Chapter 4
Library Construction for Next Generation Sequencing

Huseyin Kucuktas and Zhanjiang (John) Liu

Several major new sequencing platforms have been adopted recently, and they are collectively referred to as the next generation of DNA sequencers. A common feature among the new generation of sequencing procedures is the elimination of the need to clone DNA fragments into cloning vectors and subsequent amplification of cloned DNA in transformed *Escherichia coli* cells, and purification of DNA templates prior to sequencing. Instead, sequence templates are handled in bulk, and massively parallel sequencing allows the generation of numerous sequences simultaneously. Nonetheless, samples need to be prepared to be adaptable to the sequencing platform. In this chapter, we will describe principles and methods for sample preparation for the next generation sequencing platforms. Our focus will be on planning for the library construction either in-house or through outsourcing rather than on the detailed procedures and protocols for library construction. For detailed protocols, readers are referred to protocols for each sequencing platform.

Sample Preparation for Illumina Sequencing

For the Illumina sequencing platform, sample preparation depends on the sources of DNA or RNA and the purpose of the sequencing project. For instance, the sequencing purpose could be for genomic sequencing, genomic sequencing of multiple samples using bar-coded primers, genomic mate pair sequencing, mRNA sequencing, or sequencing for the discovery of small RNAs. Sources of DNA can be in different forms including genomic DNA or pooled PCR products. For PCR products, a minimum fragment size of 2.5 kb is required, or restriction fragments such as reduced representation libraries can be used. For RNA samples, obviously, first- and second-strand cDNA must be synthesized, and from cDNA, the operations are similar to DNA templates. Here the considerations of using genomic DNA as samples are described. The general steps for library construction is illustrated in Figure 4.1.

Next Generation Sequencing and Whole Genome Selection in Aquaculture Edited by Zhanjiang (John) Liu
 ISBN: 978-0-813-80637-2

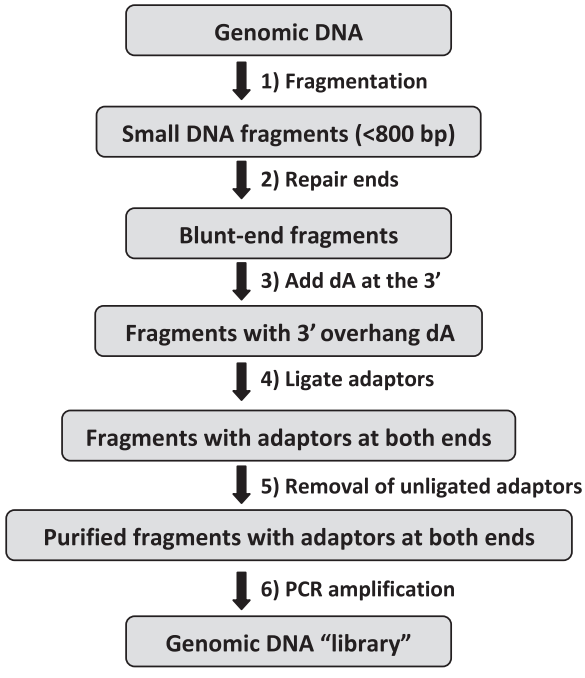

Figure 4.1 The general procedures of library preparation for next generation sequencing.

DNA Fragmentation

The first step of DNA sample preparation is fragmentation of DNA to generate fragments of various sizes. Three approaches are frequently used for DNA fragmentation: nebulization, sonication, and enzymatic fragmentation. DNA fragmentation through nebulization involves the use of a nebulizer that creates a fine mist of DNA by forcing a DNA solution through a small hole in the nebulizer unit. Several factors determine the size of the fragments using a nebulizer including the speed at which the DNA solution passes through the hole, the pressure of the gas blowing through the nebulizer, the viscosity of the solution, and the temperature.

The advantage of nebulization is that it is easy, quick, and requires only small amounts of DNA (0.5–5 μg). The disadvantage is that the distribution of the resulting DNA fragments is over a narrow range of sizes (700–1300 bp). It is difficult to obtain small fragments in the range of 200 bp.

Sonication shears DNA into small fragments through the use of hydrodynamic force by using a sonicator. A general sonication protocol can be found in Sambrook and Russell (2001), although conditions for shearing should be adjusted for each sample. Most often, for small fragment generation, high power, one single pulse, for a short period of 1–2 s should be sufficient to generate a whole range of smears of fragments. Generation of very small fragments and large fragments can be difficult with sonication. For instance, fragments smaller than 400 bp can be very difficult to achieve. Similarly, large fragment sizes such as 8 kb or 10 kb can also be difficult to achieve.

DNA fragmentation can also be achieved through the use of enzymatic fragmentation such as New England Biolabs (NEB) fragmentase. The NEBNext™ dsDNA Fragmentase™ generates dsDNA breaks in a time-dependent manner to yield 100–800-bp DNA fragments depending on the reaction time. NEBNext dsDNA Fragmentase contains two enzymes: one randomly generates nicks on dsDNA, and the other recognizes the nicked site and cuts the opposite DNA strand across from the nick, producing dsDNA breaks. The resulting DNA fragments contain short overhangs, 5′-phosphates, and 3′-hydroxyl groups. According to the NEB Web description (www.neb.com/nebecomm/products/productM0348.asp), a comparison of the sequencing results between genomic DNA prepared with the NEBNext dsDNA Fragmentase and with mechanical shearing demonstrates that the NEBNext dsDNA Fragmentase does not introduce any detectable bias during the sequencing library preparation and that no difference in sequence coverage is observed using the two methods. A major advantage of the fragmentase approach is the control of fragment size. The quantities of fragmentase and reaction times can be tested to achieve the optimal results. For instance, 15–30-min treatment of HeLa cell genomic DNA with fragmentase generated a good smear in the range of 100 bp to 2 kb.

In order to generate larger DNA fragments, the NEB fragmentase can be diluted, for example, 1 : 10 dilution with the storage buffer, and the diluted fragmentase can be tested with different reaction times to generate the desired DNA fragment sizes. The ability of generating various sizes of DNA fragments is a major strength of the enzymatic approach.

DNA End Repair

A common problem of DNA after fragmentation is the need for end polishing. This need is demanded when (1) DNA fragments are not blunt-ended with overhangs; and (2) in some cases, the 5′ end is not phosphorylated, prohibiting the fragment from ligation. An enzymatic step is required to repair the ends of DNA fragments to blunt-ended molecules with phosphate at their 5′ ends. Blunt ends can be achieved through the use of T4 DNA polymerase and *E. coli* DNA polymerase I Klenow fragment. The 3′ to 5′ exonuclease activity of these enzymes removes 3′ overhangs, and the polymerase activity fills in the 5′ overhangs. The 5′-phosphate can be added by phosphorylation reactions using polynucleotide kinase. All these functions can be achieved in a single combined reaction (Figure 4.2). For instance, one can use the end repair reagents from a commercial source such as those of NEB or Illumina. Here is a typical end repair reaction with a total volume of 100 µL:

1. Assemble the following into an Eppendorf tube:
   30 µL fragmented DNA sample
   10 µL phosphorylation buffer
   4 µL dNTP solution mix
   5 µL T4 DNA polymerase
   1 µL DNA polymerase I, Klenow fragment
   5 µL T4 polynucleotide kinase
   45 µL dH_20

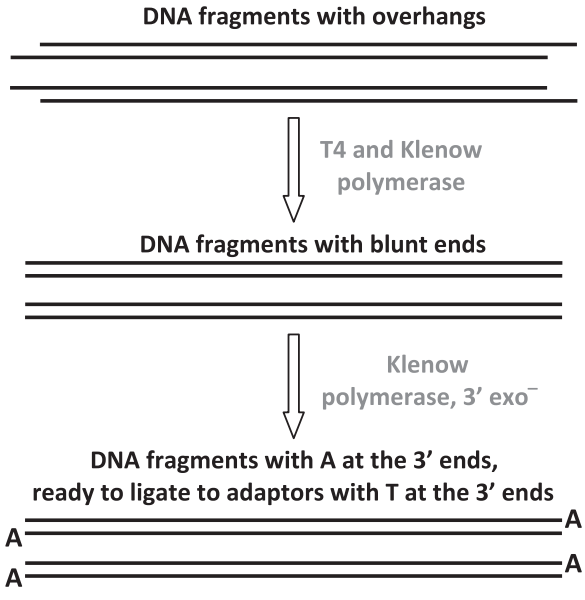

Figure 4.2 Schematic presentation of end repair of fragmented DNA during library preparation for next generation sequencing.

2. Mix gently and incubate for 30 min at 20°C.
3. Purify DNA by phenol/chloroform and ethanol precipitation or DNA purification column. Resuspend DNA in 32 µL TE buffer.

Now, after this reaction is completed, the DNA fragments are repaired to harbor blunt ends with phosphate at their 5′ ends.

Addition of an Extra Base (Adenine) at the 3′ Ends

A single base, adenine (A), needs to be added at the 3′ end of DNA fragments with repaired ends because the adaptors have a single T base overhang at their 3′ end (Figure 4.2). This can be achieved by using Klenow fragment (3′ to 5′ exo⁻). Here is a typical reaction of 50 µL:

1. Assemble the following into an Eppendorf tube:
   32 µL blunted, phosphorylated DNA
   5 µL NEBuffer 2 for Klenow exo⁻
   10 µL dATP solution
   3 µL Klenow fragment (3′ to 5′ exo⁻)
2. Mix gently and incubate for 30 min at 37°C.
3. Purify DNA by phenol/chloroform and ethanol precipitation or DNA purification column. Resuspend DNA in 32 µL TE buffer.

Now, the DNA fragments with a 3′-A are ready to be ligated to adaptors.

Ligation to Adaptors

The adaptors can be from the Illumina or custom made. Custom-made adaptors are required if one is attempting to bar-code the samples in order to sequence multiple samples in a single lane (see the section on "Indexed Libraries for Sequencing Multiple Samples in a Single Lane"). In order to ensure adaptor ligation, a 10:1 molar excess of adaptors are used as compared with DNA fragments. Here is a typical ligation reaction of 50 μL:

1. Mix the following in an Eppendorf tube:
   10 μL DNA sample
   25 μL DNA ligase buffer (2×)
   10 μL adaptor oligo mix
   5 μL DNA ligase
2. Incubate for 15 min at room temperature.
3. Purify the ligation product. Basically, an agarose gel is run to separate the DNA fragments ligated to adaptors. Regions of interest, for example, a 150–200 bp range for short template and 300–650 bp for long template, is excised from the gel. Qiagen Gel Extraction Kit (Qiagen part # 28704) and MinElute Extraction Kit (Qiagen part # 28604) works well for the purification of DNA fragments from the gel.

PCR Amplification of the Library

After the DNA fragments of the desired size are purified, they need to be enriched by PCR. This step enriches the DNA fragments with adaptors on both ends. The PCR is conducted using primers that anneal to the ends of the adaptors. Usually, a low number of PCR amplification, for example, 18 cycles, is used to avoid skewing the representation of the library. A typical PCR reaction is the following:

1. Mix the following in a 15 μL tube.
   1 μL purified DNA fragments
   25 μL Phusion DNA polymerase
   1 μL PCR primer #1
   1 μL PCR primer #2
   22 μL water
2. PCR for 18 cycles of 98°C for 10 s, 65°C for 30 s, and 72°C for 30 s, followed by a final incubation at 72°C for 5 min, and then hold at 4°C.
3. Purify the PCR products using the QIAquick PCR Purification Kit, and elute in 50 μL elution buffer.

The library is now ready to be sequenced. However, it is recommended to check the quality and quantity of the library by

- measuring its absorbance at 260 nm;
- checking the 260/280 ratio;
- running a gel to see the isolated fraction with the sizes of the original gel slice; and
- cloning a fraction into a sequencing vector, and then sequencing by Sanger sequencing.

Indexed Libraries for Sequencing Multiple Samples in a Single Lane

Sequencing multiple biological samples in a single lane is of great interest to biologists, and particularly so for aquaculture researchers, as this can significantly reduce the sequencing costs, allowing biological questions to be answered while staying within the budget. In order to sequence multiple biological samples in a single lane, obviously, sample pooling is required. However, before the samples are pooled, different sequence tags can be ligated to different samples to allow bioinformatic tracing of the sources of the samples during sequence analysis. For instance, if one is considering analysis of genetic differences of strains using SNPs, DNA from each strain can be bar-coded with sequence tags in the adaptors before pooling. Obviously, such sequences for bar coding purposes need to be positioned downstream of the sequencing primer so they will be sequenced during sequencing.

Indexes (adapter "bar codes") can be added to genomic DNA or PCR products for sequencing multiple samples (multiplexing up to 12 samples for Illumina libraries) in a single sequencing deliverable. Customers can choose to prepare their own libraries using Illumina's sample prep kits, or sample preparation can be completed at Illumina or other sources. For instance, Hudson/Alpha and Lucigen both offer indexed library services at a cost of $300–$400 per tagged library. With the Illumina's Multiplexing Sample Preparation Oligonucleotide Kit, 12 unique oligonucleotides are included to "tag" libraries for pooling in a single lane of a flow cell, or up to 96 samples can be sequenced on a single flow cell using the Genome Analyzer.

Selection of proper sequences for bar coding is important because any sequencing errors may tamper the ability to differentiate the sequences if the tags are similar. As a rule of thumb, the more the sequence differences, the better the bar coding tags are. Table 4.1 provides a guide for selecting the bar coding adaptors. Each tagging oligo is six bases in length, assuring for accurate differentiation between tags and to overcome any single-base errors that may inadvertently be introduced during PCR (www.illumina.com/products/multiplexing_sample_preparation_oligonucleotide_kit.ilmn).

The indexed libraries have various applications, in particular in the consideration of research costs. For instance, one can obtain strain-specific SNPs by sequencing DNA from 12 strains in a single lane, with samples from each strain tagged with an index adaptor. Indexed libraries are extremely useful for analysis of tissue expression profiles. RNA from each tissue can be converted to cDNA, tagged with indexed adaptors, and pooled for DNA sequencing. In a similar fashion, RNA samples with various treatments can be tagged and then pooled for sequencing in a single lane. Unless very deep sequencing is required, the construction of indexed libraries cost much less than the cost for each lane of sequencing. Therefore, pooling samples tagged with indexed adaptors is an effective way to reduce costs.

Sample Preparation for 454 Sequencing

Genome Sequencer GS FLX™ System, commonly known as the 454 sequencer, provide long sequence reads (see Chapter 3), currently with an average read length

Table 4.1 A guide for selecting the bar coding adaptors.

1	**ACTGTG**	0											
2	**AGCCAT**	5	0										
3	**ATCTCG**	4	4	0									
4	**CAGTGT**	6	6	5	0								
5	**CGAATG**	4	5	5	5	0							
6	**CTATGC**	6	6	4	3	4	0						
7	**GCTAGT**	4	5	6	4	5	5	0					
8	**GTACTG**	4	5	4	6	3	4	5	0				
9	**GTTGCA**	4	6	4	6	6	5	4	4	0			
10	**TCGCAA**	5	4	6	5	6	6	5	5	5	0		
11	**TGTAGC**	5	5	6	5	4	4	3	6	5	5	0	
12	**TTGCTC**	5	5	5	5	5	4	6	3	5	3	4	0
		ACTGTG	**AGCCAT**	**ATCTCG**	**CAGTGT**	**CGAATG**	**CTATGC**	**GCTAGT**	**GTACTG**	**GTTGCA**	**TCGCAA**	**TGTAGC**	**TTGCTC**

The more different the adaptor sequences, the less likely a mistake will be made in tracing the bar code. The 12 adaptor sequences shown on the first column are cross-compared with themselves listed on the last row of the table. The numbers in the table are the number of bases that are different among them. The table was adopted from Dr. Marta Matvienko of University of California–Davis with permission.

of 400–500 bp; but its read length is improving rapidly. Although not commercially available yet, Roche has achieved average length of over 700–850 bp recently. A single run can be achieved in approximately 10 h, generating over a million reads, thereby providing quite high throughput. Compared with the Illumina sequencing, its read length is much longer, but throughput is lower.

Sample preparation for the 454 sequencing is similar to that for the Illumina sequencing. All the protocols for 454 library construction can be found at http://454.com/my454/documentation/gs-flx-system/manuals.asp. The first step is to generate DNA fragments from genomic DNA to prepare a universal library. Library preparation can be performed by one lab technician in an afternoon without special equipment. A single library preparation can supply enough material for numerous sequencing runs of the Genome Sequencer GS FLX™ System.

Like Illumina sequencing, the 454 sequencers support the sequencing of samples from a wide range of sources including genomic DNA, PCR products, BACs, and cDNA. For genomic DNA and BACs, the first required step is fragmentation into 300–800-bp fragments. Depending on the applications, libraries can be made as shotgun libraries, paired-end read libraries, amplicon libraries, or cDNA libraries.

Protocols for shotgun and paired-end libraries have been developed by Margulies et al. (2005) and Ng et al. (2006), respectively. Shotgun libraries and sequencing are relatively straightforward, but sequences so generated do not provide any physical information. Paired-end libraries are much preferred in a genome sequencing project or any project that requires a level of sequence assembly. Because two reads are generated from each DNA segment, one each from the end of the segment, these reads are "physically linked" with the space of the size of the segments, providing scaffolding capabilities to the paired reads.

Detailed general library construction protocol is available with Roche 454 at http://454.com/downloads/my454/documentation/gs-flx/method-manuals/GS-FLX-Titanium-General-Library-Preparation-Method-Manual-%28Apr2009%29.pdf. This protocol is suitable for a general library, that is, libraries other than a paired-end or an amplicon library.

Protocols for making libraries using PCR products are available from Roche for amplicon library protocols at http://454.com/downloads/my454/documentation/gs-flx/method-manuals/GS-FLX-Titanium-Amplicon-Library-Prep-Method-Manual.pdf. Two protocols are currently available for the construction of paired-end libraries depending on the fragment size. For 3-kb fragment libraries, the protocols are available at http://454.com/downloads/my454/documentation/gs-flx/method-manuals/GS-FLX-Titanium-Paired-End-Library-Prep-3kbSpan-Method-Manual.pdf. For 8–20-kb fragment libraries, the protocols are available at http://454.com/downloads/my454/documentation/gs-flx/method-manuals/GS-FLX-Titanium-Paired-End-Library-Prep-20-8kbSpan-Method-Manual.pdf.

Construction of paired read libraries involves additional steps as compared with that of general libraries. The relatively large (3 kb, or 8–20 kb) fragments need to be ligated to adaptors to mark sequence orientation, followed by circularization. The circularized molecules with adaptors at the ligation junctions are fragmented by nebulization. Afterward, the fragmented DNA need to be end polished and ligated to library adaptors. The remaining steps are similar to the procedures for general libraries.

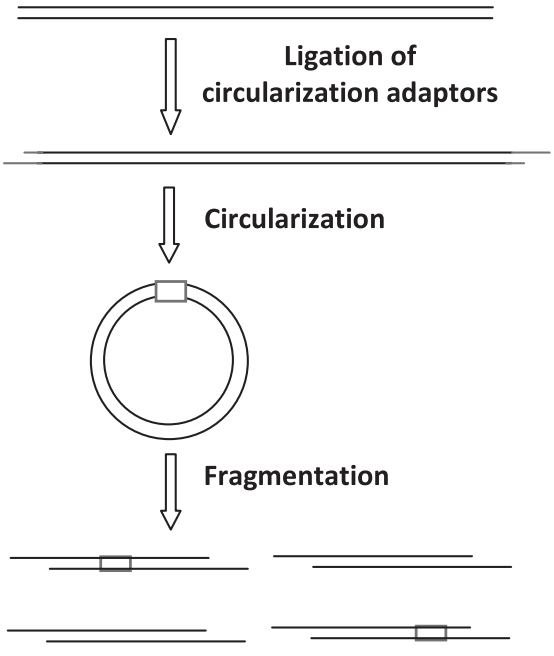

Figure 4.3 Schematic presentation of paired-end read library preparation. See color insert.

It is important to note that only a fraction of the paired reads are true paired reads. These paired reads are marked by sequencing of the adaptor junction in the circularization processes, and the orientation of the sequences are resolved by the adaptor sequences, and the distance between them are estimated by the size of the library (Figure 4.3).

The Roche 454 sequencing is based on pyrosequencing. Therefore, homopolymers would lead to a huge release of lights that may ruin the sequencing reaction recording on the sequencers. Therefore, cDNA libraries are made differently for 454 sequencing. Instead of using poly dT priming, random primer priming is used. The detailed protocol for making cDNA libraries for 454 sequencing is available at http://454.com/downloads/my454/documentation/gs-flx/method-manuals/GS-FLX-Titanium-cDNA-Rapid-Library-Preparation-Method-Manual-%28Jan2010%29.pdf. The major difference is the first steps that involve fragmentation of RNA, then synthesis of first-strand cDNA using random primers. From there, second-strand cDNA synthesis and the subsequent steps are similar to the construction of general libraries.

Sample Preparation for Sequencing Multiplexed Samples Using SOLiD Sequencing

One of the advantages of SOLiD sequencing is that you can have a high level of multiplexing, up to 96 samples in one sequencing reaction. Each of the samples can be bar-coded by a bar code sequence linked to one of the adaptors. Protocol for SOLiD sequencing provided by Applied Biosystems is very detailed, so we are not

attempting to provide a detailed protocol here, but rather concentrate on various considerations for aquaculture applications. We will use the SOLiD™ Fragment Library Barcoding Kit Module 1–16 (PN 4444836) to illustrate the concept and the procedures. Approximately 5 ug DNA is made into 100–150-bp fragments; then these fragments are linked to the adaptors, one of which carries a unique sequence identifier or bar code. The bar code tag enables multiplexed sequencing of multiple samples in a single sequencing reaction. The DNA fragments are ligated with a truncated Multiplex P1 Adaptor and a Multiplex P2 Adaptor with a bar code. The Multiplex P2 Adaptor consists of three segments: (1) an internal adaptor sequence (derived from the sequence used for mate-paired libraries); (2) a bar code decamer sequence; and (3) a standard P2 adaptor sequence. Because the Multiplex P2 Adaptor is longer than the standard P2 adaptor, the Multiplex P1 Adaptor is truncated to keep the total length of the adaptors approximately the same as for a nonbar-coded library.

The steps for making an ABI SOLiD library is similar to those described above for the Illumina or 454 sequencing platforms. The procedure involves DNA fragmentation, end repair, and adaptor ligation (Figure 4.3). The key to multiplexed sample sequencing is the use of bar codes, one for each of the samples contained in P2 adaptors. After adaptor ligation, the libraries are amplified by using the adaptor sequences as primers. Each library is quantified using qPCR, and then equal molar of each library is pooled together for multiplex sequencing. Detailed protocols for the library construction can be found at ABI's Web site: http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_059675.pdf.

Acknowledgments

Research in my laboratory is supported by grants from United States Department of Agriculture, Agriculture and Food Research Initiative (USDA AFRI) Animal Genome and Genetic Mechanisms Program, USDA National Research Initiative (NRI) Basic Genome Reagents and Tools Program, Mississippi–Alabama Sea Grant Consortium, Alabama Department of Conservation, United States Agency for International Development, National Science Foundation, and Binational Agricultural Research and Development Fund. The authors would like to thank Dr. Martha Matvienko for allowing us to use the adapter sequences shown in Table 4.1.

References

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR,

Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, and Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380.

Ng P, Tan JJ, Ooi HS, Lee YL, Chiu KP, Fullwood MJ, Srinivasan KG, Perbost C, Du L, Sung WK, Wei CL, and Ruan Y. 2006. Multiplex sequencing of paired-end ditags (MS-PET): A strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Research*, 34(12):e84.

Sambrook J and Russell D. 2001. *Molecular Cloning: A Laboratory Manual*, 3rd edn. Cold Spring Harbor Press, Cold Spring Harbor, NY.