CHAPTER SEVEN

Illumina Sequencing

Illumina sequencing is revolutionizing phylogenomics in the much the same manner as Sanger sequencing had spurred the growth of molecular phylogenetics. Although this technology was originally developed for purposes of obtaining whole genome sequences, phylogenomics researchers are primarily using it to target large numbers of specific genomic loci. This more focused approach enables researchers to acquire dozens to thousands of loci from genomes of many different individuals all in a single sequencing experiment (Lemmon and Lemmon 2013). Indeed, Illumina sequencing is proving to be a powerful means by which enormous phylogenomic datasets can be obtained.

One class of methods known as *hybrid selection* has effectively been coupled with Illumina sequencing in order to generate targeted genomic datasets (Turner et al. 2009). Hybrid selection, which is also commonly referred to as "target capture," involves the hybridization of sequence-specific oligonucleotide probes to a large number of target locations in genomes followed by isolation of these templates for a single mass sequencing run (Turner et al. 2009). For example, Faircloth et al. (2012) and Lemmon et al. (2012) developed probes for sequencing large numbers of UCE-anchored and AE-anchored loci, respectively. Large numbers (e.g., 25–100) of different PCR products from dozens to hundreds of individuals can also be sequenced using the Illumina platform (e.g., Barrow et al. 2014). These giant-sized datasets now routinely obtained by researchers would have been almost inconceivable a decade ago. Among these methods, hybrid selection represents the most powerful approach owing to its capability of generating datasets with thousands of DNA-sequence loci, which are long enough (i.e., ~250–1,500 bp) to allow for the robust inferences of gene trees. Although sequencing of PCR amplicons using Illumina sequencers is effectively limited to datasets with fewer than 100 loci, these datasets can still provide robust phylogenomic results. Other types of loci being used in phylogenomic studies such as RAD-seq loci and single-nucleotide polymorphisms (SNPs) generally provide poorly reconstructed gene trees. Thus, we will focus our attention on hybrid selection and PCR loci-based methods in this chapter because both approaches usually yield sequences long enough to produce well-constructed gene trees. For additional information about how Illumina sequencing is being used to generate other types of phylogenomic datasets the reader should consult reviews by Lemmon and Lemmon (2013), McCormack et al. (2013), and Toews et al. (2015).

7.1 HOW ILLUMINA SEQUENCING WORKS

The Illumina sequencing workflow consists of three basic steps: (1) construction of indexed sequencing libraries; (2) generation of clusters on a flow cell; and (3) sequencing of clusters. Like Sanger sequencing this workflow begins with the preparation of sequencing templates from input DNA samples such as genomic DNA, short or long PCR products, cDNAs, etc. However, while the former sequencing method uses PCR products directly as sequencing templates, the latter uses sequencing libraries. Furthermore, as organismal phylogenomic studies include DNA samples from multiple individuals or species, one library must be made for each input sample.

Thus, each library must be tagged with at least one individual-specific *index* sequence so that they can be sequenced together in a single Illumina sequencing run, a practice called *multiplexing* (Cronn et al. 2008). Library constructs may have a single index sequence, which ranges from 7 to 10 bp long (depending on the kit or protocol) or, more usually, they have dual indices. Later, during the bioinformatics processing of sequencing data, software is used to *deconvolute* the sequences (i.e., sequences are sorted by individual according to their indices).

Once the desired libraries are acquired, the next step is to generate *in situ* clonal colonies of target DNA templates called *clusters*. Each cluster comprises of ~1,000 ssDNA molecules derived from each target library fragment (Shendure and Ji 2008). As with Sanger sequencing, Illumina sequencing also relies on detecting and imaging base-specific light signals emitted from fluorescently labeled nucleotides during the sequencing process. However, because the imaging devices found on Sanger and second-generation NGS sequencers such as Illumina are not sensitive enough to detect light emissions from single molecules, large numbers of identical molecules must first be generated via PCR before they can be sequenced. Thus, when these populations of molecules are subsequently tagged with fluorescent bases and are illuminated by the sequencer's laser, the emitted light signals from each population are strong enough to be detected and imaged (Shendure et al. 2004).

The earliest version of the Illumina platform generated sequences only 25–35 bp long (e.g., Bentley et al. 2008; Cronn et al. 2008). Since then, however, Illumina has increased their read lengths up to 100–300 bp depending on the sequencer model. While the "short-read" nature of Illumina sequencing is one disadvantage compared to the longer Sanger reads, the hundreds of millions to *billions* of reads output per run by some Illumina platforms is nothing short of mind boggling. It is the massively parallel nature of this form of DNA sequencing that gives it tremendous sequencing power, which is why this and other NGS technologies are commonly referred to as "massively parallel sequencing" (Shendure et al. 2004; Mardis 2008).

In addition to increasing their read lengths, Illumina introduced an innovation called *paired-end sequencing* to facilitate the construction of *de novo* genomes particularly for organisms having large genomes (see Brown 2013 and Masoudi-Nejad et al. 2013 for further explanations). Paired-end or "PE" sequencing is a form of DNA sequencing whereby forward and reverse sequences of each target library fragment are obtained. For example, if a target library fragment is 500 bp long and 150 bp PE sequencing is performed, then a 150 bp read will be made from one end of a target library molecule while the other 150 bp read will be taken from the opposite end of the same molecule (but the latter sequence will reflect the complementary strand). In this example, a ~200 bp stretch of bases in the middle of the target would not be sequenced. However, given the massive number of reads obtained from a single Illumina run, a sufficient number of overlapping reads will usually be obtained to allow for the accurate reconstruction of each entire allele sequence regardless of their lengths (Lemmon and Lemmon 2013). Although this is similar to using forward and reverse primers to sequence a single PCR product, it is not the same thing. This is because PE reads are effectively obtained from single molecules reflecting actual chromosomal sequences, whereas forward and reverse Sanger reads based on a single PCR product are obtained from a population of molecules that often consist of multiple haplotypes (Lemmon and Lemmon 2013). This general property of NGS offers another major advantage over Sanger sequencing: nuclear haplotypes can be easily and unambiguously determined for typical phylogenomic loci (Brito and Edwards 2009; Edwards and Bensch 2009; Lemmon and Lemmon 2013; O'Neill et al. 2013). Both single-end and PE Illumina sequencing have been successfully used to generate large multilocus datasets. However, because most studies are now using PE sequencing we will only examine this type of sequencing here. The following sections in this chapter are meant to provide a simplistic overview of Illumina sequencing chemistry. Moreover, as Illumina owns the rights to this technology, they may alter any of the procedures and reagents at any time. Given the fast-paced nature of genomics, it is essential that you use procedures and reagents that are compatible with the Illumina sequencing facility you plan to use. If in doubt, contact the intended sequencing provider before a project is begun in order to minimize the risk of wasting thousands of dollars and many hours of lab time.

7.1.1 Construction of Indexed Sequencing Libraries

Many similarities exist between genomic library-making methods used in the pre-NGS era and the Illumina methodology. The key difference, however, is that the latter type of library requires special adapter sequences, which must be attached to the ends of each library DNA fragment (Figure 7.1a). The adapters contain a variety of oligo sequences required for subsequent PCR-amplification and sequencing steps (Figure 7.1b). Because of the extensive space needed to fully explain how these constructs are acquired in phylogenomic studies, we will defer this discussion until later in this chapter. However, at this time we are able to examine how these adapters function in the cluster generation and sequencing processes.

7.1.2 Generation of Clusters on a Flow Cell

The process of creating clusters is quite different from the standard PCR methods used to amplify templates for Sanger sequencing. Instead of using a thermocycler to perform the amplification step, the Illumina workflow uses a fluidics device called a *cluster station*. Furthermore, rather than conducting the amplification reaction inside plastic microtubes or PCR microplates, clusters are generated on the inside surfaces of a *flow cell* (Figure 7.2a). The flow cell, which is small enough to fit in your hand, is a glass slide that is subdivided into eight enclosed glass channels referred to "lanes" (Fedurco et al. 2006). Flow cells with fewer lanes have also been developed for some sequencing platforms but the eight-lane flow cell has been the workhorse in Illumina sequencing. Multiplexed sequencing libraries for a phylogenomics project are usually loaded into a single lane, which is not uncommonly shared with the libraries from other projects.

The top and bottom surfaces of the flow cell's lanes are carpeted by two types of oligos for PE sequencing, all of which had their 5′ ends covalently anchored or "grafted" to the flow cell while their 3′ ends remain free (Adessi et al. 2000; Fedurco et al. 2006; Bentley et al. 2008; Figure 7.2b). These oligos, which are randomly distributed across the flow cell's surfaces in a 1:1 ratio, are commonly referred to as the "P5" and "P7" flow cell oligos though they were originally named "oligo C" and "oligo D," respectively (Bentley et al. 2008). The sequence of the P5 flow cell oligo is 5′-TTTTTTTTTTAATGATACGGCGACCACCGAGAUCTACAC-3′ where U = 2-deoxyuridine, while the sequence of the P7 flow cell oligo is 5′-TTTTTTTTTTCAAGCAGAAGACGGCATACGAGoxoAT-3′ where Goxo = 8-oxoguanine (Bentley et al. 2008). The uracil and Goxo residues in the P5 and P7 flow cell oligos, respectively, are cleavage sites used during the sequencing process. An enzyme mix called uracil-specific excision reagent ("USER") is used to cut each P5 flow cell oligo at the appropriate time (Bentley et al. 2008; see below). The enzymes used in this mix are uracil DNA glycosylase (UDG) and DNA glycosylase–lyase endonuclease VIII (New England Biolabs). The former enzyme excises the uracil base while the latter breaks the phosphodiester bond thereby cleaving the oligo into two fragments minus the uracil. The enzyme used to cleave the Goxo cleavage site in the P7 flow cell oligo is called "Fpg" (Bentley et al. 2008), which stands for formamidopyrimidine DNA glycosylase (New England Biolabs). For further description about the development of the Illumina flow cell technology see Adessi et al. (2000) and Fedurco et al. (2006).

Let's now examine the process of cluster formation. First, notice that two of the adapter sequences embedded in the library construct shown in Figure 7.1b are complementary to the P5 and P7 flow cell oligos. Thus, the adapter sequences at the ends of the library strands are able to hybridize to both flow cell oligos where they are subsequently amplified into individual clusters that are affixed to the flow cell surface. Indeed, one major function of the flow cell oligos is to serve as PCR primers in a special type of PCR known as *isothermal bridge amplification* or "solid-phase DNA amplification" (Adessi et al. 2000).

At the start of the cluster generation process the double-stranded library constructs must be chemically denatured into single-stranded molecules using 0.1 M NaOH. The single-stranded molecules are then added to a flow cell lane where they are allowed to hybridize to their complementary flow cell oligos under stringent thermal conditions controlled by the cluster station (Figure 7.3a). For simplicity, Figure 7.3a shows both strands of a library construct hybridizing to their respective flow cell oligos in close proximity to each other. Note, however, it is extremely unlikely that both strands from the same construct will hybridize

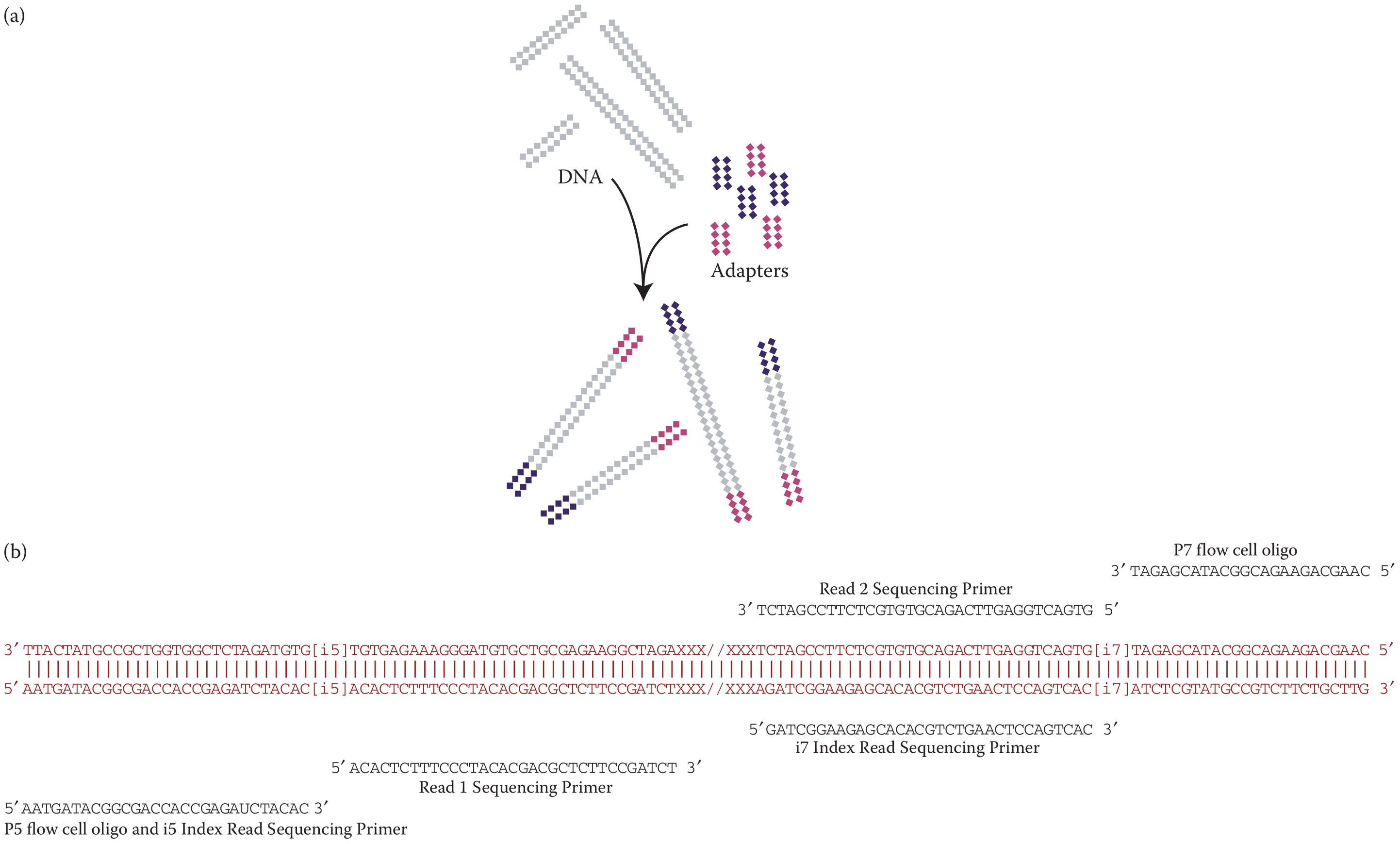

(a)
DNA
Adapters
(b)
P7 flow cell oligo
3′ TAGAGCATACGGCAGAAGACGAAC 5′
Read 2 Sequencing Primer
3′ TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG 5′
3′ TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
5′ AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
i7 Index Read Sequencing Primer
5′ ACACTCTTTCCCTACACGACGCTCTTCCGATCT 3′
Read 1 Sequencing Primer
5′ AATGATACGGCGACCACCGAGAUCTACAC 3′
P5 flow cell oligo and i5 Index Read Sequencing Primer

Figure 7.1. Construction of Illumina adapter–library fragment constructs. (a) Before target (library) DNA fragments can be sequenced, Illumina adapters must be attached to the ends of the fragments. Each library fragment must include both the P5 and P7 adapters (shown in different colors) in order to be sequenced. (Image courtesy of Illumina, Inc.) (b) Detailed look at a finished Illumina adapter–library fragment construct. The middle section of the construct contains the target DNA (a string of Xs). The // signifies that most of the target fragment length is not shown due to space limitations and the vertical dashes indicate hydrogen bonding between complementary bases. The [i5] and [i7] sequences represent two index sequences used to identify sequenced fragments after sequencing. Oligos shown above and below the construct are discussed in the main text and can be found in Table 7.1. (Oligonucleotide sequences © 2007–2010 Illumina, Inc. All rights reserved.)

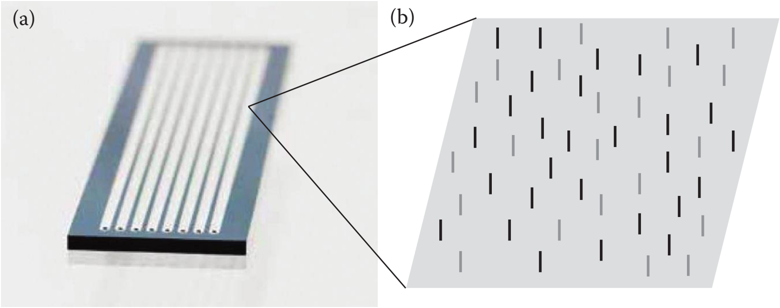


Figure 7.2. (a) An Illumina flow cell. This flow cell is a glass slide with eight lanes. (Image courtesy of Illumina, Inc.) (b) Magnified view of the flow cell surface in one of the lanes. Top and bottom (not shown) flow cell surfaces are populated with two different oligos, here shown as vertical black and gray lines, which have been covalently attached to the surface.

to adjacent flow cell oligos; in other words, each cluster generally originates from a single library strand. Once a library strand hybridizes to its complementary flow cell oligo, a primer template junction is created, which allows for synthesis of a new strand of DNA after unlabeled dNTPs and DNA polymerases are added to the flow cell (Figure 7.3b,c). Notice that the new strand, which is complementary to the original strand, is covalently attached to the flow cell surface (Figure 7.3c). Thus, when these double-stranded molecules are chemically denatured via treatment with formamide in the next step, the original library strands can be washed away leaving the new complementary strands firmly attached to the flow cell (Bentley et al. 2008; Figure 7.3d). Each flow cell-tethered strand will be the initial template molecule used to make each cluster. Next, grafted library strands fold over to hybridize or "bridge" nearby free flow cell oligos, which again creates primer–template junctions that are used in another round of DNA synthesis (Figure 7.3e). After synthesis, double-stranded bridges are formed in scattered locations across the flow cell surface (Figure 7.3f). The double-stranded bridges are then chemically denatured with formamide, which linearizes each library molecule (Figure 7.3g). After many additional synthesis cycles (35 cycles total; Bentley et al. 2008), the surface of the flow cell contains hundreds of millions of randomly distributed clusters, each of which will be sequenced (Figure 7.3g). As we will soon see, the fixed-location nature of these clusters represents one of the keys to the Illumina sequencing method. At this moment in the process, sequences matching both strands of the original library construct are now covalently attached to the flow cell; the strand anchored by the P7 flow cell oligo is hereafter referred to as the *forward strand*, whereas the stand connected to the P5 flow cell oligo is now the *reverse strand* (Figure 7.3g). In the last step before sequencing, the reverse strands are cut with USER enzymes (as explained before) and washed away leaving only the forward strands attached to the flow cell (Figure 7.3g). After removal of the reverse strands, notice that the P5 flow cell oligos remain anchored to the flow cell except that they are seven bp shorter at their 3′ ends. These P5 flow cell oligos will still be used during the PE sequencing process.

7.1.3 Sequencing of Clusters

The process of sequencing clusters is more complex than sequencing PCR products via Sanger sequencing. The Illumina PE adapters shown in Figure 7.1b contain sequences that function as annealing sites for three different sequencing primers: Read 1 and Read 2 Sequencing Primers generate sequences at each end of the library fragment (i.e., PE reads) while the i7 Index Read Sequencing Primer obtains the i7 index. A fourth sequencing primer is not needed for acquiring the i5 index sequence because the P5 flow cell oligo is used for this purpose (Figure 7.1b). Thus, there are four different sequencing passes made on the library construct in order to record the PE sequences and dual indices.

As we saw in Chapter 6 fluorescent-labeled chain terminator nucleotides are key components of the Sanger sequencing methodology. Similarly, dye-terminator nucleotides are also essential

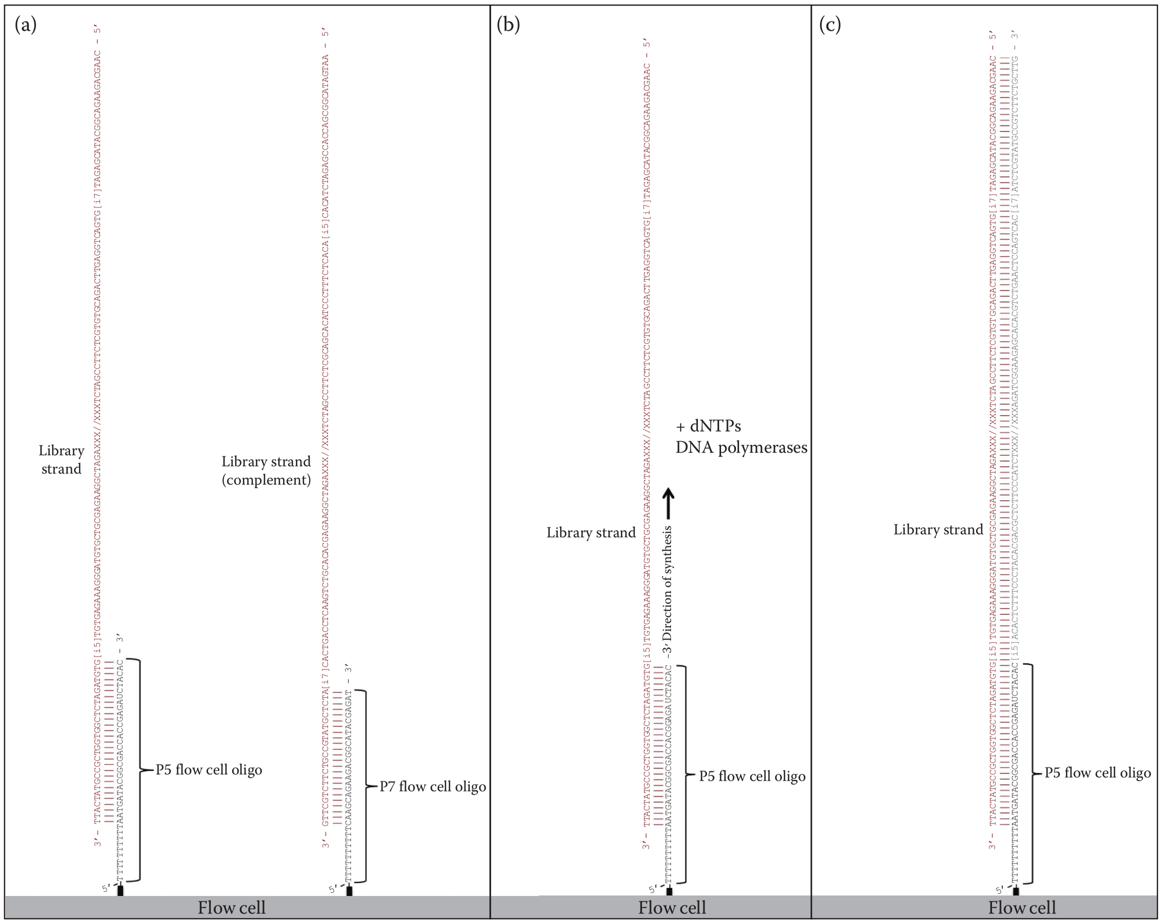


Figure 7.3. Isothermal bridge amplification. (a) Adapters at the ends of single-stranded library DNA hybridize to P5 and P7 flow cell oligos. Sequence of Xs in each strand represents target library fragments, // indicates that much of the library sequence is not shown due to space limitations, and vertical dashes signify hydrogen bonding between complementary bases. (b) Hybridization of a library strand to a flow cell oligo (only P5 flow cell oligo is shown) creates a primer–template junction, which is used for synthesizing a complementary strand. (c) A new strand is synthesized (gray bases).

(Continued)

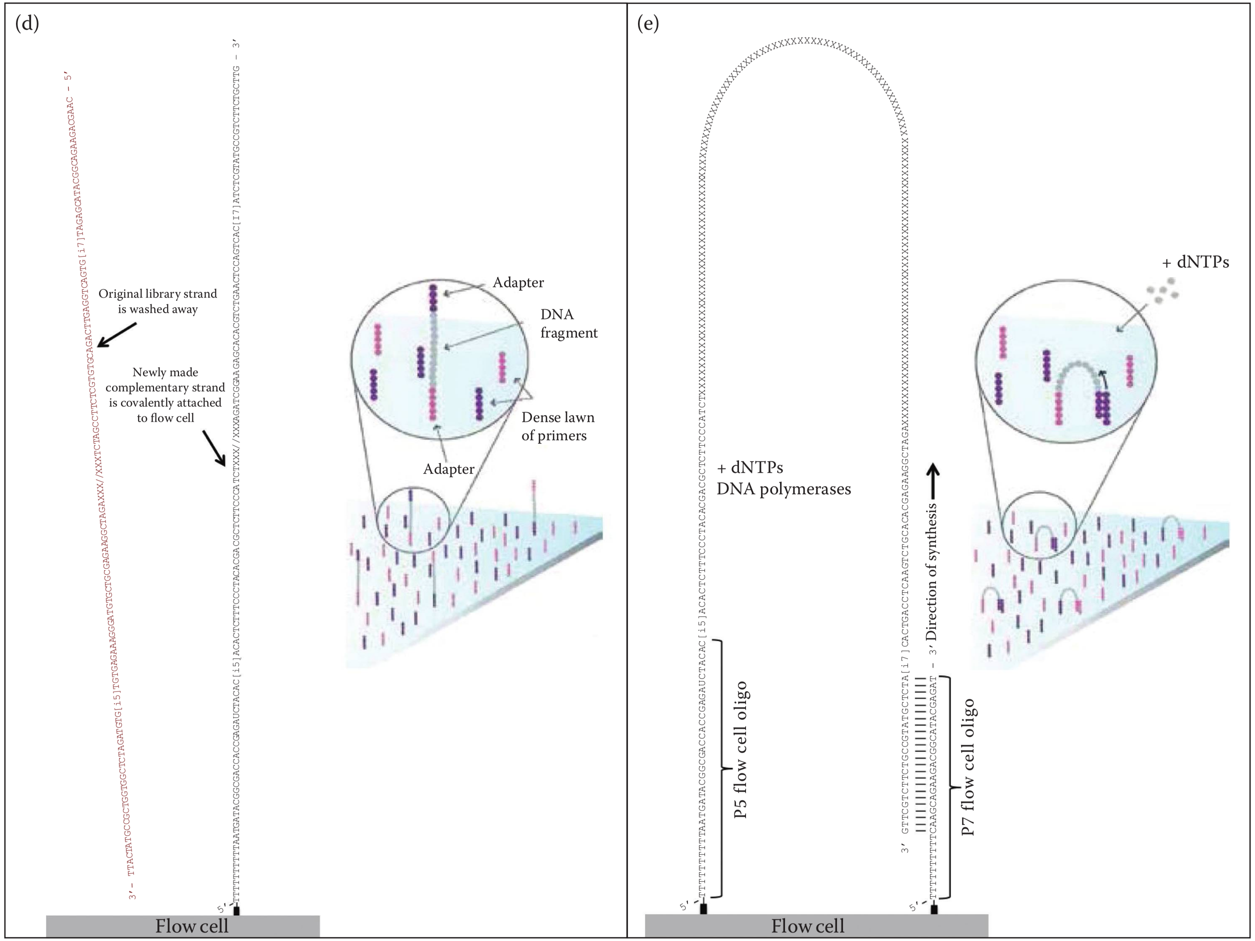

(d)
3'- TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC - 5'
5'-TTTTTTTTTTAATGATACGGCGACCACCGAGAUCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG - 3'
Original library strand is washed away
Newly made complementary strand is covalently attached to flow cell
Adapter
DNA fragment
Dense lawn of primers
Adapter
Flow cell
(e)
5'-TTTTTTTTTTAATGATACGGCGACCACCGAGAUCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCATCTXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
3' GTTCGTCTTCTGCCGTATGCTCTA[i7]CACTGACCTCAAGTCTGCACACGAGAAGGCTAGAXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
5'-TTTTTTTTTTCAAGCAGAAGACGGCATACGAGAT - 3'
P5 flow cell oligo
P7 flow cell oligo
+ dNTPs
DNA polymerases
Direction of synthesis
+ dNTPs
Flow cell

Figure 7.3. (Continued) (d) DNA duplexes are denatured using formamide and original library strands are washed away leaving complementary strands covalently attached to the flow cell. (e) Anchored library strands fold over so their adapters can hybridize or "bridge" the other flow cell oligos, PCR reagents are added, and DNA synthesis starts again. (Colored images courtesy of Illumina, Inc.)

(Continued)

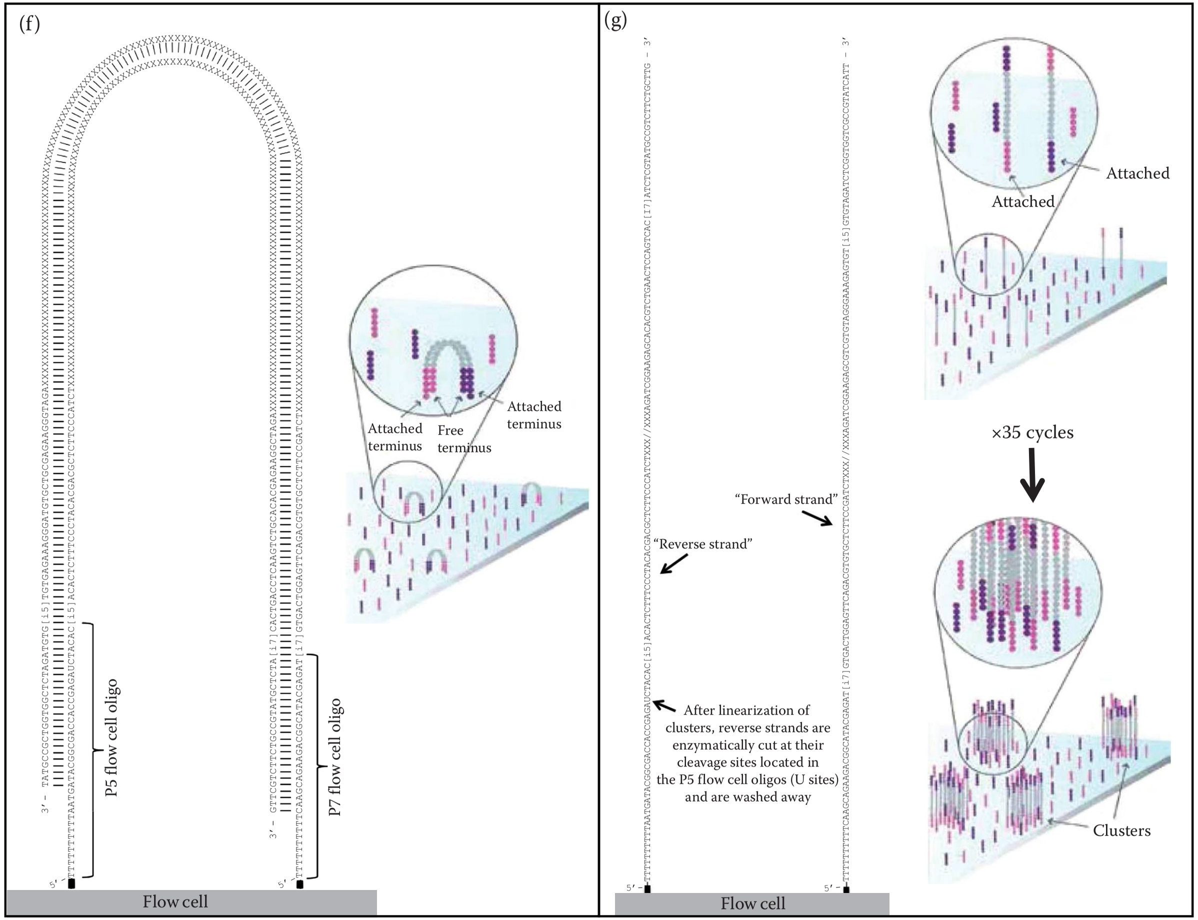


Figure 7.3. (Continued) (f) After synthesis of a new strand (gray bases), a double-stranded bridge is formed. (g) DNA duplexes are chemically denatured using 0.1 M NaOH to linearize flow-cell anchored library strands. After additional PCR cycles (35 total), clusters are generated across the flow cell surface. Following linearization of strands, the reverse strands are cleaved and washed away leaving only the forward strands attached to the flow cell. (Colored images courtesy of Illumina, Inc.)

players in the Illumina methodology. However, the design of the Illumina sequencing nucleotides differs from Sanger dye-ddNTPs in two important ways. First, the Illumina terminators can be easily "reversed" back to a normal-functioning nucleotide state after they have been incorporated into a DNA strand (i.e., 3′ hydroxyl groups can be regenerated). Secondly, their fluorophore moieties can be chemically cleaved from incorporated nucleotides after their fluorescent signals have been imaged. This modified nucleotide, which is known as a *reversible terminator*, is 3′-O-azidomethyl 2′-deoxynucleoside triphosphate (Bentley et al. 2008). The chemical structure of the reversible terminator is identical for all four bases (A, C, G, and T) except for base-specific fluors (Bentley et al. 2008). Figure 7.4a shows the chemical structure of one of these modified nucleotides, a 3′-O-azidomethyl 2′-deoxythymine triphosphate with a cleavable fluorophore. The chemical Tris(2-carboxyethyl)phosphine (TCEP) is used to cleave the fluorophore and regenerate the 3′ hydroxyl group (Figure 7.4b; Bentley et al. 2008). Thus, if the labeled reversible "T" terminator shown in Figure 7.4a becomes incorporated into a growing DNA strand, DNA synthesis will be terminated. However, following subsequent laser-detection and imaging steps, the fluorophore and 3′ blocking group can be chemically removed with regeneration of the 3′ OH group (Figure 7.4b). This now unlabeled reversible terminator can prime the addition of the next fluorescent-labeled reversible terminator.

Before the first sequencing read can be performed, all exposed 3′ ends, which include 3′ termini of the forward strands and unextended flow cell oligos, must first be blocked in order to preclude the possibility of having reversible terminators incorporated in those unwanted locations (Bentley et al. 2008). This blocking step is accomplished by using terminal transferases, which are template-independent DNA polymerases (New England Biolabs), to incorporate a single ddNTP at each available 3′ end (Figure 7.5a; Bentley et al. 2008). Next, Read 1 Sequencing Primers are annealed to the forward strand templates followed by the addition of reversible terminators and DNA polymerases (Figure 7.5a).

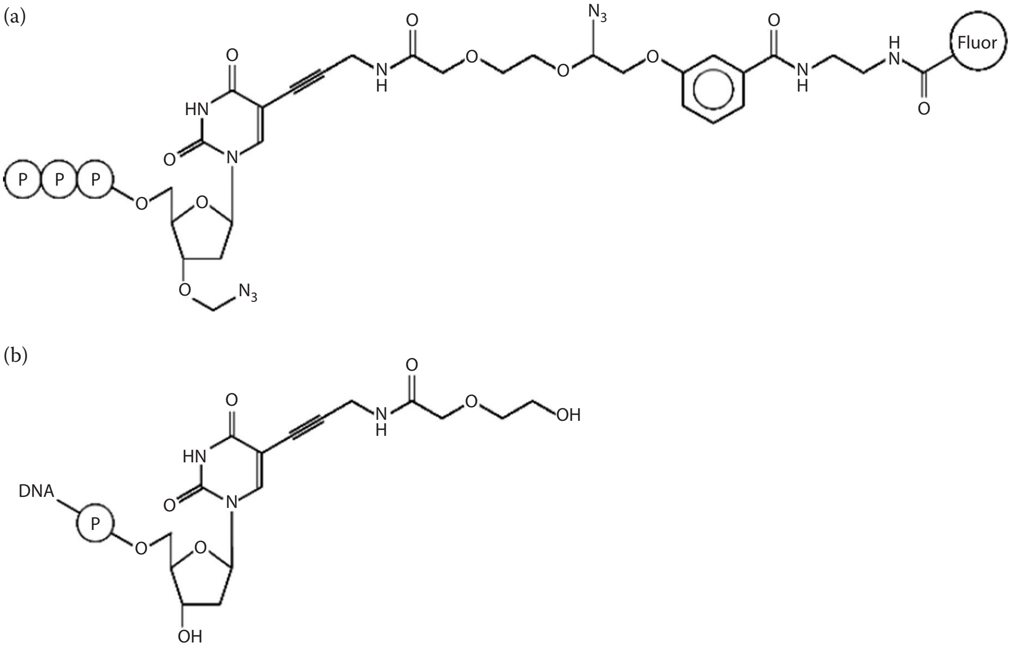


Figure 7.4. The chemical structure of a reversible terminator nucleotide. (a) Example shown is 3′-O-azidomethyl 2′-deoxythymine triphosphate with a cleavable fluorophore. The 3′ location contains a blocking group, which terminates DNA synthesis after this nucleotide has been incorporated. (b) Following cleavage of the fluorophore and 3′ blocker and regeneration of the 3′ OH group, an incorporated reversible terminator can prime the addition of a new base in the next sequencing cycle. (Reprinted from Bentley, D. R. et al. 2008. *Nature* 456:53–59. With permission.)

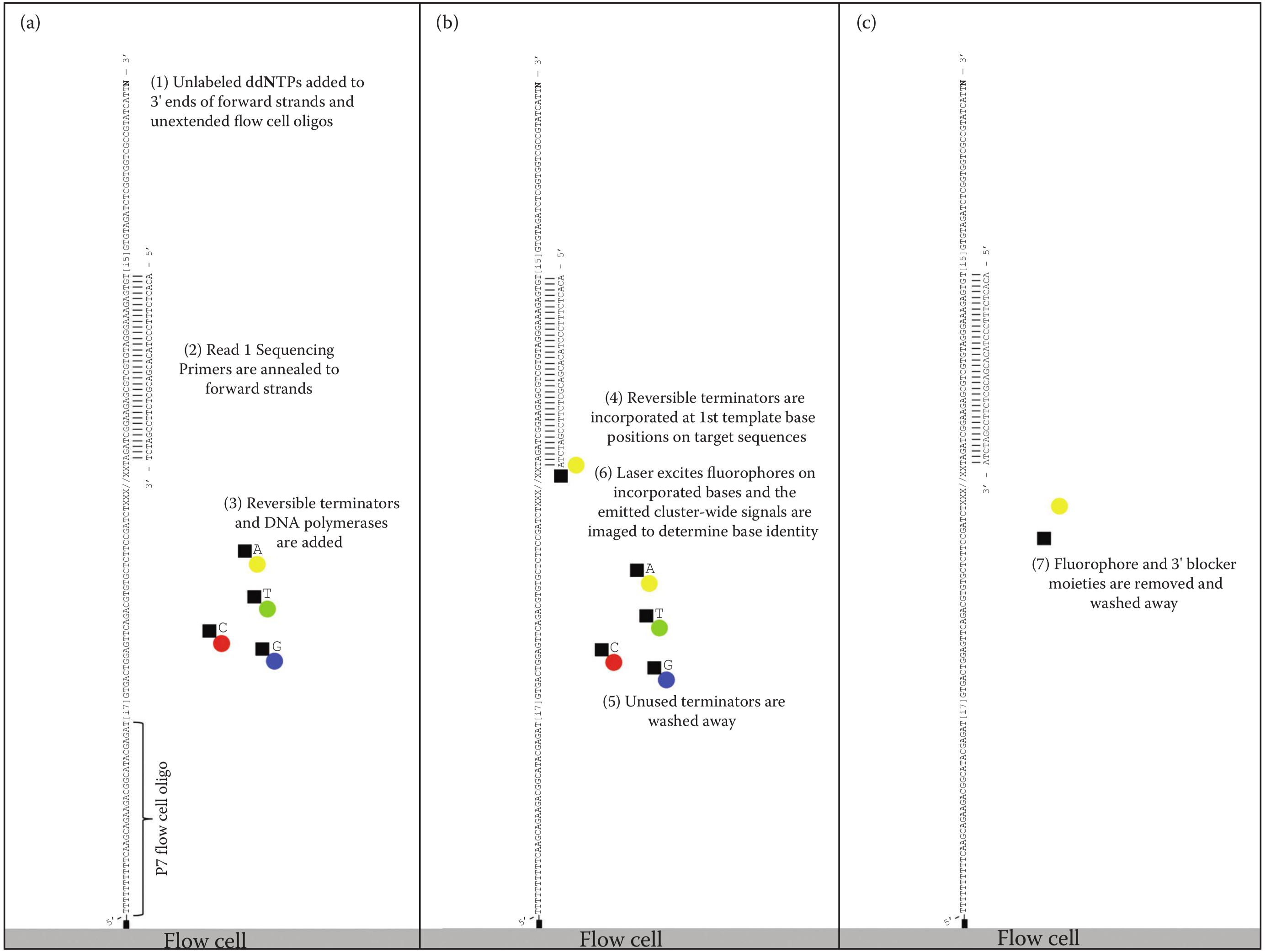
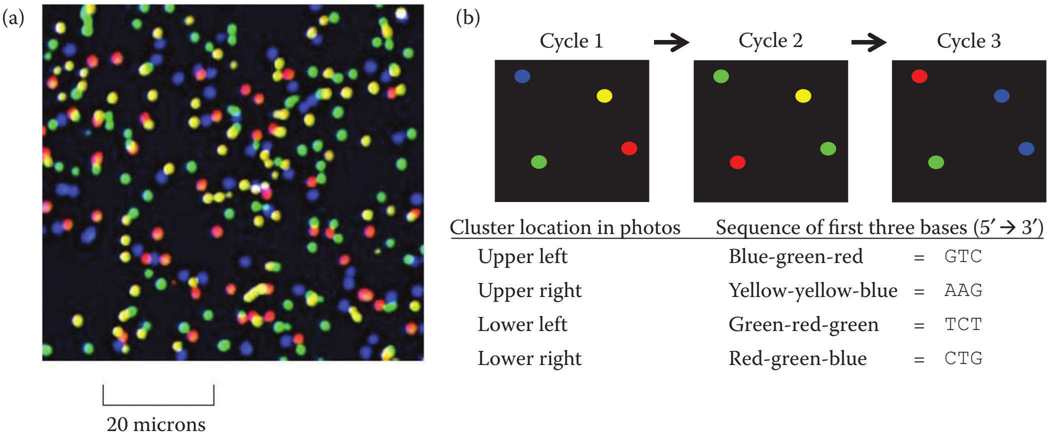

(a)
5' - TTTTTTTTTTCAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTXXX//XXTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT[i5]GTGTAGATCTCGGTGGTCGCCGTATCATTN – 3'
3' - TCTAGCCTTCTCGCAGCACATCCCTTTCTCACA - 5'
P7 flow cell oligo
(1) Unlabeled ddNTPs added to 3' ends of forward strands and unextended flow cell oligos
(2) Read 1 Sequencing Primers are annealed to forward strands
(3) Reversible terminators and DNA polymerases are added
A
T
C
G
Flow cell
(b)
5' - TTTTTTTTTTCAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTXXX//XXTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT[i5]GTGTAGATCTCGGTGGTCGCCGTATCATTN – 3'
ATCTAGCCTTCTCGCAGCACATCCCTTTCTCACA - 5'
(4) Reversible terminators are incorporated at 1st template base positions on target sequences
(6) Laser excites fluorophores on incorporated bases and the emitted cluster-wide signals are imaged to determine base identity
A
T
C
G
(5) Unused terminators are washed away
Flow cell
(c)
5' - TTTTTTTTTTCAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTXXX//XXTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT[i5]GTGTAGATCTCGGTGGTCGCCGTATCATTN – 3'
3' - ATCTAGCCTTCTCGCAGCACATCCCTTTCTCACA - 5'
(7) Fluorophore and 3' blocker moieties are removed and washed away
Flow cell

Figure 7.5. First paired-end read of the target DNA on the forward strands. (a) First, all exposed 3′ ends must be blocked with ddNTPs. Next, Read 1 Sequencing Primers are annealed to the forward strands followed by the addition of reversible terminators and DNA polymerases. The middle section of the shown forward strand contains the target DNA (a string of Xs). The // signifies that most of the target fragment is omitted due to space limitations and the vertical dashes indicate hydrogen bonding between complementary bases. (b) A reversible terminator that is complementary to the first template base in the target sequence becomes incorporated. Unused terminators are then washed away before the fluorophore-detection and imaging steps. (c) The fluorophore and 3′ blocking group on each incorporated reversible terminator are removed and washed away. This process also regenerates 3′ OH groups on the incorporated terminators enabling them to prime the addition of a new reversible terminator in the next sequencing cycle.

After a single reverse terminator has been incorporated into the first template base position of the target library fragment, the unused terminators must be washed away. In the next step, the sequencer's laser is used to excite the fluorophores attached to the incorporated terminators (Figure 7.5b). Because all of the ~1,000 strands comprising each cluster are expected to incorporate the same base at the same position, each cluster should emit a single and strong base-specific signal, which the sequencer's imaging apparatus can detect and record. After the imaging step is done, the fluorophores and 3′ blockers on the incorporated reversible terminators are removed (Figure 7.5c). This completes one cycle of sequencing. The second cycle starts with the addition of reversible terminators and DNA polymerases and finishes with the removal of the fluorophores and 3′ blockers. The total number of cycles will equal the length of each cluster read (e.g., 100–300 bp). This step-by-step process has been termed *sequencing by synthesis* (Bentley et al. 2008; Mardis 2008). When all cycles are completed to generate the complete Read 1 sequences, the sequencing products are denatured from the template strands via treatment with 0.1 M NaOH and then washed away (Bentley et al. 2008).

As alluded to earlier, the fixed locations of the clusters plays a critical role in the Illumina sequencing process. Figure 7.6a shows an image taken of a small portion of a flow cell during the sequencing process. On this image we can see a scattering of colored dots each of which corresponds to an individual cluster that was illuminated by the sequencer's laser. Using the series of images taken—one image taken per sector on the flow cell surface per cycle, the sequencer can generate a read for each cluster. For example, in Figure 7.6b we see images for the first three cycles. Owing to the fixed location of each cluster, the sequencer can record the color of each individual cluster in a sequential manner across cycles thereby producing a read for each cluster. In our example, the cluster located at the upper left corner of the images is blue in the first cycle, green in the second, and red in the third, which gives the



Cluster location in photos	Sequence of first three bases (5′ → 3′)		
Upper left	Blue-green-red	=	GTC
Upper right	Yellow-yellow-blue	=	AAG
Lower left	Green-red-green	=	TCT
Lower right	Red-green-blue	=	CTG

Figure 7.6. Imaging of illuminated clusters and the determination of their sequences. (a) Image showing clusters occupying a small area of a flow cell. Each color dot represents a cluster, which has a fixed location on the flow cell. (Reprinted from Bentley, D. R. et al. 2008. *Nature* 456:53–59. With permission.) (b) A simple example showing how sequences are elucidated for each cluster.

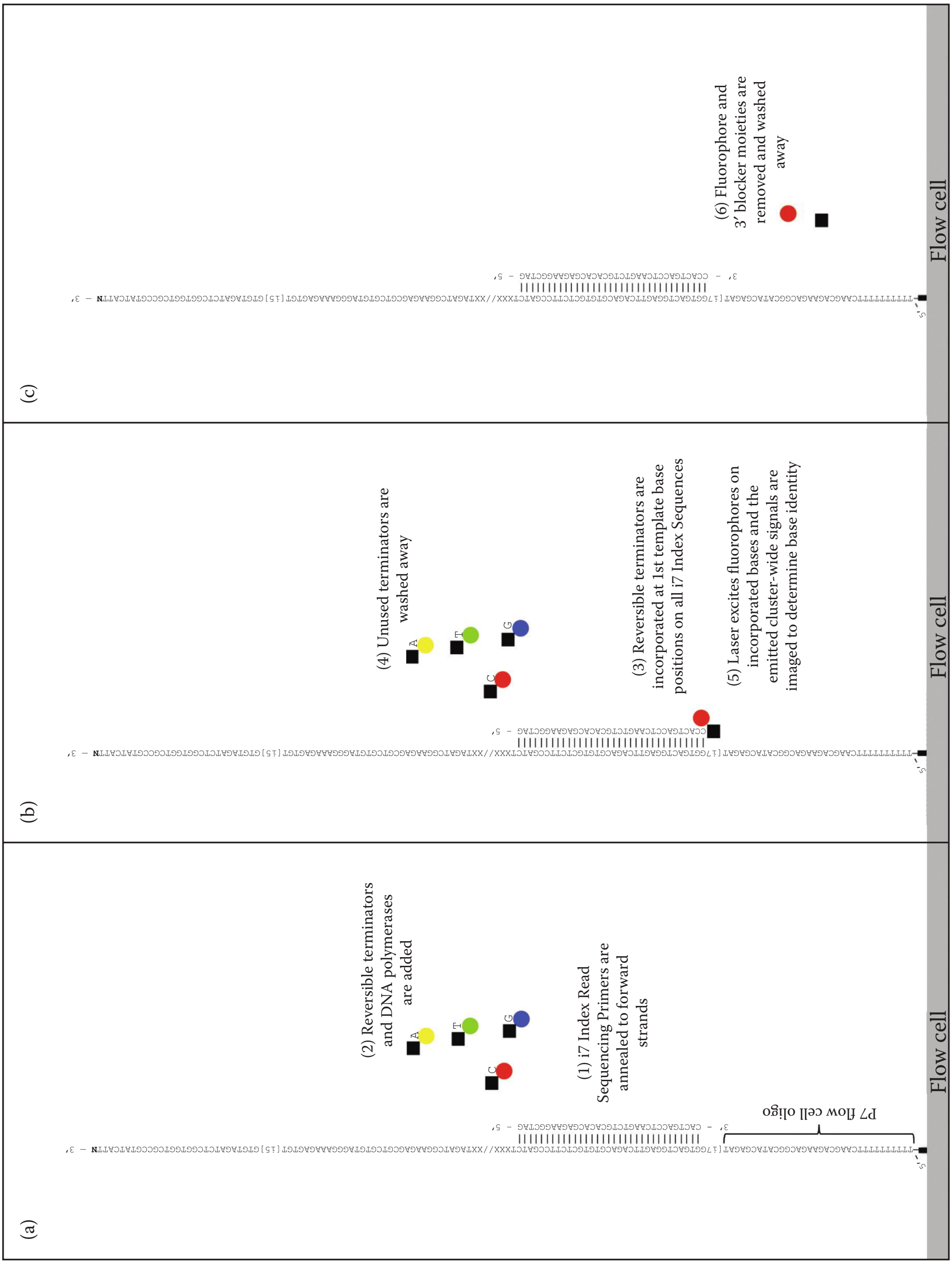

(a)
5'-TTTTTTTTTCAAGCAGAAGACGGCATACGAGAT[i7GGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTXXX//XXTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT[i5]GTGTAGATCTCGGTGGTCGCCGTATCATTN – 3'
3' - CACTGACCTCAAGTCTGCACACGAGAAGGCTAG - 5'
P7 flow cell oligo
(1) i7 Index Read Sequencing Primers are annealed to forward strands
(2) Reversible terminators and DNA polymerases are added
A
T
G
C
Flow cell
(b)
5'-TTTTTTTTTCAAGCAGAAGACGGCATACGAGAT[i7GGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTXXX//XXTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT[i5]GTGTAGATCTCGGTGGTCGCCGTATCATTN – 3'
CCACTGACCTCAAGTCTGCACACGAGAAGGCTAG - 5'
(3) Reversible terminators are incorporated at 1st template base positions on all i7 Index Sequences
(4) Unused terminators are washed away
A
T
G
C
(5) Laser excites fluorophores on incorporated bases and the emitted cluster-wide signals are imaged to determine base identity
Flow cell
(c)
5'-TTTTTTTTTCAAGCAGAAGACGGCATACGAGAT[i7GGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTXXX//XXTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT[i5]GTGTAGATCTCGGTGGTCGCCGTATCATTN – 3'
3' - CCACTGACCTCAAGTCTGCACACGAGAAGGCTAG - 5'
(6) Fluorophore and 3' blocker moieties are removed and washed away
Flow cell

Figure 7.7. First index read (i7 index) on the forward strands. (a) i7 Index Read Sequencing Primers are annealed to the forward strands followed by the addition of reversible terminators and DNA polymerases. The middle section of the shown forward strand contains the target DNA (a string of Xs). The // signifies that most of the target fragment is omitted due to space limitations and the vertical dashes indicate hydrogen bonding between complementary bases. (b) A reversible terminator that is complementary to the first template base on the i7 index sequence becomes incorporated. Unused terminators are then washed away before the fluorophore-detection and imaging steps. (c) The fluorophore and 3′ blocking group on each incorporated reversible terminator are removed and washed away. This process also regenerates 3′ OH groups on the incorporated terminators enabling them to prime the addition of a new reversible terminator in the next sequencing cycle.

sequence 5′-GTC-3′ (Figure 7.6b). In reality, the number of images analyzed for this same flow cell region would number somewhere between 100 and 300 depending on the sequencer model and thus the overall read lengths would range from 100 to 300 bp for each cluster.

The second sequence to obtain from each of the clusters will be the first index sequence (i.e., i7 index). This process begins with binding of i7 Index Read Sequencing Primers to the forward strands (Figure 7.7a). The same sequencing steps used earlier to obtain all Read 1 sequences are used and thus will not be repeated (Figure 7.7a–c). Thus, eight cycles must be performed in order to obtain the complete sequence for an 8-bp index. When sequencing of the first index is completed, the sequencing products are denatured with 0.1 M NaOH and washed away.

The second index sequence represents the third sequence to be obtained from the library molecules. Unlike the Read 1 and i7 Index Read Sequencing steps, a dedicated sequencing primer is not used to obtain the i5 index sequence. Instead, the 3′ ends of the P5 flow cell oligos are used as i5 index sequencing primers (see Figure 7.1b). Thus, the process begins with restoration of 3′ OH groups on the forward strands and P5 flow cell oligos—though only the P5 flow cell oligos will participate in the i5 index sequencing process—and folding over of the forward strands so they can hybridize with the P5 flow cell oligos (Figure 7.8a). While in this single-stranded bridge configuration, the P5 flow cell oligo is used to prime the sequencing reaction; that is, instead of adding unlabeled dNTPs and DNA polymerases to generate double-stranded bridges as was done in cluster formation, reversible terminators and DNA polymerases are added to begin the sequencing of the i5 index (Figure 7.8a). The same sequencing steps are used as before (Figure 7.8a–c). However, recall that following the step in which the reverse strands were cleaved, the P5 flow cell oligos lost the last seven bases at their 3′ ends (Figure 7.3g). Thus, the first seven cycles of the i5 Index Read will be used to sequence the bases immediately upstream of the i5 index meaning a total of 15 cycles are needed to obtain a 8-bp i5 index sequence (Figure 7.8b). After all i5 Index Read cycles have been completed, the i5 index sequencing products are denatured and washed away.

To obtain the fourth and final sequence from the library molecules—the other PE read of the target fragment, the reverse strands must first be regenerated via bridge amplification (Bentley et al. 2008). This is because the binding sites for the Read 2 Sequencing Primer are located on the reverse strand (see Figure 1b). After the clusters are regenerated and then denatured to linearize the forward and reverse strands, the Goxo sites in the P7 flow cell oligos are cleaved using Fpg enzymes, which releases the forward strands allowing them to be washed away (Bentley et al. 2008). As these bridge amplification and strand cleavage steps are essentially the same as in Figure 7.3, they will not be shown again here.

With only the reverse strands now anchored to the flow cell surface, all exposed 3′ ends must again be blocked with unlabeled ddNTPs before the Read 2 Sequencing Primers are annealed (Figure 7.9a). The remaining sequencing steps are the same as for the previous reads (Figure 7.9a–c).

7.2 METHODS FOR OBTAINING MULTIPLEXED HYBRID SELECTION LIBRARIES

The goal of this section is to introduce several methods for making multiplexed Illumina libraries, which will be followed by a discussion of the procedures for performing hybrid selection with these libraries. The set of target-loci enriched libraries obtained via these procedures will then be ready for the cluster generation and sequencing-by-synthesis steps just described.

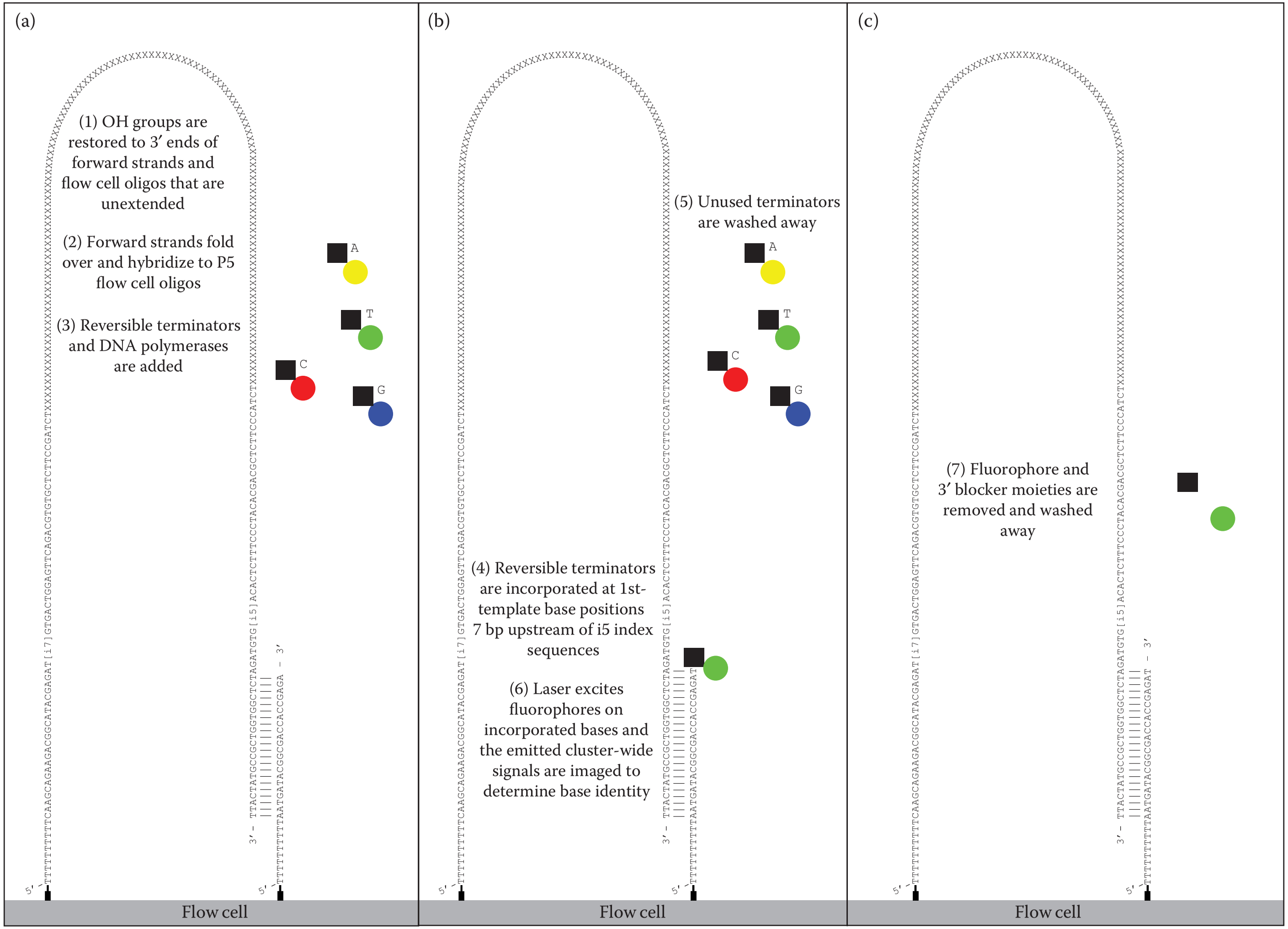

(a)
(1) OH groups are restored to 3′ ends of forward strands and flow cell oligos that are unextended
(2) Forward strands fold over and hybridize to P5 flow cell oligos
(3) Reversible terminators and DNA polymerases are added
A
T
C
G
5′ - TTTTTTTTTTCAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
3′ - TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT
5′ - TTTTTTTTTTAATGATACGGCGACCACCGAGA - 3′
Flow cell
(b)
(4) Reversible terminators are incorporated at 1st-template base positions 7 bp upstream of i5 index sequences
(5) Unused terminators are washed away
(6) Laser excites fluorophores on incorporated bases and the emitted cluster-wide signals are imaged to determine base identity
A
T
C
G
5′ - TTTTTTTTTTCAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
3′ - TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT
5′ - TTTTTTTTTTAATGATACGGCGACCACCGAGAT
Flow cell
(c)
(7) Fluorophore and 3′ blocker moieties are removed and washed away
5′ - TTTTTTTTTTCAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
3′ - TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT
5′ - TTTTTTTTTTAATGATACGGCGACCACCGAGAT - 3′
Flow cell

Figure 7.8. Second index read (i5 index) on the forward strands. (a) First, hydroxyl groups must be restored to the ends of exposed 3′ termini. Next, forward strands fold over to hybridize with truncated P5 flow cell oligos followed by the addition of reversible terminators and DNA polymerases. The middle section of the shown forward strand contains the target DNA (a string of Xs). The // signifies that most of the target fragment is omitted due to space limitations and the vertical dashes indicate hydrogen bonding between complementary bases. (b) A reversible terminator becomes incorporated adjacent to a template base 7-bp upstream of the first i5 index template base. Unused terminators are then washed away before the fluorophore-detection and imaging steps. (c) The fluorophore and 3′ blocking group on each incorporated reversible terminator are removed and washed away. This process also regenerates 3′ OH groups on the incorporated terminators enabling them to prime the addition of a new reversible terminator in the next sequencing cycle.

7.2.1 Library Preparation Approaches

Various methods for constructing Illumina libraries have been developed over the years, four of which we will examined here. The "traditional approach" reflects the methodology used in Illumina TruSeq Kits (e.g., http://www.illumina.com/products/truseq-nano-dna-library-prep-kit.html) as well as in some published non-kit protocols (e.g., Bentley et al. 2008; Bronner et al. 2014). A second method, developed by Meyer and Kircher (2010), co-opted the library-making method used for 454 sequencing (Margulies et al. 2005) for generating multiplexed sequencing libraries in a cost-effective manner. Hereafter, this will be referred to as the "Meyer and Kircher approach." A third approach, which is also based on the 454 library methods but has several important innovations that distinguish it from the Meyer and Kircher approach, was developed by Rohland and Reich (2012). Hereafter, this approach will be referred to as the "Rohland and Reich approach." A fourth approach, which uses *in vitro* transposition to construct Illumina libraries, has since become another popular method for making Illumina libraries. This method, which will be referred to as the "Nextera approach," was originally developed and sold by Epicentre Technologies (Syed et al. 2009a,b), but is now sold by Illumina (http://www.illumina.com/products/nextera_dna_library_prep_kit.html). Each of these library-making approaches can be implemented in low-throughput or high-throughput (96 sample plates) workflows.

Although these approaches to making sequencing libraries differ from each other in many ways, all are designed to convert input DNA samples into indexed libraries suitable for Illumina sequencing. In addition to commercially available library kits, a number of protocols have been published, which show, step-by-step, how these libraries can be generated. Use of these "homemade" protocols can substantially reduce the costs of making libraries—particularly if many libraries are needed. However, it must not be forgotten that the process of generating these libraries still involves an intricate series of molecular biology steps each of which must be successfully performed otherwise expensive failures can easily result if something goes wrong. Indeed, the cost of making libraries and then sequencing them on an Illumina platform is on the order of thousands of dollars. It is therefore highly recommended that researchers who do not have experience in making these libraries or do not have well-honed molecular biology skills—especially relating to genomic cloning work—should work closely with someone who is an expert and/or use commercial library kits. One advantage of kits is that they contain control reactions at different steps, which are helpful for monitoring progress and troubleshooting problems. Another advantage of these kits is that they contain ready-to-use "master mixes," which reduces the amount of pipetting and consequently speeds up the preparation process while reducing the amount of plastics consumed as well as errors (e.g., cross-contamination). Illumina provides much information on their website about making high-quality sequencing libraries (http://www.illumina.com/techniques/sequencing/ngs-library-prep.html). We will now review these four approaches to making indexed Illumina sequencing libraries.

7.2.1.1 Traditional Illumina Library Approach

In one of the earliest demonstrations of the Illumina sequencing platform, Bentley et al. (2008) used this technology to sequence the human genome. This study also provided the earliest complete and detailed description of the Illumina sequencing workflow. Since then, Illumina has offered library preparation kits, which are user-friendly compared to the non-kit methodology outlined by Bentley et al. (2008). These kits also permit the researcher to make up

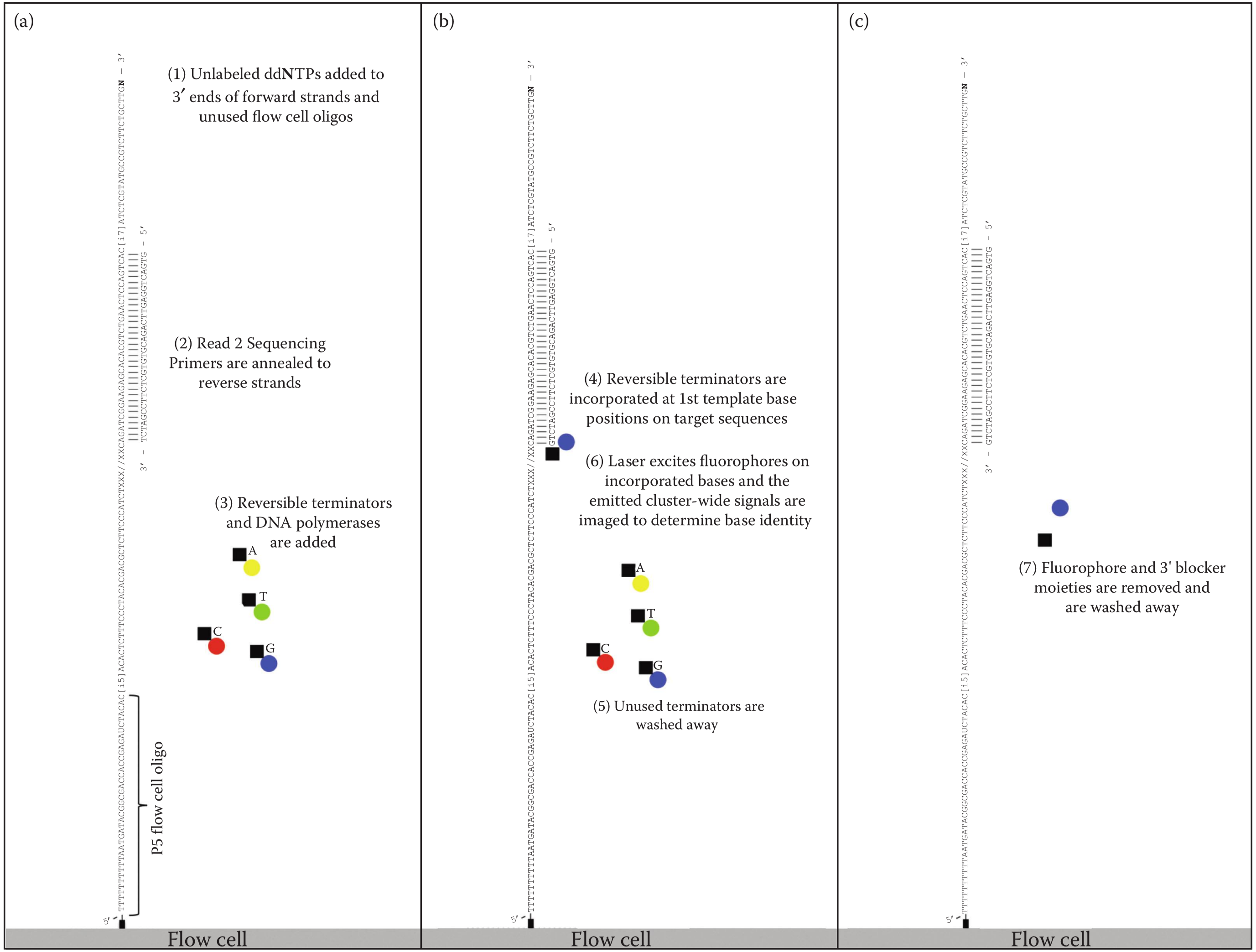

(a)
(1) Unlabeled ddNTPs added to 3′ ends of forward strands and unused flow cell oligos
(2) Read 2 Sequencing Primers are annealed to reverse strands
(3) Reversible terminators and DNA polymerases are added
A
T
C
G
P5 flow cell oligo
5′
TTTTTTTTTTAATGATACGGCGACCACCGAGAUCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXCAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTGN – 3′
3′ – TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG – 5′
Flow cell
(b)
(4) Reversible terminators are incorporated at 1st template base positions on target sequences
(6) Laser excites fluorophores on incorporated bases and the emitted cluster-wide signals are imaged to determine base identity
A
T
C
G
(5) Unused terminators are washed away
5′
TTTTTTTTTTAATGATACGGCGACCACCGAGAUCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXCAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTGN – 3′
GTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG – 5′
Flow cell
(c)
(7) Fluorophore and 3' blocker moieties are removed and are washed away
5′
TTTTTTTTTTAATGATACGGCGACCACCGAGAUCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXCAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTGN – 3′
3′ – GTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG – 5′
Flow cell

Figure 7.9. Second paired-end read of the target DNA on the reverse strands. (a) First, all exposed 3′ ends must be blocked with ddNTPs. Next, Read 2 Sequencing Primers are annealed to the reverse strands followed by the addition of reversible terminators and DNA polymerases. The middle section of the shown reverse strand contains the target DNA (a string of Xs). The // signifies that most of the target fragment is omitted due to space limitations and the vertical dashes indicate hydrogen bonding between complementary bases. (b) A reversible terminator that is complementary to the first template base in the target sequence becomes incorporated. Unused terminators are then washed away before the fluorophore-detection and imaging steps. (c) The fluorophore and 3′ blocking group on each incorporated reversible terminator are removed and washed away. This process also regenerates 3′ OH groups on the incorporated terminators enabling them to prime the addition of a new reversible terminator in the next sequencing cycle.

to 96 individually indexed libraries, which can be multiplexed before sequencing. For example, recent phylogenomic studies by Leaché et al. (2015) and McCormack et al. (2015) have used the Illumina TruSeq® Nano DNA HT Library Prep Kit and Kapa Biosystems KAPA Library Preparation Kit, respectively, in order to generate large multi-locus datasets. In addition to these kits, a number of non-kit protocols for constructing libraries have been published (e.g., Bronner et al. 2014). The traditional approach for making Illumina sequencing libraries consists of the basic steps outlined in Figure 7.10. Although both kit and non-kit approaches share most of the same steps, there are some differences between them. The main difference between the kit and non-kit based methods is that libraries are indexed during the ligation step in kits (step 4 in Figure 7.10), while the indexing step takes place during the first limited cycle PCR in non-kit protocols (step 5 in Figure 7.10). In the following discussion of the traditional approach we will focus on the Illumina kit methodology.

Step 1: Fragmentation of Input DNA Followed by SPRI For many phylogenomics applications, the starting or "input" DNA for an Illumina sequencing library usually consists of extracted genomic DNA but it can also be other forms of DNA such as short or long PCR products. Moreover, the amount of input DNA required for building an Illumina library usually varies between 100 ng and 1 μg depending on the kit or protocol being used. As the Illumina Truseq protocol points out it is extremely important to accurately quantify the input DNA prior to constructing libraries. Some researchers have reported using the UV spectrophotometers such as the Nanodrop to estimate genomic DNA concentrations for Illumina libraries when the DNA samples were free of RNA. However, other researchers have reported that UV spectrophotometric methods often yielded less reliable estimates of DNA concentrations—even when DNA samples had been treated with RNase—compared to estimates obtained using a fluorometric device such as the Qubit (F. Raposo, personal communication, 2016). Thus, when making sequencing libraries it is preferable to use a fluorometer to quantify input DNA samples. For sequencing on an Illumina platform there is an additional critical requirement: the DNA templates must be on the order of several hundred base pairs long. Accordingly, the first step in the traditional library-making process is to shear or fragment the input DNA into smaller pieces (Figure 7.10). When this procedure is performed correctly the researcher obtains a distribution of fragments that are, on average, ~200–300 bp and range ~100–1,000 bp.

Researchers have used a variety of methods for fragmenting DNA such as *nebulization*, *restriction enzymes*, *sonication*, and *acoustic shearing* (Bronner et al. 2014). Although nebulization can be used to generate a sequencing library, it suffers from one major drawback: approximately half of the input sample is lost due to vaporization, which means only a small proportion of the original sample is converted into fragments in the desired size range (Bronner et al. 2014). Another method consists of using restriction enzymes to cut up genomic DNA. The restriction enzyme method can perform well if the procedure is optimized though the G + C content of a genome may ultimately determine how well this method performs (Bronner et al. 2014). Sonication has been a very popular method for creating fragment distributions suitable for Illumina libraries. This method consists of placing a DNA sample in close proximity to a sonicator device for a specific amount of time (longer times result in shorter fragments).

However, like nebulization, use of the sonication method may result in the vast majority of a DNA sample being lost plus heat generated during the sonication procedure may damage the DNA (Bronner et al. 2014). However, one way to improve the efficiency of DNA shearing and normalize shear rates across samples is to use

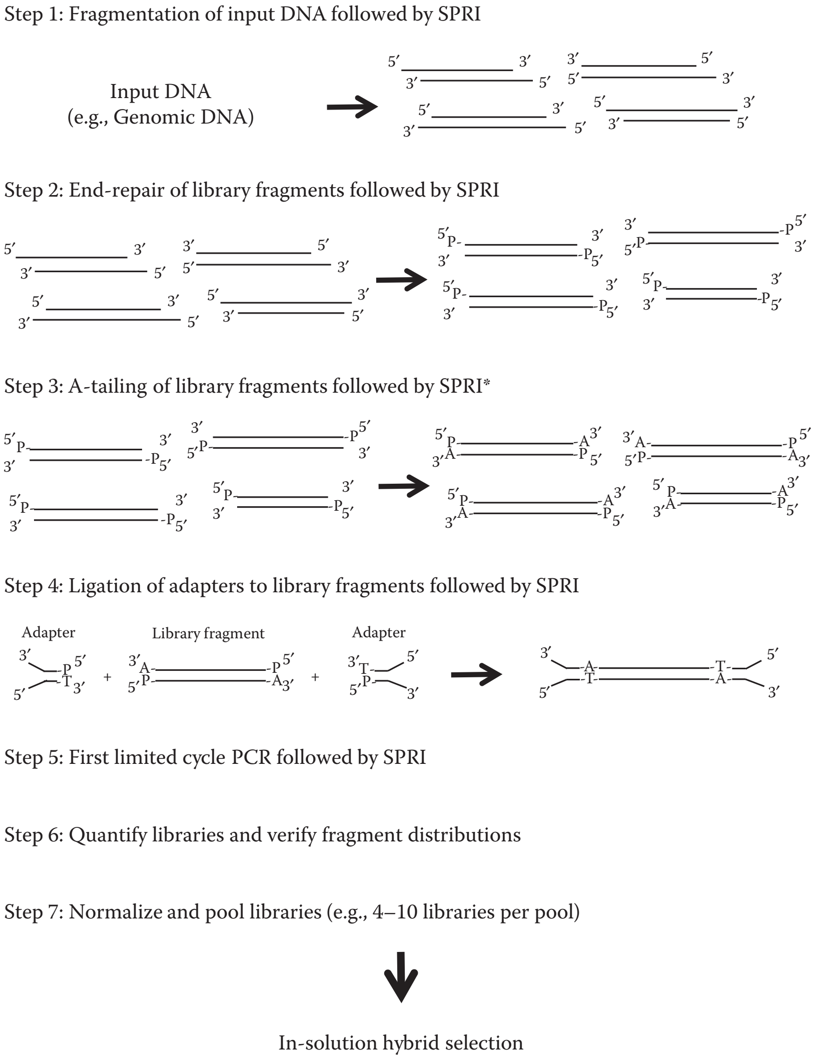

Step 1: Fragmentation of input DNA followed by SPRI
Input DNA
(e.g., Genomic DNA)
Step 2: End-repair of library fragments followed by SPRI
Step 3: A-tailing of library fragments followed by SPRI*
Step 4: Ligation of adapters to library fragments followed by SPRI
Adapter
Library fragment
Adapter
Step 5: First limited cycle PCR followed by SPRI
Step 6: Quantify libraries and verify fragment distributions
Step 7: Normalize and pool libraries (e.g., 4–10 libraries per pool)
In-solution hybrid selection

Figure 7.10. Basic workflow for producing a traditional Illumina sequencing library. Note that the DNA molecules shown in steps 1–4 are largely double-stranded. Following fragmentation of input DNA, the ends of these molecules become "ragged" in that they have 5′ or 3′ single-stranded overhangs. During the end-repair reaction a cocktail of enzymes are used to blunt the ends of the fragments (i.e., make double-stranded end-to-end) and to add a phosphate "P" to the 5′ end of each fragment. Following end-repair, an enzyme adds a single "A" nucleotide to the 3′ end of each fragment. Adapters have complementary 3′ "T" and 5′ "P" to enable ligation with library fragments. A limited cycle PCR is used to selectively amplify only fragments that have an adapter completely ligated to each end. Note that indexing of libraries can occur either during the adapter ligation (i.e., kit protocols) or PCR (i.e., non-kit protocols) steps. See main text for discussion of remaining steps. The asterisk means that the SPRI bead cleanup following the A-tailing reaction is not implemented in some kits, whereas it is used in other protocols (e.g., Fisher et al. 2011). This scheme reflects, with some modifications, the workflows found in the Illumina TruSeq Nano Kit, Bentley et al. (2008), Mamanova et al. (2010), Fisher et al. (2011), and Bronner et al. (2014).

special thin-walled sonication tubes (S. B. Reilly, personal communication, 2016). Focused acoustic shearing is another sound-based method for shearing DNA. In this method, a focused ultrasonicator focuses energy on the sample, which generates the desired narrow size range of fragments without losing a large portion of the original sample (Bronner et al. 2014). This is why acoustic shearing is now the preferred approach to fragmenting DNA. The Truseq kit and protocol by Bronner et al. (2014) both call for using the Covaris S220 or E220 focused ultrasonicators

(Covaris Inc.). An excellent description of Covaris use can be found in Fisher et al. (2011).

Depending on the fragmentation method used, it may be necessary to clean the fragmented DNA before proceeding with the protocol. Also, it is desirable to perform further size-selection procedures on freshly fragmented DNA samples in order to further narrow the fragment size distribution. Although fragments that are smaller or larger than the ideal sizes for Illumina sequencing can be removed in a double-size selection (see Rohland and Reich 2012 for a protocol that uses SPRI beads), in practice usually only the smaller-size fragments are eliminated. Such a size-selection step will enable a larger number of library fragments to create clusters on the flow cell, which can be sequenced. Methods used to accomplish these cleanups include using column-based cleanup kits, agarose-gel extraction followed by column-based cleanups, or SPRI beads. The preferred method at this time is to use SPRI beads owing to the method's efficacy compared to column-based methods, automation-friendliness, and availability of inexpensive generic SPRI beads (see Chapter 6).

Following size selection, it is a good idea to verify the correct fragment size distribution before proceeding with the library-construction process. Although this can be done on a simple 0.7% mini-agarose gel, the preferred approach is to check the results on a Bioanalyzer (Agilent Technologies). This is because the gel can only provide a range of sizes and a rough idea of concentration, whereas the Bioanalyzer can graphically show the actual size distribution and, importantly, indicate the average fragment size as well as the overall concentration (e.g., Mamanova et al. 2010). When performing this genome fragmentation step for the first time researchers should first practice their chosen protocol using an unimportant test DNA sample (e.g., chicken genomic DNA) before attempting the procedure using valuable genetic material and potentially wasting expensive kit reagents.

Step 2: End-Repair of Library Fragments Followed by SPRI The process of fragmenting input DNA creates a large number of smaller dsDNA fragments that have "ragged" ends. This means that many fragments have single-stranded overhangs at their 5′ or 3′ ends and may be missing phosphate groups at their 5′ ends (Figure 7.10). Left unrepaired, these fragments cannot be ligated to library sequencing adapters. Thus, before adapters can be covalently attached to both ends of each library fragment, they must first be enzymatically "end-repaired." Following the end-repair step, each fragment will have blunt or "polished" ends and phosphorylated 5′ termini (Figure 7.10).

The end-repair reaction consists of a cocktail of enzymes, which includes T4 DNA polymerase, Klenow DNA polymerase, and T4 polynucleotide kinase. These enzymes are the usual components of blunt-ending reactions: in the presence of dNTPs T4 DNA polymerase and Klenow DNA polymerase synthesize DNA strands in the 5′ to 3′ direction starting at the 3′ terminal base. Thus, used in tandem (or just T4 DNA polymerase) they are good at filling in gaps on ragged-ended fragments. In addition, both of these enzymes have 3′ to 5′ exonuclease activity and thus they can also remove 3′ overhangs, which also results in blunt-ended products. In addition to blunting fragments, it is critically important to the ligation process that all 5′ terminal bases have a phosphate group. Thus, T4 polynucleotide kinase is included in the end-repair step because this enzyme can phosphorylate the 5′ ends of each fragment lacking those phosphate groups. To complete the process of preparing library fragments for the adapter ligation step, a single "A" nucleotide must be attached to the 3′ ends of each fragment. However, before the next step can be performed, the end polished fragments must be purified otherwise the blunting enzymes will degrade the "A" overhangs soon after they are synthesized. Again, although different methods for cleaning end-repair reactions have been used in the past (e.g., column-based cleanups), SPRI bead cleanup is the preferred method for cleaning end-repair reactions.

Step 3: A-Tailing of Library Fragments Followed by SPRI The purpose of the A-tailing reaction is to add a single "A" nucleotide to the polished 3′ ends of fragments (Figure 7.10). These single 3′ A-overhangs are important because they prevent the formation of chimeric molecules, which can occur when blunt-ended fragments are ligated to each other. As we will see they are also complementary to the single 3′ T-overhangs on the adapters and thus facilitate ligation (Figure 7.10).

The key ingredients of an A-tailing reaction simply consist of the blunted fragments, dATPs, and Klenow DNA polymerase "exo-minus" (or exo-).

Klenow exo- is a mutant version of the Klenow DNA polymerase that has lost its 3′ to 5′ exonuclease activity and thus its function is to simply add a single "A" nucleotide to the 3′ ends of blunted fragments. Following the reaction, the A-tailed library fragments can be cleaned using SPRI beads (e.g., Fisher et al. 2011) though it should be noted that some kit and non-kit protocols do not involve this cleanup.

Step 4: Ligation of Adapters to Library Fragments Followed by SPRI The next step in the workflow consists of ligating adapters to A-tailed library fragments. This traditional library preparation approach uses a single adapter, whereas two distinct adapters are used in other Illumina library preparation methods (see below). If using a library preparation kit, then the adapter constructs are already in complete form; that is, they already contain PCR primer annealing sites, dual indices, and the P5 and P7 common adapters for attachment to flow cell oligos (see Table 7.1 for a list of these oligo sequences). As a side note: whereas adapters in kits are immediately ready for the ligation step (i.e., adapter sequences in Table 7.1 are pre-annealed to each other), non-kit protocols require the user to prepare the adapters for use. This initial construct is called a "forked adapter" because the two sequences comprising the adapter are perfectly complementary to each other for only a 12 bp stretch at one end, whereas the other portions of the two sequences are not complementary to each other (Figure 7.11a). This strange looking adapter design is effective for two reasons. First, the section of the adapter that contains noncomplementary sequences helps to ensure that ligation is directional in that only the correct end of adapters are ligated to A-tailed library fragments (Figure 7.11b). Secondly, the T-overhangs improve the efficiency of the ligation reaction not only because they are complementary to the single A-overhangs on the library fragments, but also by helping to prevent adapters from forming adapter–adapter dimers. In addition to having a 3′ T overhang at the blunt end of the adapter, the adjacent 5′ terminal base is phosphorylated, a feature that enables ligase enzymes to covalently link the 5′ ends of adapter strands to 3′ ends of library fragments (Figure 7.11a). Thus, when the ligation reaction is completed, the desired products will be those that have an adapter completely ligated to each end of a library fragment with the two adapters being inverted with respect to each other (i.e., no nicks will be present in the phosphate backbones of the construct (Figures 7.10 and 7.11b).

During the ligation reaction adapters lacking T-overhangs may become ligated to each other, which results in the formation of adapter dimers that are ~135 bp long. If these adapter dimers

TABLE 7.1

Oligos for Illumina sequencing based on the traditional approach

Name of oligo	Sequence (5′ → 3′)
i5 Index Adapter	AATGATACGGCGACCACCGAGATCTACAC [i5] ACACTCTTTCCCTACACGACGCTCTTCCGATCT
i7 Index Adapter	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC [i7] ATCTCGTATGCCGTCTTCTGCTTG
PCR Primer 1	AATGATACGGCGACCACCGA
PCR Primer 2	CAAGCAGAAGACGGCATACGA
Read 1 Sequencing Primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Read 2 Sequencing Primer	CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
i7 Index Read Sequencing Primer	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

SOURCE: Oligonucleotide sequences © 2007–2010 Illumina, Inc. All rights reserved.

NOTE: The i5 and i7 Index Adapters each contain a unique 8 bp index sequence within the [i5] and [i7] sections of each oligo, respectively. This dual-indexing system allows for the individual identification and hence multiplexing of up to 96 different libraries. Note, oligo sequences are only shown here for purposes of understanding the library-making process using Illumina TruSeq adapters. As Illumina may upgrade its kits, some or all of these oligos may be obsolete. Always check beforehand with the sequencing service provider to make sure that your Illumina libraries are compatible with their sequencing service.

(a)

```
         3′ GTTCGTCTTCTGCCGTATGCTCTA[i7]CACTGACCTCAAGTCTGCACACGAGAAGGCTAG-P 5′
                                                               |||||||||||||
5′ AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT   3′
```

(b)

```
      Annealing sites for primer 2                         A-tailed library fragment ("insert")
   3′GTTCGTCTTCTGCCGTATGCTCTA[i7]CACTGACCTCAAGTCTGCACACGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGCAGCACATCCCTTTCTCACA[i5]CACATCTAGAGCCACCAGCGGCATAGTAA 5′
                                                     |||||||||||||||||||||||||||||||||
5′AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
                                                            A-tailed library fragment ("insert")              Annealing sites for primer 2
```

Figure 7.11. Structure of the Illumina "forked adapter" and initial adapter–fragment construct. (a) The two adapter sequences are partially hybridized to each other, which gives the molecule a "forked" appearance. Note the 3′ "T" overhang and 5′ "P" (phosphate) at the end of the adapter where the bases are complementary to each other (vertical dashes indicate hydrogen bonding between complementary bases). (b) An end-repaired and A-tailed library fragment ligated to two adapters. The middle section of the construct contains the target DNA (a string of Xs) and the // signifies that most of the target fragment length is not shown due to space limitations. For Illumina sequencing fragments are often in the 200–600 bp range depending on the protocol. Adapters contain annealing sites for only one of the two PCR primers (PE PCR Primer 2.0) used in the first limited cycle PCR. (Adapter sequences and PCR primers were obtained from Bentley, D. R. et al. 2008. *Nature* 456:53–59, and are also found in Table 7.1.) Note, these sequences are now obsolete in the Illumina product line though some non-kit protocols use them (e.g., Bronner et al. 2014).

are not removed prior to the PCR (next step in workflow), then they will be coamplified with the desired adapter-ligated library fragments. In this scenario, both adapter dimers and target DNA molecules would be capable of copopulating the flow cell thereby diminishing the quality of sequencing data. A variety of cleanup methods can be used to purify ligation products (Bronner et al. 2014). However, the SPRI bead method has been shown to outperform one column-based method in terms of eliminating adapter dimers from ligation reactions (see Figure 4 in Quail et al. 2008). The ability of SPRI beads to effectively clean ligation reactions in addition to the simplicity of the procedure makes this the best cleanup option.

Before moving on to the next step in the workflow, it is essential to verify that the library fragments were successfully ligated to the adapters. One way to assess the outcome of a ligation reaction is to compare a 1 μL sample of preligation reaction mix with 1 μL of postligation reaction side by side on a Bioanalyzer; the correctly ligated products (i.e., adapter–fragment–adapter molecules) are expected to be ~135 bp longer than unligated library fragments (Bronner et al. 2014).

Step 5: First Limited Cycle PCR (or "prehybridization PCR") Followed by SPRI These initial adapter–fragment constructs must be further modified before they can be made suitable for cluster generation and sequencing. Thus, a special type of PCR called *limited cycle PCR* is used to complete the constructs (Figure 7.10). This is called limited cycle PCR because it only involves 6–18 cycles. Note that in the entire library-making and in-solution hybrid selection workflow there are two limited cycle PCRs that must be done. The first occurs immediately after the ligation reaction (this step) while the second will be performed later in the workflow as we will see. These PCRs are commonly referred to as "enrichment PCR" or "indexing PCR" (when indices are added to the adapters via PCR). Thus to avoid ambiguity we will refer to these as the "first limited cycle PCR" and "second limited cycle PCR."

The ligation reaction purification step is primarily needed to eliminate unused adapters and especially adapter dimers. However, "purified" adapter ligation reaction will still consist of a variety of ligation products such as library fragments with an adapter ligated at each end (i.e., the correct construct; see Figure 7.11b) as well as library fragments with only one adapter or no adapters at all. Moreover, some of these products may have one or more nicks still present in their phosphate backbone due to the vagaries of a ligase reaction (e.g., missing phosphate groups on the 5′ ends of adapters or fragments). It is essential to use PCR in order to enrich the library with the correct adapter constructs. If this is not done, then library fragments containing no adapters or only a single adapter—both of which cannot generate clusters on a flow cell—will compete in the hybridization reaction thereby reducing the capture efficiency of target DNA molecules (Fisher et al. 2011; Rohland and Reich 2012). Let's now examine the mechanics of this first limited cycle PCR.

Notice that in Figure 7.11b the initial adapter construct, perhaps oddly, only shows annealing sites for one of the two PCR primers listed in Table 7.1. Where are the annealing sites for the other primer (i.e., PCR Primer 1)? The reaction starts off in a manner different from standard PCR, which is due to the ingenious design of the forked adapters. During this PCR the adapter constructs must first be resolved before they can be amplified. "Resolving the constructs" refers to the PCR-mediated process that converts initial adapter–fragment constructs into DNA molecules that are double-stranded end-to-end and exhibit the proper configuration for all downstream processes in the workflow. Understanding how this PCR works can be quite confusing owing to its unconventional nature. Thus, to appreciate exactly how the forked adapter functions in PCR we will now closely examine the first two cycles in this PCR.

The limited cycle PCR is similar to standard PCR in that each cycle consists of the familiar denaturation → primer annealing → primer extension steps. Thus, at the start of the first cycle the reaction is heated to ~95°C in order to denature the adapter–fragment constructs into single strands (Figure 7.12). The reaction is then cooled down to the primer-annealing temperature at which time only one of the two PCR primers (i.e., PCR Primer 2) anneals to each strand. Next, the samples are heated to 72°C so that new strands can be synthesized during this primer-extension step (Figure 7.12). Notice that the constructs are now completed by the end of the first cycle. The second cycle begins by denaturing the DNA into single strands once again before proceeding to the second primer-annealing step. Because DNA synthesized in the first cycle included the annealing

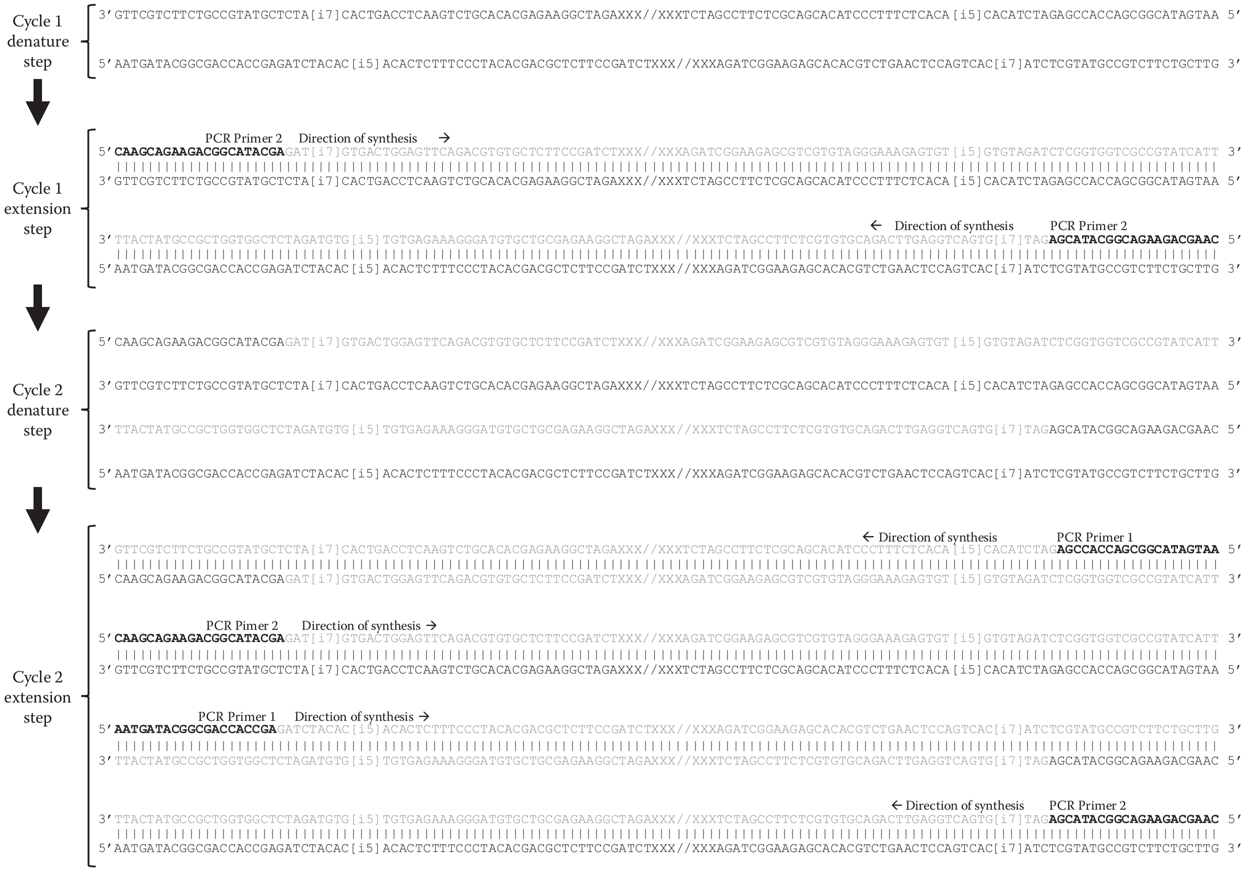

Cycle 1 denature step
3′ GTTCGTCTTCTGCCGTATGCTCTA[i7]CACTGACCTCAAGTCTGCACACGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGCAGCACATCCCTTTCTCACA[i5]CACATCTAGAGCCACCAGCGGCATAGTAA 5′
5′ AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
Cycle 1 extension step
PCR Primer 2
Direction of synthesis →
5′ CAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT[i5]GTGTAGATCTCGGTGGTCGCCGTATCATT 3′
3′ GTTCGTCTTCTGCCGTATGCTCTA[i7]CACTGACCTCAAGTCTGCACACGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGCAGCACATCCCTTTCTCACA[i5]CACATCTAGAGCCACCAGCGGCATAGTAA 5′
← Direction of synthesis
PCR Primer 2
3′ TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
Cycle 2 denature step
5′ CAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT[i5]GTGTAGATCTCGGTGGTCGCCGTATCATT 3′
3′ GTTCGTCTTCTGCCGTATGCTCTA[i7]CACTGACCTCAAGTCTGCACACGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGCAGCACATCCCTTTCTCACA[i5]CACATCTAGAGCCACCAGCGGCATAGTAA 5′
3′ TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
Cycle 2 extension step
← Direction of synthesis
PCR Primer 1
3′ GTTCGTCTTCTGCCGTATGCTCTA[i7]CACTGACCTCAAGTCTGCACACGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGCAGCACATCCCTTTCTCACA[i5]CACATCTAGAGCCACCAGCGGCATAGTAA 5′
5′ CAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT[i5]GTGTAGATCTCGGTGGTCGCCGTATCATT 3′
PCR Pimer 2
Direction of synthesis →
5′ CAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT[i5]GTGTAGATCTCGGTGGTCGCCGTATCATT 3′
3′ GTTCGTCTTCTGCCGTATGCTCTA[i7]CACTGACCTCAAGTCTGCACACGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGCAGCACATCCCTTTCTCACA[i5]CACATCTAGAGCCACCAGCGGCATAGTAA 5′
PCR Primer 1
Direction of synthesis →
5′ AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
3′ TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
← Direction of synthesis
PCR Primer 2
3′ TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′

Figure 7.12. How the first limited cycle PCR completes adapter–fragment library constructs using the traditional Illumina approach. The denaturation and primer-extension steps in the first two cycles of the reaction are illustrated (primer-annealing steps not shown due to space limitation). The initial adapter constructs only contain annealing sites for PCR Primer 2. Annealing sites for the PCR Primer 1 become available after the first cycle and complete adapter–fragment library constructs are produced at the end of the second cycle. The 8-bp i5 and 8-bp i7 index sequences are represented by [i5] and [i7], respectively. The middle section of each strand contains the target DNA (a string of Xs). The // signifies that most of the target fragment is omitted due to space limitation and the vertical dashes indicate hydrogen bonding between complementary bases. Boldface sequences represent newly annealed primer in the same cycle and gray letters represent DNA synthesized in this reaction.

sites for the other PCR primer (i.e., PCR Primer 1) both primers can now anneal to templates and be extended (Figure 7.12). Therefore the constructs can be amplified in an exponential manner over the remaining cycles. The completed adapter–fragment construct, which was shown in Figure 7.1b, now has all sequences in their correct configuration including the P5 and P7 common adapters for the flow cell oligos, the i5 and i7 indices, Read 1 and Read 2 Sequencing Primers, and the i7 Index Read Sequencing Primer. The final procedure of this first limited PCR step is to purify the PCR products using SPRI beads.

Regarding the number of cycles used in this PCR there is an important tradeoff to consider: a larger number of cycles will generate more unique genomic fragments for sequencing but it will also duplicate many fragments and create biases in the distribution of fragments particularly if many PCR cycles are used (Fisher et al. 2011; Bronner et al. 2014). In general, it is advisable to use the fewest number of PCR cycles when enriching libraries. Bronner et al. (2014) provide some rough guidelines for specifying the number of PCR cycles depending on the amount of input DNA: 8 cycles if 500 ng is used, 12 cycles if 200–500 ng, 14 cycles if 50–200 ng, and 18 cycles if <50 ng is input.

Step 6: Quantify Libraries and Verify Fragment Distributions Typically in phylogenomic studies the libraries are pooled together before they are used in the hybrid selection phase of the workflow. To do this, however, the concentration of each library must be estimated and its fragment distribution verified. The most accurate way to estimate the concentration of a sequencing library is to perform qPCR using primers that anneal to the flow cell oligo sites on the adapters (Bronner et al. 2014). However, because qPCR is expensive, concentration estimates based on the Bioanalyzer are usually adequate for normalizing and pooling libraries (Bronner et al. 2014; A. D. Gottscho, personal communication, 2016). Many other studies have used fluorometric devices (e.g., Qubit) to estimate library concentrations. Although it may be possible to check the size range of adapter-ligated fragments on an agarose gel, this approach is not ideal because it does not show what the average adapter–fragment size is. Thus, Bioanalyzer assays represent the preferred means for examining fragment distributions of libraries. Bioanalyzer electropherograms can also indicate whether or not adapter dimers were amplified (Bronner et al. 2014). The average fragment sizes will depend largely on the input library fragment distribution at the beginning of the workflow. In any case, better cluster generation results will be obtained if the average adapter–fragment size matches the target size specified in the cluster generation kit that will be used.

Step 7: Normalize and Pool Libraries (e.g., 4–10 Libraries per Pool) In most phylogenomic studies, owing to the large number of libraries made (e.g., >25), all indexed libraries will ultimately be pooled together before being sequenced in at least one flow cell lane. However, at this point in the workflow a decision must be made as to how many indexed libraries will be combined together *for each hybridization reaction*. This is because of a tradeoff: the higher the number of libraries being "captured" in a single hybridization reaction the lower the per library hybridization efficiency. This means that pools containing a larger number of libraries will be more susceptible to "missing data" problems as different loci in different individuals will have missing sequences. Although the opposite extreme—using a single library per hybridization reaction—will yield the best capture results, this strategy is costly in terms of reagents. Another important factor to consider is genome size because larger genomes have fewer copies of each target (Lemmon et al. 2012; Portik et al. 2016; S. B. Reilly, personal communication, 2016). For example, for frogs, which have very large genomes, hybridization pools consisting of 4–6 libraries have worked well (Portik et al. 2016; S. B. Reilly, personal communication, 2016). For organisms with smaller genomes, the recent trend has been for researchers to group around 6–10 indexed libraries into individual hybridization pools (e.g., Faircloth et al. 2013, 2015; Smith et al. 2013; McCormack et al. 2015). Using the obtained concentration estimates, each library is first normalized so that all libraries will have equal representation in the resulting sequencing data. Thus, following normalization and deciding pool sizes libraries are combined in equimolar ratios to form each pool. Once this step is completed, library pools are ready to enter the hybrid selection workflow followed by cluster generation and sequencing (Figure 7.10). If problems are encountered while preparing traditional Illumina sequencing libraries, the reader can consult

Bronner et al. (2014) and Illumina Truseq protocol for troubleshooting tips.

If we look back at the entire traditional library preparation workflow in Figure 7.10, we see that there are four to five discrete SPRI bead cleanup steps. Fisher et al. (2011) developed a fantastic innovation to improve the efficiency and efficacy of this workflow, which relates to these cleanup steps. They developed a method called *with-bead SPRI*, which means that the SPRI beads added to the DNA samples following the fragmentation step *remain* with the input DNA throughout the library-making procedure. In between enzymatic reactions, only the PEG/NaCl buffer is exchanged (Figure 7.13). The library fragments are only eluted from the beads following the cleanup of PCR products made in the first limited cycle PCR. Thus, the cleanup steps are linked together rather than occurring as discrete steps in the workflow (Fisher et al. 2011).

There are many significant benefits to using this with-bead SPRI workflow. First, fewer beads are consumed, which saves a considerable amount of money. Second, this approach is automation friendly and thus large numbers of libraries can be constructed using liquid-handling robots or by humans with multichannel pipettes. Third, owing to the efficiency of the SPRI method relative to column-based purification kits as well as the fewer liquid transfer steps, 80%–90% of the DNA is recovered versus 50%–60% recovery rate for column-based cleanups (Fisher et al. 2011).

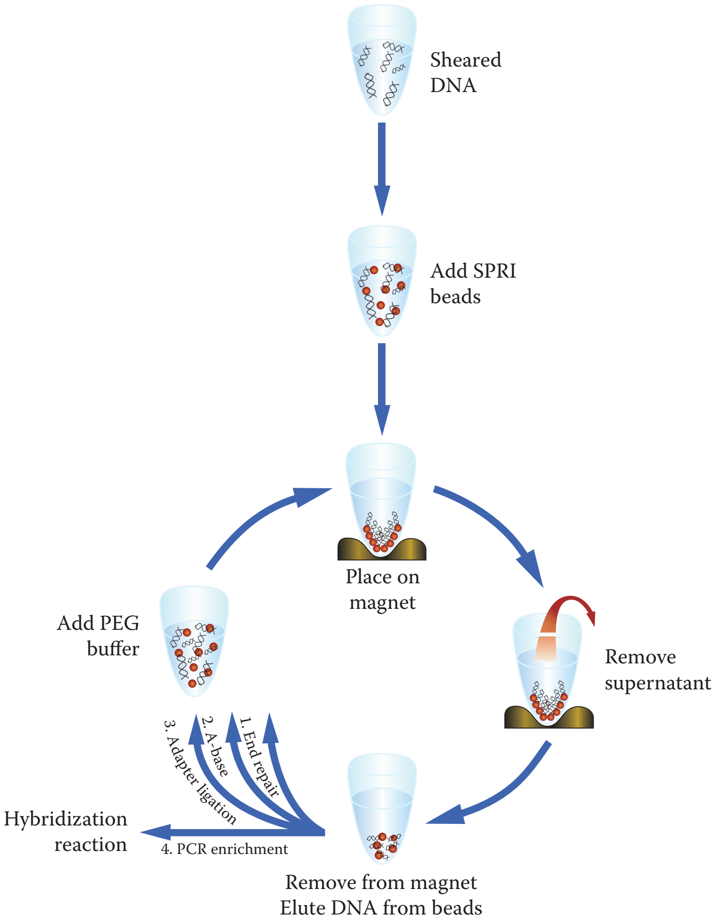


Figure 7.13. Fisher "with-bead" library preparation workflow. In contrast with earlier SPRI-bead clean-up protocols, which require the addition of new aliquots of beads at the start of each enzymatic clean-up step, the Fisher et al. method only requires the addition of fresh beads immediately after the DNA fragmentation step. In subsequent clean-up steps, the PEG buffer (and the not the beads) are exchanged. See main text and Fisher et al. (2011) for further explanation. (Reprinted from Fisher, S. et al. 2011. *Genome Biol* 12:1–15. With permission.)

Because of this improvement in the DNA recovery rate, less input DNA is needed to start a library (e.g., 100 ng can be used to obtain good hybrid selection results) and fewer PCR cycles (e.g., 6–8) are needed in the first limited cycle PCR. A fourth benefit is that the with-bead SPRI workflow can increase the diversity of a library (i.e., increase the number of unique genomic fragments) 12-fold. Moreover, the with-bead method can be used for any of the library-making approaches described in this chapter. With all these advantages it is not surprising that researchers have adopted this modification in their library-making protocols (e.g., Rohland and Reich 2012; Faircloth et al. 2015; McCormack et al. 2015).

7.2.1.2 Meyer and Kircher Library Approach

Meyer and Kircher (2010) developed a protocol for easily and inexpensively making large numbers of indexed Illumina sequencing libraries. This approach, which is based on the 454 sequencing library construction methods of Margulies et al. (2005), has been a popular method for generating enormous phylogenomic datasets (e.g., Lemmon et al. 2012; Portik et al. 2016). The basic workflow for the Meyer and Kircher approach is outlined in Figure 7.14. This approach shares many of the same steps with the traditional library preparation approach (compare with Figure 7.10) and thus we will not discuss in detail every step as we did before.

Like the traditional library preparation method, the Meyer and Kircher approach begins with the fragmentation of an input DNA sample but it does not involve an SPRI bead cleanup of the fragmentation products. However, as an SPRI bead cleanup can be used to good effect in this step (i.e., eliminate fragments too small for sequencing; Fisher et al. 2011), this can be considered an optional cleanup. Next, library fragments are end-repaired followed by an SPRI bead cleanup (Figure 7.14). Instead of A-tailing the blunt-ended and phosphorylated library fragments as is done in the traditional method, the blunt-ended fragments are ligated to nonphosphorylated adapters in a blunt-ended ligation reaction. Another difference between methods is that the Meyer and Kircher approach uses two different adapters (Figure 7.14), which are called “Adapter P5” and “Adapter P7,” whereas the traditional approach uses a single forked adapter (Figures 7.10 and 7.11a). Thus, while a ligation reaction involving forked adapters will only produce one type of adapter–fragment–adapter construct, a ligation reaction with the Adapter P5 and Adapter P7 combo can generate three different adapter–fragment–adapter products: Adapter P5–library fragment–Adapter P5, Adapter P5–library fragment–Adapter P7, and Adapter P7–library fragment–Adapter P7. As you may have just realized, only the ligation products containing both Adapter P5 and Adapter P7 can create clusters on a flow cell. However, as we will see below, ligation products having two Adapter P5s and two Adapter P7s will not affect the quality of the completed library.

Let’s take a closer look at the structure of Adapters P5 and P7. Table 7.2 shows the three different oligos (i.e., IS1, IS2, and IS3) that are used to construct Adapters P5 and P7. Notice that each of these three oligos is fortified with four phosphorothioate bonds on their 5′ and 3′ ends (Table 7.2). These special bonds are added to prevent 3′ to 5′ exonuclease digestion of the oligos. If these bases are not protected, then they would be vulnerable to being digested from contaminating exonucleases or from the blunting enzymes used in the end-repair reaction. Adapters P5 and P7 are made by hybridizing oligo IS1 with IS3 and IS2 with IS3, respectively (Figure 7.15a,b). The two strands of each adapter type are complementary to each other but notice that one strand on each adapter has a long 5′ overhang (Figure 7.15a,b). The existence of these overhangs ensures that each adapter can be ligated to the library fragment in only one way (Margulies et al. 2005). Unlike the Illumina forked adapter (see Figure 7.11a), Adapters P5 and P7 do not have phosphates at their 5′ ends. This means that ligase enzymes cannot seal the phosphate backbone where an unphosphorylated 5′ adapter end is positioned adjacent to the 3′ end of a library fragment. In contrast, ligase does seal the backbone where the 3′ end of an adapter comes into contact with the phosphorylated 5′ end of a library fragment. The ligation reaction thus produces an initial adapter construct with two “nicks” (or gaps) present in the phosphate backbone along with two 5′ overhangs (Figures 7.14 and 7.15c). Before this initial adapter construct can be subjected to the first limited cycle PCR, a “fill-in” reaction must be performed in which an enzyme called *Bst* DNA polymerase, Large Fragment, is used to fill-in the missing bases and seal both nicks (Margulies

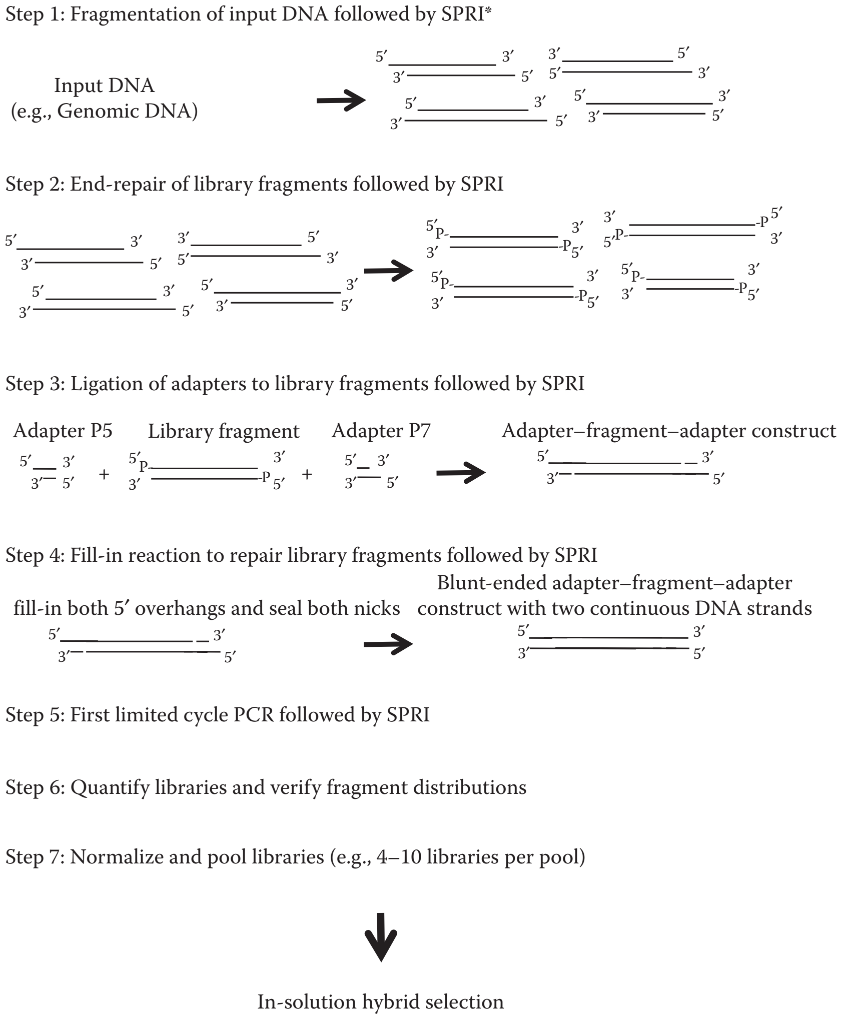

Step 1: Fragmentation of input DNA followed by SPRI*
Input DNA
(e.g., Genomic DNA)
Step 2: End-repair of library fragments followed by SPRI
Step 3: Ligation of adapters to library fragments followed by SPRI
Adapter P5
Library fragment
Adapter P7
Adapter–fragment–adapter construct
Step 4: Fill-in reaction to repair library fragments followed by SPRI
fill-in both 5′ overhangs and seal both nicks
Blunt-ended adapter–fragment–adapter construct with two continuous DNA strands
Step 5: First limited cycle PCR followed by SPRI
Step 6: Quantify libraries and verify fragment distributions
Step 7: Normalize and pool libraries (e.g., 4–10 libraries per pool)
In-solution hybrid selection

Figure 7.14. Basic workflow for producing an Illumina sequencing library using the Meyer and Kircher approach. Note that the DNA molecules shown in steps 1–4 are largely double-stranded. Following fragmentation of input DNA, the ends of these molecules become “ragged” in that they have 5′ or 3′ single-stranded overhangs. During the end-repair reaction a cocktail of enzymes are used to blunt the ends of the fragments (i.e., make double-stranded end-to-end) and to add a phosphate “P” to the 5′ end of each fragment. Following end-repair, two different adapters (Adapters P5 and P7) are ligated to the blunt-ended library fragments. In Step 4 a fill-in reaction adds bases to complement the 5′ overhangs and to seal the two nicks in the phosphate backbone (shown as gaps in the strands). Next, a limited cycle PCR is used to selectively amplify only fragments that have both types of adapters completely ligated the ends and to complete the adapter constructs by adding sample-specific indices, sequences complementary to flow cell oligos and sequencing primers. The asterisk means the SPRI bead cleanup following fragmentation is not included in the actual Meyer and Kircher (2010) protocol, but it can be used at this step to eliminate fragments too small for sequencing. See main text for descriptions of the other steps.

et al. 2005; Meyer and Kircher 2010). The resulting products consist of adapter–fragment constructs that are double-stranded end-to-end (i.e., without nicks) and contain the annealing sites for both PCR primers used in the first limited cycle PCR (Figure 7.15d; Table 7.2).

The adapter–fragment constructs are now ready to be used as templates in the first limited cycle PCR (Figure 7.14). Although this first PCR is used to “enrich” the library with the correct adapter–fragment constructs in both the traditional Illumina library preparation and Meyer and Kircher approaches, there are some differences in the PCR procedure between these two methods. While the traditional kit-based approach uses PCR to resolve the adapter constructs using a pair

TABLE 7.2

Oligos for Illumina sequencing using the Meyer and Kircher approach

Name of oligo	Sequence (5′ → 3′)
IS1_adapter.P5	A*C*A*C*TCTTTCCCTACACGACGCTCTTCCG*A*T*C*T
IS2_adapter.P7	G*T*G*A*CTGGAGTTCAGACGTGTGCTCTTCCG*A*T*C*T
IS3_adapter.P5 + P7	A*G*A*T*CGGAA*G*A*G*C
IS4_indPCR.P5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT
Indexing PCR primer	CAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGT
IS5_reamp.P5	AATGATACGGCGACCACCGA
IS6_reamp.P7	CAAGCAGAAGACGGCATACGA
IS7_short_amp.P5	ACACTCTTTCCCTACACGAC
IS8_short_amp.P7	GTGACTGGAGTTCAGACGTGT
BO1.P5.F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-Pho
BO2.P5.R	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT-Pho
BO3.P7.part1.F	AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-Pho
BO4.P7.part1.R	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-Pho
BO5.P7.part2.F	ATCTCGTATGCCGTCTTCTGCTTG-Pho
BO6.P7.part2.R	CAAGCAGAAGACGGCATACGAGAT-Pho
Read 1 Sequencing Primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Read 2 Sequencing Primer	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Index Read Sequencing Primer	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

NOTE: Oligos IS1 to IS3 are used to make the initial adapters for the ligation reaction (asterisk = phosphorothioate bond); IS4 and the indexing PCR primer are used in the first limited cycle PCR; IS5 and IS6 are used to amplify a library following target capture and IS7 and IS8 can be used to amplify a library prior to the indexing PCR; BO1-06 are blocking oligos for target capture (Pho = 3′ phosphate); and Read 1, Read 2, and Index Read Sequencing Primers are used to sequence the completed library construct. Note, the [i7] sequence embedded in the indexing primer represents the 7-bp index sequence (see Supplementary File in Meyer and Kircher 2010 for a list of 228 different indexing PCR primers). (All oligos are from Meyer, M. and M. Kircher. 2010. *Cold Spring Harbor Protocols*. doi: 10.1101/pdb.prot5448.) Note, these oligos are shown here only to help illustrate this library construction approach. Readers should consult the original source by Meyer and Kircher (2010) for the actual step-by-step protocol as well as listing of all required oligos for making multiplexed Illumina libraries.

of typical nontailed PCR primers, the Meyer and Kircher approach instead uses 5′-tailed PCR primers in this PCR in order to complete each construct by adding additional sequence elements including indices. Because PCR is used to index libraries in this fashion—rather than adding them to the constructs during the ligation step—that is why this is sometimes called an indexing PCR. Note, there are traditional non-kit library protocols (e.g., Bronner et al. 2014) that also use tailed PCR primers in the first limited cycle PCR to complete the constructs similar to the Meyer and Kircher approach but we will not consider those methods here.

Let's now examine the first limited cycle PCR in the Meyer and Kircher workflow. In contrast to the traditional approach, the initial adapter construct in the Meyer and Kircher approach contains annealing sites for both PCR primers (Figure 7.15d). Thus, if we look at the first three cycles of a limited PCR using the Meyer and Kircher adapter–fragment constructs we see that both primers (IS4 and indexing PCR primer) anneal to their templates during the first cycle (Figure 7.16). By the end of the second cycle, we see that full-length adapter–fragment strands are produced and that completed double-stranded constructs only appear at the end of the third cycle. The remaining cycles are used to amplify the number of these completed constructs. Figure 7.17 shows the final sequencing-ready library construct, which

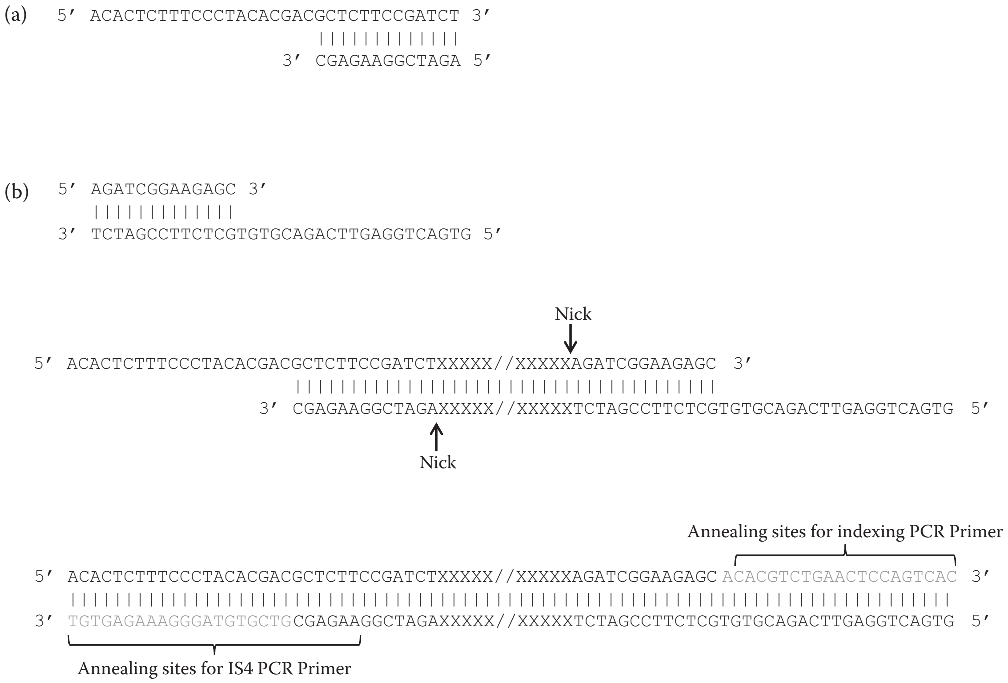

(a)
5′ ACACTCTTTCCCTACACGACGCTCTTCCGATCT 3′
3′ CGAGAAGGCTAGA 5′
(b)
5′ AGATCGGAAGAGC 3′
3′ TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG 5′
Nick
5′ ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXX//XXXXXAGATCGGAAGAGC 3′
3′ CGAGAAGGCTAGAXXXXX//XXXXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG 5′
Nick
Annealing sites for indexing PCR Primer
5′ ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXX//XXXXXAGATCGGAAGAGC ACACGTCTGAACTCCAGTCAC 3′
3′ TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXXXX//XXXXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG 5′
Annealing sites for IS4 PCR Primer

Figure 7.15. Structure of the Meyer and Kircher Illumina adapters and initial adapter–fragment constructs. (a) Adapter P5 consists of the IS1 (above) and IS3 (below) oligos hybridized to each other (vertical dashes indicate hydrogen bonding between complementary bases). (b) Adapter P7 consists of the IS3 (above) and IS2 (below) oligos hybridized to each other. Note, phosphorothioate bonds not shown (see Table 7.2). (c) Following the ligation reaction, the initial adapter–fragment construct has long 5′ overhangs at each end as well as two nicks in the phosphate backbone at two junctions where the adapters and fragment meet. (d) A fill-in reaction repairs this initial construct by sealing the two nicks and filling in bases in a 5′ to 3′ direction (gray bases represent newly synthesized DNA). The result is a double-stranded and blunt-ended adapter–fragment construct. The middle section of the adapter–fragment constructs contains the target DNA (a string of Xs). The // signifies that most of the target fragment is omitted due to space limitation. Adapters contain annealing sites for IS4 and indexing PCR primers, which are used in the first limited cycle PCR. (Adapter sequences and PCR primers were obtained from Meyer, M. and M. Kircher. 2010. *Cold Spring Harbor Protocols*. doi: 10.1101/pdb.prot5448, and are also found in Table 7.2.)

contains the following elements: P5 and P7 flow cell adapters, 7-bp sample-specific index, annealing sites for the PCR primers in the second limited cycle PCR (IS5 and IS6 Primers), annealing sites for the Read 1 and Read 2 Sequencing Primers, and annealing sites for the Index Read Sequencing Primer.

Earlier we saw that the Meyer and Kircher ligation products consist of three types of adapter–fragment–adapter constructs. Given that only one of the three construct types can create clusters on the flow cell, should we be concerned that PCR might coamplify all three constructs and thus adversely affect sequencing data? Fortunately, this is not a problem thanks to an interesting phenomenon known as the *suppression PCR effect* (Siebert et al. 1995; Diatchenko et al. 1996; Lukyanov et al. 2007). Following the denaturation phase of PCR, the single-stranded templates cool down toward the primer-annealing temperature. Strands that have inverted terminal repeat sequences—in this case an adapter at one end and the reverse complement of this same adapter at the opposing end—are likely to form intramolecular terminal duplexes (secondary structures) that resemble “panhandles” (Siebert et al. 1995; Diatchenko et al. 1996; Lukyanov et al. 2007; Meyer and Kircher 2010). For ligation products having only a single adapter type on both ends, this means that the panhandle duplex will consist of a P5 or P7 sequence duplexed to its reverse complement sequence found at the other end of the strand. Thus, fragments having only a single adapter type will not be amplified in this reaction; instead, only the fragments having

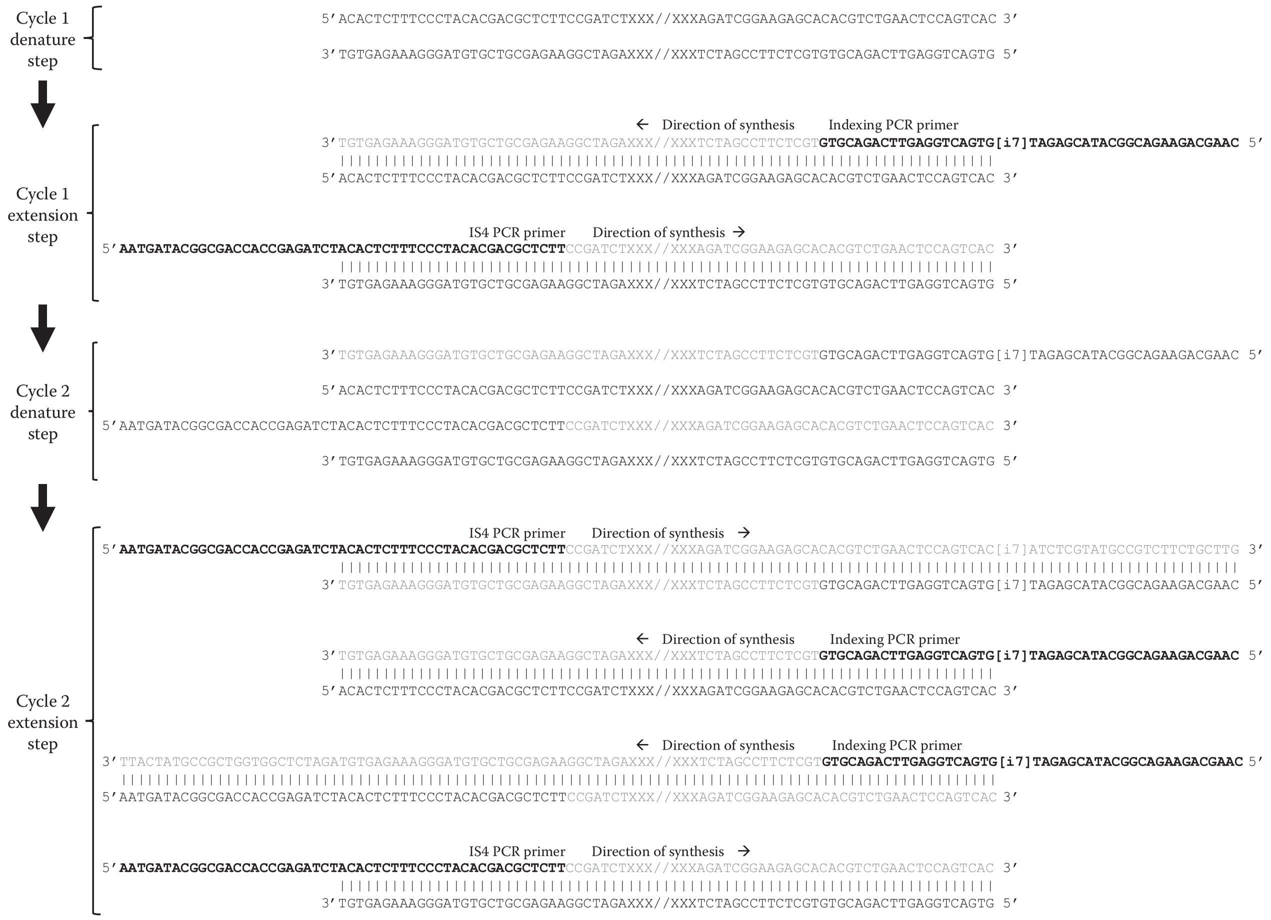

Cycle 1 denature step
5′ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
3′TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG 5′
Cycle 1 extension step
Direction of synthesis
Indexing PCR primer
3′TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
IS4 PCR primer
Direction of synthesis
5′AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
3′TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG 5′
Cycle 2 denature step
3′TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
5′AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
3′TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG 5′
Cycle 2 extension step
IS4 PCR primer
Direction of synthesis
5′AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
3′TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
Direction of synthesis
Indexing PCR primer
3′TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
Direction of synthesis
Indexing PCR primer
3′TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
IS4 PCR primer
Direction of synthesis
5′AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
3′TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG 5′

Figure 7.16. How the first limited cycle PCR completes adapter–fragment library constructs using the Meyer and Kircher approach. The denaturation and primer-extension steps in the first three cycles of a limited PCR are illustrated (primer-annealing steps not shown due to space limitation). *(Continued)*

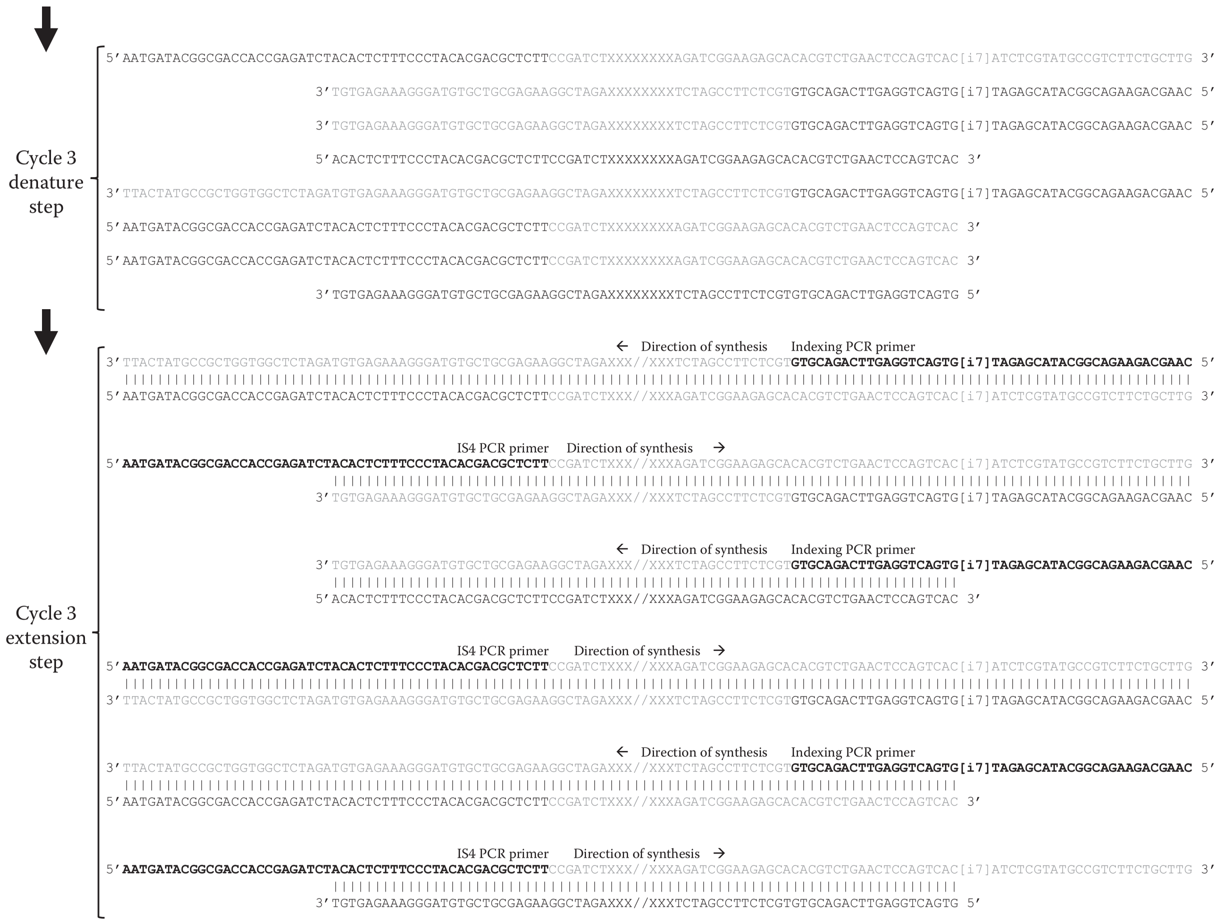

Cycle 3
denature
step
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXXXXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
3′ TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXXXXXXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
3′ TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXXXXXXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXXXXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
3′ TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXXXXXXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXXXXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXXXXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
3′ TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXXXXXXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG 5′
Cycle 3
extension
step
← Direction of synthesis
Indexing PCR primer
3′ TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
IS4 PCR primer
Direction of synthesis →
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
3′ TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
← Direction of synthesis
Indexing PCR primer
3′ TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
IS4 PCR primer
Direction of synthesis →
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
3′ TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
← Direction of synthesis
Indexing PCR primer
3′ TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
IS4 PCR primer
Direction of synthesis →
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
3′ TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG 5′

Figure 7.16. (Continued) The initial adapter constructs contain annealing sites for both the IS4 PCR primer and indexing PCR primer. Completed adapter–fragment constructs appear only at the end of the third cycle. The 7-bp index sequence is represented by [i7]. The middle section of the adapter–fragment constructs contains the target DNA (a string of Xs). The // signifies that most of the target fragment is omitted due to space limitation and the vertical dashes indicate hydrogen bonding between complementary bases. Boldface sequences represent newly annealed primer in the same cycle and gray letters represent DNA synthesized in this reaction.

```
                                                                                                                P7 flow cell oligo
                                                                                                       3′ TAGAGCATACGGCAGAAGACGAAC 5′

                                                                          Read 2 Sequencing Primer                IS6_reamp.P7
                                                                 3′ TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGT 5′  3′ AGCATACGGCAGAAGACGAAC 5′

3′TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXX//XXXTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG[i7]TAGAGCATACGGCAGAAGAC GAAC 5′
  |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
5′AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXX//XXXAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTG CTTG 3′

                        5′ ACACTCTTTCCCTACACGACGCTCTTCCGATCT 3′   5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCAC 3′
                                 Read 1 Sequencing Primer                 Index Read Sequencing Primer
5′AATGATACGGCGACCACCGA 3′
      IS5_reamp.P5

5′AATGATACGGCGACCACCGAGAUCTACAC 3′
          P5 flow cell oligo
```

Figure 7.17. Completed Meyer and Kircher adapter–fragment construct for Illumina sequencing. The construct contains annealing sites for the P5 and P7 flow cell oligos, PCR primers for amplifying the library after target capture (i.e., primers IS5 and IS6), Read 1 and 2 Sequencing Primers, and the Index Read Sequencing Primer. The 7-bp index sequence is represented by [i7]. The middle section of the adapter–fragment construct contains the target DNA (a string of Xs). The // signifies that most of the target fragment is omitted due to space limitation and the vertical dashes indicate hydrogen bonding between complementary bases. (Oligos are from Meyer, M. and M. Kircher. 2010. *Cold Spring Harbor Protocols* doi: 10.1101/pdb.prot5448, and are also found in Table 7.2.)

both types of adapters will be amplified, which is the desired result (Meyer and Kircher 2010). When PCR is used to selectively amplify products in this manner it is referred to as *suppression PCR* (Siebert et al. 1995; Diatchenko et al. 1996; Lukyanov et al. 2007; Adey et al. 2010).

Following the limited cycle PCR, the PCR products are cleaned using SPRI beads. At this time, the libraries are quantified, normalized, and pooled in the same manner already described for the traditional approach.

7.2.1.3 Rohland and Reich Library Approach

The Rohland and Reich (2012) approach to making Illumina sequencing libraries resembles the Meyer and Kircher approach because it is also based on the 454 library preparation methods developed by Margulies et al. (2005). Indeed, their library-making workflow consists of the same basic steps found in the Meyer and Kircher workflow (see Figure 7.14) except that the former approach uses an SPRI bead cleanup step following fragmentation of the DNA, whereas the latter does not (see Figure 1 in Rohland and Reich 2012 for a detailed workflow). There are, however, a number of important differences that distinguish these approaches. First, Rohland and Reich (2012) developed a high-throughput workflow that can enable academic laboratories to generate thousands of sequencing libraries (192 indexed libraries/day) at far lower cost/library (one to two orders of magnitude) compared to commercial library kits. Another major innovation of the Rohland and Reich approach concerns the use of novel adapters, which can greatly enhance the quality of phylogenomic datasets.

Table 7.3 lists the oligos that comprise these adapters as well as other oligos required for the library and in-solution hybrid selection workflows. Unlike other Illumina adapter designs discussed in this chapter, their P5 adapter contains a 6-bp internal "barcode" sequence (Table 7.3). The barcode is used to individually mark libraries so that they can be grouped together into hybrid

TABLE 7.3
Oligos for Illumina sequencing using the Rohland and Reich approach

Name of oligo	Sequence (5′ → 3′)
Barcoded P5 adapter	CTTTCCCTACACGACGCTCTTCCGATCT [barcode]
Barcoded P5 adapter complement	[barcode] AGATCGGAA
PE P7 adapter	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
PE P7 adapter complement	AGATCGGAAGAGC
PreHyb-PE_F	CTTTCCCTACACGACGCTCTTC
PreHyb-PE_R	CTCGGCATTCCTGCTGAACC
Sol-PE-PCR_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC
Index PE Primer	CAAGCAGAAGACGGCATACGAGAT [index] CGGTCTCGGCATTCCTGCTGAACC
Index PE Sequencing Primer	GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG
PE Read 1 Sequencing Primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
PE Read 2 Sequencing Primer	CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
Univ_Block_P5	AGATCGGAAGAGCGTCGTGTAGGGAAAG
Univ_Block_P7	AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

SOURCE: Rohland, N. and D. Reich. 2012. *Genome Res* 22:939–946, and © 2007–2010 Illumina, Inc. All rights reserved.

NOTE: The first four oligos are used to make the two adapters for the ligation reaction; PreHyb-PE_F and PreHyb-PE_R are used to amplify the initial adapter-library fragment construct in the first limited cycle PCR; Read 1, Read 2, and Index Read Sequencing Primers are used to sequence the completed library construct, and Univ_Block P5 and P7 are blocking oligos for target capture. A 6-bp barcode sequence is used to individually mark libraries prior to making hybridization pools. A 7-bp index sequence is added during the second limited PCR to distinguish pools (see Supplementary File in Rohland and Reich 2012 for a list of 16 different indexing PCR primers). All oligos are from Rohland and Reich (2012). Note, these oligos are shown here only to help illustrate this library construction approach. Readers should consult Rohland and Reich (2012) for the actual protocol as well as listing of all required oligos for making multiplexed Illumina libraries. Always check beforehand with the sequencing service provider to make sure that your Illumina libraries are compatible with their sequencing service.

selection pools. We will return to this subject later and discuss the significance of this barcode sequence and how it can lead to better quality sequencing data.

When each pair of complementary adapter sequences is annealed to the other, they form two distinct adapters (Figure 7.18a,b). Like the Meyer and Kircher adapters, both adapters do not have phosphorylated 5′ ends, which is by design in order to prevent the formation of adapter dimers and reduce oligo costs (Rohland and Reich 2012). The single-stranded 5′ overhangs on the adapters (Figure 7.18a,b) also ensure that they can only be ligated to library fragments in the correct manner.

After the library fragments have been end-repaired/phosphorylated and SPRI-cleaned, they are ligated to the adapters. The desired ligation products are similar to those produced in a ligation reaction with the Meyer and Kircher adapters (compare Figures 7.15c and 7.18c). Accordingly, a nick-sealing and fill-in reaction using *Bst* DNA Polymerase, Large Fragment, must be performed in order to create constructs that are double-stranded from end-to-end (Figure 7.18d). Notice that this adapter construct contains the annealing sites for both PCR primers, which are used in the first limited cycle PCR (a.k.a. suppression or prehybridization PCR; Figure 7.18d) and thus both primers can begin synthesizing new strands in the first PCR cycle (not shown).

In contrast to the Meyer and Kircher approach, the first limited cycle PCR in the Rohland and Reich approach does not involve the use of 5′ tailed PCR primers. As you may recall, the Meyer and Kircher approach uses tailed primers in the first PCR to index and complete sequencing-ready constructs (see Figure 7.16). The Rohland and Reich constructs on the other hand, are not

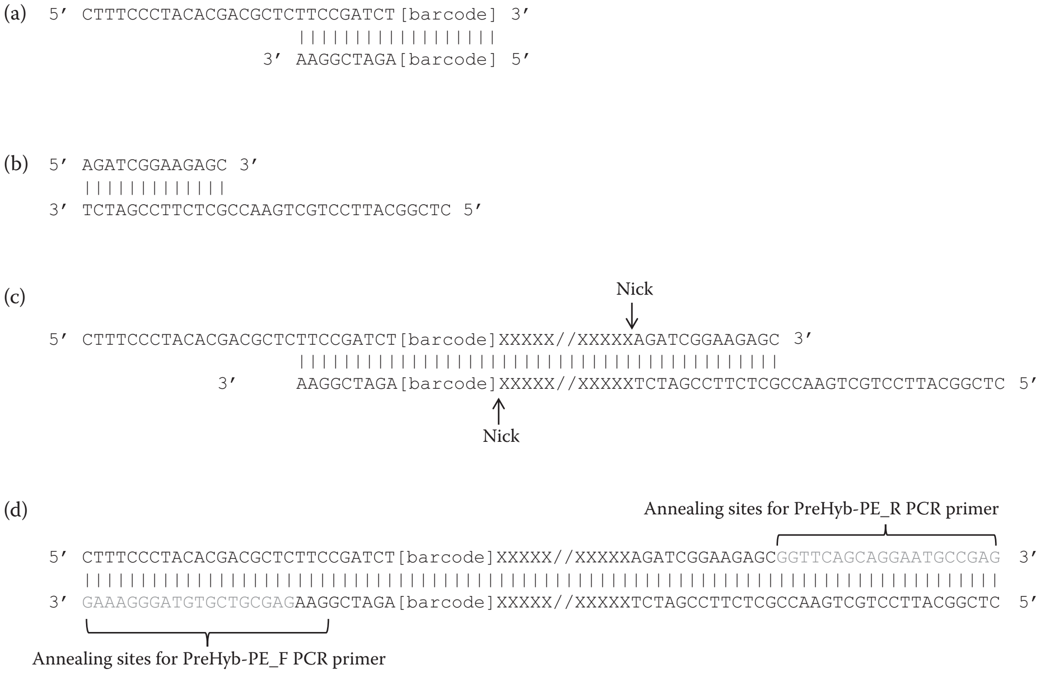

(a)
5′ CTTTCCCTACACGACGCTCTTCCGATCT[barcode] 3′
3′ AAGGCTAGA[barcode] 5′
(b)
5′ AGATCGGAAGAGC 3′
3′ TCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTC 5′
(c)
Nick
5′ CTTTCCCTACACGACGCTCTTCCGATCT[barcode]XXXXX//XXXXXAGATCGGAAGAGC 3′
3′ AAGGCTAGA[barcode]XXXXX//XXXXXTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTC 5′
Nick
(d)
Annealing sites for PreHyb-PE_R PCR primer
5′ CTTTCCCTACACGACGCTCTTCCGATCT[barcode]XXXXX//XXXXXAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG 3′
3′ GAAAGGGATGTGCTGCGAGAAGGCTAGA[barcode]XXXXX//XXXXXTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTC 5′
Annealing sites for PreHyb-PE_F PCR primer

Figure 7.18. Structure of the Rohland and Reich Illumina adapters and initial adapter–fragment constructs. (a) One adapter contains a 6-bp "barcode" sequence that is used to mark individual libraries prior to making hybridization pools (vertical dashes indicate hydrogen bonding between complementary bases). (b) The second adapter does not contain barcode or index sequences and resembles one of the Meyer Kircher adapters (see Figure 7.15b). (c) The initial ligation product contains two nicks in the phosphate backbone and two long 5′ overhangs. Before this construct can be used in the first limited cycle PCR, the nicks must be repaired and the overhangs filled in. (d) Following the fill-in reaction, the adapter–fragment construct is double-stranded end-to-end and contains annealing sites for the PreHyb-PE_F and PreHyb-PE_R PCR Primers, which are used in the first limited cycle PCR. The middle section of the adapter–fragment construct contains the target DNA (a string of Xs). The // signifies that most of the target fragment is omitted due to space limitation. (Adapter sequences and PCR primers were obtained from Table 7.3.)

augmented during the first limited PCR; they are merely enriched with the correct adapter constructs (i.e., library fragments having both initial adapters). Thus, following the first limited cycle PCR, these constructs lack the sequences that enable them to hybridize with flow cell oligos (i.e., P5 and P7 common adapters). The rationale for making these incomplete or "truncated" adapter constructs (Rohland and Reich 2012) prior to the in-solution hybridization reaction(s) will be explained below. Only during the second limited cycle PCR will the captured truncated constructs be completed through the addition of P5 and P7 common adapters, a second index sequence, and sites for the various sequencing primers.

The remaining steps of this workflow, which are standard to all library methods, include quantifying and verifying libraries, normalizing libraries, and making prehybridization pools. As mentioned before, the current norm in phylogenomic studies is to group four to ten indexed libraries into a single pool. Each of these pools, in turn, will be used in a separate hybridization reaction (see below).

7.2.1.4 Nextera Library Approach

An alternative to the ligase-based methods for preparing Illumina sequencing libraries involves the use of the "Nextera" library preparation kit. Nextera kits for Illumina sequencing were initially developed and sold by Epicentre Biotechnologies (Madison, Wisconsin; Syed et al. 2009a,b). Illumina subsequently acquired Epicentre in 2011 and has since further refined and diversified the Nextera kits for a number of different applications (see the Illumina website: http://www.illumina.com/products/nextera_dna_library_prep_kit.html). Numerous helpful guides dealing with Nextera library preparation can be found on the Illumina Nextera website pages. At the present time, Nextera kits are available in low- or high-throughput formats and up to 96 sample-specific dual-indices can be employed. In phylogenomic studies, Nextera kits have been used to generate the initial whole-genome sequencing libraries, which were then used as input libraries for kit-based (e.g., SureSelect Custom Probe Kit, Agilent, Inc.) in-solution hybrid selection and sequencing (e.g., Faircloth et al. 2013; Meiklejohn et al. 2016).

Nextera kits have a number of advantages over other library preparation methods. First, the amount of starting material (input genomic DNA) needed for Nextera library preparation is far less than the required amount for other library approaches. For example, in an exome capture study, Nextera only requires an input amount of 50 ng of DNA, whereas a traditional library would require ~3 μg (Adey et al. 2010). Moreover, while the Nextera kits have been optimized for input of 50 ng of starting material, the Nextera XT kit has been optimized for input of only 1 ng of total genomic DNA per reaction. Even lower input amounts are possible, as Adey et al. (2010) found that inputs as low as 10 pg (0.01 ng) of genomic DNA can be used! A second advantage Nextera has is that fewer steps are involved compared to other library preparation methods. This not only conserves consumables (e.g., plastics) and labor but also reduces the risk of errors (e.g., contamination) during the library preparation procedures. The Nextera library making workflow is illustrated in Figure 7.19. We will now review the Nextera methodology.

Step 1: Tagmentation of input DNA followed by SPRI As with the previously discussed methods of library preparation a sample of purified DNA must be acquired (note that RNA can be present because it does not interfere with the tagmentation reaction). For example, in the Nextera kit approach the required amount of genomic DNA employed as starting material is only 50 ng. However, as pointed out in the Illumina Nextera protocol, the kit has been optimized for input of 50 ng of DNA. If a larger amount of DNA is used then the resulting library fragment distribution will consist of fragments that are, on average, larger than desired, whereas if <50 ng is used, then the size distribution will shift toward smaller-sized fragments. Thus, it is important to use the exact amount of input DNA specified in the protocol. Ligase-mediated library methods, on the other hand, are more robust to the input DNA amount with larger amounts generally yielding better sequencing results.

Preparation of an Illumina library using the Nextera approach is similar to the other approaches in that the starting DNA must be broken into fragments small enough for Illumina sequencing and special adapters must be ligated to the ends of each fragment. However, the Nextera approach uses a radically different means during the initial steps to accomplish this. Recall that the previous library-making methods require the

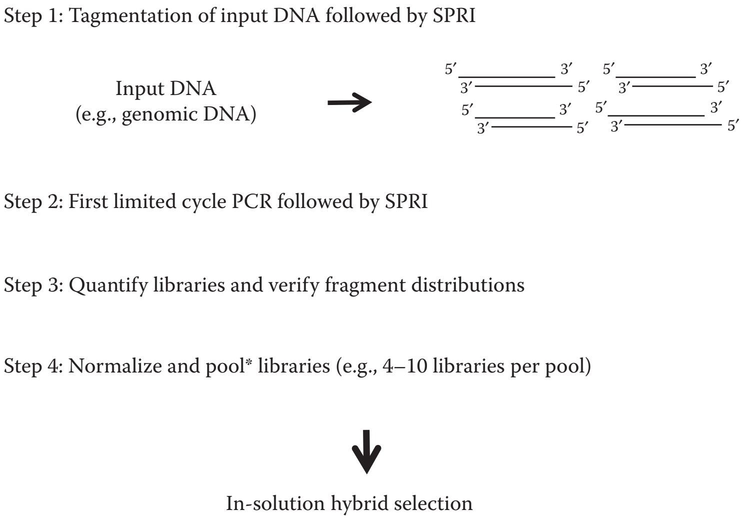

Step 1: Tagmentation of input DNA followed by SPRI
Input DNA
(e.g., genomic DNA)
5′
3′
3′
5′
Step 2: First limited cycle PCR followed by SPRI
Step 3: Quantify libraries and verify fragment distributions
Step 4: Normalize and pool* libraries (e.g., 4–10 libraries per pool)
In-solution hybrid selection

Figure 7.19. Basic workflow for producing an Illumina sequencing library using the Nextera approach. This method uses transposases to simultaneously fragment input genomic DNA and attach adapters to the ends of library fragments. This process is called "tagmentation." The initial adapter–fragment constructs are then subjected to the first limited cycle PCR before they are cleaned using SPRI beads. See main text for discussion of remaining steps.

starting DNA material to be fragmented (e.g., via sonication), end-repaired, A-tailed (traditional approach only), and ligated to adapters. In contrast, the Nextera method only involves a one-step 5-minute enzymatic reaction—a process called *tagmentation*, which results in adapter-ligated fragments ready for the first limited cycle PCR step (Figure 7.19).

The tagmentation reaction uses *in vitro* DNA transposition to simultaneously fragment the input DNA into sizes appropriate for sequencing on Illumina platforms and ligating or "tagging" adapter molecules to the ends of each fragment. This is why there are no steps in this protocol that involve manual shearing of input DNA followed by separate ligation reactions. The Nextera kit is optimized so that adapters are ligated, on average, every 400–500 bp thereby creating the ideal adapter-ligated fragment distribution for the Illumina platforms. Before we continue discussing Nextera procedures, let's first review aspects of *in vitro* transposition involving the Tn5 transposase—the key player in this method.

Wildtype Tn5 transposase carries out "cut and paste" transposition in prokaryotes (Steiniger-White et al. 2004). Tn5 transposons are DNA transposons that are excised from and reinserted back into the host genome with little specificity for insertion locations (Reznikoff 2003). Briefly, the mechanism of *in vivo* transposition consists of the following steps: (1) two separate transposase subunits work together to excise a Tn5 transposon from the host DNA; (2) the two subunits form a transposase homodimer that is bound to the transposon; and (3) this newly formed "synaptic complex" then binds to a new target location in the host DNA and inserts the transposon (see Steiniger-White et al. 2004 for more details).

The wildtype form of Tn5 is not useful for *in vitro* transposition owing to its exceedingly low activity (Reznikoff 2003). However, researchers have successfully used a "hyperactive" (mutated) version of the Tn5 transposase to conduct *in vitro* transposition experiments (Reznikoff 2008). Normally, a Tn5 synaptic complex will have a single dsDNA molecule (the transposon) bound to the homodimer with each end of the DNA molecule being inserted into an active site of each transposase subunit (i.e., a loop of DNA is formed from one subunit to the other). When this complex attacks a target DNA molecule, the end result will be insertion of the entire transposon into the target DNA (see Figure 1 in Steiniger-White et al. 2004).

The ends of the Tn5 transposon contain critical sequence elements important for transposition; that is, they are essential for recognition by the transposases, synaptic complex formation, and the insertion of the transposon back into the target DNA (Steiniger-White et al. 2004). These

```
   19                  1
5′ AGATGTGTATAAGAGACAG 3′   ←  19 bp ME sequence (transferred strand)
   |||||||||||||||||||
3′ TCTACACATATTCTCTGTC 5′   ←  Complement (nontransferred strand)
```

Figure 7.20. The 19-bp ME sequence for Tn5 transposase. (Modified from Figure 1 in Bhasin, A. et al. 2000. *J Mol Biol* 302:49–63.)

19-bp sequence ends, which are called "outside end" (OE) and "inside end" (IE) sequences, are nearly identical to each other (see Figure 1 in Bhasin et al. 2000). A 19-bp "hybrid" between OE and IE elements is called the "mosaic end" (ME) sequence (Figure 7.20). As with the OE and IE sequences, this ME sequence contains the essential chemical features for enabling synaptic complex formation and subsequent transposition with target DNA. In particular, the first nucleotide position (GC base pair) of the ME (Figure 7.20), which is in contact with the active site of a transposase subunit, will be positioned closest to where a staggered double-stranded cut in the target DNA takes place during the transposition reaction. However, only one of the ME strands (i.e., transferred strand; see Figure 7.20) will be ligated to a target strand; that is, the 3′ end of this transferred strand (i.e., "G" nucleotide at position (1) will be covalently joined to the 5′ end of the newly fragmented target DNA. Researchers discovered that if transposase synaptic complexes are assembled *in vitro* using *two* free ME sequences instead of the single longer Tn5 transposon, then the ultimate result of transposition will be fragmentation of the target DNA with the ME sequences ligated to each fragment (Reznikoff 2008). Figure 7.21 shows the crystal structure of a Tn5 synaptic complex with

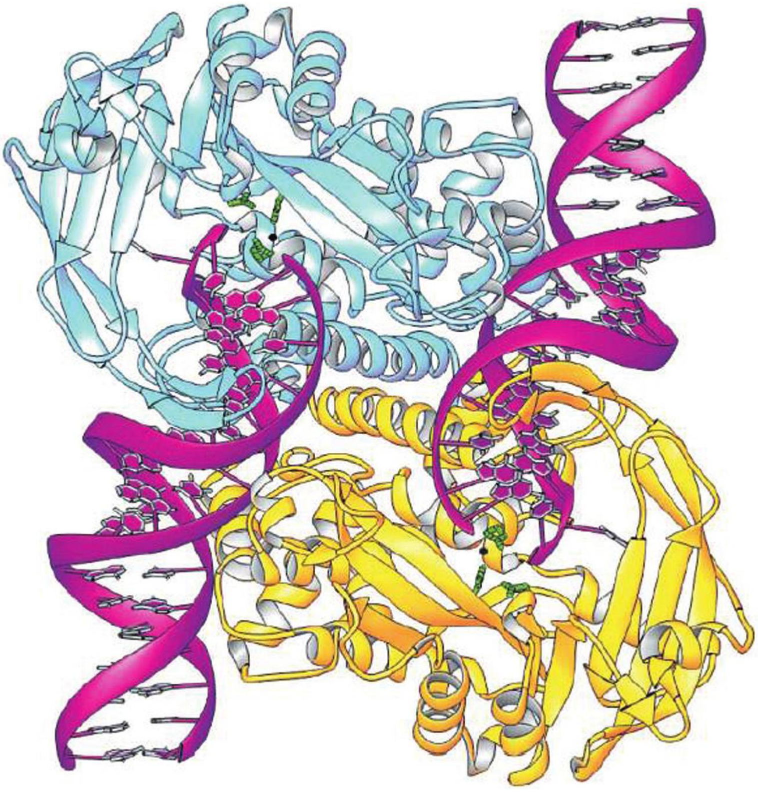


Figure 7.21. Crystal structure of Tn5 transposase/DNA complex. The transposase synaptic complex is comprised of a homodimer between two transposase subunits (yellow and blue ribbons) with two DNA transposon ends (purple) protruding from the active sites of the subunits. Catalytic residues in the active site are shown as green ball-and-stick structures and the associated Mn^{2+} ion is black. (Reprinted from Davies, D. R. et al. 2000. *Science* 289:77–85. With permission.)

TABLE 7.4

Oligos for Illumina sequencing using the Nextera approach

Name of oligo	Sequence (5′ → 3′)
Index 1 Read Primer	CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG
Index 2 Read Primer	AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTC
Read 1 Sequencing Primer	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
Read 2 Sequencing Primer	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

SOURCE: Oligonucleotide sequences © 2007–2010 Illumina, Inc. All rights reserved.

NOTE: Index 1 and 2 Read Primers are used in the first limited cycle PCR; and Read 1 and Read 2 Sequencing Primers are used to sequence the completed library construct. The i5 and i7 Indices each contain a unique 8-bp index sequence. This dual-indexing system allows for the individual identification and hence multiplexing of up to 96 different libraries. Note, oligo sequences are only shown here for purposes of understanding the library-making process using Illumina Nextera approach. As Illumina may upgrade its kits, some or all of these oligos may be obsolete. Always check beforehand with the sequencing service provider to make sure that your Illumina libraries are compatible with their sequencing service.

two transposon sequences protruding away from the homodimer. We will now see how this experimental system evolved into a novel and simple NGS library preparation method.

Researchers took advantage of the free 5′ ends of each ME sequence by tailing them with platform-specific adapter sequences (Syed et al. 2009a,b). The Illumina Nextera adapter sequences are listed in Table 7.4 along with other essential oligos for this library-making approach. We can see each of these 5′-tailed adapter or "transposon" sequences on the left side of Figure 7.22. Notice that the right half of each sequence is comprised of the critical 19-bp ME sequence, whereas the left half contains the sites necessary for the first limited cycle PCR in Step 2 (Figure 7.19). Although only the transferred strand of each adapter sequence will be ligated to the end of a library fragment via the transposition mechanism, a double-stranded adapter sequence is required for assembly of the Tn5 synaptic complex.

Figure 7.23 shows a tagmentation reaction involving ME-flanked adapter sequences. Initially, when synaptic complexes are assembled *in vitro*, there is randomness with respect to which particular ME-flanked adapters become annealed to the transposase subunits. This means that some complexes will have one of each adapter type such as the complexes having one blue and one orange adapter in Figure 7.23a, while the other two classes of complexes in the solution (not shown) will have only one adapter type (i.e., all blue or all orange adapters). In the next step of the reaction (Figure 7.23b), the two synaptic complexes that are shown attack the target DNA where they will create staggered double-stranded breaks in the target DNA and initiate strand-transfer from ME-flanked

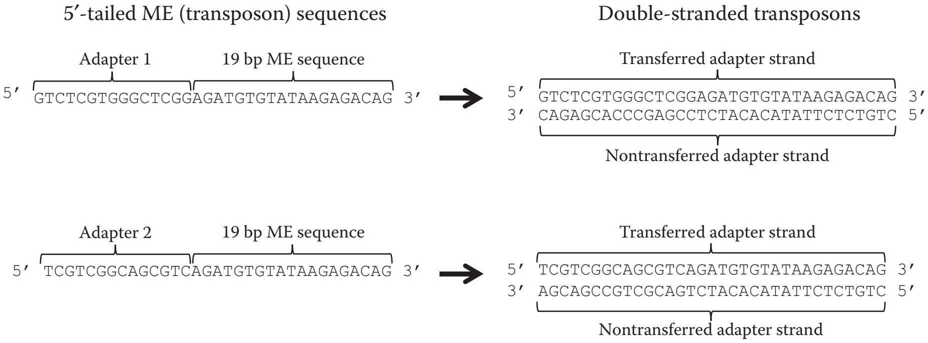

5′-tailed ME (transposon) sequences
Double-stranded transposons
Adapter 1
19 bp ME sequence
5′ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG 3′
Transferred adapter strand
5′ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG 3′
3′ CAGAGCACCCGAGCCTCTACACATATTCTCTGTC 5′
Nontransferred adapter strand
Adapter 2
19 bp ME sequence
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG 3′
Transferred adapter strand
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG 3′
3′ AGCAGCCGTCGCAGTCTACACATATTCTCTGTC 5′
Nontransferred adapter strand

Figure 7.22. Nextera adapter sequences. Two ME sequences with 5′ tail sequences consisting of an Illumina-specific adapters 1 and 2 (left side). These tailed-ME sequences must be double-stranded "transposons" (right side) in order to form synaptic complexes with Tn5 transposases. Note that only one of the two strands of each transposon (i.e., transferred adapter strand) will be transferred to a library fragment during tagmentation. (Oligonucleotide sequences © 2007–2010 Illumina, Inc. All rights reserved.)

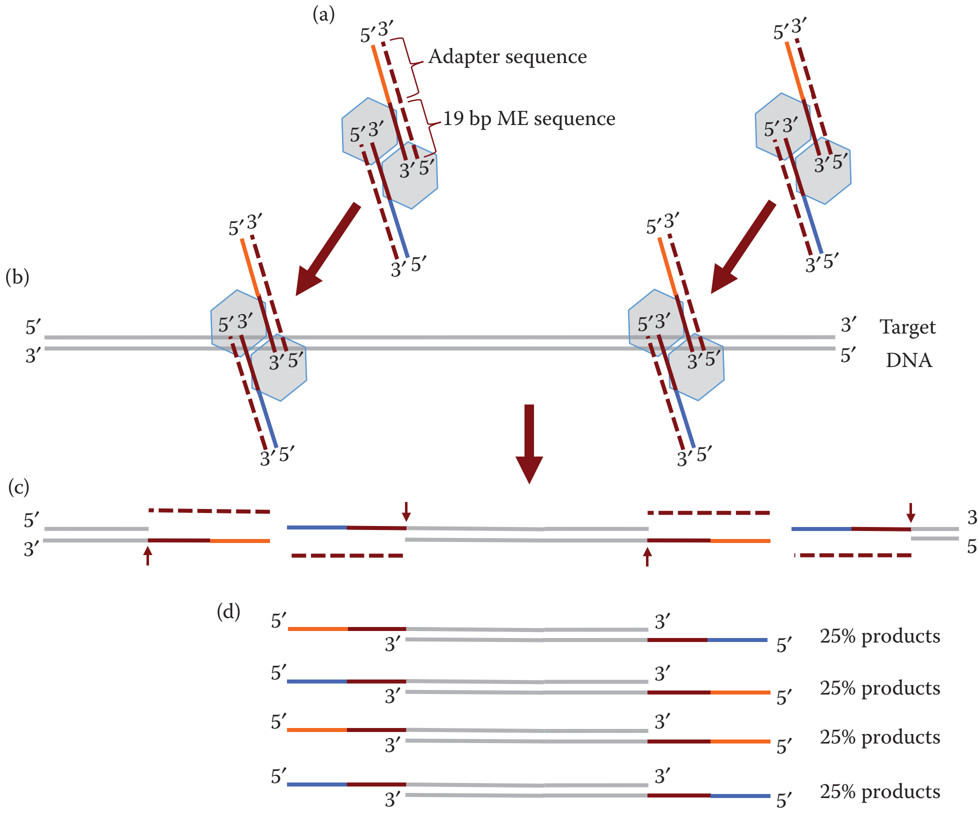


Figure 7.23. Tagmentation reaction involving two synthetic Tn5 transposase synaptic complexes and target DNA. (a) Each synaptic complex is comprised of a transposase homodimer (two hexagons) and two ME-flanked adapters. Each ME sequence terminates in the active site of a subunit (near center of each hexagon) while the adapter ends (blue and orange ends) are free and not in contact with the subunits. Solid color DNA strands are the strands that will be transferred (covalently attached) to the target DNA, while the (dashed) complementary strands are the nontransferred strands and thus will not be incorporated into target DNA. (b) Each synaptic complex independently attacks the target DNA. Upon binding, a staggered double-stranded cut is made and the 3′ ends of the ME sequences (transferred strands) subsequently become ligated to the 5′ end of the target DNA strands. (c) When the reaction is finished, the target DNA is fragmented and tagged ("tagmented") with ME-flanked adapters. Small arrows indicate the ligation points between ME-flanked adapters and target DNA. (d) Products of tagmentation. The topmost product (blue and orange ends) is the complete product shown in step (c), while the bottom two represent the other products present in the reaction. (Parts of this figure are based on Figure 2 in Steiniger-White, M. et al. 2004. *Curr Opin Struct Biol* 14:50–57 and Figure 1 in Syed, F. et al. 2009. *Nat Methods* 6:10; Syed, F. et al. 2009. *Nat Methods* 6:11.)

adapters to target DNA. Strand transfer occurs when the 3′ ends of the transferred (adapter) strands are ligated to the 5′ ends of the target DNA (Figure 7.23c). Just like the ligation reactions in the Meyer–Kircher and Rohland–Reich approaches, which also use two different adapters, there will also be an element of randomness concerning which adapter is ligated to the end of a particular fragment (Figure 7.23d). Accordingly, a limited cycle (suppression) PCR will be used to selectively amplify the library fragments having both adapters (Figure 7.19; Syed et al. 2009a,b; Adey et al. 2010).

Performing the tagmentation reaction step is simple as it consists of first adding the input DNA, enzyme buffer, and transposase enzyme/adapter complexes to a reaction tube followed by a 5-minute incubation period at 55°C in a thermocycler. At the conclusion of this *in vitro* transposition reaction, the transposases remain bound to the target DNA (Reznikoff 2003) and thus they must be removed from the tagmentation products and discarded otherwise they can interfere with downstream steps. In the past, the Illumina Nextera protocol dictated the use of column-based DNA cleanups for this purpose. However, these kits now rely on SPRI bead cleanups as they perform better than column-based cleanups (e.g., Faircloth et al. 2012).

Step 2: First Limited Cycle PCR Followed by SPRI As with the traditional and Meyer and Kircher approaches, limited cycle PCR is used to (1) complete the initial adapter constructs by adding the sample-specific indices, P5 and P7 common adapters, and annealing sites for sequencing primers; and (2) selectively amplify the correct tagmentation products (i.e., library fragments with both kinds of adapters). The Nextera kit uses a dual index system with each index being 8-bp long. Thus, one PCR primer contains the Index 1 or "i7 index" sequence while the other primer contains the Index 2 or "i5 index" sequence (Table 7.4). There are 12 and 8 different indices for the i7 and i5 index schemes, respectively, which means 96 different libraries can ultimately be pooled together into a single Illumina flow cell.

The desired tagmentation product consists of a library fragment with two 5′ overhangs each of which corresponds to a different adapter sequence (Figure 7.24a). However, inspection of the tagmentation product (Figure 7.24a) and the tailed PCR primers that will be used to amplify that product in the first limited cycle PCR (Figure 7.24b) reveals that this initial adapter–fragment construct lacks annealing sites for these primers. Instead, both adapter sequences ligated to the library fragment are actually the *same* sequence as the two primer sequences. How then, does the PCR proceed?

The trick is to begin the thermocycling program with a 3–5 minute extension step at 72°C to allow DNA polymerases a chance to fill-in the missing bases just before the normal PCR cycling begins (i.e., cycle = denaturation → annealing → extension steps). This crucial detail cannot be overlooked while performing a limited cycle PCR involving tagmentation products—to reiterate, the thermocycling program *must* begin with this 72°C step and *not* the usual 94–98°C denaturation step! Once this first extension step is completed the initial adapter constructs are double-stranded molecules from end to end and, importantly, now contain the needed annealing sites for both PCR primers (Figure 7.25). After the initial extension step has generated the PCR primer annealing sites on the adapter–fragment constructs, the thermocycler program then subjects them to at least five normal PCR cycles in order to complete these constructs and amplify their number. Similar to the limited cycle PCRs in the traditional and Meyer and Kircher approaches, fully formed constructs do not appear until after the

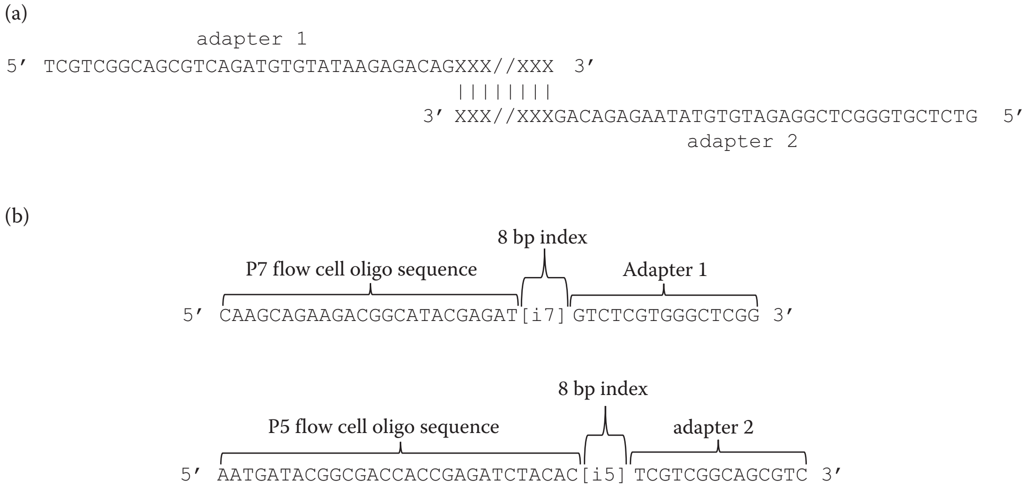

(a)
adapter 1
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXX 3′
||||||||
3′ XXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG 5′
adapter 2
(b)
8 bp index
P7 flow cell oligo sequence
Adapter 1
5′ CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG 3′
8 bp index
P5 flow cell oligo sequence
adapter 2
5′ AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTC 3′

Figure 7.24. Tagmentation product and PCR primers used in the first limited cycle PCR. (a) The desired initial tagmentation product contains adapter 1 and adapter 2 sequences (vertical dashes indicate hydrogen bonding between complementary bases). The middle section of the adapter–fragment construct contains the target DNA (a string of Xs). The // signifies that most of the target fragment is omitted due to space limitation. (b) Two PCR primers, called "Index 1 Read Primer" (top) and "Index 2 Read Primer" (bottom), have 5′ tail sequences, which are used to add sequences matching the P5 and P7 flow cell oligos and to incorporate dual 8-bp indices into adapter–fragment constructs during the first limited cycle PCR. Oligos are found in Table 7.4. (Oligonucleotide sequences © 2007–2010 Illumina, Inc. All rights reserved.)

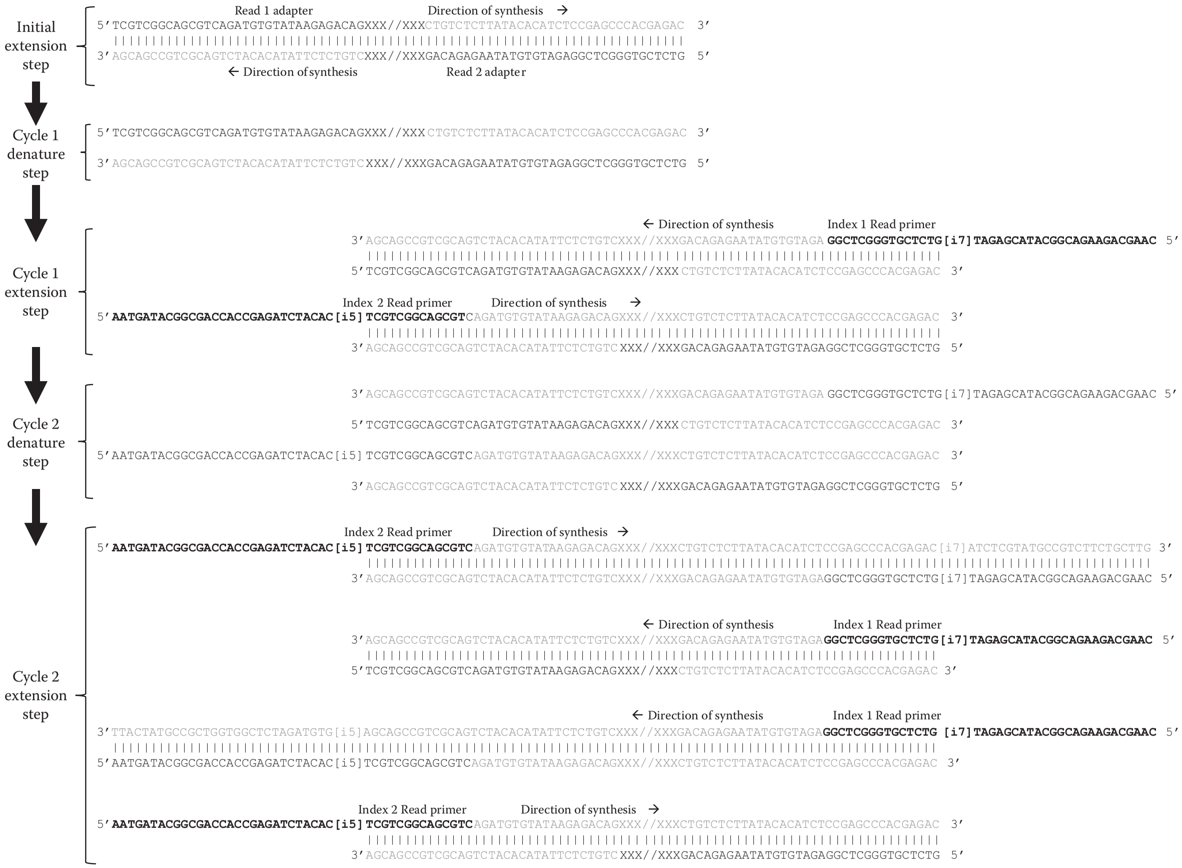

Initial extension step
Read 1 adapter
Direction of synthesis →
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
3′ AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG 5′
← Direction of synthesis
Read 2 adapter
Cycle 1 denature step
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXX CTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
3′ AGCAGCCGTCGCAGTCTACACATATTCTCTGTC XXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG 5′
Cycle 1 extension step
← Direction of synthesis
Index 1 Read primer
3′ AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGA GGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXX CTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
Index 2 Read primer
Direction of synthesis →
5′ AATGATACGGCGACCACCGAGATCTACAC[i5] TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
3′ AGCAGCCGTCGCAGTCTACACATATTCTCTGTC XXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG 5′
Cycle 2 denature step
3′ AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGA GGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXX CTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
5′ AATGATACGGCGACCACCGAGATCTACAC[i5] TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
3′ AGCAGCCGTCGCAGTCTACACATATTCTCTGTC XXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG 5′
Cycle 2 extension step
Index 2 Read primer
Direction of synthesis →
5′ AATGATACGGCGACCACCGAGATCTACAC[i5] TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
3′ AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
← Direction of synthesis
Index 1 Read primer
3′ AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
← Direction of synthesis
Index 1 Read primer
3′ TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG [i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′ AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
Index 2 Read primer
Direction of synthesis →
5′ AATGATACGGCGACCACCGAGATCTACAC[i5] TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
3′ AGCAGCCGTCGCAGTCTACACATATTCTCTGTC XXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG 5′

Figure 7.25. How the first limited cycle PCR completes adapter–fragment library constructs using the Nextera approach. The denaturation and primer-extension steps in the first three cycles of a limited PCR are illustrated (primer-annealing steps not shown due to space limitation). The initial adapter constructs does not contain annealing sites for the Index 1 and 2 Read Primers and thus the first step is an extension step to allow polymerases to fill in the missing bases. Completed adapter–fragment constructs appear only at the end of the third cycle. The 8-bp i5 and 8-bp i7 index sequences are represented by [i5] and [i7], respectively. The middle section of each strand contains the target DNA (a string of Xs). The // signifies that most of the target fragment is omitted due to space limitation and the vertical dashes indicate hydrogen bonding between complementary bases. Boldface sequences represent newly annealed primer in the same cycle and gray letters represent DNA synthesized in this reaction. *(Continued)*

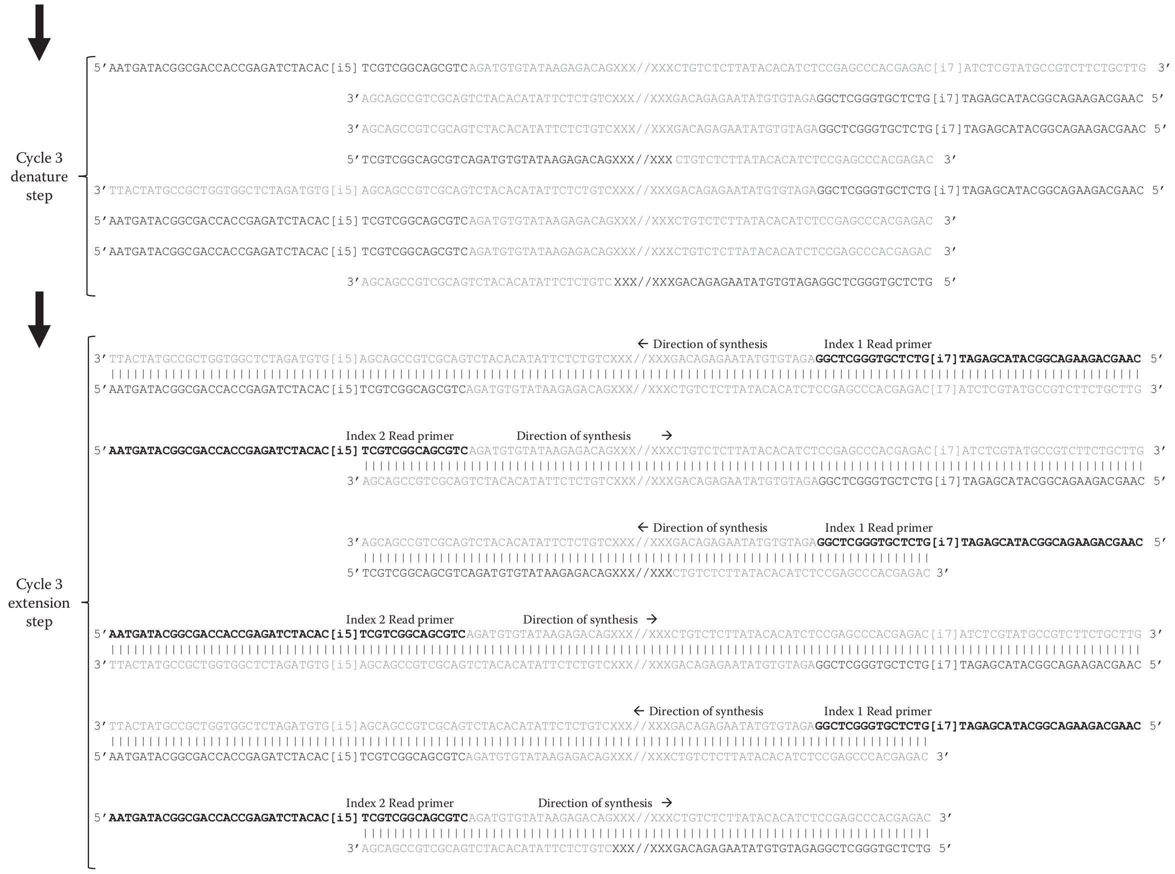

Cycle 3 denature step
5′AATGATACGGCGACCACCGAGATCTACAC[i5] TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
3′AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
3′AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXX CTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
3′TTACTATGCCGCTGGTGGCTCTAGATGTG[i5] AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′AATGATACGGCGACCACCGAGATCTACAC[i5] TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
5′AATGATACGGCGACCACCGAGATCTACAC[i5] TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
3′AGCAGCCGTCGCAGTCTACACATATTCTCTGTC XXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG 5′
Cycle 3 extension step
← Direction of synthesis
Index 1 Read primer
3′TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC[I7]ATCTCGTATGCCGTCTTCTGCTTG 3′
Index 2 Read primer
Direction of synthesis →
5′AATGATACGGCGACCACCGAGATCTACAC[i5] TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
3′AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
← Direction of synthesis
Index 1 Read primer
3′AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
Index 2 Read primer
Direction of synthesis →
5′AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′
3′TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
← Direction of synthesis
Index 1 Read primer
3′TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
5′AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
Index 2 Read primer
Direction of synthesis →
5′AATGATACGGCGACCACCGAGATCTACAC[i5] TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC 3′
3′AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG 5′

Figure 7.25. (Continued)

third cycle is finished (Figure 7.25). A total of five normal PCR cycles are required when using the Illumina Nextera kit. When the program stops, the adapter–fragment constructs are complete as they now contain the required P5 and P7 common adapters, annealing sites for the Read 1, Read 2, and i7 Read Sequencing Primers, and two sample-specific indices (Figure 7.26). Note, the P5 flow cell oligo not only serves to tether an adapter–fragment strand during the start of cluster generation but it also acts as the i5 Index Read Sequencing Primer. The actual sequence of the i7 Index Read Sequencing Primer is proprietary has not been disclosed by Illumina but its approximate location in the construct is shown in Figure 7.26. Illumina has also not disclosed information about the DNA polymerase(s) they include in their Nextera kits for this limited PCR step. However, Picelli et al. (2014) used the KAPA HiFi DNA polymerase (KAPA Biosystems), which synthesizes DNA in a 5′ → 3′ direction and has 3′ → 5′ exonuclease (i.e., proofreading) capability, to successfully amplify tagmentation products. Following the limited cycle PCR, the completed adapter–fragment constructs must be purified. The Illumina Nextera protocol uses SPRI beads to purify the PCR products.

Step 3: Quantify Libraries and Verify Fragment Distributions After PCR purification, the libraries must be quantified and verified (Figure 7.19). As with the other library preparation methods the concentration of each library can be estimated using qPCR (best method) or using a Bioanalyzer. A Bioanalyzer can also be used to verify that each library exhibits a fragment size distribution within the desired range. Libraries made using the Illumina Nextera kit without any additional size-selection steps are expected to display a size range of ~200–1,000 bp.

Step 4: Normalize and Pool Libraries (e.g., 4–10 Libraries per Pool) After libraries have been quantified and passed Bioanalyzer inspections, they can be normalized and grouped together with four to ten libraries per hybridization pool (Figure 7.19). Note, some researchers have decided to not index and pool their Nextera libraries before hybrid selection (e.g., Faircloth et al. 2012), which means that more labor and consumables must be expended in order to generate large multilocus datasets. However, as we will soon see there are advantages and disadvantages to indexing and pooling Nextera libraries prior to hybrid selection.

Although the commercial Nextera kits offer the simplest and most reliable method for constructing Illumina sequencing libraries, their expense may limit the number of libraries a given project or laboratory can produce. To help reduce the cost per library, Picelli et al. (2014) developed a "kit-free" protocol to making Tn5 transposition-mediated libraries for Illumina sequencing. This method requires additional laboratory equipment and technical expertise for molecular cloning and thus it will be more challenging than using kits. Nonetheless, because it is possible to produce vast amounts of the key reagents such as the Tn5 transposases, laboratories can make their own premade Tn5 synaptic complexes ready for producing a large number of sequencing libraries.

In their protocol, Picelli et al. (2014) detail methods for producing in-house Tn5 transposases using *E. coli* expression vectors. Once the transposase enzyme subunits are produced and purified, they can be assembled into functional synaptic complexes (i.e., Tn5 homodimers annealed to two ME-flanked adapter sequences). This approach also gives the researcher the option of making custom adapter sequences (Picelli et al. 2014).

One of the important findings in the study by Picelli et al. (2014) concerns the importance of using the proper type and concentration of a "molecular crowding agent" in the tagmentation reactions. One of the problems encountered during the early years of the transposition-mediated NGS library construction was that the researcher had little control over the fragment sizes found in the resulting libraries (Adey et al. 2010). Indeed, Adey et al. (2010) determined that without any optimized buffers or special size-selection step the observed fragment distribution averaged about 100 bp ± 47 bp. Note that the minimum fragment size from Tn5-mediated transposition is sharply delineated at ~38 bp, a limit that is likely due to steric interactions between competing synaptic complexes attacking the same stretch of target DNA in adjacent locations (Adey et al. 2010). Although these fragment sizes are smaller than the desired 400–500 bp range for Illumina sequencers, Adey et al. (2010) correctly predicted that further optimization of reaction buffers could increase the sizes of the fragments and produce a more desirable size range.

As we discussed in Chapter 6, a solution of PEG can be used in a simple and inexpensive method

```
                                                                                                                            P7 flow cell oligo
                                                                                                                  3′TAGAGCATACGGCAGAAGACGAAC 5′
                                                                                        Read 2 Sequencing Primer
                                                                               3′GACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG 5′

3′  TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]AGCAGCCGTCGCAGTCTACACATATTCTCTGTCXXX//XXXGACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC 5′
    |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
5′  AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXX//XXXCTGTCTCTTATACACATCTCCGAGCCCACGAGAC[i7]ATCTCGTATGCCGTCTTCTGCTTG 3′

                                                                              5′ i7 Index Read Sequencing Primer 3′
                                   5′TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG 3′
                                           Read 1 Sequencing Primer
  5′AATGATACGGCGACCACCGAGAUCTACAC 3′
 P5 flow cell oligo and i5 Index Read Sequencing Primer
```

Figure 7.26. Completed Nextera adapter–fragment construct for Illumina sequencing. The construct contains annealing sites for the P5 and P7 flow cell oligos, Read 1 and 2 Sequencing Primers, and an i7 Read Sequencing Primer (sequence not shown). The 8-bp i5 and 8-bp i7 index sequences are represented by [i5] and [i7], respectively. The middle section of the construct contains the target DNA (a string of Xs). The // signifies that most of the target fragment length is not shown due to space limitations and the vertical dashes indicate hydrogen bonding between complementary bases. (Oligonucleotide sequences © 2007–2010 Illumina, Inc. All rights reserved.)

for purifying and size-selecting DNA fragments such as PCR products and sequencing library fragments. Picelli et al. (2014) experimented with various types and concentrations of PEG in tagmentation reactions. Importantly, they showed how PEG in particular can modulate fragment size in libraries depending on the molecular weight of each PEG polymer type and its concentration. Their results showed that the Illumina Nextera XT kit produced the best library results as the fragment distribution ranged from 200 to 1,000 bp and was centered over the desired sizes of 400–500 bp (Figure 7.27). Interestingly, the authors nearly duplicated these results by using their in-house Tn5 transposase enzymes and a buffer containing a 5% solution of PEG 8000 (Figure 7.27). Other concentrations and crowding agents not only yielded inferior library results, but also the exclusion of any crowding agent caused a complete failure (Figure 7.27). These results are important because achieving the optimal size range of adapter-ligated fragments in a tagmentation reaction will lead to more efficient library construction, even with picogram-scale input DNA (Picelli et al. 2014).

As pointed out earlier, Nextera kits now use SPRI beads to clean tagmentation products and thus this method may also perform well with the Picelli et al. protocol. An alternative way to accomplish the task of stripping the transposases away from the tagmentation products was discovered by Picelli et al. (2014). These authors found that a 0.1% SDS "stripping buffer" was effective at purifying tagmentation products as evidenced by their good quality PCR results. Further experimentation by these authors revealed that a buffer with a concentration of 0.2% SDS gave the highest PCR yields while a 0.3% SDS concentration caused a complete PCR failure (Picelli et al. 2014). Besides its simplicity, the stripping buffer approach has the added advantage of being well suited for a high-throughput sequencing library workflow (Picelli et al. 2014). It remains to be seen which of the two methods—SPRI beads or the stripping buffer—should be the preferred approach to cleaning tagmentation products.

7.2.2 In-Solution Hybrid Selection

The attainment of whole-genome or "shotgun" sequencing libraries is a major achievement toward the goal of acquiring large multilocus datasets for phylogenomic studies. However, before these libraries can be sequenced they must be reduced down to the subset of library fragments representing the set of target loci (e.g., exons, anchored loci, etc.). Once these target DNA fragments have been isolated they can be prepared for sequencing on an Illumina platform.

Gnirke et al. (2009) developed a method called *in-solution hybrid selection*, which has since become the standard approach for obtaining phylogenomic datasets consisting of hundreds to thousands of loci for dozens to hundreds of individuals. In-solution hybrid selection typically employs a complex mixture of biotinylated RNA probes, which are 60–120 bp long

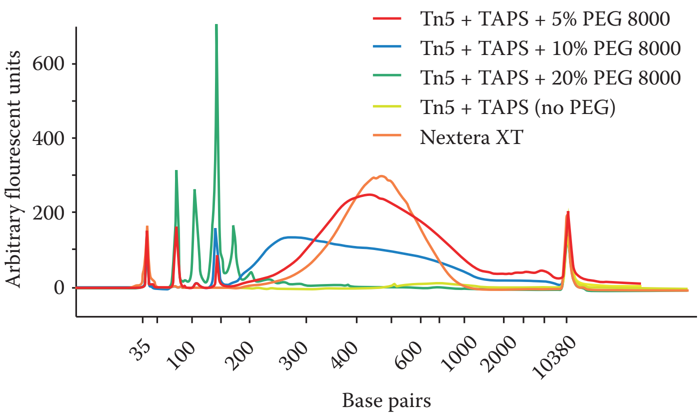


Figure 7.27. Effects of different crowding agents on tagmentation reactions. The Bioanalyzer electropherogram shows the results of tagmentation reactions involving different concentrations of PEG 8000, no PEG, or using an Illumina Nextera XT kit. (Reprinted from Picelli, S. et al. 2014. *Genome Res* 24:2033–2040. With permission.)

(Lemmon and Lemmon 2013) and are complementary to a set of target genomic loci. During an in-solution hybridization reaction, which is performed under stringent thermal conditions, RNA probes hybridize with complementary single-stranded (target) library molecules. The biotin moieties, which are incorporated into each probe at the time of oligonucleotide synthesis via transcription, play a key role together with streptavidin-coated magnetic beads in capturing RNA–DNA hybrids while in solution. Following a "pull down" step in which a magnetic separator is used to immobilize the RNA–DNA hybrids, the nontarget DNA fragments are washed away, which leaves behind target DNA for sequencing. The in-solution hybrid selection method has been analogized to the practice of fishing whereby RNA "baits" are used to catch target DNA fragments (i.e., "the catch") from a "pond" of whole-genome library fragments (Gnirke et al. 2009; Lemmon and Lemmon 2013).

Although there are many similarities between in-solution and "on-array" (i.e., microarray) approaches, one big difference between the two is that the former uses a vast excess of oligonucleotide probes relative to library templates, while the latter employs a reverse strategy (Gnirke et al. 2009; Mamanova et al. 2010). The high probe-to-template ratios that characterize in-solution hybridization reactions has several advantages over the on-array approach including: (1) improved probe–template hybridization efficiency; (2) much less library template is required for hybrid selection reactions; and (3) hybridization reactions can be conducted in 96-well microplates on thermocyclers and do not require specialized equipment (Gnirke et al. 2009; Mamanova et al. 2010). Thus, library inputs for in-solution hybrid selection experiments only need to be in the 50–500 ng range rather than necessitating microgram quantities of library DNA.

Figure 7.28 illustrates the basic workflow of an in-solution hybrid selection experiment, which requires at least 2 days to complete (Blumenstiel et al. 2010). Note that Figure 7.28 is simplistic in that it only illustrates the major steps involved in hybrid selection. Other crucial steps such as DNA purification steps (e.g., using SPRI beads) immediately follow some of the steps shown in Figure 7.28 and thus it is critically important to follow actual hybrid selection protocols unless certain modifications are clearly warranted and are well understood by the researcher. Although hybrid selection experiments are now routinely performed using commercially available kits, additional helpful references containing protocols, information about key step-modifications, and troubleshooting tips can be found (e.g., Blumenstiel et al. 2010; Mamanova et al. 2010; Fisher et al. 2011; Rohland and Reich 2012). We will now review the major steps in the in-solution hybrid selection workflow.

Step 1: Acquire set of Biotinylated RNA "Baits" The initial step of an in-solution hybrid selection experiment is to obtain a set of biotinylated RNA "baits" (Figure 7.28). These probe sets are acquired in two different ways. First, many phylogenomic studies especially those involving vertebrates can likely use an existing probe set. For example, a number of probe sets for phylogenomic studies have been designed for anchored loci such as UCE-anchored loci (http://ultraconserved.org) and AE loci (http://anchoredphylogeny.com). The total number of different genomic loci targeted by these probe sets varies but at this time it ranges from ~500 (Lemmon et al. 2012) to more than 5,000 loci (McCormack et al. 2015). Thus, if a probe set that is suitable for a planned study already exists, then an RNA bait kit can simply be custom-ordered from a company that produces these kits. The SureSelect Kit (Agilent) and MYbaits kit (MYcroarray, Inc.) represent two of the more popular RNA bait kits at the present time. The alternative way to acquire a probe set is to design it yourself and then order a custom-made kit from one of the aforementioned suppliers. In principle, a set of RNA baits can target any portion of a genome (e.g., anonymous loci) and thus a researcher is free to design a probe set comprised of any aggregate of genome-wide loci. In Chapter 8, we will visit the issue of how these probe sets are designed.

Step 2: Heat-Denature the "Pond" and "Blocking" DNA Each pool of indexed libraries made during the library preparation workflow is used as input whole-genome library DNA (i.e., "pond DNA") in a hybridization reaction. However, before the RNA baits can be hybridized to the target DNA, the pond DNA and special DNA additives called "blocking DNA" must first be heat denatured (in the absence of the RNA baits) in a thermocycler, then allowed to cool down to the hybridization temperature (Figure 7.28). This is a necessary procedure because all blocking DNAs must be

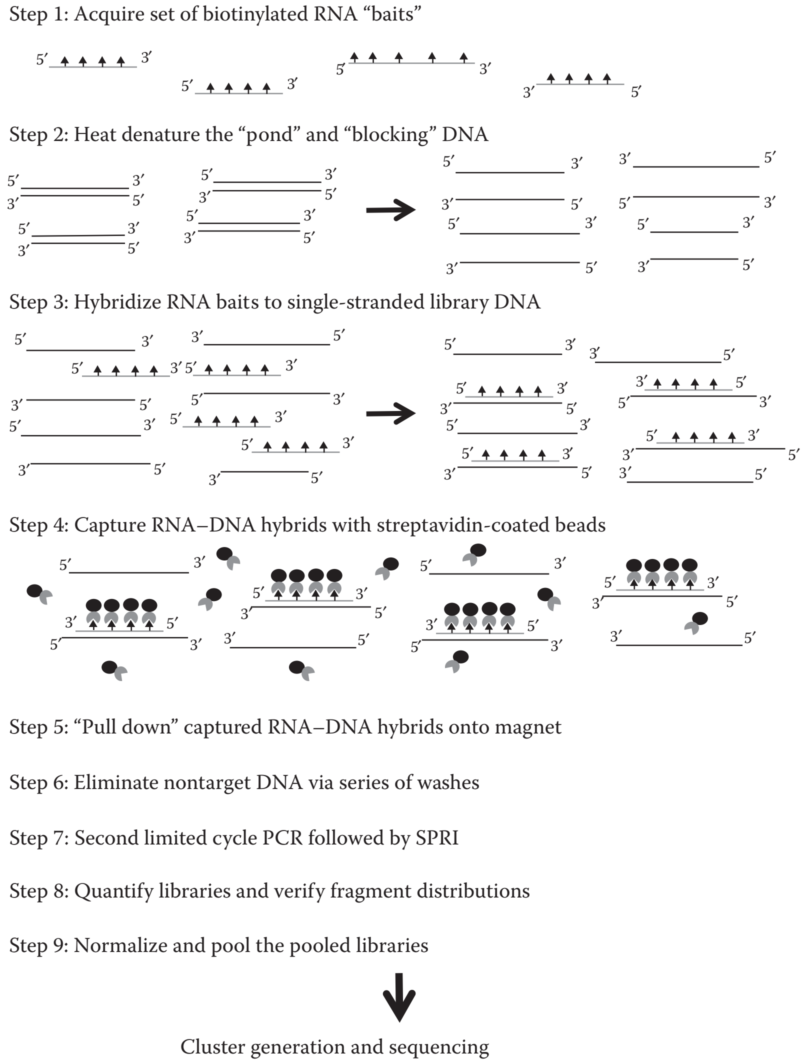

Step 1: Acquire set of biotinylated RNA "baits"
Step 2: Heat denature the "pond" and "blocking" DNA
Step 3: Hybridize RNA baits to single-stranded library DNA
Step 4: Capture RNA–DNA hybrids with streptavidin-coated beads
Step 5: "Pull down" captured RNA–DNA hybrids onto magnet
Step 6: Eliminate nontarget DNA via series of washes
Step 7: Second limited cycle PCR followed by SPRI
Step 8: Quantify libraries and verify fragment distributions
Step 9: Normalize and pool the pooled libraries
Cluster generation and sequencing

Figure 7.28. Basic workflow for in-solution hybrid selection. See main text for descriptions of each step.

single-stranded and free of secondary structures in order to function properly before the RNA baits are placed in the pond DNA mixture to initiate the hybridization reaction. The rationale and use of blocking DNA is discussed below. Thus, the first thermocycler step is a "denaturation step" of 95–98°C for 5 minutes followed by a "hybridization step" of 65°C for about 24 hours.

Pond DNA is a heterogeneous mixture of genomic elements some of which are the target DNA molecules. Because the hybridization reaction is performed at a high level of stringency for DNA duplex formation (i.e., at high temperatures), in principle RNA baits should only anneal to their complementary library DNA strands suspended in the solution. These target DNA molecules can then be harvested and sequenced. However, even at these high temperatures, other unwanted molecular hybrids between "off-target" and "on-target" library strands can also form. If off-target library strands are harvested along with the target library strands, then the output sequence data will consist of many "by-catch" (to continue the fishing analogy) sequences that do not map to any of the desired loci.

Hodges et al. (2009) described two scenarios whereby off-target library strands can be captured during a hybridization reaction. In the first, if a library DNA strand contains a stretch of sites that are complementary to an RNA bait *and* an adjacent

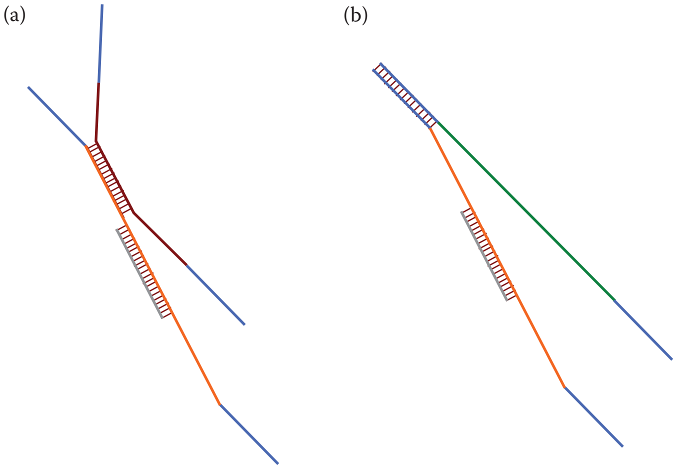


Figure 7.29. Two "by-catch" scenarios whereby nontarget library fragments may be captured during an in-solution hybridization reaction. (a) In one scenario, nonspecific hybridization can occur between two different library fragments (orange and black) when stretches of repetitive DNA are found in both fragments. (b) In the other scenario, complementary adapter sequences (blue) can anneal to each other regardless of their respective fragment DNA sequences (orange and green). In both scenarios, if the target DNA fragment is also hybridized to an RNA probe (gray), then both target and nontarget fragments will be captured and sequenced. (Modified after Figure 2 of Hodges, E. et al. 2009. *Nat Protoc* 4:960–974. With permission.)

stretch of bases consisting of repetitive DNA, then a target strand can simultaneously anneal to its correct RNA bait and to the repetitive DNA found in an off-target library strand (Figure 7.29a). In this scenario, both the target and off-target library strands will be captured and likely sequenced. A second type of unfavorable intermolecular interaction that can occur during a hybridization reaction consists of having the adapter sequence on a target DNA strand anneal to the complementary adapter sequence found on a nontarget DNA strand (Figure 7.29b). Again, if a RNA bait is annealed to a target strand whose adapter sequence is duplexed with the adapter sequence found on an off-target strand, then both target and off-target strands will be captured and probably sequenced. This phenomenon whereby off-target library strands are captured and sequenced because they had annealed to adapters on target library strands during the hybridization reaction is called "daisy chaining" (Mamanova et al. 2010; see Figure 2 in Rohland and Reich 2012). As we will see below, daisy chaining can drastically reduce the efficiency of a hybridization reaction.

There are two main strategies aimed at minimizing the by-catch problem in hybridization reactions. One strategy is to use a "blocking DNA mix" in the hybridization reaction (Hodges et al. 2009) while the other is to construct libraries with shorter adapter sequences before hybrid selection (Faircloth et al. 2012; Rohland and Reich 2012). We will first consider the use of blocking DNA additives and later in this chapter we will discuss the effects of adapter length on in-solution hybridization efficiency.

"Blocking DNA mix" consists of different types of DNA that are used to competitively anneal to stretches of repetitive DNA and adapters on pond DNA strands in order to block these sites from catching unwanted library strands. A type of DNA called Cot-1 DNA, which is a repeat-rich DNA that has been selected for its high affinity for repetitive DNA, is used to neutralize stretches of repetitive DNA found in pond library strands (Hodges et al. 2009). Although it was originally thought that Cot-1 DNA should be obtained from the same (or similar) species as species represented by the libraries (Hodges et al. 2009), in practice it may not matter which Cot-1 DNA is used as blocking DNA because repetitive DNA can evidently anneal to a variety of nonhomologous sequences. For example, Faircloth et al. (2015) found that hybrid selection of hymenopteran libraries was more efficient when chicken Cot-1 DNA was used as the repetitive DNA blocker compared to actual hymenopteran Cot-1 DNA. Thus, human or chicken Cot-1 DNA has been used as the repetitive DNA blocker in many vertebrate phylogenomic studies. Salmon sperm is another additive that is sometime used to help minimize unfavorable molecular interactions during the hybridization reaction. Another strategy that has worked well is to mix different Cot-1 DNAs. For example, good hybridization results were obtained for several reptile and amphibian species by using a mix consisting of 1/3 human Cot-1 DNA, 1/3 chicken Cot-1, and 1/3 salmon sperm Cot-1 (S. B. Reilly, personal communication, 2016). Similarly Portik et al. (2016) used a mix comprised of 1/3 human Cot-1, 1/3 mouse Cot-1, and 1/3 chicken Cot-1 DNA in their hybrid selection of pooled frog libraries. A third and critical component to the blocking mix is comprised of forward and reverse complements to the adapter sequences. It is essential to match the adapter blocking oligos as closely as possible with the adapters used to make the library constructs. It is therefore common practice to replace the kit-supplied adapter blocking mix with a custom-made adapter blocking mix that better matches the adapters used in the libraries. For example, Lemmon et al. (2012) replaced the adapter blocking mix contained in the SureSelect kit (i.e., Block #3) with a custom-made indexed blocking mix (see Geller 2011 for oligo sequences and recipe) appropriate for indexed Illumina libraries. Custom-designed adapter blocking oligos can also be found in Meyer and Kircher (2010) and Rohland and Reich (2012). Faircloth et al. (2015) designed blocking oligos for their custom-made adapters, each of which included a string of 10 inosines to block the 10-bp index sequences. Similarly, Leaché et al. (2015) used a custom blocking mix matched to the Illumina Truseq Nano adapters, which also used inosines in place of the index bases. The recent study by Portik et al. (2016) evaluated the performance of three different types of adapter blocking oligos on hybridization results. These included the adapter blocking oligos included in the MYbaits kit, short blocking oligos that do not protect the index bases, and the *xGen* blocking oligos made by IDT. Based on their results, the authors concluded that the *xGen* blocking oligos performed much better than the other oligos included in their test. Accordingly, given that hybridization reactions are typically carried out with commercial baits kits, careful consideration should be given to which adapter blocking mix should be used.

How much pond DNA is input into a hybrid selection experiment? Typically, the total amount of input pond DNA used for the hybrid selection procedure usually ranges between 100 and 500 ng (e.g., Blumenstiel et al. 2010) and protocols usually presume that the pond DNA itself is at a concentration of 100 ng/μL. The original in-solution hybrid selection study of Gnirke et al. (2009) used 500 ng of pond DNA together with 500 ng of RNA baits. However, these authors noted that they obtained comparable results regardless of whether their inputs were 100 ng or 500 ng each of pond and baits (see also Faircloth et al. 2015). Given the high cost of baits and occasions when pond concentrations are low, it may be desirable to use the minimum amount of baits allowed by the protocol (e.g., 100 ng). However, for organisms with very large genomes (e.g., amphibians), it is preferable to use at least 250 ng of each library in a hybridization reaction; thus if four frog libraries are pooled together, then 1 μg total of library DNA would comprise a hybridization pool (S. B. Reilly, personal communication, 2016).

Step 3: Hybridize RNA Baits to Single-Stranded Library DNA After the pond DNA and blocking DNAs have been denatured, the thermocycler lowers the temperature down to the standard hybridization temperature of 65°C. For the next 16–48 hours the hybridization reaction will occur at this temperature, which is high enough to ensure that RNA baits largely only anneal to complementary stretches of DNA within the library DNA strands (Figure 7.28). The process of setting up hybridization reactions should be done exactly according to the baits protocol in order to minimize the occurrence of by-catch and nonspecific RNA–DNA hybrids during the reaction. Note, to achieve the best possible capture results, it is critically important to also do the following: (1) carefully mix, by pipetting, your library pool and hybridization mix and avoid creating bubbles or splashing liquid on the sides of the reaction tubes and (2) be sure to heat the thermocycler lid during the hybridization reaction (S. B. Reilly, personal communication, 2016).

To complete the reaction setup, the following procedures must be performed in the correct sequence. When the thermocycler's heat block arrives at 65°C, the machine must be temporarily paused. This is to allow some time for the Cot-1 DNA and adapter blocking oligos to hybridize with repetitive DNA and adapters on the library DNA strands, respectively, before the

RNA baits are mixed with the pond DNA. At this time, the hybridization buffer mix (i.e., hybridization buffer, RNA baits, and RNase block) must be prewarmed at 65°C thus they are placed into the thermocycler's heat block near to the tubes containing the libraries. After the preheating period is completed, the hybridization buffer mix is transferred into the pond DNA tubes while all tubes are still sitting in the heat block. For safety reasons it should be kept in mind that the thermocycler's heated block is hot enough to cause severe burns and thus extreme caution should be exercised. After all reagents have been mixed together the lid of the thermocycler is closed and program is resumed to allow the hybridization reaction to proceed.

Step 4: Capture RNA–DNA Hybrids with Streptavidin-Coated Beads When the hybridization reaction is completed the next step is to "capture" each of the RNA–DNA hybrids while they are still in solution. This is accomplished by adding streptavidin-coated paramagnetic beads to the hybridization reactions. The streptavidin-coated beads are able to capture the RNA–DNA hybrids because streptavidin (a bacterial protein) has an exceptionally high affinity for the biotin moieties (a form of vitamin B) that are incorporated into the RNA baits (Figure 7.28). The streptavidin–biotin interaction is evidently one of the strongest noncovalent biological interactions (Holmberg et al. 2005).

Step 5: "Pull Down" Captured RNA–DNA Hybrids onto Magnet Next, a magnetic separator is used to "pull down" the beads attached to the RNA–DNA hybrids, which immobilizes them against the tube wells near the magnet. The supernatant contains nonhybridized DNA and is therefore removed and discarded (Figure 7.28).

Step 6: Eliminate Nontarget DNA via Series of Washes Although a hybridization reaction may produce a sufficient number of the target RNA–DNA hybrids, these reactions may also generate a number of hybrids involving nontarget DNA. If nontarget library fragments having both the P5 and P7 common adapters are not somehow eliminated, then they will populate the flow cell along with the target DNA resulting in lower quality sequence data. Fortunately, these nonspecific hybrids are weakly bound together owing to their nonhomologous sequence matches and thus they can be disassociated via a series of stringency washes (Figure 7.28).

The first wash step is a "low stringency" wash in which a wash buffer consisting of nuclease-free water, saline sodium citrate (SSC), and sodium dodecyl sulfate (SDS) is incubated with the hybridization reactions at room temperature. This first wash buffer helps to strip away nontarget DNAs that are weakly bound to RNA baits or are sticking to the inside of the tube. At the end of the brief incubation period the supernatant containing these unwanted DNAs is removed using a pipette and discarded. The second wash is a "high stringency" wash, which is needed in order to dissociate the more stable nonspecific RNA–DNA hybrids. This wash is more stringent than the first for two reasons. First, the second wash buffer contains a 10-fold lower concentration of SSC (salt buffer) than is found in the first wash buffer. Secondly, the second wash is conducted at a much higher temperature (65°C) than was used during the first wash (room temperature). Following this second wash step, the collection of captured DNA fragments, which at this time are still bound to the biotinylated RNA baits on the magnet, must be amplified in a second limited cycle PCR.

Step 7: Second Limited Cycle PCR Followed by SPRI Following the hybrid selection and wash steps, a second (or "post hybridization") limited cycle PCR must be performed (Figure 7.28). If the captured library constructs are already sequencing-ready (i.e., are indexed and contain P5 and P7 common adapters), then this PCR is merely used to enrich the library with the captured library strands so that a sufficient number of sequencing templates can be added to the flow cell. However, if the captured constructs do not contain all the required elements for sequencing, then this PCR is used to (1) complete the constructs using 5′ tailed PCR primers in order to add one or two indices as well as the P5 and P7 common adapters and (2) amplify these newly completed constructs.

Before these PCRs can be set up and run, captured DNA templates must be prepared. The original method for accomplishing this involved using a solution of sodium hydroxide (NaOH) to chemically break the hydrogen bonds linking together the RNA and DNA strands (e.g., Blumenstiel et al. 2010). Thus, the tubes (or plate) containing the hybridization reactions are removed from the magnet before a solution of NaOH is added to the reactions. Next, a neutralization buffer consisting of Tris–HCL is added, and, after a brief incubation period, the magnetic separator is used to immobilize the bead–RNA bait complexes. Supernatants containing the captured ssDNAs are then transferred to clean tubes (or plate) where they are cleaned using SPRI beads followed by elution with nuclease-free water (Blumenstiel et al. 2010). The captured DNA samples are then ready to serve as PCR templates.

Fisher et al. (2011) developed a far simpler alternative method for preparing the captured DNA. Instead of employing the standard chemical-denaturing and SPRI bead cleanup approach, the RNA–DNA hybrids are simply allowed to air dry while they are on the magnet. This drying procedure is initiated immediately after the removal and disposal of the second wash buffer (i.e., high stringency buffer) while the tubes (or plate) are on magnet. Next, the magnet is removed and nuclease-free water is pipetted into the same hybridization tubes (or plate) in order to resuspend the dried RNA–DNA hybrids. The other PCR reagents are then added before the samples are placed into a thermocycler to conduct the PCR. Fisher et al. (2011) termed this procedure *off-bead PCR* because target DNA strands are released from the streptavidin bead-RNA bait complexes during the first heat-denaturation step of the PCR. After the off-bead PCR is completed, the magnet is used to pull down and immobilize the streptavidin beads while the supernatant containing the PCR products is saved to other clean tubes (or plate). The PCR products are cleaned using a standard SPRI bead cleanup.

The off-bead PCR method has significant advantages over the NaOH-denaturing approach. Besides the simplicity and automation friendly aspects of the method, Fisher et al. (2011) observed an approximately 3-fold increase in captured product yield following off-bead PCR compared to the NaOH-based approach. This method represents a major improvement in the hybrid selection methodology and thus it has been used in some phylogenomic studies (e.g., Faircloth et al. 2015) and is now implemented in some RNA bait kits (e.g., MYbaits and SureSelect kits).

This second limited cycle PCR typically consists of 15–20 normal cycles. Like the first limited cycle PCR, only PCR proofreading (high fidelity) DNA polymerases should be used. One interesting difference between this PCR and other PCRs already discussed is that the templates are single-stranded at the start because the RNA baits only capture one library strand at a time. However, this does not present a problem because one of the two PCR primers can readily synthesize complete complementary strands during the first cycle. Thus, starting in the second cycle there will be two complementary strands that can be copied because both will have annealing sites for each primer. Thenceforth, target templates can be copied at an exponential rate over the course of the remaining PCR cycles.

As you will recall, the library constructs generated by the Rohland and Reich (2012) approach contained a 6-bp barcode sequence, which allows for the pooling of multiple libraries prior to hybridization. Because these library constructs have truncated adapters and thus are not yet in complete form, 5′ tailed primers must be used in the second limited cycle PCR in order to add a 7-bp index sequence to the P7 side of the construct as well as the P5 and P7 common adapters. Figure 7.30 shows the first three cycles of a second limited cycle PCR involving a Rohland and Reich truncated-adapter construct. As the reaction mixture is heated toward the first denaturation temperature, the RNA baits become dissociated from the library DNA strands and thus the first extension step will effectively only involve one of the two complementary strands. However, at the conclusion of the first cycle, library strands containing the binding sites for both PCR primers will be ready to participate in DNA synthesis starting in the second cycle (Figure 7.30). By the end of third cycle, fully formed constructs start to appear among the reaction products (Figure 7.30). When the PCR is finished, the products will be largely comprised of fully-formed adapter constructs derived from the original captured library DNAs (Figure 7.31). Notice that the PE Read 1 Sequencing Primer is used to sequence the 6-bp barcode *and* one end of the target fragment (Figure 7.31). Thus, the first six bases to be sequenced using the PE Read 1 Sequencing Primer will actually be the barcode followed by the target fragment. The 7-bp index is, however, sequenced using its own sequencing primer (i.e., Index PE Sequencing Primer; Figure 7.31).

Step 8: Verify and Quantify Libraries Following the second limited cycle PCR, each posthybridization library pool must be verified and quantified. Verification of libraries is generally done two ways. First, a Bioanalyzer is used to confirm the size range distribution for each pool. The electropherogram should show a distribution of

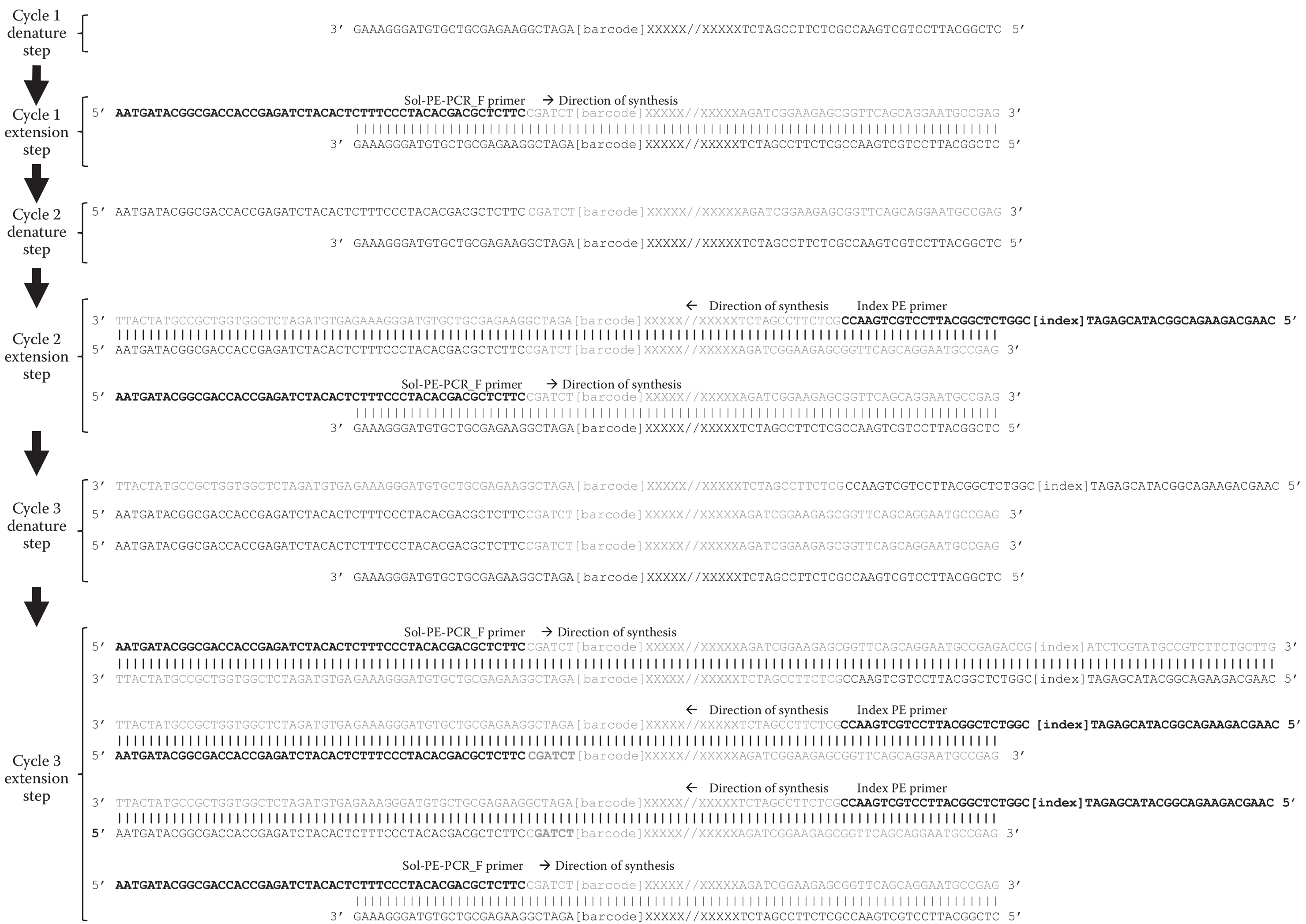

Cycle 1 denature step
3′ GAAAGGGATGTGCTGCGAGAAGGCTAGA[barcode]XXXXX//XXXXXTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTC 5′
Cycle 1 extension step
Sol-PE-PCR_F primer
→ Direction of synthesis
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT[barcode]XXXXX//XXXXXAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG 3′
3′ GAAAGGGATGTGCTGCGAGAAGGCTAGA[barcode]XXXXX//XXXXXTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTC 5′
Cycle 2 denature step
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC CGATCT[barcode]XXXXX//XXXXXAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG 3′
3′ GAAAGGGATGTGCTGCGAGAAGGCTAGA[barcode]XXXXX//XXXXXTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTC 5′
Cycle 2 extension step
← Direction of synthesis
Index PE primer
3′ TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA[barcode]XXXXX//XXXXXTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTCTGGC[index]TAGAGCATACGGCAGAAGACGAAC 5′
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT[barcode]XXXXX//XXXXXAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG 3′
Sol-PE-PCR_F primer
→ Direction of synthesis
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT[barcode]XXXXX//XXXXXAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG 3′
3′ GAAAGGGATGTGCTGCGAGAAGGCTAGA[barcode]XXXXX//XXXXXTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTC 5′
Cycle 3 denature step
3′ TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA[barcode]XXXXX//XXXXXTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTCTGGC[index]TAGAGCATACGGCAGAAGACGAAC 5′
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT[barcode]XXXXX//XXXXXAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG 3′
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT[barcode]XXXXX//XXXXXAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG 3′
3′ GAAAGGGATGTGCTGCGAGAAGGCTAGA[barcode]XXXXX//XXXXXTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTC 5′
Cycle 3 extension step
Sol-PE-PCR_F primer
→ Direction of synthesis
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT[barcode]XXXXX//XXXXXAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG[index]ATCTCGTATGCCGTCTTCTGCTTG 3′
3′ TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA[barcode]XXXXX//XXXXXTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTCTGGC[index]TAGAGCATACGGCAGAAGACGAAC 5′
← Direction of synthesis
Index PE primer
3′ TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA[barcode]XXXXX//XXXXXTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTCTGGC [index]TAGAGCATACGGCAGAAGACGAAC 5′
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC CGATCT[barcode]XXXXX//XXXXXAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG 3′
← Direction of synthesis
Index PE primer
3′ TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA[barcode]XXXXX//XXXXXTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTCTGGC[index]TAGAGCATACGGCAGAAGACGAAC 5′
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT[barcode]XXXXX//XXXXXAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG 3′
Sol-PE-PCR_F primer
→ Direction of synthesis
5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT[barcode]XXXXX//XXXXXAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG 3′
3′ GAAAGGGATGTGCTGCGAGAAGGCTAGA[barcode]XXXXX//XXXXXTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTC 5′

Figure 7.30. How the second limited cycle PCR completes adapter–fragment library constructs using the Rohland and Reich approach. The denaturation and primer-extension steps in the first three cycles of a limited PCR are illustrated (primer-annealing steps not shown due to space limitation). Because this PCR is begun with a single adapter–fragment strand that was captured in the hybridization reaction, only one of the two primers can bind to the template in the first cycle. However, in the second cycle binding sites are available for both PCR primers. Completed adapter–fragment constructs appear only at the end of the third cycle. The 7-bp index sequence is represented by (index). The middle section of the adapter–fragment constructs contains the target DNA (a string of Xs). The // signifies that most of the target fragment is omitted due to space limitation and the vertical dashes indicate hydrogen bonding between complementary bases. Boldface sequences represent newly annealed primer in the same cycle and gray letters represent DNA synthesized in this reaction.

fragments with the peak centered between 250 and 500 bp depending on initial size distribution generated at the start of the library making process and taking into account adapter lengths. The other critical verification assay that should be performed is to evaluate hybrid selection performance via qPCR; in other words, verify that target molecules were selectively captured and nontarget molecules depleted following each hybridization reaction (Hodges et al. 2009; Bi et al. 2012; Peñalba et al. 2014; Faircloth et al. 2015).

In this "relative qPCR" procedure (Faircloth et al. 2015) a small subset of target and nontarget loci from random genomic locations are subjected to qPCR in order to compare the amounts of specific target and nontarget templates present in pre- versus posthybridization pools. To perform this procedure, custom locus-specific PCR primers for 2–3 target loci and 2–3 nontarget loci are first designed (Chapter 8 discusses primer design) to amplify a 100–150 bp section of each target and nontarget locus (Hodges et al. 2009; Peñalba et al. 2014; S. B. Reilly, personal communication, 2016). These primers should be 20 bp long, have melting temperatures (Tm) between 58 and 60°C, and must be tested beforehand using regular PCR procedures to ensure that they robustly amplify the correct-sized products and do not produce other artifacts such as primer dimers (Hodges et al. 2009; Peñalba et al. 2014). Next, each test locus is qPCR-amplified in duplicate or triplicate with templates supplied from pre- and posthybridization library pools (see Hodges et al. 2009 and Peñalba et al. 2014 for protocols). Thus, if duplicate reactions are set up, then each locus will be amplified in two replicates using prehybridization pool template and in two replicate reactions using posthybridization template. After qPCR is completed, within-locus C_T values output from qPCR analyses are compared to each other. The C_T value (cycle threshold) is defined as the cycle number in which the detected amount of fluorescently labeled products is first distinguishable from a horizontal threshold value (i.e., background fluorescence "noise"). The C_T value is useful because it is correlated with the amount of starting template. Thus, if hybrid selection was successful, then C_T values for posthybridization library templates should be shifted 6–10 cycles to the left of C_T values for prehybridization library templates; in contrast, C_T values for nontarget loci should show the opposite with posthybridization C_T values shifted 4–10 cycles to the right of prehybridization C_T values (Hodges et al. 2009; Peñalba et al. 2014; S. B. Reilly, personal communication, 2016). We can easily visualize differences between C_T values by examining plots of qPCR amplification curves for each locus (Figure 7.32). Notice that the C_T values in the top plot of Figure 7.32 are located near the 20 and 30 cycle marks, while the values in the bottom plot are approximately at 18 and 28 cycles. Because the within-locus results show that posthybridization C_T values are shifted to the left of prehybridization C_T values by at least five cycles, this hybrid selection reaction can be considered successful (Hodges et al. 2009; Peñalba et al. 2014; S. B. Reilly, personal communication, 2016). Though not shown here, amplification curves for nontarget loci should show evidence that they were depleted in posthybridization pools.

Once each posthybridization pool has passed the relative qPCR test, they must be quantified. Although a Bioanalyzer or Qubit fluorometer can be used to quantify each library pool, the most accurate way to quantify library pools is to use qPCR. However, unlike the relative qPCR we just discussed, an Illumina qPCR quantification kit should be used, which includes primers that anneal to the adapters.

Step 9: Normalize and Pool the Pooled Libraries After quantification, library pools are ready to be pooled together in equimolar ratios (i.e., "pooling of pools"). The number of final pools will depend on the number of Illumina lanes that will be used

```
                                                                                                                                      P7 flow cell oligo
                                                                                                                              3'TAGAGCATACGGCAGAAGACGAAC 5'
                                                                                        PE read 2 Sequencing Primer
                                                                                3' TCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTCTGGC 5'

3' TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA [barcode]XXXXX//XXXXXTCTAGCCTTCTCGCCAAGTCGTCCTTACGGCTCTGGC [index]TAGAGCATACGGCAGAAGACGAAC 5'
   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||||||||||||||||||
5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT[barcode] XXXXX//XXXXXAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG [index]ATCTCGTATGCCGTCTTCTGCTTG 3'

                                                                                 5' GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG 3'
                                                                                          Index PE Sequencing Primer
                          5'ACACTCTTTCCCTACACGACGCTCTTCCGATCT 3'
                                  PE Read 1 Sequencing Primer
5' AATGATACGGCGACCACCGAGAUCTACAC 3'
           P5 flow cell oligo
```

Figure 7.31. Completed Rohland and Reich adapter–fragment construct for Illumina sequencing. The construct contains annealing sites for the P5 and P7 flow cell oligos, PE Read 1 and 2 Sequencing Primers, and the Index PE Sequencing Primer. Notice that the 6-bp barcode is sequenced together with the Read 1 sequence. The 7-bp index sequence is represented by (index). The middle section of the adapter–fragment construct contains the target DNA (a string of Xs). The // signifies that most of the target fragment is omitted due to space limitation and the vertical dashes indicate hydrogen bonding between complementary bases. (Adapter sequences and primers were obtained from Rohland, N. and D. Reich. 2012. *Genome Res* 22:939–946, and are also found in Table 7.3.)

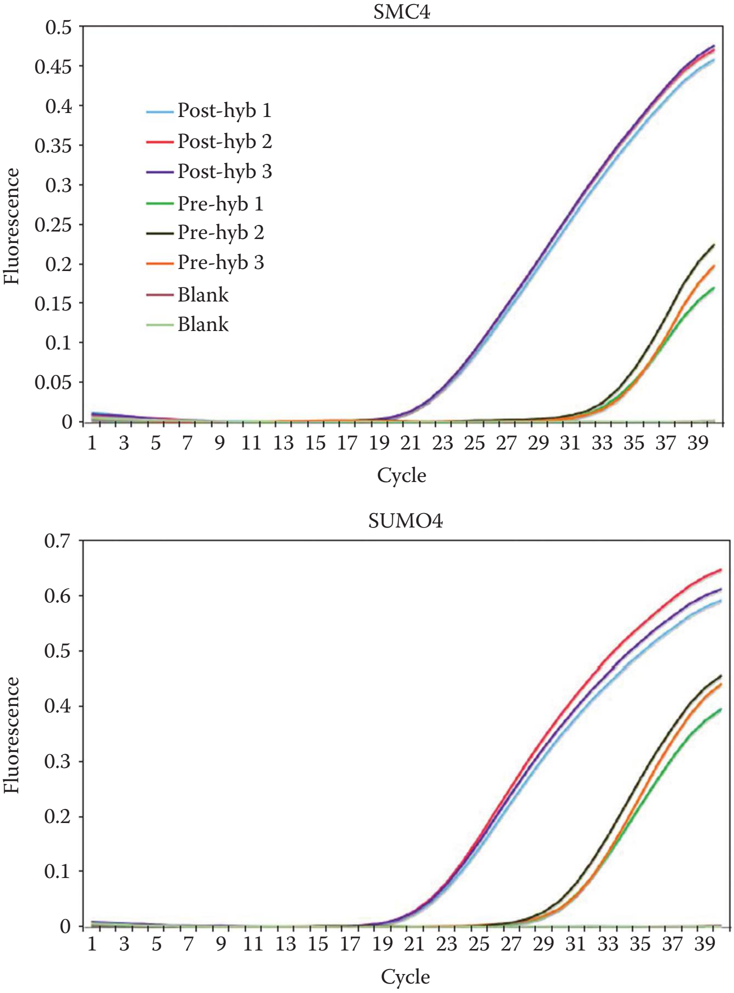


Figure 7.32. qPCR analyses showing levels of hybrid selection efficiency for two targeted loci. The top amplification plot shows qPCR results for the first target locus, which was amplified using the pre- and posthybridization pools as templates (curves pre-hyb 1–3 and post-hyb 1–3, respectively). Three replicates each of pre- and posthybridization qPCRs and two replicates of negative controls (blanks) were performed. The bottom amplification plot shows the results for the other targeted locus. The ~6–10 cycle shifts of the post-hyb curves to the left of the pre-hyb curves indicates that the hybrid selection reactions were successful. The x-axis shows the qPCR cycle number and y-axis corresponds to the accumulation of products (amount of detected fluorescence). This procedure can also be performed using 2–3 nontarget loci in order to verify their depletion following the hybridization reactions (results not shown). (Reprinted from Hodges, E. et al. 2009. *Nat Protoc* 4:960–974. With permission.)

though for phylogenomic studies this will usually mean that a single pool will be needed.

7.2.3 Indexing, Pooling, and Hybrid Selection Efficiency Revisited

Because phylogenomic studies often include genetic samples from dozens to hundreds of individuals, performing pooled hybridization reactions can greatly economize RNA baits (which are very expensive) as well as save labor in the lab. Thus, libraries must be indexed prior to the hybridization step if they are to be grouped together into hybridization pools. Although most of the steps in the library-making and in-solution workflows are comparable across studies, one aspect of this work that does vary concerns the indexing and pooling steps. Table 7.5 shows eight different indexing and pooling schemes that could be done using the methods described in this chapter. Schemes

TABLE 7.5
Eight possible indexing and pooling schemes for the in-solution hybrid selection workflow

Scheme	Adapter ligation	1st PCR	Pooling?	Hybrid capture	2nd PCR	Pooling?	Sequencing	References
A	Indexing		Yes	Capture each pool		No		x
B	Indexing		Yes	Capture each pool		Yes		1
C	Indexing		No	Capture each library		Yes		x
D		Indexing	Yes	Capture each pool		No		x
E		Indexing	Yes	Capture each pool		Yes		2, 3, 4
F		Indexing	No	Capture each library		Yes		x
G			No	Capture each library	Indexing	Yes		5, 6
H	Indexing		Yes	Capture each pool	Indexing	Yes		7

NOTE: The top row with the arrows shows the sequence of major steps in the in-solution hybrid selection workflow (not all steps shown). The indexing step can be performed during adapter ligation (A–C), during the first limited cycle PCR (D–F), or during the second limited cycle PCR (G and H). The only exception is scheme H in which indexes are added at two separate times (see main text). Pooling of libraries can only take place after the individual libraries have been indexed and following the first or second PCR. Below, the term "adapter ligation" is used to mean traditional ligase-mediated or transposase-mediated adapter attachment (i.e., tagmentation) and "hybrid capture" refers to whether an individual hybridization reaction includes a single library (i.e., "capture each library") or a pool of libraries (i.e., "capture each pool"). References: 1 = Faircloth et al. (2015); 2 = Lemmon et al. (2012); 3 = Smith et al. (2013); 4 = Meiklejohn et al. (2016); 5 = Meyer and Kircher (2010); 6 = Faircloth et al. (2012); 7 = Rohland and Reich (2012). x = unknown if studies have used this scheme.

A–C, which tend to be implemented in library kits, do the indexing step at the time of adapter ligation. The next three (D–F), which are usually done in non-kit protocols, index libraries during the first limited cycle PCR. Scheme G delays the indexing step until the second limited cycle PCR and H introduces index steps during the ligation and second PCR steps. Interestingly, at least half of these schemes have been used in studies showing that none of them is evidently dominant over the others in practice. What are the advantages and disadvantages to each of these schemes?

The schemes in which pooling is done in two stages—first to make hybrid capture pools followed by "pooling the pools" (B, E, H)—are certainly better for studies that have large numbers of samples (e.g., >30 libraries). However, for studies with a smaller number of libraries, making only the hybrid pools (A, D) will still be beneficial from the standpoint of conserving RNA baits. If hybrid selection is performed on single libraries, which are indexed before or after hybrid selection (C, F, G), then hybridization efficiencies are expected to be better than the results obtained from pooled samples. A drawback to the C, F, G approaches is that they are not practical for studies with large numbers of libraries. Scheme H may, at first glance, appear peculiar—after all, why does this approach, which was developed by Rohland and Reich (2012), introduce an index into the adapter–fragment construct during the ligation and second limited cycle PCR steps?

During the ligation step in the Rohland and Reich library-making approach, one adapter consisting of a 6-bp barcode sequence and binding sites for a nontailed PCR primer is ligated to one end of the library fragment while the other adapter only contains sites for the other nontailed PCR primer (Figure 7.18). A second (7-bp) index is added along with the missing sequence elements during the second limited cycle PCR (Figures 7.30 and 7.31). Thus, the initial barcoding of libraries permits the pooling of samples for hybridization using minimal-length adapters, whereas the index added later is used to distinguish pools. The combination of barcode and index sequences makes this approach amenable to high-throughput hybrid selection and sequencing projects (Rohland and Reich 2012). However, what is the significance of these short adapters?

Recall our discussion from earlier in this chapter concerning the potentially adverse effects of daisy chaining on hybridization reactions. Hodges et al. (2009) surmised that long adapter sequences—those containing the P5 and P7 common adapters, index sequences, and binding sites for PCR and sequencing primers—are at high risk of hybridizing to complementary adapter sequences, which could result in the capture and sequencing of unwanted library molecules

TABLE 7.6
Comparison of Illumina adapter lengths for adapter constructs used in pooled hybridization reactions

Types of adapters	Length of P5 adapter	Length of P7 adapter
Illumina TruSeq Nano HT	70 (8)	66 (8)
Meyer and Kircher (2010)	58	65 (7)
Rohland and Reich (2012)	34 (6)	33
Illumina Nextera	70 (8)	66 (8)

NOTE: Shown are the lengths (in bp) of adapters for several different adapter types made for Illumina library construction. Number within parentheses represents the number of bases for the index (when present). Adapter lengths are for adapters used in hybrid selection reactions and they are based on Figures 7.1b, 7.17, 7.18d, and 7.26.

(Figure 7.29b). Rohland and Reich (2012) tested this hypothesis and found that libraries with their "truncated" adapters, which are approximately half the length of full-length adapters (Table 7.6), had 2-fold higher hybridization efficiencies compared to libraries with long adapters. Hybrid efficiency here refers to the percentage of reads that mapped to target loci. Thus, daisy chaining appears to have a severe effect on the efficiency of hybrid selection reactions when long adapters are used.

One strategy that can be used with library-making approaches that produce long indexed adapters prior to the hybridization step is to simply not add the P5 and P7 common adapters and indices until the second limited cycle PCR (scheme G in Table 7.5; Meyer and Kircher 2010; Faircloth et al. 2012). However, this approach is not practical for studies with large numbers of libraries. As pointed out earlier, careful selection of adapter blocking oligos may also help ameliorate the daisy chaining problem but additional studies are needed to determine if an optimal indexing and pooling scheme exists. On reflection, perhaps this is not a big problem after all, as recent studies (e.g., McCormack et al. 2015) show that it is still possible to collect thousands of loci despite the use of long adapters in hybridization reactions.

7.3 COST-EFFECTIVE METHODS FOR OBTAINING MULTIPLEXED TARGETED-LOCI LIBRARIES

In-solution hybrid selection represents the best approach for obtaining large phylogenomic datasets consisting of hundreds to thousands of DNA sequence loci. However, a major drawback is that this approach is likely to be prohibitively expensive for many smaller phylogenomic laboratories owing to the high cost of obtaining commercially made probe kits. Fortunately, there are a number of other more cost-effective approaches for utilizing Illumina sequencing to generate phylogenomic datasets. One of these methods is a form of in-solution hybrid selection, which uses biotinylated PCR amplicons as baits instead of RNA (Maricic et al. 2010; Peñalba et al. 2014). In this method, which has been termed *sequence capture using PCR-generated probes* or "SCPP" (Peñalba et al. 2014), PCR baits can be easily generated in any laboratory at low cost (i.e., mainly the cost of the primers) thereby obviating the need to purchase expensive bait kits (Maricic et al. 2010; Peñalba et al. 2014). Another approach, which is known as *parallel tagged amplicon sequencing* or "PTS," involves the sequencing of pooled adapter-tagged PCR products (O'Neill et al. 2013). Although these methods could potentially be used to produce datasets comprised of hundreds of loci, they are practical for obtaining datasets of relatively modest sizes (e.g., 20–100 loci). The cost-effectiveness of these methods, however, means that smaller labs can obtain phylogenomic datasets consisting of (at least) dozens of loci for dozens to hundreds of individuals. We will now review these two approaches.

7.3.1 Sequence Capture Using PCR-Generated Probes (SCPP)

Maricic et al. (2010) described a method for using PCR product baits to capture target loci fragments from indexed Illumina libraries that had been pooled together. The authors illustrated the utility of this method, which can be applied to any set of target loci for which PCR products can be produced, by generating a sufficient number of reads from a single flow cell lane to construct 46 complete mitochondrial genomes. As can be seen in Figure 7.33, their workflow largely resembles the in-solution workflow we examined earlier in this chapter. This overview of the workflow shows that the PCR baits and indexed sequencing libraries (i.e., the "pond") must first be prepared before the hybrid selection reaction can take place (Figure 7.33).

To begin preparation of the PCR baits, PCR products for the target sequences in at least one exemplar individual must be generated. In the Maricic et al.

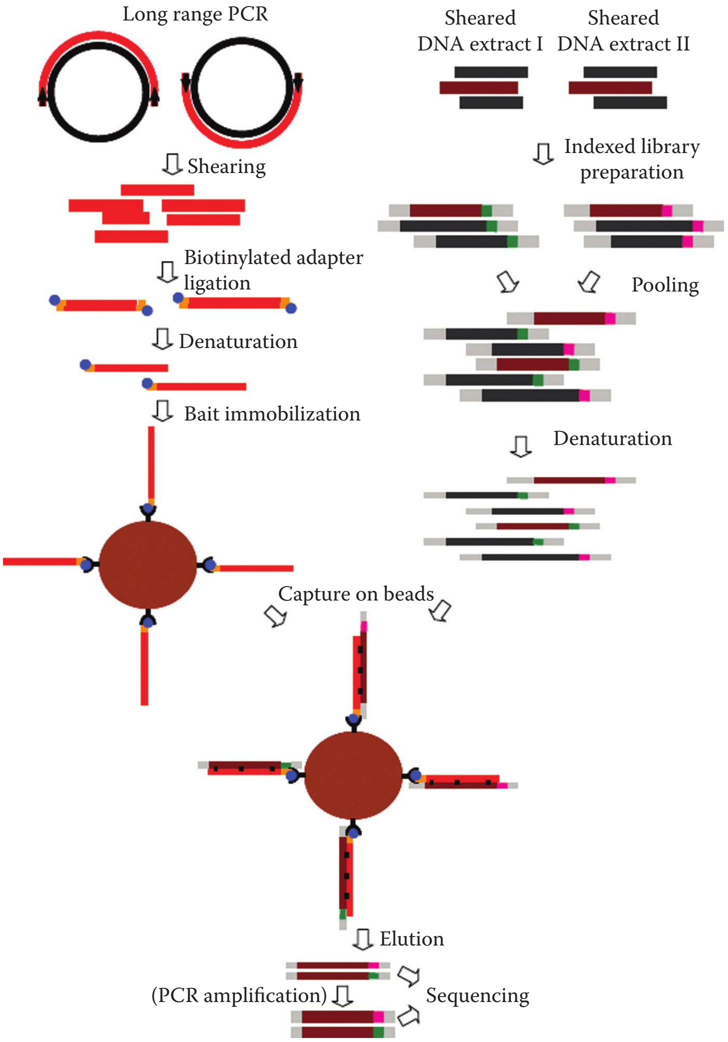


Figure 7.33. Overview of the Maricic et al. SSCP workflow applied to the sequencing of entire human mitochondrial genomes. The preparation of baits (top left) begins with the generation of two long PCR products that cover the entire mitochondrial genome. The PCR products are cleaned then pooled before they are sheared into fragments. Next, biotinylated adapters are ligated to the fragmented PCR products followed by a denaturation step to generate single-stranded bait DNA (light red) for immobilization on streptavidin-coated beads. Indexed libraries are constructed (top right) then pooled before being denatured into ssDNA. The hybridization reaction (bottom) proceeds after the bead-immobilized PCR baits are combined with the pooled library DNA. Captured library fragments (dark red) are eluted and directly sequenced or they are PCR-amplified before the sequencing step. Color codes for library DNAs: Indices are green and pink while the adapters are light gray. Thick lines indicate dsDNA while thin lines show ssDNA. (Reprinted from Maricic, T. et al. 2010. *PLoS One* 5:e14004. With permission.)

(2010) study, the authors used long PCR to produce two long mitochondrial DNA fragments that covered the entire mitochondrial genome for one of the genetic samples. Next, they SPRI-cleaned the fragments and then quantified their concentrations using a Nanodrop. In the following step, the two PCR products were combined together in equimolar ratios before subjecting the pool to a sonication treatment in order to generate a fragment size range of 150–850 bp. The fragments were biotinylated by ligating biotin-labeled adapters to each fragment. Following a column-based purification step, the biotinylated baits were immobilized on streptavidin-coated beads (Figure 7.33).

The library making process began with DNA extraction from the 46 samples (Figure 7.33).

Next, all DNA extracts were subjected to a sonication treatment to create fragment sizes between 150 and 850 bp. A portion of each fragmented sample was then used as input DNA to generate indexed Illumina libraries via the Meyer and Kircher (2010) approach discussed earlier. Following the first limited cycle PCR, the concentration of each library was measured using a Nanodrop before all libraries were grouped together, in equimolar amounts, to form a hybridization pool.

The hybridization reaction procedure began with the preparation of a hybridization mixture consisting of pond DNA, blocking oligo mix, Agilent blocking mix, and Agilent hybridization buffer. The mixture was then heat-denatured at 95°C to generate single-stranded adapter-ligated library fragments (Figure 7.33). The mixture was incubated at a temperature that allowed the blocking oligos and DNAs to bind to their targets (65°C) before bead-bound PCR baits were added to complete the hybridization reaction mixture. The hybridization reaction was incubated at 65°C for 48 hours. Later, nonhybridized DNA was discarded and a series of wash steps, which increased in stringency with each step, were used to further eliminate nonspecific hybrids. Captured DNA fragments were released from baits (i.e., eluted) via a treatment with NaOH. Next, the captured DNA was quantified using qPCR. If a sufficient amount of template was obtained from the hybridization reaction, then the next step is cluster generation followed by sequencing. However, if insufficient template is found, then a second limited cycle PCR must be used to further enrich the library pool. If a second limited cycle PCR is used, then it is essential to use the fewest number of cycles possible in order to minimize the number of duplicated fragments as well as to lower the risk of recombination between amplicons representing different individuals (Maricic et al. 2010). In addition to obtaining enough sequence reads to assemble all 46 mitochondrial genomes, the authors noted that they found no evidence of numts contamination in their sequence data.

Peñalba et al. (2014) modified the Maricic et al. (2010) method to make it more suitable for phylogenomic studies involving groups of organisms with varying levels of intraclade divergences. These modifications were primarily for improving capture efficiencies of target loci particularly when libraries with varying phylogenetic distances to the PCR baits are being captured. Peñalba et al. (2014) also used standard library-making approaches (i.e., Illumina Truseq kit and Meyer-Kircher [2010] approach).

The first key modification involved diversifying intralocus sequences within PCR bait sets. Maricic et al. (2010) had generated their PCR bait set from long PCR products obtained from a single individual. While this approach is expected to work well for studies operating at intraspecific levels or recently diverged species, hybridization efficiencies would likely suffer in studies involving species with moderate to high divergences. Thus, Peñalba et al. (2014) broadened the hybridization capabilities of their PCR bait sets by including amplicons obtained from species that bracketed levels of sequence divergences found within each study group. For each locus, good quality PCR products obtained from different species are pooled together and then electrophoresed in an agarose gel. Next, each target band is excised from the gel and purified using a gel-purification kit. Though this procedure is much more laborious than the SPRI-bead based purification used in the Maricic et al. protocol (Figure 7.34), it should yield pure target amplicons by eliminating all smaller and larger PCR artifacts.

After all loci have been extracted from the gel and purified, they are pooled together in equimolar ratios. Another difference between the Maricic et al. and Peñalba et al. methods is that the former used sonication to shear the long PCR products into smaller fragments, while the latter leaves the PCR amplicons intact (Figure 7.34). The pooled amplicons are biotinylated the same way in both protocols (i.e., via ligating biotinylated adapters to the ends of amplicons), but the Peñalba et al. approach involves making two aliquots of PCR baits for each project (Figure 7.34). In both protocols, the next step is to denature PCR baits to make them single-stranded and then immobilize them on streptavidin-coated magnetic beads.

Peñalba et al. (2014) made several modifications to the hybridization reaction procedures in an effort to enhance capture efficiencies of target sequences. The first of these changes was to add Cot-1 DNA to the hybridization mix in order to reduce the chances of capturing nontarget library DNA containing repetitive DNA (Figure 7.34). The second change was to use a longer hybridization time (72 hours) and a touch-down PCR program that begins at the standard hybridization

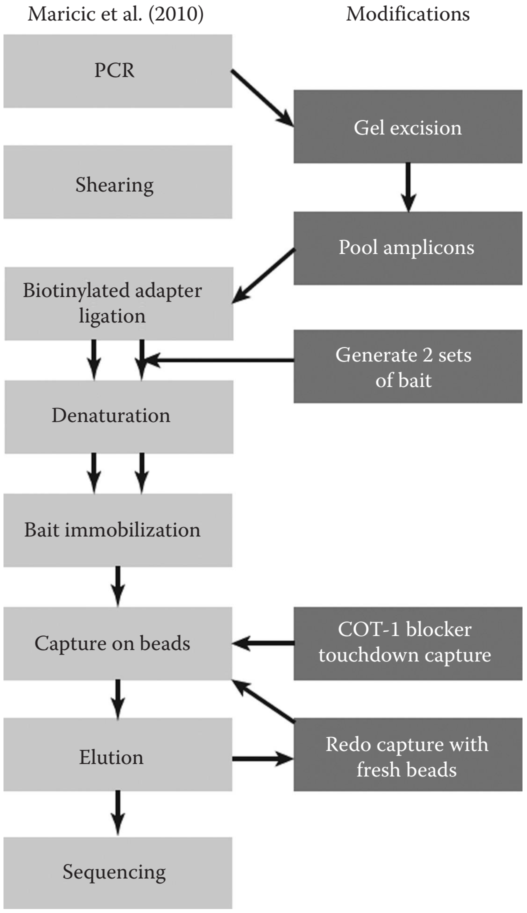


Figure 7.34. Overview of the Peñalba et al. modified SSCP workflow. The original SSCP workflow developed by Maricic et al. (2010) is shown on the left in the light gray rectangles while the modified steps used by Peñalba et al. (2014) are shown in the dark gray rectangles to the right. Note, this workflow only shows the PCR bait preparation and hybridization procedures and not the library-making methods, which are standard. (Reprinted from Peñalba, J. V. et al. 2014. *Mol Ecol Resour* 14:1000–1010. With permission.)

temperature of 65°C and slowly decreases in temperature down to a final temperature of 60°C. By gradually lowering the stringency of the hybridization temperature, it is believed that capture efficiency of more divergent sequences (relative to the bait sequences) can be increased. The third change was to redo the capture with a set of fresh PCR baits. In this procedure, the eluate from the first hybridization reaction is subjected to a second hybridization reaction using the second (fresh) aliquot of PCR baits (Figure 7.34). As discussed earlier, it is also important to perform a relative qPCR assay in order to evaluate the success of SCPP reactions. Peñalba et al. (2014) provide a detailed explanation of this procedure involving SCPP reactions.

The modified SCPP method allows for capitalization of primers in the literature or new primers can be designed and used (Chapter 8). This method is therefore well suited for small labs that already have many primer sets optimized for their study systems (Peñalba et al. 2014). In practical terms, this method can be used to generate phylogenomic datasets of modest sizes (25–100 target loci) for nonmodel organisms involving shallow-level divergences (e.g., phylogeography) or deeper level divergences (Peñalba et al. 2014).

7.3.2 Parallel Tagged Amplicon Sequencing

The study by O'Neill et al. (2013) showed how PTS can be used to generate phylogenomic datasets consisting of nearly 100 nuclear loci. These authors used 454 sequencing to obtain haplotype sequences from 95 independent loci representing 93 individual tiger salamanders. Although it is possible to obtain datasets with a larger number of loci, notice that their study required a minimum of 8,835 PCRs! In another study on amphibians, Barrow et al. (2014) used PTS with Illumina sequencing to generate haplotype sequences for 27 nuclear and three mitochondrial loci from 44 individual chorus frogs. While PTS may not be practical for harvesting hundreds or thousands of loci from genomes, this method does have at least one advantage over in-solution hybrid selection methods when anonymous loci are used.

As you will recall, one of the desirable properties of anonymous loci is that they are believed to often meet the neutrality assumption. Moreover, in Chapters 2 and 3, we learned that a large fraction of all transposable elements in genomes are no longer functional and thus these sequences—provided they are not linked to functional sites—can represent ideal anonymous loci. However, researchers have long expressed worries about designing PCR-based loci containing repetitive DNA because of concerns about amplifying and sequencing paralogous copies. Thus, repetitive DNA is usually masked or not given consideration for designing new anonymous loci. For reasons that will be made clear in Chapter 8, PCR primers, when properly designed, can faithfully obtain the same "orthologous" sequences even when these sequences contain one or more transposable elements. This high level of specificity means that special measures for avoiding repetitive DNA such as designing PCR primers in regions of the genome not containing repetitive DNA need not be used for PTS. In contrast, RNA or PCR product probes used in in-solution hybrid selection studies will, if they contain repetitive DNA sequences, be more prone to obtaining multiple copies of loci thereby potentially leading to the creation of datasets that are contaminated with with paralogous sequences. Even if the paralogy problem didn't exist or if you were to use anonymous loci free of repetitive DNA, there is another reason why it would not be practical to use in-solution hybridization methods that use RNA probes to generate anonymous loci datasets. Unless a given anonymous loci bait set were to be applied to closely related species groups, given the lack of sequence conservation typical of anonymous loci, a new bait set would need to be developed and purchased for each study group, a practice that would likely be cost-prohibitive for most labs. Thus, PTS used with Illumina sequencing is a cost-effective method for generating datasets consisting of dozens of single-copy, neutral, and independent loci for different organismal groups.

REFERENCES

Adessi, C., G. Matton, G. Ayala et al. 2000. Solid phase DNA amplification: Characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res* 28:e87–e87.

Adey, A., H. G. Morrison, X. Xun et al. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biol* 11:R119.

Barrow, L. N., H. F. Ralicki, S. A. Emme, and E. M. Lemmon. 2014. Species tree estimation of North American chorus frogs (Hylidae: *Pseudacris*) with parallel tagged amplicon sequencing. *Mol Phylogenet Evol* 75:78–90.

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

Bhasin, A., I. Y. Goryshin, M. Steiniger-White, D. York, and W. S. Reznikoff. 2000. Characterization of a Tn5 pre-cleavage synaptic complex. *J Mol Biol* 302:49–63.

Bi, K., D. Vanderpool, S. Singhal, T. Linderoth, C. Moritz, and J. M. Good. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics* 13:403.

Blumenstiel, B., K. Cibulskis, S. Fisher et al. 2010. Unit 18.4 Targeted exon sequencing by in-solution hybrid selection. *Curr Protoc Hum Genet* 18:1–18.

Brito, P. H. and S. V. Edwards. 2009. Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica* 135:439–455.

Bronner, I. F., M. A. Quail, D. J. Turner, and H. Swerdlow. 2014. Improved protocols for Illumina sequencing. *Curr Protoc Hum Genet* 79:18.2.1–18.2.42.

Brown, S. M. 2013. *Next-Generation DNA Sequencing Informatics*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Cronn, R., A. Liston, M. Parks, D. S. Gernandt, R. Shen, and T. Mockler. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res* 36:e122–e122.

Diatchenko, L., Y. F. Lau, A. P. Campbell et al. 1996. Suppression subtractive hybridization: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc Natl Acad Sci USA* 93:6025–6030.

Edwards, S. and S. Bensch. 2009. Looking forwards or looking backwards in avian phylogeography? A comment on Zink and Barrowclough 2008. *Mol Ecol* 18:2930–2933.

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:e65923.

Fedurco, M., A. Romieu, S. Williams, I. Lawrence, and G. Turcatti. 2006. BTA: A novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res* 34:e22–e22.

Fisher, S., A. Barry, J. Abreu et al. 2011. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol* 12:1–15.

Geller, E. 2011. High-throughput indexed library preparation and pooled Agilent exome enrichment for Illumina sequencing platform. http://openwetware.org/images/1/10/Geller_exome.pdf (accessed May 9, 2016).

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.

Hodges, E., M. Rooks, Z. Xuan et al. 2009. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc* 4:960–974.

Holmberg, A., A. Blomstergren, O. Nord, M. Lukacs, J. Lundeberg, and M. Uhlen. 2005. The biotin–streptavidin interaction can be reversibly broken using water at elevated temperatures. *Electrophoresis* 26:501–510.

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

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, E. M. and A. R. Lemmon. 2013. High-throughput genomic data in systematics and phylogenetics. *Annu Rev Ecol Evol Syst* 44:99–121.

Lukyanov, S. A., K. A. Lukyanov, N. G. Gurskaya, E. A. Bogdanova, and A. A. Buzdin. 2007. Selective suppression of polymerase chain reaction and its most popular applications. In *Nucleic Acids Hybridization Modern Applications*, 29–51. Springer: Netherlands.

Mamanova, L., A. J. Coffey, C. E. Scott et al. 2010. Target-enrichment strategies for next-generation sequencing. *Nat Methods* 7:111–118.

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

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

Maricic, T., M. Whitten, and S. Pääbo. 2010. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One* 5:e14004.

Masoudi-Nejad, A., Z. Narimani, and N. Hosseinkhan. 2013. *Next Generation Sequencing and Sequence Assembly: Methodologies and Algorithms* (Vol. 4). New York, Heidelberg, Dordrecht, London: Springer Science & Business Media.

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

McCormack, J. E., W. L. Tsai, and B. C. Faircloth. 2015. Sequence capture of ultraconserved elements from bird museum specimens. *Mol Ecol Resour* doi: 10.1111/1755-0998.12466.

Meiklejohn, K. A., B. C. Faircloth, T. C. Glenn, R. T. Kimball, and E. L. Braun. 2016. Analysis of a rapid evolutionary radiation using ultraconserved elements: Evidence for a bias in some multispecies coalescent methods. *Syst Biol*. doi: 10.1093/sysbio/syw014.

Meyer, M. and M. Kircher. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*. doi: 10.1101/pdb.prot5448.

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

Peñalba, J. V., L. L. Smith, M. A. Tonione et al. 2014. Sequence capture using PCR-generated probes: A cost-effective method of targeted high-throughput sequencing for nonmodel organisms. *Mol Ecol Resour* 14:1000–1010.

Picelli, S., Å. K. Björklund, B. Reinius, S. Sagasser, G. Winberg, and R. Sandberg. 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res* 24:2033–2040.

Portik, D. M., L. L. Smith, and K. Bi. 2016. An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Mol Ecol Resour* doi: 10.1111/1755-0998.12541.

Quail, M. A., I. Kozarewa, F. Smith et al. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5:1005–1010.

Reznikoff, W. S. 2003. Tn5 as a model for understanding DNA transposition. *Mol Microbiol* 47:1199–1206.

Reznikoff, W. S. 2008. Transposon Tn 5. *Annu Rev Genet* 42:269–286.

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

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

Shendure, J., R. D. Mitra, C. Varma, and G. M. Church. 2004. Advanced sequencing technologies: Methods and goals. *Nat Rev Genet* 5:335–344.

Siebert, P. D., A. Chenchik, D. E. Kellogg, K. A. Lukyanov, and S. A. Lukyanov. 1995. An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Res* 23:1087.

Smith, B. T., M. G. Harvey, B. C. Faircloth, T. C. Glenn, and R. T. Brumfield. 2013. Target capture and massively parallel sequencing of ultraconserved elements (UCEs) for comparative studies at shallow evolutionary time scales. *Syst Biol*. doi: 10.1093/sysbio/syt061.

Steiniger-White, M., I. Rayment, and W. S. Reznikoff. 2004. Structure/function insights into Tn5 transposition. *Curr Opin Struct Biol* 14:50–57.

Syed, F., H. Grunenwald, and N. Caruccio. 2009a. Optimized library preparation method for next-generation sequencing. *Nat Methods* 6(10).

Syed, F., H. Grunenwald, and N. Caruccio. 2009b. Next-generation sequencing library preparation: Simultaneous fragmentation and tagging using *in vitro* transposition. *Nat Methods* 6(11).

Toews, D. P., L. Campagna, S. A. Taylor et al. 2015. Genomic approaches to understanding population divergence and speciation in birds. *Auk* 133:13–30.

Turner, E. H., S. B. Ng, D. A. Nickerson, and J. Shendure. 2009. Methods for genomic partitioning. *Annu Rev Genomics Hum Genet* 10:263–284.