeq SE75–PE150	Davey et al. (2011), Davey et al. (2013)
Target enrichment by sequence capture	500	Probes (information to design and cost)	30×	HiSeq or NextSeq PE100–PE150	Mamanova et al. (2010), Mertes et al. (2011)

Table that comprises relevant experimental design parameters, to carefully consider before applying this methodology

Table QG3.2 Available software recommendations

Program	Phylogenetic method	Multiple accessions	Missing data	Requirements	Command line vs. GUI	Computational time
BEST (Liu 2008)	Sequence alignment	Can assign accessions to species	Accepts missing data	None	Command line	Very slow (months)
*BEAST (Heled and Drummond 2010)	Sequence alignment	Multiple accessions recommended	Does not allow missing data for a given species	Priors	GUI	Very slow (months)
SVDquartets (Chifman and Kubatko 2014)	Sequence alignment	Cannot assign accessions to species	Accepts missing data	None	Command line	Very fast (hours to days)
STEM (Kubatko et al. 2009)	Summary method	Cannot assign accessions to species	Accepts missing data	Must have an estimate of individual gene evolution relative to each other	Command line	Fast (days)
MP-EST (Liu et al. 2010)	Summary method	Can assign accessions to species	Accepts missing data	All gene trees must be rooted	Both	Fast (days)
ASTRAL (Mirarab et al. 2014)	Summary method	Cannot assign accessions to species	Accepts missing data	Gene trees need to be fully resolved (can be unrooted)	Command line	Very fast (hours)

Table displaying a selection of the recommended software available for the computational analysis of data yielded by this technique

References

Bao S, Jiang R, Kwan WK, Wang BB, Ma X, Song YQ (2011) Evaluation of next-generation sequencing software in mapping and assembly. J Hum Genet 56:406–414

Bayzid MD, Warnow T (2013) Naïve binning improves phylogenomic analyses. Bioinformatics 29:2277–2284

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

Blumenstiel B, Cibulskis K, Fisher S, DeFelice M, Barry A et al. (2010) Targeted exon sequencing by in-solution hybrid selection. Curr Protoc Hum Genet Chapter 18: Unit 18.4.

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

Carpenter ML, Buenrostro JD, Valdiosera C et al (2013) Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. Am J Hum Genet 93:852–864

Catchen J, Hohenlohe P, Bassham S, Amores A, Cresko W (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124–3140

Chifman J, Kubatko L (2014) Quartet inference from SNP data under the coalescent model. Bioinformatics 30:3317. doi:10.1093/bioinformatics/btu530

Comer JR, Zomlefer WB, Barrett CF, Davis JL, Stevenson DW, Heyduk K, Leebens-Mack J (2015) Resolving relationships within the palm subfamily Arecoideae (Arecaceae) using plastid sequences derived from next-generation sequencing. Am J Bot 102:888–899

Cummings N, King R, Rickers A, Kaspi A, Lunke S, Haviv I, Jowett JBM (2010) Combining target enrichment with barcode multiplexing for high throughput SNP discovery. BMC Genomics 11:641

Davey JW, Blaxter ML (2010) RADSeq: next-generation population genetics. Brief Funct Genomics 9:416–423

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

Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML (2013) Special features of RAD Sequencing data: implications for genotyping. Mol Ecol 22:3151–3164

Dermitzakis ET, Reymond A, Antonarakis SE (2005) Conserved non-genic sequences—an unexpected feature of mammalian genomes. Nat Rev Genet 6:151–157

Derti A, Roth FP, Church GM, Wu C-T (2006) Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. Nat Genet 38:1216–1220

Duarte JM, Wall PK, Edger PP, Landherr LL, Ma H, Pires JC, Leebens-Mack J, dePamphilis CW (2010) Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis*, and *Oryza* and their phylogenetic utility across various taxonomic levels. BMC Evol Biol 10:61

Easton DF, Rharoah PDP, Antoniou AC et al (2015) Gene-panel sequencing and the prediction of breast-cancer risk. N Engl J Med 372:2243–2257

Eaton DAR (2014) PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. Bioinformatics 30:1844. doi:10.1093/bioinformatics/btu121

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107:1–15

Ekblom R, Wolf JBW (2014) A field guide to whole-genome sequencing, assembly, and annotation. Evol Appl 7(9):1026–1042

Enk JM, Devault AM, Kuch M, Murgha YE, Rouillard JM, Poinar HN (2014) Ancient whole genome enrichment using baits built from modern DNA. Mol Biol Evol 31:1292–1294

Faircloth BC (2016) PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics 32:786-788. doi:10.1093/bioinformatics/btv646

Faircloth BC, Glenn TC (2012) Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. PLoS One 7:e42543. doi:10.1371/journal.pone.0042543

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 61:717–726

Faircloth BC, Branstetter MG, White ND, Brady SG (2015) Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol Ecol Resour 15:489

Feng YJ, Liu QF, Chen MY, Liang D, Zhang P (2016) Parallel tagged amplicon sequencing of relatively long PCR products using the Illumina HiSeq platform and transcriptome assembly. Mol Ecol Resour 16:91. doi:10.1111/1755-0998.12429

Fisher S, Barry A, Abreu J, Minie B, Nolan J et al (2011) A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol 12:R1

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhue C, Pudlo P, Cornuet JM, Estoup A (2012) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol Ecol 22:3165–3178

Glenn TC, Nilsen R, Kieran TJ, Finger JW Jr, Pierson TW, García-De-Leon FJ, del Rio Portilla MA, Reed K, Anderson JL, Meece JK, Alabady M, Belanger M, Faircloth BC (2016) Adapterama I: universal stubs and primers for thousands of dual-indexed Illumina Nextera and TruSeqHT compatible libraries (iNext & iTru). bioRxiv

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652

Grover CE, Salmon A, Wendel JF (2011) Targeted sequence capture as a powerful tool for evolutionary analysis. Am J Bot 99(2):312–319

Haas BJ, Gevers D, Earl AM et al (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res 21:494–504

Hahn DA, Ragland GJ, Shoemaker DD, Denlinger DL (2009) Gene discovery using massively parallel pyrosequencing to develop ESTs for the fleshy fly *Sarcophaga crassipalpis*. BMC Genomics 10:234. doi:10.1186/1471-2164-10-234

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

Heyduk K, Trapnell DW, Barnett CF, Leebens-Mack J (2016) Estimating relationships within *Sabal* (Arecaceae) through multilocus analyses of sequence capture data. Biol J Linnean Soc 17(1):106–120

Huang H, Knowles LL (2014) Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. *Syst Biol* doi: 10.1093/sysbio/syu046

Keane TM, Goodstadt L, Danecek P, White MA, Wong K et al (2011) Mouse genome variation and its effect on phenotypes and gene regulation. Nature 477:289–294

Kubatko LS (2009) Identifying hybridization events in the presence of coalescence via model selection. Syst Biol 58:478–488

Kubatko LS, Carstens BC, Knowles LL (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973

Lemmon EM, Lemmon AR (2013) High-throughput genomic data in systematics and phylogenetics. Annu Rev Ecol Evol Syst 44:99–121

Li Y, Zhao S, Ma J, Li D, Yan L, Li J, Qi X, Guo X et al (2013) Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. BMC Genomics 14:579. doi:10.1186/1471-2164-14-579

Liu L (2008) BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24:2542–2543

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. doi:10.1186/1471-2148-10-302

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotype to genome typing. Nat Rev Genet 4:981–994. doi:10.1038/nrg1226

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH et al (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7:111–118

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

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013a) Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol 66:526–538

McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT (2013b) A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS One 8:e54848. doi:10.1371/journal.pone.0054848

McCormack JE, Tsai WLE, Faircloth BC (2015) Sequence capture of ultraconserved elements from bird museum specimens. Molecular Ecology Resources doi: 10.1111/1755-0998.12466

Meiklejohn KA, Danielson MJ, Faircloth BC, Glenn TC, Braun EL, Kimball RT (2014) Incongruence among different mitochondrial regions: a case study using complete mitogenomes. Mol Phylogenet Evol 78:314–323

Mertes F, ElSharawy A, Sauer S, van Helvoort JMLM, van der Zaag PJ, Franke A, Nilsson M, Lehrach H, Brookes AJ (2011) Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief Funct Genomics 10(6):374–386

Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc 2010: pdb prot5448

Mirarab S, Reaz R, Bayzid MS, Zimmerman T, Swenson MS, Warnow T (2014) ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548

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

Peñalba JV, Smith LL, Tonione MA, Sass C, Hykin SM, Skipwith PL, McGuire JA, Bowie RCK, Moritz C (2014) Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms. Mol Ecol 14(5):1000–1010

Puritz JB, Matz MV, Toonen RJ, Weber JN, Bolnick DI, Bird CE (2014) Demystifying the RAD fad. Mol Ecol 23(24):5937–5942

Raposo do Ameral F, Neves LG, Resende MF Jr, Mobili F, Miyaki CY, Pellegrino KC, Biondo C (2015) Ultraconserved elements sequencing as a lowcost source of complete mitochondrial genomes and microsatellite markers in non-model amniotes. PLoS One 10:e0138446

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res 22:939–946

Rubin BER, Ree RH, Moreau CS (2012) Inferring phylogenies from RAD sequence data. PLoS One 7:1–12

Shearer EA, Hildebrand MS, Ravi H, Joshi S, Guiffre AC, Novak B, Happe S, LeProust EM, Smith RJH (2012) Pre-capture multiplexing improves efficiency and cost-effectiveness of targeted genomic enrichment. BMC Genomics 13:618

Sims D, Sudbery I, Ilot NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15:121–132

Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT (2014) Target capture and massively parallel sequencing of ultraconserved elements (UCEs) for comparative studies at shallow evolutionary time scales. Syst Biol 63(1):83–95

Stephen S, Pheasant M, Makunin IV, Mattick JS (2008) Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol Biol Evol 25:402–408

Stephens JD, Rogers WL, Heyduk K, Cruse-Sanders JM, Determann RO, Glenn TC, Malmberg RL (2015a) Resolving phylogenetic relationships for the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. Mol Phylogenet Evol 85:76–87

Stephens JD, Rogers WL, Mason CM, Donovan LA, Malmberg RL (2015b) Species tree estimation of diploid *Helianthus* (Asteraceae) using target enrichment. Am J Bot 102:921–941

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

Wang Y, Qian PY (2009) Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. PLoS One 4:e7401. doi:10.1371/journal.pone.0007401

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

Wang Y, Ghaffari N, Johnson CD, Braga-Neto UM, Wang H, Chen R, Zhou H (2011) Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. BMC Bioinformatics 12:S5. doi:10.1186/1471-2105-12-S10-S5

Weitmeier K, Straub SCK, Cronn RC, Fishbein M, Schmickl R, McDonnell A, Liston A (2014) Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. Appl Plant Sci 2:1400042. doi:10.3732/apps.1400042

Xu J, Zhao Q, Du P, Xu C, Wang B, Feng Q, Liu Q, Tang S, Gu M, Han B, Liang G (2010) Developing high throughput genotyped chromosome segment substitution lines based on population whole-genome re-sequencing in rice (*Oryza sativa* L.). BMC Genomics 11:656. doi:10.1186/1471-2164-11-656

Yu Y, Nakhleh L (2015) A distance-based method for inferring phylogenetic networks in the presence of incomplete lineage sorting. Bioinform Res Appl 9096:378–389

Yu Y, Cuong T, Degnan JH, Nakhleh L (2011) Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Syst Biol 60:138–149

Zhu Y, Bergland AO, González J, Petrov DA (2012) Empirical validation of pooled whole genome population re-sequencing in *Drosophila melanogaster*. PLoS One 7:e41901. doi:10.1371/journal pone.0041901