SURVEY AND SUMMARY

Capturing the ‘ome’: the expanding molecular toolbox for RNA and DNA library construction

Morgane Boone1,2,*, Andries De Koker1,2 and Nico Callewaert1,2,*

1Center for Medical Biotechnology, VIB, Zwijnaarde 9052, Belgium and 2Department of Biochemistry and Microbiology, Ghent University, Ghent 9000, Belgium

Received November 05, 2017; Revised February 05, 2018; Editorial Decision February 22, 2018; Accepted February 23, 2018

ABSTRACT

All sequencing experiments and most functional genomics screens rely on the generation of libraries to comprehensively capture pools of targeted sequences. In the past decade especially, driven by the progress in the field of massively parallel sequencing, numerous studies have comprehensively assessed the impact of particular manipulations on library complexity and quality, and characterized the activities and specificities of several key enzymes used in library construction. Fortunately, careful protocol design and reagent choice can substantially mitigate many of these biases, and enable reliable representation of sequences in libraries. This review aims to guide the reader through the vast expanse of literature on the subject to promote informed library generation, independent of the application.

INTRODUCTION

Next generation sequencing technologies have undeniably changed the scientific landscape in biology. The fast-paced methodological progress driving many of the developments in the field has not only been the result of exceptional advances in sequencing chemistry, detection systems and data-processing or analysis methods (1), but also of innovations in the area of sequencing library construction. The paramount role of library construction is often underappreciated, yet it shapes both outcome and inference: the library protocol should meticulously capture the specific molecules of interest, yet minimize unwanted fragments or biases in order to ensure accurate interpretation (‘garbage in is garbage out’). Additionally, a higher quality library usually maximizes the useful sequencing read output and facilitates data processing. Indeed, in the past few years, the number of studies reporting (and in many cases, addressing) the impact of the choice of specific enzymes, reagents, reaction conditions or overall protocols on the resulting library quality have grown exponentially, and there is renewed interest in the development of molecular biology tools designed to overcome these biases.

In addition to libraries for sequencing purposes, many proteome-wide functional assays, for instance assessing protein interactions (2,3), protein localization (4), post-transcriptional regulation (5) or drug activity (6), also rely on pooled or arrayed nucleic acid libraries as input. Fortunately, some of these libraries can now be accurately synthesized at relatively low cost, or one can rely on available collections of full-length and validated open reading frames (ORFs) on plasmids (7), short hairpin or small interfering RNA libraries (8) and guide RNA libraries for CRISPR screens (9). In several other cases, however, such as for very large libraries or libraries with custom requirements, high-quality libraries still need to be generated. Coding sequence fragment libraries are a prominent example (10–13).

Many researchers can (and do) resort to the use of commercial kits to capture the desired nucleic acid species into a workable library of molecules. While there are numerous suppliers for sequencing library construction, and the resulting libraries are often of reasonable quality for standard sequencing experiments (e.g. transcriptome sequencing), it is generally acknowledged that these conventional procedures allow little room to tailor the library toward the specific needs of the researcher, especially when the research question calls for a non-standard approach. Additionally, there is always a lag between the description of a new method and its commercialization.

The goal of this review is to provide an in-depth yet application-independent overview of current and state-of-the-art technical developments in the field, guiding the reader through the vast expanse of tools that can be used

*To whom correspondence should be addressed. Tel: +1 415 476 4636; Email: morgane.boone@ucsf.edu
Correspondence may also be addressed to Nico Callewaert. Tel: +32 9 331 3630; Email: nico.callewaert@ugent.vib.be
Present address: Morgane Boone, Department of Biochemistry and Biophysics, UCSF, San Francisco, CA 94158, USA.

© The Author(s) 2018. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
to turn a pool of nucleic acids into a library that can be sequenced or assayed using other means. We here summarized the principal insights in this fast-paced discipline, expanding on newly published studies and aspects not covered in previous reviews (14–16).

### STARTING WITH RNA

The plethora of different types of libraries all converge to dealing with either DNA or RNA (which is, eventually, almost always converted into amplifiable DNA). The starting point in RNA procedures are mostly total RNA or poly(A)$^+$-RNA transcripts, but can extend to in vitro-transcribed (IVT) RNA, various types of non-coding RNAs, ribosome footprints, tRNAs, crosslinked RNA or modified RNA. For each of these subsets, dedicated protocols (17–23) or commercial kits exist for their purification—these are beyond the scope of this review and will not be detailed further. Nevertheless, the downstream steps for most of these molecules are generally the same.

### Ribosomal RNA depletion

Ribosomal RNA (rRNA) makes up more than 80–90% of the total RNA pool of all cells (24–26). In most applications, this large fraction is irrelevant to the question of interest. While downstream computational filtering of reads mapping to rRNA genes is always an option, these molecules take up unnecessary sequencing space, needlessly inflate screening scale when assaying libraries for expression and can reduce the overall sensitivity of the assay in question. As a consequence, rRNA depletion methods have received considerable attention, and the advantages and disadvantages of commonly used procedures are well studied.

Poly(A)$^+$-tailed RNA selection via hybridization capture using oligo(dT)-coupled beads (or variations on this theme) has been very powerful to extract protein-coding mRNA transcripts from the total RNA pool, passively depleting it from rRNA and immature or incompletely processed heterogeneous nuclear RNA (27). The most obvious downside of this method is the counterselection for all other poly(A)$^+$-negative RNAs which might potentially be of interest, many of them small non-coding RNAs transcribed by RNA polymerase III (small nucleolar RNAs [ snoRNAs], several microRNAs, U6 spliceosomal RNAs, the SRP RNA component, among others) (28). The poly(A)$^+$-negative transcripts of bimorphic genes (that produce both classically poly(A)$^+$-tailed as well as non-tailed mRNAs) are also missed in this situation, which is likely the reason why their distinct roles have been overlooked for many years (29). Histone mRNAs are also known to lack a poly(A)$^+$-tail, just like the HEG1 and DUX mRNAs (23), although a recent study reported the detection of 28 histone cluster genes in the poly(A)$^+$ RNA fraction, arguably resulting from incorrect 3' processing (27). Additionally, although bacteria can tag mRNAs with poly(A)$^+$-tails for the purpose of degradation (30), bacterial transcripts generally lack these tails and consequently, this strategy is not applicable in bacteria. In contrast, the 13 proteins encoded by the mitochondrial genome in eukaryotes that produce 'prokaryote-like' polycistronic, intron- and capless mRNAs are nevertheless also poly(A)$^+$-tailed by a mitochondrion-specific poly(A)$^+$-polymerase (27,30,31). For the purpose of rRNA depletion, poly(A)$^+$ selection is effective but not complete: even after several rounds, at least 0.3% of all sequencing reads map to rRNA genes (27). Many of these rRNAs contain poly(A)$^+$-stretches in their sequence. Moreover, the enrichment for poly(A)$^+$ transcripts can lead to a bias in sequence coverage through differential binding to oligo(dT), as was recently assessed by sequencing of IVT-arrayed cDNA libraries (18). Finally, for degraded RNA (especially in formalin-fixed, paraffin-embedded (FFPE) samples), poly(A)$^+$ selection will only recover the 3' portion of the transcript.

Active removal of rRNA sequences using a mixture of sequence-specific probes immobilized on beads (e.g. Ribo-Zero (Illumina) and RiboMinus (Thermo Fisher)) is a popular alternative compatible with the recovery of poly(A)$^+$-negative RNA, as it offsets many of the disadvantages of poly(A)$^+$-selection. However, remaining contaminating rRNA is also of concern, to a variable extent but generally more so than in poly(A)$^+$ selection (27,32,33). Active ribodepletion using these methods can also affect sequencing coverage, especially of those genes with stretches sharing similarity with rRNA sequences (18,26). Of the most popular commercial reagents, the Ribo-Zero kit seems to be less susceptible to this coverage skewing than the RiboMinus kit, most likely because of the more stringent hybridization requirements (34). For mRNA abundance measurement in Saccharomyces cerevisiae, results obtained with the Ribo-Zero kit, compared to RiboMinus or poly(A)$^+$-selection, correlated the most with total RNA data (34). Enzymatic methods for active ribodepletion have also gained popularity. As such, abundant DNA sequences (like cDNAs derived from rRNAs) can be digested non-specifically using the Kamchatka crab duplex-specific nuclease (DSN) (35,36), even in a single-cell setting (37) (see below in the 'Normalization' section). Similarly, RNA bound to specific DNA oligos can be digested by the heteroduplex-specific RNase H (38). Of all the common active ribodepletion methods, the RNase H method came out as overall best performer by most measures in a recent comparative study, leading to the highest rRNA depletion efficiency and the lowest coverage, especially of those genes with stretches sharing similarity with rRNA sequences (18,26). Another promising newcomer is DASH (depletion of abundant sequences by hybridization), in which ribodepletion is obtained through enzymatic digestion by recombinant Cas9 and rRNA-specific guides (39). DASH could effectively deplete mitochondrial ribosomal sequences in low-input RNA-seq libraries, reportedly out-performing several commercial RNase H-based and Ribo-Zero ribodepletion kits in performance, cost and input requirements (39).

An alternative tactic that has been used for the purpose of ribodepletion is selective random hexamer priming. By computationally subtracting rRNA-complementary hexamers from a random hexamer primer library before synthesis, the Raymond lab generated a 749 not-so-random hexamer library that could indeed selectively prime the non-rRNA transcriptome under high salt conditions (40). Leveraging the tolerance of reverse transcriptase (RT) for one or two mismatches at the priming site, the number of primers can even be reduced to below 50 while still broadly covering
the transcriptome (41) and requiring only limited quantities (50 pg) of RNA with careful primer design (42). This method can also be expanded to deplete other abundant transcripts (see below in the ‘Normalization’ section) or to reduce priming artefacts (41,42). Although the selective random hexamer strategy has been used with success in RNA-seq (43), the observation that still more than 10% of reads mapped to (cytoplasmic) RNA (40,41) makes this method much less efficient, and thus less advisable, for ribosomal depletion compared to the methods cited above.

In all, when the input RNA amount is not limiting, poly(A) selection seems on par with active ribodepletion methods like RNase H-based or DASH, and it is mostly the RNA species of interest (mRNA, non-coding RNA) that will dictate which approach is the most appropriate. However, it is important to note that none of these strategies are compatible with the minute amounts of RNA extracted from a single cell. Instead, current single-cell RNA-seq library construction methods almost exclusively rely on direct oligo(dT)-based priming (not hybridization-based physical selection) of extracted RNA to simultaneously deplete ribosomal species and prime the mRNA for reverse transcription (44–50). In one recent report, poly(A)-negative transcripts from single cells could be detected by combining oligo(dT)-priming with selective random hexamer priming and strand displacement (RamDA-Seq, Random Displacement Amplification Sequencing) (51).

RNA fragmentation

Fragmentation is a requirement for most sequencing libraries, as uniform sizing of molecules is important for optimal performance of most ‘second-generation’ sequencing instruments. This is not only due to restrictions in read length, but also because amplification (both in solution and solid-phase) favors smaller fragments over longer ones. In addition to the observation that RNA hydrolysis is more straightforward and less prone to sequence bias than DNA fragmentation, it can mitigate some of the biases that can be introduced during the conversion to cDNA by RTs (see below). As such, RNA fragmentation reduces random priming bias during cDNA synthesis, likely by limiting secondary structure formation, and enables a more equal coverage of the 5′ and 3′ transcript ends (52).

Taking advantage of the nucleophilicity of the 2′-hydroxyl group of RNA, simple heating and addition of catalytic metal ions that act as Bronsted bases to abstract the 2′-OH proton, like Zn²⁺ or Mg²⁺, is sufficient for efficient hydrolysis (53,54). The resulting fragment ends are a mix of 5′-hydroxyl groups, 3′ phosphates, but also 2′ phosphates and 2′,3′-cyclic phosphates (55), which can be problematic for certain downstream enzymatic steps (predominantly for RNA ligation). Consequently, such chemical fragmentation is often followed by T4 polynucleotide kinase treatment, resolving cyclic or 3′ or 2′ phosphates back to 2′ and 3′ OH groups and phosphorylating 5′ ends (56–58). Because chemical shearing is quick and efficient, and size distributions can easily be optimized by changing incubation time, it has become more widespread than mechanical methods, such as sonication, for RNA fragmentation.

Enzymatic digestion with the double-strand-specific RNase III is also an alternative, and has the advantage that it generates 5′-phosphate and 3′-hydroxyl ends more compatible with direct RNA ligation. Although the enzyme has a preference for double-stranded RNA (dsRNA), single-stranded RNA (ssRNA) can also be cleaved by modulating the salt and RNA concentration (59). However, digestion with RNase III is not completely random (60), a feature that does not really seem to affect coding region expression measurements in RNA-seq, but does substantially lead to under-representation of specific classes of non-coding RNA (61,62).

cDNA generation

Reverse transcriptase. RNA requires conversion to DNA for most applications, whether it is for cloning or for sequencing. Direct sequencing of RNA has been reported (63–65) and is still an area of intense research, but is not as advanced and robust yet as the sequencing of DNA. RTs are RNA-dependent 5′→3′ DNA polymerases and can be found in all domains of life with roles in various different biological processes, although they are generally believed to have evolved from a single ancient enzyme (66). Most currently commercially available RTs are derived from retroviral RTs, either from Moloney Murine Leukemia Virus (MMuLV or MMLV), or from the Avian Myeloblastosis Virus, and show various improvements in terms of processivity, thermostability or lack of RNase H activity—factors that all affect the reliability with which RNA libraries can be converted to cDNA. Processivity issues can lead to under-representation of 5′ ends of long RNAs, such as unfragmented mRNA transcripts. Highly structured or GC-rich RNAs, such as tRNAs, are notoriously difficult to reverse transcribe, and many efforts have been directed towards increasing RT thermostability to allow for template secondary structure melting and specific primer binding at elevated temperatures (67). Modifications can also inhibit RT (68), and its RNase H activity is often undesirable as it can degrade long RNA molecules before complete cDNA synthesis has taken place, which is why several commercially available RTs have mutated RNase H domains.

Despite these efforts, however, reverse transcription remains a significant source of bias during library generation. A principal aspect of all RTs is the intrinsic lack of 3′→5′ exonuclease or ‘proofreading’ activity. Error rates are high compared to DNA polymerases, and vary between 1/9000 and 1/30000 depending on the assay and enzyme, compared to 10⁻⁶–10⁻³ for DNA polymerases (69–71). While this is less of an issue for small RNA library construction, and can be mitigated in sequencing library construction by including more technical replicates, it remains difficult to analyze RNA sequence polymorphisms (72,73) and can be problematic in assays that rely on expression of the molecule. In addition to the RT’s low processivity (Figure 1A) and relatively high error rate, several artefactual activities have been reported as well. As such, intrinsic DNA-dependent DNA polymerase activity can lead to spurious second-strand DNA during first-strand synthesis, leading to artificial antisense sequences (64,74–76) (Figure 1B). Reportedly, the addition of actinomycin D, which
Figure 1. Undesired activities during cDNA synthesis. (A) The processivity of retroviral RTs is generally limited, which is problematic for complete reverse transcription of long RNAs. Secondary structures (gray) or modifications like 2′-O-methylation (indicated by *) in the RNA template can further impede full retrotranscription. Black = cDNA strand with annealing primer (random, oligo(dT) or specific). (B) Artefactual antisense products can be formed due to DNA-dependent DNA polymerase activity of RT during first-strand synthesis. This can occur through looping or repriming of the first cDNA strand. (C) During template switching, the RT repositions itself (and the synthesized first cDNA strand) further downstream of the same template, or a new one, during synthesis, leading to gapped synthesis of cDNA of intra-molecular fusions. (D) MMLV RTs have terminal transferase activity with a preference for template-independent cytosine addition. (E) cDNA synthesis with tailed primers. If the tail (blue) is unprotected, the Y-bifurcation formed is susceptible to the nuclease activity of DNA polymerase I during second strand synthesis, leading to incomplete incorporation in the final product. This can be mitigated by including phosphorothioate bonds or buffering bases (*) in the primer tail.

binds deoxyguanosines, can suppress this activity (77,78). Template switching, in which the RT and cDNA dissociate from the RNA template and reanneal to a different stretch, creates chimeric sequences, false deletions and inexistnet splice variants (79,80) (Figure 1C); 1–7% of all reads show evidence for this phenomenon (64). MMLV RTs are known to add additional bases at the 3′ end of the newly synthesized cDNA strand (81) (Figure 1D). The latter feature has been turned into an asset in some cDNA synthesis protocols, such as the SMART (switching mechanism at the 5′ end of the RNA template) method, in which the dC tail preferentially appended by RT is used for hybridization with an oligo(G)-containing primer for second strand synthesis (82). However, this terminal transferase activity of RTs is undesired in expression libraries as the extra bases could interfere with the reading frame and could result in proteins with extra amino acids. Finally, MMLV-derived RTs can be sensitive to 2′-O-methyl modifications in RNA (83) (Figure 1A), which can be an issue for mammalian piwiRNA or plant microRNA reverse transcription (84).

Two recent promising developments deal with several of these issues at once. The first has come forth from the study of maturase RTs, an alternative class of RTs found in non-long-terminal-repeat retrotransposons (66) and in intron-encoded proteins of group II introns (85). The Lambowitz group focused on bacterial mobile group II intron RTs, which have evolved to reverse transcribe very structured group II intron RNAs (86). Known as TGIRTs, or Thermostable Group II Intron RTs, these RTs have higher thermostability, higher processivity and about 2-fold higher fi-
delity than the commercial golden standard retroviral RTs (SuperScript III) (86). They can also read through modified bases, and while the template switching frequency remains the same (about 0.14% of reads), the resulting deletions are only rarely internal (87). The authors also discovered that RNA–DNA duplexes with single 3′ N-overhangs can be used to directly couple the cDNA strand to an adaptor sequence (86,87) (see also Figure 2C). The method has been broadly adopted, also for the sequencing of highly structured tRNAs (21,88–90). Another exceptionally processive and highly soluble maturase RT was recently discovered in *Eubacterium rectale* (91). While this ‘MarathonRT’ remains to be validated in a next-generation sequencing context, the observation that it can reverse translate a 5 kb transcript with less background than TGIRT make it especially promising for long-read sequencing technologies such as PacBio (92).

A second advancement, reported by the Ellington group, is the modified direction evolution of a high-fidelity thermostable DNA polymerase to enable reverse transcription with proofreading (71). The final reverse transcription xenopolymerase (RTX) has a 3- to 10-fold lower error rate than MMLV RT (3.7*10^-5 versus 1.1*10^-4), remains thermostable and processive, and was shown to be completely compatible with RNA-seq, leading to nearly identical coverage and expression profiles as an established RT (71).

**Priming.** RTs require a primer for first strand cDNA synthesis. Unless a sequence-specific primer can be used (e.g. in the case of TGIRTs or after RNA ligation, see below), the standard approach relies on either oligo(dT) or random primers. Homopolymer stretches, mostly poly(A), can be added to substrates without poly(A)-tail to enable oligo(dT) priming (93). The *Escherichia coli* poly(A) polymerase, the most often used tailing enzyme in these approaches, is however significantly affected by terminal stemloop structures (94,95) and to a lesser extent by 3′ nt identity (84) of the substrate, although both features can be minimized by adapting reaction conditions (increased temperature and reaction times). Nevertheless, the addition of bases can be problematic if the products are to be cloned for expression downstream in the procedure, as it may disrupt the frame or add unwanted codons. Poly(A)-tailning can also obscure the identity of the 3′ base of each template fragment, as an original 3′ adenosine may be mistaken for the synthetic poly(A)-tail. Moreover, as most vertebrate piRNAs and plant miRNAs carry 2′-O-methyl groups at their 3′ ends instead of 2′-OH (96,97), and these ends are poor substrates for poly(A) polymerases (84), the method is not suited to capture these types of RNAs.

A frequently used alternative is random priming. Primers as short as 6 bp are capable of sequence-specific RNA binding (98). Consequently, for random priming, random hexamers or heptamers are most commonly employed. In comparison with oligo(dT) priming, the random approach was shown to enable more equal sequence coverage across mRNA transcripts in early RNA-seq studies, especially after RNA fragmentation attenuate structure formation (52). Nevertheless, random primer annealing is prone to skewing; one meta-analysis of several RNA-seq experiments revealed that nucleotide frequencies of the 13 first nucleotides of each read were clearly diverging from the expected 1:1:1:1 A:C:G:T ratio in a manner that correlated with the type of primer used (random or not) (99). While there is a role for thermodynamic preferences toward GC-rich sequences, the actual skew depends on the composition of the transcriptome and also on motif preferences of the exact RT and polymerase used during cDNA synthesis (99,100). This positional bias can be corrected for in silico (99).

Simple random priming does not retain strand information, however. To do so, it is possible to tag random primers (or oligo(dT) primers after fragmentation) with specific sequences (and for instance, add a restriction site or barcode). These tails reportedly only modestly influence priming (40,100,101), although a rigorous systematic assessment is lacking. It is important to note that these non-hybridizing tags of random primers are sensitive to nucleolytic degradation, which can lead to inactivation of incorporated restriction sites and loss of directionality (100–102) (Figure 1E). This phenomenon has been attributed to the 5′→3′ exo- and endonuclease activity of DNA polymerase I during second strand synthesis, which has a particular preference for single-stranded DNA (ssDNA) in bifurcated duplex structures (103,104). The incorporation of nuclease-resistant phosphorothioate bonds (100) or additional bases that buffer the tag sequence (101) can counter this effect. Alternatively, the DNA polymerase I can be replaced by the 5′→3′ exo- Klenow fragment, a protolytic product of the *E. coli* DNA polymerase I which only retains polymerase and 3′→5′ exonuclease activity, but this requires the availability of a second primer binding site for second strand synthesis and full degradation of the RNA template (40).

How sensitive are these methods for the generation of single-cell libraries? As alluded to above, the greatest strength of oligo(dT)-based priming is its ability to combine ribodepletion and priming of mRNA for reverse transcription in a single step, which is why this strategy has become by far the most widespread starting point for single-cell transcriptome library synthesis (44–50). The Huang lab has however shown that tagged random priming can also be accommodated to minute input amounts without massively amplifying rRNA; the authors speculate that the mild lysis conditions and specific reverse transcription procedure likely contribute to this effect (105).

**RNA ligation.** A popular alternative to oligo(dT) or (tagged) random primers is the ligation of adaptors at the RNA level prior to cDNA synthesis. Crucially, this method preserves the directionality of RNA molecules and is thus a stranded approach, provided that the necessary end groups are protected. Combined with an rRNA-masking oligo, RNA ligation can also be used in a single-cell setup (106).

In general, single-stranded adaptors are sequentially ligated, first to the 3′ end of the RNA molecule, and before or after cDNA synthesis, to the other end (107) (Figure 2A and B). In order to avoid domination of circular or concatamerized products, without having to resort to extensive dephosphorylation/rephosphorylation reactions, most protocols rely on a C-terminally truncated form of T4 RNA ligase 2, trRnl2, which has lost the ability to use free adenosine triphosphate (ATP) to catalyze ligation reactions (108). Using pre-adenylated DNA adaptors (App-adaptor) (Figure
Figure 2. Common strategies for RNA adaptor ligation. (A) RNA substrates with 5′ phosphates and 3′-OH can be sequentially ligated with a 5′ pre-adenylated (App), 3′ blocked (x) DNA adaptor using truncated Rnl2 (ideally the K227Q R55K mutant), and a 5′ unphosphorylated, 3′ hydroxylated RNA adaptor with Rnl1. Sometimes the primer for reverse transcription is added before 5′ adaptor ligation. (B) In RNA/DNA ligation, RNA substrates with 3′-OH are ligated to a 3′ adaptor as in A, but no blocking is required. After reverse transcription by RT and degradation of the RNA strand, the 3′-OH of the resulting cDNA strand is ligated to a 5′ preadenylated, 3′ blocked DNA adaptor using the 5′ App DNA/RNA ligase (Mth K97A). (C) In TGIRT-mediated addition, RNA templates are immediately reverse transcribed and adaptor ligated via TGIRT and a double-stranded, single random overhang adaptor. Ligation of the other adaptor can be done as in B. (D) CircLigase can be used to circularize single-stranded cDNA molecules that were ligated to a bifunctional adaptor on one side using either RNA ligation or TGIRT-type methods, followed by reverse transcription. After circularization, the adaptor can serve as starting point for PCR to regenerate linear molecules with a different adaptor on both sides. (E) In hybridization-based RNA ligation, RNA templates are ligated to adaptors with randomized single-stranded overhangs, and then reverse transcribed.

2A and B), free 3′-OH RNA ends can be adaptor ligated, effectively avoiding circularization (109). Adaptor–adaptor concatamers are avoided as the enzyme requires 3′ RNA, not DNA, ends, although in practice, 3′ adaptor ends are nevertheless often blocked (e.g. –NH₂, three-carbon or six-carbon spacers) for the 5′ ligation reaction. The trRnl2 does tend to deadenylate the App-adaptors and to subsequently adenylate the substrate RNA molecule, leading to substrate concatamers and circles; the K227Q point mutant lacks this activity, leading to less side products (110). The mutation does slightly affect ligation efficiency, but this has been mitigated using a compensatory R55K mutation (leading to ‘tr-Rnl2 K227Q R55K’). A related pre-adenylation dependent enzyme, Mth K97A, derived from the Methanobacterium thermoautotrophicum RNA ligase, has the added advantage of thermostability, facilitating the melting out of potentially inhibitory RNA structures in the template (111). The enzyme does show a preference for A and C at the third nucleotide from the ligation site (112).

After 3′ adaptor ligation, the 5′ adaptor can either be ligated to the 5′ end of the RNA before first strand synthesis, or to the 3′ end of the resulting cDNA strand after first strand synthesis. In the former scenario, the RNA substrate 5′ phosphate is linked to the 5′ RNA adaptor’s 3′ hydroxyl
by the ss T4 RNA ligase 1 (Rnl1) (113) (Figure 2A). To avoid side products, the substrate's 3' end should be blocked, and the adaptor should not be phosphorylated at the 5' end. As the Rnl1 is much more a single-strand specific ligase than Rnl2, often the DNA primer for reverse transcription, which anneals to the 3' adaptor, is added even before the 5' adaptor ligation step. This also reduces undesired products caused by excess unligated 3' adaptor.

Alternatively, the 5' adaptor can be ligated to the first strand cDNA after degradation of the RNA strand, for instance through alkaline treatment (69) or RNase H digestion (Figure 2B). Provided that the 5' adaptor (DNA) is 5' adenylated and 3' blocked, the ATP-independent thermostable Mth K97A (sometimes referred to as the 5' App DNA/RNA Ligase) is used for this, as it has better ssDNA ligation activity than (tr)Rnl2 (111).

Both 3' and 5' end RNA (or ssDNA) ligation biases are significant and have been extensively documented, mostly in the context of small RNA sequencing (72,73,95,112,114–117). Using synthetic equimolar pools of more than 900 different miRNAs, the Brett Robb lab measured that differences in ligation efficiencies between single molecules can introduce up to 10 000-fold abundance variation, independent of polymerase chain reaction (PCR) biases (118). Although initially, this bias was often attributed to primary structures in ligating efficiencies between single molecules can cause by excess unligated 3' end.

RNA ligation can be a challenge when substrate RNA molecules are modified at their 5' or 3' ends. Under the right conditions, 2'-O-methyl groups are not an issue for tr-Rnl2 (84). In contrast, 3' end 2', 3'-cyclic phosphates are not ligatable. For resolution of unwanted 2', 3'-cyclic phosphates, as arises after divalent cation or ribozyme, RNase A, RNase T1 or RNase 1 activity, treatment with wild-type T4 polynucleotide kinase in acidic conditions is sufficient, as mentioned before. For 5' end ligation of RNA molecules that lack a regular 5' phosphate, enzymatic treatment with tobacco acid pyrophosphatase to remove cap structures or with T4 PNK to phosphate 5'-OH ends, can be necessary (123,124).

Second strand synthesis. Second strand synthesis is generally performed using the very efficient and versatile classical Gubler and Hoffman method (125), or one-tube versions that are offered commercially. Principally, the method combines E. coli RNase H digestion, which creates nicks in the RNA strand of the RNA–DNA duplex after first-strand synthesis, E. coli DNA pol I, which can use these nicked sites as primer for 5'→3' DNA synthesis while displacing and degrading the RNA in the same direction through its 5'→3' activity, and E. coli DNA ligase, which ligates the nicks. Overhangs are degraded through the 5'→3' and 3'→5' nuclease activities of the DNA pol I, leaving blunt ended DNA.

Although this classical Gubler and Hoffman second strand synthesis method is not intrinsically strand specific, the polarity of transcripts can be retained by replacing dTTP with dUTP in the second strand synthesis reaction. The introduction of uracil blocks high-fidelity amplification of the second strand in the PCR step (126), and combined with the appropriate adaptors (see below), all amplified molecules will consequently have the same orientation. Alternatively, the uracil-containing strand can be degraded using a mixture of uracil–DNA glycosylase and DNA glycosylase-lyase endonuclease VIII (NEB's USER) before PCR. The method is popular and efficient, and it performed best among several other strand-specific methods for RNA-seq with regard to a variety of criteria, including evenness of coverage and strand specificity (127).

If specific sequences are incorporated prior to second strand synthesis, for instance through RNA ligation or SMART-type template switching, double-stranded DNA (dsDNA) can be generated from the single-stranded cDNA through PCR amplification. This approach is sensitive and suitable for second strand generation in single-cell setups (47,48,106).

STARTING WITH DNA

In many applications, DNA is the starting point of the library synthesis. This can be genomic DNA, immunoprecipitated DNA such as in ChiP (chromatin immunoprecipitation) or MeDIP (methylated DNA immunoprecipitation), targeted sequence captured DNA or any other method where a specific subset of sequences requires library synthesis. Alternatively, existing DNA collections such as the human ORFeome (7) can also be used as source material. Several fragment libraries for yeast-two-hybrid screening have...
been constructed from such collections, by PCR amplification of ORFs and titrated exonuclease digestion for progressive removal of vector end sequences (13,128).

**DNA fragmentation**

DNA fragmentation is required for short-read sequencing library construction when starting from molecules longer than the required platform range. Additionally, fragmentation is also an intrinsic part of fragment library generation for expression or protein–protein interaction screening. Compared to RNA, the double-stranded configuration and lower reactivity of the deoxyribose in DNA makes it more difficult to hydrolyze. Hence, one generally resorts to physical shearing methods using sonication, nebulization or acoustic shearing; or to enzymatic methods.

With sonication or nebulization, the size range tends to be wide and difficult to adapt, resulting in low yields; sample heating in the process may additionally lead to DNA damage and strand dissociation (129–131). The Covaris method of focused acoustics is considered best-in-class, with low sample loss, tunable DNA size ranges and high reproducibility (130). Fragmentation using either of these three methods nevertheless results in the preferential cleavage at CG dinucleotides (132), suggesting this is perhaps a typical attribute of physical shearing of DNA. Whatever the origin, this preference thus introduces a form of bias at an early step in the procedure.

Early reports (from 2006) employing DNase I digestion to randomly fragment DNA described the method as essentially bias-free (133,134). The DNase I endonuclease is often used in DNase hypersensitivity assays for chromatin analysis, and in transcription-factor footprinting methods. However, closer inspection of several hypersensitivity sequencing datasets revealed a clear preference for sites with cytosines at the −2 position of the cut site (135). The latest generation of fragmenting enzymes or enzyme blends (such as the NEB Fragmentase, with a nicking enzyme and an endonuclease cleaving the opposite strand) perform well in comparison, being less susceptible to sequence bias (136) and giving more consistent results than sonication or nebulization (137). Size range can easily be customized by modifying the DNA-to-enzyme ratio and digestion time, and as the resulting products are blunt ended, no end repair step is needed downstream.

Random priming of DNA material has been done as well (138). While short random hexamers and heptamers give satisfactory results for RNA, longer primers are required to offset competition for annealing with the opposite strand when working with dsDNA (139). The incorporation of a hairpin structure in the 5′ portion of the random primer has been reported to substantially reduce the number of byproducts due to random primer self-anneling in ChIP-seq libraries (140). Nevertheless, the strategy is far from ideal for the generation of random fragments, as it tends to be less efficient and more sequence-biased than other methods.

Methods in which uracil is doped into the DNA to enable fragmentation have been popular for protein fragment expression screening (141). Amplicon libraries can be amplified in a PCR with the regular four dNTPs and low amounts of dUTP. Fragmentation can then be induced at the doped sites, by uracil–DNA glycosylase digest for abasic site generation, nicking at these sites by the apurinic/apyrimidinic endonuclease IV and the generation of a double-strand break by the cleavage of the strand opposite the nick by S1 nuclease (10,142). Others have used a combination of endonuclease V and Mn2+ to induce double-strand breaks after uracil doping (143,144). The size distribution of the fragments can be manipulated by modulating the dUTP/dTTP ratio (10). Note that using this strategy, AT-rich regions will be more prone to cleavage compared to GC-rich regions, as more break-inducing dUTPs are incorporated (144).

**Adaptor ligation to DNA**

Depending on the fragmentation method, in most cases, ends of dsDNA need to be repaired or ‘polished’ to blunt ends before downstream processing. Polishing involves digestion with enzymes that fill in 5′ overhangs and remove 3′ overhangs; T4 DNA polymerase (sometimes combined with Klenow fragment) is mostly used for this purpose (145). Generally, this is combined with T4 polynucleotide kinase to phosphorylate 5′ ends that lack phosphates. To ligate the adaptors, ultrapure T4 DNA ligase preparations can also boost ligation efficiencies (130). The most popular adaptor design combines template phosphorylation and 3′ tailing with a single nucleotide (usually A, although G-tailing is efficient as well), followed by ligation with a single T (or C) -tailed, Y-shaped adaptor (146) (Figure 3). This combination maximizes the ligation efficiency by avoiding blunt-end ligation, while effectively sidestepping template concatamerization and adaptor dimer for-
and 3′ oxyuridines, treatment can degrade the strand after 3′ endonuclease (148). As the strand with the overhang is doped with deoxyuridines, USER treatment can cleave the strand after ligation and the resulting single-stranded adaptor-ligated DNA can be amplified (148). In another example, commercialized by Swift Biosciences, dsDNA is ligated to the individual strands of the Y-shaped adaptor in a sequential reaction (149). In the first ligation, a semi-single stranded 3′ blocked adaptor is ligated to one strand only of the dsDNA molecule. USER treatment can then degrade the non-ligated strand due to the presence of deoxyuridines, consequently allowing the next adaptor strand to anneal and ligate (149). In a third example, a combination of dUTP-doped forward and regular reverse primers can be used to amplify DNA, and USER treatment asymmetrically releases one strand of one of the adaptors on the molecule, which is then ligated to a 5′ blocked single-stranded oligo (150). This ‘reshaping’ of adaptors on DNA has been used to resolve problematic instances of intramolecular hairpin formation due to adaptor complementarity, which precludes Ion Torrent sequencing (150).

The ligation-based schemes with the Y-shaped or hairpin adaptors mentioned above are efficient, and the formation of side products is strongly reduced. Nevertheless, the procedure requires much sample-handling and is incompatible with very limited inputs (e.g. DNA from single cells). In contrast, the clever ‘tagmentation’ approach, which uses an engineered hyperactive Tn5 transposase for simultaneous DNA fragmentation and tag (or adaptor) insertion, is fast and suited for low input amounts (151). A general point of concern for tagmentation, however, is insertion bias. Although negligible for DNA sequencing of human genomes, the skews are significant in GC-rich, small genomes or when using PCR products as a starting material (151,152).

More difficult input sample types require adapted protocols. Highly degraded DNA, especially from ancient or FFPE samples, has a higher proportion of ssDNA and the input material is often only available in trace amounts. Single-strand compatible methods include the Swift Biosciences approach of sequential ligation as outlined above (149), but tailing of the ssDNA to enable priming and dsDNA generation has also been used (153,154). The Meyers lab has developed a method based on ssDNA ligation of single-stranded biotinylated adaptors using CircLigase, which avoids loss of material during purification as the sample is bound to streptavidin-coated beads (155). A recently improved version of this approach, ‘ssDNA2.0’, replaces the adaptors with splinted adaptors and the ligase with T4 DNA Ligase, and was shown to be superior for ancient DNA sequencing library preparation (156).

Capturing methylation

Analyzing the methylation status of the genome requires the construction of libraries of methylated DNA. The golden standard for genome-wide profiling of 5′-methylcytosines (5mC), the most established DNA methylation mark, relies on chemical treatment of (generally fragmented) DNA with bisulfite (157). Bisulfite deaminates unmethylated cytosines (C) to uracils (U) while leaving 5′-methylcytosines intact (158). As such, comparing bisulfite-treated and untreated samples reveal loci with unconverted, and hence methylated, cytosines. While powerful, the use of bisulfite has several important repercussions. First, efficient amplification of bisulfite-treated DNA requires a polymerase that can tolerate the presence of unnatural deoxyuridines, and cope well with the now more abundant AT-rich regions (see section ‘Amplification’). The current best performer in that regard is considered to be the KAPA HiFi Uracil+ DNA polymerase (BioRxiv: http://dx.doi.org/10.1101/165449), which has a mutated uracil-binding pocket to avoid stalling at uracils. Second, bisulfite treatment can also result in the loss of cytosine bases and subsequent DNA breakage at the resulting abasic sites, consequently inducing DNA fragmentation (159). As this especially affects regions of unmethylated C-rich sequences, this can significantly skew sequence representation and estimation of methylation levels, although a reduction of denaturation temperatures and bisulfite concentration can limit these effects (BioRxiv: http://dx.doi.org/10.1101/165449).

The ligation of adaptors is therefore also not arbitrary in bisulfite protocols. Because of the aforementioned degradation issue with bisulfite, pre-bisulfite ligation (160,161) leads to sequence bias (BioRxiv: http://dx.doi.org/10.1101/165449) and requires relatively high input amounts. In addition, it necessitates adaptor synthesis with full cytosine to-5′-methylcytosine replacement in order to avoid uracil conversion of the adaptor (160,161). The more recent post-bisulfite ligation strategies exploit bisulfite-induced degradation for fragmentation and only attach adaptor sequences...
after bisulfite treatment, for example using random primer extension (post-bisulfite adapter tagging or PBAT) (162–164) or hexamer-guided partially single-stranded adapters (SPlimited Ligation Adaptor Tagging—SPLAT) (165). These methods are substantially less bias-inducing compared to pre-bisulfite ligation (BioRxiv: http://dx.doi.org/10.1101/165449) and have pushed the starting material limit down to the nanogram and even single-cell (163,164) range.

Although the above whole-genome bisulfite sequencing methods allow for full genome-scanning of methylation status, only a fraction of the genome is generally (differentially) methylated, and it can be more efficient and cost-effective to focus on methylene-relevant regions instead of whole genomes. One strategy involves the digestion of genomic DNA with methylation-insensitive restriction enzymes that recognize CG-rich sites, such as CCGG in the case of MspI, thereby enabling enrichment of regions with high CpG content. Combined with bisulfite treatment of digested and size-selected fragments, such reduced bisulfite representation sequencing (RRBS) allows the monitoring of a reproducible subset of CpG islands in genomes (166,167). Enrichment for certain sites can be modulated through careful selection of the restriction enzyme (168). Although powerful and amenable to single-cell studies (169), all RRBS methods are currently critically depend on some form of size selection to maximize their enrichment factor, and thus are incompatible with highly fragmented circulating cell-free DNA (170,171). Further innovations in RRBS protocols will address these limitations (De Koker et al., in preparation).

Alternatives to bisulfite-based strategies focus on pull-down of methylene-relevant regions using methyl-binding domains (172,173) or 5mC-binding antibodies (174). These methods, however, require more input DNA than PBAT, SPLAT and RRBS, and do not have single-basepair resolution of methylation status.

**AMPLIFICATION**

Although PCR is an extremely powerful technique, it is well known that the amplification of pools of molecules with different sequences and lengths, as occurs in libraries, can result in serious distortion of relative abundances, with under-representation or over-representation of particular sequences. Extremely GC-rich or GC-poor templates are generally difficult to amplify, while short sequences are preferentially amplified. Stochastic effects account for part of the bias as well (175). Additionally, errors can accumulate in templates, often at low-complexity regions, and side products resulting from overamplification, such as concatamers or self-primed chimeric sequences (176), are common. However, the extent of these issues can be attenuated by careful optimization of PCR conditions and polymerase choice. For instance, the monitoring of PCR cycle number to remain in the exponential phase was shown to substantially reduce the number of overamplification products (177–179) and to reduce effects of bias toward shorter sequences (180). Carrying out the reaction on beads in emulsion (emulsion PCR) also reduces the number of chimeras, as single molecules are amplified in individual compartments, which reduces cross-priming (181). The addition of compounds such as betaine can largely prevent the under-representation of GC-rich templates, but it does not improve bias against AT-rich sequences (182). The opposite is true for TMAC (tetramethyl ammonium chloride) (183).

Aside from PCR cycle number, the biggest impact comes from the polymerase used. Quail et al. systematically compared polymerase performance for sequencing library amplification over a range of different contexts, revealing considerable differences in fidelity, yield, sequence-sensitivity and processivity between the 23 polymerases tested (184). The KAPA HiFi enzyme, engineered for increased affinity towards DNA via directed evolution, came out as best performer, as it has the unique ability to amplify the most difficult (AT- or GC-rich) templates. The sequencing results of pools amplified by KAPA HiFi closely matched those of PCR-free libraries (184). The KAPA polymerase also surpassed the acclaimed Q5 high-fidelity polymerase (NEB), whose processivity has been enhanced through fusion with an additional DNA binding domain, in terms of accuracy and proportion of chimeric molecules (185). However, this high fidelity may come at a cost: the authors of the latter study also observed the surprising ability of both KAPA and Q5 enzymes to edit primer sequences (4% of primed molecules), leading to the unwanted amplification of sequences with small primer mismatches.

It is possible to generate libraries without the need for amplification, although the high sample input amounts (up to 5 μg) limit the breadth of applications of such amplification-free methods. The Turner lab demonstrated the superiority of PCR-free sequencing library construction using simple ligation of Y-shaped adaptors that contain all the necessary sequences required for Illumina sequencing, in the sequencing of extremely AT- or GC-rich bacterial genomes (186). Similarly, adapter-ligated RNA libraries do not have to be amplified for RNA-seq using the FRT-seq method (Flowcell Reverse Transcription Sequencing), in which reverse transcription is performed on the Illumina flow cell prior to bridge amplification and sequencing (187).

However, when the input material is limited, such as in the extreme case of single-cell sequencing, many researchers resort to (semi-)linear amplification methods to amplify the material while minimizing artifacts. Because of the exponential aspect of PCR, errors quickly propagate and biases are exacerbated; this cumulative effect is less extreme for linear methods relying on the T7 RNA polymerase or strand-displacing enzymes such as the BstI or Φ29 polymerases. The bacteriophage T7 RNA polymerase methods rely on *in vitro* transcription of DNA molecules encoding a T7 promoter, a system routinely used for microarray sample preparation (188,189) (Figure 4A). As each DNA molecule is templated multiple times, but the resulting RNA products are not, polymerase errors are not propagated. Both single-cell ChIP-seq and RNA-seq libraries have been generated using this method (49,190–192). The downside of this approach is that the T7 polymerase is prone to premature termination on low complexity sequences, and if temperatures are reduced to counteract this problem, yield is affected (191). Strand displacement enzymes have been a popular alternative, especially in the context of whole genome amplification (WGA), and to a lesser extent, whole transcriptome...
amplification. As such, in MDA (multiple strand displacement amplification), DNA is amplified in an isothermal reaction using a random primer and the φ29 polymerase (Figure 4B), a very processive enzyme that can generate fragments up to 10 kb from a single template (193). The most efficient templates are either large, linear molecules or circularized molecules (194). As a result, MDA has been successfully applied in various settings, from low-input or single-cell RNA-seq after circularization of cDNA (195,196) to the sequencing of single bacteria in clinical samples (197), or of single tumor cells (198). Despite catalyzing efficient amplification (which is technically not linear), and its high fidelity and very low sequence bias, ∼6% of molecules are chimeras, and amplification bias can still occur due to primer binding skew (199–203). Other strand displacement enzymes used in MDA-type setups include the BstI polymerase and derivatives (204), and a synthetic fusion of the T7 DNA polymerase (3′→5′ exominus) with the processivity-enhancing thioredoxin (marketed as Sequenase), which has successfully been used for low-input ChIP-seq (140) and single-
cell RNA-seq (196). In another technique, the MALBAC (multiple annealing and looping-based amplification cycles) method, a strand-displacing enzyme such as BstI is used to generate overlapping fragments from a template using cycles of gradually increasing temperatures and template looping, followed by limited PCR (205) (Figure 4C). The quasilinear amplification step in MALBAC would reportedly result in vastly higher coverage, a lower allele dropout rate, and a higher reproducibility than MDA for WGA (205,206), although the error rate is lower in MDA due to the higher fidelity of the δ29 polymerase (207). Recently, the method has been adapted for single-cell RNA-seq (208).

NORMALIZATION

Multiple applications benefit from the removal or normalization of abundant nucleic acid sequences, beyond rRNA-derived molecules, in libraries. The large dynamic range of eukaryotic transcriptomes, which spans over four orders of magnitude (209,210), entails that highly expressed transcripts are strongly over-represented in transcriptome libraries. This can be problematic for rare transcript discovery (such as infrequent splicing events) in RNA-seq, and it also needlessly scales the size of the library to be screened in approaches relying on RNA as input material but for which transcript abundance information does not need to be retained, such as cDNA expression libraries. Abundant repetitive or organellar sequences in eukaryotic genomes can be a nuisance for some applications, complicating de novo genome assembly and alignment (211). Moreover, the sequencing of microbially infected clinical samples (212), of rare (mutated) tumor DNA or RNA in a background of healthy cells, or of fetal cells in a background of abundant maternal cells (213) all represent examples where depletion of unwanted high-abundant RNA or DNA species could substantially increase detection sensitivity.

Historically, these issues have been addressed in several ways: repetitive sequences, which are often hypermethylated (214,215), have been removed with methylation-specific or methylation-sensitive restriction enzyme systems (216,217), and abundant transcript sequences could be subtracted by hybridization with biotinylated or bead-immobilized driver sequences (218,219). Most often, however, normalization relied on the second-order kinetics of nucleic acid renaturation after denaturation (DNA concentration ~ rehybridization rate); a feature exploited intensely in the context of C0t analysis (initial DNA concentration x time) to estimate size, complexity and repetitiveness of genomes before sequencing became the norm (220,221). As abundant DNA sequences reassociate faster than rare ones after denaturation, any method that can reliably separate dsDNA from ssDNA could enrich for low-abundant sequences—most commonly, this was achieved using hydroxyapatite chromatography (222,223). All the above methods proved to be rather labor-intensive (some required substantial skill) and were therefore less suited for higher-throughput studies. The discovery and characterization of a DSN isolated from the hepatopancreas of the Kamchatka crab (Paralithodes camtschaticus), however, enabled simple and robust digestion of double-stranded abundant species (224–226) (Figure 5). The DSN enzyme displays a high specificity for DNA in dsDNA or RNA–DNA hybrids of 10 bp or longer, only very little activity on ssDNA, and does not cleave ss or dsRNA, nor does it seem to have any apparent sequence specificity (226,227). As such, it has been efficiently deployed for normalization of cDNA or RNA-seq libraries (224,228–230), reaching up to a 1000-fold reduction in abundance differences (225); but also for genomic DNA normalization (231,232); the removal of specific transcripts (224); and, as mentioned above, ribodepletion (35–37). Additionally, DSN’s ability to discriminate single mismatches in DNA duplexes has successfully been put to use for SNP detection (227). The Michelmore group characterized the overall effect of DSN-based normalization through deep sequencing of DNA and RNA libraries, concluding that, for the conditions tested, substantial but not complete abundance equalization was obtained, and that not all sequences seem equally prone to DSN digest (232). Predictably, GC-content plays a role, as high GC% stimulates rehybridization. The addition of TMAC, known to normalize GC and AT pair reannealing rates as exploited in several other applications (183,233–236), could improve this bias and lead to enhanced normalization of AT-rich genes, but it also negatively affected overall normalization efficiency (232). Our own observations suggest that for adaptor-ligated libraries, adaptor sequence can also substantially influence the efficiency of DSN normalization (BioRxiv: http://doi.org/10.1101/241349).

The CRISPR-associated nuclease Cas9 can also be used for similar normalization purposes. DASH could effectively enrich for a rare mutant variant of the KRAS gene in synthetic gDNA mixtures with a guide sequence against wild-type KRAS, mimicking the situation where rare cancer cells need to be detected in a pool of normal cells (39). This innovative CRISPR-based application can likely easily be extended to remove any combination of sequences of interest from a variety of libraries, as long as good and specific guide RNAs can be designed. Thus, it is anticipated that DASH could complement hybridization-based normalization for sequences that are less efficiently depleted using DSN.

BARCODES, MOLECULAR TAGS AND FRAMESHIFTS

Despite the high technical reproducibility of next-generation sequencing technologies, batch-to-batch variation effects can still be of concern. Multiplexing samples for sequencing by sample barcoding is a common and recommended approach to reduce part of this variation, while at the same time increasing cost efficiency—provided that the barcodes are well-designed (237). The main culprit for the observed variability between samples, even identical ones, is mostly the multistep library preparation. As such, the earlier samples are barcoded and pooled in the procedure, the better. For single-cell methods, such parallelization provides the additional benefit of increasing total sample amount (238). Shishkin et al. recently implemented barcode incorporation during RNA ligation for pooled multiplexed RNA-seq library construction (‘RNAtag-seq’) (239). Similarly, barcodes have been incorporated during cDNA synthesis before pooling (240). Considering the
Figure 5. Normalization of DNA abundance with DSN. Adaptor-ligated DNA pools with abundant molecules (black) and rare molecules (red) are subjected to denaturation and controlled slow renaturation at high temperature. Abundant molecules rehybridize faster. This pool of mixed dsDNA and ssDNA is then digested by DSN, which targets duplexes, resulting in unhybridized, single-stranded, low-abundant molecules remaining. A final PCR step enables recovery of these molecules to dsDNA.

Figure 6. Resolving the issue of low diversity amplicon sequencing on Illumina platforms using frameshifting nucleotides. (A) Schematic representation of the sequencing of different molecules with identical starting sequence (e.g. a common primer binding site used for amplification before the addition of Illumina adaptors). Illumina adaptor sequences are represented by Xs. Each molecule symbolizes a sequence cluster on the flow cell. At each cycle, an identical base is read in all clusters, interfering with cluster identification. (B) As in A, but here sequences have been amplified with a mix of primers containing additional frameshifting sequences of different lengths. As such, the nucleotide composition at each position in the different clusters is more diverse, enabling more reliable cluster identification. The actual first base of the common region is interrogated at different cycles for each cluster.

sequence or structural preferences of the various enzymes used during library preparation, it must be noted that exact barcode sequences or their location in the final sequence may also represent a source of bias. miRNA expression profiles, for instance, are known to be significantly skewed when barcodes are introduced adjacent to the ligation site during RNA ligation, but not during PCR amplification (115, 241).

Aside from barcoding individual samples, another relatively recent development involves the tagging of individual molecules in single samples through the incorporation of degenerate regions in adaptors or PCR primers before PCR. Such molecular tags (MTs, or unique molecular identifiers, UMIs) have been tremendously useful to differentiate identical molecules originating from the same PCR template (PCR duplicates), and those that were present at the onset of the library preparation (242–246). Sequences with the same UMI can be summarized into consensus sequences, and as such, in applications where the counting of sequences is important, the outcome is less skewed by PCR bias (247) or sequencing errors. UMIs have been successfully applied in the detection of rare variant molecules (248), to accurately profile immune repertoires (249, 250) or to quantify mRNA levels from single cells (45, 46, 48–50), as PCR amplification noise and sequencing errors often obscure these efforts. It has been noted that UMI-based correction does
require very high read depths, and that errors in the MTs or barcodes are an issue that should be taken into account (251–253). As a final note, for libraries of amplicons intended for Illumina sequencing, it may be convenient to introduce sequences of varying lengths just upstream of the first ampiclon bases to be sequenced. Illumina platforms strongly rely on the equality of base distributions in the first few cycles for phasing and cluster calling; the sequencing of libraries where the first position is the same in all clusters on the flow cell is therefore very inefficient (254) (Figure 6A). This issue can be bypassed by designing custom sequencing primers (255), but this may require thorough optimization, and is incompatible with paired-end sequencing using older versions of the Illumina control software. Alternatively, mixing in one or more samples with a more random base distribution, such as the PhiX 174 genome, can resolve the problem, but this makes that amplicon samples can never fully benefit from the full chip capacity. Others have reported a custom Illumina sequencing protocol, ‘dark sequencing’, where in a first run clusters are identified in late cycles (after the non-random bases) and the first bases of the sample are sequenced in a second ‘run’ (256). The preferred method, however, involves the incorporation of ‘frameshifting bases’, basically a pool of sequences of varying lengths that are added to the PCR primers. As such, the first sequenced base of each amplicon is different for the different neighbouring clusters (Figure 6B). This strategy has successfully been integrated in several 16S metagenome studies (246,257,258), and ensures full exploitation of the flowcell capacity.

CONCLUSION

Most molecular manipulations during library preparation introduce some form of bias, resulting in a skewed representation of the original molecules. This can affect accurate quantification, lead to false results, or mask potentially interesting patterns. The nature, source and impact of these library preparation biases in various settings has been subject to intense research in the past decade, and steadily, strategies to address some of these issues are emerging. As such, TGI-RTs and the reverse RTx are showing promise in replacing the inherently more error-prone retroviral RTs, and the benefits of internal randomization of adaptors during RNA ligation have become clear. For DNA amplification, the KAPA HiFi enzyme still tops the charts when it comes to PCR, and with careful PCR cycle number monitoring and the incorporation of MTs, PCR-related data distortions can be attenuated. Linear amplification methods such as MDA and MALBAC are being increasingly used, especially in single-cell setups. The implementation of nucleases such as the DSN or Cas9 for library normalization opens up the prospect of capturing rare molecules in complex samples. These valuable insights should help the researcher to make informed choices when it comes to library generation.

While protocols or enzymes of some commercial kits are generally updated with time, these adaptations often lag behind current knowledge; customizing the library preparation is almost always a better option and generally leads to libraries of superior quality. With continuous effort, it is expected that better enzymes or even simple protocol changes will continue to improve such procedures, enabling more accurate systematic assessment of genome, transcriptome and proteome function.

FUNDING

Ghent University Biţzeron Onderzoeksfonds PhD Fellowship (to M.B.); Research Foundation Flanders (FWO) personal PhD Fellowships (to M.B., A.D.K.); FWO Research Grant [G.0276.13N] (to N.C.); European Research Council Consolidator Grant [616966] (to N.C.). Funding for open access charge: VIB Institutional Grant Funding. Conflict of interest statement. None declared.

REFERENCES


64. Cameron, V. and Uhlenbeck, O.C. (1977) Structural effects of base modifications in messenger RNA.


