CHAPTER ONE

Introduction

The great evolutionary geneticist, Theodosius Dobzhansky, famously wrote "*Nothing in biology makes sense except in the light of evolution*" (Dobzhansky 1973). One area in evolutionary biology that has shed much light on biological phenomena is the field of molecular phylogenetics. Phylogenetic trees inferred from molecular genetic data have led to quantum leaps in our understanding about molecular evolution and the *Tree of Life*. The Tree of Life Project is a worldwide collaboration of evolutionary biologists that aims to elucidate the evolutionary history for all life found on Earth (Maddison et al. 2007; http://tolweb.org/tree/). Another important initiative is the Open Tree of Life (http://opentreeoflife.org/). Advances in DNA sequencing capability, computers, and bioinformatics from the late 1970s through the 1990s spurred the rapid growth of molecular phylogenetics (Hillis et al. 1996). An outgrowth of this field, which began slowly in the 1990s but later blossomed into its own field due to the emergence and explosive growth of genomics, is the discipline of *phylogenomics*. Given that substantial overlap obviously exists between molecular phylogenetics and phylogenomics, we should ask the following questions: *What is phylogenomics and how does it differ from molecular phylogenetics?*

1.1 WHAT IS PHYLOGENOMICS?

Before we further consider a definition for phylogenomics, let's first examine a traditional definition of molecular phylogenetics. The field of molecular phylogenetics can be defined as follows: *molecular phylogenetics is the discipline concerned with using phylogenetic methodology on molecular genetic data to infer evolutionary phylogenies or "trees" to elucidate the evolutionary relationships and distances or divergence times among DNA sequences, amino acid sequences, populations, species, or higher taxa*. The vast majority of molecular phylogenetic studies have been based on DNA sequence data, typically representing one to several genes, though other types of molecular genetic data such as amino acid sequences are also used. Molecular phylogenies, especially those inferred for a single gene, are commonly called *gene trees*.

In contrast to molecular phylogenetics, "phylogenomics" is more difficult to define for two reasons. First, some methodological and conceptual overlap exists between them—namely both fields rely on phylogenetic methodology to infer phylogenies from molecular data. Secondly, researchers have used the term phylogenomics to characterize different types of studies. We will now take a closer look at how researchers have used the term phylogenomics before we settle on a definition to follow in this book.

1.1.1 The Early View of Phylogenomics

Eisen (1998a) originally coined the term phylogenomics and defined this discipline as the prediction of gene function and study of gene and genome evolution using molecular phylogenies in conjunction with modern comparative methods. For example, in an early phylogenomic study, Eisen (1998b) first inferred the gene tree among members (amino acid sequences) of the *MutS* family of proteins, a group of proteins important for recognition and repair of DNA mismatches caused by errors during DNA replication. He then used this tree to investigate the evolutionary diversification of this gene family by looking at *MutS* homologs found within and among genomes across the Tree of Life. Phylogenetic-based methods are not only superior to similarity-based methods for

predicting the functions of unknown genes, but they allow a researcher to split genes into orthologous and paralogous subfamilies and identify key events in the histories of gene families such as gene divergences, lateral gene transfers, and gene losses (Eisen et al. 1995; Eisen 1998a,b).

Shortly thereafter, O'Brien and Stanyon (1999) used the term phylogenomics differently, as they mentioned "comparative phylogenomics" to describe studies using comparative gene maps for a number of closely related species combined with cladistic analysis to reconstruct ancestral genomes (e.g., Haig 1999). Although these two uses of the term phylogenomics were both applied to the study of molecular or genomic evolution, these studies nonetheless differed from each other in terms of data, analytical methods, and study goals.

1.1.2 An Expanded View of Phylogenomics

The purview of phylogenomics broadened further during the early 2000s soon after the genome era commenced, as the rapidly increasing volumes of genomic data—including some fully sequenced eukaryotic genomes such as the human genome—allowed researchers to dramatically scale up sizes of their datasets in phylogenetically based evolutionary studies. It was during this time that researchers could begin using phylogenetic methodology to analyze enormous genome-wide datasets for addressing problems ranging from genome evolution to reconstructing the Tree of Life (Eisen and Fraser 2003; Rokas et al. 2003; Delsuc et al. 2005; Philippe et al. 2005).

The viewpoint that phylogenomics is a discipline comprised of two main areas of inquiry—one concerned with questions in molecular and genomic evolution and the other focused on the evolutionary history of organisms—was subsequently reinforced at the first phylogenomics symposium (Philippe and Blanchette 2007) and in the first book focused on phylogenomics (Murphy 2008). Aside from the dramatic growth in phylogenomic studies since that time, little has changed regarding this dichotomy of research goals. Thus, at the present time we can think of phylogenomics as comprising two major subdisciplines: *molecular phylogenomics* and *organismal phylogenomics*. Accordingly, we may broadly define "phylogenomics" as *the field of study concerned with using genome-wide data to infer the evolution of genes, genomes, and the Tree of Life*. What primarily differentiates molecular phylogenetics from phylogenomics is that the latter field often uses much larger or "computer bursting" datasets and gene trees as independent units of analyses in evolutionary studies.

1.2 ANATOMY OF GENE TREES

Gene trees that have been inferred from DNA sequence data represent fundamental units of analysis in various types of phylogenomic studies. The basic anatomy of a gene tree is illustrated in Figure 1.1. In this figure we see the genealogical relationships among a sample of five DNA sequences labeled a through e for a single gene. Except for some specialized cases (e.g., studying microorganisms in a laboratory), the *true* gene tree cannot be known for a given set of DNA sequences. Instead, the genealogical history must be *inferred* or reconstructed using phylogenetic methodology.

The structure of an inferred gene tree is defined by its branching pattern or "topology" and length of each branch. For example, the tree in Figure 1.1 has four *nodes* (labeled 1 through 4). The bottom-most node (node 4) represents the *root* of the gene tree. The root node is particularly important because it represents the *most recent common ancestor* or "MRCA" of the five DNA sequences and therefore provides directionality or time's arrow along the tree (Figure 1.1). Similarly, we can describe node 1 as the MRCA of a and b, node 2 is the MRCA of a, b, and c, and node 3 is the MRCA of d and e. The other major structural features of gene trees are its *branches*, which represent *lines of descent*.

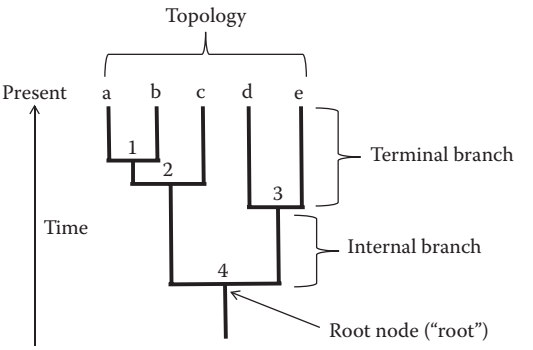

Figure 1.1. An example of a rooted gene tree for five DNA sequences labeled (a–e). In this tree there are four nodes, five terminal branches, and four internal branches. The placement of the root in this tree (node 4) gives the tree a direction with respect to time and thus ancestor–descendant relationships can be inferred.

Branches (i.e., the vertical lines in the tree) can be subdivided into two categories: *terminal branches*, which connect the tips (observed DNA sequences a–e) with nodes below them. For example, the branches connecting node 1 to sequences a and b represent two terminal branches. Likewise, the branch between node 2 and sequence c is also a terminal branch. The other type of branch is known as an *internal branch*. There are a total of four internal branches in this tree and they are found between nodes 1 and 2, nodes 2 and 4, nodes 3 and 4, and below the root node. The lengths of the branches in gene trees can indicate rates of molecular evolution or evolutionary time depending on how the tree is constructed. Note that the tree in Figure 1.1 represents one of the 105 possible different topologies that could be generated for five labeled tips and a tree that is completely bifurcating (i.e., each node has exactly two descendent branches connected to it) and has a root (Felsenstein 2004). The numbers of unique rooted tree topologies becomes shockingly high as the number of sequences increases. For example, a perusal of Table 3.1 in Felsenstein (2004) shows that for only 10 sequences there are more than 34 million different rooted bifurcating trees and a mind boggling 2.75×10^{76} different trees for 50 sequences! As the focus of this book is on acquiring phylogenomic data, we will not delve into the details on how these trees are made. Readers wanting to learn about methods for inferring phylogenetic trees using molecular data should consult the following references: Hillis et al. (1996), Felsenstein (2004), and Lemey et al. (2009).

1.3 GENE TREES VERSUS SPECIES TREES

When a molecular biologist reconstructs a gene tree for a particular gene family, the interpretations of the resulting tree are clear-cut: the tree shows the inferred evolutionary relationships among the *molecules* (amino acid or DNA sequences) used to make the tree and may also display the rates or timing of lineage divergences. In other words, what is gleaned from a gene tree in this type of phylogenomic study is the evolution of the molecular sequences themselves. In contrast, when a single gene tree is used to infer a phylogeny of populations or species, as has been done innumerable times in traditional molecular phylogenetics, then the researcher is extrapolating an *organismal phylogeny* from a molecular phylogeny (Maddison 1995, 1997). In Tree of Life studies, an organismal phylogeny is more commonly referred to as a *species tree* because it shows the branching relationships and times of divergence among populations or species (Maddison 1995, 1997). Although a gene tree may to some extent mirror a species tree, it is important to realize that gene trees and organismal trees are not the same thing. Thus, a researcher who uses a single gene tree to infer a species tree hopes that they match each other (i.e., are congruent).

Schematic examples of gene and species trees are shown side-by-side in Figure 1.2. Let's first consider the meaning of the topology in each evolutionary tree. The gene tree in Figure 1.2a shows the evolutionary relationships between two DNA sequences (haplotypes) sampled from one gene in two species, whereas the species tree in Figure 1.2b shows the evolutionary relationships between the two species. In addition to the topologies, the divergence times in each type of tree are also fundamentally different from each other. Divergence times derived from a gene tree represent *gene divergence times*, whereas in the species tree the divergence times represent *population* or *species divergence times* (Figure 1.2). The MRCA for the two haplotypes (i.e., gene divergence) corresponds to a single individual in the ancestral population, while the MRCA for the two species is the ancestral species. Thus the gene tree shows the inferred genealogical history of the sampled DNA sequences, whereas the latter exhibits the evolutionary history for populations or species. It can be a perilous practice to naively equate a gene tree with a species tree because, even if a gene tree is reconstructed without error, its topology may be incongruent with the corresponding species tree (Hudson 1983, 1992; Tajima 1983; Maddison 1995, 1997; Rosenberg 2002; Felsenstein 2004).

Notice in Figure 1.2 that the widths of the branches differ between gene and species trees. The branches of a gene tree are thin lines of constant width (Figure 1.2a) while the branches of a species tree are much wider branches. In some cases, the branch widths are drawn in this manner simply to distinguish the schematic of a gene tree from a species tree. In other cases, the widths of each branch in a species tree are drawn to be proportional to the *effective population sizes* for that population at particular points in time (Figure 1.2b). Thus wider branches represent larger effective

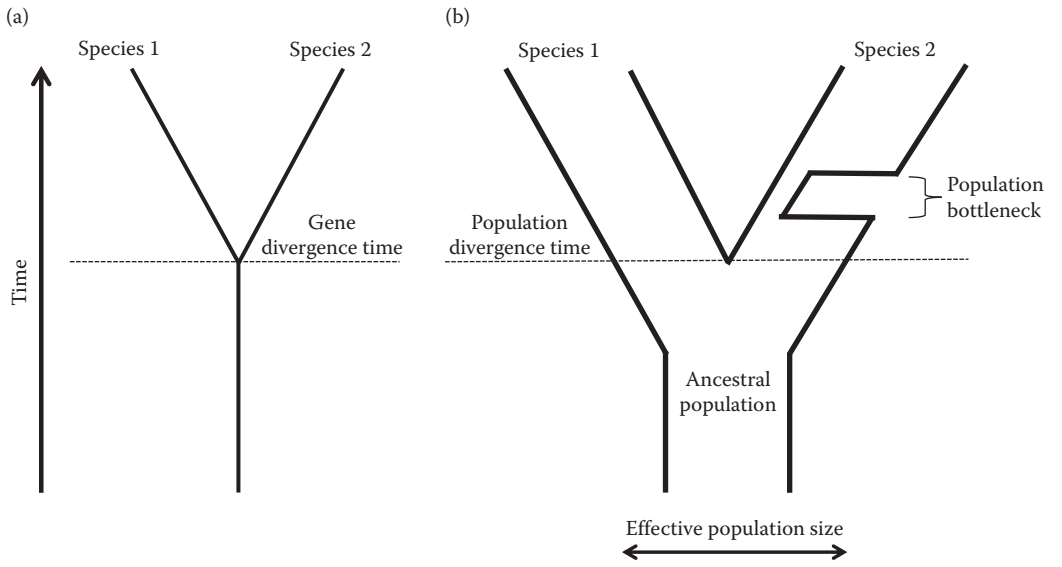

Figure 1.2. A comparison between the anatomy of a gene tree versus a species tree. (a) Shown is a gene tree that depicts the evolutionary history of two haplotypic lineages from two species. The timing of divergence between lineages is called gene divergence time. (b) A species tree showing the evolutionary history of two sister species. The timing of divergence between lineages in a species tree is referred to as the population divergence or speciation time. The branch widths in a species tree are drawn in proportion to the effective population sizes, which may vary through time. In this example, the ancestral population of species 2 is shown to have suffered a population bottleneck.

population sizes than do thinner branches. For example, a species tree with a branch having the same width from node to node means that the long-term effective population sizes have remained constant in size (e.g., ancestral populations of Species 1 in Figure 1.2b). In contrast, the ancestral populations may have undergone fluctuations in size such as from a population bottleneck (e.g., ancestral populations of Species 2 in Figure 1.2b). In contrast, line width in a gene tree has no meaning. Despite the potential lack of congruence between gene and species trees, phylogenomic methods have been developed to address these issues, which in turn, are enabling researchers to generate more accurate and robust estimates of species trees than was ever possible using single gene trees (Knowles and Kubatko 2010).

1.4 PHYLOGENOMICS AND THE TREE OF LIFE

The task of reconstructing the Tree of Life represents a monumental undertaking. In the realm of phylogenomics, there are several levels of study that are contributing to this effort. First, at the shallowest levels in the Tree of Life (i.e., "recent" speciation events), the use of phylogenomic datasets in conjunction with phylogeographic methodology is helping to enumerate the true numbers of extant species as well as provide insights into the history of their formation. *Phylogeography* is the study of how past demographic processes and environmental forces have contributed to speciation and shaped the genetic structure of contemporary populations and species (Avise 2000). Thus, phylogeography provides insights about the temporal and geographical aspects of speciation in recent species radiations and therefore this field shares a close connection to the Tree of Life. On a larger scale, the application of phylogenomic methods for inferring species trees is assisting with the reconstruction of the topology and branch lengths of the Tree of Life (Knowles and Kubatko 2010). Lastly, the use of *DNA barcoding* (Hebert et al. 2003), whether based on single organellar genes, entire organellar genomes, or even vast numbers of nuclear loci, is providing biologists with a powerful and simple tool for identifying known and possible new species. We will now briefly introduce each of these approaches.

Phylogenomics and phylogeography—During the first two decades of phylogeography beginning in the 1980s, the vast majority of phylogeographic studies relied on gene trees that were inferred

from mitochondrial DNA sequences. However, in later years two factors dramatically increased the sophistication of this discipline. First, the commencement of the genomics era provided researchers with far greater access to genomic sequences, which, in turn, enabled them to mine genomes for large numbers of DNA sequence-based loci (e.g., Chen and Li 2001). A second critical factor was the development of software implementing coalescent theory-based Bayesian and maximum likelihood statistical methods (e.g., Yang 2002; Hey and Nielsen 2004; Hey 2010). With the newfound availability of more voluminous datasets and bioinformatics tools, phylogeographers could for the first time conduct more robust multilocus analyses, which were capable of yielding more accurate and precise estimates of key phylogeographic parameters such as divergence times between populations (Yang 2002; Jennings and Edwards 2005).

Phylogenomic methods for estimating species trees—Various approaches for inferring species trees have been employed in phylogenomic studies, some of which make use of individual gene trees while others do not. These methods fall into two main categories: *DNA sequence-based approaches* and *whole-genome features* (Delsuc et al. 2005). By far, most phylogenomic studies have relied on DNA sequence datasets for inferring species trees.

Within the category of DNA sequence-based approaches, three primary methods have been used to estimate species trees: (1) the coalescent-theory or independent gene tree approach (Edwards et al. 2005; Knowles and Kubatko 2010); (2) the "supermatrix" approach (Delsuc et al. 2005; Edwards et al. 2005; Philippe et al. 2005); and (3) the "supertree" approach (Delsuc et al. 2005; Philippe et al. 2005). The independent gene tree approach directly estimates a species tree by analyzing the gene trees in a *coalescent theory* framework (Edwards et al. 2005; Edwards 2009). Coalescent theory is a population genetics-based theory that models, in a probabilistic manner, how historical demography (e.g., population sizes, gene flow) of ancestral populations can influence the distributions of gene trees found in contemporary populations and species (Felsenstein 2004; Wakeley 2009). The study of Australian grass finches by Jennings and Edwards (2005) provides an early example in which a species tree was inferred from a coalescent-based analysis of a genome-wide dataset.

The supermatrix approach basically involves joining together, end-to-end, multiple loci to generate long contiguous sequences with one sequence representing each species thereby resulting in a supermatrix of phylogenetic characters. This method, which is also commonly referred to as *concatenation*, can involve the joining together of tens, hundreds, or thousands of loci from throughout a genome. Once a supermatrix is assembled, the investigator can then use phylogenetic methods to infer a species tree. Two early studies on eukaryotes that used this approach with genome-wide datasets include the work on hominoids by Chen and Li (2001) and on yeasts by Rokas et al. (2003).

Lastly, in the supertree approach, the individual trees, which can be based on molecular data, morphological data, or both, are first independently estimated (usually these are already published) and then a single "supertree" is constructed by essentially stitching the overlapping gene trees into a single composite tree (e.g., Hinchliff et al. 2015).

The "whole-genome features" approach for inferring species trees includes a variety of different methods that rely on using architectural features of genomes or rare genomic changes as phylogenomic characters from which species trees can be inferred (Delsuc et al. 2005; Edwards et al. 2005). Phylogenomic datasets making use of rare genomic changes may include data on: (1) gene order or "synteny" (Delsuc et al. 2005; Edwards et al. 2005); (2) retroposon insertion sites (Shedlock and Okada 2000; Okada et al. 2004; Shedlock et al. 2004; Edwards et al. 2005); or (3) "DNA strings" or "genomic signatures" (Deschavanne et al. 1999; Karlin and Burge 1995; Edwards et al. 2002; Delsuc et al. 2005). Although these methods may help resolve particular deep branches in the Tree of Life, their use thus far has been limited.

DNA barcoding and mitogenomics—In DNA barcoding studies, a single gene sequence, which is typically a mitochondrial gene, is used to identify existing and possible new species (e.g., Hebert et al. 2003; Guarnizo et al. 2015; Jennings et al. 2016). The increase in DNA sequencing power and availability of bioinformatics tools now enables biologists to sequence and annotate entire mitochondrial genomes far faster and simpler than before. Moreover, many recently published mitogenomes essentially represent "byproducts"

of next-generation sequencing (NGS) studies with a focus on nuclear DNA sequences (e.g., Souto et al. 2014; Amaral et al. 2015). This surge in mitogenomes is providing molecular biologists with a treasure trove of data from which to study mitochondrial diseases, evolution of mitochondria, and the Tree of Life. Thus, the nascent field of "mitogenomics" is expected to continue to grow and become a major area of inquiry in phylogenomics.

1.5 SEQUENCING WORKFLOWS TO GENERATE PHYLOGENOMIC DATA

At the present time there are two major workflows for generating phylogenomic datasets: *Sanger sequencing* and *NGS*. Although numerous methodological differences exist between these workflows, both utilize the same basic three steps: (1) "DNA extraction," which is the isolation and purification of genomic DNA from tissues; (2) mass-copying or "amplification" of *target DNA templates* for each locus; and (3) sequencing these target templates using a DNA sequencing platform. We will now briefly look at the differences between Sanger and NGS workflows.

1.5.1 Sanger Sequencing Workflow

The Sanger method for DNA sequencing, named in honor of the English biochemist Frederick Sanger, has been the primary method for sequencing DNA over the past three decades. Sanger and his coworkers published a groundbreaking paper (Sanger et al. 1977) that described a new method for sequencing DNA. For this accomplishment, Sanger was awarded his second Nobel Prize in Chemistry and the method has since gone on to revolutionize biology and medicine.

The three main steps of the Sanger sequencing workflow are (1) isolation and purification of genomic DNA; (2) amplification of target templates using the *polymerase chain reaction* or "PCR"; and (3) sequencing of the PCR products using fluorescently labeled chain-terminating dideoxynucleotides in an automated sequencing machine. In Chapter 4, we will learn how DNA is isolated and purified before we take a detailed look at PCR and Sanger methodologies in Chapters 5 and 6, respectively.

Through the years, various improvements to the Sanger workflow have enhanced its high-throughput capabilities while lowering the cost per sequence. "Throughput" simply refers to the number of DNA sequences that are essentially generated at the same time. These advances made it easier and cheaper for researchers to generate phylogenomic datasets of modest sizes (~1–10 loci). However, the ability of researchers to generate larger phylogenomic datasets such as hundreds of loci or entire genome sequences, is still beyond the reach of most laboratories that rely solely on Sanger sequencing mainly because of cost limitations (i.e., cost of individual NGS experiments is relatively expensive compared to Sanger sequencing).

1.5.2 NGS Workflow

During the mid- to late 2000s several different next-generation sequencing or "NGS" methods and sequencing machines appeared on the genomics scene (Mardis 2008; Shendure and Ji 2008), which set the stage for a dramatic change in the nature of phylogenomics data acquisition. In September of 2005, two papers unveiled the first two NGS methods each of which presented powerful new methods for sequencing genomes: the "polony sequencing" method of Shendure et al. (2005) and the "pyrosequencing" method of Margulies et al. (2005). The sequencing technology presented by Margulies et al. (2005) was incorporated into the first commercial NGS platform called the "Roche/454" (Mardis 2008). Illumina introduced another type of NGS platform a few years later, which was based on a "sequencing by synthesis" technology (Mardis 2008). Bentley et al. (2008) provided an early demonstration of the genome-sequencing ability of the Illumina platform. A fourth NGS method, which used a "sequencing by ligation" approach (McKernan et al. 2009), was incorporated into a sequencing platform called the "SOLiD" (Mardis 2008).

These NGS platforms are often referred to as *second-generation* sequencing platforms while Sanger sequencing is considered a *first-generation* platform (Glenn 2011; Niedringhaus et al. 2011; Hui 2014). However, as is the typical theme in evolution whether it is organismal- or consumer product-based, one to a few of the many novel forms survives the initial diversification event and thrives thereafter. Indeed, from among the flurry of early NGS platforms, the Illumina platform has

emerged to be, by far, the dominant NGS platform used in phylogenomics (Videvall 2016). Newer NGS technologies are appearing on the horizon any one of which might become the future dominant sequencing platform. These include the *third-generation* NGS platforms such as the Ion Torrent by (Life Sciences), PacBio (Pacific Biosciences), and Complete Genomics while a so-called *fourth-generation* NGS platform by Oxford Nanopore is in development (Glenn 2011; Niedringhaus et al. 2011; Hui 2014). Given Illumina's current dominance in the NGS market, in Chapter 7 we will focus on how this platform is being used to generate phylogenomic datasets.

The Illumina platforms have provided a substantial boost to phylogenomic studies in a number of ways. Although this technology was originally developed for sequencing entire genomes in a more efficient manner than Sanger sequencing, researchers have devised various methods for isolating target templates for large numbers of loci, which could then be sequenced. Thus, instead of using PCR to directly generate the intermediary products for a small number of loci (~1–10), the researcher prepares a *sequencing library* that represents a targeted portion of a genome, which can potentially contain the templates for dozens, hundreds, or thousands of loci from an individual's genome. Researchers can further dramatically increase the throughput of single Illumina runs by pooling multiple libraries together for sequencing. This kind of brute sequencing power is producing phylogenomic datasets that are orders of magnitude larger than those generated using Sanger sequencing (e.g., Faircloth et al. 2012; Jarvis et al. 2014; Prum et al. 2015).

NGS technology has impacted phylogenomic data acquisition in at least two other important ways. First, researchers can now generate partial or complete genome sequence datasets for "non-model organisms" far faster and cheaper than using the traditional shotgun cloning-Sanger workflow—the approach that was used to sequence the human genome (Venter et al. 2001). This easier access to substantial amounts of genome data means that researchers can rapidly and easily design new phylogenomic loci using bioinformatics methods rather than having to resort to laborious DNA cloning methods (Thomson et al. 2010). In Chapter 8, we will learn more about developing loci for use in either the Sanger or NGS workflows.

1.5.3 Is Sanger Sequencing Still Relevant in Phylogenomics?

Despite the recent surge in popularity of NGS, Sanger sequencing is still the most widely used method today. Still, as the hurdles to acquiring phylogenomic data using NGS become fewer, at some point in upcoming years some form of NGS technology will inevitably replace the Sanger method as the primary sequencing method used by phylogenomics researchers. Given this eventuality, it is fair to ask the following question: *Is Sanger sequencing still relevant in phylogenomics?*

In the short term, say for at least another 5–10 years, the answer to this question is clearly "yes." Despite the dramatically increased sequencing power offered by NGS, the Sanger method still retains some advantages over NGS such as cheaper costs of data collection for small-scale projects, longer sequence read lengths, and simple computer analysis of raw sequence data. The high cost per run for an NGS sequencer, which generally runs in the thousands of dollars, means that Sanger sequencing still has an important role for smaller projects or, at the very least, for filling in sequence "gaps" caused by incomplete coverage in NGS sequences. Another current difficulty with using NGS to generate phylogenomic data is that the construction of some types of sequencing libraries requires considerable molecular biology skills and specialized equipment not normally found in a Sanger-equipped laboratory. Further, the excitement of these new NGS-based methods for acquiring phylogenomic data can quickly subside for those who are not accustomed to performing bioinformatics analysis of NGS data. Once obtained, NGS data poses a number of terrific challenges owing to difficulties arising from storage, manipulation, and filtering of raw sequence files. Not only that, but such data manipulation must be done using UNIX-based "super" computers that can handle *terabytes* of genomic data (McCormack et al. 2013). At the present time only *one* run of an Illumina sequencer can generate more than a one-half terabyte of sequence data! In contrast, even the highest throughput of Sanger sequence data cannot overwhelm the most basic laptop computer using a Mac or Windows operating system. Thus, while NGS offers tremendous promise for the future, it does not represent a practical sequencing solution for all researchers and projects at the present time.

Even in the longer term after NGS becomes the standard method of phylogenomics data acquisition, there will be good reasons for all phylogenomics researchers to have a solid understanding of Sanger sequencing. First, genomic and nucleotide databases such as GenBank already contain an enormous volume of DNA sequence data generated via this technology. Therefore, researchers who mine sequence databases for bioinformatics purposes should have some knowledge about how those Sanger sequences were generated and understand their characteristics and limitations, which we will review in Chapter 6. Given these considerations, Sanger sequencing will likely remain the primary method for obtaining DNA sequence data by the majority of small labs and perhaps many of the larger labs until the various pre- and postdata NGS hurdles are reduced or eliminated.

Regardless of when NGS relegates Sanger sequencing to a minor role in phylogenomics, this issue of "obsoleteness" touches upon a broader phenomenon in molecular biology that we will now consider. When one looks more closely at the history of molecular biology, one sees the pattern of new technologies (i.e., lab methods and reagents) helping to spur scientific revolutions only to be rendered obsolete some years later as they become replaced by newer higher-performing technologies. What is most interesting about this fact is that these supposedly obsolete methods or reagents are often later co-opted to play essential new roles in these the "cutting-edge" methods. My favorite example of this is the Sanger method for sequencing DNA: Sanger et al. (1977) simply combined pre-existing reagents and methods into a novel methodology for determining the sequence of DNA molecules. Another example is the enzyme DNA ligase, which was independently discovered by five groups of researchers in 1967 (Lehman 1974) and subsequently became a key component in recombinant DNA technology throughout the 1970s and 1980s. Once PCR largely replaced molecular cloning as a means for generating target DNA templates for Sanger sequencing in the late 1980s (except for shotgun sequencing of genomic libraries), ligase was no longer relevant for routine DNA sequencing, correct? The answer is no! The SOLiD (Mardis 2008) NGS platform, works by a *sequencing-by-ligation* principle (McKernan et al. 2009), *and*, the method of ligating DNA molecules together using DNA ligase is an essential step in the construction of many types of NGS sequencing libraries! A third example comes from PCR. As the use of Sanger sequencing workflow declines, does this mean that PCR and the design of PCR primers will be less important in the era of NGS? Again, the answer is a resounding no! Achieving a thorough knowledge about PCR is not only essential for successful Sanger sequencing, but PCR has also found important new niches in many NGS-based methods as we will see in Chapter 7. This phenomenon of rediscovering methods is a recurring theme in molecular biology. The important message here is that it is worthwhile for you to learn all of these methods because future cutting-edge technologies may co-opt existing technologies to produce important new methods for acquiring phylogenomic data.

1.6 THE PHYLOGENOMICS LABORATORY

We will close this first chapter with a brief discussion about what the modern phylogenomics laboratory looks like and see, perhaps surprisingly, how easily they can be set up. Large universities, natural history museums, and other research institutes typically have spacious and expensive molecular genetic laboratories full of various machines (e.g., automated DNA sequencers), equipment, and lab technicians where phylogenomics projects are undertaken. Although historically these sophisticated well-funded facilities were a prerequisite for conducting technically challenging and costly projects that involved gene cloning, high-throughput Sanger sequencing, or NGS-based sequencing, today a small-budget lab can be established to handle such genome-scale projects. Speaking from my own experiences, even a dirty old small storage room can be converted into a clean, shiny, and high-throughput phylogenomics laboratory. All one needs is a small workspace and the minimal equipment. How is this possible?

Recall from the previous section that the process of acquiring phylogenomic data can be divided into three basic steps:

DNA extraction → template acquisition → DNA sequencing

Although DNA extraction is a method common to all labs, methods for acquiring DNA templates

for sequencing vary. Laboratories only equipped for PCR and Sanger sequencing (Chapters 5 and 6) are easy to set up and can be done so at low cost. Figure 1.3 shows a layout of what a minimalistic phylogenomics lab would resemble. If space and funding allow, additional pieces of equipment to the lab can further enhance its capabilities. For example, adding a second thermocycler and electrophoresis box (to be used with the same voltage box) can increase the lab's throughput. If a second electrophoresis apparatus is placed within the pre-PCR area apparatus (with dedicated single-channel and multichannel pipettes), then purified DNA extracts can be evaluated within a "clean" part of the lab, which would further reduce the risk of contamination. Adding a *Qubit* fluorometer (Life Technologies), *BioAnalyzer* (Agilent Technologies), and qPCR machine would enable researchers to construct NGS libraries. Finally, a −80°F or "ultra-cold" freezer can be used for long-term storage of genetic samples (e.g., tissues and DNA extracts *but not PCR products due to contamination concerns*) and reagents (e.g., concentrated primer stocks, enzymes, etc.). We will discuss strategies for minimizing contamination risks (Chapters 4 and 5) and maximizing the sample throughput capability of labs (Chapters 4 through 6).

You may have noticed the absence of an automated sequencing machine as part of the equipment repertoire in Figure 1.3. This is because of a fantastic innovation—*outsourcing* of DNA sequencing to another laboratory. Once the sequencing templates—regardless whether they are PCR products or NGS libraries—are prepared, they can be outsourced to another laboratory that performs the sequencing using their Sanger or NGS machines. The advent of outsourcing of DNA sequencing obviates the need for all laboratories to have their own expensive sequencing machines and technicians to run them, which effectively reduces the costs of setting up a phylogenomics lab from hundreds of thousands (or more) dollars down to tens of thousands of dollars. In addition to the cost savings, this outsourcing option represents a great equalizer because the *per capita* output of high-quality research projects from a small

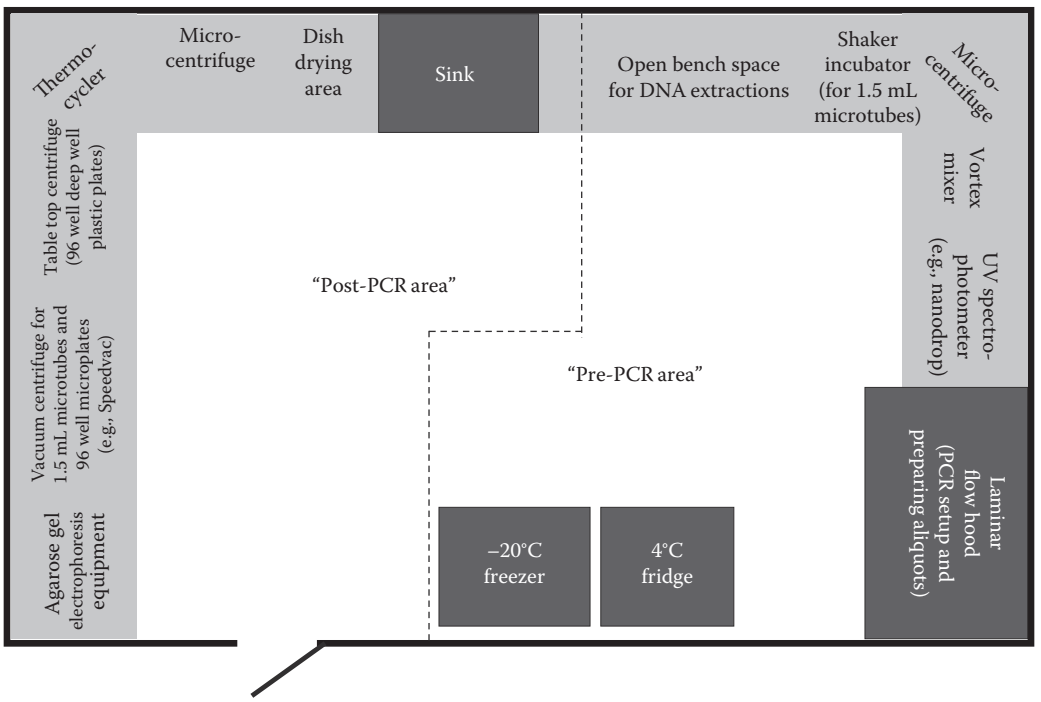

Figure 1.3. Diagram showing layout of a basic phylogenomics lab. This illustration is meant to give an idea of how a small room can be converted into a DNA extraction and PCR lab. The lab is divided into "pre-PCR" and "post-PCR" areas for purposes of minimizing the risk of contamination. Additional equipment can be included to further increase the capabilities of the lab. Shelves (not shown) for storing sterile plastics (kits, microtubes, microplates, pipette tips, etc.) can be placed on the walls above the work benches.

lab can now be more comparable with the larger well-established laboratories found in universities and research institutes.

REFERENCES

Amaral, F. R., L. G. Neves, M. F. Resende Jr et al. 2015. Ultraconserved elements sequencing as a low-cost source of complete mitochondrial genomes and microsatellite markers in non-model amniotes. *PloS One* 10:e0138446.

Avise, J. C. 2000. *Phylogeography: The History and Formation of Species*. Cambridge: Harvard University Press.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59.

Chen, F. C. and W.-H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* 68:444–456.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* 6:361–375.

Deschavanne, P. J., A. Giron, J. Vilain, G. Fagot, and B. Fertil. 1999. Genomic signature: Characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol* 16:1391–1399.

Dobzhansky, T. 1973. Nothing in biology makes sense except in the light of evolution. *Am Biol Teach* 35:125–129.

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

Edwards, S. V., B. Fertil, A. Giron, and P. J. Deschavanne. 2002. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Syst Biol* 51:599–613.

Edwards, S. V., W. B. Jennings, and A. M. Shedlock. 2005. Phylogenetics of modern birds in the era of genomics. *Proc R Soc Lond B: Biol Sci* 272:979–992.

Eisen, J. A. 1998a. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* 8:163–167.

Eisen, J. A. 1998b. A phylogenomic study of the MutS family of proteins. *Nucleic Acids Res* 26:4291–4300.

Eisen J. A. and C. M. Fraser. 2003. Phylogenomics: Intersection of evolution and genomics. *Science* 300:1706–1707.

Eisen, J. A., K. S. Sweder, and P. C. Hanawalt. 1995. Evolution of the SNF2 family of proteins: Subfamilies with distinct sequences and functions. *Nucleic Acids Res* 23:2715–2723.

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.

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

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

Guarnizo, C. E., A. Paz, A. Muñoz-Ortiz, S. V. Flechas, J. Méndez-Narváez, and A. J. Crawford. 2015. DNA barcoding survey of anurans across the eastern cordillera of Colombia and the impact of the Andes on cryptic diversity. *PLoS One* 10:e0127312.

Haig, D. 1999. A brief history of human autosomes. *Philos Trans R Soc Lond B: Biol Sci* 354:1447–1470.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. deWaard. 2003. Biological identifications through DNA barcodes. *Proc R Soc Lond B: Biol Sci* 270:313–321.

Hey, J. 2010. Isolation with migration models for more than two populations. *Mol Biol Evol* 27:905–920.

Hey, J. and R. Nielsen. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167:747–760.

Hillis, D. M., B. K. Mable, and C. Moritz. 1996. *Molecular Systematics*, 2nd edition. Sunderland: Sinauer.

Hinchliff, C. E., S. A. Smith, J. F. Allman et al. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc Natl Acad Sci USA* 112:12764–12769.

Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37:203–217.

Hudson, R. R. 1992. Gene trees, species trees, and the segregation of ancestral alleles. *Genetics* 131:509–513.

Hui, P. 2014. Next generation sequencing: Chemistry, technology and applications. *Top Curr Chem* 336:1–18.

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.

Jennings, W. B. and S. V. Edwards. 2005. Speciational history of Australian Grass Finches (*Poephila*) inferred from thirty gene trees. *Evolution* 59:2033–2047.

Jennings, W. B., H. Wogel, M. Bilate, R. O. L. Salles, and P. A. Buckup. 2016. DNA barcoding reveals species level divergence between populations of the microhylid frog genus *Arcovomer* (Anura: Microhylidae) in the Atlantic Rainforest of southeastern Brazil. *Mitochondrial DNA* Part A 27:3415–3422.

Karlin, S. and C. Burge. 1995. Dinucleotide relative abundance extremes: A genomic signature. *Trends Genet* 11:283–290.

Knowles, L. L. and L. S. Kubatko, eds. 2010. *Estimating Species Trees: Practical and Theoretical Aspects*. Hoboken: John Wiley and Sons.

Lehman, I. R. 1974. DNA ligase: Structure, mechanism, and function. *Science* 186:790–797.

Lemey, P., M. Salemi, and A. Vandamme, eds. 2009. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, 2nd edition. Cambridge: Cambridge University Press.

Maddison, D. R., K. S. Schulz, and W. P. Maddison. 2007. The tree of life web project. *Zootaxa* 1668:19–40.

Maddison, W. P. 1995. Phylogenetic histories within and among species. In *Experimental and Molecular Approaches to Plant Biosystematics. Monographs in Systematic Botany* eds. P. C. Hoch, and A. G. Stephenson, 53:273–287. St. Louis: Missouri Botanical Garden.

Maddison, W. P. 1997. Gene trees in species trees. *Syst Biol* 46:523–536.

Mardis, E. R. 2008. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 9:387–402.

Margulies, M., M. Egholm, W. E. Altman et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380.

McCormack, J. E., S. M. Hird, A. J. Zellmer, B. C. Carstens, and R. T. Brumfield. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenet Evol* 66:526–538.

McKernan, K. J., H. E. Peckham, G. L. Costa et al. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* 19:1527–1541.

Murphy, W. J. 2008. *Phylogenomics: Methods in Molecular Biology*. New York: Humana Press.

Niedringhaus, T. P., D. Milanova, M. B. Kerby, M. P. Snyder, and A. E. Barron. 2011. Landscape of next-generation sequencing technologies. *Anal Chem* 83:4327–4341.

O'Brien, S. J. and R. Stanyon. 1999. Phylogenomics: Ancestral primate viewed. *Nature* 402:365–366.

Okada, N., A. M. Shedlock, and M. Nikaido. 2004. Retroposon mapping in molecular systematics. In *Mobile Genetic Elements: Protocols and Genomic Applications*, ed. P. Capy, 189–226. New York: Humana Press.

Philippe, H. and M. Blanchette. 2007. Overview of the first phylogenomics conference. *BMC Evol Biol* 7:1.

Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005. Phylogenomics. *Annu Rev Ecol Evol Syst* 541–562.

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.

Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. *Theor Popul Biol* 61:225–247.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74:5463–5467.

Shedlock, A. M. and N. Okada. 2000. SINE insertions: Powerful tools for molecular systematics. *Bioessays* 22:148–160.

Shedlock, A. M., K. Takahashi, and N. Okada. 2004. SINEs of speciation: Tracking lineages with retroposons. *Trends Ecol Evol* 19:545–553.

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

Shendure, J., G. J. Porreca, N. B. Reppas et al. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309:1728–1732.

Souto, H. M., P. A. Ruschi, C. Furtado, W. B. Jennings, and F. Prosdocimi. 2014. The complete mitochondrial genome of the ruby-topaz hummingbird *Chrysolampis mosquitus* through Illumina sequencing. *Mitochondrial DNA* 27:769–770.

Tajima, F. 1983. Evolutionary relationships of DNA sequences in finite populations. *Genetics* 105:437–460.

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

Venter, J. C., M. D. Adams, E. Myers et al. 2001. The sequence of the human genome. *Science* 291:1304–1351.

Videvall, E. 2016. Results of the molecular ecologist's survey on high-throughput sequencing blog. *The Molecular Ecologist* blog. http://www.molecularecologist.com/2016/04/results-of-the-molecular-ecologists-survey-on-high-throughput-sequencing/ (accessed April 11, 2016).

Wakeley, J. 2009. *Coalescent Theory: An Introduction* (Vol. 1). Greenwood Village: Roberts & Company Publishers.

Yang, Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics* 162:1811–1823.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com