CHAPTER NINE

Future of Phylogenomic Data Acquisition

Prior to the arrival of molecular biology kits and outsourcing of Sanger sequencing, researchers had to be skilled in all aspects of phylogenomic data acquisition—from DNA extractions to preparing sequencing gels for an automated sequencing machine. This arduous and low throughput workflow meant that it was not unusual to wait many months or even years before the desired dataset was in hand. Thus, the welcome arrival of outsourcing for the Sanger sequencing step greatly simplified the process of obtaining DNA sequence data. Outsourcing is having a similar effect on NGS. During the early years of NGS, researchers needed to prepare their own sequencing libraries before outsourcing the sequencing step (e.g., "buying a lane" on an Illumina sequencer). Now, many NGS facilities offer to perform library preparations—for nontarget partial genome sequencing, target capture (e.g., UCE-anchored loci), or whole genome sequencing—as well as the final sequencing step. Thus, researchers who opt to outsource all NGS work only need to perform DNA extractions in order to generate enormous datasets. Whether researchers prepare their own libraries or not does not change the fact that it now only takes weeks to obtain large datasets via the NGS route. By spending less time in the laboratory obtaining data, researchers can invest more time in the analysis of their datasets. Indeed, the long-term trend in phylogenomics has been a reduction in time for amassing sequence data, which has had the favorable consequence of allowing researchers more time to conduct analyses. Can this trend be extended further? Given the fast pace of genomics and biotechnology, we should ponder what the future holds for DNA sequence data acquisition.

9.1 THE IMPENDING FLOOD OF GENOMES

Jarvis et al. (2014) published a massive phylogenomic study of birds based on 48 assembled and annotated genomes that spanned the avian tree of life. This was hailed as a landmark achievement in tree of life studies (Callaway 2014; Zhang et al. 2014). However, this feat required many collaborators from different laboratories, access to substantial funding, and years of hard work. Despite the advent of NGS technologies more than a decade ago, obtaining large genomes such as those for vertebrates still remains a difficult and expensive task. That only 259 vertebrate genomes had been sequenced and assembled up to and including the year 2014 (OBrien et al. 2014) bespeaks the challenges of acquiring whole genome sequences.

For a phylogenomic study, the current high costs of genomes means that a tradeoff exists between the numbers of genomes that can be sequenced versus numbers of species that can be sampled. Prum et al. (2015) conducted a large-scale study of the avian tree of life with this tradeoff in mind. Instead of acquiring whole genome sequences for each sampled species—as was done by Jarvis et al. (2014), these authors used a target capture method to obtain 259 anchored loci from 198 avian species distributed throughout the avian tree of life. The main advantage of the approach advocated by Prum et al. (2015) is that it allows for more extensive taxon sampling, which is expected to improve the accuracy of gene trees (Graybeal 1998; Heath et al. 2008; Townsend and Lopez-Giraldez 2010). Thus, subsampling genomes for many loci using target capture methods is currently the best (or

only) strategy for phylogenomic studies involving many species. However, at some point in the future the cost of obtaining full genomes may fall low enough to trigger a major change in how DNA sequence datasets are acquired.

How much does a new genome sequence cost? In early 2014, Illumina announced that it was making a sequencing platform called the HiSeq X Ten that is expected to sequence human genomes for under 1,000 USD (Hayden 2014). This cost estimate is being made possible by improvements to Illumina's existing NGS technology. There has also been some speculation that the cost of outsourcing a genome (for any species) will soon be in the low thousands of dollars (Hayden 2014). While it is impossible to know how low the cost of full genome sequencing will ultimately reach, it is not far-fetched to believe that the cost will eventually plunge down into the low hundreds of dollars if not lower. Human medicine is a powerful driver of genomics and biotechnology and thus the cost of obtaining full genome sequences will likely continue to fall as a result of competition among sequencing technologies.

The appearance of a novel genome sequencing platform, which has low labor and reagent costs combined with software that automatically assembles genomes, could make the $100–$400 genome a reality at some point in coming years. In this scenario, a researcher could send dozens or more of purified DNA samples to a genome sequencing facility where all sequencing and genome assemblies are performed. Later, perhaps within days or 1–2 weeks, the researcher would be able to download all genome data files and begin phylogenomic analyses. Even if this futuristic scenario does not become reality in the near future, a large number of vertebrate genomes may still soon be available. OBrien et al. (2014) predicted that the number of sequenced vertebrate genomes will grow to more than 10,000 within 5–10 years. This estimate was based on assumptions that existing NGS technologies and bioinformatics tools for assembling genomes may see modest improvements and therefore it does not take into consideration what might happen if a new and higher performing NGS platform becomes available. This impending flood of genomes is going to vigorously stimulate a large number of new full genome-based studies of the vertebrate tree of life, which, in turn, will enable *in silico* phylogenomics to come of age (Costa et al. 2016).

9.2 *IN SILICO* ACQUISITION OF PHYLOGENOMIC DATASETS

As we saw in Chapters 7 and 8, available genome sequences are enabling researchers to easily design large numbers of phylogenomic loci. Once sets of target capture probes or PCR primers are designed, laboratory methods are required to obtain DNA sequence data from the genomes of all sampled individuals. Although this hybrid *in silico*-laboratory workflow for obtaining datasets represents a quantum-leap improvement over earlier methods, the ultimate approach to acquiring a dataset is to use 100% *in silico* methods whereby all data are directly obtained from computer files containing complete genome sequences. This approach, however, requires complete genome sequences for each individual or species in the study, which explains why there have been few such studies to date.

Rokas et al. (2003) conducted a completely *in silico* phylogenomic study involving eight species of yeasts whose genomes had already been sequenced. Using only a computer, these authors extracted DNA sequences representing 106 orthologous loci from each of the genomes before performing analyses of yeast phylogeny. In similar fashion, Faircloth et al. (2012) and McCormack et al. (2012) took advantage of existing full genome data in order to generate large datasets consisting of UCE-anchored loci for mammals. Jarvis et al. (2014) also used computer-based methods to construct datasets based on UCE-anchored loci but their data were derived from genomes that they themselves had earlier sequenced and assembled. More recently, Costa et al. (2016) conducted a fully *in silico*-based study of hominoid evolution, which was based on a large number of anonymous and AE-anchor loci extracted from available genome sequences.

The 100% *in silico* approach to data acquisition offers several advantages over other methods. First, given complete genome data for each individual or species in a study, the amount of time required to construct multiple loci datasets is on the order of days or less. A second advantage, at least over some of the alternative methods, is that selected loci can be verified as being single-copy in genomes. A third advantage is that *a priori* physical distance thresholds between loci located on the same chromosomes can be used to ensure that each locus is genealogically independent

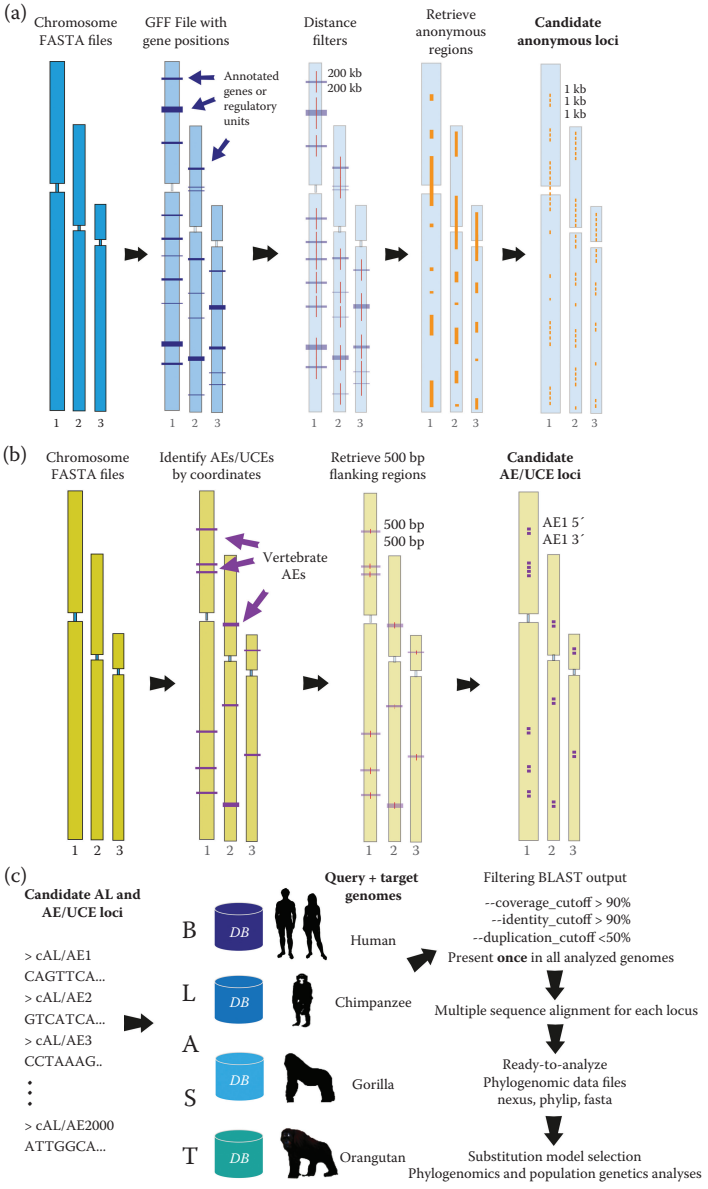

Figure 9.1 A software pipeline for acquiring ready-to-analyze phylogenomic datasets consisting of anonymous or anchor (AE/UCE) loci (Costa et al. 2016, fig. 1). The legend reads, "ALFIE software pipeline. (a) Anonymous loci (AL) finding module: User inputs complete genome sequences in a FASTA format and a general feature format (GFF) file for the query genome. Program first applies a user-defined 'distance filter,' which removes all known functional elements + flanking sequences of user-specified lengths (purple color blocks). Remaining (presumably neutral) intergenic regions (orange color blocks), called candidate ALs, are retrieved and cut into consecutive segments of user-defined length and saved in FASTA files. (b) Anchor loci (AE/UCE) finding module: User inputs genome sequences in FASTA format. Program finds locations of target AEs/UCEs in a reference human genome with a coordinate file that currently contains 512 vertebrate AEs (included in package). Module retrieves flanking regions with user-defined length (e.g., 500 bp). User also specifies distance (in base pairs) between flanking sequences and their AEs/UCEs. Paired flanking sequences (i.e., candidate AE/UCE loci) are saved in FASTA files. (c) Downstream analyses: AL or AE/UCE candidate loci are used as query sequences in BLAST searches against target genomes. Single-copy loci are retained and subsequently aligned. A user-specified distance filter retains loci that are likely independent from other sampled loci. Each pair of AE/UCE flanking sequences is concatenated to form independent loci. Lastly, ALFIE outputs ready-to-analyze data sets." (Reprinted from Costa, I. R., F. Prosdocimi, and W. B. Jennings. 2016. *Genome Res* 26:1257–1267.)

from other sampled loci (Sachidanandam et al. 2001; O'Neill et al. 2013; Leaché et al. 2015; Costa et al. 2016; see Table 3.1 for additional examples). Finally, the computer-based process of generating DNA sequence datasets can be automated from start to finish: software can be used to extract target loci sequences from all sampled genomes, perform multiple sequence alignments, and output ready-to-analyze datasets.

The recent work by Costa et al. (2016) provides us with an example of how phylogenomic data acquisition can be fully automated. In this study, the authors developed a software pipeline called *ALFIE*, which is for Anonymous/Anchor Loci FIndEr (Figure 9.1). After the user inputs full genome data for each individual or species and specifies values for various distance filters (or thresholds), desired locus length, etc., the program seamlessly performs the following steps: (1) extracts the maximum number of single-copy, presumably neutral, and independent anonymous loci from an annotated genome (Figure 9.1a) or obtains a set of predefined AE or UCE loci from an unannotated genome sequence (Figure 9.1b); (2) extracts orthologous sequences for all anonymous or AE/UCE loci from other complete genomes; (3) constructs multiple sequence alignments for each locus; and (4) outputs ready-to-analyze datasets in various commonly used formats such as NEXUS, PHYLIP, and FASTA (Figure 9.1c). These authors bench tested this software using complete genome data from the well-studied extant hominoids (humans, chimpanzees, gorillas, and orangutans). In less than 3 hours, this software output a 1.2 Mb-sized dataset consisting of 292 anonymous loci (each ~1 kb long) while only 13 minutes was needed to output a dataset consisting of 242 AE loci of similar lengths. Although this type of study can only be done using fully sequenced genomes for all study individuals, this example shows the exciting full potential of *in silico* data acquisition.

REFERENCES

Callaway, E. 2014. Flock of geneticists redraws bird family tree. *Nature* 516:297.

Costa, I. R., F. Prosdocimi, and W. B. Jennings. 2016. In silico phylogenomics using complete genomes: A case study on the evolution of hominoids. *Genome Res* 26:1257–1267.

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

Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst Biol* 47:9–17.

Hayden, E. C. 2014. Is the $1,000 genome for real? *Nature*. doi: 10.1038/nature.2014.14530.

Heath, T. A., S. M. Hedtke, and D. M. Hillis. 2008. Taxon sampling and the accuracy of phylogenetic analyses. *J Syst Evol* 46:239–257.

Jarvis, E. D., S. Mirarab, A. J. Aberer et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346:1320–1331.

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

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

OBrien, S. J., D. Haussler, and O. Ryder. 2014. The birds of Genome 10K. *GigaScience* 3:32.

O'Neill, E. M., R. Schwartz, C. T. Bullock et al. 2013. Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Mol Ecol* 22:111–129.

Prum, R. O., J. S. Berv, A. Dornburg et al. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526:569.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804.

Sachidanandam, R., D. Weissman, S. C. Schmidt et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–933.

Townsend, J. P. and F. Lopez-Giraldez. 2010. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. *Syst Biol*. doi: 10.1093/sysbio/syq025.

Zhang, G., E. D. Jarvis, and M. T. P. Gilbert. 2014. A flock of genomes. *Science* 346:1308–1309.