Review Article

Library preparation methods for next-generation sequencing: Tone down the bias

Erwin L. van Dijk\textsuperscript{a,}, Yan Jaszczyszyn\textsuperscript{b}, Claude Thermes\textsuperscript{a}

\textsuperscript{a}Centre de Génétique Moléculaire – CNRS, Avenue de la Terrasse, 91198 Gif sur Yvette, France
\textsuperscript{b}Plateforme Intégrée IMAGIF – CNRS, Avenue de la Terrasse, 91198 Gif sur Yvette, France

Article Information

Article Chronology:
Received 25 October 2013
Received in revised form 7 January 2014
Accepted 8 January 2014
Available online 15 January 2014

Keywords:
Next-generation sequencing
Library preparation
Bias
DNA-seq
RNA-seq

Abstract

Next-generation sequencing (NGS) has caused a revolution in biology. NGS requires the preparation of libraries in which (fragments of) DNA or RNA molecules are fused with adapters followed by PCR amplification and sequencing. It is evident that robust library preparation methods that produce a representative, non-biased source of nucleic acid material from the genome under investigation are of crucial importance. Nevertheless, it has become clear that NGS libraries for all types of applications contain biases that compromise the quality of NGS datasets and can lead to their erroneous interpretation. A detailed knowledge of the nature of these biases will be essential for a careful interpretation of NGS data on the one hand and will help to find ways to improve library quality or to develop bioinformatics tools to compensate for the bias on the other hand. In this review we discuss the literature on bias in the most common NGS library preparation protocols, both for DNA sequencing (DNA-seq) as well as for RNA sequencing (RNA-seq). Strikingly, almost all steps of the various protocols have been reported to introduce bias, especially in the case of RNA-seq, which is technically more challenging than DNA-seq. For each type of bias we discuss methods for improvement with a view to providing some useful advice to the researcher who wishes to convert any kind of raw nucleic acid into an NGS library.

Contents

Introduction ......................................................................................... 13
DNA-sequencing .................................................................................. 13
  Fragmentation .................................................................................. 13
  Size selection .................................................................................. 13
  PCR (1): improvements of existing methods ........................................... 14
  PCR (2): alternative methods for library amplification ....................... 15
RNA-sequencing .................................................................................. 15
  rRNA depletion ............................................................................. 17
  RNA fragmentation ........................................................................ 17

© 2014 Elsevier Inc. All rights reserved.
**Introduction**

Over the recent years, NGS technology has become an essential tool for nearly all fields of biological research [1]. It enables parallel sequencing of millions of small DNA fragments for low per-base costs in a short time. Besides genome (re)sequencing, NGS provides accurate information on the composition of complex (c)DNA samples, making it the method of choice for most, if not all, genomic applications, such as transcriptome analysis (RNA-Seq), metagenomics, or profiling of methylated DNA (MeDip-seq), or DNA-associated proteins (ChIP-Seq). New applications for NGS appear frequently and furthermore, NGS applications are under intense scrutiny to produce even more and better quality data.

NGS requires the conversion of the source nucleic acid material into standard libraries suitable for loading onto a sequencing instrument. A wide variety of NGS library preparation protocols exist, but they all have in common that (fragments of) DNA or RNA molecules are fused with adapters that contain the necessary elements for immobilization on a solid surface and sequencing. In addition, size selection steps are often performed and libraries are usually amplified by PCR (Fig. 1). It is well understood that the quality of sequencing data depends highly upon the quality of the sequenced material. Therefore, the library construction process should guarantee a high molecular recovery of the original fragments (low bias and high complexity) in order to achieve the most genomic coverage with the least amount of sequencing. Importantly however, most popular library preparation protocols being used today may introduce serious biases in sample composition, which poses technical challenges and may lead to the misinterpretation of NGS data.

General methods of NGS sample preparation have been discussed previously in an excellent review [2]. Important new insights on biases in library preparation have appeared since, and here we review the current literature on biases in NGS libraries for ‘DNA-seq’ applications (e.g. genomic DNA-seq, ChIP-seq, exome sequencing) or ‘RNA-seq’ applications (e.g. transcriptome analysis, small RNA-seq). The focus will be on solutions to reduce bias and to ameliorate library quality for the Illumina platform, but most of the general principles will also apply to the other systems.

**DNA-sequencing**

The starting material for DNA-sequencing (DNA-seq) is generally double-stranded DNA in the form of isolated genomic DNA or chromatin (ChIP-seq). This DNA or chromatin is fragmented, followed by immunoprecipitation and removal of DNA-bound proteins in the case of ChIP-seq. These steps are followed by end-repair and adapter ligation, and usually a size selection step to remove free adapters and to select molecules in the desired size range (Fig. 1). Next, PCR amplification is often performed to generate sufficient quantities of template DNA to allow accurate quantification and to enrich for successfully adaptered fragments. PCR can also serve to add additional adapter sequence using tailed primers, resulting in template molecules that contain all the necessary elements for bridge amplification on the flowcell surface and for sequencing. Below we will discuss the different steps of the workflow that have been implicated in bias introduction. These are the steps of fragmentation, size selection and especially PCR; end-repair and adapter ligation do not appear to introduce detectable bias [3]. A summary with suggestions for improvement is shown in Table 1.

**Fragmentation**

DNA shearing is typically achieved either by mechanical force through nebulization or sonication, or by enzymatic digestion. Whereas fragmentation of naked DNA has not been considered a major source of bias, chromatin sonication for ChIP-seq has been shown to be non-random, with euchromatin being sheared more efficiently than heterochromatin [4]. As a result, DNA fragments of the selected size for library preparation (~200 base pairs) will preferentially contain euchromatin DNA, while heterochromatin is underrepresented. To solve this problem, Mokry and colleagues developed a double-fragmentation ChIP-seq protocol [5]. After conventional crosslinking and immunoprecipitation, chromatin is de-crosslinked and sheared a second time to concentrate fragments in the optimal size range for NGS. This approach not only reduces bias against heterochromatin DNA, but also increases the yield of material.

**Size selection**

Many current protocols use solid-phase reversible immobilization (SPRI) beads, sold by Beckman Coulter as AMPure beads, for size selection. The use of SPRI beads provides a fast and efficient method to enrich for DNA molecules of a selected size range. However, gel extraction is still commonly used as it allow for a more precise size selection. Quail and colleagues identified that melting agarose gel slices by heating to 50 °C in chaotropic salt buffer decreased the representation of AT-rich sequences. This possibly reflects a higher affinity of spin columns for double-stranded DNA, as strands with a high AT content are most likely to become denatured during this step and may not re-anneal [6]. A simple solution to this problem is to melt the gel slices in the
supplied buffer at room temperature (18–22°C), considerably reducing GC bias.

**PCR (1): improvements of existing methods**

PCR amplification is known to introduce bias in sample composition, due to the fact that not all fragments in the mixture are amplified with the same efficiency. GC-neutral fragments are amplified more efficiently than GC-rich or AT-rich fragments, and as a result fragments with high AT- or GC content may become underrepresented or are completely lost during library preparation [7]. This effect severely complicates the sequencing of extremely AT-rich genomes, such as that of a human malaria parasite (*Plasmodium falciparum*—average AT content, ~80% [8]), or extremely GC-rich genomes such as the herpes B virus (average GC content, ~75% [9]), or profiling of highly GC-rich CpG islands of cancer genomes. PCR bias can also be a great obstacle towards accurately evaluating community structure in metagenomics studies as it causes errors in the estimation of the mean relative abundance of each detected operational taxonomic unit [10].

Various solutions have been proposed to reduce PCR bias. A library preparation protocol that avoids PCR amplification altogether was initially proposed by Kozarewa and colleagues [11]. Contrary to the original Illumina protocols, these authors proposed to ligate adapters that contain all necessary elements for bridge amplification on Illumina flowcells, eliminating the need for PCR to add these sequences afterwards. Indeed, the current Illumina TruSeq adapters are composed in this way, and PCR-free library preparation kits are now available from Illumina. This protocol effectively reduces bias, but a drawback is that is relatively large quantities (>1 μg) of input material are required. Thus, if input amounts of DNA are low, some form of unbiased amplification is necessary.

Several laboratories have compared different PCR polymerases and conditions to minimize amplification bias. The most comprehensive analysis was performed by Quail and colleagues [12]. These authors tested a large series of polymerases for their ability to amplify libraries from microbial genomes with GC contents ranging from 19.3% to 67.7%. Strikingly, many of these enzymes performed better than Phusion, the standard polymerase for Illumina library preparation. The best overall enzyme was Kapa HiFi (Kapa Biosystems); genome coverage using this enzyme was far more uniform than that with Phusion and was remarkably close to results achieved without PCR for all tested GC contents. Three other reports confirm that Phusion polymerase has a relatively strong GC bias [3,13,14]. Dabney and Meyer in addition
MDA is based on the observation that the is termed multiple displacement amplification (MDA) [17]. The subsequent steps differ among protocols of fragmenting chromatin and/or when a large number of amplification cycles is needed it is important to choose a different PCR polymerase. The best overall performing polymerase appears to be Kapa HiFi. Additives such as betaine or TMAC may help to further improve coverage of AT-base-pair regions. However, it is not completely unbiased and error-free [20]. One issue with the amplification of single genomes is uneven coverage due to the stochastic priming and amplification at the early stages of the reaction, causing variation in the amplification of different regions. A possible solution is the use of molecular crowding agents such as trehalose or PEG, which can provide more homogeneous amplification via the volume reduction effect. Therefore, MDA has become the method of choice for whole genome amplification (WGA) from single cells [19]. MDA has been reported to perform better than PCR for WGA, even though it is not completely unbiased and error-free [20].

In summary, alternative methods for DNA amplification have been developed that may perform better than standard Phusion polymerase for the amplification of genomes with extreme base composition. However, the best overall performing method may be classical PCR amplification with Kapa HiFi polymerase. In the extreme case of small input amount, the single cell, MDA may be the preferred amplification method.

### RNA-seq

RNA-seq methodologies have provided extremely powerful tools for qualitative and quantitative transcriptome analysis in both prokaryotes and eukaryotes. RNA-seq procedures usually start with ribosomal RNA (rRNA) depletion or mRNA enrichment, as the large majority of cellular total RNA is rRNA. For eukaryotic transcriptomes, polyadenylated mRNAs are commonly extracted using oligo-dT beads, or alternatively rRNAs can be selectively depleted (see below: “rRNA depletion”). For prokaryotes, as their mRNAs are not polyadenylated, there is only the second option. Subsequently, RNA is often fragmented to generate reads that cover the entire length of the transcripts (in some protocols fragmentation occurs after double stranded cDNA synthesis, see below). The subsequent steps differ among protocols. The standard Illumina approach relies on randomly primed

### Table 1 – Sources of bias in DNA-seq library preparation and suggestions for improvement.

<table>
<thead>
<tr>
<th>Description</th>
<th>Suggestions for improvement</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fragmentation of chromatin</td>
<td>Second fragmentation after IP to concentrate fragments in the optimal size range</td>
<td>Mokry et al 2010 [5]</td>
</tr>
<tr>
<td>Bias due to more efficient sonication of euchromatin than heterochromatin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size selection</td>
<td>Melt agarose gel slices at room temperature</td>
<td>Quail et al 2008 [6]</td>
</tr>
<tr>
<td>Bias due to heating agarose gel slices in chaotropic salt buffer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias due to preferential amplification of templates with neutral GC</td>
<td>2. For AT-rich genomes, use lower extension temperatures and/or use the PCR additive TMAC</td>
<td>Oyola et al 2012 [13]</td>
</tr>
<tr>
<td></td>
<td>3. For GC-rich genomes, use extended denaturation times and/or use the additive betaine</td>
<td>Aird et al 2011 [3]</td>
</tr>
<tr>
<td></td>
<td>4. For the amplification of minute quantities of genomic DNA (single cell), use MDA rather than PCR</td>
<td>Dabney et al 2012 [14]</td>
</tr>
</tbody>
</table>

### PCR (2): alternative methods for library amplification

In addition to PCR, various other methods have been developed to amplify DNA. One of these, recombinase polymerase amplification (RPA), is like PCR, an exponential amplification method, but it is based on a strand-displacement polymerase that operates at low temperature [15]. Another method is linear amplification for deep sequencing (LADS), which is an in vitro transcription system whereby the T7 polymerase transcribes DNA molecules of a library multiple times under isothermal conditions followed by reverse transcription, resulting in a linearly amplified library [16]. Both methods appear to perform better than standard Phusion polymerase amplification for templates with extreme base composition. However, they both generate relatively high levels of chimerics and duplicate reads, and classical PCR with Kapa HiFi was found to perform better for overall library amplification [13].

A different linear amplification method that is worth mentioning is termed multiple displacement amplification (MDA) [17]. MDA is based on the observation that the α29 DNA polymerase, which was originally used for isothermal rolling circle amplification of circular DNA, can also amplify linear genomic DNA in a cascading strand displacement reaction [18]. MDA is an extremely powerful amplification method, allowing microgram quantities of DNA to be obtained from femtograms of starting material. For this reason, MDA has become the method of choice for whole genome amplification (WGA) from single cells [19]. MDA has been reported to perform better than PCR for WGA, even though it is not completely unbiased and error-free [20]. One issue with the amplification of single genomes is uneven coverage due to the stochastic priming and amplification at the early stages of the reaction, causing variation in the amplification of different regions.

In summary, alternative methods for DNA amplification have been developed that may perform better than standard Phusion polymerase for the amplification of genomes with extreme base composition. However, the best overall performing method may be classical PCR amplification with Kapa HiFi polymerase. In the extreme case of small input amount, the single cell, MDA may be the preferred amplification method.
double-stranded cDNA synthesis followed by the same steps as done for DNA-seq, i.e. end-repair, ligation of dsDNA adapters, and PCR amplification (Fig. 2). A major drawback of this method is that strand information is not preserved. For this reason, many different methods have been developed for strand-specific RNA-seq, which fall into two main classes. One class of methods relies on marking one strand by chemical modification. The dUTP second strand marking method follows basically the same procedure as the classical protocol except that dUTP is incorporated during second strand cDNA synthesis, preventing this strand from being amplified by PCR. Most current transcriptome library preparation kits follow the dUTP method. The second class of strand-specific methods relies on attaching different adapters in a known orientation relative to the 5’ and 3’ ends of the RNA transcript. The Illumina ligation method is a well-known example of this class and is based on sequential ligation of two different adapters. Most current small RNA library preparation kits follow the RNA ligation method.

RNA-seq protocols are technically more challenging than DNA-seq protocols and are often biased procedures. Common types of bias include low complexity (many reads with the same starting point), uneven coverage across different regions of transcription units, and antisense artifacts in the case of stranded libraries.
In addition, aberrant measurements of transcript levels may occur. A series of strand-specific RNA-seq methods was tested or these types of bias in a comprehensive comparative analysis [25]. The dUTP second strand marking and the Illumina RNA ligation methods were identified as the best overall performing protocols. The SMART, Hybrid, and especially the NSR methods were found to be relatively biased procedures, with the SMART method having a particularly high level of antisense artefacts.

In the following section we will examine the sources of bias in the various steps of RNA-seq protocols. A summary with suggestions for improvement is presented in Table 2.

**rRNA depletion**

Two main categories of rRNA depletion methods exist: (1) subtractive hybridization with probes specific for rRNA or mRNA (oligo-dT, for eukaryotes) or (2) selective degradation of rRNAs and other 5’ monophosphate RNAs by exonucleases, while mRNAs are protected by their 5’ cap structures, or triphosphates in prokaryotes. A comparative study identified subtractive hybridization using rRNA-specific probes as the method that introduced the least bias in relative transcript abundance [26]. A consistent drop in mRNA fidelity was observed with the exonuclease method, potentially due to exonuclease targeting partially degraded mRNAs that have lost their protective 5’ moiety. This notion was supported by the observation that highly expressed genes with short half-lives were preferentially lost after exonuclease treatment. In addition, exonuclease treatment tends to be less efficient in rRNA depletion than subtractive hybridization, presumably due to stable secondary structures that block exonuclease progression (personal communication).

**RNA fragmentation**

For longer transcripts, a fragmentation step is necessary to obtain reads covering the entire length of the RNAs. While in the original dUTP protocol double-stranded cDNAs generated from intact RNA are fragmented [22], the Illumina ‘TruSeq Stranded mRNA Sample Prep Kit’ (based on the dUTP protocol) as well as the standard Illumina protocol and the RNA ligation method use RNA fragmentation prior to cDNA synthesis. Different techniques of RNA fragmentation have been described, among which RNase III digestion and chemical zinc-induced hydrolysis. The Illumina protocols use zinc-mediated cleavage, while other protocols, such as the “Total RNA Seq kit” (Life Technologies), are based on RNase III. In contrast to zinc-mediated RNA cleavage, which has no sequence or structure specificity, RNase III is known to cleave double-stranded RNA in a sequence-and structure-specific manner [27]. A recent study compared the effects of RNase III and zinc fragmentation on transcript level quantification and reassembly from RNA-seq data [28]. RNA cleavage by RNase III was less homogeneous than zinc cleavage, leading to a reduced autocorrelation between neighboring nucleotides across the transcriptome. Although measurement of mRNA abundance led to similar results with the two fragmentation methods, two major classes of abundant non-coding transcripts were found to be underrepresented when RNase III was used. Importantly, transcript computational reassembly was severely affected when RNase III was used, leading to a reduced number of reassembled transcripts, with less well defined 5’ and 3’ ends. Together, the results show that chemical zinc-mediated cleavage is more appropriate than RNase III for transcriptome reassembly from RNA-seq data.

**Random primer hybridization**

Various methods, including the classical Illumina protocol and the dUTP protocol, use random primers to initiate cDNA synthesis.

### Table 2 – Sources of bias in RNA-seq library prep and suggestions for improvement.

<table>
<thead>
<tr>
<th>Description</th>
<th>Suggestion for improvement</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA extraction using Trizol: selective loss of GC poor or highly structured small RNAs at low RNA concentrations</td>
<td>Use similar RNA concentrations for samples that are to be compared or avoid Trizol extraction altogether. Use alternative protocols such as the MirVana miRNA isolation kit.</td>
<td>Kim et al., 2012 [43]</td>
</tr>
<tr>
<td>Ribosomal RNA (rRNA) depletion/ mRNA enrichment: bias due to exonuclease targeting partially degraded mRNAs</td>
<td>Use subtractive hybridization rather than exonuclease treatment to deplete rRNAs</td>
<td>He et al., 2010 [26]</td>
</tr>
<tr>
<td>RNA fragmentation by RNase III: not completely random, leading to reduced complexity</td>
<td>Use chemical treatment (e.g. zinc) rather than RNase III for RNA fragmentation</td>
<td>Wery et al., 2013 [28]</td>
</tr>
<tr>
<td>Random hexamer priming bias: not completely random</td>
<td>A read count reweighing scheme was proposed that adjusts for the bias and makes the distribution of reads more uniform</td>
<td>Hansen et al., 2010 [29]</td>
</tr>
<tr>
<td>Reverse transcription: antisense artefacts (especially in the SMART and the NSR methods)</td>
<td>Add actinomycin D to the reaction (not possible for the SMART method)</td>
<td>Perocchi et al., 2007 [38]</td>
</tr>
<tr>
<td>Adapter ligation bias due to substrate preferences of T4 RNA ligases</td>
<td>Use adapters with random nucleotides at the extremities to be ligated.</td>
<td>Levin et al., 2010 [25], Jayaprakash et al., 2011 [34], Sorefan et al., 2013 [31], Sun et al., 2011 [35], Zhuang et al., 2012 [33], Munro and Robb, 2010 [46]</td>
</tr>
<tr>
<td>Reduced ligation efficiency due to RNA modifications</td>
<td>If necessary, perform the appropriate enzymatic treatments to generate a 5’ monophosphate and a 3’ OH. For RNAs with a 2’O-methyl modification use adjusted ligation reaction conditions</td>
<td></td>
</tr>
</tbody>
</table>
Unlike the name suggests, a study by Hansen and colleagues [29] demonstrates that random primer hybridization is not random, but instead shows a preference for specific nucleotide compositions. This leads to a significant bias in the nucleotide composition at the start of sequencing reads, which influences the uniformity of the location of the reads along expressed transcripts. This may explain the observation that NSR libraries have low complexity (see above; [25]). A biochemical solution has not (yet) been found, but the authors propose a bioinformatics tool in the form of a reweighing scheme that adjusts for the bias and makes the distribution of the reads more uniform.

**Adapter ligation**

In the Illumina RNA ligation method, RNA fragments are ligated with an pre-adenylated 3′ adapter using a truncated form of T4 RNA ligase 2 (Rnl2), followed by ligation of a 5′ adapter using T4 Rnl1. The ligation products are subsequently reverse transcribed and amplified by PCR. This method is also widely used for small RNA library preparation (see below). Rnl1 and Rnl2 are two different families of RNA end-joining enzymes and have been shown to have different substrate specificities. Initially, a SELEX-type approach was used to show that Rnl1 prefers single stranded substrates [30]. More recently, Sorefan and colleagues [31] followed a similar approach to characterize substrate preferences of Rnl1 and Rnl2; randomized 21 nt RNA oligonucleotides were used to prepare a library following the Illumina RNA ligation method. After sequencing, secondary structures of the 1000 most abundant sequences with the 5′ and 3′ adapters were predicted and compared with the predicted structures of 1000 random sequences. At the ligation site between the 5′ adapter and the random RNA, only 25% of the obtained sequences was predicted to base-pair, against 49% in the control data set. This result confirms a preference for single stranded ligation sites for Rnl1. In contrast, the same analysis performed for Rnl2 revealed a preference of Rnl2 for single stranded nucleotides downstream of the ligation site but double stranded nucleotides upstream of the ligation site.

It can be anticipated that these substrate preferences will cause bias, and indeed the RNA ligation method yields less even and continuous coverage than the dUTP method [32], presumably due to this issue. Several recent publications confirm this ligation bias [31,33–36]. As a solution, several groups propose to randomize the 3′ end of the 5′ adapter and the 5′ end of the 3′ adapter [33–35]. The strategy is based on the hypothesis that a population of degenerate adapters would average out the sequencing bias because the slightly different adapter molecules would form stable secondary structures with a more diverse population of RNA sequences. Indeed, the above-mentioned independent studies claim that such degenerate adapters yield less biased libraries, which lends support to this hypothesis.

**Reverse transcription**

All common RNA-seq protocols involve reverse transcription. A known feature of reverse transcriptases is that they have a tendency to generate spurious second strand cDNA based on their DNA-dependent DNA polymerase activity [37]. This can confound sense versus antisense transcript determination and may thus cause problems when strand-specificity is required. The dUTP and the RNA ligation protocols do not appear to be significantly affected by this phenomenon, but other widely used protocols do contain antisense artifacts, most notably the SMART method [25]. Actinomycin D, a compound that specifically inhibits DNA-dependent DNA synthesis, has been proposed as an agent to eliminate antisense artifacts [38]. Indeed, actinomycin D improved the strand-specificity of the NSR method ([25]; see above). Unfortunately however, actinomycin D cannot be used for the SMART method, because it inhibits both second-strand synthesis and template switching [39].

**PCR amplification**

The problem of PCR bias and a discussion of methods to improve PCR fidelity has been described above. What is worth adding here is that an amplification-free RNA-seq protocol has been developed [40]. In this method, RNA fragments are ligated with adapters as in the RNA ligation method. Subsequently however, ligation product are directly loaded onto an Illumina flowcell (this requires a modified 3′ adapter that contains a sequence complementary to the oligonucleotides on the flowcell) and reverse transcription takes place in situ. Next, a standard cluster amplification step is carried out. Thus, the PCR amplification step during library preparation is avoided.

**Small RNA-sequencing**

Technical issues with small RNA library preparation have been discussed in an excellent recent review [41]. In addition, as the most common small RNA library preparation protocols follow the RNA ligation method, the same RNA ligation bias problems as discussed above apply to small RNA library preparation. Two additional issues that are important to mention here are: (I) bias due to RNA extraction procedures, and (II) bias due to RNA modifications.

**RNA extraction**

In an intriguing recent paper, Kim and colleagues report that their original observation that a specific subset of microRNAs is decreased when cells are grown at low density [42] is actually due to RNA extraction artefacts [43]. They find that at low RNA concentrations, specific miRNAs are lost during classical RNA extractions using Trizol reagent. These miRNAs share as common features that they are either GC poor or highly structured. This leads the authors to hypothesize that as miRNAs are precipitated inefficiently, they require large RNAs as carriers. The miRNAs interact with large RNAs through base-pairing and highly structured or GC poor miRNAs base-pair less efficiently with large RNAs. As a result, these RNAs are selectively lost at low RNA concentrations. The important take home message from this work is that “total” RNA preparations may actually be biased due to unequal precipitation efficiencies. If miRNA levels are to be compared among samples, it is recommended to use similar amounts of material of each sample for RNA extraction. Another option is to avoid Trizol extraction and to use the MirVana kit (Ambion – Life Technologies) for example to extract miRNAs.
RNA modifications

T4 RNA ligases can efficiently catalyze the formation of a 3' to 5' phosphodiester bond between a 3' hydroxyl group and a 5' monophosphate group [44]. Mature miRNAs and siRNAs from mammals, nematodes and insects have a monophosphate at their 5' end and 2', 3' hydroxyl groups at their 3' ends [41]. However, siRNAs from insects and plants and miRNAs from plants have a 2'-O-methyl modification at their 3' ends [45]. RNAs with this modification are ligated with a strongly reduced efficiency, resulting in an under-representation in small RNA libraries [46,47]. However, modified reaction conditions, using optimal PEG concentrations and overnight incubation at 16 °C instead of one hour at 28 °C can significantly improve 3' adapter ligation of such modified RNA to a level equivalent to that of unmodified RNAs [46].

Conversely, some small RNAs may not have a 5' monophosphate; for instance, secondary siRNAs in nematodes have a triphosphate at their 5' end [48]. Such RNAs can be treated with alkaline phosphatase to yield a 5' hydroxyl group, followed by poly nucleotide kinase treatment yielding a 5' monophosphate. Alternatively, some small RNA species may contain a 5' cap structure [49]. Such RNAs can be treated with tobacco acid pyrophosphatase (TAP), which hydrolyzes the phosphoric acid anhydride bonds in the triphosphate bridge of the cap structure, releasing the cap nucleoside and generating a 5' monophosphate terminus on the RNA molecule [50].

Summary and practical considerations

Library preparation protocols are often biased procedures, which is a major concern for NGS data quality. A notorious source of bias is PCR amplification. This is especially a problem for the sequencing of extreme genomes or when absolute or relative quantification within a single sample is required, such as in metagenomics studies. PCR free protocols have been developed, but these require large amounts of input material. Amplification is therefore necessary when amounts of input material are limited. A comprehensive analysis revealed significant differences in bias introduction among polymerases [12]. Kapa HiFi was found to be the best overall performing enzyme; Kapa-amplified libraries had genome coverage close to PCR-free libraries. This enzyme is therefore a better choice than the traditional Phusion polymerase.

While PCR is the principal source of bias in DNA-seq, in RNA-seq protocols practically all steps are potential sources of bias. Good care should therefore be taken when choosing a protocol or a kit for a given RNA-seq application. The dUTP second strand marking method was identified as the best performing strand-specific protocol, closely followed by the Illumina RNA ligation method [25]. It is therefore not surprising that most commercially available transcriptome analysis kits follow the dUTP protocol, and small RNA library preparation kits are often based on the RNA ligation protocol. Nevertheless, even though most current kits are based on the same principle, some important differences exist. First, not all kits use the same reagent for RNA fragmentation. For example the “Total RNA Seq” kit from Life Technologies uses RNase III, which leads to lower complexity libraries than chemical fragmentation. Second, kits based on RNA ligation use different adapters and different ligation reaction conditions. In addition, the presence of PEG in the 3' ligation reaction reaction should improve ligation with 2'-O-methyl RNAs. Third, kits differ in the use of PCR polymerases. While the Illumina kits use Phusion polymerase, other kits are based on different polymerases that may introduce less bias. It would be good to check if the polymerase from a given kit has been included in the comparative analyses described above, and if necessary replace this polymerase by Kapa HiFi.

In conclusion, recent reports provide important information on how to ameliorate library quality. Simple modifications of existing protocols are often sufficient. While future work will continue to bring further improvements, important progress has been made, allowing one to prepare libraries that more faithfully represent the starting nucleic acid material and that yield higher quality NGS data.

Reference


