CHAPTER EIGHT

Developing DNA Sequence Loci

Once a researcher decides which type of locus or loci will be needed for a phylogenomic study, the next problem to solve is how to begin the process of generating a DNA sequence dataset. Many types of DNA sequence loci are PCR-based, which of course means they require pairs of DNA primers to target particular genomic regions for sequencing. As we saw in Chapters 6 and 7, PCR products can be used as input templates for Sanger sequencing or NGS (e.g., Illumina sequencing). However, we have not considered how PCR primers are designed in the first place.

During the early years of molecular phylogenetics and phylogenomics a researcher could acquire PCR primers for DNA sequence loci via three methods. In the first method, which was the only way to develop novel genomic loci, a researcher had to use genomic cloning methods. This method, however, had major drawbacks because it required researchers to have considerable expertise in molecular biology skills and access to a sophisticated molecular biology laboratory. This cloning-based technique was therefore not a practical option for many researchers. The second method consisted of using existing *universal primers* borrowed from published studies. The original universal primers developed by Kocher et al. (1989) enabled large numbers of researchers to sequence the same mitochondrial genes in many different metazoan species. Unfortunately, the numbers of universal primers available at that time in the literature—and indeed for many years thereafter—was limited to a small number of mitochondrial and nuclear genes. The successful use of these primer sets also depended on a trial and error process of repeated PCR testing of different primers in the laboratory. Such *cross-species PCR* experiments did not always produce successful amplifications owing to mismatches between primers and templates. "Cross-species PCR" is a PCR experiment in which a pair of primers developed from the genome of one species is used (hopefully) to amplify the same locus in the genome of a different species. The third method for acquiring loci consisted of custom-designing universal primers. As this approach depended on the availability of existing DNA sequences obtained from other organisms, it largely limited researchers to obtain sequence data for preexisting loci. Despite the limitations of the universal primer approach, these special primers represented one of the key innovations that helped spur the early growth of molecular systematics (Palumbi 1996) and they are still vital today. A striking example is the pair of mitochondrial universal primers originally developed for invertebrates more than 20 years ago (Folmer et al. 1994), which later launched the DNA barcoding program (Hebert et al. 2003). Many DNA barcoding studies—even on vertebrates—are still using these original universal primers.

Two decades later, the situation concerning primer availability has dramatically improved due to the existence of online primer databases and advent of genome-enabled methods for developing new loci (Thomson et al. 2010). Later in this chapter, we will be covering the topic of designing primers for new loci including how to make universal primers. For many types of studies, however, it may be unnecessary to make new primers because the requisite primers can often be obtained from published sources. It is not uncommon for researchers to locate suitable primers from previously published work performed on shared or closely related study organisms. In recent years, journals such as *Molecular Ecology Resources* and *Conservation Genetics Resources* have published articles offering sets of newly designed primers. Researchers may also find primers in public online databases such the *Molecular Ecology Resources* (http://tomato.bio.trinity.edu/) and *Barcode of Life Database* or "BOLD" (http://www.boldsystems.org/) websites. Thus, at the start of a study it may be worth your time to search the literature or online databases for existing primers, an approach that can save you time and money in attempting to design your own primers. Literature sources not only contain the actual primer sequences, but other useful information can also be found such as expected PCR product sizes, optimal annealing temperatures, and thermocycling profiles. It is important to understand that success of a cross-species PCR experiment is dependent upon how the primers were designed as well as the number and nature of base differences between the primers and target templates—even a single mismatch, in one of the two primers, can be sufficient to cause PCR failure. Although obtaining primers from published sources is the simplest method of primer acquisition, this approach may not be effective for many studies.

Researchers equipped with knowledge on how to design new loci are not constrained by the availability of primers in the literature. Thus they are free to design their own custom primers thereby greatly expanding the number of research possibilities. Also, the capability to design custom primers empowers researchers to obtain higher quality PCR results when only poorly functioning primers are available. In order to design new PCR-based loci a researcher must have access to one or more genomic template sequences for the target locus or loci and then design primers according to well-established rules of primer design. The existence of vast amounts of genomic sequences in online databases as well as NGS sequencing are providing researchers with templates for primer design while the primer design step can be satisfied by using primer design software or by manual design methods. As we will see in this chapter, primer design software can perform remarkably well for making some types of primers but not for all types of primers. Accordingly, because some types of primers must still be manually constructed, it is essential for a researcher to obtain a solid knowledge of primer design rules, which we will now review before discussing the various methods for developing phylogenomic loci.

8.1 PRIMER DESIGN THEORY

As we already know the forward and reverse primers are located in the regions immediately flanking the target sequence. Thus, if we know the flanking sequences, then we might be able to design a pair of primers that can amplify the target in a PCR experiment. Although it is possible that a researcher could design a good-functioning set of primers without consideration of the molecular properties of candidate primers except for observing the 18 bp minimum length rule we learned in Chapter 5, it is more likely that such primers would fail to some degree in PCR. This is because there are a number of potential design flaws that can seriously impair the proper functioning of a primer. What are the potential flaws that could afflict a primer? In general, these flaws consist of primers interacting with themselves (i.e., forming secondary structures or "hairpins"), interacting with other primers (e.g., forming primer dimers), or amplifying nontarget stretches of the genome. To address this problem, primer design rules have been developed. These well-established rules enable a researcher (or primer design software) to narrow down a list of candidate primers to those that are most likely to function well in PCR. Although there is no guarantee that a newly designed primer will work well in PCR, you can greatly increase your chances of making good primers by practicing careful primer design.

8.1.1 Rules of Primer Design

As PCR began to take off in the late 1980s, the need for an effective means of designing new PCR primers arose. Researchers realized that primers that effectively amplify their target loci could be developed if certain criteria were followed during the design process. These criteria were largely based on the chemistry of duplex formation between two primers, between primer and DNA template, and also whether or not primers could form secondary or "hairpin" structures. To meet this need, Rychlik and Rhoads (1989) developed the first software (http://oligo.net/) for primer design, which implemented these selection criteria. A short time later, Innis and Gelfand (1990) formalized a set of primer design "rules," which, together with the work by Rychlik and Rhoads (1989), has formed the foundation of primer design methodology. Another popular primer design program called *Primer3* (Rozen and Skaletsky 1999; Koressaar and Remm 2007; Untergasser et al. 2012; http://bioinfo.ut.ee/primer3/), which was based on the program *Primer 0.5* (Lincoln et al. 1991), has also greatly facilitated the task of choosing new primers and many other similar programs are available on the internet (Abd-Elsalam 2003). Although the "default" settings in primer design software often produces excellent results, many of these programs allow the user to alter the default settings to decrease or increase the stringency of searches. Changing any of the options usually affects the number of possible primers output from searches. We will now examine the most commonly applied rules. These rules are not given in order of importance though a brief consideration of which rules can be relaxed is given at the end of this discussion.

Rule 1: Primers Should Be 18–28 bp Long As was shown in Chapter 5, primers for routine PCR should be at least 18 bases long in order to achieve sufficient target template specificity for most PCR applications. Primers shorter than this minimum size may be susceptible to the problem of *mis-priming*. Mis-priming occurs when a primer anneals to nontarget genomic locations long enough for DNA polymerase to begin synthesis of a new DNA strand. The severity of mis-priming ranges from relatively minor effects such as consumption of reagents (e.g., dNTPs) to the deleterious consequences of coamplification of double-stranded PCR artifacts. The presence of multiple bands in an agarose gel is evidence of a mis-priming problem. As we will see below, other design defects can cause mis-priming problems in PCR.

What about the upper limit to the length of primers? Although the upper length limit of a primer is less constrained, primers should be less than 29 bases to allow for rapid annealing to the template (Innis and Gelfand 1990). In most studies, primers range from 18 to 25 bases long. Unless you have reason to do otherwise, for routine PCR applications it is best to choose primers that are close to 20 bases long with a search range of 19–22 bases.

Rule 2: Primers Should Have 40%–60% G + C Content Primer that have >60% G + C content may cause a number of PCR problems including an increased incidence of mis-priming, the formation of secondary structures called "hairpins," and stable primer–primer duplexes called "dimers." For this reason the G + C content of a primer should be kept as low as possible. Innis and Gelfand (1990) recommended an upper limit of 60% for the G + C content of a primer. The lower limit to G + C content seems to be less important. Suggested lower limits range from 50% (Innis and Gelfand 1990) down to 40% or even lower (Rychlik 1995). Note that for some types of A + T-rich templates (e.g., noncoding nuclear DNA) the primers will likely have a low G + C content therefore primers having a G + C content in the 30%–40% range may represent the only options. As we will see below under other primer design rules, the base sequence and total number of bases determine the optimal temperature at which a primer will anneal to its target template.

Rule 3: Avoid Complementarity at the 3′ End of the Primer In Chapters 5 and 6, we briefly considered what primer dimers are and how they can impact the quality of a DNA sequence. We will now look at the nature of primer dimers in more detail and see why it is critically important to ensure that the 3′ end of the primer does not contain a sequence of bases that could either make it self-complementary or complementary to the 3′ end of the other primer in the reaction. When a primer is self-complementary at the 3′ end it can lead to the formation of a type of primer dimer called a *homodimer*, whereas a 3′ end duplex formed between forward and reverse primers is called a *heterodimer*. Figure 8.1a shows an example of a homodimer, which formed because the 3′ end of one primer contained a 5′-CCGG-3′ sequence. There are two important consequences of this duplex formation. First, notice that *two* primer–template junctions are created, which could allow DNA polymerase to start extending DNA strands at both ends of the duplexed molecule. This can result in the production of double-stranded PCR products that are slightly shorter than the two primers joined end to end (Figure 8.1b). The second consequence of this dimer product formation is that binding sites for forward and reverse primers are synthesized, meaning each strand of this product can serve as a template in all future cycles (Figure 8.1b). Thus, as the reaction mixture is heated toward the denaturation temperature in the first cycle, the genomic DNA and newly formed dimer products denature into single-stranded templates and thus become available for primers during the initial annealing

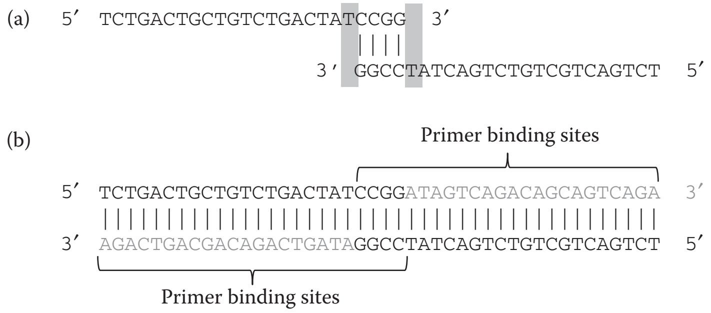

(a)
5′ TCTGACTGCTGTCTGACTATCCGG 3′
3′ GGCCTATCAGTCTGTCGTCAGTCT 5′
(b)
Primer binding sites
5′ TCTGACTGCTGTCTGACTATCCGGATAGTCAGACAGCAGTCAGA 3′
3′ AGACTGACGACAGACTGATAGGCCTATCAGTCTGTCGTCAGTCT 5′
Primer binding sites

Figure 8.1. Formation of a homodimer in PCR. (a) Self-complementarity at the 3′ end of a primer causes formation of homodimers. Two primer–template junctions are formed in the homodimer (gray rectangles). (b) DNA polymerase (not shown) synthesizes DNA (gray bases) resulting in the amplification of a 44 bp PCR product. In subsequent cycles each of the strands can serve as a template for both primers ("primer binding sites"). Vertical dashes represent hydrogen bonding between complementary bases.

step. Given the high concentration of unincorporated primer in the reaction, the dimer products will proliferate exponentially and compete with the target products for reagents. When the reaction is completed, there is a great abundance of at least two PCR products: the target product and the primer dimers. Unless the primer dimers are somehow removed from the completed reaction and discarded before sequencing, they too will generate extension products in a cycle sequencing reaction with the end result being the diminished quality of DNA sequence chromatograms.

However, as the 4-bp duplex in Figure 8.1a cannot form during the annealing temperatures of PCR experiments (i.e., this duplex is only stable at temperatures less than 10°C), how then is it possible for the primer dimer to become replicated throughout the PCR process? The answer to this question highlights the critical nature of the first step of the first PCR cycle—specifically the time from when all the reagents are combined in the PCR tube until the first denaturation step is reached. Because the PCR tubes are initially at a cooler temperature (4–25°C) from the moment all the reagents are combined until the start of the first cycle, these conditions could enable dimer formation between primers because *Taq* polymerase can still function at these lower temperatures. At these lower temperatures *Taq* can begin extending off the primers and produce short double-stranded products prior to the first denaturation step. When the first denaturation step is reached, both the genomic DNA and newly created primer dimer templates become single stranded. Thus by the time the first annealing step is reached, there will be *two* sets of templates available for the primers, which ultimately results in coamplification of target and primer dimer templates.

Not all primer–primer duplexes are necessarily detrimental to PCR. When checking a particular primer or primer pair using primer design software (or by visual inspection), it is possible that the analysis will suggest possible primer–primer duplex formation involving internal bases or the 5′ends (Figure 8.2). Although these duplexes may temporarily form before the first denaturation step, at no time can DNA polymerase synthesize DNA on these dimers simply because neither of these dimers contains primer–template junctions. Therefore, these particular dimers cannot lead to the mass production of double-stranded PCR artifacts. It is conceivable that the formation of these dimers can potentially lower the efficiency of a PCR reaction in other ways. For example, if the primer–primer duplex is stable enough—even at typical PCR annealing temperatures, then these primer–primer interactions might decrease the target product yield by depriving the target templates of available primers. However, as we will see below, this is unlikely to occur on the basis of predicted primer–primer duplex stabilities.

Thus far, we have been considering a vague notion of "stability" of primer–primer or primer–DNA template duplexes, but what factors determine the stability of a DNA double helix? While hydrogen bonding between complementary bases on adjacent strands contributes to helix

```
(a)      5′  ATACACTGCCGGTGACTATCATG  3′
                     ||||
      3′  GTACTATCAGTGGCCGTCACATA  5′


(b)                     5′ GGCCACTGCTGTCTGACTATCATG  3′
                           ||||
    3′ CATGATAGTCAGACAGCAGTCCGG  5′
```

Figure 8.2. Two primer–primer duplexes that have little or no effect on PCR efficiency. (a) Internal self-complementarity may lead to duplex formation between two primers but not to homodimer formation. (b) Self-complementarity at the 5′ end of a primer may also lead to duplex formation between two primers but not to homodimer formation. Note in both cases a primer–template junction cannot be formed therefore DNA synthesis cannot occur. Vertical dashes represent hydrogen bonding between complementary bases.

stability, other factors play significant roles in determining the molecular stability of duplexes; these include entropic forces caused by water molecules that are displaced to the outside of the helix and especially by base stacking interactions of adjacent bases (Watson et al. 2014). Consider two dsDNA molecules of equal length. Given that each base pair follows Watson–Crick pairing rules and each duplex experiences comparable entropic forces, which duplex is expected to be the more stable one? It is often thought that base composition of a sequence plays a large role in the stability of a duplex, as sequences with a large number of G + C bases tend to be more stable than sequences containing fewer of these bases. Thus, one might conclude that the duplex with the higher G + C content might be the more stable duplex of the two. Although base composition correlates with duplex stability, it is not base composition *per se* that determines duplex stability. Rather, stability is largely determined by the thermodynamic interactions of adjacent or nearest-neighbor bases on the same strand—in other words, stability is determined by the primary nucleotide sequence (Breslauer et al. 1986). Thus, stability of a dsDNA molecule is primarily due to the sum of its nearest-neighbor interactions. This implies that longer DNA duplexes will be more stable than shorter duplexes all else equal. Recognition of this fundamental idea, which has since been formalized as *nearest-neighbor thermodynamic theory*, has provided researchers with an effective means for predicting the stability of DNA duplexes and has proved useful to applications such as PCR and DNA sequencing. An important extrinsic factor that determines duplex formation between two complementary ssDNA molecules is temperature. We saw in Chapters 5 and 6 that a thermocycler is used to modulate the thermal environment of the PCR reaction between denaturing DNA duplexes at high temperatures and facilitating proper primer–template duplexes at lower temperatures. Thus, it is the length and sequence of a primer together that ultimately determine its optimal annealing temperature to target genomic sequences in PCR.

The relevant metric for quantifying the stability of short DNA duplexes is the Gibb's Free Energy or "ΔG" (Breslauer et al. 1986; Rychlik 1995), which was named in honor of Josiah Willard Gibbs, a scientist who made seminal contributions to physics, chemistry, and mathematics (Josiah Willard Gibbs, Wikipedia). Predicted ΔG values, which can be positive or negative, are in units of kcal/mol. Duplex formation is energetically favored when ΔG is negative and the strength of stability for this duplex increases as ΔG becomes more negative. Because the stability of a DNA duplex is determined by the identity of its nearest-neighbor bases, the stability of a particular duplex can simply be calculated by summing the free energies of each pair of bases in a sequence. Breslauer et al. (1986) provided a thermodynamic library for all ten unique nearest-neighbors (Table 8.1). This library consists of the free energies for each possible dinucleotide that can be found in a Watson–Crick DNA duplex. Inspection of the free energy parameters for each nearest-neighbor pair reveals that the most stable dinucleotide groups are CG/GC, CC/GG, GC/CG, and GG/CC (i.e., their respective ΔG values are −3.6, −3.1, −3.1, and −3.1), while in contrast all the pairings involving A's and T's have less negative ΔG's and are therefore relatively less stable (Table 8.1). Other

TABLE 8.1
Nearest-neighbor free energy (ΔG) parameters for all 10 dinucleotide pairs at 25°C

Interaction	ΔG
AA/TT	1.9
AT/TA	1.5
TA/AT	0.9(1.0)
CA/GT	1.9
GT/CA	1.3
CT/GA	1.6
GA/CT	1.6
CG/GC	3.6
GC/CG	3.1
GG/CC	3.1

SOURCE: Data from Table 8.2 of Breslauer, K. J. et al. 1986. *Proc Natl Acad Sci USA* 83:3746–3750. With permission.

NOTE: ΔG values are in units of kcal/mol. For example, the nearest-neighbor group AT/TA is interpreted as a 5′–AT–3′ dinucleotide that is hydrogen bonded to its complementary sequence 3′–TA–5′.

workers subsequently refined these parameters (see SantaLucia 1998), but the values in Table 8.1 are good approximations for evaluating stabilities of short DNA duplexes.

Breslauer et al. (1986) and SantaLucia (1998) provided equations for estimating the total ΔG of a duplex. Their equations compute the total ΔG by summing the free energies of each nearest-neighbor dinucleotide then correcting this estimate using additional terms in the equations (see Breslauer et al. 1986 and SantaLucia 1998 for details). For purposes of primer design, we will follow Rychlik (1995) and use a modified equation that omits the correction terms. This modified equation predicts the total uncorrected free energy ΔG for a given dimer duplex at 25°C:

$$\Delta G = \sum \Delta G_i, \qquad (8.1)$$

where i is each nearest-neighbor group.

Using Equation 8.1 together with Table 8.1 we can estimate the stability of the 3′ end duplex shown in Figure 8.1a. If we momentarily ignore the part of each primer not involved in the primer–primer interaction and focus only on the bases involved in the duplex, then the duplex consists of double-stranded 4-bp stretch of DNA (Figure 8.1a). Notice there are three different nearest-neighbor groups each of which will contribute to the total uncorrected ΔG value for the duplex. Using Table 8.1 we can estimate ΔG for this duplex:

$$\Delta G = \Delta G(CC) + \Delta G(CG) + \Delta G(GG)$$
$$\Delta G = -3.1 + (-3.6) + (-3.1)$$
$$\Delta G = -9.8 \text{ kcal/mol}$$

The above calculation follows the example in Rychlik (1995) and can easily be done without a calculator. Now that we have estimated the ΔG for the 3′ end homodimer duplex in Figure 8.1a, we need to address the following question: *if we used a primer that forms homodimers with a duplex stability of ΔG = –9.8 kcal/mol, then what is the expected impact on PCR product yield?* Rychlik (1995) showed the dependence of PCR product yield on 3′ duplex stability of primer dimers, which is shown in Figure 8.3. The curve in Figure 8.3 suggests that our ΔG of –9.8 kcal/mol may severely diminish PCR product yield. Rychlik's results show that once the ΔG for a 3′ end duplex exhibits higher stability (i.e., more negative) than ΔG = –4 kcal/mol, we can expect PCR product yields to be sharply reduced. Primers with negative ΔG 3′ stability values should be avoided because, even if the formation of the unwanted duplexes does not lead to the formation and accumulation of primer dimers, these partially stable duplexes may interfere with the synthesis of target products (Rychlik 1995). Thus, it is best to design primers that are free of 3′ end dimer issues.

There are occasions when, for lack of other primer options, a researcher must use a pair of primers that produce primer dimers along with

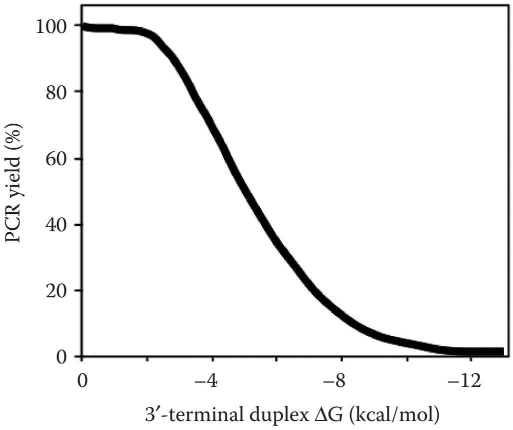


Figure 8.3. Dependence of PCR yield on the ΔG of 3′-terminal duplexes. (Reprinted from Rychlik, W. 1995. *Mol Biotechnol* 3:129–134. With permission.)

a well-amplified target product. In this situation, there are three strategies that can be used to obtain high-quality target sequences. First, if the primer dimers are homodimers and the product is short enough (<700 bp), then simply sequencing with the primer not involved in the dimer formation will likely yield a high-quality sequence. If the PCR reaction produces heterodimers, then sequencing the amplicons in both directions can solve the problem. Although each sequence will have low-quality base calls in the first ~40–50 bases, they will have higher quality base calls (≥Q20 Phred scores) for the remainder of the sequences. This will allow the researcher to construct a high-quality contig sequence using both chromatograms. A second strategy is to use methods of PCR product cleanup (Chapter 6) that are effective at removing primer dimers from target products (e.g., PEG or some commercial spin-column kits). These methods can be effective but they can be costly in terms of sample processing time and cost of kits. A third method attempts to prevent the formation of primer dimers during PCR via the "hot start" method (Chou et al. 1992). It should now be clearer exactly *how* the hot start method can prevent the formation of primer dimers. These harmful DNA templates can only form during the brief time period prior to the first denaturation step in PCR because the remainder of the PCR steps will be carried out at temperatures that prohibit the initial formation of primer–primer duplexes. Thus, if *Taq* is added to the reaction at a temperature well above the permissible temperatures for the formation of short DNA duplexes, then it is likely that the 3′ end duplexes will not form and thereby preclude the mass production of alternative priming sites for *Taq*.

Rule 4: Avoid Runs of Three or More C's and G's at the 3′ End of Primer When designing primers it is critically important to ensure that the bases at the 3′ end of a primer should not contain runs of 3 or more "G" or "C" bases otherwise mis-priming will likely occur (Innis and Gelfand 1990). For example, in Figure 8.4a we see a primer with a "sticky" or relatively stable 3′ end (consisting of a 4-bp CGCG sequence) hybridizing to a nontarget template location. Mis-priming occurs when the 3′ terminal end of the primer duplexes with a nontarget strand long enough to allow the DNA polymerase to synthesize a new strand of DNA (Figure 8.4b). In contrast, primers with relatively less stable 3′ ends, but highly stable internal sections or 5′ ends, are not likely to mis-prime (Rychlik 1995).

How sticky must a 3′ end of a primer be for mis-priming to occur? Rychlik (1995) suggested that the 3′ end stability of a primer could be quantified by calculating free energy of the 3′ end *pentamer*, which we will refer to as $\Delta G_{\text{pentameter}}$. Note, that here we are specifically referring to the stability of the last five bases located at the 3′ end of the primer. Based on empirical data, Rychlik (1995) showed that primers with relatively high stabilities at the 3′ pentameter ($\Delta G_{\text{pentameter}} = -9$ to −12 kcal/mol or lower) performed poorly in PCR owing to mis-priming problems, whereas primers with low to medium 3′ end stabilities ($\Delta G_{\text{pentameter}} = -5$ to −9 kcal/mol) performed better. In Figure 8.5, we see 20 primers of the same

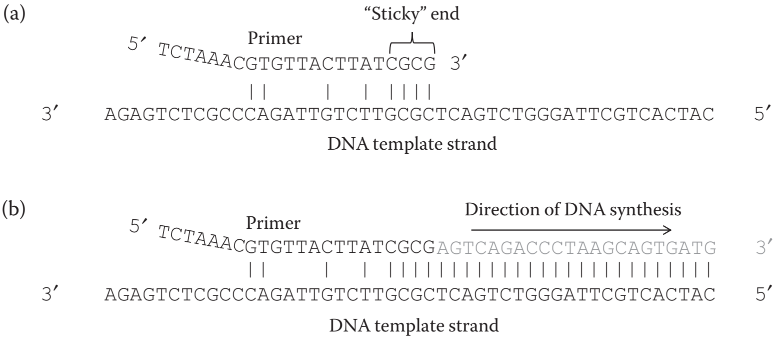

(a)
"Sticky" end
Primer
5′ TCTAAACGTGTTACTTATCGCG 3′
3′ AGAGTCTCGCCCAGATTGTCTTGCGCTCAGTCTGGGATTCGTCACTAC 5′
DNA template strand
(b)
Direction of DNA synthesis
Primer
5′ TCTAAACGTGTTACTTATCGCGAGTCAGACCCTAAGCAGTGATG 3′
3′ AGAGTCTCGCCCAGATTGTCTTGCGCTCAGTCTGGGATTCGTCACTAC 5′
DNA template strand

Figure 8.4. Mis-priming caused by primers with "sticky" 3′ ends. (a) Partial annealing of primer to nontarget location (i.e., genomic DNA or PCR amplicon). The resulting partial duplex contains a primer–template junction, which can be extended by DNA polymerase. (b) DNA polymerase (not shown) begins synthesizing new strand of DNA (gray bases). Vertical dashes represent hydrogen bonding between complementary bases.

length but with different 3′ end pentamers. On the right side of Figure 8.5 are shown the ΔG values for each pentamer. As you will notice, when the G + C content increases in the pentamers, the end stabilities also tend to increase (Figure 8.5). Accordingly, if we designate a critical threshold for primer end stability of $\Delta G_{\text{pentameter}} = -9$ kcal/mol for specific (ideal) primers, then we can see that primers with a total G + C count of two or less at their 3′ ends are within this range. Some primers with a G + C count of three located within the 3′ pentamer also have acceptable levels of end stability but it depends on their nearest-neighbors (e.g., GACTC and TAGGG both have 3′ ends that less stable than ΔG = −9 kcal/mol threshold; Figure 8.5). Rychlik (1995) suggested that a primer with a low stability 3′ end but higher stability middle section or 5′ end is a more specific primer and less susceptible to mis-priming problems.

Although incidences of mis-priming can be reduced if the 3′ primer end does not contain too many G's and C's, some researchers have suggested that having a terminal 3′ base consisting of a G or C—called a "GC clamp"—may help ensure proper priming (e.g., Kwok et al. 1990). Provided that the 3′ end pentamer of a primer exhibits low to medium stability, then such a GC clamp may improve primer performance (Rychlik 1995). However, empirical evidence provided by Haas et al. (1998) in addition to the great abundance non-GC clamp primers published in the literature suggests that a primer could have any of the four possible bases at the 3′ terminus and still function properly.

Rule 5: Avoid Palindromic Sequences within the Primer There is another form of self-complementary sequence that, if present in a primer, can cause serious PCR problems. If a primer contains a *palindromic sequence*, then it is self-complementary to itself and can therefore form a secondary structure or *hairpin* structure (Rychlik and Rhoads 1989; Innis and Gelfand 1990). A hairpin can form when a primer contains a palindromic sequence with a string of least three loop-forming bases located in the middle of the palindromic sequence (Figure 8.6a). Note, it is sterically impossible for a loop to form with less than three bases (Freier et al. 1986; Rychlik and Rhoads 1989). By forming a single-stranded loop structure (the bases in the loop do not hydrogen bond to other bases), the primer can fold over on itself and form a base-paired duplex—called the *stem*—with the other half of the palindromic sequence (Figure 8.6b). If the hairpin involves the 3′ end of the primer then a primer–template junction can form leading to self-priming and hence primer extension (Figure 8.6c). Just

0/5 GC	#1	5′..................TATAT 3′	ΔG = –4.9
	#2	5′..................TAATA 3′	ΔG = –5.3
	#3	5′..................TATTA 3′	ΔG = –5.4
	#4	5′..................TAATT 3′	ΔG = –6.3
1/5 GC	#5	5′..................TAATC 3′	ΔG = –6.0
	#6	5′..................GAATA 3′	ΔG = –6.0
	#7	5′..................TATTC 3′	ΔG = –6.0
	#8	5′..................TAATG 3′	ΔG = –6.3
2/5 GC	#9	5′..................TACAC 3′	ΔG = –5.6
	#10	5′..................GAATC 3′	ΔG = –6.6
	#11	5′..................TATCC 3′	ΔG = –7.1
	#12	5′..................TATGG 3′	ΔG = –7.5
3/5 GC	#13	5′..................GACTC 3′	ΔG = –6.1
	#14	5′..................TAGGG 3′	ΔG = –8.7
	#15	5′..................TACGC 3′	ΔG = –9.1
	#16	5′..................TAGCG 3′	ΔG = –9.3
4/5 GC	#17	5′..................GACGC 3′	ΔG = –9.7
	#18	5′..................TGGCC 3′	ΔG = –11.2
	#19	5′..................TCCGG 3′	ΔG = –11.3
	#20	5′..................TGCGC 3′	ΔG = –11.8

Figure 8.5. Stabilities of 3′ terminal pentamers of primers as a function of G + C base composition. For example, primers #9–12 each have a G + C composition of 2/5 in the 3′ end of their primer sequence. Only the 3′ end pentamer sequence is shown for each primer (dots represent bases hidden from view). Stabilities are measured as ΔG (free energy) values.

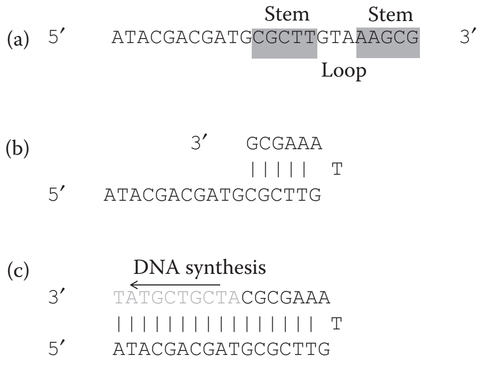


Figure 8.6. Potential consequence of a primer hairpin in PCR. (a) Primer containing a palindromic sequence (two gray rectangles). (b) Primer folds over on itself to form a hairpin involving the 3′end. This hairpin is characterized by a 3-bp loop, 5-bp stem, and a $\Delta G_{\text{hairpin}} = -5.0$ kcal/mol. (c) Because this hairpin forms a primer–template junction, self-priming can occur.

like primer dimers, hairpins are likely to form and subsequently self-prime DNA synthesis only when the reaction mixture is at permissible (low) temperatures (i.e., at the time when all PCR reagents are mixed together and before the first denaturation step). In such a scenario, the hairpin-forming primer can be rendered nonfunctional in the PCR, because most if not all the hairpin-forming primer will form PCR artifacts and not target products. Not all primer hairpins are problematic, as hairpins that only affect the 5′ end or internal part of the primer will likely not affect PCR efficiency because they do not contain primer–template junctions and therefore cannot self-prime (Rychlik 1995).

As with primer dimers, the stability of duplex formation is the key to whether or not synthesis of PCR artifacts can occur due to primer hairpins. Thus, the stability of a hairpin duplex on primers can be evaluated using ΔG but a term characterizing the free energy of the loop (ΔG_{loop}) must be included in the calculation because the loop is a destabilizing influence on the duplex (Rychlik 1995). Approximate free energies for loops ranging in sizes from 3 to 8 bases are the following: three base loop = 5.2 kcal/mol, four base loop = 4.5 kcal/mol, five base loop = 4.4 kcal/mol, six base loop = 4.3 kcal/mol, seven base loop = 4.1 kcal/mol, and eight base loop = 4.1 kcal/mol (Rychlik 1995). Thus, ΔG of the hairpin duplex ($\Delta G_{\text{hairpin}}$) can readily be calculated using Equation 8.1 but with an ΔG_{loop} term included

$$\Delta G_{\text{hairpin}} = \sum \Delta G_i + \Delta G_{\text{loop}}, \tag{8.2}$$

where i is each nearest-neighbor group.

Again, this approximation of hairpin stability can be considered as an "uncorrected" estimate because we are omitting the initiation terms from the equation (see SantaLucia 1998 for full equation). Using Equation 8.2, Table 8.1, and the ΔG_{loop} for a 3-base loop, the stability of the hairpin in Figure 8.6b is estimated to be

$$\Delta G_{\text{hairpin}} = \Delta G(\text{GC/CG}) + \Delta G(\text{CG/GC}) + \Delta G(\text{GA/CT}) + \Delta G(\text{AA/TT}) + \Delta G_{\text{loop}}$$

$$\Delta G_{\text{hairpin}} = -3.1 + (-3.6) + (-1.6) + (-1.9) + 5.2$$

$$\Delta G_{\text{hairpin}} = -5.0 \text{ kcal/mol}$$

How bad is a hairpin with this level of stability? In general, negative $\Delta G_{\text{hairpin}}$ values should be avoided but primers with stabilities as high as $\Delta G_{\text{hairpin}} = -3.0$ kcal/mol may function properly in PCR (Rychlik 1995). Observing that the hairpin structure in Figure 8.6b involves the 3′ end and exhibits duplex stability exceeding the −3.0 kcal/mol threshold, this primer should not be used in PCR. Because PCR artifacts due to hairpins form prior to the initial denaturation step in PCR (like primer dimers), the hot start method (Chou et al. 1992) may be effective at preventing the initial formation of these products and thus allow the reaction to produce only target products. However, designing a hairpin-free primer offers a simpler approach.

Rule 6: Primers Should Have a Tm 50–65°C The melting temperature or "Tm" of a primer is a commonly used measure of the stability between the primer and its template when they are hybridized together via hydrogen bonding into a duplex molecule. The Tm is defined as the temperature at which half of the potential binding sites are bound with primer (Palumbi 1996). Knowing the Tm of the primers allows you to predict the optimal annealing temperature for thermocycling. Because low annealing temperatures can decrease primer specificity thereby leading to the production of PCR artifacts (i.e., a mis-priming problem), higher annealing temperatures should be used (e.g., ≥50°C) when possible. Innis and Gelfand (1990) suggested that primers should be designed to have Tm's in the 55–80°C range. However, in practice the most commonly used Tm's are in the 50–65°C range.

The most accurate way to estimate the Tm of a primer involves using nearest-neighbor thermodynamic theory. Many primer design software packages use this approach to estimate the Tm of primers. However, this method is difficult to do by hand such as we did earlier when estimating the ΔG for short primer–primer DNA duplexes. A less accurate "rule of thumb" method for estimating the Tm of a primer simply considers the length of a primer and weighs the four bases differently to reflect their general stabilities in nearest-neighbor interactions. Suggs et al. (1981), Thein and Wallace (1986), and Palumbi (1996) provided this simple equation for estimating the Tm of a primer:

$$\mathrm{Tm} = 2°\mathrm{C} \times (\mathrm{A}+\mathrm{T}) + 4°\mathrm{C} \times (\mathrm{G}+\mathrm{C}) \qquad (8.3)$$

This simple formula for estimating the Tm of a primer, which only requires counts of each of the four bases in a primer, has been widely used by researchers.

Rule 7: Each Primer in a Pair Should Have Tm < 5°C of Each Other It is critically important for a pair of primers to have similar melting temperatures because the primer with the higher Tm may mis-prime if the annealing temperature is set too low. Alternatively, if the annealing temperature is matched to the primer with the higher Tm, then the temperature may be too high for the lower Tm primer causing it to not anneal at all. Therefore, the rule of thumb is that the two primers should have Tm values within 5°C of each other (Innis and Gelfand 1990).

Rule 8: Avoid Mismatched Bases at the 3′ End of the Primer If a primer is annealed to the correct template locus but one or more bases are mismatched, then the consequences on PCR yield vary from having little effect to a more drastic effect. In a study that focused on the effects of such mismatches on PCR yield, Kwok et al. (1990) found that single mismatches had no significant effect on PCR yield as long as they were located internally within the primer-template duplex and did not involve the sensitive 3′ end of the primer. In contrast, their findings indicated that if two of the final four bases in the 3′ end were mismatched, then PCR yield was substantially lowered, especially if the terminal 3′ base was mismatched. In cases where only the 3′ terminal base was mismatched, PCR yield varied depending on mismatch: a 100-fold reduction resulted from A:G, G:A, and C:C mismatches, a 20-fold reduction from a A:A mismatch, but all other mismatches largely had little effect on yield under normal PCR conditions (Kwok et al. 1990). One surprising finding by the authors concerned the discovery that T:G, T:C, and T:T mismatches had little effect on PCR yield. This suggests that designing a primer with a "T" at the 3′ terminal base may be a good strategy if the primer is to be used in cross-species studies (Palumbi 1996). If the primer is embedded within a coding sequence, then a common strategy is to make sure that the 3′ terminal base corresponds to a 2nd codon position site (e.g., Backström et al. 2008). Owing to the lack of variability observed at 2nd codon sites, this strategy virtually guarantees that a mismatch involving the 3′ terminal base will not happen.

8.1.2 Final Comments about Primer Design Rules

Anytime new primers are designed without the aid of a primer selection software program, the candidate primers should be carefully evaluated to see how well they meet the aforementioned rules of design. The use of bioinformatics software tools can greatly facilitate the analysis of primers. For example, the *OligoAnalyzer 3.1* tool by Integrated DNA Technologies® (http://www.idtdna.com/calc/analyzer; Owczarzy et al. 2008), which I routinely use, is very useful for analyzing characteristics of primers, including homodimers, heterodimers, hairpins, 3′ end pentamers, Tm, and G + C content. Kibbe (2007) also provides a web-based bioinformatics tool for examining the properties of primers.

When using primer selection software, occasionally the program will not be able to suggest many or any possible primer pairs even when the default settings are used. In these cases, changing some of the settings can lessen the stringency of the search, which usually increases the output of suggested primer pairs. The first rules to relax are (in order): rules #1 (primer length), #2 (G + C%), and #6 (Tm). For example, if you first search for primers that are 19–22 bases long, have a 40%–60% G + C content, and have Tm's 55–60°C, but are unable to find any candidate primers as a result, then it is time to broaden the search. In the second search, the aforementioned parameters could be expanded to include primers of: 18–25 bases, G + C content 30%–60%, and Tm's 50–65°C. If it is likely that the primers and template will match each other perfectly, then these relaxed criteria should still enable you to select good primers. If on the other hand, you suspect that mismatches might exist between the primers and template, then primer length and Tm should remain at the higher end of the search ranges to ensure a high level of specificity in PCR.

8.1.3 Testing New Primers in the Lab

Whenever new PCR primers are designed, they must be bench tested in the laboratory. The testing procedure amounts to performing PCR reactions and varying the annealing temperature until the thermal "window" in which the primers are able to amplify the correct product is found. Sometimes this thermal window at the annealing step spans 5–10°C (or more) and sometimes it is <1°C. This annealing temperature parameter can only be empirically determined for each untested primer pair and therefore it is a trial and error process. However, to reduce the number of test PCRs that must be performed, manufacturers of many thermocyclers have incorporated a thermal gradient feature during the annealing step. For example, if a researcher is evaluating annealing temperatures that range between 50 and 60°C, the PCR tubes containing identical reagents are placed across the 12 wells in the thermocycler's metal heating block where the temperature gradient is generated. This gradient feature not only helps the researcher identify the optimal annealing temperature for each primer pair, but it can also obviously reduce the amount of time expended for testing new primers. Occasionally, some new primers never perform well in PCR and so they should either be redesigned or discarded. However, assuming that the template sequences used to make the primers were of appropriate quality and primer design software used (or manually designed), then most if not all new primer pairs should generate the expected PCR products. Table 8.2 shows the success rates for new sets of primers from a variety of studies that used single genomic templates together with primer design software to develop new PCR-based anonymous loci. These data show that success rates are generally high, which varied between 47% and 86% (Table 8.2). Note that little if any PCR optimization (e.g., experimenting with different annealing temperatures) was done in some of these studies and therefore it is likely that at least some of the success rates shown in Table 8.2 represent underestimates of the true number of good-functioning primers.

8.2 PRIMER AND PROBE DESIGN APPROACHES

The rapidly progressing genomics and bioinformatics fields are removing the barriers to developing phylogenomic loci. In recent years, a number of published studies have proffered a flurry of new methods for designing new loci. These newer methods not only allow researchers to target particular loci without resorting to challenging and time-consuming genomic cloning approaches, but some methods are capable of generating hundreds or thousands of loci for organismal phylogenomic studies. Many of these newer approaches involve the use of Perl or Python scripts as part of

TABLE 8.2
PCR success rates for sets of anonymous loci primers involving studies of vertebrates

Organisms	# Primer sets tested	# Primer sets with successful PCR	% Success	Study
Birds	17	8	47	Amaral et al. (2012)
Lizards	15	12	80	Bertozzi et al. (2012)
Snakes	17	12	71	Bertozzi et al. (2012)
Birds	35	30	86	Jennings and Edwards (2005)
Lizards	77	50	65	Rosenblum et al. (2007)
Turtles	96	73	76	Thomson et al. (2008)

NOTE: A pair of primers is considered functional if the pair can successfully amplify the correct locus in a DNA sample from the same species (i.e., a single amplification band of the expected size in an agarose gel is sufficient evidence). Primers in all listed studies were designed using primer design software.

their pipelines. Thus, in order to implement these procedures with the least amount of troubles, it is helpful to have some knowledge about basic programming in the Perl and Python languages. Before a particular locus development method can be chosen, however, the researcher must first decide which type of locus or loci will be needed. Locus design methods can be divided into two basic approaches: (1) *single template approaches* and (2) *multiple homologous templates approaches*.

8.2.1 Single Template Approaches for Developing PCR-Based Loci

For studies involving highly similar sequences such as those involving a single gene family, a single species, or a complex of closely related species, a *single template approach* is generally used. This is because these types of sequences generally show low levels of nucleotide variation and hence mismatches at primer binding sites are expected to be too few to adversely impact a PCR. Thus, only a single representative sequence is needed to serve as the template from which primer design software can find appropriate forward and reverse primers.

Various approaches have been used for obtaining a single template sequence to design new primers. Until recently, most studies relied on genomic cloning to make new loci. However, simpler and more efficient methods have been developed since the dawn of the genomics era, which make use of existing genomic resources (e.g., clone libraries), partial genome sequences generated from NGS platforms, or use lab-free computer searches of whole genome sequences available on public databases. We will examine these methods below beginning with a discussion of genomic cloning methods. Although cloning methods have been superseded by newer genomics-based methods, important insights have emerged from that work that can benefit locus development endeavors in the future.

8.2.1.1 Single Template Methods Using Genomic Cloning Methods

Although this method is no longer used to generate phylogenomic loci, it is being presented here for two reasons. First, I believe it is of historical interest to see how novel loci (e.g., anonymous loci) were developed prior to the time of genomics and NGS. Secondly, there are some important implications from these early cloning-based phylogenomic studies that are instructive to modern genome-enabled and NGS approaches to loci development, particularly pertaining to anonymous loci.

The following represents a simplified description of the gene cloning process. Interested readers should consult Sambrook et al. (1989) for detailed protocols. The first major step in cloning is to construct a genomic library from an individual of the study species. To construct a genomic library, the following steps must be performed: (1) obtain a sample of purified genomic DNA from an individual of the reference species; (2) break the genomic DNA into shorter fragments using sonication or restriction enzymes; (3) run the fragmented DNA through an agarose gel and excise a gel slice that contains a subset of the DNA fragments (e.g., fragments in 1–2 kb range); (4) purify the size-selected DNA; (5) ligate the fragments into plasmid cloning vectors; (6) transform the recombinant vectors into *E. coli* bacteria; and (7) screen the bacterial colonies for the "clones" containing recombinant vectors. At this point in the cloning process, the next step will depend on the type of locus that the researcher wishes to develop. If the researcher would like to design primers that amplify a particular gene or gene family, then a radioactively labeled oligonucleotide probe must be used to hybridize only to the recombinant plasmids that contain a copy of that gene. The recombinant clones retained from this hybridization procedure are then selected for sequencing, which will hopefully yield the template + flanking sequences for the locus of interest. The final step is simple: the obtained template sequence is input into primer design software, which then outputs the primer sequences.

Alternatively, if the researcher wants to develop anonymous loci, then a random subset of the recombinant plasmids are selected for sequencing. Once a number of different plasmid inserts have been sequenced, the researcher can design a set of new anonymous loci simply by using primer design software to produce one primer pair nested within each insert sequence. However, if these anonymous loci are destined for use in phylogenomic analyses that assume each locus has the properties of being selectively neutral, genealogically independent from other sampled loci, and single-copy, then additional filtering steps are needed prior to the final primer design step.

To evaluate whether or not a given insert sequence meets the neutrality assumption, two approaches can be tried. First, a BLAST search of the Genbank database can be used to query the genome of the most closely related species to see if a candidate anonymous locus matches a known conserved part of the genome such as a protein-coding gene (Jennings and Edwards 2005). Candidate anonymous loci that show no "hits" with the reference genome are tentatively considered to be from noncoding regions of the genome and hence are deemed more likely to meet the neutrality assumption. Another strategy is to use bioinformatics software to scan each sequence for an "ORF" that spans hundreds of bases because sequences that have such long ORFs may represent protein-coding regions. Sequences with long ORFs can be discarded as a precautionary measure.

The independent loci assumption is thought to be largely satisfied if anonymous loci are obtained from random genomic locations. This is because such loci would be found on different chromosomes or, if found on the same chromosome, they would likely be found far enough apart from each other that their gene trees would effectively be independent of each other. Evidence that this approach works as desired comes from the study of Jennings and Edwards (2005), who used blunt-end genomic cloning methods to develop 30 anonymous loci for Australian grass finches. When these loci are mapped to the chromosomes of the zebra finch, a grass finch species that had its genome fully sequenced (Warren et al. 2010), we can see that these loci are scattered across the genome as expected (Figure 8.7). Although some loci appear to be close together (e.g., on Chromosomes 1, 4, 4A, 7, and 19), the minimum physical distances between neighboring loci is on the order of hundreds of thousands of bases and, in some cases, millions of bases apart. Given these rather large physical distances (see Table 3.1), it is likely that all sampled loci have independent gene trees.

Lastly, in an attempt to satisfy the single-copy locus assumption, Karl and Avise (1993) used laboratory procedures to identify and then eliminate from consideration all vector insert sequences containing repetitive DNA. This procedure was deemed necessary because it was originally believed that anonymous loci based on genomic sequences containing replicative transposons would likely violate the single-copy assumption (i.e., PCR would coamplify many copies of a single locus or different copies in different individuals). Since that time, many published methods for developing anonymous loci include a filtering step to eliminate sequences containing unwanted repetitive DNA (e.g., Chen and Li 2001; Shaffer and Thomson 2007; Thomson et al. 2008; Bertozzi et al. 2012; Lemmon and Lemmon 2012).

Owing to the high copy numbers of some transposable elements in many eukaryote genomes, there is certainly sufficient reason to be concerned about the possible adverse effects of repetitive DNA on locus development. However, these measures also can eliminate large parts of the genome from locus development consideration thereby limiting the number of loci for phylogenomic studies. Thus, an important question to ask is: *will anonymous loci with repetitive DNA likely violate the single-copy assumption?*

We can evaluate this question using the set of anonymous loci for the grass finches from the study of Jennings and Edwards (2005) because these authors did not include a repetitive DNA filtering step during loci development. First, we can get an idea about which of these loci are comprised of transposable elements by using the *CENSOR* software tool (Kohany et al. 2006) found on the online database *Repbase* (http://www.girinst.org/censor/index.php; Jurka et al. 2005). The results revealed that 10 of the 30 loci showed no evidence of containing any repetitive DNA. Of the remaining 20 loci, 16 contained a single transposon, three loci had two transposons embedded within their sequences, and one locus has three transposons. The transposable elements found amongst these loci consisted of four DNA transposons, 13 LTR retrotransposons, and eight non-LTR retrotransposons. As you will recall from Chapter 2, retrotransposons are represented by many large gene families, which can comprise large fractions of eukaryote genomes. Despite the threat posed by retrotransposons, several lines of evidence suggest that they do not pose problems with these particular anonymous loci. First, sequencing of multiple clones per PCR product did not reveal more than two alleles per individual (Jennings and Edwards 2005). Secondly, when sequences of each locus were used in a BLAST search against

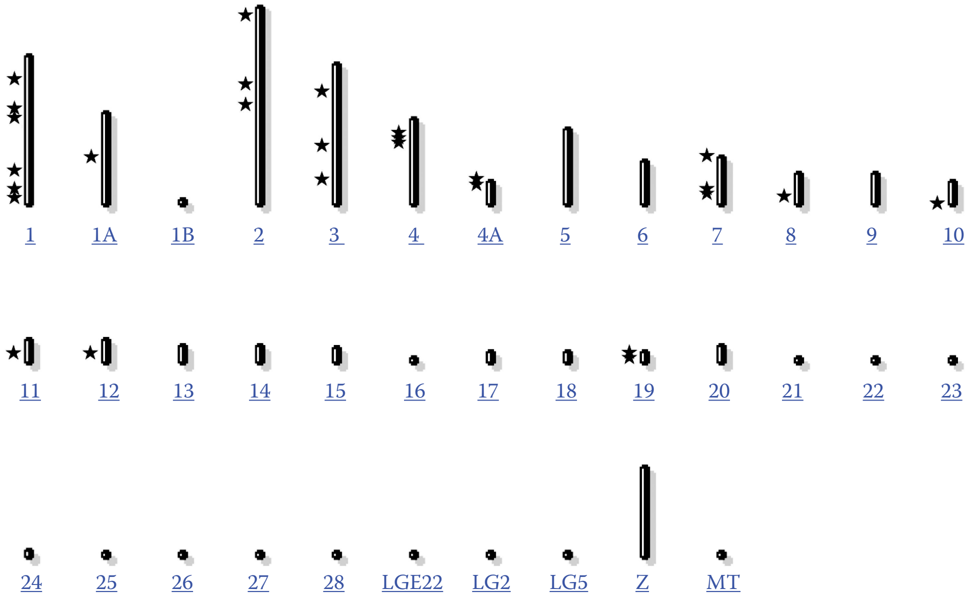


Figure 8.7. Chromosomal locations of 27 anonymous loci in the genome of the zebra finch. Loci were obtained from the genome of a closely related species of Australian grass finch (see Jennings and Edwards 2005). Each star indicates the chromosomal position of an anonymous locus. This figure was generated using the zebra finch genome (annotation release 101) on the NCBI genome browser site (www.ncbi.nlm.nih.gov/genome/browse/).

the zebra finch genome, 28/30 loci had only a single strongly supported genomic location. Results of *in silico* PCR of these loci against the zebra finch genome on the UCSC Genome Bioinformatics site (http://genome.ucsc.edu/cgi-bin/hgPcr) also did not reveal convincing evidence of any multiple copy loci. Thus, the vast majority of these loci appear to agree with the single-copy assumption despite the prevalence of transposable elements in their sequences. How can we explain this anomalous finding? Further reflection about the nature of replicative transposable elements and PCR principles reveals a simple explanation and shows why we need not worry about the possible adverse effects of repetitive DNA on the design of PCR-amplifiable anonymous loci.

To see why it may not be a problem to develop anonymous loci containing transposable elements, first consider a retroelement copy that resides in a particular chromosomal location of the zebra finch genome. If we were to place *both* PCR primers *within* this element as shown in Figure 8.8a, then it is possible that multiple copies of this element would subsequently be amplified in a single PCR reaction or different copies amplified from different individual genomes among PCR reactions—both scenarios would violate the single-copy assumption. However, recall that these genomic parasites are believed to insert themselves into random genomic locations—such as inside noncoding DNA, other transposable elements, coding regions, and introns. This means that the genomic landscape flanking each inserted retrotransposon is expected to be random with respect to the adjacent retrotransposon. Thus, if one primer is placed inside the element and the other primer is positioned *outside* of the element in the flanking region as shown in Figure 8.8b, then a subsequent PCR would not only amplify a single-copy of the element from a given genome, but also the same "orthologous" element would always be amplified in any closely related genome as well. Thus, as long as both PCR primers are not placed inside the *same* element, then the PCR is expected to generate only a single amplicon and thus satisfy the single-copy assumption in anonymous loci studies. The key here is that although each individual primer site might have multiple locations in a genome (especially for the primer within the retroelement), there is very likely only a single genomic location where *both* priming sites oppose each other on separate DNA strands and are capable of generating an amplicon that matches the

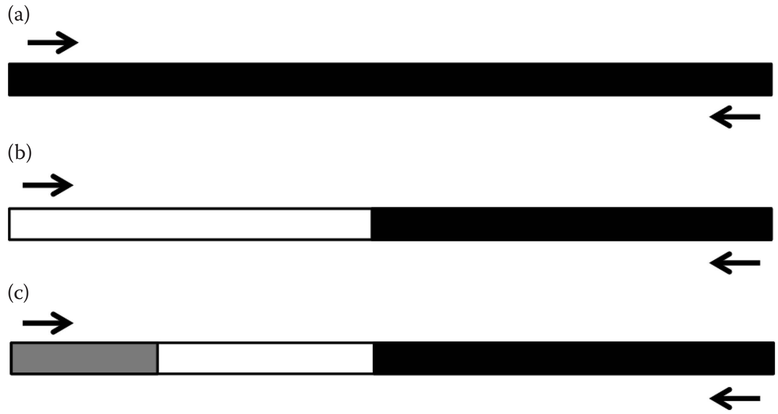


Figure 8.8 Forward and reverse PCR primers in relation to transposons embedded in a genome. (a) Forward and reverse primers (arrows) are positioned within a single transposon (black bar); (b) forward primer (above the white bar) is located outside the transposon while the reverse primer is inside (below the black bar); and (c) forward primer is located in one transposon (gray bar) and reverse primer is inside an adjacent but different transposon (black bar). The two shown transposons have independent insertion histories.

expected size. Given this principle, we can envision a third scenario shown in Figure 8.8c in which the forward and reverse primers are both inside *different* transposable elements; that is, both elements were inserted into the genomic DNA independently of each other regardless of their identity (i.e., independent insertion histories). Thus, like the scenario in Figure 8.8b, this third scenario is not expected to pose a problem in PCR. The scenarios illustrated in Figure 8.8b and 8.8c can explain the empirical results of Thomson et al. (P. 522, 2008) who observed no qualitative differences in PCR performance or DNA sequence characteristics between their apparently repeat-free anonymous loci versus a single locus that contained repetitive DNA. Returning to the grass finch example, of the 20 anonymous loci that contained one or more transposons, we see that only a single anonymous locus corresponds to the scenario in Figure 8.8a. In contrast, 19 loci matched the scenario in Figure 8.8b. Because a number of loci contained two or three distinct transposons, scenario "C" could have easily been observed if both primers were positioned in two separate transposons, but, by accident, this did not occur.

There is another good reason to expect that most anonymous loci will not violate the single-copy assumption. Recall from Chapter 2 that most retrotransposons are free to accumulate mutations from the moment they become inserted into the host genome. This means it is only a matter of time before primer-binding sites in one transposable element (or any other nonfunctional DNA not influenced by sites under selection) become mutated to the point that the primer cannot perform in PCR. Thus, primers developed for one neutral locus will likely only function for that same locus in other closely related genomes.

These findings suggest that most anonymous loci obtained from a genomic library will likely meet the single-copy assumption regardless of whether they contain transposable elements or not. This result has important implications, as we will see below when we consider NGS and bioinformatics approaches to anonymous loci development. Instead of the traditional practice of discarding candidate loci sequences that contain repetitive DNA, it may actually be advantageous to develop anonymous loci containing transposable elements. First, a larger number of genealogically independent anonymous loci can potentially be developed from a genome. Secondly, because most retroelements are no longer functional once they are inserted into a host genome, they will accumulate mutations that are not weeded out by natural selection and therefore they can be considered to meet the assumption of neutrality. The only issue with anonymous loci containing transposable elements would be their possible high GC content, which may result in elevated substitution rates at remaining CpG sites. However, a simple method for correcting this problem is to delete CpG sites from the multiple sequence alignment.

This would result in minimal loss of data and the remaining sites would likely have a single homogeneous substitution rate that reflects the natural spontaneous rate (Yang 2002; Jennings and Edwards 2005; Costa et al. 2016).

If a researcher still wants to eliminate possible multicopy loci from a sample of new loci, then a new and much simpler method can be used to remove these potentially poor quality loci. Instead of using *RepeatMasker* (Smit et al. 2004; Tarailo-Graovac and Chen 2009) to identify and discard template sequences containing repetitive DNA (which would likely be many sequences), a better strategy would be to use *RepeatMasker* (http://www.repeatmasker.org/; Smit et al. 2004) or the CENSOR software tool (Kohany et al. 2006) found on the *Repbase* site (http://www.girinst.org/censor/index.php; Jurka et al. 2005) to check each template sequence *after* the primer design step to determine whether or not the forward and reverse primers reside within the *same* transposable element. Any template sequence in which both primers are nested in the same element would be eliminated or redesigned prior to purchase and testing of new primers. This filtering step should ensure that all new loci meet the single-copy assumption. If the entire genome sequence of a closely related species is available, then *in silico* PCR and BLAST can be used to verify the single-copy nature of newly developed loci.

8.2.1.2 Single Template Methods Using Available Genomics Resources

During the early years of the genomics era the number of fully sequenced genomes in Genbank dramatically increased. At this time, which was just before NGS approaches to sequencing genomes entered the scene, the primary method for sequencing genomes consisted of shotgun cloning of genomic DNA, sequencing (Sanger) the ends of cloned inserts to a coverage at least 6× of the genome being sequenced, and then using bioinformatics tools to assemble the genome sequences (Venter et al. 2001). Forward and reverse vector (e.g., M13) sequencing primers were used to collect only the ends of the insert sequence because often the inserts were too long to be fully sequenced—particularly when the cloning vectors were bacterial artificial chromosomes (BAC) vectors, which could accommodate enormous 100–500 kb long chromosomal fragments. These genomic libraries were not only useful for sequencing entire genomes and obtaining particular chromosomal regions of interest to investigators (e.g., for gene evolution studies), but they could also be used for development of large numbers of genomic loci (Edwards et al. 2005; Shaffer and Thomson 2007; Thomson et al. 2008). By co-opting these genomics resources, researchers could have at their disposal the DNA templates for easily designing dozens or hundreds of different genomic loci without having to resort to genomic cloning methods.

For example, Shaffer and Thomson (2007) demonstrated the efficacy of this approach by designing a set of new anonymous loci for turtles using a BAC-end sequence library for the painted turtle (*Chrysemys picta*). "BAC-end" sequences are acquired simply by using the BAC-sequencing primers to sequence (via the Sanger method) the first 700–1,000 bases of the insert DNA molecule from one or both ends of the insert. Since BAC-end "reads" are generated from randomly selected BAC clones, they represent the sequences of anonymous chromosomal fragments and are, as such, from anonymous locations in the genome. To maximize the chance that new loci are single-copy and do not consist of repetitive genomic elements, Shaffer and Thomson (2007) used bioinformatics tools (*RepeatMasker*; Smit et al. 2004) to first filter away BAC-end sequences containing repetitive elements. Note, for reasons just discussed with the grass finch example, this step may not drastically improve the quality of the loci and thus it may be omitted. In the final step, primer design software is used to design a pair of primers from each retained template sequence followed by testing of the new primers on a panel of species that span the desired phylogenetic distances (e.g., intraspecific populations, family, order, etc.).

8.2.1.3 Single Template Methods Using NGS Partial Genome Data

More recently, NGS-based approaches for developing anonymous loci have been developed, which represent significant advances in loci development. Bertozzi et al. (2012) used 454 sequencing to generate a large number of random genomic sequences from low coverage libraries for a lizard and a snake species. After obtaining their raw sequencing data, the authors used a custom bioinformatics workflow to isolate high-quality sequence data, which could be used for candidate anonymous loci. This workflow included a number of filtering steps designed to discard sequences containing low quality scores, repetitive DNA, and known protein-coding regions. In the final step, the candidate loci sequences were fed into primer design software. In another NGS-based study, Lemmon and Lemmon (2012) presented a similar workflow for generating new anonymous loci but used an Illumina platform for generating raw sequence reads. Although both of these NGS studies included a filtering step to remove sequences containing repetitive DNA, this step can be replaced with the newer filtering strategy already suggested—i.e., eliminating candidate loci that have both primers within the same transposable element (see Figure 8.8a). These NGS-based approaches represent the best current methods for developing large numbers (tens to hundreds) of anonymous loci in species that have not yet had their genomes fully sequenced.

8.2.1.4 Single Template Methods Using Whole Genome Sequences

The previously described loci design methods involve a combination of laboratory work to generate candidate loci sequences (via Sanger or NGS platforms) followed by the use of bioinformatics tools to filter the sequences until a final set of single templates are acquired for the final primer design step. This loci development workflow can be made far simpler if a complete and annotated genome sequence is available for one of the target study organisms, as this would mean omitting the expensive and time-consuming laboratory step.

Chen and Li (2001) took advantage of the newly sequenced human genome when they pioneered an *in silico* method of anonymous locus development. Using human genome data and a computer-based workflow, these workers designed PCR primers to amplify 53 single-copy, presumably neutral, and independent anonymous loci. Once they obtained their set of primers, they used PCR and Sanger sequencing to obtain homologous sequences from the genomes of chimpanzee, gorilla, and orangutan to complement their human sequences. Their anonymous locus development workflow consisted of the following steps: (1) use a computer to find and retain all intergenic sequences; (2) from each intergenic region, retain one 2–20 kb segment if it is at least 5 kb from known functional genes in both directions—this step attempts to find sequences that are not strongly influenced by the possible effects of genetic hitchhiking or background selection (i.e., to satisfy the neutrality assumption); (3) use *RepeatMasker* (Smit et al. 2004) to mask repetitive DNA in each retained 2–20 kb segment (i.e., to satisfy the single-copy locus assumption); (4) retain each block of unmasked sequence ≥800 bp for use as a template sequence; and (5) design one PCR primer pair for each retained template sequence.

The main drawback with this approach is that only a miniscule number of organisms have had their genomes fully sequenced and even fewer have had their genomes annotated. Therefore, this approach has limited utility at the present time. However, in the future when more and more complete genomes become available, these *in silico*-based methods will become popular for designing large numbers of anonymous loci (and other types of loci).

8.2.2 Multiple Homologous Template Approaches for Designing PCR-Based and Anchor Loci

Primers designed using only a single template sequence usually work well for intraspecific studies or for studies involving a group of closely related species. However, when a researcher uses these same primers in attempts to amplify orthologous sequences in more evolutionarily distant taxa such as those in other genera, families, etc., then PCR will often fail. Such failures can be attributed to the fact that the primers were developed without consideration about the degree of evolutionary conservatism of the primer annealing sites.

As was discussed in Chapters 2 and 3, many genomic sites are undoubtedly free from evolutionary constraints (i.e., are not conserved) because they have no known function (e.g., intergenic DNA) or they cannot function (e.g., most retrotransposons) and hence there is nothing for natural selection to maintain (Graur et al. 2013). Thus, these sites are able to accumulate mutations (indels or substitutions) without any consequence for the organism. An implication of this is that the longer the time of divergence between two species, the more likely that the primer annealing sites will become mutated to the point that the primers will no longer function properly in cross-species PCR experiments.

A number of empirical studies have noticed this fall off in cross-species amplification success as the phylogenetic distance increases between primers and DNA being tested. Rosenblum et al. (2007), who developed 77 anonymous loci for the eastern fence lizard, observed that amplification success in cross-species PCR declined with increasing genetic distance. Similarly, Thomson et al. (2008) observed a similar drop in PCR success while testing 96 anonymous loci primer pairs on various turtle species; in particular, 50% of primer pairs had failed in PCR when divergences between the primers and species being tested moved into the 70–130 million year divergence range (see Figure 5 in Thomson et al. 2008). These authors further pointed out that for earlier microsatellite studies involving birds and mammals (Primmer et al. 1996) and fishes (Carreras-Carbonell et al. 2008), that this 50% failure mark (i.e., 50% of the loci fail to amplify) was reached when distances were only in the first tens of millions of years (Thomson et al. 2008). As to why anonymous loci for turtles are apparently able to perform well for cross-species PCRs spanning longer divergences than for birds, mammals, and fishes (i.e., >70 million years vs. 10s of millions of years, respectively), Thomson et al. (2008) hypothesized that this could be attributed to slower substitution rates in turtles. The microsatellite study on turtles by FitzSimmons et al. (1995) provides evidence supporting the hypothesis of Thomson et al. (2008), but this problem needs further study.

The aforementioned studies show that for some types of loci such as anonymous loci and microsatellites it is only a matter of time before one or more primer annealing sites become mutated in a manner that precludes successful PCR amplification. This also means that attempts to use primers developed for one species (or species group) on another species must be a trial and error process in the lab—the chance of success being inversely correlated with degree of evolutionary divergence between the sequences.

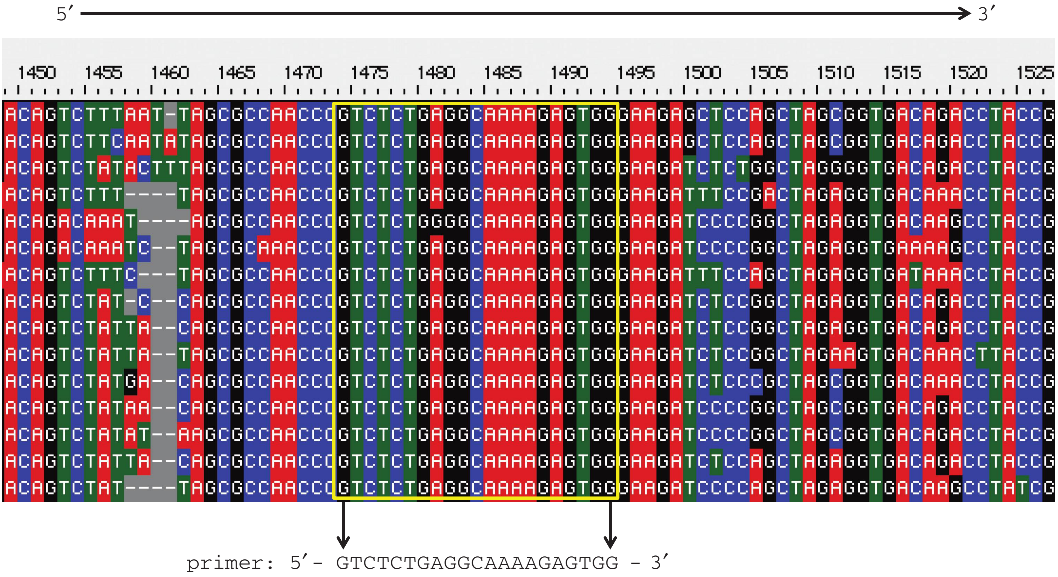
Of course, not all types of primers are vulnerable to the problem of primer sites decay because the degree of evolutionary conservatism of primer sites varies among types of loci. Universal primers are positioned on highly conserved sites of genomes, which can allow researchers to obtain homologous sequences for taxa spanning wide phylogenetic distances such as across genera, families, and orders. Universal primers have been developed for many nuclear loci such as exon and EPIC loci where the primer annealing sites are located within highly conserved portions of exons.

8.2.2.1 Designing Universal Primers by Comparative Sequence Analysis

In order to identify evolutionarily conserved sites suitable for primer placement, at least two homologous sequences must be aligned and compared site by site. Importantly, these sequences must span the evolutionary distances of the taxa for which the primers will be applied (Shaffer and Thomson 2007). Although two sequences may be sufficient for designing a pair of universal primers, the researcher can obtain a better understanding about within site variation at candidate primer locations by examining larger numbers of sequences in a multiple sequence alignment.

A visual inspection of a sequence alignment consisting of homologous sequences obtained from a number of different species can reveal blocks of nucleotide sites that are more evolutionarily conserved than other regions. An example of this type of analysis is shown in Figure 8.9, which shows an alignment of 15 sequences representing various hypothetical species of a study group. The block of highly conserved sites contained within the yellow square in Figure 8.9 can be used to design a new universal primer. Let's now take a closer look at the process of designing a universal primer.

First, note that the sequences in Figure 8.9 are shown in a 5′ → 3′ direction and that the target sequence is to the right (not shown). Thus, we are looking for a suitable location for a new "forward" primer. Notice that the sites (columns) on the left and right sides include many variable sites. The left side is particularly messy as it contains several sites with indels, which have resulted in the placement of a several uncertain gaps in the alignment. On the right side, the sites appear to be well aligned and gap-free but they still contain a number of variable sites. However, the region near the middle that includes sites #1463–1499 is the block that contains the fewest variable sites and so this stretch of sequence seems to represent the best option for designing a new forward primer.



```
ACAGTCTTTAAT-TAGCGCCAACCCGTCTCTGAGGCAAAAGAGTGGGAAGAGCTCCAGCTAGCGGTGACAGACCTACCG
ACAGTCTTCAATATAGCGCCAACCCGTCTCTGAGGCAAAAGAGTGGGAAGAGCTCCAGCTAGCGGTGACAGACCTACCG
ACAGTCTATACTTTAGCGCCAACCCGTCTCTGAGGCAAAAGAGTGGGAAGATCTCTGGCTAGGGGTGACAGACCTACCG
ACAGTCTTT----TAGCGCCAACCCGTCTCTGAGGCAAAAGAGTGGGAAGATTTCCGACTAGAGGTGACAAACCTACCG
ACAGACAAAT----AGCGCCAACCCGTCTCTGGGGCAAAAGAGTGGGAAGATCCCCGGCTAGAGGTGACAAGCCTACCG
ACAGACAAATC--TAGCGCAAACCCGTCTCTGAGGCAAAAGAGTGGGAAGATCCCCGGCTAGAGGTGAAAAGCCTACCG
ACAGTCTTTC---TAGCGCCAACCCGTCTCTGAGGCAAAAGAGTGGGAAGATTTCCAGCTAGAGGTGATAAACCTACCG
ACAGTCTAT-C--CAGCGCCAACCCGTCTCTGAGGCAAAAGAGTGGGAAGATCTCCGGCTAGAGGTGACAGACCTACCG
ACAGTCTATTA--CAGCGCCAACCCGTCTCTGAGGCAAAAGAGTGGGAAGATCTCCGGCTAGAGGTGACAGACCTACCG
ACAGTCTATTA--TAGCGCCAACCCGTCTCTGAGGCAAAAGAGTGGGAAGATCTCCGGCTAGAAGTGACAAACTTACCG
ACAGTCTATGA--CAGCGCCAACCCGTCTCTGAGGCAAAAGAGTGGGAAGATCTCCCGCTAGCGGTGACAAACCTACCG
ACAGTCTATAA--CAGCGCCAACCCGTCTCTGAGGCAAAAGAGTGGGAAGATCCCCGGCTAGCGGTGACAGACCTACCG
ACAGTCTATAT--AAGCGCCAACCCGTCTCTGAGGCAAAAGAGTGGGAAGATCCCCGGCTAGCGGTGACAGACCTACCG
ACAGTCTATTA--CAGCGCCAACCCGTCTCTGAGGCAAAAGAGTGGGAAGATCTCCAGCTAGAGGTGACAGACCTACCG
ACAGTCTAT----TAGCGCCAACCCGTCTCTGAGGCAAAAGAGTGGGAAGATCCCCAGCTAGAGGTGACAAGCCTATCG
```

```
primer: 5′- GTCTCTGAGGCAAAAGAGTGG - 3′
```

Figure 8.9. Design of a nondegenerate universal primer within a conserved block of DNA sequences. Shown is a multiple sequence alignment for 15 hypothetical species in a study group. Sites #1474–1494 (area within yellow rectangle) were used to design a new forward primer, which is shown below the alignment. Note that site #1481 is the only observed variable site within the primer design region. Sequences were aligned by eye using the software Se–Al (Rambaut 2007) and the 5′ → 3′ polarity of the sequences is from left to right.

Accordingly, sites #1474–1494 (delineated by yellow) were selected for the new primer. Notice that within this particular stretch of sequence there is only a single variable site (site #1481; A/G transition). This variable site will likely be acceptable to be included because single variable sites in the middle or at the 5′ end of a primer generally do not adversely affect PCR. Because this variable site consists of only two different bases among all sequences and only one sequence shows a "G" nucleotide, we can just use the most prevalent base at this position ("A"). The newly designed primer is shown below the alignment in Figure 8.9. Using the online program *Oligoanalyzer* 3.1 (http://www.idtdna.com/calc/analyzer; Owczarzy et al. 2008) we can check the physical characteristics of this primer:

21 bp long (satisfies Rule 1)

G + C% = 52.4% (satisfies Rule 2)

No stable homodimers at 3′ end (satisfies Rule 3)

$\Delta G_{\text{pentamer}} = -7.96$ kcal/mol (satisfies Rule 4)

$\Delta G_{\text{hairpin}} = -1.05$ kcal/mol (satisfies Rule 5)

Tm = 56°C (satisfies Rule 6)

No mismatches at 3′ end of primer (satisfies Rule 8)

If a "reverse" primer is also designed to pair with this forward primer, then Rule 3 (no stable heterodimers) and Rule 7 (Tm between primers <5°C) can also be evaluated. Although it may be tempting to add one additional base to the 3′ end of the primer (site #1495), the effect of adding this extra "G" base would be to increase the "stickiness" of the 3′ end of the primer (i.e., decrease $\Delta G_{\text{pentamer}}$ to −9.43 kcal/mol), which would violate Rule 4. Thus, it is preferable to not include this extra base.

In the aforementioned example, nearly all sites within the chosen primer sequence were invariable (conserved). However, it is often not possible to find such a highly conserved stretch of sites in which a new primer can be placed—remember that we need to find a contiguous stretch of sites that is *at least* 18 bp long in order to design a new primer. Figure 8.10 shows a different stretch of bases from the genomes of the same 15 hypothetical species used in the Figure 8.9 example. First, notice that the left and right sides again do not contain blocks of invariant sites long enough to place a primer. The least variable region,

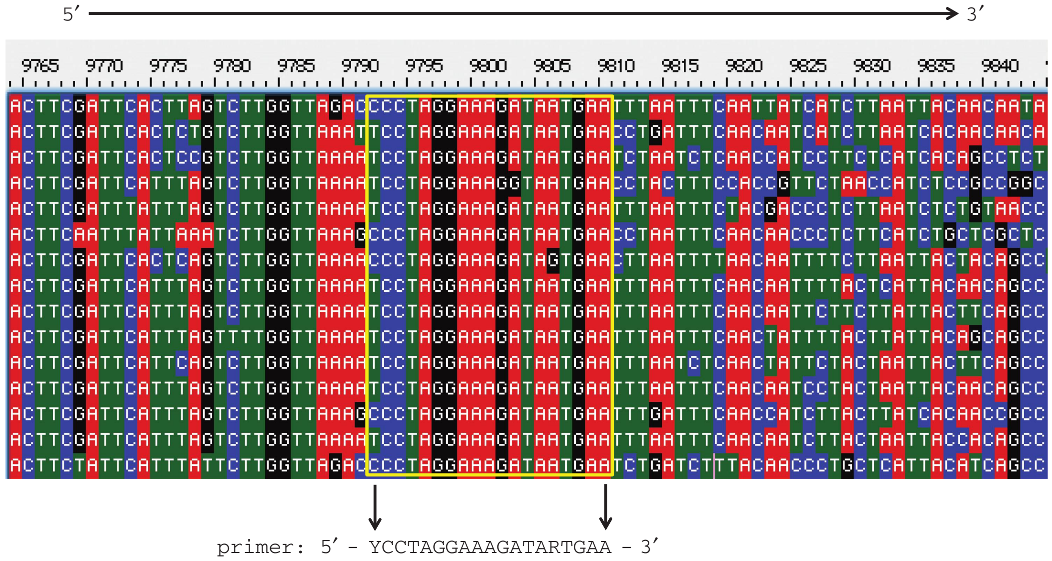

5′
3′
9765
9770
9775
9780
9785
9790
9795
9800
9805
9810
9815
9820
9825
9830
9835
9840
ACTTCGATTCACTTAGTCTTGGTTAGACCCCTAGGAAAGATAATGAATTTAATTTCAATTATCATCTTAATTACAACAATA
ACTTCGATTCACTCTGTCTTGGTTAAATTCCTAGGAAAGATAATGAACCTGATTTCAACAATCATCTTAATCACAACAACA
ACTTCGATTCACTCCGTCTTGGTTAAAATCCTAGGAAAGATAATGAATCTAATCTCAACCATCCTTCTCATCACAGCCTCT
ACTTCGATTCATTTAGTCTTGGTTAAAATCCTAGGAAAGGTAATGAACCTACTTTCCACCGTTCTAACCATCTCCGCCGGC
ACTTCGATTTATTTAGTCTTGGTTAAAATCCTAGGAAAGATAATGAATTTAATTTCTACGACCCTCTTAATCTCTGTAACC
ACTTCAATTTATTAAATCTTGGTTAAAGCCCTAGGAAAGATAATGAACCTAATTTCAACAACCCTCTTCATCTGCTCGCTC
ACTTCGATTCACTCAGTCTTGGTTAAAACCCTAGGAAAGATAGTGAACTTAATTTTAACAATTTTCTTAATTACTACAGCC
ACTTCGATTCATTTAGTCTTGGTTAAAATCCTAGGAAAGATAATGAATTTAATTTCAACAATTTTACTCATTACAACAGCC
ACTTCGATTCATTTAGTCTTGGTTAAAATCCTAGGAAAGATAATGAATTTAATTTCAACAATTCTTCTTATTACTTCAGCC
ACTTCGATTCATTTAGTTTTGGTTAAAATCCTAGGAAAGATAATGAATTTAATTTCAACTATTTTACTTATTACAGCAGCC
ACTTCGATTCATTCAGTCTTGGTTAAAATCCTAGGAAAGATAATGAATTTAATCTCAACTATTCTACTAATTACTTCAGCC
ACTTCGATTCATTTAGTCTTGGTTAAAATCCTAGGAAAGATAATGAATTTAATTTCAACAATCCTACTAATTACAACAGCC
ACTTCGATTCATTTAGTCTTGGTTAAAGCCCTAGGAAAGATAATGAATTTGATTTCAACCATCTTACTTATCACAACCGCC
ACTTCGATTCATTTAGTCTTGGTTAAAATCCTAGGAAAGATAATGAATTTAATTTCAACAATCTTACTAATTACCACAGCC
ACTTCTATTCATTTATTCTTGGTTAGACCCCTAGGAAAGATAATGAATCTGATCTTTACAACCCTGCTCATTACATCAGCC
primer: 5′ - YCCTAGGAAAGATARTGAA - 3′

Figure 8.10. Design of a degenerate universal primer within a conserved block of DNA sequences. Shown is a multiple sequence alignment for 15 hypothetical species in a study group. Sites #9792–9810 (area within yellow rectangle) were used to design a new forward primer, which is shown below the alignment. This primer contains two degenerate sites ("Y" and "R") and therefore it is a 4-fold degenerate primer. Sequences were aligned by eye using the software Se–Al (Rambaut 2007) and the 5′ → 3′ polarity of the sequences is from left to right.

which includes sites #9792–9810, is found in the center (Figure 8.10). This particular 19 bp stretch of sequence, which is outlined in yellow, might suffice for placing a forward primer, but notice that it includes three variable sites (#9792, 9803, and 9806). However, because each of these three sites exhibits simple transition type substitutions (T/C, A/G, and A/G, respectively), it may be possible to design the primer using the most prevalent base at each of the variable sites (i.e., "T," "A," and "A," respectively). However, a primer that is only 19 bp long may not perform well in PCR if it has this many mismatches with the genomic DNA templates (longer primers can better tolerate mismatches).

One effective strategy for accommodating variable sites in a candidate primer location is to design a special type of universal primer called a *degenerate primer*. A degenerate primer is, in reality, a mixture of primers that are usually the same length but exhibit different bases at one or more sites. The rationale for using a mixture of primers is that during a PCR reaction at least one of the primers in the mixture will be the correct match (or close enough) with the target template for amplification to occur. Some mismatched bases can be tolerated as long as there are not too many of them and provided that they are at the 5′ end or in the center of the primer; mismatches located at the 3′ of the primer should be avoided. Thus, we can design a degenerate primer using the sequences in Figure 8.10 by designating each of the three variable sites with the appropriate IUPAC ambiguity code (see Table 6.4). This means that site #9792 is now indicated as "Y" to represent a "T" or "C" base; site #9803 as "R" to represent an "A" or "G"; and site #9806 as "R" to represent an "A" or "G." Because there are two possible bases at each degenerate site, this would be known as an "8-fold" degenerate primer (2 × 2 × 2) meaning that the actual primer mixture used in PCR would consist of eight different primer sequences representing all possible sequences. Similarly, if another primer had four degenerate sites consisting of "R," "Y," "D," and "N" codes at each degenerate site (Table 6.4), then this degenerate primer would exhibit 2 × 2 × 3 × 4 = 48-fold level of degeneracy and hence contain 48 unique primers! On the other hand, a primer with no degenerate sites would be 0-fold degenerate and thus consist of a single primer.

When designing degenerate primers it is critically important to remember that a tradeoff exists here: the more degenerate the primer (i.e., larger numbers of variant primers), the greater the chance of amplifying nontarget products including paralogous sequences (Chenuil et al. 2010). For example, Chenuil et al. (2010) found that their EPIC loci primers, which were developed for obtaining orthologous sequences throughout the metazoan tree of life, often performed poorly when the primers exhibited greater than 8-fold level of degeneracy. Primers of varying levels of degeneracy can still perform well in PCR, but it is advisable to minimize the levels of degeneracy in a primer whenever possible. In the example shown in Figure 8.10 we might elect to not make site #9803 a degenerate site because a single mismatch in the middle of the primer will likely not affect the primer's performance. Thus, the preferable strategy is to minimize levels of degeneracy in this primer, which gives us the 4-fold degenerate primer shown at the bottom of Figure 8.10. Using the *Oligoanalyzer* 3.1 tool (http://www.idtdna.com/calc/analyzer; Owczarzy et al. 2008) we can examine this primer's physical characteristics:

19 bp long (satisfies Rule 1)

G + C% = 36.8% (satisfies Rule 2)

No stable homodimers at 3′ end (satisfies Rule 3)

$\Delta G_{\text{pentamer}} = -6.82$ to -6.95 kcal/mol (satisfies Rule 4)

$\Delta G_{\text{hairpin}} = -0.02$ kcal/mol (satisfies Rule 5)

Tm = 44.4–47.8°C (violates Rule 6)

No mismatches at 3′ end of primer (satisfies Rule 8).

Again, if a "reverse" primer is also designed to pair with this forward primer, then Rule 3 (no stable heterodimers) and Rule 7 (Tm between primers <5°C) can also be evaluated. Note that because there is a mixture of primers involved in one degenerate primer (four different primer sequences), some of the primer's characteristics (above) are presented as minimum and maximum values. Also, you may have noticed that the 5′ terminal base located at site #9792 is variable and thus maybe you wondered if it could be omitted from the primer to further reduce the level of degeneracy. Doing this, however, would lower Tm of the primer down to the 42.4–44.6°C range, which might be too low for successful PCR. Remember that the Tm difference between forward and reverse primers should be <5°C (Rule 7). Primers with Tm values in this range can work well but it is best to strive for making primers with a Tm at least 50°C (Rule 6). These fine-tuning steps during the primer design process, which include deciding the exact span of sites to include and levels of degeneracy, represent a critical part of universal primer design.

8.2.2.2 Multiple Homologous Template Approaches Using Whole Genome Sequences

In an effort to address the longstanding paucity of available DNA sequence loci for phylogenomic studies concerned with inferring higher level organismal phylogenies, a number of innovative new genome-enabled methodologies for developing large numbers of exon, EPIC, and anonymous loci have been proffered in recent years. Similar to the single template whole genome methods, these *in silico*-based methods take advantage of existing whole genome sequences in databases such as ENSEMBL to find appropriate genomic regions for loci development. The main difference between the single and multiple template approaches is that the latter requires at least two whole genome sequences so that conserved primer sites can be identified.

An example of this approach is illustrated in the study by Li et al. (2007) who constructed a bioinformatics pipeline to find a large number of exon loci that could be useful in higher-level studies of ray-finned fishes, a clade that comprises about half of all extant vertebrate species. Figure 8.11 shows the bioinformatics pipeline developed by Li et al. (2007). First, the genome of the reference species (zebrafish) is downloaded from the ENSEMBL database before it is searched for all likely exonic segments of DNA, which consist of ORFs >800 bp. The resulting ORFs (probable exons) are then BLAST searched against the reference genome in an effort to filter out those ORFs that are not single-copy. The single-copy exons are then BLAST searched against a second query genome to narrow the list of exons to only those that are shared (conserved) between both genomes. The next step is to align homologous exon sequences and design primers.

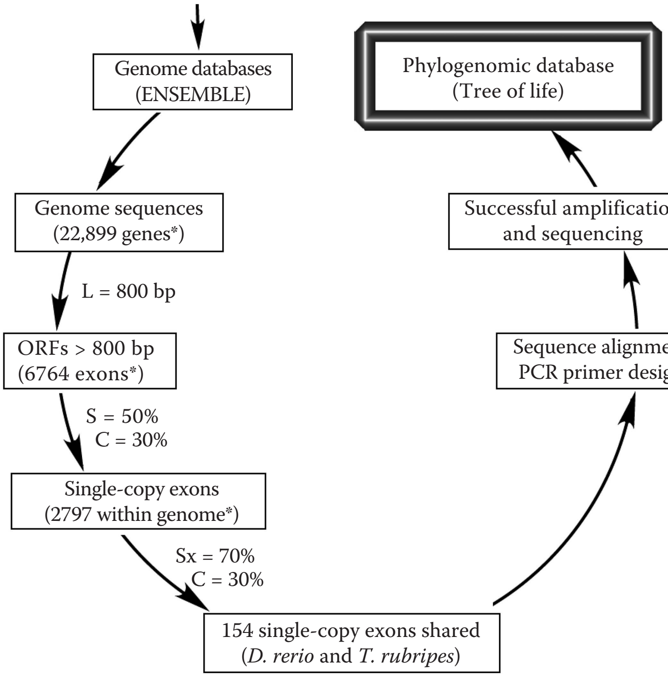


Figure 8.11. Bioinformatics pipeline for *in silico* development of large numbers of exon loci using one annotated whole genome sequence and at least one other whole genome sequence. (Reprinted from Li, C. et al. 2007. *BMC Evol Biol* 7:44. With permission.)

Interestingly, this last step is the only "low tech" step as the primer design is done by eye—similar to that described earlier for designing universal primers. If the sequences are divergent enough from each other, yet still alignable, then the primers will likely require one or more degenerate sites. If additional homologous sequences obtained from other species can be included in the alignment, then additional variation can be accounted for in the primers (i.e., the degeneracy of the primers is increased) making them even more universal in their ability to amplify the target locus on widely divergent species. Thus, even though this locus design approach can be largely accomplished using complete genome data and bioinformatics software, the actual design of the primers still relies on the investigator's knowledge about the rules of primer design. Comparable methods for designing large numbers of exon loci that can be used on phylogenetically diverse taxa come from the studies by Townsend et al. (2008) and Portik et al. (2012). Whole genome-based methods have also been devised for developing large numbers of EPIC loci (Backström et al. 2008; Chenuil et al. 2010; Li et al. 2010) and anonymous loci (Peng et al. 2009; Wenzel and Piertney 2015).

8.2.2.3 Designing Anchor Loci Probes Using Whole Genome Sequences

If you are planning to undertake a phylogenomic study that will require RNA bait sets for in-solution hybrid selection, then you should first check to see if RNA probe sets that could be used in your study already exist. For example, there are a number of UCE- and AE-anchored loci probe sets already available for vertebrates (e.g., Faircloth et al. 2012, 2013; Lemmon et al. 2012; Prum et al. 2015).

If suitable RNA bait sets do not yet exist for your study group, then you will need to develop your own probe set. Probes used for custom RNA bait kits are designed using complete genome data. Once designed, the new bait set can be produced from a company that offers this service. For example, custom RNA bait kits can be ordered from Agilent, Inc. (SureSelect kits) and from Macroarray (MYbaits kits).

How are RNA bait sets made? For highly conserved loci such as UCE- and AE-anchored loci, multiple genomes (at least two), which bracket the evolutionary divergences in the clade of interest, are first aligned to each other so that highly conserved regions can be identified as candidate loci regions. From this pool of candidate regions, probe sequences are designed in the most-conserved stretches of sites and possibly in less-conserved flanking sequences (see below). Because the types of RNA probes used in in-solution hybrid selection reactions are usually 120 bp long, which is shorter than typical phylogenomic loci, a number of additional probes are often designed in order to catch all adapter-ligated library fragments that contain some portion of a target locus sequence. Thus, for each candidate locus region, the initial probe is usually placed in the center of the most-conserved stretch of sites (e.g., middle of a UCE) and then additional probes are designed in the regions on both sides of the first probe. This practice of using multiple probes per probe region is termed *tiling*. Different tiling schemes have been used such as placing probes in a nonoverlapping or sequential manner to cover each protein-coding exon (Gnirke et al. 2009). Alternatively, and more commonly, probes are placed in overlapping configurations to some degree. For example, in some UCE-anchored loci studies (e.g., Faircloth et al. 2012, 2015), a tiling density of 2× was used whenever UCEs were >180 bp long; that is, 120 bp-long probes were tiled such that adjacent probes overlapped each other by 60 bp. In another UCE-anchored loci study, Faircloth et al. (2013) used a 4× tiling density. A much denser tiling strategy was employed in the first AE-anchored loci study (Lemmon et al. 2012), as the authors positioned a new probe every five bp along each distinct probe region. This meant that each probe region was represented by as many as 25 largely overlapping probes. In a more recent study using AE-anchored loci, Prum et al. (2015) used a 1.5× tiling density. Although probe-tiling density may be optimized for different groups of organisms, the enormous multilocus datasets generated in each of the aforementioned studies attests to the efficacy of using even a nonoptimized tiling density.

For additional details about how UCE-anchored probe sets are made see Faircloth et al. (2012, 2013, 2015) and the website (http://ultraconserved.org). Additionally, software called UCE-Probe-Design-Program or "UPDP" has been developed to help researchers design UCE probe sets (see http://github.com/faircloth-lab/uce-probe-design). For more information about how AE-anchored loci are developed see Lemmon et al. (2012), Prum et al. (2015), and the following website for additional information (http://anchoredphylogeny.com).

REFERENCES

Abd-Elsalam, K. A. 2003. Bioinformatic tools and guideline for PCR primer design. *Afr J Biotechnol* 2:91–95.

Amaral, F. R., S. V. Edwards, and C. Y. Miyaki. 2012. Eight anonymous nuclear loci for the squamate antbird (*Myrmeciza squamosa*), cross-amplifiable in other species of typical antbirds (Aves, Thamnophilidae). *Conserv Genet Resour* 4:645–647.

Backström, N., S. Fagerberg, and H. Ellegren. 2008. Genomics of natural bird populations: A gene-based set of reference markers evenly spread across the avian genome. *Mol Ecol* 17:964–980.

Bertozzi, T., K. L. Sanders, M. J. Sistrom, and M. G. Gardner. 2012. Anonymous nuclear loci in non-model organisms: making the most of high throughput genome surveys. *Bioinformatics* 28:1807–1810.

Breslauer, K. J., R. Frank, H. Blöcker, and L. A. Marky. 1986. Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci USA* 83:3746–3750.

Carreras-Carbonell, J., E. Macpherson, and M. Pascual. 2008. Utility of pairwise mtDNA genetic distances for predicting cross-species microsatellite amplification and polymorphism success in fishes. *Conserv Genet* 9:181–190.

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

Chenuil, A., T. B. Hoareau, E. Egea et al. 2010. An efficient method to find potentially universal population genetic markers, applied to metazoans. *BMC Evol Biol* 10:276.

Chou, Q., M. Russell, D. E. Birch, J. Raymond, and W. Bloch. 1992. Prevention of pre-PCR mis-priming and primer dimerization improves low-copy-number amplifications. *Nucleic Acids Res* 20:1717–1723.

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

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

Faircloth, B. C., M. G. Branstetter, N. D. White, and S. G. Brady. 2015. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Mol Ecol Resour* 15:489–501.

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

Faircloth, B. C., L. Sorenson, F. Santini, and M. E. Alfaro. 2013. A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). *PLoS One* 8:p.e65923.

FitzSimmons, N. N., C. Moritz, and S. S. Moore. 1995. Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. *Mol Biol Evol* 12:432–440.

Folmer, O., M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek, R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotechnol* 3:294.

Freier, S. M., R. Kierzek, J. A. Jaeger et al. 1986. Improved free-energy parameters for predictions of RNA duplex stability. *Proc Natl Acad Sci USA* 83:9373–9377.

Gibbs, Josiah Willard. Wikipedia Page. http://en.wikipedia.org/wiki/Josiah_Willard_Gibbs (accessed September 5, 2015).

Gnirke, A., A. Melnikov, J. Maguire et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnol* 27:182–189.

Graur, D., Y. Zheng, N. Price, R. B. Azevedo, R. A. Zufall, and E. Elhaik. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* 5:578–590.

Haas, S., M. Vingron, A. Poustka, and S. Wiemann. 1998. Primer design for large scale sequencing. *Nucleic Acids Res* 26:3006–3012.

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

Innis, M. A. and D. H. Gelfand. 1990. Chapter 1. Optimization of PCRs. In *PCR Protocols: A Guide to Methods and Applications*, eds. M. A. Innis, D. H. Gelfand, J. J. Sninsky, and T. J. White, 3–12. New York: Academic Press.

Innis, M. A., D. H. Gelfand, J. J. Sninsky, and T. J. White, eds. 1990. *PCR Protocols: A Guide to Methods and Applications*. New York: Academic Press.

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

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. Repbase update; a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110:462–467.

Karl, S. A. and J. C. Avise. 1993. PCR-based assays of mendelian polymorphisms from anonymous single-copy nuclear DNA: Techniques and applications for population genetics. *Mol Biol Evol* 10:342–361.

Kibbe, W. A. 2007. OligoCalc: An online oligonucleotide properties calculator. *Nucleic Acids Res* 35:W43-W46.

Kocher, T. D., W. K. Thomas, A. Meyer et al. 1989. Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. *Proc Natl Acad Sci USA* 86:6196–6200.

Kohany, O., A. J. Gentles, L. Hankus, and J. Jurka. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and censor. *BMC Bioinf* 7:474.

Koressaar, T. and M. Remm. 2007. Enhancements and modifications of primer design program Primer3. *Bioinformatics* 23:1289–1291.

Kwok, S., D. E. Kellogg, N. McKinney, D. Spasic, C. Levenson, and J. J. Sninsky. 1990. Effects of primer–template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies. *Nucleic Acids Res* 18:999–1005.

Lemmon, A. R., S. A. Emme, and E. M. Lemmon. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst Biol* 61:727–744.

Lemmon, A. R. and E. M. Lemmon. 2012. High-throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. *Syst Biol* 61:745–761.

Li, C., G. Ortí, G. Zhang, and G. Lu. 2007. A practical approach to phylogenomics: The phylogeny of ray-finned fish (Actinopterygii) as a case study. *BMC Evol Biol* 7:44.

Li, C., J.-J. M. Riethoven, and L. Ma. 2010. Exon-primed intron-crossing (EPIC) markers for non-model teleost fishes. *BMC Evol Biol* 10:90.

Lincoln, S. E., M. J. Daly, and E. S. Lander. 1991. *PRIMER*: A computer program for automatically selecting PCR primers. MIT Center for Genome Research and Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts,

2142. ftp://ftp.broadinstitute.org/distribution/software/Primer0.5/readme.txt (accessed September 17, 2016).

Owczarzy, R., A. V. Tataurov, Y. Wu et al. 2008. IDT SciTools: A suite for analysis and design of nucleic acid oligomers. *Nucleic Acids Res* 36:W163–W169.

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.

Peng, Z., N. Elango, D. E. Wildman, and V. Y. Soojin. 2009. Primate phylogenomics: Developing numerous nuclear non-coding, non-repetitive markers for ecological and phylogenetic applications and analysis of evolutionary rate variation. *BMC Genomics* 10:247.

Portik, D. M., P. L. Wood Jr., J. L. Grismer, E. L. Stanley, and T. R. Jackman. 2012. Identification of 104 rapidly-evolving nuclear protein-coding markers for amplification across scaled reptiles using genomic resources. *Conserv Genet Resour* 4:1–10.

Primmer, C. R., A. P. Møller, and H. Ellegren. 1996. A wide-range survey of cross-species microsatellite amplification in birds. *Mol Ecol* 5:365–378.

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

Rambaut, A. 2007. *Se-Al, version 2.0 a11*. Edinburgh: University of Edinburgh.

Rosenblum, E. B., N. M. Belfiore, and C. Moritz. 2007. Anonymous nuclear markers for the eastern fence lizard, *Sceloporus undulatus*. *Mol Ecol Notes* 7:113–116.

Rozen, S. and H. Skaletsky. 1999. Primer3 on the WWW for general users and for biologist programmers. In *Bioinformatics methods and protocols*, ed. S. Misener, 365–386. New York: Humana Press.

Rychlik, W. 1995. Selection of primers for polymerase chain reaction. *Mol Biotechnol* 3:129–134.

Rychlik, W. and R. E. Rhoads. 1989. A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and *in vitro* amplification of DNA. *Nucleic Acids Res* 17:8543–8551.

Sambrook J., E. F. Fritsch, and T. Maniatis. 1989. *Molecular Cloning: A Laboratory Manual*, 2nd edition. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

SantaLucia, J., Jr. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA* 95:1460–1465.

Shaffer, H. B. and R. C. Thomson. 2007. Delimiting species in recent radiations. *Syst Biol* 56:896–906.

Smit, A., R. Hubley, and P. Green. 2004. *RepeatMasker Open-3.0*. Available from http://www.Repeatmasker.org.

Suggs, S. V., T. Hirose, E. H. Miyake, M. J. Kawashima, K. I. Johnson, and R. B. Wallace. 1981. In *Developmental Biology Using Purified Genes*, ed. D. D. Brown, 23:683–693. New York: Academic Press.

Tarailo-Graovac, M. and N. Chen. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* 25:4.10.1–4.10.14.

Thein, S. L. and R. B. Wallace. 1986. The use of synthetic oligonucleotides as specific hybridization probes in the diagnosis of genetic disorders. In *Human Genetic Diseases: A Practical Approach*, ed. K. E. Davis, 33–50. Herndon: IRL Press.

Thomson, R. C., A. M. Shedlock, S. V. Edwards, and H. B. Shaffer. 2008. Developing markers for multilocus phylogenetics in non-model organisms: A test case with turtles. *Mol Phylogenet Evol* 49:514–525.

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

Townsend, T. M., R. E. Alegre, S. T. Kelley, J. J. Wiens, and T. W. Reeder. 2008. Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: An example from squamate reptiles. *Mol Phylogenet Evol* 47:129–142.

Untergasser, A., I. Cutcutache, T. Koressaar et al. 2012. Primer3 – new capabilities and interfaces. *Nucleic Acids Res* 40:e115.

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

Warren, W. C., D. F. Clayton, H. Ellegren et al. 2010. The genome of a songbird. *Nature* 464:757–762.

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.

Wenzel, M. A. and S. B. Piertney. 2015. In silico identification and characterisation of 17 polymorphic anonymous non-coding sequence markers (ANMs) for red grouse (*Lagopus lagopus scotica*). *Conserv Genet Resour* 7:319–323.

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

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