Published in final edited form as:
*Biotechniques*. ; 56(2): 61–passim. doi:10.2144/000114133.

# Library construction for next-generation sequencing: Overviews and challenges

**Steven R. Head**[1], **H. Kiyomi Komori**[2], **Sarah A. LaMere**[2], **Thomas Whisenant**[2], **Filip Van Nieuwerburgh**[3], **Daniel R. Salomon**[2], and **Phillip Ordoukhanian**[1]

[1]NGS and Microarray Core Facility, The Scripps Research Institute, La Jolla, CA

[2]Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA

[3]Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium

## Abstract

High-throughput sequencing, also known as next-generation sequencing (NGS), has revolutionized genomic research. In recent years, NGS technology has steadily improved, with costs dropping and the number and range of sequencing applications increasing exponentially. Here, we examine the critical role of sequencing library quality and consider important challenges when preparing NGS libraries from DNA and RNA sources. Factors such as the quantity and physical characteristics of the RNA or DNA source material as well as the desired application (i.e., genome sequencing, targeted sequencing, RNA-seq, ChIP-seq, RIP-seq, and methylation) are addressed in the context of preparing high quality sequencing libraries. In addition, the current methods for preparing NGS libraries from single cells are also discussed.

### Keywords

deep sequencing; DNA; RNA; library preparation; next-generation sequencing; RNA-seq; DNA-seq; ChIP-seq; RIP-seq

Over the past five years, next-generation sequencing (NGS) technology has become widely available to life scientists. During this time, as sequencing technologies have improved and evolved, so too have methods for preparing nucleic acids for sequencing and constructing NGS libraries (1,2). For example, NGS library preparation has now been successfully demonstrated for sequencing RNA and DNA from single cells (3–11).

Address correspondence to Steven R. Head, Director, NGS and Microarray Core Facility, The Scripps Research Institute, La Jolla, CA. shead@scripps.edu.

**Competing Interests**
The authors DRS, SRH, and PO are founding scientists and consultants to Transplant Genomics Inc.

**Author contributions**
SRH, PO, and DRS wrote and edited the paper. HKK contributed to the Methylseq section, SALM contributed to the ChIP-seq section, TW contributed to the RIP-seq section, and FVN contributed to the sections on Nextera and mate-pair library sections.

To purchase reprints of this article, contact: biotechniques@fosterprinting.com

Fundamental to NGS library construction is the preparation of the nucleic acid target, RNA or DNA, into a form that is compatible with the sequencing system to be used (Figure 1). Here, we compare and contrast various library preparation strategies and NGS applications, focusing primarily on those compatible with Illumina sequencing technology. However, it should be noted that almost all of the principles discussed in this review can be applied with minimal modification to NGS platforms developed by Life Technologies, Roche, and Pacific Biosciences.

## Fragmentation/Size selection

In general, the core steps in preparing RNA or DNA for NGS analysis are: (*i*) fragmenting and/or sizing the target sequences to a desired length, (*ii*) converting target to double-stranded DNA, (*iii*) attaching oligonucleotide adapters to the ends of target fragments, and (*iv*) quantitating the final library product for sequencing.

The size of the target DNA fragments in the final library is a key parameter for NGS library construction. Three approaches are available to fragment nucleic acid chains: physical, enzymatic, and chemical. DNA fragmentation is typically done by physical methods (i.e., acoustic shearing and sonication) or enzymatic methods (i.e., non-specific endonuclease cocktails and transposase tagmentation reactions)(12). In our laboratory, acoustic shearing with a Covaris instrument (Covaris, Woburn, MA) is typically done to obtain DNA fragments in the 100–5000 bp range, while Covaris g-TUBEs are employed for the 6–20 Kbp range necessary for mate-pair libraries. Enzymatic methods include digestion by DNase I or Fragmentase, a two enzyme mix (New England Biolabs, Ipswich MA). Comparisons of NGS libraries constructed with acoustic shearing/sonication versus Fragmentase found both to be effective (13). However, Fragmentase produced a greater number of artifactual indels compared with the physical methods. An alternative enzymatic method for fragmenting DNA is Illumina's Nextera tagmentation technology (Illumina, San Diego, CA) in which a transposase enzyme simultaneously fragments and inserts adapter sequences into dsDNA. This method has several advantages, including reduced sample handling and preparation time (12).

Desired library size is determined by the desired insert size (referring to the library portion between the adapter sequences), because the length of the adaptor sequences is a constant. In turn, optimal insert size is determined by the limitations of the NGS instrumentation and by the specific sequencing application. For example, when using Illumina technology, optimal insert size is impacted by the process of cluster generation in which libraries are denatured, diluted and distributed on the two-dimensional surface of the flow-cell and then amplified. While shorter products amplify more efficiently than longer products, longer library inserts generate larger, more diffuse clusters than short inserts. We have successfully sequenced libraries with Illumina instruments up to 1500 bases in length.

Optimal library size is also dictated by the sequencing application. For exome sequencing, more than 80% of human exomes are under 200 bases in length (14). We run 2 × 100 paired-end reads and our exome sequencing libraries typically contain insert sizes of approximately 250 bases in length as a compromise to match the average size of most exons while sequencing without overlapping read pairs. The size of an RNA-Seq library is also determined by the applications. We typically do basic gene expression analysis using single-end 100 base reads. However, for analysis of alternative splicing or determination of transcription start and stop sites, we employ 2 × 100 base paired-end reads. In most instances, the RNA will be fragmented before conversion into cDNA. This is typically done through the use of controlled heated digestion of the RNA with a divalent metal cation (magnesium or zinc). The desired length of the library insert can be adjusted by increasing or decreasing the time of the digestion reaction with good reproducibility.

In a recent study of seven different RNA-seq library preparation methods (15), the majority involve some sort of fragmentation of the mRNA prior to adapter attachment. The two that do not use a hexamer priming method (16) or in the case of the SMARTer Ultra Low RNA Kit (Clontech, Mountain View, CA)(17), a full length cDNA is synthesized with a fixed 3′ and 5′ sequence added so that the entire cDNA library (average 2 kb in length) can be amplified in long distance PCR (LD-PCR). This amplified double-stranded cDNA is then fragmented by acoustic shearing to the appropriate size and used in a standard Illumina library preparation (involving end-repair and kination, A-tailing and adapter ligation, followed by additional amplification by PCR).

A second post-library construction sizing step is commonly used to refine library size and remove adaptor dimers or other library preparation artifacts. Adapter dimers are the result of self-ligation of the adapters without a library insert sequence. These dimers form clusters very efficiently and consume valuable space on the flow cell without generating any useful data. Thus, we typically use either magnetic bead-based clean up, or we purify the products on agarose gels. The first works in most instances for samples where sufficient starting material is available. When sample input is limiting, more adapter dimer products are often generated. In our experience, bead-based methods may not perform optimally in this situation and combining bead-based with agarose gel purifications may be necessary.

In the case of microRNA (miRNA)/ small RNA library preparation, the desired product is only 20–30 bases larger than the 120 bp adaptor dimers. Therefore, it is critical to perform a gel size selection to enrich the libraries as much as possible for the desired product. This resolution of separation is not feasible using beads. Alternatively, we often create large library inserts (1 kb) combined with longer reads (2 × 300 base paired-end) and no PCR amplification for de novo assembly of bacterial genomes. To optimize the value of the data generated for de novo assembly, it is necessary to do careful gel-based size selections to ensure uniform insert size.

## NGS library construction using fragmented/size selected DNA

There are several important considerations when preparing libraries from DNA samples, including the amount of starting material and whether the application is for resequencing (in which a reference sequence is available to align reads to) or de novo sequencing (in which the reads will need to be assembled to create a new reference sequence). Library preparations can be susceptible to bias resulting from genomes that contain unusually high or low GC content and approaches have been developed to address these situations through careful selection of polymerases for PCR amplification, thermocycling, conditions and buffers (18, 19–21).

Library preparation from DNA samples for sequencing whole genomes, targeted regions within genomes (for example exome sequencing), ChIP-seq experiments, or PCR amplicons (see below) follows the same general workflow. Ultimately, for any application, the goal is to make the libraries as complex as possible (see below).

Numerous kits for making sequencing libraries from DNA are available commercially from a variety of vendors. Competition has driven prices steadily down and quality up. Kits are available for making libraries from microgram down to picogram quantities of starting material. However, one should keep in mind the general principle that more starting material means less amplification and thus better library complexity.

With the exception of Illumina's Nextera prep, library preparation generally entails: (*i*) fragmentation, (*ii*) end-repair, (*iii*) phosphorylation of the 5′ prime ends, (*iv*) A-tailing of the 3′ ends to facilitate ligation to sequencing adapters, (*v*) ligation of adapters, and (*vi*) some number of PCR cycles to enrich for product that has adapters ligated to both ends (1) (Figure 1). The primary differences in an Ion Torrent workflow are the use of blunt-end ligation to different adapter sequences.

Once the starting DNA has been fragmented, the fragment ends are blunted and 5′ phosphorylated using a mixture of three enzymes: T4 polynucleotide kinase, T4 DNA polymerase, and Klenow Large Fragment. Next, the 3′ ends are A-tailed using either Taq polymerase or Klenow Fragment (exo-). Taq is more efficient at A-tailing, but Klenow (exo-) can be used for applications where heating is not desired, such as preparing mate-pair libraries. During the adapter ligation reaction the optimal adapter:fragment ratio is ~10:1, calculated on the basis of copy number or molarity. Too much adapter favors formation of adapter dimers that can be difficult to separate and dominate in the subsequent PCR amplification. Bead or column-based cleanups can be performed after end repair and A-tail reactions, but after ligation we find bead-based cleanups are more effective at removing excess adapter dimers.

To facilitate multiplexing, different barcoded adapters can be used with each sample. Alternatively, barcodes can be introduced at the PCR amplification step by using different barcoded PCR primers to amplify different samples. High quality reagents with barcoded adapters and PCR primers are readily available in kits from many vendors. However, all the components of DNA library construction are now well documented, from adapters to enzymes, and can readily be assembled into "home-brew" library preparation kits.

An alternative method is the Nextera DNA Sample Prep Kit (Illumina), which prepares genomic DNA libraries by using a transposase enzyme to simultaneously fragment and tag DNA in a single-tube reaction termed "tagmentation" (Figure 2)(22). The engineered enzyme has dual activity; it fragments the DNA and simultaneously adds specific adapters to both ends of the fragments. These adapter sequences are used to amplify the insert DNA by PCR. The PCR reaction also adds index (barcode) sequences. The preparation procedure improves on traditional protocols by combining DNA fragmentation, end-repair, and adaptor-ligation into a single step. This protocol is very sensitive to the amount of DNA input compared with mechanical fragmentation methods. In order to obtain transposition events separated by the appropriate distances, the ratio of transposase complexes to sample DNA is critical. Because the fragment size is also dependent on the reaction efficiency, all reaction parameters, such as temperatures and reaction time, are critical and must be tightly controlled.

Sequencing the genomes of single cells has been recently reported by several group (11,23–26). The current strategy utilizes whole genome amplification with multiple displacement amplification (MDA). MDA is based on the use of random primers with phi29, a highly processive strand displacing polymerase (27). While this technique is capable of generating enough amplified material to construct sequencing libraries, it suffers from considerable bias, created by nonlinear amplification. A recent report demonstrated a significantly improved method of MDA by adding a quasi-linear preamplification step that reduced bias (10). A technology platform based on small compartmentalization and microfluidics can be used to facilitate library preparation from up to 96 single cells per run is offered by Fluidigm (South San Francisco, CA).

## NGS library construction using RNA

It is important to consider the primary objective of an RNA sequencing experiment before making a decision on the best library protocol. If the objective is discovery of complex and global transcriptional events, the library should capture the entire transcriptome, including coding, noncoding, anti-sense and intergenic RNAs, with as much integrity as possible. However, in many cases the objective is to study only the coding mRNA transcripts that are translated into the proteins. Yet another objective might be to profile only small RNAs, most commonly miRNA, but also small nucleolar RNA (snoRNA), piwi-interacting RNA (piRNA), small nuclear RNA (snRNA), and transfer RNA (tRNA). While we will endeavor to describe the principles of RNA sequencing libraries in this review, it is not possible to explain all of the different protocols available. Interested readers should research the many options (Table 1) themselves.

One of the first and earliest successes in applying NGS to RNA-seq was in the case of miRNA (28,29). The protocols for preparing miRNA sequencing libraries are surprisingly simple and are usually performed in a one-pot reaction (Figure 3). The fact that miRNAs are found in their native state with a 5′ terminal phosphate allows the use of ligases to selectively target miRNAs.

In the first step of the Illumina protocol (Figure 3A), an adenylated DNA adapter with a blocked 3′ end is ligated to the RNA sample using a truncated T4 RNA ligase 2. This enzyme is modified to require the 3′ adapter substrate to be adenylated. The result is that fragments of other RNA species in the total RNA sample are not ligated together in this reaction; only the pre-adenylated oligo-nucleotide can be ligated to free 3′ RNA ends. Moreover, since the adapter is 3′ blocked, it cannot serve as a substrate for self-ligation. In the next step, a 5′ RNA adapter is added along with ATP and RNA ligase 1. Only RNA molecules whose 5′ ends are phosphorylated will be effective substrates for the ligation reaction. After this second ligation, a reverse transcription (RT) primer is hybridized to the 3′ adapter and a RT-PCR amplification is performed (usually 12 cycles). Due to the small but predictable size of the miRNA library (120 bases of adapter sequence plus the miRNA insert of ~20–30 bases), the library or a pooled sample composed of multiple barcoded libraries can be run on a gel and size selected. The gel size selection is critical due to the presence of adapter dimer side products created during the ligation reaction as well as higher molecular weight products generated from ligation of other non-miRNA RNA fragments containing 5′ phosphate groups (e.g., tRNA and snoRNA). This library preparation method results in an oriented library such that the sequencing always reads from the 5′ end to the 3′ end of the original RNA species. The principle of miRNA sequencing on the Ion Torrent platform is similar (Figure 3B). Ion Torrent uses dual duplex adapters that ligate to the miRNA's 3′ and 5′ ends in a single reaction, followed by RT-PCR. This general library prep approach can also be used to create a directional RNA-seq library from any RNA substrate.

One major limitation in miRNA library construction arises when the amount of input RNA is low (e.g., <200 ng total RNA); short adapter dimers compete in the RT-PCR reaction with the desired product, adapters, and miRNA inserts. When too many adapter dimers are present they stream up the gel during the size selection step and contaminate the product bands. To minimize this problem, many commercial miRNA library preparation kits now incorporate various strategies to suppress adapter dimer formation.

For mRNA sequencing libraries, methods have been developed based on cDNA synthesis using random primers, oligo-dT primers, or by attaching adapters to mRNA fragments followed by some form of amplification. mRNA can be primed by random oligomers or by an anchored oligo-dT to generate first strand cDNA. If random priming is used, the rRNA must first be removed or reduced. rRNA can be removed using oligonucleotide probe-based reagents, such as Ribo-Zero (Epicenter, Madison, WI) and RiboMinus (Life Technologies, Carlsbad, CA). Alternatively, poly-adenylated RNA can be positively selected using oligo-dT beads.

It is often desirable to create libraries that retain the strand orientation of the original RNA targets. For example, in some cases transcription creates anti-sense RNA constructs that may play a role in regulating gene expression (30). In fact, long noncoding RNA (lncRNA) analysis depends on directional RNA sequencing (31). Methods for preparing directional RNA-seq libraries are now readily available (15). The concept is to perform the cDNA reaction and remove one of the two strands selectively, by incorporating dUTP into the second strand cDNA synthesis reaction. The uracil-containing strand can then be removed enzymatically (32) (NEBNext Ultra Directional RNA Library Prep Kit for Illumina) or prevented from further amplification with a PCR polymerase that cannot recognize uracil in the template strand (Illumina TruSeq Stranded Total RNA kit). In addition, actinomycin D is frequently added to the first strand cDNA synthesis reaction to reduce spurious antisense synthesis during the first strand synthesis reaction (33).

An alternative and hybrid method utilizes random or anchored oligo-dT primers with an adapter sequence on the 5′ end of the primer to initiate first strand cDNA synthesis. Next, in a procedure called template switching (shown in Figure 4B), a 3′ adapter sequence is added to the cDNA molecule (17). This method has a distinct advantage in that the first strand cDNA molecule can be PCR amplified directly without second strand synthesis using the unique sequence tag put on the 3′ end by the template switching reaction. A 5′ unique sequence tag is also introduced by standard priming in the first strand synthesis.

The strategic design of the primers used for cDNA synthesis is a powerful strategy for making RNA-seq libraries. For example, rRNA sequences can be avoided by including strategically designed primers that target rRNA but do not allow subsequent amplification. A commercial kit (NuGEN Ovation RNA-seq; San Carlos, CA) combines SPIA nucleic acid amplification technology (34) with primers used in the first strand cDNA synthesis that are designed to suppress amplification of rRNA sequences. Another method was reported in which all 4096 possible hexamer sequences were screened against rRNA sequences to identify and eliminate perfect matches. A pool of 749 hexamers remained that was then used to prime the first strand cDNA synthesis reaction. The result was a drop in rRNA reads from 78% to 13% in the sequencing data (16). Finally, a method called DP-seq (7) was developed, in which the amplification of a majority of the mouse transcriptome was accomplished using a defined set of 44 heptamer primers. This primer sequence design selectively suppressed the amplification of highly expressed transcripts, including rRNA, and provided a reliable estimation of low abundance transcripts in a model of embryonic development.

Recently methods for preparing RNA-seq libraries from single cells have been reported (Figure 4)(3–5,8,9). One strategy utilizes polynucleotide tailing of the first strand cDNA (Figure 4A)(5,8), which can be combined with a template switching reaction (Figure 4B) (4,9). The end result is a first strand cDNA product that can be amplified by universal PCR primers. The version shown in Figure 4B has been incorporated into a commercially available kit (SMARTer Ultra Low RNA Kit; Clontech). An alternative approach called CEL-Seq incorporates a T7 promoter sequence at the 5′ end of the cDNA, followed by linear amplification using in vitro transcription (Figure 4C)(3).

A typical cell has approximately 10 pg of total RNA and may contain only 0.1 pg of poly-adenylated RNA. Thus, these approaches all require some sort of whole-transcript amplification to generate enough material to make a sequencing library (5). The downside of such extensive amplification is the generation of significant technical noise, and this problem has yet not been solved (35).

Finally, ribosomal footprinting can reveal the pool of cellular mRNA transcripts undergoing translation at any point in time (36,37). The protocol involves treating cell lysates with RNase, leaving behind only the 30-nucleotide region protected by each ribosome. Ribosomes are then purified by sucrose density gradient centrifugation, and the co-purified mRNA fragments are extracted from the ribosomes. Another novel application of RNA sequencing is SHAPE-Seq (Selective 2′-hydroxyl acylation analyzed by primer extension) (38), which is used to probe the secondary structure of RNA via acylating reagents that preferentially modify unpaired bases. When the modified RNA and an unmodified control undergo RT using specific primers, the resulting cDNA fragments can be sequenced and compared to reveal nucleotide level base pairing information.

## ConsiderationsinNGSlibrary preparation: Complexity, bias, and batch effects

The main objective when preparing a sequencing library is to create as little bias as possible. Bias can be defined as the systematic distortion of data due to the experimental design. Since it is impossible to eliminate all sources of experimental bias, the best strategies are: (*i*) know where bias occurs and take all practical steps to minimize it and (*ii*) pay attention to experimental design so that the sources of bias that cannot be eliminated have a minimal impact on the final analysis.

The complexity of an NGS library can reflect the amount of bias created by a given experimental design. In terms of library complexity, the ideal is a highly complex library that reflects with high fidelity the original complexity of the source material. The technological challenge is that any amount of amplification can reduce this fidelity. Library complexity can be measured by the number or percentage of duplicate reads that are present in the sequencing data (39). Duplicate reads are generally defined as reads that are exactly identical or have the exact same start positions when aligned to a reference sequence (40). One caveat is that the frequency of duplicate reads that occur by chance (and represent truly independent sampling from the original sample source) increases with increasing depth of sequencing. Thus, it is critical to understand under what conditions duplicate read rates represent an accurate measure of library complexity.

Using duplicate read rates as a measure of library complexity works well when doing genomic DNA sequencing, because the nucleic acid sequences in the starting pool are roughly in equimolar ratios. However, RNA-seq is considerably more complex, because by definition the starting pool of sequences represents a complex mix of different numbers of mRNA transcripts reflecting the biology of differential expression. In the case of ChIP-seq the complexity is created by both the differential affinity of target proteins for specific DNA sequences (i.e., high versus low). These biologically significant differences mean that the number of sequences ending up in the final pool are not equimolar.

However, the point is the same—the goal in preparing a library is to prepare it in such a way as to maximize complexity and minimize PCR or other amplification-based clonal bias. This is a significant challenge for libraries with low input, such as with many ChIP-seq experiments or RNA/DNA samples derived from a limited number of cells. It is now technologically possible to perform genomic DNA and RNA sequencing from single cells. The key point is that the level of extensive amplification required creates bias in the form of preferential amplification of different sequences, and this bias remains a serious issue in the analysis of the resulting data. One approach to address the challenge is a method of digital sequencing that uses multiple combinations of indexed adapters to enable the differentiation of biological and PCR-derived duplicate reads in RNA-seq applications (41,42). A version of this method is now commercially available as a kit from Bioo Scientific (Austin, TX).

When preparing libraries for NGS sequencing, it is also critical to give consideration to the mitigation of batch effects (43–45). It is also important to acknowledge the impact of systematic bias resulting from the molecular manipulations required to generate NGS data;

for example, the bias introduced by sequence-dependent differences in adaptor ligation efficiencies in miRNA-seq library preparations. Batch effects can result from variability in day-to-day sample processing, such as reaction conditions, reagent batches, pipetting accuracy, and even different technicians. Additionally, batch effects may be observed between sequencing runs and between different lanes on an Illumina flow-cell. Mitigating batch affects can be fairly simple or quite complex. When in doubt, consulting a statistician during the experimental design process can save an enormous amount of wasted money and time.

There are many ways to minimize bias during library preparation. Within a single experiment, we aim to start with samples of similar quality and quantity. We also use master mixes of reagents whenever possible. One particularly egregious source of bias is from amplification reactions such as PCR; it is well documented that GC content has a substantial impact on PCR amplification efficiency. We recommend PCR enzymes such as Kapa HiFi (Kapa Biosystems, Wilmington, MA) or AccuPrime Taq DNA Polymerase High Fidelity (Life Technologies) that have been shown to minimize amplification bias resulting from extremes of GC content. It was recently reported that, for particularly high GC targets, a 3 min initial denaturation time with subsequent PCR melt cycles extended to 80 s can significantly reduce amplification bias (18). We use as few amplification cycles as necessary, but it is critical that every sample within an experiment is amplified the same number of cycles. In miRNA library preparation protocols, ligase enzymes have been shown to contribute a high level of sequence-dependent bias (46,47). One group found that addition of three degenerate bases to the 5′ end of the 3′ adapter and the 3′ end of the 5′ adapter significantly reduced this ligation bias (48). A miRNA library prep kit that incorporates three degenerate bases on the 5′ adapter is commercially available through Gnomegen (San Diego, CA).

In addition to enzymatic steps, bias can be reduced in purification steps by pooling barcoded samples before gel or bead purification. In the case of miRNA-seq libraries, we first run the individual libraries on an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA) to quantitate the miRNA peaks. We use this information to create barcoded library pools of up to 24 samples and then perform gel purification in a single lane of an agarose gel to avoid sizing variation between samples.

## Sample preparation for NGS applications: Targeted and amplicon sequencing

Targeted sequencing allows investigators to study a selected set of genes or specific genomic elements; for example, CpG islands and promoter/enhancer regions (reviewed in References 49). A common application of targeted sequencing is exome sequencing and high quality kits are commercially available; SureSelect (Agilent Technologies), SeqCap (Roche NimbleGen, Madison, WI) and TruSeq Exome Enrichment Kit (Illumina). All three capture methods are based on probe hybridization to enrich sequencing libraries made from whole genome samples (51,52). Life Technologies has commercialized an alternative approach based on highly multiplexed, PCR-based AmpliSeq technology. There are options to customize all these products and investigators can design capture or PCR probes for target regions covering from thousands to millions of bases within a genome.

Hybridization capture approaches generally work well but can suffer from off-target capture and struggle to effectively capture sequences with high levels of repetition or low complexity (i.e., the Human Histocompatibility Locus region). The PCR-based AmpliSeq method is more efficient with lower amounts of DNA (53). It should also be noted that probes are based on a reference sequence, and variations that substantially deviate from the reference, as well as significant insertion/deletion mutations, are not always going to be identified.

Another targeted sequencing method, developed by Raindance (Billerica, MA) uses microdroplet PCR and custom-designed droplet libraries (54,55). The nature of micro-droplet emulsion PCR significantly decreases PCR amplification bias (56). Microdroplet PCR allows the user to set up $1.5 \times 10^6$ micro-droplet amplifications in a single tube in under an hour. The droplet libraries are designed based on 500 bp amplicons, and a single custom library can target from 2000 to 10,000 different amplicons covering up to $5 \times 10^6$ bases.

Amplicon sequencing involves making NGS libraries from PCR products. This form of targeted sequencing is more appropriate for applications such as microbiomic experiments where community composition is analyzed by surveying 16S rRNA sequences in complex bacterial mixtures (57), analysis of antibody diversity (58) and T cell receptor gene repertoires (50), and facilitating the process of identifying and selecting high value aptamers in a SELEX protocol (59). To highlight the flexibility of amplicon sequencing, a recent study used the method to analyze the incorporation of unnatural nucleotides during DNA synthesis (60).

Sequencing of short amplicons also makes obtaining entire sequences possible in either a single read or using a paired-end read design. Here, adapters can be added directly to the ends of the amplicons and sequenced to retain haplotype information essential for reconstructing antibody or T cell receptor gene sequences as well as identifying species in micro-biome projects.

However, it is often necessary to design longer amplicons for targeted sequencing applications. In this case, the PCR products need to be fragmented for sequencing. Amplicons can be fragmented as-is using acoustic shearing, sonication, or enzymatic digestion. Alternatively, they can be first concatenated into longer fragments using ligation followed by fragmentation. One problem associated with amplicon sequencing is the presence of chimeric amplicons generated during PCR by PCR-mediated recombination (61). This problem is exacerbated in low complexity libraries and by overamplification. A recent study identified up to 8% of raw sequence reads as chimeric (62). However, the authors were able to decrease the chimera rate down to 1% by quality filtering the reads and applying the bioinformatic tool, Uchime (63). The presence of the PCR primer sequences or other highly conserved sequences presents a technical limitation on some sequencing platforms that utilize fluorescent detection (i.e., Illumina). This can occur with amplicon-based sequencing such as microbiome studies using 16S rRNA for species identification. In this situation, the PCR primer sequences at the beginning of the read will generate the exact same base with each cycle of sequencing, creating problems for the signal detection hardware and software. This limitation is not an issue with Ion Torrent systems (not fluorescence-based) and can be addressed on Illumina systems by sequencing multiple different amplicons in the same lane whenever possible. An alternative strategy we employ is to use several PCR primers during PCR of a specific amplicon. Each primer has a different number of bases (typically 1–3 random bases) added to the 5′ end to offset/ stagger the order of sequencing when adapters are ligated to the amplicons.

## Sample preparation for NGS applications: Mate pair sequencing and other strategies

The objective of de novo sequencing is to use algorithms to produce a novel genome assembly that can serve as a reference for future experiments. Closing contigs and scaffolds into a cohesive genome map can be a remarkably challenging task. Because of this, de novo assemblies require some of the highest quality (i.e., least biased, most representative) sequencing libraries of any NGS application.

We routinely use three library preparation strategies to maximize assembly efficiency: (*i*) libraries comprised of long inserts (~1 kb insert sizes), (*ii*) no PCR amplification in library preparation, and (*iii*) mate-pair libraries with long distance spacing (5–20 kb) between reads. While it has so far proven impossible to build mate-pair libraries without PCR amplification, long insert libraries can easily be constructed without PCR if sufficient DNA is available (2). Such long insert libraries are created by careful shearing of genomic DNA. We find that the final data quality is greatly improved if sheared ~1 kb DNA is first size selected on a 1% agarose gel to narrow the size distribution as much as possible. This step minimizes the possibility for small fragments to concatenate during the adapter ligation step that increases the risk of chimeric read pairs impeding the data assembly process.

Mate-pair libraries are constructed by circularization of input DNA that has been fragmented to a size of >2 kb. Typically, insert size measures between 2 and 20 kb. We developed a mate-pair protocol using Cre-Lox recombination instead of blunt end circularization (64). In this method, a biotin-labeled LoxP sequence is created at the junction site from the end ligation of two LoxP adapters. This strategy allows junctions to be identified without using a reference genome. The location of the LoxP sequence in the reads distinguishes true mate-paired reads from spurious paired-end reads using the bioinformatics tool, Deloxer (64). A similar approach improves upon this method by allowing longer insert sizes (up to 22 kb) (65). Illumina also provides a transposome-based protocol that requires only a small amount of input material (~1 μg) and allows barcoded multiplexing of up to 12 samples per lane.

A significantly more complicated protocol generates mate-pair reads with approximately 40 kb spacing using a unique fosmid vector design (Lucigen NxSeq 40 kb Mate-Pair Cloning Kit; Middleton, WI). The phage packaging mechanism selects for DNA fragments of ~40 kb, which are packaged into phage particles in vitro by bacteriophage Lambda packaging extract followed by transfection into *Escherichia coli* for replication. Experience in fosmid preparation and replication is a definite plus before taking on this protocol.

## Sample preparation for NGS applications: ChIP-seq

Chromosome immunoprecipitation sequencing (ChIP-seq) is now a well-established method for evaluating the presence of histone modifications and/or transcription factors on a genome-wide scale. Histone modifications are an important part of the epigenomic landscape and are thought to help regulate the recruitment of transcription factors and other DNA modifying enzymes. The precise biological role of histone modifications is still poorly understood, but genome-wide studies using ChIP-seq are beginning to provide important insights into their patterns and purpose.

Originally developed as a low-throughput PCR-based assay, the introduction of NGS technology has allowed ChIP-seq to be efficiently applied on a genome wide scale (Figure 5). The general principle of this assay involves immunoprecipitation of specific proteins along with their associated DNA. The procedure usually requires DNA-protein crosslinking with formaldehyde followed by fragmentation of the chromatin using micro-coccal nuclease (MNase) and/or sonication. Specific antibodies are used to target the protein or histone modification of interest, at which point the DNA is purified and subjected to high throughput sequencing. The sequencing results should be compared with a proper control. Data from a successful ChIP-seq should be enriched for the sequences that were crosslinked to the targeted protein/ modified histone.

There has been some discussion on the best controls for ChIP-seq. Rabbit IgG has been used as a control for non-specific antibody binding, but these antisera typically don’t control well for the non-specific cross-reactivity that is present with the use of affinity-purified antibodies. Thus, an aliquot of the input DNA pool after fragmentation but before immunoprecipitation has become more commonplace as the control for ChIP-seq. Additionally, input controls appear to give a better estimation of biases that result from chromatin fragmentation and sequencing (66).

ChIP-seq has a number of technical challenges that require consideration and more standardization to facilitate cross-study analysis. In particular, antibody quality is a large factor affecting the outcome of ChIP-seq experiments. The ENCODE (Encyclopedia Of DNA Elements; www.genome.gov/10005107) and Roadmap consortia (NIH Roadmap Epigenomics Mapping Consortium) have set forth procedures for assessing antibody quality, including dot blot immunoassays against histone tail peptides to evaluate binding specificity and cross-reactivity (67). Some of the technical procedures used in ChIP-seq studies have a direct impact on downstream ChIP-seq library preparation and the resulting sequencing data (40,66,68,69). For example, the formaldehyde crosslinking typically used in ChIP-seq experiments is particularly important for studying transcription factors, but it appears to result in lower resolution and increases the likelihood of non-specific interactions (40). Resolution was recently addressed for DNA binding proteins with the use of lambda exonuclease to digest the 5′ ends at a fixed distance from the crosslinked protein, thus greatly reducing contaminating non-specific DNA (66). Additionally, the use of formaldehyde crosslinking has been shown to protect DNA from micrococcal nuclease digestion, so sonication is now the preferred method of fragmentation when using ChIP-seq in the assessment of DNA binding proteins. Conversely, micrococcal nuclease is known to digest the linker regions between nucleosomes, so it remains the preferred method for chromatin fragmentation when studying histone modifications (68). Regardless of fragmentation method, if successful the DNA insert plus the sequencing adapters should be ~300 bp. We routinely do bead-based purifications after sequencing adapter ligation and again after the PCR step in the library protocol in order to minimize sample losses.

One of the greatest technical issues in ChIP-seq has been the requirement for large amounts of starting material (68). Typically, 1 million to 20 million cells are required per IP in order to acquire sufficient material for sequencing. These amounts are particularly difficult to achieve for primary cells, progenitor cells, and clinical samples. This remains an area that will benefit greatly from improved sequencing library preparation methods from very small quantities of relatively short fragments of DNA. To date, most methods attempting to ameliorate the large amount of starting material required for ChIP-seq have required whole genome amplification or extensive PCR amplification. However, the recently introduced Nano-ChIP-seq method allows for starting amounts down to 10,000 cells by using custom primers with hairpin structures and an internal *Bci*VI restriction site (66,70). In another recent development, ChIP-seq for the transcription factor ERalpha was successfully performed with an input of only 5000 cells by using single tube linear amplification (LinDA). This approach uses an optimized T7 RNA polymerase IVT-based protocol, which was demonstrated to be robust and reduced amplification bias due to GC content (66).

It is especially challenging to study a novel DNA binding protein or histone modification for which there are no commercial antibodies. The approach required in these cases usually entails the use of transient or stable expression of the protein of interest with a tag that can be targeted (such as a His or FLAG tag). The drawback of this approach is the need for extensive controls to ensure that the fusion protein is localized properly and that interactions are not affected by steric hindrance or non-endogenous expression levels (67).

## Sample preparation for NGS applications: RIP-seq/CLIP-seq

Transcription of primary RNAs begins a complex process involving the recognition of intron/exon junctions, splicing and alternative splicing, addition of poly(A) tails, transport to the cytoplasm, entry into ribosomes, processing of various non-coding RNAs, and the generation of signals for RNA degradation. One powerful tool for studying these events, and the proteins that control them, is RIP-seq, where protein complexes assembled at different sites on the RNA molecules are immunoprecipitated and then the RNA bound to them is purified and sequenced (Figure 6)(71).

RNA binding proteins (RBPs) recognize ribonucleic acid motifs including specific sequences, single-stranded backbones, secondary structures, and double-stranded RNA (72,73). These interactions involve all types of RNAs and occur at every step from transcription to degradation (74). Many steps in the post-transcriptional processing of messenger RNA overlap, resulting in multiple RBP complexes bound to a transcript at any given moment in its existence (75). RIP-seq can be done with protein-specific antibodies or by expressing tagged versions of the RBPs of interest. Furthermore, RIP-seq provides the ability to characterize the function of an RBP in a specific cell type and/or cell state based on the population of bound RNAs (76–78).

The amount of starting total RNA needed for a successful RIP-seq experiment is significantly greater than that required for RNA-seq. First, the amount of RNA bound by any given RBP is highly variable but always only a fraction of the original pool and often a very minor fraction. Second, depending on the target RBP, a nuclear lysate may be required, necessitating an even greater amount of starting material (79). Another technical challenge is the tendency of RNA to non-specifically bind proteins. We address this limitation by preclearing the lysate with an isotype control antibody bound to beads. Non-specific DNA binding is also a challenge. DNase I treatment should be performed multiple times throughout the protocol (i.e., during lysate preparation, post-TRIZOL separation, and library preparation). The duration of the IP step can vary from 2 h to overnight. Longer incubation times can increase the percentage of pulled down protein; however, non-specific RNA binding is also increased, resulting in additional noise. RIP-purified RNA can be taken directly into standard library protocols suitable for low input, short fragment samples. We have had good success with the ScriptSeq-v2 RNA-Seq Library Preparation Kit (Epicenter) with our RIP-seq samples.

A variation of RIP-seq is crosslinking and immunoprecipitation (CLIP-seq) followed by digestion of the RNA sequences not protected by the RBP complexes. This procedure is used to identify the specific binding sites and flanking sequences of RBPs. In the original CLIP protocol, the starting material was crosslinked by exposure to UV radiation (80). Prior to immunoprecipitation, the prepared lysate is digested with RNase, limiting the RNA populations to those regions protected by the bound RBPs. Next, there is a multistep protocol to radiolabel the RBP-bound RNA, separate the samples by SDS-PAGE, visualize the RNA-protein complex by radiography, and excise the desired region (~5–30 kDa above the target RBP's molecular weight). Finally, the RBP is digested with proteinase K, linkers are ligated to the remaining RNA fragments, and a library is constructed for sequencing (81,82). Control samples are required to account for crosslinking efficiency, RNase digestion, and non-specific RNA binding (83).

Recent modifications to the CLIP-seq protocol include individual-nucleotide resolution CLIP (iCLIP)(84) and photo-activatable-ribonucleoside-enhanced CLIP (PAR-CLIP)(85). In iCLIP, an adapter ligation step is replaced with an intramolecular circularization step that has increased reaction efficiency and the added ability to identify the site of crosslinking (individual nucleotide resolution)(84). In PAR-CLIP, a ribonucleoside analog (4-SU or 6-SG) is added to the media prior to UV-crosslinking. The irradiation step binds the ribonucleoside analog to the RBP in addition to changing the base's identity. Following the standard CLIP-seq protocol, the photoactivated crosslinked sites can be identified by locating single base mismatches or indels when compared with the whole RNA-seq data (86).

## Sample preparation for NGS applications: Methylseq

A fundamental mechanism of the epigenetic regulation of gene activity is DNA methylation. This is rapidly being recognized as a critical feature of disease states where simple genetic inheritance is not sufficient to explain the complexity of the phenotypes encountered in clinical medicine. In principle, DNA methylation changes also reflect the history of the organism, not just the genetic inheritance.

Methylation of the 5 position of cytosine (5mC) is the most common form of DNA methylation, with 60%–80% of the 28 million CpG dinucleotides in the human genome being methylated (87,88). While genome-wide hypomethylation has been linked to increased rates of mutation and chromosomal instability, hypermethylation of promoters inhibits gene transcription (89). DNA methylation is also essential for genetic imprinting, suppression of transposable elements, and X chromosome inactivation (90). Aberrant DNA methylation is associated with many diseases including cancer, autoimmune diseases, inflammatory diseases, and metabolic disorders (91–94).

Early studies were limited to investigating DNA methylation in a few genes at a time or generating a non-specific but global estimation of methylation. Recent advances in high throughput sequencing have dramatically increased both the throughput and resolution of such studies. There are three major methods for studying DNA methylation with NGS platforms: (*i*) restriction enzyme (RE) based, (*ii*) targeted enrichment, and (*iii*) bisulfite sequencing (Figure 7). Each of these methods has advantages and disadvantages that must be weighed according to the researcher’s needs and budget.

Methylation sensitive restriction enzyme sequencing (MRE-seq) relies on restriction enzymes that are sensitive to CpG methylation (Figure 7A)(95,96). The most commonly used REs are the methylation-sensitive *Hpa*II and its methylation-insensitive isoschizomer *Msp*I (97). A method called HELPseq (*Hpa*II tiny fragment enriched by ligation mediated PCR) utilizes both of these enzymes to analyze genome-wide methylation profiles (98). A sample is digested with each enzyme, and the resulting fragments are sequenced separately. The *Msp*I digested reference sample not only allows for a point of comparison for methylation but also controls for misinterpretation of *Hpa*II not cutting due to single nucleotide polymorphisms (SNPs)(97). Other RE-based methods, such as methyl-sensitive cut counting (MSCC), methylation-specific digital sequencing (MSDS), and modified methylation-sensitive digital karyotyping (MMSDK) rely on other methylation sensitive REs (97). RE-based methods are limited in their scope by the fixed number of digestion sites present in the genome, which skews the view of CpG methylation to these particular sites, and its accuracy is dependent upon complete digestion with high fidelity (67).

Affinity enrichment of methylated DNA requires either antibodies specific for methylated DNA (MeDIP) or other proteins capable of binding methylated DNA (MBDseq)(Figure 7B) (95,97,98). Specifically, the methyl binding domain (MBD)-containing proteins MeCP2, MBD1, MBD2, and their binding partner MBD3L1 have been used to immunoprecipitate methylated DNA (98). While such immunoprecipitation methods are not limited by sequence specificity, they tend to preferentially pull down regions that are heavily methylated and miss genomic areas with sparse methylation. Moreover, sequencing of the recovered material gives the researcher an idea of the areas that are methylated, but does not reveal which individual bases are methylated.

Treatment of DNA with sodium bisulfite results in the chemical conversion of unmethylated cytosine to uracil while methylated cytosines are protected (Figure 7C)(99). Bisulfite conversion coupled with shotgun sequencing was first performed in *Arabidopsis thaliana* by two research groups who coined the methods BS-seq (100) and MethylC-seq (101). MethylC-seq was also used to create the first human single base resolution map of DNA methylation (87). While BS-seq/MethylC-seq is widely considered the gold standard in methylome analysis, it requires significant read depth (30× coverage)(67). It remains expensive and not easily applied to the large sample sizes needed for clinical investigations. Recently, it was shown that only ~20% of CpGs are differentially methylated across 30 human cells and tissues, suggesting that 80% of the CpG methylation in whole genome sequencing is not informative (88). To reduce the cost and complexity of data associated with whole genome bisulfite sequencing, recent methods have sought to couple enrichment methods with bisulfite sequencing. The capture and targeted sequencing of specific regions identified in the genome to be enriched for CpG methylation sites such as islands, shores, gene promoters, and differentially methylated regions (DMRs) can be accomplished using a commercially available kit from Agilent Technologies (SureSelect$^{XT}$ Methyl-Seq Target Enrichment). Alternatively, bisulfite conversion of DNA isolated by MeDIP or MBD pull downs allows for single base resolution to be achieved by these methods. Sequence-specific binding to beads (51) followed by bisulfite treatment or binding of bisulfite-converted DNA to bisulfite padlock probes (BSPPs)(102) has also been demonstrated to be an effective method for enriching potentially methylated regions. Our group developed a method for targeted bisulfite sequencing using microdroplet PCR with custom-designed droplet libraries (55). This technique relies on the unbiased amplification of bisulfite treated DNA with region-specific primers. All of these enrichment methods retain the single base pair resolution that is so advantageous for bisulfite sequencing while vastly reducing the amount of sequencing required. However, it is important to note that bisulfite treatment of DNA leads to DNA instability and loss of product; thus, many of these methods require more input DNA than the non-bisulfite conversion-based methods.

The recent discovery that 5-hydroxymethyl-cytosine (5hmC)(103) is an intermediate of the demethylation of 5mC to cytosine has opened a whole new area of study into the mechanics of DNA methylation and epigenetic regulation. Studies revealed that the Ten-Eleven Translocation (TET) family of proteins facilitate demethylation of 5mC to cytosine through three intermediates, 5hmC, 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC). Bisulfite treatment converts 5fC and 5caC to uracil, but cannot convert 5mC or 5hmC. Thus, bisulfite sequencing cannot distinguish between 5mC and 5hmC (67). In order to detect these novel methylation intermediates, new techniques have been developed. The first efforts either involved antibodies specific for 5hmC (hMeDIP-seq) or chemical modification of 5hmC (67). More recent advances toward single-base resolution sequencing of 5hmC are oxidative bisulfite sequencing (oxBS-seq)(104) and TET-assisted bisulfite sequencing (TAB-seq) (105). Single-molecule real-time (SMRT) DNA sequencing (Pacific Biosciences, Menlo

Park, CA) has been introduced as another method to sequence 5hmC (106). SMRT sequencing relies on the kinetics of polymerase incorporation of individual nucleotides, allowing for direct detection of these modified cytosines (106). Most recently, antibody-based immunoprecipitation methods (107,108) and chemical modification methods have been developed to allow for sequencing of 5fC (109).

The tremendous and rapid evolution of NGS technologies and protocols has generated both amazing opportunities for science and significant challenges. We believe that the transformational power of deep sequencing has already been clearly demonstrated in basic science. It is poised to advance into clinical medicine, creating a new generation of molecular diagnostics based on DNA sequencing, RNA sequencing, and epigenetics.

## Acknowledgments

This research was supported by NIH grants: R24 GM09879 (SRH), U19 A1063603 (DRS, SRH, PO, HKK), U01 GM094653 (DRS, TW), U01 AI084146 (DRS, SRH) and U54 AI008353 (SALM), and by a JDRF postdoctoral fellowship (HKK). This paper is subject to the NIH Public Access Policy.

## References

1. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. A large genome center's improvements to the Illumina sequencing system. Nat Methods. 2008; 5:1005–1010. [PubMed: 19034268]
2. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods. 2009; 6:291–295. [PubMed: 19287394]
3. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012; 2:666–673. [PubMed: 22939981]
4. Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012; 30:777–782. [PubMed: 22820318]
5. Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, Ueda HR. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol. 2013; 14:R31. [PubMed: 23594475]
6. Pan X, Durrett RE, Zhu H, Tanaka Y, Li Y, Zi X, Marjani SL, Euskirchen G, et al. Two methods for full-length RNA sequencing for low quantities of cells and single cells. Proc Natl Acad Sci USA. 2013; 110:594–599. [PubMed: 23267071]
7. Bhargava V, Ko P, Willems E, Mercola M, Subramaniam S. Quantitative transcriptomics using designed primer-based amplification. Scientific reports. 2013; 3:1740. [PubMed: 23624976]
8. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009; 6:377–382. [PubMed: 19349980]
9. Islam S, Kjallquist U, Moliner A, Zajac P, Fan JB, Lonnerberg P, Linnarsson S. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011; 21:1160–1167. [PubMed: 21543516]
10. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science. 2012; 338:1622–1626. [PubMed: 23258894]
11. Wang J, Fan HC, Behr B, Quake SR. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. Cell. 2012; 150:402–412. [PubMed: 22817899]
12. Marine R, Polson SW, Ravel J, Hatfull G, Russell D, Sullivan M, Syed F, Dumas M, Wommack KE. Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. Appl Environ Microbiol. 2011; 77:8071–8079. [PubMed: 21948828]

13. Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. PLoS ONE. 2011; 6:e28240. [PubMed: 22140562]
14. Sakharkar MK V, Chow T, Kangueane P. Distributions of exons and introns in the human genome. In Silico Biol. 2004; 4:387–393. [PubMed: 15217358]
15. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods. 2010; 7:709–715. [PubMed: 20711195]
16. Armour CD, Castle JC, Chen R, Babak T, Loerch P, Jackson S, Shah JK, Dey J, et al. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. Nat Methods. 2009; 6:647–649. [PubMed: 19668204]
17. Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. Biotechniques. 2001; 30:892–897. [PubMed: 11314272]
18. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011; 12:R18. [PubMed: 21338519]
19. Seguin-Orlando A, Schubert M, Clary J, Stagegaard J, Alberdi MT, Prado JL, Prieto A, Willerslev E, Orlando L. Ligation bias in illumina next-generation DNA libraries: implications for sequencing ancient genomes. PLoS ONE. 2013; 8:e78575. [PubMed: 24205269]
20. Dabney J, Meyer M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. Biotechniques. 2012; 52:87–94. [PubMed: 22313406]
21. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, Macinnis B, et al. Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. BMC Genomics. 2012; 13:1. [PubMed: 22214261]
22. Adey A, Morrison Asan HG, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 2010; 11:R119. [PubMed: 21143862]
23. Fan HC, Wang J, Potanina A, Quake SR. Whole-genome molecular haplotyping of single cells. Nat Biotechnol. 2011; 29:51–57. [PubMed: 21170043]
24. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011; 472:90–94. [PubMed: 21399628]
25. Gundry M, Li W, Maqbool SB, Vijg J. Direct, genome-wide assessment of DNA mutations in single cells. Nucleic Acids Res. 2012; 40:2032–2040. [PubMed: 22086961]
26. Hou Y, Fan W, Yan L, Li R, Lian Y, Huang J, Li J, Xu L, et al. Genome analyses of single human oocytes. Cell. 2013; 155:1492–1506. [PubMed: 24360273]
27. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. Genome Res. 2001; 11:1095–1099. [PubMed: 11381035]
28. Vigneault F, Sismour AM, Church GM. Efficient microRNA capture and bar-coding via enzymatic oligonucleotide adenylation. Nat Methods. 2008; 5:777–779. [PubMed: 19160512]
29. Buermans HP, Ariyurek Y, van Ommen G, den Dunnen JT, At'Hoen P. New methods for next generation sequencing based microRNA expression profiling. BMC Genomics. 2010; 11:716. [PubMed: 21171994]
30. Morris KV, Vogt PK. Long antisense non-coding RNAs and their role in transcription and oncogenesis. Cell Cycle. 2010; 9:2544–2547. [PubMed: 20581457]
31. Hangauer MJ I, Vaughn W, McManus MT. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet. 2013; 9:e1003569. [PubMed: 23818866]
32. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 2009; 37:e123. [PubMed: 19620212]

33. Perocchi F, Xu Z, Clauder-Munster S, Steinmetz LM. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. Nucleic Acids Res. 2007; 35:e128. [PubMed: 17897965]

34. Kurn N, Chen P, Heath JD, Kopf-Sill A, Stephens KM, Wang S. Novel isothermal, linear nucleic acid amplification systems for highly multiplexed applications. Clin Chem. 2005; 51:1973–1981. [PubMed: 16123149]

35. Bhargava VH, Head SR, Ordoukhanian P, Mercola M, Subramaniam S. Technical variations in low-input RNA-seq methodologies. Sci Rep. 2014; 4:3678. [PubMed: 24419370]

36. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc. 2012; 7:1534–1550. [PubMed: 22836135]

37. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. Genome-wide annotation and quantitation of translation by ribosome profiling. Curr Protoc Mol Biol. 2013; Chapter 4(Unit 4): 18. [PubMed: 23821443]

38. Mortimer SA, Trapnell C, Aviran S, Pachter L, Lucks JB. SHAPE-Seq: High-Throughput RNA Structure Analysis. Curr Protoc Chem Biol. 2012; 4:275–297. [PubMed: 23788555]

39. Parkinson NJ, Maslau S, Ferneyhough B, Zhang G, Gregory L, Buck D, Ragoussis J, Ponting CP, Fischer MD. Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. Genome Res. 2012; 22:125–133. [PubMed: 22090378]

40. Gilfillan GD, Hughes T, Sheng Y, Hjorthaug HS, Straub T, Gervin K, Harris JR, Undlien DE, Lyle R. Limitations and possibilities of low cell number ChIP-seq. BMC Genomics. 2012; 13:645. [PubMed: 23171294]

41. Fu GK, Hu J, Wang PH, Fodor SP. Counting individual DNA molecules by the stochastic attachment of diverse labels. Proc Natl Acad Sci USA. 2011; 108:9026–9031. [PubMed: 21562209]

42. Shiroguchi K, Jia TZ, Sims PA, Xie XS. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci USA. 2012; 109:1347–1352. [PubMed: 22232676]

43. Taub MA, Corrada Bravo H, Irizarry RA. Overcoming bias and systematic errors in next generation sequencing data. Genome Med. 2010; 2:87. [PubMed: 21144010]

44. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010; 11:733–739. [PubMed: 20838408]

45. Lauss M, Visne I, Kriegner A, Ringner M, Jonsson G, Hoglund M. Monitoring of technical variation in quantitative high-throughput datasets. Cancer Inform. 2013; 12:193–201. [PubMed: 24092958]

46. Zhuang F, Fuchs RT, Sun Z, Zheng Y, Robb GB. Structural bias in T4 RNA ligase-mediated 3′-adapter ligation. Nucleic Acids Res. 2012; 40:e54. [PubMed: 22241775]

47. Hafner M, Renwick N, Brown M, Mihailovic A, Holoch D, Lin C, Pena JT, Nusbaum JD, et al. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA. 2011; 17:1697–1712. [PubMed: 21775473]

48. Sorefan K, Pais H, Hall AE, Kozomara A, Griffiths-Jones S, Moulton V, Dalmay T. Reducing ligation bias of small RNAs in libraries for next generation sequencing. Silence. 2012; 3:4. [PubMed: 22647250]

49. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010; 7:111–118. [PubMed: 20111037]

50. Fischer N. Sequencing antibody repertoires: the next generation. MAbs. 2011; 3:17–20. [PubMed: 21099370]

51. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009; 27:182–189. [PubMed: 19182786]

52. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet. 2007; 39:1522–1527. [PubMed: 17982454]

53. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013; 155:27–38. [PubMed: 24074859]

54. Tewhey R, Warner JB, Nakano M, Libby B, Medkova M, David PH, Kotsopoulos SK, Samuels ML, et al. Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol. 2009; 27:1025–1031. [PubMed: 19881494]

55. Komori HK, LaMere SA, Torkamani A, Hart GT, Kotsopoulos S, Warner J, Samuels ML, Olson J, et al. Application of microdroplet PCR for large-scale targeted bisulfite sequencing. Genome Res. 2011; 21:1738–1745. [PubMed: 21757609]

56. Hori M, Fukano H, Suzuki Y. Uniform amplification of multiple DNAs by emulsion PCR. Biochem Biophys Res Commun. 2007; 352:323–328. [PubMed: 17125740]

57. Fouts DE, Pieper R, Szpakowski S, Pohl H, Knoblach S, Suh MJ, Huang ST, Ljungberg I, et al. Integrated next-generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury. J Transl Med. 2012; 10:174. [PubMed: 22929533]

58. Cheng J, Torkamani A, Grover RK, Jones TM, Ruiz DI, Schork NJ, Quigley MM, Hall FW, et al. Ectopic B-cell clusters that infiltrate transplanted human kidneys are clonal. Proc Natl Acad Sci USA. 2011; 108:5560–5565. [PubMed: 21415369]

59. Hoon S, Zhou B, Janda KD, Brenner S, Scolnick J. Aptamer selection by high-throughput sequencing and informatic analysis. Biotechniques. 2011; 51:413–416. [PubMed: 22150332]

60. Malyshev DA, Dhami K, Quach HT, Lavergne T, Ordoukhanian P, Torkamani A, Romesberg FE. Efficient and sequence-independent replication of DNA containing a third base pair establishes a functional six-letter genetic alphabet. Proc Natl Acad Sci USA. 2012; 109:12005–12010. [PubMed: 22773812]

61. Lahr DJ, Katz LA. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. Biotechniques. 2009; 47:857–866. [PubMed: 19852769]

62. Schloss PD, Gevers D, Westcott SL. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS ONE. 2011; 6:e27310. [PubMed: 22194782]

63. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011; 27:2194–2200. [PubMed: 21700674]

64. Van Nieuwerburgh F, Thompson RC, Ledesma J, Deforce D, Gaasterland T, Ordoukhanian P, Head SR. Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. Nucleic Acids Res. 2012; 40:e24. [PubMed: 22127871]

65. Peng Z, Zhao Z, Nath N, Froula JL, Clum A, Zhang T, Cheng JF, Copeland AC, et al. Generation of long insert pairs using a Cre-LoxP Inverse PCR approach. PLoS ONE. 2012; 7:e29437. [PubMed: 22253722]

66. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet. 2012; 13:840–852. [PubMed: 23090257]

67. Rivera CM, Ren B. Mapping human epigenomes. Cell. 2013; 155:39–55. [PubMed: 24074860]

68. Northrup DL, Zhao K. Application of ChIP-Seq and related techniques to the study of immune function. Immunity. 2011; 34:830–842. [PubMed: 21703538]

69. Rawlings JS, Gatzka M, Thomas PG, Ihle JN. Chromatin condensation via the condensin II complex is required for peripheral T-cell quiescence. EMBO J. 2011; 30:263–276. [PubMed: 21169989]

70. Adli M, Bernstein BE. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. Nat Protoc. 2011; 6:1656–1668. [PubMed: 21959244]

71. Jayaseelan S, Doyle F, Tenenbaum SA. Profiling post-transcriptionally networked mRNA subsets using RIP-Chip and RIP-Seq. Methods. 2013

72. Chen Y, Varani G. Protein families and RNA recognition. FEBS J. 2005; 272:2088–2097. [PubMed: 15853794]

73. Draper DE. Protein-RNA recognition. Annu Rev Biochem. 1995; 64:593–620. [PubMed: 7574494]

74. Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett. 2008; 582:1977–1986. [PubMed: 18342629]

75. Müller-McNicoll M, Neugebauer KM. How cells get the message: dynamic assembly and function of mRNA-protein complexes. Nat Rev Genet. 2013; 14:275–287. [PubMed: 23478349]

76. Salton M, Elkon R, Borodina T, Davydov A, Yaspo ML, Halperin E, Shiloh Y. Matrin 3 binds and stabilizes mRNA. PLoS ONE. 2011; 6:e23882. [PubMed: 21858232]

77. Sephton CF, Cenik C, Kucukural A, Dammer EB, Cenik B, Han Y, Dewey CM, Roth FP, et al. Identification of neuronal RNA targets of TDP-43-containing ribonucleoprotein complexes. J Biol Chem. 2011; 286:1204–1215. [PubMed: 21051541]

78. Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, et al. Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell. 2010; 40:939–953. [PubMed: 21172659]

79. Hart T, Komori HK, Lamere S, Podshivalova K, Salomon DR. Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics. 2013; 14:778. [PubMed: 24215113]

80. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. CLIP identifies Nova-regulated RNA networks in the brain. Science. 2003; 302:1212–1215. [PubMed: 14615540]

81. Ule J, Jensen K, Mele A, Darnell RB. CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods. 2005; 37:376–386. [PubMed: 16314267]

82. Jensen KB, Darnell RB. CLIP: cross-linking and immunoprecipitation of in vivo RNA targets of RNA-binding proteins. Methods Mol Biol. 2008; 488:85–98. [PubMed: 18982285]

83. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature. 2008; 456:464–469. [PubMed: 18978773]

84. König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol. 2010; 17:909–915. [PubMed: 20601959]

85. Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T, Lin C, Holoch D, Lim C, Tuschl T. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. Methods. 2008; 44:3–12. [PubMed: 18158127]

86. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell. 2010; 141:129–141. [PubMed: 20371350]

87. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009; 462:315–322. [PubMed: 19829295]

88. Gifford CA, Ziller MJ, Gu H, Trapnell C, Donaghey J, Tsankov A, Shalek AK, Kelley DR, et al. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell. 2013; 153:1149–1163. [PubMed: 23664763]

89. Shanmuganathan R, Basheer NB, Amirthalingam L, Muthukumar H, Kaliaperumal R, Shanmugam K. Conventional and nanotechniques for DNA methylation profiling. J Mol Diagn. 2013; 15:17–26. [PubMed: 23127612]

90. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003; 33(Suppl):245–254. [PubMed: 12610534]

91. Cheung HH, Lee TL, Rennert OM, Chan WY. DNA methylation of cancer genome. Birth Defects Res C Embryo Today. 2009; 87:335–350. [PubMed: 19960550]

92. Ehrlich M. DNA hypomethylation in cancer cells. Epigenomics. 2009; 1:239–259. [PubMed: 20495664]

93. Grolleau-Julius A, Ray D, Yung RL. The role of epigenetics in aging and autoimmunity. Clin Rev Allergy Immunol. 2010; 39:42–50. [PubMed: 19653133]

94. Villeneuve LM, Natarajan R. The role of epigenetics in the pathology of diabetic complications. Am J Physiol Renal Physiol. 2010; 299:F14–F25. [PubMed: 20462972]

95. Fouse SD, Nagarajan RO, Costello JF. Genome-scale DNA methylation analysis. Epigenomics. 2010; 2:105–117. [PubMed: 20657796]

96. Huang HC, Zheng S, VanBuren V, Zhao Z. Discovering disease-specific biomarker genes for cancer diagnosis and prognosis. Technol Cancer Res Treat. 2010; 9:219–230. [PubMed: 20441232]

97. Suzuki M, Greally JM. Genome-wide DNA methylation analysis using massively parallel sequencing technologies. Semin Hematol. 2013; 50:70–77. [PubMed: 23507485]

98. Umer M, Herceg Z. Deciphering the epigenetic code: an overview of DNA methylation analysis methods. Antioxid Redox Signal. 2013; 18:1972–1986. [PubMed: 23121567]

99. Shapiro R, Cohen BI, Servis RE. Specific deamination of RNA by sodium bisulphite. Nature. 1970; 227:1047–1048. [PubMed: 5449768]

100. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature. 2008; 452:215–219. [PubMed: 18278030]

101. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008; 133:523–536. [PubMed: 18423832]

102. Diep D, Plongthongkum N, Gore A, Fung HL, Shoemaker R, Zhang K. Library-free methylation sequencing with bisulfite padlock probes. Nat Methods. 2012; 9:270–272. [PubMed: 22306810]

103. Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. Nature. 2013; 502:472–479. [PubMed: 24153300]

104. Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, Balasubramanian S. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science. 2012; 336:934–937. [PubMed: 22539555]

105. Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A, Li X, Dai Q, et al. Base-resolution analysis of 5-hydroxymethylcytosine in the Mamm. Genome Cell. 2012; 149:1368–1380.

106. Song CX, Clark TA, Lu XY, Kislyuk A, Dai Q, Turner SW, He C, Korlach J. Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. Nat Methods. 2012; 9:75–77. [PubMed: 22101853]

107. Raiber EA, Beraldi D, Ficz G, Burgess HE, Branco MR, Murat P, Oxley D, Booth MJ, et al. Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. Genome Biol. 2012; 13:R69. [PubMed: 22902005]

108. Shen L, Wu H, Diep D, Yamaguchi S, D'Alessio AC, Fung HL, Zhang K, Zhang Y. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. Cell. 2013; 153:692–706. [PubMed: 23602152]

109. Song CX, Szulwach KE, Dai Q, Fu Y, Mao SQ, Lin L, Street C, Li Y, et al. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. Cell. 2013; 153:678–691. [PubMed: 23602153]

110. External RNA Controls Consortium. Proposed methods for testing and selecting the ERCC external RNA controls. BMC Genomics. 2005; 6:150. [PubMed: 16266432]

111. Mäder U, Nicolas P, Richard H, Bessieres P, Aymerich S. Comprehensive identification and quantification of microbial transcriptomes by genome-wide unbiased methods. Curr Opin Biotechnol. 2011; 22:32–41. [PubMed: 21074401]

112. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10:57–63. [PubMed: 19015660]

113. Eswaran J, Horvath A, Godbole S, Reddy SD, Mudvari P, Ohshiro K, Cyanam D, Nair S, et al. RNA sequencing of cancer reveals novel splicing alterations. Scientific reports. 2013; 3:1689. [PubMed: 23604310]

114. Zhao C, Waalwijk C, de Wit PJ, Tang D, van der Lee T. RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen Fusarium graminearum. BMC Genomics. 2013; 14:21. [PubMed: 23324402]

115. Röther S, Meister G. Small RNAs derived from longer non-coding RNAs. Biochimie. 2011; 93:1905–1915. [PubMed: 21843590]

116. Saxena A, Carninci P. Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. BioEssays: news and reviews in molecular, cellular and developmental biology. 2011; 33:830–839.

117. Saxena A, Carninci P. Whole transcriptome analysis: what are we still missing? Wiley interdisciplinary reviews. Systems biology and medicine. 2011; 3:527–543. [PubMed: 21197667]

118. Wang H, Chung PJ, Liu J, Jang IC, Kean M, Xu J, Chua NH. Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. Genome Res. 2014 (In press.).

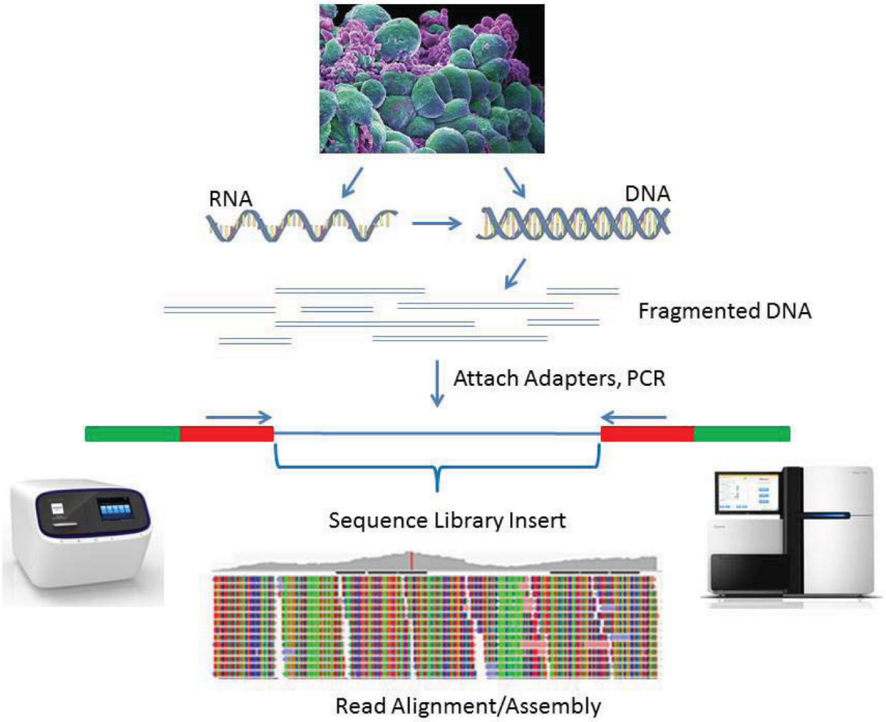

**Figure 1. Basic workflow for NGS library preparation**

RNA or DNA is extracted from sample tissue/cells and fragmented. RNA is converted to cDNA by reverse transcription. DNA Fragments are converted into the library by ligation to sequencing adapters containing specific sequences designed to interact with the NGS platform, either the surface of the flow-cell (Illumina) or beads (Ion Torrent). The next step involves clonal amplification of the library, by either cluster generation for Illumina or microemulsion PCR for Ion Torrent. The final step generates the actual sequence via the chemistries for each technology. One difference between the two technologies is that Illumina allows sequencing from both ends of the library insert (i.e., paired end sequencing). *Cell photograph courtesy of Annie Cavanagh, Wellcome Images.*

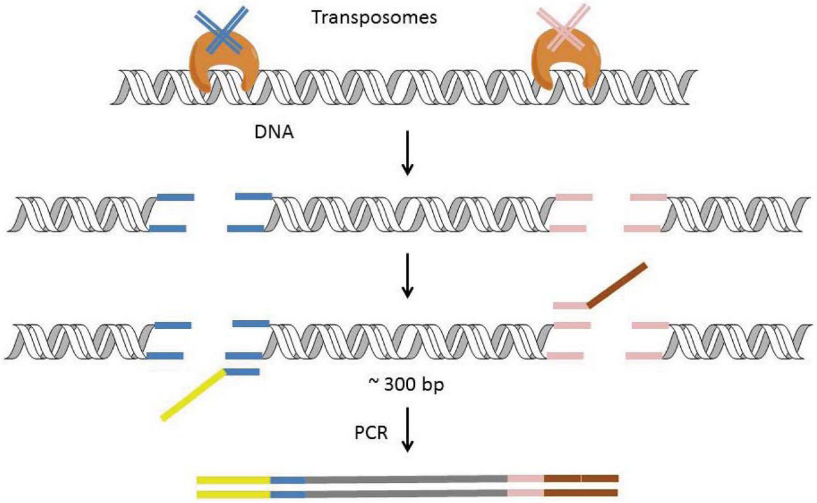

**Figure 2. DNA library preparation using a transposase-based method (Nextera) developed by Illumina**

The transpososome complex comprises an engineered transposase pre-loaded with two double-stranded sequencing adapters. The transpososome simultaneously fragments the DNA and inserts the adapters. The full Illumina adapter sequences are completed during subsequent PCR cycling, after which the library is ready for quantitation and loading onto the flow cell.

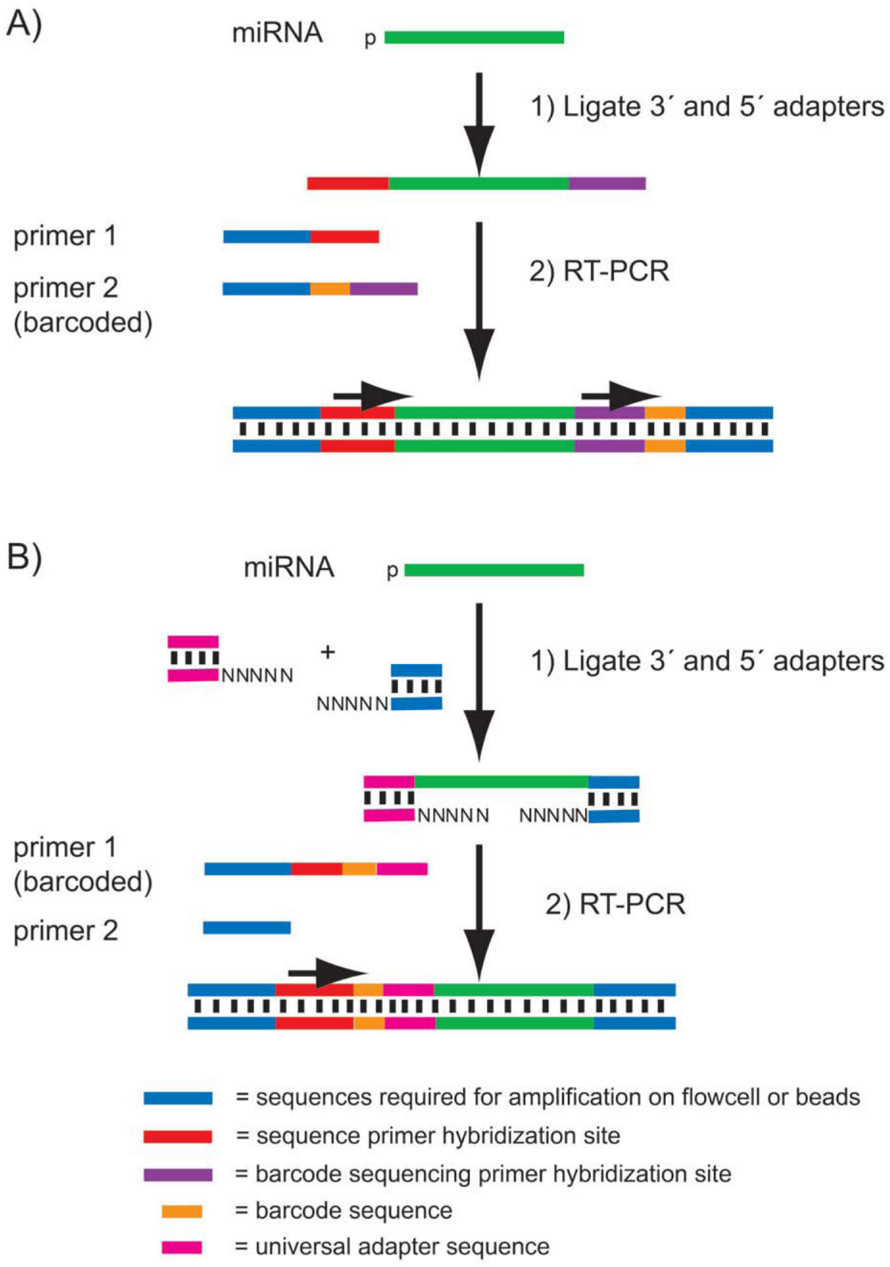
A)
miRNA
p
1) Ligate 3´ and 5´ adapters
primer 1
primer 2
(barcoded)
2) RT-PCR
B)
miRNA
p
NNNNN
+
NNNNN
1) Ligate 3´ and 5´ adapters
NNNNN
NNNNN
primer 1
(barcoded)
primer 2
2) RT-PCR
= sequences required for amplification on flowcell or beads
= sequence primer hybridization site
= barcode sequencing primer hybridization site
= barcode sequence
= universal adapter sequence

**Figure 3. Library preparation workflow for miRNA-seq**

A) The Illumina workflow ligates a 3′ adenylated DNA adapter to the 3′ end of miRNA in a total RNA sample. Then, an RNA adapter is ligated to the 5′ end of the miRNA. The doubled-ligated products are RT-PCR amplified to introduce barcodes for multiplex applications and generate sequencing libraries. The first read sequences the insert miRNA; a second and separate sequencing read is necessary to sequence the barcode. B) Ion Torrent’s workflow uses an RNA ligase to attach 5′ and 3′ adapters composed of hybrid RNA-DNA duplexes. An RT-PCR reaction amplifies the sample and introduces the barcodes to the library construct. In this method, the barcode and the miRNA insert are sequenced in a single read.

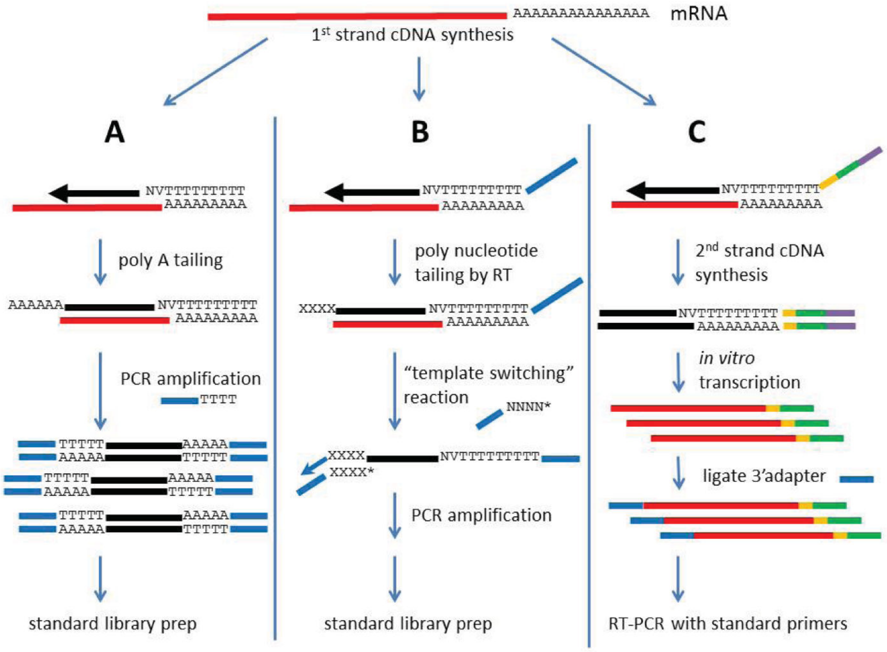

**Figure 4. Approaches for preparing RNA-seq libraries from single cells**

A) Poly-adenylated RNA is reverse transcribed with an anchored oligo-dT primer carrying a universal primer sequence at its 5′ end. Next, poly-nucleotide tailing is used to add a poly(A) tail to the 3′ end of the cDNA. This cDNA can now be amplified with universal PCR primers containing an oligo-dT sequence at the 3′ end. Amplified cDNA can then be used in a standard DNA library construction protocol. B) An anchored oligo-dT primer initiates cDNA synthesis and adds a universal primer sequence. Next, the cDNA is polynucleotide tailed by the RT, producing a 3′ overhanging tail. Template switching is initiated on the 3′ end of the cDNA by hybridization of a second universal primer sequence containing complementary bases at its 3′ end. The template switching oligonucleotide is 3′ blocked (*) to prevent extension by the polymerase, whereas the 3′ end of the cDNA is extended to copy the second universal primer sequence onto the end of the cDNA. The cDNA can now be amplified by PCR. The PCR products created are then taken into a standard library protocol. C) cDNA synthesis is initiated using a barcoded (orange) and anchored oligo-dT primer containing an Illumina adapter sequence (green) and T7 promoter sequence (purple) at the 5′ end. After second strand cDNA synthesis, the fully duplex T7 promoter element is used to initiate in vitro transcription and generate cRNA copies of the cDNA with the 5′ Illumina adapter and barcode. Finally, a second Illumina adapter is ligated to the 3′ end of the cRNA. Doing a final RT-PCR amplification completes the construction of the library.

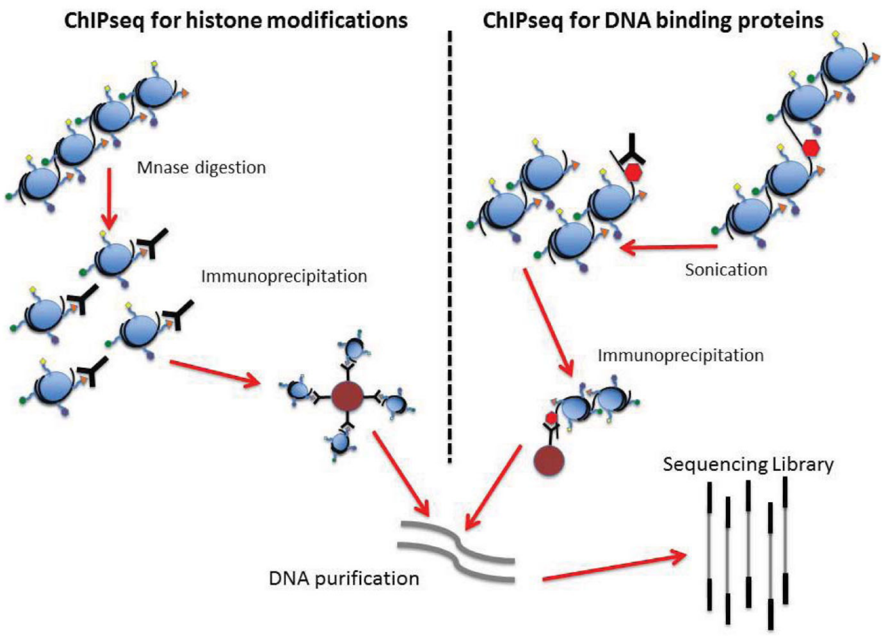

**Figure 5. ChIP-seq procedure for detecting sequences at the sites of histone modifications or the recognition sequences of DNA binding proteins**

Chromatin is crosslinked, fragmented either by micrococcal nuclease digestion or by sonication, and then incubated with antibodies for either the histone modification or protein of interest. Immunoprecipitation is performed using either Protein A or Protein G beads. After washing, the DNA is uncrosslinked, eluted from the beads and purified, at which point the DNA can be taken into standard DNA library construction protocols.

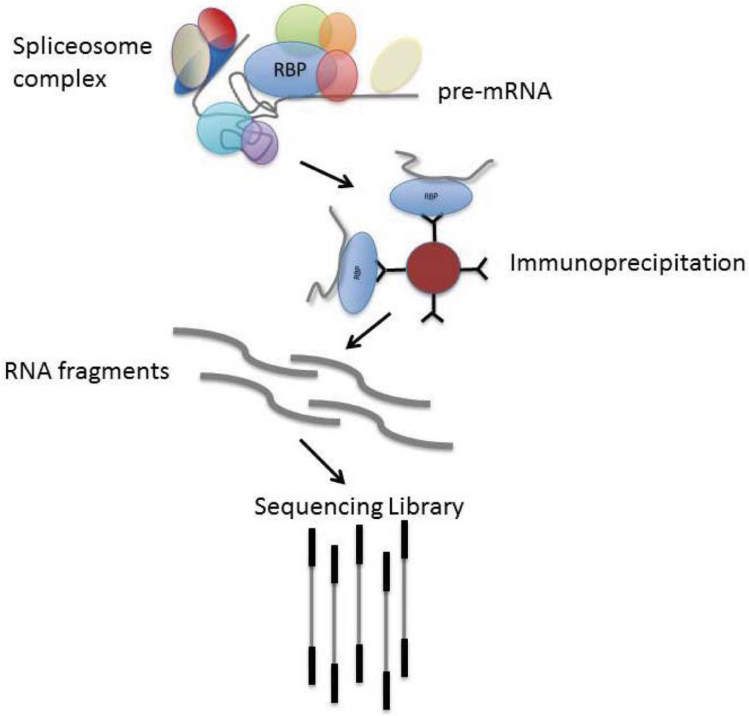

**Figure 6. RNA immunoprecipitation (RIP-seq) done by targeting RNA binding proteins (RBPs)**

The basic principle of RIP-seq is immunoprecipitation of RBPs that are bound to target RNA molecules. The RNA molecules are then purified and a sequencing library is created. In some protocols, the RBP complex is chemically crosslinked to the target RNA; that crosslinking must be reversed after immunoprecipitation. We have found that crosslinking is not necessary for simple RIP-seq where the objective is to identify the RNA molecules bound by RBP, but it is required for CLIP-seq protocols that are used to identify the specific sequence motifs for RBP binding. The immunoprecipitation step can be done with antibodies directed at the specific RBP of interest, or the RBP can be tagged and expressed in the cells under study.

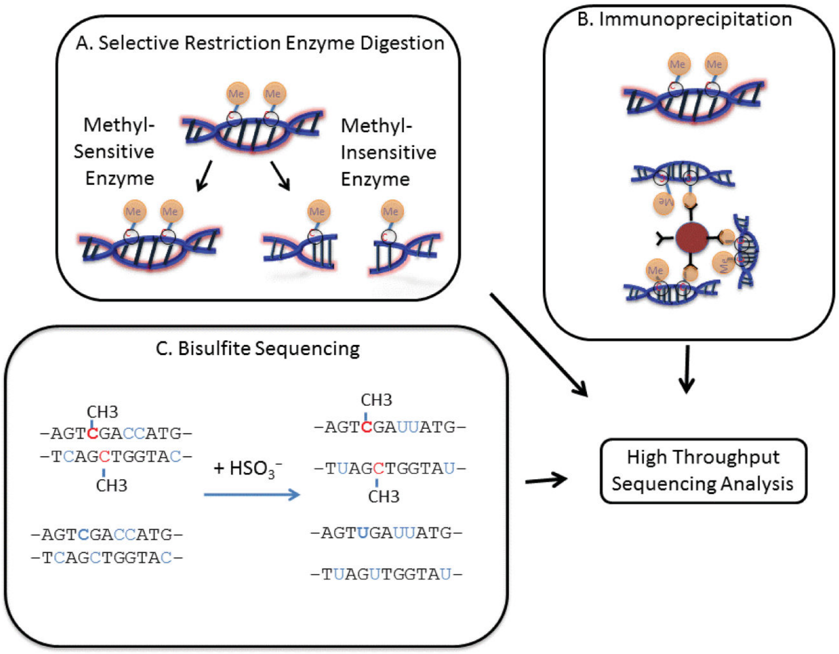

**Figure 7. Approaches for the study of CpG methylation epigenetics (Methylseq)**

A) A combination of methyl-sensitive and methyl-insensitive restriction enzymes can be used to selectively identify and compare the CpG methylation status of specific regions of sequence. B) Antibodies that specifically recognize methylated cytosines can be used to immunoprecipitate DNA fragments, followed by deep sequencing. C) Chemical treatment of DNA with sodium bisulfite results in the conversion of unmethylated cytosines to uracils. In contrast, methylated cytosines are protected. Subsequently, deep sequencing of these libraries reveals the methylation status of individual nucleotides.

## Table 1

Approaches for RNAseq.

| Objective | Principles of approach | References |
|---|---|---|
| **Gene expression** | Target poly(A) mRNAs (enrich or selectively amplify). To quantify expression new methods are available based on 3′ sequence tags or combinatorial barcodes to remove duplicate reads. Short read runs (50–100 bp) concentrating on 3′ sequence can be sufficient and save considerable resources. One option is to spike in the ERCC synthetic standards for quantification (110). | (36,111,112) |
| **Alternative splicing** | Target exon/intron boundaries by either doing long read sequencing (>300 bp) or paired end read sequencing (≥ 2 × 100). In the case of paired end sequencing, the insert size is typically larger and/or variable in size. | (113,114) |
| **miRNA (or small RNAs)** | Target short reads using size selection purification because miRNAs are in the 18–23 bp range. piRNAs, snoRNAs, tRNAs are all under 100 bps. | (115) |
| **Non-coding RNA** | Directional RNA sequencing is critical. | (116,117) |
| **Anti-sense RNA** | Consider combining mRNA expression with directional sequencing to reveal the subset of transcripts representing the anti-sense orientation and correlate these with gene expression changes. | (30,118) |
| **Single cell sequencing** | Requires special strategies to start with picogram quantities of input RNA and allow extensive whole transcriptome amplification.<br>Critical challenge is the technical noise created by amplification. | (3,5,7,10,11) |