CHAPTER SIX

Sanger Sequencing

In Chapter 5, we saw that a single PCR is capable of generating a very large number of copies for a locus of interest. Why then are so many replicate templates needed for the Sanger sequencing method? The reason is because enough copies of a particular locus are needed in order to generate sufficient signal for the laser detection system of a modern DNA sequencing machine to record the sequence of a PCR product. Of course in order for the sequencer to accomplish this imaging task, the templates must be labeled with a fluorescent dye, which can be detected by the machine's laser.

In recent years, it has become cheaper and easier to use Sanger sequencing thanks to further methodological improvements, enhancements to sequencing machines, and especially the availability of outsourcing of DNA sequencing. These factors have simplified the sequencing process to the extent that a researcher can now generate a wealth of DNA sequence data while having only a fuzzy understanding about the molecular mechanisms underlying Sanger sequencing. This is remarkable considering that the modern Sanger methodology is a medley of molecular biology techniques. However, as with other molecular methods such as PCR, a researcher who has a solid knowledge of the details underlying each of these components to the Sanger methodology can achieve a higher level of success in obtaining the desired sequence data.

6.1 PRINCIPLES OF SANGER SEQUENCING

6.1.1 The Sanger Sequencing Concept

The ingenious DNA sequencing method developed by Frederick Sanger and colleagues capitalized on an earlier discovery made by other molecular biologists. Atkinson et al. (1969), who were looking at how nucleotide analogs might be used to investigate the functions of DNA polymerases, demonstrated that an analog to the naturally occurring 2′-dTTP known as a 2′-,3′-dideoxythymidine triphosphate (ddTTP) could terminate the synthesis of a DNA strand after it had been incorporated into the same strand. Accordingly, they described ddNTPs as being *chain growth terminators*. An example of a ddNTP is shown in Figure 6.1. If you compare the usual dNTP (Figure 5.2) with the ddNTP shown in Figure 6.1 you will see that these nucleoside triphosphates differ only by the latter lacking a 3′-hydroxyl group; that is, there is only a hydrogen atom located at the 3′ carbon position on the sugar moiety of a ddNTP. However, as Atkinson et al. (1969) pointed out, this minor structural difference can have a dramatic functional consequence on DNA synthesis: a ddNTP *can* be added to a growing strand of DNA in normal fashion (like dNTPs) but it *cannot* act as a primer for another ddNTP or dNTP therefore synthesis of that strand abruptly terminates.

Sanger et al. (1977) exploited the chain terminating property of ddNTPs to develop their method of DNA sequencing. Their method consisted of two main steps. First, they performed *in vitro* DNA synthesis using a single primer, template DNA, DNA polymerase (Klenow fragment), and importantly, a *mixture* of dNTPs and ddNTPs. By using a mix of dNTPs and ddNTPs in the same reaction they could generate a distribution of variably sized but nested primer *extension products* (Watson et al. 2014). In the second step, they fractionated the extension products on a denaturing acrylamide gel. Because they had labeled some ddNTPs with a radioactive isotope (^{32}P) prior to

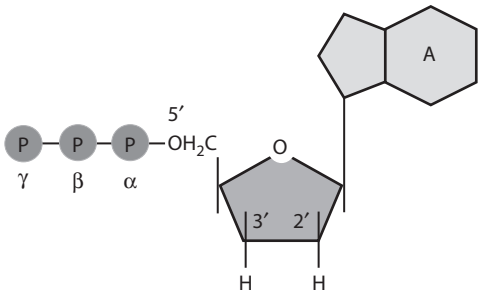

Figure 6.1. Chemical structure of a 2′-, 3′-dideoxynucleoside triphosphate (ddNTP). The example shown here is 2′-, 3′-dideoxyadenosine triphosphate (ddATP). The lack of an oxygen atom at the 3′ position on the ribose sugar represents the only difference between this ddATP and the dATP shown in Figure 5.2.

the sequencing reactions, they could visualize banding patterns in the gel that corresponded to the relative positions of a particular base type. Four separate sequencing reactions—one for each base type (A, G, C, and T)—had to be performed in order to elucidate the desired sequence.

Although this new chain termination method of DNA sequencing was superior to other proposed methods for DNA sequencing at that time (e.g., the method of Maxam and Gilbert 1977), the Sanger method would not come of age until more than a full decade later. Similar to the modernization of PCR, the 1980s were a period of tremendous innovation of new DNA sequencing refinements many of which dramatically improved the Sanger methodology. What were these advances that improved the Sanger method? A timeline for the modernization of the Sanger method is shown in Table 6.1. One of the difficulties with the Sanger method in its earliest implementation was that radioactive isotopes had to be used to "label" the ddNTPs to allow for visualization of the DNA sequence via autoradiographs. However, in 1985 researchers from Caltech published a paper (Smith et al. 1985) showing how sequence visualization could be accomplished using fluorescent-labeled primers (or "dye-primers") and a laser detector. This advance not only removed the health hazard posed by working with radioactive isotopes, but it enabled the production and first trial of a semi-automated DNA sequencing machine in 1986 (Smith et al. 1986). A year later, another research group unveiled the first use of fluorescent-labeled ddNTPs (or "dye-terminators") in DNA sequencing (Prober et al. 1987). Although the use of radioactive-labeled nucleotides and primers and dye-primers would continue to be used in Sanger sequencing into the mid-1990s, they would all be superseded by dye-terminators in DNA sequencing within a few years thereafter (Table 6.1).

The critical contributions from PCR to Sanger sequencing cannot be overstated. First, PCR provided a simple method for generating template DNA for Sanger sequencing that was far superior to cloning-based methods (Innis et al. 1988). Secondly, researchers dramatically improved the methodology for generating the Sanger sequencing extension products by co-opting PCR components including *Taq* polymerase and thermocyclers (Innis et al. 1988; Carothers et al. 1989; Murray 1989). This PCR-like sequencing reaction, which is called *cycle sequencing* or "sequencing PCR," uses a linear polymerase reaction to generate the extension products (Hillis et al. 1996). We will soon see why this is a linear rather than an exponential production process. The development of fluorescent dye-terminators

TABLE 6.1
Timeline for the modernization of Sanger sequencing

Year	Advance	References
1977	Chain-termination method for DNA sequencing published	Sanger et al. (1977)
1985	Fluorescent-labeled primers introduced	Smith et al. (1985)
1986	First semi-automated DNA sequencer introduced	Smith et al. (1986)
1987	Fluorescent-labeled ddNTPs introduced	Prober et al. (1987)
1989	Cycle sequencing using *Taq* polymerase introduced	Murray (1989); Carothers et al. (1989)
1996	ABI 377 automated "slab gel" DNA sequencer introduced	Applied Biosystems Wikipedia
1999	ABI 3700 automated "capillary" DNA sequencer introduced	Applied Biosystems Wikipedia

and cycle sequencing together with advances in computers collectively made possible the existence of the first fully automated DNA sequencing machines by the middle 1990s and a short time later the more advanced capillary sequencers (Table 6.1). The capillary DNA sequencer has remained largely unchanged since then and they are still being used despite the increasing popularity of newer NGS platforms. Readers interested in learning more about the history of DNA sequencing as well as the principles and methods not discussed here should read the comprehensive account in Hillis et al. (1996).

6.1.2 Modern Sanger Sequencing

6.1.2.1 Cycle Sequencing Reaction

Recall that the first essential principle of the Sanger method is to use a single primer, a DNA template, a mixture of dNTPs and ddNTPs, and DNA polymerase to produce a nested set of extension products. Accordingly, before the cycle sequencing reaction can be set up, the researcher must first use PCR to obtain good quality templates for sequencing. However, before PCR products can be sequenced they must be "cleaned," which means that unincorporated dNTPs and primers must be removed or inactivated otherwise they will interfere with the sequencing reaction. We will see why this cleaning step is necessary as well as learn about the techniques for cleaning PCR products later in this chapter. The cleaned PCR products are then used as templates in a cycle sequencing reaction.

Although regular PCR and cycle sequencing share many similarities, important differences between the two methods also exist. For example, the cycle sequencing reaction requires a mixture of dNTPs and fluorescent-labeled ddNTPs (dye terminators) in approximately a 100:1 ratio and only a *single* primer, which is called the *sequencing primer*, is used in the reaction. The sequencing primer, which directs the sequencing of one strand of the PCR product, is typically one of the original PCR primers used to generate the PCR product that is being sequenced. However, other types of sequencing primers also exist, which we will see later in this chapter. The computer program that the thermocycler uses to execute a cycle sequencing reaction is similar to a typical PCR program, as it subjects the reactants to the same denaturation, annealing, and extension steps repeated for 35 or more cycles. However, cycle sequencing only requires a brief annealing time of 5 seconds because a large concentration of the template, which perfectly matches the sequencing primer, is present. A longer extension time of 4 minutes per cycle is also used to help ensure that the extension products are properly formed (i.e., terminated by a ddNTP).

Let's now take a detailed look at the cycle sequencing reaction once it has begun in a thermocycler. After each denaturation step the double-stranded PCR products become single stranded. As the single-stranded amplicons cool down to the set annealing temperature (most often at 50°C), the sequencing primers bind to their complementary sites on one of the PCR strands as is shown in Figure 6.2a. The temperature of the reaction mixture is then raised to 60°C so that *Taq* polymerase can begin incorporating nucleotides starting at the 3′ end of the primer where the primer–template junction is initially located. Because there are 100 times more dNTPs than ddNTPs, the former type of nucleotide is more likely to be incorporated into the growing strand. However, synthesis stops once a ddNTP becomes incorporated into the strand (Figure 6.2b). In this example, synthesis of the strand was terminated after a ddGTP was added. The strand that includes the sequencing primer plus the newly synthesized DNA including the ddNTP at the 3′ end is defined as the extension product. This is a good moment to reflect on why unincorporated dNTPs leftover from the PCR must be removed prior to cycle sequencing. If unused dNTPs are not removed, then they will dilute the ddNTPs to the point that far fewer extension products will be made leading to a low quality sequence.

In our earlier discussion, we learned that cycle sequencing generates a linear increase in products, which is in contrast to the exponential process of PCR. Let's now see why this must be the case. During the first cycle of a cycle sequencing reaction, a single PCR product first denatures into two single strands. Only one of these strands will be used as template throughout all 35 cycles because only a single sequencing primer is used. The other PCR strand is not used in this reaction. If the researcher wishes to sequence the other strand, then a second cycle sequencing reaction must be performed in a separate PCR tube that includes the appropriate sequencing primer.

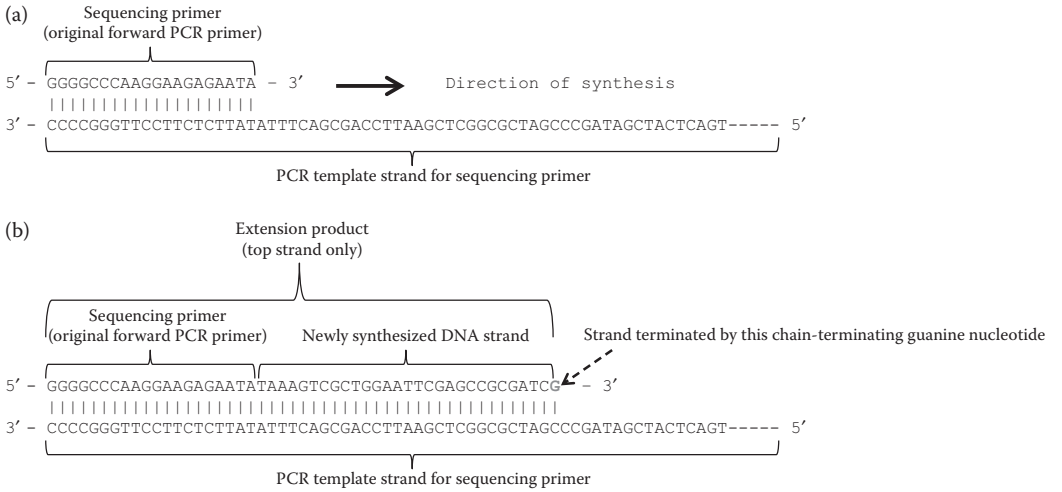

Figure 6.2. Annealing and extension steps in a cycle sequencing reaction. (a) Annealing step: forward sequencing primer (original forward PCR primer) anneals to the stretch of complementary sites on the PCR template strand. (b) Extension step: *Taq* polymerase (not shown) extends the newly synthesized strand until a ddNTP is incorporated into the strand. Once incorporated into the nontemplate strand, the newly added nucleotide (now ddGMP), which is shown in gray, terminates synthesis. Vertical dashes between bases indicate hydrogen bonding between Watson–Crick base pairs. The dashes at the 3′ end of the PCR template (lower) strand indicate that the PCR template strand is truncated on the right due to lack of space in the figure.

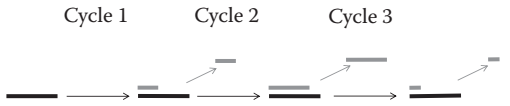

Figure 6.3. Synthesis of extension products during the first three cycles of a cycle sequencing reaction. After the extension step in each cycle, a new extension product (gray) can be formed from each PCR amplicon though the number of templates (black) does not change. The lengths of the extension products are variable owing to the random nature of ddNTP incorporation.

Following the first cycle, one extension product per PCR amplicon will be produced as shown in Figure 6.3. In the second cycle, the denaturation step denatures the PCR template-extension product duplex, which makes the PCR template available for a new sequencing primer during the annealing step. After the second cycle, a second extension product is made from a PCR template, and so on. Thus, each PCR template can potentially be reused in each cycle and the number of templates remains fixed throughout the process (Figure 6.3).

6.1.2.2 Gel Electrophoresis of Extension Products

The second essential principle of the Sanger method is to separate the extension products by their size using high-resolution gel electrophoresis. Polyacrylamide is used as the separation matrix in sequencing, rather than agarose, because it has the ability to separate DNA fragments differing in length by a single nucleotide. In order for this to perform as desired, the extension products must travel through the gel matrix in a single-stranded and completely linear state. The mobility of extension products through a gel will be adversely affected if the extension products are either duplexed with a PCR template or if they have any secondary structures (e.g., hairpin loops, etc.). Thus, immediately prior to electrophoresing the extension products, they are denatured into single strands using high heat (>80°C) and mixed with chemicals (e.g., formamide) to prevent the formation of secondary structures.

Once the extension products made in a sequencing reaction are loaded into a capillary sequencer, they are then fractionated by length in the gel-filled capillaries; that is, they are size-sorted using electrophoresis. As the extension products pass by the machine's laser in order of their length with shorter products appearing first, the fluorescent-dyed ddNTP of each extension product is recorded. Figure 6.4 illustrates a simplified example of capillary sequencing. In Figure 6.4,

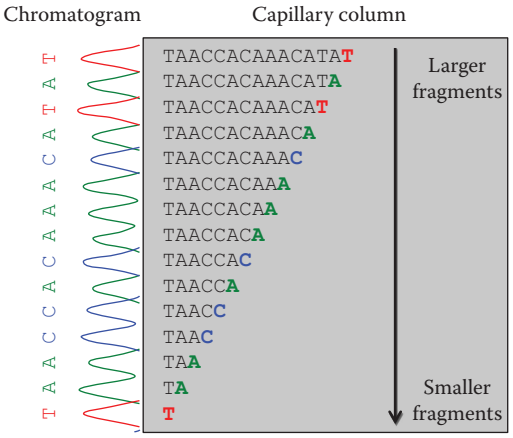

Figure 6.4. Electrophoresis of extension products in a capillary sequencer and recording of the sequence. Extension products are separated according to their length in the capillary column with smaller fragments moving through the gel matrix at a faster rate than longer fragments. The sequencer's laser detects the fluorescence emitted by the ddNTP located at the 3′ end of each extension product and displays the DNA sequence in a chromatogram. Extension products are shown in a 5′ → 3′ orientation.

each sequence in the capillary column represents a group of many thousands of identical extension products that pass by the laser over a short time interval thereby reporting the same type of ddNTP (i.e., A, C, G, or T base). When a group of such fragments passes by the laser a signal of sufficient strength is detected by the machine and recorded as a colored peak corresponding to one of the bases (Figure 6.4). As a consequence of the nested nature of the extension products, during electrophoresis the size-specific groups travel in an orderly progression through the gel matrix with the shortest products at the front owing to their rapid mobility through the matrix. After the last (and longest) of the extension products passes by the laser, the sequencer's computer software assembles the data into a DNA sequence. Looking at Figure 6.4, it should be apparent why only a single sequencing primer can be used in a cycle sequencing reaction: if two primers were used (one to sequence each opposing strand), then two sets of nonoverlapping extension products would be produced thereby producing ruined sequence data.

Capillary sequencers are capable of outputting high-quality sequence data between 700 and 1,000 bp per *read*. A raw sequence "read" is the length of the sequence in bp that a sequencer can accurately generate. As you can see in Figure 6.4, the sequencer determines the sequence using the information gleaned from laser-detection of the fluorescent-extension products. The name of the output file generated by an automated sequencer is called a "*chromatogram*," "*electropherogram*," or "*trace*," an example of which is shown in Figure 6.5. The chromatogram shows the sequence or *base calls* (A, T, G, and C) from one DNA strand. Each base call is based on a distinct colored peak in the chromatogram with the four base types being distinguished by color: adenine (green), thymine (red), guanine (black), and cytosine (blue) as is shown in Figure 6.5. For example, as a group of identical-length extension products passes by the laser, the instrument records each product's fluorescent ddNTP label and then tabulates the hundreds or thousands of independent values into distribution; a peak that resembles a single symmetrical distribution signifies that all the products of that length shared the same type of ddNTP and therefore the base call is correct. When the sequencing computer is unable to call a particular base due to weak or conflicting signals, an "N" symbol will be used to indicate an unknown base.

6.1.2.3 Sequence Data Quality

Chromatogram files represent an important milestone not only because they contain the raw sequence data needed for analyses, but because they also provide the researcher with an effective means for determining the quality of each base call in a sequence. It is crucial for a researcher to

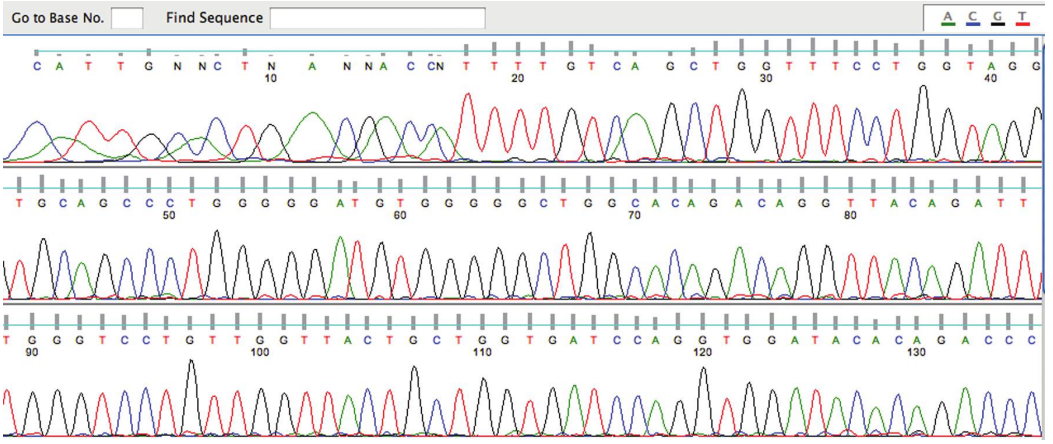

Figure 6.5. Chromatogram of a sequenced PCR product. Each colored peak represents a different base with thymine (red), guanine (black), cytosine (blue), and adenine (green). The sequencer records an "N" for ambiguous base calls when the quality of the sequence is poor. The vertical gray bars above each peak are the quality scores. Vertical gray bars that rise above the horizontal line are considered good quality base calls, whereas bars remaining below the line are judged poor quality.

distinguish poor versus good quality sequence data before proceeding with any computer analyses of those data. Chromatograms are rich in information, which can be used by the researcher to troubleshoot problems. Numerous software packages are available that allow the user to edit base calls, reverse complement a chromatogram, compute quality scores, and export a text version of the sequence for use in downstream analyses.

How is the quality of each base call evaluated on a chromatogram? Both qualitative and quantitative methods have been developed for assessing the quality of base calls. First, let us consider how to evaluate base calls using qualitative criteria. In Figure 6.6 we see a sequence chromatogram, which shows the first 140 base calls. Notice that the first half of the sequence is comprised of messy-looking or indistinct peaks, whereas the latter half is represented by a sequence of symmetrical peaks. Good quality sequence is generally comprised of *single distinct peaks* that are *evenly spaced from each other*, which is what we see from base calls #66–140. In contrast, base calls #1–65 are of poor quality. Another measure used to judge base calls is the *quality score*, which is shown on the chromatogram as a vertical gray bar above each peak. Quality scores that rise above the thin horizontal line are deemed to be good (acceptable) quality base calls, whereas bars that lie below the line indicate poor (unacceptable) quality base calls. In Figure 6.6 notice that the quality scores before base call #66 are consistently poor while the latter bases are all good—as we would expect had we only examined the colored peaks. Another characteristic you see in chromatograms concerns the unevenness of peak height, as some peaks are higher than others. This facet of a chromatogram should be ignored because this pattern reflects the idiosyncrasies of the cycle sequencing chemistry rather than any underlying reality about the organism's DNA.

Because visually scanning chromatograms to evaluate the quality of base calls is a tedious and slow process, researchers developed a bioinformatics approach that can evaluate the quality of each base call in an automated fashion. This was accomplished with the development of the program *Phred* (Ewing and Green 1998; Ewing et al. 1998), which computes a *Phred* quality score or "Q-score" for each base call on individual chromatograms. The Q-score reflects the probability of an incorrectly called base. For example, a Q-score of 20 or "Q20" is a commonly used threshold for discriminating poor versus acceptable base calls: scores below Q20 are considered poor quality, while at or above this value are deemed as good quality base calls. A score of Q20 means there is a 1 in 100 chance that the called base is incorrect. A Q-score of 10 means there is a 10% chance that a base call was incorrect, a Q30 is 1 in a 1,000 chance of such an error, and

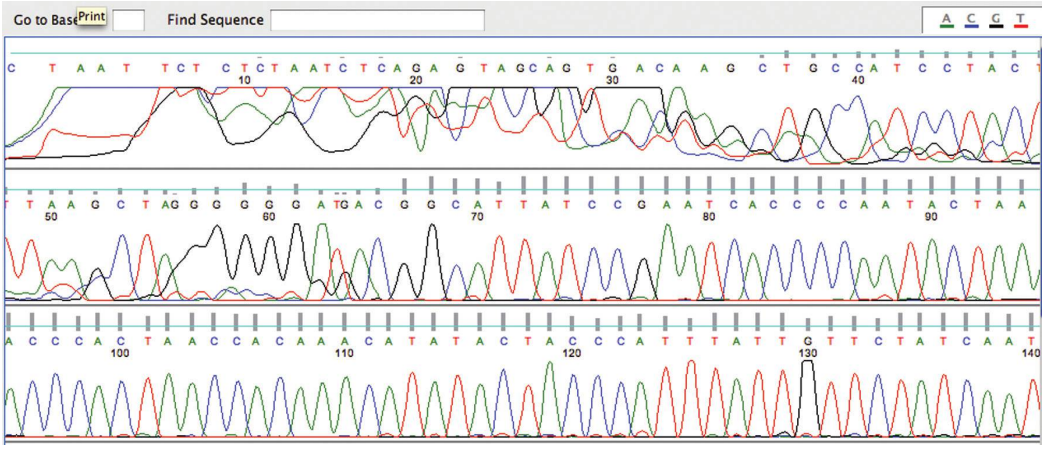

Figure 6.6. Chromatogram of a sequenced PCR product showing low-quality base calls in the first ~65 base positions of the sequence. Although the sequencer called a number of bases in this beginning part of the sequence, these base calls cannot be trusted because of the lack of evenly spaced single peaks, which is also reflected by the low quality scores.

Q60 implies there is a 1 in a million chance that a base was not correctly called.

It is not unusual for a chromatogram to show poor base calls for the first ten to thirty bases as well as the bases beyond approximately the 700th base call (e.g., Figure 6.5). Even in the highest quality sequences the first ~10 base calls are usually of poor quality simply due to the vagaries of the Sanger sequencing technology. If the first part of a sequence is needed for a study, then the same PCR product should be sequenced from the other direction (using the other primer) so that a reverse complement of the second sequence can be compared to the first sequence. Conducting a second cycle sequencing reaction of the same PCR product for purposes of sequencing the other strand is also desirable if the PCR products are longer than 700 bp or if there are concerns about quality of the first sequence. It is common practice for researchers to sequence each PCR product from both directions. Sequencing from both directions may be required to obtain whole sequences particularly if a long microsatellite region occurs in the sequence, as such repeats tend to cause problems when sequencing.

Chromatograms can also exhibit poor base calls beyond the norm just described, which can result in sequence data being either unusable or partially usable at best. There are various causes for such poor results. For example, problems can arise if unpurified PCR products (i.e., containing unincorporated primer and unincorporated dNTPs) or products containing nontarget amplicons are used in a cycle sequencing reaction. Primer dimers are one common type of PCR artifacts that can interfere with sequencing and unfortunately they are difficult to remove from PCR products. Inclusion of primer dimers in a cycle sequencing reaction results in poorer quality sequence data, as the chromatogram will show too much signal during the first 30–60 bases (i.e., poor base calls) and less signal downstream resulting in shorter read lengths. Other types of PCR artifacts include nonspecific PCR products caused by having the primers annealing to nontarget areas of the genome. PCR artifacts are generally caused by poorly performing primers (Chapter 8).

The use of uncleaned extension products can also lead to poor sequence data because unincorporated dye-terminators will interfere with the sequencing process. If dye-terminators are not removed following the cycle sequencing reaction, then so-called "dye blobs" can appear on the chromatogram. Dye blobs, which are colored lines that "float" above other (often distinct) peaks and span one or several (or more) base positions, can prevent the accurate calling of the first 30–100 bases of a sequence as well as in other places on the chromatogram. As we will see later in this chapter, there are other causes of problematic chromatograms that are not due to the factors just described but instead are due to limitations of the Sanger method.

6.2 SANGER SEQUENCING PROCEDURES

Following the successful amplification of a target locus the remaining steps toward getting those products sequenced will involve (1) PCR product purification; (2) cycle sequencing; (3) cleanup and resuspension of extension products; and (4) loading cleaned extension products into an automated sequencer. If you plan to outsource your DNA sequencing, then you can skip all or some of these steps. For example, one option is to simply send your unpurified PCR products by mail to the sequencing provider where their technicians will perform all four steps. Alternatively, many researchers prefer to only perform step 1 before sending their products to be sequenced or they complete steps 1–3 before delivering their products to be sequenced. If you do choose to outsource your sequencing, then be sure to check with the sequencing provider on their requirements for sample submission. Let's now take a closer look at each of these steps.

6.2.1 Purification of PCR Products

There are various methods available for cleaning up PCR products. It is important to note that each method has advantages and disadvantages and so the best method for a given situation will depend on one or more factors. These factors include the quality of the PCR results, cost of reagents or kits, and the lab time required to perform the procedures.

6.2.1.1 Exo-SAP Treatment of PCR Products

Werle et al. (1994) developed a simple PCR purification method that only consists of two steps: pipetting a two-enzyme mixture into original PCR tubes or plate followed by a brief thermal incubation period to allow the enzymes to perform their functions. Although the Exo-SAP method is considered a form of PCR product "purification," it does not actually purify anything. Instead, this method enzymatically destroys unused primers and inactivates unused dNTPs. One of the enzymes is **Exo**nuclease I, which destroys single-stranded DNA. This enzyme is thus used to degrade unincorporated primers. The second enzyme is **S**hrimp **A**lkaline **P**hosphatase or "SAP," which dephosphorylates the excess dNTPs thereby preventing these molecules from participating in DNA synthesis reactions.

Exo-SAP has significant advantages over other methods including its low cost/sample, fast preparation time without the need of centrifugation (a 96 plate of PCR products can be treated in much less than an hour), and some commercial Sanger sequencing facilities will perform the Exo-SAP procedure on a 96 plate for very low cost/sample—even as low as $10 USD per 96 sample plate ($0.10/sequence). The primary disadvantage is that Exo-SAP cannot destroy or inactivate primer dimers or other double-stranded PCR artifacts when they are also present with the target products. If these double-stranded nontarget products are much shorter in length than the target product (e.g., the artifacts are <100 bp and the target is >500 bp), then you can still acquire quality sequence data. The numbers of ambiguous base calls in the beginning of a chromatogram will be determined by the length of the nonspecific PCR products. For example, most primer dimers are ~40–60 bp long, thus the first 40–60 bp of the sequence will consist of unreadable base calls. If the reverse sequence can be obtained, and if it reaches the opposite end of the product, then you can usually recover the lost bases and obtain the entire target sequence. Also, if the nontarget products have both PCR primers incorporated at the ends, then both the forward and reverse sequences will be affected. However, occasionally such products are formed only from one of the primers and so in those cases only one sequence will suffer while the other will be unaffected. The Exo-SAP enzyme mixture can be purchased as a ready-to-use "kit" or you can save money and buy the enzymes and chemicals as separate components and prepare the mix yourself.

6.2.1.2 Spin Column and Vacuum Manifold Kits for PCR Product Purification

One of the earliest available "user friendly" methods for preparing PCR products for Sanger sequencing consisted of the "spin column" and "vacuum manifold" kits. The former method consists of adding the PCR products to membrane-containing plastic spin columns, which are subsequently treated using various proprietary buffers and multiple rounds of centrifugation at high speed (~14,000 rpm). The final round of centrifugation leads to the recovery of the purified PCR products. The vacuum method also uses plastic membrane-containing columns but the

buffers and DNA are drawn through the column using vacuum suction instead of centrifugation. Both types of kits yield PCR products that are free of unused primers, dNTPs, and, depending on the kit, also primer dimers and short nontarget products. These kits can be used in either single tube or 96-sample formats. The main drawback is the higher cost/sample.

6.2.1.3 20% PEG 8000 Precipitation of PCR Products

Polyethylene glycol (PEG) has long been a useful chemical in molecular biology with many different applications. PEG acts as a "molecular crowding agent," which has been used to dramatically improve the efficiencies of enzymatic reactions (Zimmerman and Pheiffer 1983) and size-select DNA molecules in solutions (Lis and Schleif 1975; Lis 1980; Rosenthal et al. 1993). In phylogenomics, PEG is a simple and inexpensive method used to purify PCR products and, as we will see in Chapter 7, size-select DNA fragments in NGS libraries.

The PEG method purifies PCR products through precipitation of target PCR molecules. A 20% solution of PEG 8000 (PEG molecular weight = 8000) and sodium chloride is used to fractionate the DNA in a PCR according to molecular mass (length in bp) with DNA molecules longer than 100–200 bp precipitating out of solution while *all* types of smaller products—single- and double-stranded—remain in the supernatant (Lis and Schleif 1975; Lis 1980; Rosenthal et al. 1993). Thus, after the supernatant containing most of the unused primers and nonspecific products is discarded, the remaining DNA in the tube should only be the target product. The remaining steps consist of two separate ethanol washes of the PCR products, drying of the products, and resuspension of the products using pure water. Other than its negligibly low cost per sample, the main advantage of this method is that it not only removes unused primers and dNTPs, but it also works superbly to remove *all* nontarget nucleic acid molecules such as primer dimers and nontarget PCR products that are <200 bp long (when using 20% PEG). Another advantage of this method is that for weakly amplified PCR products the target PCR product can be further concentrated if a smaller volume of water or buffer is used to resuspend the DNA pellet in the last step. Thus one strategy for obtaining a sufficient amount of target PCR product for samples that are weakly amplified is to redo a PCR with a much larger volume. The PCR products are then cleaned using the PEG procedure and the desired DNA is resuspended with a smaller volume of water thereby yielding a higher concentration sufficient for Sanger sequencing. Although 20% PEG works well when purifying PCR products of typical size (>200 bp), this method can be modified to purify products that are in the 100–200 bp range. To do this, one must optimize the concentration of the PEG solution for a given PCR product size (i.e., a less concentrated solution will retain smaller sized nucleic acids).

Of the methods described so far, the PEG precipitation method is the most inexpensive in terms of cost/sample, but it is substantially more labor intensive unless a high-throughput protocol for 96 sample PCR/sequencing plates is followed. Although the PEG 8000 method may seem like an old-fashioned molecular biology protocol, this method enjoys some major advantages over more modern kit-based methods of PCR purification and it plays an important role in the following PCR clean up method.

6.2.1.4 Solid-Phase Reversible Immobilization Beads

Hawkins et al. (1994) reported that DNA can reversibly bind to the surface of paramagnetic beads in a solution with a high concentration of PEG 8000 and salt. When these carboxyl-coated beads (1 μm diameter) are bound to DNA and are later placed close to a magnet, they can effectively isolate the target DNA, which is then washed, dried, and eluted. This method was termed *solid-phase reversible immobilization* or "*SPRI*" because the beads, which are the "solid phase," and the DNA can be first bound together and then unbound simply by exchanging buffer solutions (Hawkins et al. 1994). Using this principle DeAngelis et al. (1995) developed an SPRI-based method for cleaning PCR products.

The SPRI method is easily applied to 96 PCR products in a microplate. Typically, SPRI beads and PEG buffer are added at 1.8× the amount of PCR products (e.g., 36 μL SPRI beads/buffer + 20 μL PCR product). A synopsis of the procedure follows. First, the paramagnetic beads are brought to room temperature before they are mixed with PCR products and a 20% PEG 8000/2.5 M NaCl

buffer solution. During a brief incubation period the beads and DNA bind to each other in solution. The PCR plate is then placed onto a magnet, which "pulls" the bead/DNA complexes to the sides of the plate's wells nearest to the magnet. The supernatant, which still contains unused primers and dNTPs as well as primer dimers (if present), is then pipetted away and discarded. Next, the beads are washed twice with 80% ethanol (70% is often used as well). The beads and PCR products are allowed to air-dry for several minutes before the microplate is removed from the magnet. Water or elution buffer is added to the plate's wells, which liberates the PCR products from the beads. The microplate is then placed back onto the magnet, which immobilizes the beads thereby allowing the supernatant containing the purified PCR products to be saved for sequencing. Interested readers should consult the original step-by-step protocol in DeAngelis et al. (1995) for more details.

In order for the SPRI bead cleanup procedure to yield the best results it is essential to do the following. First, ensure that the beads are brought to room temperature and are fully resuspended (e.g., using a vortexer) before combining them with the PCR products. Second, use the pipette to thoroughly mix the beads and PCR products in order to obtain maximal binding efficiency. Third, use only freshly made ethanol in the wash steps (usually this will be at a 70% concentration). Fourth, while the beads are "on magnet," do not disturb them while adding the wash buffers; that is, carefully and gently pipette the liquid into each well and not directly onto the beads. Lastly, be sure that all ethanol has evaporated from the wells before adding the water or elution buffer otherwise the ethanol will inhibit subsequent enzymatic steps (e.g., PCR).

The SPRI method has many advantages. Hawkins et al. (1994) and DeAngelis et al. (1995) noted that SPRI-based methods are readily amenable to high-throughput sample workflows (i.e., highly automatable). Furthermore, there are no centrifugation or filtration steps and it is relatively inexpensive on a cost/sample basis. Another major advantage is that SPRI clean up of PCR products results in elimination of all small DNA molecules including primer dimers and other short nonspecific products (Lis and Schleif 1975; Lis 1980; Rosenthal et al. 1993; Quail et al. 2009).

Although the SPRI method was deemed early on to be an inexpensive method (DeAngelis et al. 1995), the high cost of commercially available SPRI beads has limited its use thus far especially given other alternatives such as Exo-SAP. However, Rohland and Reich (2012) showed that it was possible to prepare batches of "homemade" SPRI beads and PEG/NaCl buffer at low cost, which perform as well as AMPure XP beads (Agencourt, Beckman Coulter), the market standard (Faircloth 2012; Ford 2012; Rohland and Reich 2012). The high performance of homemade SPRI beads/buffer is not surprising given the early observations by Hawkins et al. (1994) that the DNA-binding efficiency is 100% when beads are in excess and remains high at around 80% following the washing steps. The Rohland-Reich (2012) SPRI preparation protocol, which was elaborated into a detailed step-by-step protocol by Faircloth and Glenn (2011), can be found on the following website: https://ethanomics.wordpress.com/2012/08/05/homemade-ampure-xp-beads/. Given the simplicity, effectiveness, and newfound availability of inexpensive SPRI beads/buffer, this method is certainly going to be used far more frequently in various phylogenomic methods (Faircloth 2012). Indeed, as we will see in Chapter 7, SPRI technology is being extensively used in next generation sequencing workflows.

6.2.1.5 Gel Purification of PCR Products

Agarose gel electrophoresis can also be used to isolate target PCR products, particularly those that are "messy" in that they contain many different nontarget products. Similar to the PEG 8000 method, though via a different mechanism, agarose electrophoresis can separate different DNA molecules by their length, which means unused primer, dNTPs, and nontarget PCR products can be physically separated from the target product. In this procedure, the entire PCR is first run through a 2% agarose gel. Following electrophoresis, the target bands are excised from the gel and the DNA fragments recovered using a gel purification kit (i.e., spin column format such as the QiaQuick kit, Qiagen). Although this method is quite effective at isolating the target DNA, it is not practical for general use owing to high costs in terms of purchasing gel purification kits, consumption of other lab supplies (e.g., plastics, agarose, etc.), and the amount of time the procedure requires.

6.2.1.6 *Which PCR Product Purification Method Is Best?*

If we look at Table 6.2 for a comparison of the aforementioned methods, we can immediately see why Exo-SAP has become the PCR purification method of choice: it is inexpensive and fast. The only drawback to Exo-SAP is that it is unable to destroy or remove small dsDNA artifacts such as primer dimers and PCR products that are less than 200 bp. However, if your target product is less than ~800 bp long, then you can overcome this problem simply by sequencing your product in both directions; that is, each sequence can fill in the correct base calls rendered ambiguous by the cosequencing of the artifacts. Given that the SPRI beads method is also quick and easy along with the relatively newfound protocol for making homemade SPRI beads and buffer, this method of PCR clean up now represents a highly attractive option especially when primer dimers are a concern and thus its use may surge in the future. Although some spin column and vacuum manifold kits can effectively remove these nonspecific artifacts, the cost per sample is far higher than using Exo-SAP or homemade SPRI beads and buffer. PEG 8000 and gel purification can effectively remove short PCR artifacts in a cost-effective manner but they are labor intensive compared to the other methods. For most routine PCRs, Exo-SAP followed by SPRI beads represent the best available options.

6.2.2 Setting Up Cycle Sequencing Reactions

Setting up cycle sequencing reactions is similar to setting up a PCR. You only need to combine purified PCR product template, sequencing primer, and a sequencing master mix containing dNTPs, dye terminators, buffer, and *Taq* polymerase. Note, that the concentration of the sequencing primer is usually less than the concentration of a PCR primer. In Chapter 5, we saw that a typical working concentration for a PCR primer is 10 μM, but for cycle sequencing it is usually in the 3–5 μM range. Some PCR or cycle sequencing protocols specify the primer concentration in units of picomoles/μL (pmol/μL). Or protocols may state that the total amount of sequencing primer should be X pmol per cycle sequencing reaction. Thus, it is useful to keep in mind the following conversion: 1 μM = 1 μmole/L = 1 pmol/μL. Cycle sequencing reactions can be performed in individual 0.2 mL plastic tubes or 96-well microplates. The sequencing master mix is sold by the manufacturer of the automated sequencer. Although some differences exist between thermocycling programs for PCR versus cycle sequencing, both involve the basic denaturation, annealing, and extension steps and a similar number of cycles (30–40 cycles). However, be sure to use the correct thermocycling program for a cycle sequencing reaction.

6.2.3 Purification of Extension Products

When the cycle sequencing reactions are completed, the extension products must be purified to remove unincorporated dye terminators and primers. Several methods are commonly used including ethanol (EtOH) precipitation and sephadex (G-50 grade) spin columns. The EtOH method is advantageous because of its lower cost per sample, however, the quality of sequence data tends to be more variable than using spin columns. Sephadex columns are more expensive

TABLE 6.2
Comparison of methods for purifying PCR products

Method of PCR purification	Reagent costs (per sample)	Labor cost (per sample)	Reduces or eliminates small dsDNA artifacts
Exo-SAP	Low	Low	No
Spin column	High	Low	Yes
Vacuum manifold	High	Low	Yes
PEG 8000	Low	High	Yes
SPRI beads[a]	Low	Low	Yes
Gel purification	High	High	Yes

[a] Indicates that the SPRI beads and associated PEG/NaCl buffer are homemade (see section 6.2.1.4).

than EtOH but this method consistently yields the best quality sequence data. Lab costs can be reduced if the sephadex powder is purchased separately and the plastic columns are reloaded for use. The sephadex method can be used for single tubes or in a 96 plate format. The final step in sample preparation prior to loading onto a capillary sequencer is resuspending the dried purified extension products with a resuspension or "sequencing" buffer.

6.2.4 Sequencing in a Capillary Sequencer: Do-It-Yourself or Outsource?

Once the extension products have been purified and resuspended in the appropriate sequencing buffer they are ready to be loaded into a capillary sequencer. After the sequencer door is shut, there is little more to do than specify a number of things on the desktop computer, which controls the sequencer. At the conclusion of a sequencing run, the computer creates a folder containing all sequence files (i.e., chromatograms). These machines are capable of sequencing 96 samples in less than a day.

Although capillary sequencers are easy to operate, they cost hundreds of thousands of US dollars (USD) to purchase plus demand yearly maintenance costs on the order of tens of thousands more USD. This means that many labs are simply unable to afford them. However, with the introduction of commercially available Sanger sequencing services starting in the early 2000s, this outsourcing option suddenly made the power of DNA sequencing available to anyone. This represented a key innovation because it liberated labs from having to purchase and maintain expensive sequencers and hire technicians to run and maintain the machines. In effect, a person could set up his or her own DNA lab in a small room for mere tens of thousands of dollars simply by creating a clean work environment equipped for DNA extractions and PCR; the PCR products can be sent by mail for sequencing with sequence data downloaded from the Internet. This is why today most DNA labs outsource their DNA sequencing needs—even sending their PCR products overseas to be sequenced. In addition to increased availability of DNA sequencing services, the cost per sample has dropped as well. For example, in 1997 the cost for one sequence, which included PCR cleanup, cycle sequencing, and running on a capillary sequencer, was approximately $16 USD, whereas by 2007 the price dropped down to at least $3 USD per sequence. If you choose to send your PCR products via international mail, then be sure to follow all rules and regulations for export and import of your samples as laws vary by country.

6.3 HIGH-THROUGHPUT SANGER SEQUENCING

In the DNA extraction and PCR chapters we discussed high-throughput strategies for increasing the number of samples that are processed at the same time. We will now consider two additional high-throughput strategies that can dramatically facilitate the sequencing of large numbers of PCR products and save money.

6.3.1 Sequencing 96 Samples on Microplates

Regardless whether you performed your PCRs in a single 96 sample plate or in individual tubes it is still worthwhile to submit your PCR products for sequencing on a 96-well microplate. This can be a particularly good strategy if you intend to outsource your samples for sequencing. Companies that offer a Sanger sequencing service often use robotic equipment to process samples for sequencing. Thus having your samples already arranged on a plate will greatly facilitate the handling of your samples. Note that usually "skirted" microplates are used in this case but check with the provider to be certain. The outsourcing sequencing option avails you with at least two choices. You can simply send a microplate full of unpurified PCR products, which their technicians can treat with Exo-SAP, perform the cycle sequencing reactions, and then run on a capillary sequencer. Alternatively, you can perform the PCR cleanup yourself then send a plate of cleaned PCR products for sequencing. In any case, sending your PCR products on a 96 well plate, instead of single tubes, should reward you with significant price discounts. A last important tip: before you send a plate by mail, be *extra careful* about sealing your plates—if in doubt, contact the sequencing lab for advice on which plate brands to use and proper shipping methods. Plates that are not properly sealed can leak during shipment resulting in partial or entire loss of your samples.

6.3.2 Adding Sequencing Primer "Tails" to PCR Primers

The plate-based approach to sequencing PCR products is easy and works well provided two conditions are satisfied: (1) the sequencing primer is known to work properly in the cycle sequencing reaction and (2) you use the same sequencing primer in all 96 cycle sequencing reactions that will be on the same plate. The first condition is satisfied only after a primer has been tested in a cycle sequencing reactions and proven to yield good quality sequences. Although perhaps the majority of typical PCR primers that function well in PCR may also work well in sequencing, this is not always the case. The second condition is easily satisfied if all the PCR products used the same pair of PCR primers. For example, if you are going to sequence 96 products with the forward PCR primer, then you can simply use a multichannel pipette to dispense the forward primer to all 96 cycle sequencing reactions. Alternatively, you can add the sequencing primer to a separate 96 well plate and send both the PCR and primer-containing plates to the sequencing provider (some sequencing providers even allow you to simply send the sequencing primer in a 1.5 mL microcentrifuge tube). However, if more than one pair of primers was used in the PCRs, then special care must be used to ensure that the correct sequencing primers are used in the cycle sequencing reactions. If you will be sending your PCR plate to a sequencing provider so that they can perform the cycle sequencing reactions, then you will need to prepare a special primer plate that has the sequencing primers placed correctly in the primer plate so that they can later be combined with their appropriate PCR templates in the sequencing reactions. If a given plate contains PCR products generated from many different primer pairs, then obviously this greatly complicates the process due to the primer plate preparation. Indeed, making a primer plate with multiple sequencing primers is a tedious and time-consuming task that can easily lead to errors that result in failed sequences.

To address these dilemmas, a nice trick was developed to make plate-based sequencing easy and fast. This method involves adding an actual *proven* sequencing primer to the 5′ end of a normal PCR primer at the time the primer is synthesized. This results in a compound or "tailed" primer that has a sequencing primer located at the 5′ half and a PCR primer located on the other (3′) half. Although in principle the sequencing primer could be any primer that is known to function without any problems during DNA sequencing, in practice most researchers have used so-called "M13 universal primers." M13 universal primers were originally developed for sequencing insert DNA in cloning vectors for shotgun DNA sequencing (Messing et al. 1981; Vieira and Messing 1982; Table 6.3). However, more recently these special sequencing primers have been co-opted for use as PCR product sequencing primers (e.g., Dinauer et al. 2000; Ivanova et al. 2007). Although any of the primers listed in Table 6.3 could be useful sequencing tails, it is recommended to carefully read Section 6.3.2.3, which explains potential pit

TABLE 6.3
List of M13 universal primers for DNA sequencing

Primer name	Sequence (5′ → 3′)
M13 forward (–20)	GTAAAACGACGGCCAGT
M13 forward (–21)	TGTAAAACGACGGCCAGT
M13 forward (–40)	GTTTTCCCAGTCACGAC
M13 forward (–41)	GGTTTTCCCAGTCACGAC
M13 reverse (–27)	CAGGAAACAGCTATGAC
M13 reverse (–29)	CAGGAAACAGCTATGACC
M13 reverse (–20)	GCGGATAACAATTTCACACAGG
M13 reverse (–49)	GAGCGGATAACAATTTCACACAGG

NOTE: Similar primers are grouped together and aligned to facilitate pairwise sequence comparisons. The original M13 universal sequencing primers are from the work of Messing et al. (1981) and Vieira and Messing (1982).

falls associated with using M13 tails and how to avoid them.

6.3.2.1 How an M13-Tailed Primer Functions in PCR

In order to better understand how an M13-tailed PCR primer can simplify Sanger sequencing, we must first consider how an M13-tailed primer functions in PCR. Recall the PCR primers we examined in Figure 5.7, which generated a 540 bp product. Let's say we wish to add the M13 forward (–21) and M13 reverse (–29) sequencing primers listed in Table 6.3 to these PCR primers. By convention the forward PCR primer will have an M13 forward (–21) tail in our example (Figure 6.7a), whereas the reverse PCR primer will have an M13 reverse (–29) tail (Figure 6.7b). Although adding such primer tails to proven PCR primers may lead to unwanted primer–primer interactions with the end result being poor PCR results, thankfully this seems to seldom occur.

Let's now look at what happens when these tailed primers are involved in synthesis during the extension phase of a PCR. Once the M13-tailed primers have annealed to their respective target templates (which were synthesized in some previous cycle), DNA polymerase begins synthesizing new strands of DNA starting at the 3′ end of the primers (Figure 6.8a). At the end of this extension phase, new double-stranded products have been made each of which is 576 bp long (Figure 6.8b). Thus when M13-tailed primers are used, the PCR should yield target products that are ~38–50 bp longer than products generated using the original (tail-less) PCR primers; the difference explained by the combined lengths of the forward and reverse tail primers. An important thing to notice is that not only have the entire M13-tailed primers been incorporated into each newly synthesized strand, but priming sites for the M13 sequencing primers are also present in the products, which will later be used during the cycle sequencing reaction. Accordingly, during the first several cycles of PCR when the reaction mixture is dominated by target templates located on genomic DNA, the true PCR primers (not the tails) drive the synthesis of new strands because the genomic templates naturally do not contain binding sites for the M13 tails. However, in later cycles when the mixture is dominated by synthetic target templates—which now contain the target templates for the original PCR primers + their attached M13-tails, the entire M13-tailed primer can now bind to the templates and drive synthesis.

6.3.2.2 Cycle Sequencing and M13 Primer Tails

From our earlier discussion of sequencing nontailed PCR products, remember that a nontailed PCR product will have, at each end of the double-stranded molecule, strings of sites that are complementary to the forward and reverse PCR primers (see Figures 5.7 and 6.2 for review). These primer-binding sites are then used in the cycle sequencing reaction when one of the PCR primers is used as a sequencing primer (see Figure 6.2). If M13-tailed primers are used to generate PCR products, then the double-stranded amplicons will have sites complementary to both the PCR primer and M13 primer tail, as is shown in Figure 6.9a. However, in the downstream

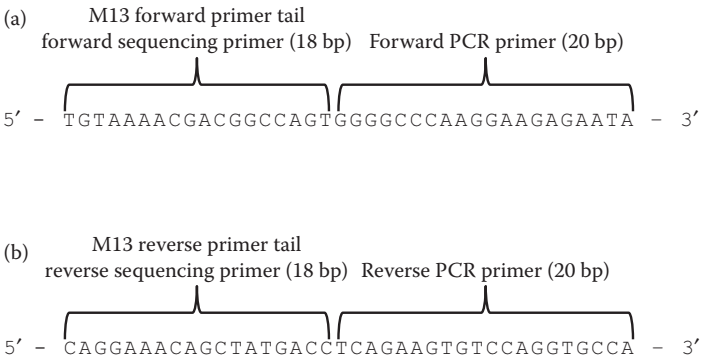

Figure 6.7. M13 universal primer tails added to normal PCR primers. (a) The forward sequencing ("tail") primer is added to the 5′ end of the forward PCR primer. (b) The reverse sequencing ("tail") primer is added to the 5′ end of the reverse PCR primer.

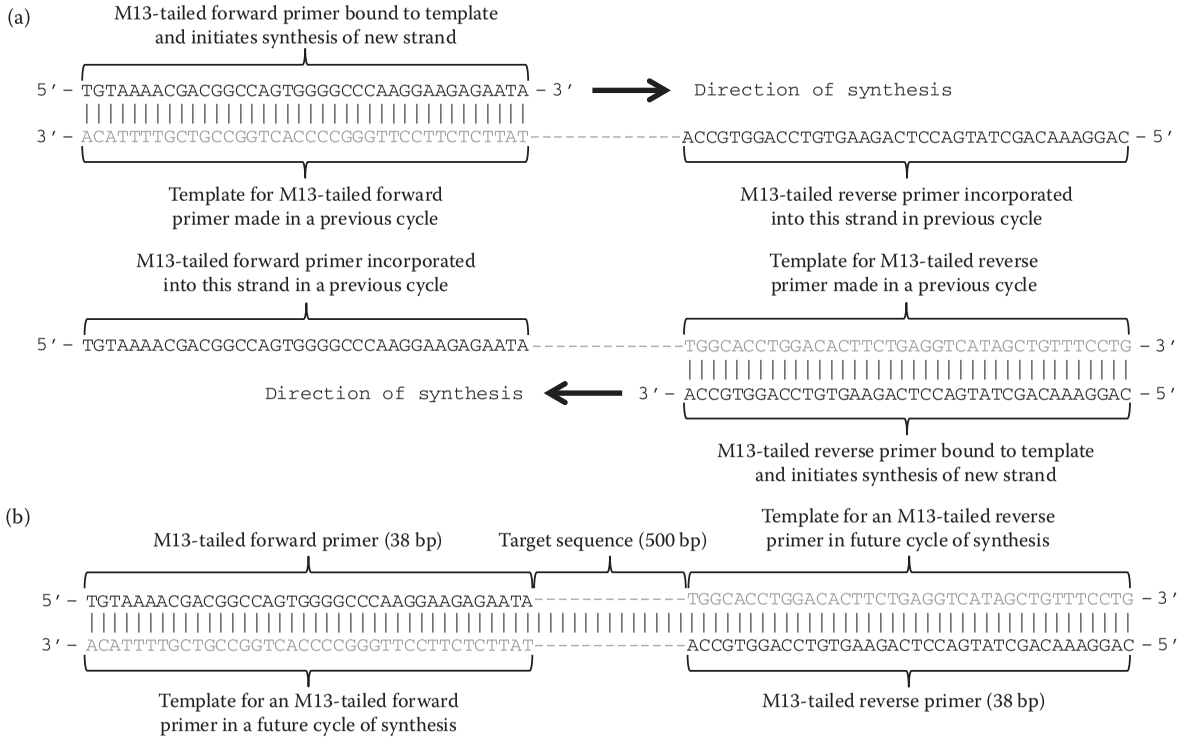

Figure 6.8. Synthesis and anatomy of a 576 bp PCR product using M13-tailed PCR primers. (a) Synthesis of two products from two complementary strands during the extension phase (after many cycles). (b) Anatomy of an M13-tailed PCR product. For clarity the target sequence is indicated as dashes. Note, most of target sequence is not shown due to lack of space in the figure. Vertical bars show hydrogen bonding between complementary bases. Gray represents strands of DNA synthesized during PCR and the M13-tailed forward and reverse primers are black. PCR primers are same as shown in Figure 6.7.

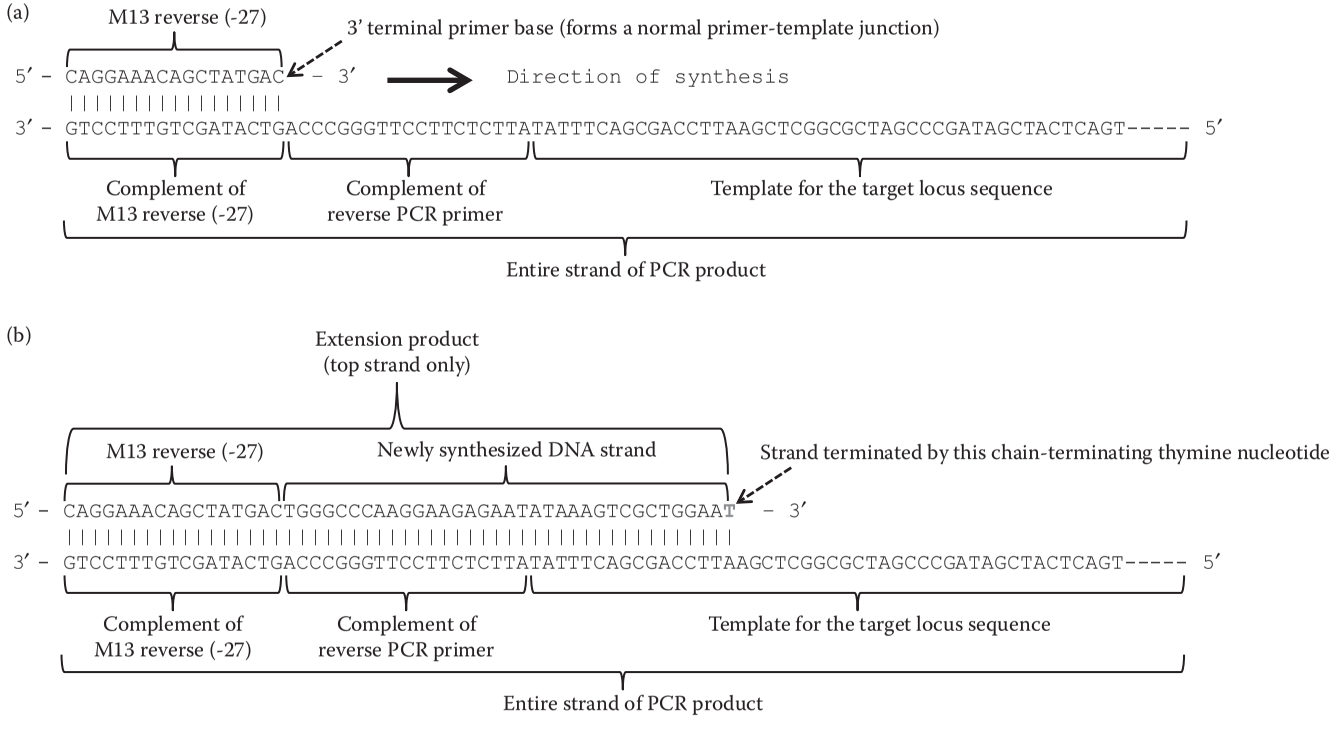

Figure 6.9. Synthesis of an extension product using an M13 universal primer. (a) An M13 reverse (–27) sequencing primer anneals to its complementary sequence on one strand of the PCR product. The match between primer and template is perfect because the same M13 primer was used as a "tail" in the PCR. Notice that the 3′ terminal base of the M13 primer forms a normal primer–template junction from which *Taq* polymerase can start synthesizing the extension product. (b) Formation of the extension product is completed once a ddNTP becomes incorporated into the new strand. Here, a chain-terminating thymine (shown in in gray) abruptly stopped synthesis.

cycle sequencing reaction, only the correct M13 primer should be used as the sequencing primer (e.g., Figure 6.9a). An interesting consequence of sequencing with an M13 primer is that the extension product will include the actual PCR primer sequence ahead of the target sequence (Figure 6.9b), and hence a chromatogram of this sequence may show part or all of the primer at the start of the sequence. Other than this difference between nontailed and tailed PCR products, the process of generating extension products is the same. Thus when the M13 sequencing primer anneals to its appropriate location on the PCR template strand (Figure 6.9a), *Taq* polymerase will start extending the new strand using dNTPs as substrates until one of the four possible ddNTPs (a ddTTP in this case) becomes incorporated at the 3′ end of the strand (Figure 6.9b). It is critical to use forward and reverse M13 tails for the forward and reverse PCR primers, respectively or only a single tailed primer per PCR. If instead the same M13 tail is used on both forward and reverse PCR primers, then two conflicting sets of extension products will be in the cycle sequencing reaction with the result being a ruined sequence.

6.3.2.3 *On the Importance of Matching Sequencing Primers*

As we saw in Table 6.3 there are a variety of M13 sequencing primers that could be used as PCR primer tails. Note that some of these sequencing primers only differ by one base such as the M13 forward (–20) and M13 forward (–21) primers. Although you may have some flexibility in which M13 primers you use as PCR primer tails, the important thing is for you to *exactly* match your M13 tails to the sequencing primers that you will use to sequence those same products. If you choose one type of M13 primer to be a PCR primer tail and later sequence your PCR products with a different M13 primer, then you run the risk of having a failed sequencing reaction due to mismatches between the PCR product and sequencing primer. Even sequencing primers that only differ by one base from your PCR tail primer *could* be the difference between success and failure. Needless to say, having all 96 of your sequences fail is a costly mistake to make. We will now take a closer look at the consequences of using a PCR primer tail that differs from the sequencing primer.

Let's say a researcher used an M13 reverse (–29) primer as a tail attached to a reverse PCR primer, but then used an M13 reverse (–27) primer in the cycle sequencing reaction to sequence this product. Notice in Table 6.3 that these two M13 primers only differ by a single base at the 3′ end. In this particular scenario, which is depicted in Figure 6.10a, despite the sequencing primer being one base shorter than the stretch of complementary sites created by the M13 reverse (–29) primer, a normal primer–template junction is formed. Assuming the sequencing primer is still able to firmly hybridize to the binding sites, the DNA polymerases in the reaction should have no problem generating extension products and thus the reaction is expected to proceed in normal fashion.

Now let's see what happens if we change the scenario by reversing the aforementioned tail and sequencing primers. As you can see in Figure 6.10b, the M13 reverse (–29) sequencing primer is one base longer at the 3′ end than the stretch of complementary binding sites generated by the M13 reverse (–27) PCR tail. Because the 3′ terminal primer base is a cytosine base, which is not complementary to the adenine base on the PCR template strand (Figure 6.10b), this creates a rather unfortunate mismatch. This is probably one of the worst kinds of mismatches at the primer–template junction because, as noted by Palumbi (1996), cytosine cannot form any hydrogen bonds with adenine. The consequence of this mismatch will be severe because the efficiency of primer extension by the DNA polymerases will be so low that the end result will be failed sequences (i.e., useless chromatograms). I have observed this same type of failure before. To confirm that the problem was indeed the wrong sequencing primer, the same plate of M13 reverse (–27) tailed PCR products was sequenced again but this time using the same M13 primer. Nearly all 96 of the resulting sequences were high quality in the second sequencing attempt thus verifying the cause of the earlier failure.

In the scenario we just looked at, the abnormal primer–template junction was due to a mismatch between the 3′ terminal primer base and the first template base. What if by chance the first template base was a guanine instead of an adenine? As you can see in Figure 6.10c, in this scenario the primer–template junction would of course be normal because of the Watson–Crick base pairing

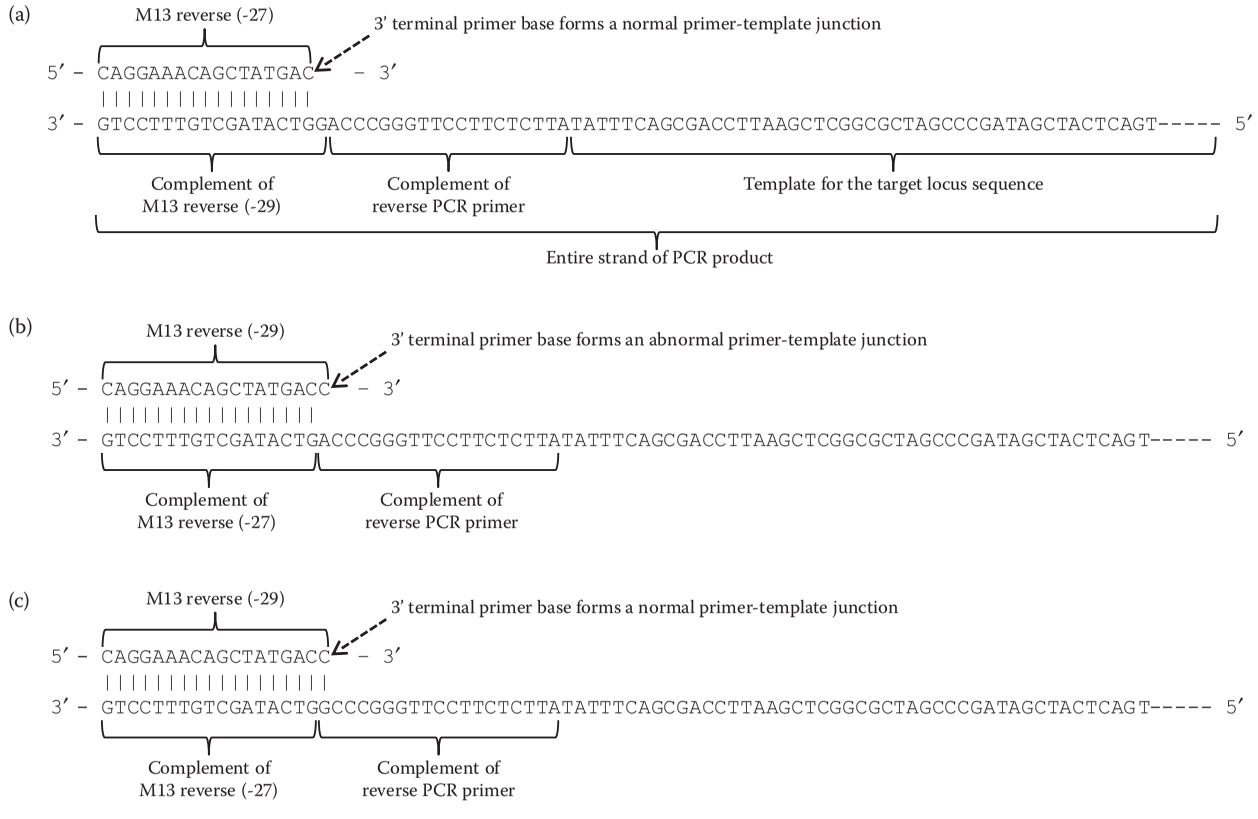

Figure 6.10. Consequences of not matching M13 primers in the PCR and cycle sequencing reactions. (a) Using an M13 reverse (–27) sequencing primer to sequence a PCR product containing an M13 reverse (–29) primer tail. Despite the sequencing primer having one less base at its 3′ end than the complement to the other M13 primer, a normal primer–template junction can form. (b) Using an M13 reverse (–29) sequencing primer to sequence a PCR product containing an M13 reverse (–27) primer tail. An abnormal primer–template junction arises between the cytosine base located at the 3′ terminal end of the primer and the adjacent adenine base in the template base position. (c) Same as previous but the template base is now shown as a guanine base. In this scenario, a normal primer–template junction can form. Note, the 5′ end of the PCR template strands are truncated (at the dashes) due to space limitation in the figure. Vertical dashes represent hydrogen bonding between complementary base pairs and M13 universal primers shown are from Table 6.3.

between cytosine and guanine. Thus, a researcher who uses a sequencing primer that differs from the tailed primer could still obtain excellent sequencing results but this would be by dumb luck. If the same researcher changes PCR primers and repeats the same tail-sequencing primer mismatch, then it is likely that eventually a major sequencing failure would occur. Note also that a primer tail and sequencing primer can differ from each other at the 5′ end and still produce the desired sequences provided the correct primer–template junction is formed at their 3′ ends.

The critical message here is to use the exact same M13 primers for both tailing your PCR primers and as the sequencing primers in order to avoid sequencing problems. If you are outsourcing your Sanger sequencing and you request that the sequencing provider uses their stock of M13 sequencing primers to sequence the PCR products, then it will be your responsibility to ensure that the M13 sequencing primers they use will exactly match your M13-tailed PCR products.

6.3.2.4 Benefits of Using M13-Tailed Primers

There are several advantages to using M13-tailed PCR products. First, M13 primers are proven to sequence well, whereas not all PCR primers sequence well even if they perform well in PCR. Secondly, PCR products that were generated using different pairs of PCR primers (i.e., multilocus study) can all be sequenced together on the same microplate as long as they have the same primer tails. Thirdly, if outsourcing an entire microplate of tailed PCR products, then many commercial sequencing facilities will supply the M13 sequencing *free of charge* thus saving you the expense of purchasing your own M13 primers. Lastly, recall that earlier in this chapter we saw that the first 10–30 bases of a Sanger sequencing chromatogram typically are of low quality and not reliable. However, if we instead sequence a tailed PCR product, then the extension products from the cycle sequencing reaction will represent the sequence of the PCR primer and not the locus of interest. Thus, the poor quality base calls in the first part of a chromatogram will represent primer sequence (which is unimportant) and thus the target sequence will likely be represented by higher quality base calls. This means that for PCR products that are less than about 700 bp and free of primer dimers and other coamplified products, it may be possible to obtain high quality sequences from just one sequence read per PCR product.

6.4 HAPLOTYPE DETERMINATION FROM SANGER SEQUENCE DATA

Obtaining DNA sequence data from haploid loci such as mitochondrial genes using the Sanger methodology is straightforward. However, Sanger sequencing of diploid (or worse, polyploid loci) presents a serious complication to the investigator. What is this complication? First, consider the simple case of sequencing any haploid locus such as a mitochondrial locus. In this scenario the output chromatograms (from forward or reverse sequencing primers) should show high-quality single peaks for each nucleotide site in the locus as in Figure 6.5. Now think about sequencing any diploid autosomal locus. If the individual being sequenced happens to be homozygous at all sites in this locus, then the resulting chromatogram should resemble the basic form observed in Figure 6.5. However, what if the individual is instead heterozygous for any of the sites? What are the implications for Sanger sequencing in such a scenario? Let's now look at this important issue in detail.

6.4.1 PCR Amplification and Sanger Sequencing of Diploid or Polyploid Loci

When a haploid locus from the mitochondrial genome or human Y-chromosome is amplified via PCR, millions of copies of a *single* allele or haplotype are made. This assumes, of course, that paralogous copies are not coamplified with the orthologous haplotype. However, it is important to realize that when a diploid or polyploid locus is similarly amplified, *all* orthologous haplotypes will likely be copied. This means that copies of *multiple haplotypes will be produced in a single PCR* (Clark 1990).

For example, let's consider two different human autosomal loci that were PCR-amplified in two separate reaction tubes. Again we will consider the simplest scenario in which no paralogous copies of these loci occur elsewhere in the genome. Note that without reference to comparable sequence data from the parents, we cannot hope to ascertain which alleles in our PCR tubes were inherited from the mother and which were

from the father but these details are unimportant. However, given that each cell contains one genomic copy from each parent, we can assume that the two *parental haplotypes* each account for approximately 50% of the total PCR products in each reaction tube (Figure 6.11). In our example, the parental haplotypes for each locus are heterozygous but in different ways. In the first locus, the two haplotypes differ only by a single SNP site but are otherwise the same length (Figure 6.11a), whereas in the second locus the two haplotypes differ by their length due to the presence of an indel (Figure 6.11b). While it is generally not a problem to amplify each parental haplotype for a given locus in a PCR, such PCR products do present additional complications when they are sequenced via the Sanger method.

Recall from Chapter 5 we considered the simplest case of PCR and Sanger sequencing, which is to amplify and sequence a haploid locus such as a mitochondrial or human Y-chromosome locus. When such loci are sequenced, you expect to see chromatograms consisting of only single and well-defined peaks. Likewise, a diploid locus that is homozygous in an individual should also generate a chromatogram exhibiting only single peaks even though there were, in actuality, *two different sets of extension products* produced—one set from each parental haplotype. Such an occurrence is not a concern for us because the two parental haplotypes are the same.

Now let's return to our earlier example from Figure 6.11 and consider what would happen if we sequence each of the heterozygous PCR products. When the PCR products from Figure 6.11a are cycle sequenced we would expect the reaction to generate two sets of extension products reflecting each of the two parental haplotypes in a 50:50 manner as shown in Figure 6.12. As these extension products become sorted by their size in the capillary sequencer, the products of equal length will each have their terminal (dye terminator) nucleotide "read" at the same time leading to a base call for that site on the chromatogram. All homozygous extension products will have matching signals at each site and therefore only single peaks will appear on the chromatogram. However, for the extension products with the polymorphic site (#380) located at the terminal labeled nucleotides, two overlapping peaks (green and black) will show on the chromatogram at that position reflecting the 50:50 mixture of terminally located adenine and guanine bases (Figure 6.12). In this situation, the computer may specify the base reflecting the taller of the two peaks, as was the case here, or it may instead call an "N" for unknown base. We thus see that sequencing a PCR product that is heterozygous for a SNP site can still

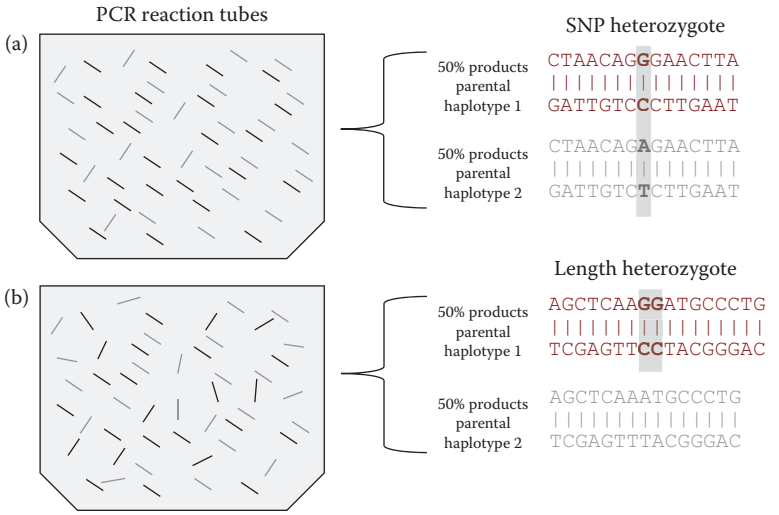

Figure 6.11. PCR of a diploid locus results in a mixture of parental products. (a) PCR involving a locus heterozygous for a single SNP site. One parental product (black) has a G:C base pair while the other parental product (light gray) has an A:T base pair. (b) PCR on a locus heterozygous for a 2-bp indel. The black parental product contains two consecutive G:C base pairs, which are missing from the light gray parental product. Polymorphic sites are shaded in dark gray and vertical dashes represent hydrogen bonding between complementary bases.

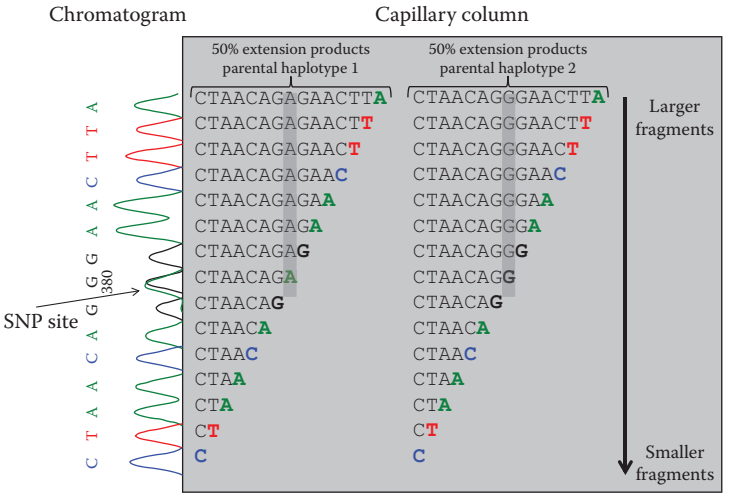

Figure 6.12. Sanger sequencing of a PCR product that is heterozygous at a single SNP site. The dark shaded rectangle shows the SNP site and represents the only polymorphic site on this sequence. SNP is located at site #380 on the adjacent chromatogram where a green and black double peak is located. A guanine base (black peak) was called but an adenine base (green peak) is also evident.

produce a high quality chromatogram. Indeed, even though quality scores for these heterozygous sites are usually low, the two overlapping peaks are usually distinct enough that the existence of each allelic base is not questionable.

When evaluating chromatograms generated for diploid loci one does not expect to see a large number of SNPs clumped together; instead they tend to be dispersed as single polymorphic sites that occur every tens or hundreds of sites along a sequence. Thus, if you observe two, three, or more double-peaks that occur on successive sites, then most likely they are not SNPs, but instead are artifacts of poor quality sequence. This is often observed on one chromatogram based on one sequencing primer, while the chromatogram containing the sequence generated from the other sequencing primer shows only sharp and distinct single peaks in the locations of the double-peaks seen on the other chromatogram.

A cycle sequencing reaction of the PCR products from Figure 6.11b will likewise produce two sets of extension products, which will be sorted by their length in the capillary sequencer (Figure 6.13). Notice that the extension products representing each parental haplotype are homozygous for the first seven sites (beginning with the smallest fragments) resulting in the formation of single peaks and unambiguous base calls on the chromatogram (Figure 6.13). However, most of the sites after the seventh site (#231) have double peaks and about half of the bases were called as "N." It might be tempting to explain the double peaks in this chromatogram as multiple heterozygous SNP sites but this would be unlikely because SNPs are usually infrequently encountered in sequences. This chromatogram pattern showing a number of high-quality single peaks suddenly changing to a long series of double peaks is the hallmark of an indel present in one parental haplotype but not in the other (i.e., a length heterozygote). If we compare the extension products for the two parental haplotypes, we see that the indel—a G–G present in the first parental haplotype, but not in the second parental haplotype—has shifted the two haplotypes out of alignment so that the extension products are out of phase with each other for the latter half of the sequence. If you imagine the G–G being removed from the first haplotype then the two sequences would be homozygous and a clean chromatogram with single peaks would result. How abundant are indels? For any given Sanger chromatogram of a diploid locus at most you are likely to see only one or two separate indels. A sequence read will be most affected if an indel is present early in the chromatogram. If this is the case, one strategy to recovering the sequence is to sequence the PCR product using the other primer. We have just learned that it is quite possible to obtain good quality PCR products (i.e., faithful copies of both parental haplotypes) but, if they are heterozygous for one or more indels, then these

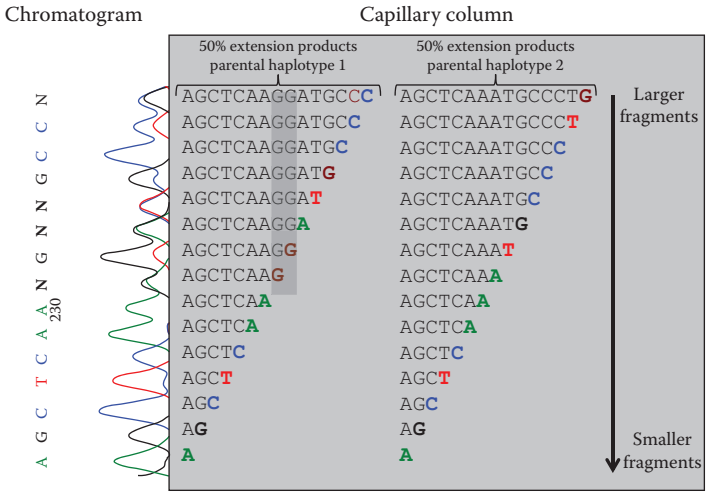

Figure 6.13. Sanger sequencing of a PCR product that is heterozygous for a 2-bp indel. Dark gray is the indel shown as two consecutive guanine bases present in the first haplotype but not the second.

products will not sequence well owing to current limitations of Sanger sequencing. Fortunately, newer NGS methods are able to circumvent this problem.

6.4.2 Multiple Heterozygous SNP Sites and Haplotype Sequences

When a single heterozygous SNP site or *singleton* is observed on a chromatogram, then the two parental haplotypes can be unambiguously determined. For example, in Figure 6.14 we see the sequence for a diploid PCR product that was recorded from a chromatogram (not shown). The chromatogram contained one site with a double peak corresponding to a guanine base and an adenine base. The presence of this singleton does not present a problem because the two parental haplotypes can be resolved with certainty—each is identical at every site except for the SNP site, which has an adenine base on one haplotype and a guanine base on the other. However, for some types of loci (e.g., anonymous loci) it is not uncommon for a chromatogram spanning hundreds of bases showing evidence of two or more heterozygous SNP sites. In these cases, the haplotypes from such PCR products cannot be determined without using outside information (Clark 1990). Figure 6.15 shows a case with two heterozygous SNP sites on the same sequence indicated by two sites that exhibited two overlapping and differently colored peaks. The base calls from the chromatogram (not shown) showed a heterozygous SNP site indicating that one haplotype has an adenine base (green peak) while the other has a guanine base (black peak). The second SNP site similarly showed two double peaks with one corresponding to a thymine base (red peak) and the

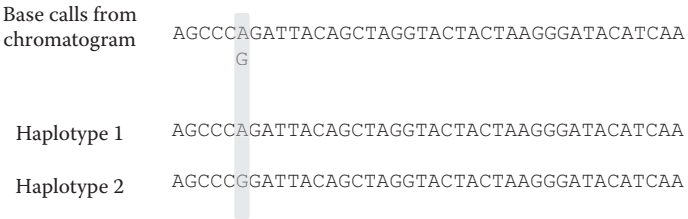

Figure 6.14. Haplotype determination from a Sanger sequence containing a "singleton." Sequence data were obtained from a chromatogram of a directly sequenced PCR product that was heterozygous for a single SNP. The top sequence, which was read directly from the base calls of a chromatogram (not shown), showed a single "double peak" indicating that both adenine and guanine bases were present at the same site. Because this is a singleton SNP, both haplotype sequences are easily resolved. The singleton and corresponding distinct haplotype bases are shaded by the dark gray rectangle.

Figure 6.15. Haplotype determination from a Sanger sequence containing multiple heterozygous SNPs. The top sequence was obtained from a chromatogram of a directly sequenced PCR product that was heterozygous for two sites. One SNP is evident as an adenine and guanine double peak while the second SNP site was shown by a thymine and cytosine double peak. There are four possible haplotype sequences but only two are correct. In the absence of outside information it is impossible to unambiguously resolve the haplotype sequences.

other a cytosine base (blue peak). With no other information we cannot determine with certainty the two haplotype sequences because there are *four different possible haplotypes* that can be inferred from these data yet only two of them are correct (Figure 6.15). This problem becomes more complicated with an increasing number of heterozygous SNP sites: three SNPs = eight possible haplotypes; four SNPs = 16 possible haplotypes; and so on.

In order to circumvent this problem, it has been common practice by researchers to simply label each heterozygous SNP site with the appropriate IUPAC ambiguity code (Table 6.4) or eliminate those problematic sites from the dataset. However, doing this reduces the information content in the sequences, which can adversely impact downstream analyses especially in phylogeographic studies. As we will see in Section 6.4.3, methods are available that allow for recovery of the complete haplotype sequences from PCR products that are heterozygous for SNPs or indels.

6.4.3 Methods for Obtaining Nuclear Haplotype Sequences from Sanger Sequence Data

A number of methods have been developed to allow investigators to "resolve" or "phase" PCR haplotypes from Sanger sequence data. These methods fall into two general categories: (1) using laboratory procedures to physically isolate each of the unique amplicons *prior* to Sanger sequencing and (2) using bioinformatics algorithms to infer each PCR haplotype from a directly sequenced PCR product. Although the physical separation methods are capable of generating high-quality haplotype sequences, their laborious time-consuming nature presents a significant obstacle to obtaining large phylogenomic datasets. Thankfully, bioinformatic approaches

TABLE 6.4
IUPAC nucleic acid codes by the Nomenclature Committee of the International Union of Biochemistry

Bases	Symbols	IUPAC code
Adenine	A	A
Guanine	G	G
Cytosine	C	C
Thymine	T	T
Adenine or guanine (purine)	A or G	R
Cytosine or thymine (pyrimidine)	C or T	Y
Cytosine or adenine	C or A	M
Thymine or guanine	T or G	K
Thymine or adenine	T or A	W
Cytosine or guanine	C or G	S
Cytosine, thymine, or guanine	C, T, or G	B
Adenine, thymine, or guanine	A, T, or G	D
Adenine, thymine, or cytosine	A, T, or C	H
Adenine, cytosine, or guanine	A, C, or G	V
Adenine, cytosine, guanine, or thymine	A, C, G, or T	N

SOURCE: From Moss, G. P. 1984. Nomenclature for incompletely specified bases in nucleic acid sequences. http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html (retrieved March 2, 2015).

provide a superior way to efficiently resolve PCR haplotype sequences.

6.4.3.1 Physical Isolation of PCR Haplotypes prior to Sequencing

Single-Strand Conformational Polymorphism—A variety of laboratory methods have been used to isolate the two parental haplotypes (when they are heterozygous) in a completed PCR mixture. One early method that has been successfully used is called *single-strand conformational polymorphism* or "SSCP" (Friesen 2000; Sunnucks et al. 2000). In brief, PCR product amplicons are first converted into single-stranded molecules via heat denaturation then are electrophoretically run through a "native" polyacrylamide gel. A native gel is simply a gel that does not contain any denaturant chemicals (e.g., those that are found in sequencing gels). Under these conditions single-stranded DNA has a tendency to form secondary structures such as loops and hairpins, which, in turn, influences their mobility through a gel. Heterozygous haplotypes usually form different secondary structures and hence each haplotype travels through the gel at a different rate. Thus, rather than separating DNA molecules by length, as is often done, amplicons consisting of heterozygous haplotypes are separated primarily by their two-dimensional secondary structures. At the conclusion of electrophoresis, the individual gel bands containing each amplicon haplotype are excised and sequenced. Although this method may well produce the best results of all haplotype resolving methods, a huge drawback is the large labor cost and use of hazardous chemicals.

PCR Cloning—Another method that has been successfully used for isolating amplicon haplotypes is PCR cloning, which is also called "TA cloning" (Jennings and Edwards 2005). In brief, PCR products are first "A-tailed" meaning that a single adenine is added to the 3′ ends of each double-stranded amplicon. This is accomplished by combining the PCR products from a single reaction with *Taq* polymerase and dATPs then incubating the mixture in a thermocycler at 70°C for about 10 minutes. The A-tailed amplicons are then efficiently ligated into plasmid cloning "T-vectors," which are plasmids having single thymine overhangs on the 3′ ends of the two vector strands. The recircularized plasmids are then transformed into *E. coli* bacteria and grown overnight into individual clone-colonies (i.e., a single bacterium containing a plasmid eventually gives rise to millions of other bacteria each containing the identical plasmid). The colonies containing recombinant plasmids are identified via blue-white screening and then the plasmids are isolated away from the individuals in each colony. The purified plasmids are then used as sequencing templates and vector sequencing primers (e.g., M13 universal primers) are used to sequence the inserted haplotype amplicons. One advantage of this method is that since each haplotype amplicon is separately sequenced, one does not have to be concerned about resolving haplotypes or about having indels wrecking all or part of a sequence (see Figure 6.2 in Jennings and Edwards 2005). If enough clones are sequenced, then paralogous loci can also be detected. Although TA cloning kits are available for this purpose, the amount of labor involved makes this an impractical method for generating large phylogenomic datasets especially with superior NGS-based approaches that are now available (Chapter 7).

6.4.3.2 Statistical Inference of Haplotypes from Sanger Sequence Data

Clark (1990) recognized the problem of inferring PCR haplotype sequences for diploid loci using Sanger sequence data. Moreover, he also suggested ways that haplotype sequences could be inferred. By locating sequences containing "singletons," a number of different haplotypes can be quickly identified. Using these known haplotypes, a process of subtraction can be used in which the known haplotypes are used to help resolve remaining ambiguous haplotypes. This method, which is done by the "eye" can work reasonably well provided a sufficient number of individuals have been sequenced and the number of polymorphic sites is not too large. Once the number of polymorphic sites becomes too large, the process of elucidating haplotypes by the eye becomes unwieldy. However, algorithmic methods for statistically inferring haplotypes from Sanger data have made this bioinformatics approach very popular. Stephens et al. (2001) developed such a Bayesian program, which they called PHASE, for inferring haplotypes from Sanger data. Other bioinformatics software developed to compliment PHASE, such as SEQPHASE (Flot 2010), have

further enhanced the user-friendliness of this bioinformatics approach to phasing haplotypes. Recent examples of studies using this methodology include Reilly et al. (2012) and Gottscho et al. (2014). However, one problem this bioinformatics method cannot effectively solve is the indel problem. If only a single heterozygous indel is observed in a chromatogram, then it might be possible to elucidate the downstream "frame-shifted" haplotypes using some software programs that allow for editing chromatograms. For example, Lee and Edwards (2008) used this approach. However, if multiple heterozygous indels are present in one PCR product, then there is nothing that can be done to fix the sequence—it will likely be unusable. Another potential problem with the bioinformatics method is that incorrectly inferred haplotypes will artificially recombine the two haplotypes, which may lead the investigator to incorrectly conclude that one or more historical recombination events occurred and not realize that the cause was bioinformatics error. In other words, false positive results might be obtained from recombination tests such as the four gamete test (Hudson and Kaplan 1985; Chapter 3). Nonetheless, the bioinformatics method for inferring haplotypes is still the best available method for inferring haplotypes from Sanger sequence data. For additional discussion about the use of PHASE in population studies see Garrick et al. (2010).

REFERENCES

Atkinson, M. R., M. P. Deutscher, A. Kornberg, A. F. Russell, and J. G. Moffatt. 1969. Enzymatic synthesis of deoxyribonucleic acid. XXXIV. Termination of chain growth by a 2′,3′ dideoxyribonucleotide. *Biochemistry* 8:4897–4904.

Carothers, A. M., G. Urlaub, J. Mucha, D. Grunberger, and L. A. Chasin. 1989. Point mutation analysis in a mammalian gene: Rapid preparation of total RNA, PCR amplification of cDNA, and Taq sequencing by a novel method. *BioTechniques* 7:494–496.

Clark, A. G. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol* 7:111–122.

DeAngelis, M. M., D. G. Wang, and T. L. Hawkins. 1995. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res* 23:4742.

Dinauer, D. M., R. A. Luhm, A. J. Uzhiris, D. D. Eckels, and M. J. Hessner. 2000. Sequence-based typing of HLA class II DQB1. *Tissue Antigens* 55:364–368.

Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8:186–194.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res* 8:175–185.

Faircloth, B. 2012. A penny for your method: AMPure Substitute. The Molecular Ecologist blog. http://www.molecularecologist.com/2012/01/a-penny-for-your-method-ampure-substitute/ (accessed May 11, 2016).

Faircloth, B. and T. Glenn. 2011. Serapure. Unpublished protocol based on the original Rohland and Reich (2012) homemade SPRI beads protocol (see also the Ford 2012 reference).

Flot, J. F. 2010. SeqPHASE: A web tool for interconverting PHASE input/output files and FASTA sequence alignments. *Mol Ecol Res* 10:162–166.

Ford, E. 2012. Homemade AMPure XP beads. Ethanomics, Everything Ethan knows about biology—the Ethan-ome. https://ethanomics.wordpress.com/2012/08/05/homemade-ampure-xp-beads/ (accessed May 16, 2016).

Friesen, V. L. 2000. Introns. In *Molecular Methods in Ecology*, ed. A. J. Baker, 274–294. Oxford: Blackwell.

Garrick, R. C., P. Sunnucks, and R. J. Dyer. 2010. Nuclear gene phylogeography using PHASE: Dealing with unresolved genotypes, lost alleles, and systematic bias in parameter estimation. *BMC Evol Biol* 10:118.

Gottscho, A. D., S. B. Marks, and W. B. Jennings. 2014. Speciation, population structure, and demographic history of the Mojave Fringe-toed Lizard (*Uma scoparia*), a species of conservation concern. *Ecol Evol* 4:2546–2562.

Hawkins, T. L., T. O'Connor-Morin, A. Roy, and C. Santillan. 1994. DNA purification and isolation using a solid-phase. *Nucleic Acids Res* 22:4543.

Hillis, D. M., B. K. Mable, A. Larson, S. K. Davis, and E. A. Zimmer. 1996. Chapter 9. Nucleic acids IV: Sequencing and cloning. In *Molecular Systematics*, 2nd edition, eds. D. M. Hillis, C. Moritz, and B. K. Mable, 321–381. Sunderland: Sinauer.

Hudson, R. R. and N. L. Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111:147–164.

Innis, M. A., K. B. Myambo, D. H. Gelfand, and M. A. Brow. 1988. DNA sequencing with *Thermus aquaticus* DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. *Proc Natl Acad Sci USA* 85:9436–9440.

Ivanova, N. V., T. S. Zemlak, R. H. Hanner, and P. D. N. Hebert. 2007. Universal primer cocktails for fish DNA barcoding. *Mol Ecol Notes* 7:544–548.

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

Lee, J. Y. and S. V. Edwards. 2008. Divergence across Australia's Carpentarian Barrier: Statistical phylogeography of the Red-backed Fairy Wren (*Malurus melanocephalus*). *Evolution* 62:3117–3134.

Lis, J. T. 1980. Fractionation of DNA fragments by polyethylene glycol induced precipitation. *Methods Enzymol* 65:347–353.

Lis, J. T. and R. Schleif. 1975. Size fractionation of double-stranded DNA by precipitation with polyethylene glycol. *Nucleic Acids Res* 2:383–389.

Maxam, A. M. and W. Gilbert. 1977. A new method for sequencing DNA. *Proc Natl Acad Sci USA* 74:560–564.

Messing, J., R. Crea, and P. H. Seeburg. 1981. A system for shotgun DNA sequencing. *Nucleic Acids Res* 9:309–321.

Moss, G. P. 1984. Nomenclature for incompletely specified bases in nucleic acid sequences. http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html (retrieved March 2, 2015).

Murray, V. 1989. Improved double-stranded DNA sequencing using the linear polymerase reaction. *Nucleic Acids Res* 17:8889.

Palumbi, S. R. 1996. Chapter 7. Nucleic acids II: The polymerase chain reaction. In *Molecular Systematics*, 2nd edition, eds. D. M. Hillis, C. Moritz, and B. K. Mable, 205–247. Sunderland: Sinauer.

Prober, J. M., G. L. Trainor, R. J. Dam et al. 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 238:336–341.

Quail, M. A., H. Swerdlow, and D. J. Turner. 2009. Improved protocols for the Illumina genome analyzer sequencing system. *Curr Protoc Hum Genet* 18–2.

Reilly, S. B., S. B. Marks, and W. B. Jennings. 2012. Defining evolutionary boundaries across parapatric ecomorphs of Black Salamanders (*Aneides flavipunctatus*) with conservation implications. *Mol Ecol* 21:5745–5761.

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

Rosenthal, A., O. Coutelle, and M. Craxton. 1993. Large-scale production of DNA sequencing templates by microtitre format PCR. *Nucleic Acids Res* 21:173–174.

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

Smith, L. M., S. Fung, M. W. Hunkapiller, T. J. Hunkapiller, and L. E. Hood. 1985. The synthesis of oligonucleotides containing an aliphatic amino group at the 5′ terminus: Synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res* 13:2399–2412.

Smith, L. M., J. Z. Sanders, R. J. Kaiser et al. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature* 321:674–679.

Stephens, M., N. J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68:978–989.

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

Vieira, J. and J. Messing. 1982. The pUC plasmids: An M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* 19:259–268.

Watson, J. D., T. A. Baker, S. P. Bell, A. Gann, M. Levine, and R. Losick. 2014. *Molecular Biology of the Gene*, 7th edition. New York: Pearson Education, Inc.

Werle, E., C. Schneider, M. Renner, M. Völker, and W. Fiehn. 1994. Convenient single-step, one tube purification of PCR products for direct sequencing. *Nucleic Acids Res* 22:4354–4356.

Zimmerman, S. B. and B. H. Pheiffer. 1983. Macromolecular crowding allows blunt-end ligation by DNA ligases from rat liver or *Escherichia coli*. *Proc Natl Acad Sci USA* 80:5852–5856.