Artifacts and biases of the reverse transcription reaction in RNA sequencing

JASPER VERWILT,[1,2,3] PIETER MESTDAGH,[1,2,3] and JO VANDESOMPELE[1,2,3]

[1]OncoRNALab, Cancer Research Institute Ghent, 9000 Ghent, Belgium
[2]Department of Biomolecular Medicine, Ghent University, 9000 Ghent, Belgium
[3]Center for Medical Genetics, Ghent University, 9000 Ghent, Belgium

ABSTRACT

RNA sequencing has spurred a significant number of research areas in recent years. Most protocols rely on synthesizing a more stable complementary DNA (cDNA) copy of the RNA molecule during the reverse transcription reaction. The resulting cDNA pool is often wrongfully assumed to be quantitatively and molecularly similar to the original RNA input. Sadly, biases and artifacts confound the resulting cDNA mixture. These issues are often overlooked or ignored in the literature by those that rely on the reverse transcription process. In this review, we confront the reader with intra- and intersample biases and artifacts caused by the reverse transcription reaction during RNA sequencing experiments. To fight the reader's despair, we also provide solutions to most issues and inform on good RNA sequencing practices. We hope the reader can use this review to their advantage, thereby contributing to scientifically sound RNA studies.

Keywords: bias; RNA sequencing; reverse transcription

INTRODUCTION

The study of RNA has propelled advances in understanding biochemical processes in health and disease and assisted in developing novel biomarkers and RNA-targeted therapies. RNA sequencing (RNA-seq) can characterize large numbers of RNA molecules in parallel and has quickly become the prevailing method to study the transcriptome of tissues and cells in various organisms. In its most widely adopted form, the quantification and characterization of RNA heavily rely on a reverse transcription (RT) step. The RNA is used as a template during the RT reaction to generate a complementary DNA (cDNA) strand. Each RT reaction uses at least four components: the template RNA, one or more oligonucleotide primers, a reverse transcriptase enzyme (RTase), and an RT buffer. This step has become so widely adopted that researchers do not hesitate to treat it like a comfortable black box: *RNA in, DNA out*. However, several studies challenged this mindset and revealed its flaws. Some researchers have gone against the current, showing how the RT reaction is more intricate than anticipated and depends on a subtle interplay of its components. As we will illustrate in this review, discrepancies can arise between the RNA and the resulting cDNA due to several factors. We group these inconsistencies into the generation of faulty molecules that differ in sequence from the template RNA ("RT artifacts"); and quantitative changes between nucleic acid fragments in the transcribed cDNA compared to the template RNA input ("RT bias"). The latter can be partitioned into intrasample bias (some [parts of] transcripts are more likely to be reverse transcribed than others) and intersample bias (due to inconsistencies between preanalytical variables or protocol choices). With this review, we hope to provide the community with a concise overview of the pitfalls and concerns inherent to the RT reaction, with a focus on RNA-seq experiments. We hope to engage researchers to include appropriate control measures and think twice about their novel or unexpected results: things are not always what they seem.

REVERSE TRANSCRIPTION BIASES

The RNA-seq community has thoroughly described the causes of quantification biases in RNA-seq experiments—such as GC content, transcript length, and sequence base

Corresponding author: jasper.verwilt@ugent.be
Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.079623.123. Freely available online through the *RNA* Open Access option.

composition—and developed numerous strategies to account for them (Oshlack and Wakefield 2009; Hansen et al. 2010; Li et al. 2010; Srivastava and Chen 2010; Roberts et al. 2011; Zheng et al. 2011). However, the RT reaction is hardly ever considered a source of these biases. In this section, we aim to uncover the biases originating from the RT reaction itself. All factors of RT can introduce individual biases, but they might also arise from interactions between multiple components. For clarity, we chose not to group the biases by cause but by whether they affect results in an individual sample (*intra*) or between samples (*inter*). Figure 1 provides an overview of the different intrasample biases. Table 1 summarizes these biases and provides recommendations to overcome them.

INTRASAMPLE BIASES

Reverse transcriptase-RNA bias

To understand the reverse transcriptase-RNA bias, it is critical to understand that contemporary RTases are engineered versions of retroviral RTases, of which the wild-type versions have several characteristics helping with gene regulation and viral replication. First, retroviruses rely on an RNA secondary structure-dependence of RTases for their gene regulation and evolution (for review, see Smyth et al. 2018). Therefore, RT should be inherently dependent on RNA secondary structure. Second, retroviral RTases contain an RNase H moiety that is responsible for hydrolyzing the RNA of the formed cDNA:RNA duplex. This subunit is indispensable for retroviral replication as it frees the cDNA for the necessary "jumps" of the RTase within the genomic RNA. In vitro, however, RNase H activity can result in premature hydrolysis of the template RNA and interruption of the RT, introducing a negative bias toward longer transcripts (Kotewicz et al. 1988).

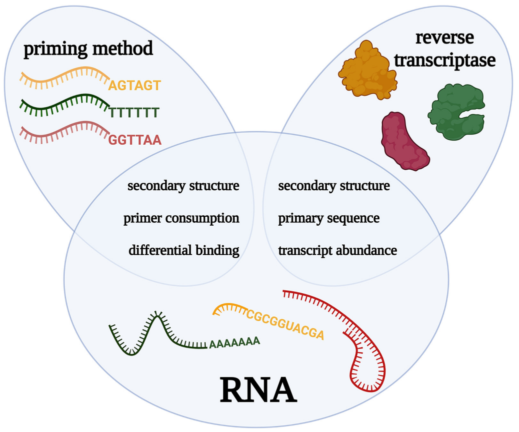

FIGURE 1. Overview of reverse transcription components and characteristics at the heart of intrasample quantification biases. Graphical overview of the intrasample biases akin to the reverse transcription reaction. The blue ellipses of "priming method" and "reverse transcriptase" overlap with the "RNA" circle as it is from the interaction of these factors that the biases originate. This figure was created using BioRender.

Commercially available RTases differ in sensitivity, specificity, reproducibility, and yield (Levesque-Sergerie et al. 2007; Sieber et al. 2010; Lindén et al. 2012; Zucha et al. 2020). These dissimilarities are inherent to the RTases due to their unequal capability of dealing with low template concentrations, specific RT conditions (such as temperature or buffer composition), and intensely structured RNAs. Therefore, before applying an RTase in the test tube, it is crucial to recognize its specific characteristics and tendencies. Most RTases have an RNase H subdomain responsible for hydrolyzing the RNA of the formed DNA:RNA duplex. Kotewicz et al. (1988) isolated an RTase lacking RNase H activity. Such recombinant enzymes synthesize cDNA from long RNA transcripts more efficiently (Kotewicz et al. 1988). However, discussions remain on the influence of the RNase H moiety of the RTase on the output of the RT reaction (Zucha et al. 2020). Enzymes with diminished or absent RNase H activity, such as Superscript IV (Thermo Fisher) and Maxima H Minus (Thermo Fisher), outperform others in terms of sensitivity, yield, and precision (Sieber et al. 2010; Zucha et al. 2020). RTases display a range of competencies in dealing with RNA folding and primary sequence, indicating an enzyme–structure and enzyme–sequence interaction (Brooks et al. 1995; Ståhlberg et al. 2004; Bustin et al. 2015; Minshall and Git 2020). More specifically, more than 100-fold cDNA yield differences can arise, purely derived from the RTase's handling of secondary structure (Ståhlberg et al. 2004). At higher RT temperatures, the RNA molecules dissociate, disrupting secondary structures. Therefore, enzymes have now been engineered to be more thermostable and which are able to operate at higher RT temperatures (Zucha et al. 2020). Consequently, we advise using thermostable RTases to mitigate RNA secondary structure biases.

Some researchers have demonstrated that the primary sequence of the RNA can introduce a bias during RT (Zheng et al. 2011). Several models exist that aim to remove this bias from the data. Li et al. (2010), for example, used the first several bases of the sequencing read (including the ones after the RT primer binding site) to predict up to 52% of the RNA abundance differences between transcripts. The random primer sequence only partially explains this bias: it is also observed in random primer-independent RNA-seq data, suggesting that the bases downstream from the priming site also play a role (Schwartz et al. 2011). However, because these studies rely on a PCR step as well, we cannot specifically pinpoint these biases to the RT reaction.

TABLE 1. Overview of reverse transcription (RT) issues summarizing the cause, effect, and associated solution

Cause	Effect	Solution
Intrasample bias		
RNA secondary structure	Underrepresentation of structured RNA	• Increase temperature • Use heat stable RTase • Bioinformatic normalization (Oshlack and Wakefield 2009; Hansen et al. 2010; Li et al. 2010; Srivastava and Chen 2010; Roberts et al. 2011; Zheng et al. 2011)
RTase sequence preference	Underrepresentation of GC-poor and long RNA	Bioinformatic normalization (Oshlack and Wakefield 2009; Hansen et al. 2010; Li et al. 2010; Srivastava and Chen 2010; Roberts et al. 2011; Zheng et al. 2011)
RTase sensitivity	Overrepresentation of abundant transcripts	Assess linearity and analytical sensitivity of RTase
Priming site accessibility	Underrepresentation of structured RNA or part of transcript	• Increase temperature • Use heat stable RTase • Bioinformatic normalization
Differential priming affinity	Overrepresentation of GC-rich RNA	Bioinformatic normalization (Oshlack and Wakefield 2009; Hansen et al. 2010; Li et al. 2010; Srivastava and Chen 2010; Roberts et al. 2011; Zheng et al. 2011)
Differential primer consumption	Underrepresentation of low abundant transcripts with similar sequence as high abundant transcripts	
Artifacts		
Template-switching	False-positive trans-splicing, fusion genes, poly(A) location, mRNAs and exitrons	• Direct RNA sequencing • RNase H- enzyme • Adjust RT temperature • Bioinformatic filtering
Modification-induced errors	False-positive SNP and RNA-editing	Bioinformatic filtering
Mispriming	• Off-target transcription • False-positive SNPs and RNA-editing	• Remove primed bases from read • Check for off-targets • Optimize annealing temperatures • RT-independent verification
Internal priming	False positive poly(A) locations and mRNAs	• Anchored primers • Bioinformatic filtering • TGIRT-based RT • RT-independent verification
Primer-independent priming	Number of different incorrect transcripts	No-primer control reaction

Primer-RNA bias

Depending on the RNA of interest, researchers can use different priming methods: oligo(dT) for reverse transcribing (mainly) mRNA (see below), targeted primers for transcripts of choice, or short random primers for any RNA. Still, these methods do not produce an unbiased cDNA representation of the RNA of interest. RNA molecules are highly structured and display numerous structural elements (Das 2021; Townshend et al. 2021). RNA secondary or tertiary structures can prevent priming. Researchers have exploited this observation by hybridizing short oligonucleotides to determine the structure of an RNA molecule of interest through the oligonucleotides' binding capabilities (Lewis and Doty 1970; Wrede et al. 1978). From this observation, we can hypothesize that linear, lowly structured RNAs dominate the cDNA pool. Mir and Southern (1999) presented the secondary structure bias in microarrays, where less structured RNAs preferentially bind to the array. Upon RT, this bias induces seemingly differential abundance among transcripts with contrasting secondary structures within the same sample due to primer inaccessibility (Mir and Southern 1999; Ståhlberg et al. 2004; Bustin et al. 2015). Even within the same transcript, the choice of the targeted region can result in markedly differing abundance estimations (Kuo et al. 1997). Again, researchers can use thermostable RTases combined with a high RT temperature to partially overcome RNA folding-dependent biases (Malboeuf et al. 2018; Minshall and Git 2020). Still, there is always the chance that highly structured regions refold during primer annealing. Alternatively, researchers could use a thermostable group II intron reverse transcriptase (TGIRT)-based protocol. This approach uses a DNA:RNA hybrid RT primer with one overhanging random deoxynucleotide. Since the hybrid primer can only prime a single RNA nucleotide at the 3′-end (using the overhanging nucleotide), the priming is minimally dependent on the secondary structure of the target (Xu et al. 2019; Begik et al. 2023).

Random primers have other issues. First, Minshall and Git (2020) discuss that specific random primers are consumed by the abundant transcripts they associate with, thereby leaving less of these primers for the remaining transcripts and, thus, introducing a bias. Second, although random primers are random in their sequence, they are not random in their RNA-binding capacities and introduce a sizeable bias into transcript abundance (Hansen et al. 2012; Minshall and Git 2020). Hansen et al. (2012) were the first to describe this bias and propose a correcting model. This model uses a random primer-specific weight for each read and adjusts the abundance levels accordingly to improve quantification accuracy. These phenomena show that random priming of total RNA samples often unreliably quantifies the RNA transcripts of interest (Lekanne Deprez et al. 2002).

Similarly, gene-specific primers can have contrasting binding capabilities (due to targeted primary sequence and structure) and the relative abundance of multiple transcripts in such a study should be interpreted cautiously.

INTERSAMPLE BIASES

Intersample biases arise from the inconsistent use of reagents or discrepancies between RNA samples, such as RNA quantity, integrity, and purity (absence of proteins, DNA, enzymatic inhibitors, complexing agents, or nucleases; Fleige and Pfaffl 2006). Standardized execution and reporting (following MIQE-like guidelines; Bustin et al. 2009), accurate data normalization, use of equally qualitative samples, use of identical reagents and reagent quantities, and proper quality control checkpoints can mostly alleviate these biases. Because of the comparative nature of intersample analyses, such as differential expression analyses, intrasample biases are canceled out when taking the appropriate measures, ensuring unbiased conclusions on transcript abundance differences. We will not be expanding on these biases in this review.

REVERSE TRANSCRIPTION ARTIFACTS

Instead of presenting a divergent cDNA pool in terms of transcript abundances, RT can also generate inaccurate cDNA sequences, that is, cDNA molecules structurally different from their RNA templates. These are referred to as RT artifacts. In this section, we will discuss the primary artifacts (Fig. 2) and provide means to reduce them to a minimum (also summarized in Table 1).

Template switching

Retroviruses are dependent on RTases for replication. The RTase scans the RNA sequence and synthesizes the complementary DNA (cDNA) sequence. However, the cDNA-RTase complex might bind to another RNA molecule with a sequence complementary to the cDNA's 3′-end mid-synthesis and proceed with synthesizing cDNA using the new RNA strand as a template. This phenomenon is called template switching (TS) and results in large deletions (intramolecular TS) or fused cDNA molecules (intermolecular TS) (Luo and Taylor 1990). Deletions resulting from intramolecular TS—which have recently been dubbed "falsitrons" (as in "false exitrons"; Schulz et al. 2021)—are typically characterized by the absence of a canonical splice site (Cocquet et al. 2006) and the presence of direct repeats flanking the deleted region (which often has a strong secondary structure bringing the repeats close together) (Pathak and Hu 1997). TS is certainly not a novel concept (Gilboa et al. 1979) and has been described in multiple retroviruses, such as HIV (DeStefano et al. 1992), but also in mitochondria (Sellem et al. 2000). This phenomenon is not confined to in vivo environments but also occurs when RTases are used in vitro. Unfortunately, these artifacts are hard to distinguish from real biological deletions, and care is advised when making claims about novel isoforms originating from a deletion. Such apparent deletions are caused by template switching, and there have been multiple efforts to quantify the extent of the problem in published isoforms (Geiszt et al. 2004; Cocquet et al. 2006; Schulz et al. 2021). In addition, when TS occurs on internal poly(A) stretches during oligo(dT)-based RT, artifacts similar to internal priming can be formed (Balázs et al. 2019). Lastly, TS events can resemble trans-splicing or back-splicing (on which circRNA formation relies) (McManus et al. 2010; Yu et al. 2014) or apparent fusion genes (Houseley and Tollervey 2010). There are, however, multiple options to reduce TS. First, the researcher can use a RTase lacking or having reduced RNase H activity. RNase H is essential to template switching as it hydrolyzes the RNA of the DNA:RNA duplex during RT, freeing up the cDNA and allowing it to find another RNA sequence to bind (Luo and Taylor 1990). Some discussion exists to what extent RNase H-negative RTases can still perform TS (Luo and Taylor 1990; Garces and Wittek 1991;

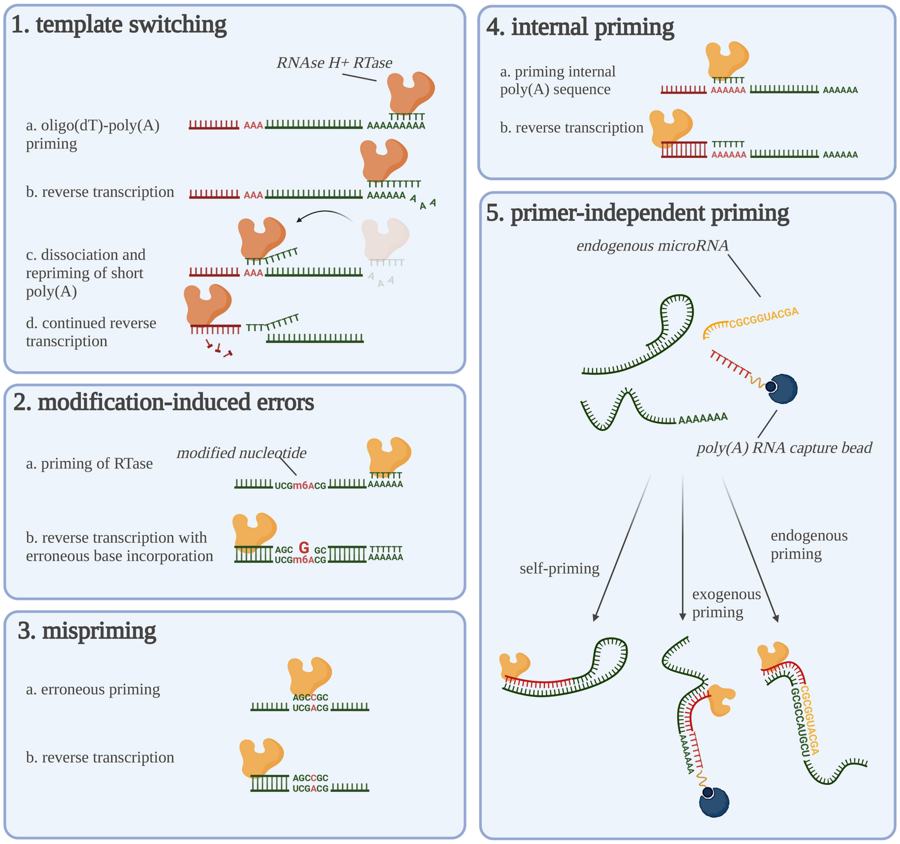

FIGURE 2. Overview of reverse transcription artifacts. Graphical overview of the five discussed priming artifacts. The reverse transcriptases are shown as yellow proteins (note the difference between yellow and orange reverse transcriptases). This figure was created using BioRender.

Zaphiropoulos 2002; Zeng and Wang 2002; Zhu et al. 2002). Second, one can tweak some of the parameters of the RT reaction. Short reaction times, low RTase concentration, and low reaction temperature reduce the extent of TS (Ouhammouch and Brody 1992). TS is also modulated by the concentration of the acceptor RNA molecule (to which the RTase "jumps"), the number of nucleotides overlapping between the two RNA molecules (Luo and Taylor 1990), and the size of the "falsitron" (Delviks and Pathak 1999), but these are typically out of the researcher's hands. Last, suspicious results should always be validated by using a non-RT-dependent technique such as northern blot (Houseley and Tollervey 2010), direct RNA sequencing (Schulz et al. 2021), in situ hybridization, or RNA-templated DNA ligation.

Modification-induced errors

Even when the RTase correctly primes the transcripts of interest, it can still introduce errors. Actually, the errors generated by RTases contribute to the high mutational rate in retroviruses driving evolution in vivo (Mizutani and Temin 1976; Roberts et al. 1988; Svarovskaia et al. 2003). As RNA–cDNA discrepancies are usually not desired in RNA characterization, researchers have developed higher-fidelity RTases for in vitro applications (for review, see Svarovskaia et al. 2003). Several studies have quantified these errors for various RTases. When these errors happen randomly, they tend not to present significant problems. When these RTase errors always occur at the same position, the phenomenon becomes more problematic. Single-nucleotide polymorphism (SNP) calling from RNA-seq data, for example, requires a robust detection of a specifically altered nucleotide at a specific position. Without any accompanying data on the sequence of the native RNA and DNA, it is impossible to differentiate between RNA-editing events, genuine SNPs, and RT-induced sequence errors. Often, when these errors occur at specific positions, they result from the RTase misinterpreting modified RNA bases. Potapov et al. (2018) analyzed the mistakes made by three different RTases on modified and unmodified RNA. They noted that the presence of a pseudouridine (Ψ), 5-hydroxylmethyluridine (hm5U), or N^6-Methyladenosine (m6A) nucleotides significantly increased the error rate of the RTase. To add to the problem, they discovered that specific RNA modifications differentially influence different RTases. ProtoScript II and AMV RTases, for example, significantly increase adenosine substitutions (Potapov et al. 2018).

Researchers should apply various approaches to avoid making an incorrect assumption about a specific nucleotide and its potential modifications. Machine learning approaches have been developed to infer RNA modifications from RNA sequencing data, with varying success (Ryvkin et al. 2013; Werner et al. 2020; Tan et al. 2021).

Mispriming

Mispriming occurs when oligonucleotides prime noncomplementary RNA sequences, introducing mismatches in the resulting cDNA, typically in the first few nucleotides of the sequencing read (Zhang and Byrne 1999; Piskol et al. 2013; Van Gurp et al. 2013). Mispriming can arise when using oligo(dT) primers (as is the case with internal poly(A) stretches), when targeting a specific transcript of interest or sequencing adapters (leading to off-target effects) (Shivram and Iyer 2018), or when using random primers (Van Gurp et al. 2013). It happens nonrandomly with different nucleotides having differential sensitivity for mispriming (uracil residues are prone to mispriming, while cytosine residues prevent mispriming) (Van Gurp et al. 2013). This problem can often be alleviated by computationally trimming a set number of bases from the 5′ read end to remove the priming site from the final read.

Internal priming

mRNA quantification often relies on oligo(dT) RT primers that target the 3′-end of polyadenylated mRNAs. This type of priming is increasingly being used due to the rising popularity of cost-effective 3′-end sequencing at the tissue or single-cell level. However, nothing withholds the RTase and its primer from binding to an internal poly(A)-rich sequence (i.e., any poly(A) sequence not at the end of the transcript), thereby generating RT artifacts that can be falsely identified as mRNA molecules (i.e., when [contaminating] DNA is primed) or as alternative polyadenylation sites (Nam et al. 2002). However, a more recent article, challenged the hypothesis that these artifacts originate from internal priming by implying the internal poly(A) sequences are often too short for preferential binding; instead, they result from template switching after poly(A)-tail binding and dissociation (Balázs et al. 2019). Although this problem can (in part) be avoided using anchored oligo(dT) primers (Nam et al. 2002), these are not readily used and do not come standard in most RNA sequencing kits. The generated artifacts can also be filtered out post-sequencing during the bioinformatic analysis by inspecting the neighborhood of the transcript for genomic poly(A)-rich regions (Beaudoing et al. 2000; Tian et al. 2005; Zhang et al. 2005; Shepard et al. 2011; Graber et al. 2013; Wilkening et al. 2013). Recently, researchers developed specialized programs designed to remove internal priming artifacts and generated a list of transcripts prone to incorrect quantification due to internal priming (Svoboda et al. 2022). Of note, these artifacts are not always useless as RNA velocity analysis exploits them to determine the number of nonspliced transcripts (La Manno et al. 2018). To overcome internal priming during the library preparation, researchers can opt to use a TGIRT-based protocol (as explained above). In such setups, the RT primer exclusively binds to the 3′-end of the RNA transcript (Xu et al. 2019; Begik et al. 2023). Note this approach incorporates all transcripts instead of focusing on polyadenylated mRNA.

Primer-independent priming

Even without a purposely administered primer, RTases produce transcribed cDNA (Bernstein et al. 1983; Agranovsky 1992; Peyrefitte et al. 2003; Khraiwesh et al. 2010; Tuiskunen et al. 2010; Moison et al. 2011; Piché and Schernthaner 2018; Timofeeva and Skrypina 2018). These artifacts can have three origins: RNA self-priming (Bernstein et al. 1983; Tuiskunen et al. 2010; Timofeeva and Skrypina 2018), priming by endogenous nucleic acids (Khraiwesh et al. 2010; Timofeeva and Skrypina 2018), or priming by (contaminating) exogenous nucleic acids (such as transfer RNAs [tRNAs] or DNA present in RTase preparations) (Agranovsky 1992; Goldenberger and Altwegg 1995; Moison et al. 2011; Piché and Schernthaner 2018). Self-priming occurs when secondary structures are formed, creating double-stranded regions which the RTase recognizes as primed regions. The RT products generated from the double-stranded regions may be incorrectly classified as pseudogenes (Bernstein et al. 1983) or antisense products (Tuiskunen et al. 2010)—the latter is especially problematic during the strand-specific priming of viral RNA genomes, where strand-specific detection is critical. Finally, endogenous DNA fragments or RNA, such as microRNAs or tRNA (fragments), can initiate RT when they bind to other RNA transcripts (Agranovsky 1992; Khraiwesh et al. 2010). Exogenous nucleic acids from an upstream step or reagents can start the RT reaction. Piché and Schernthaner (2018) uncovered an intriguing example of this phenomenon: carry-over oligo(dT) oligonucleotides originating from mRNA extraction primed poly(A) tails during a primer-free RT reaction. Additionally, multiple available RTase solutions have repeatedly been shown to contain contaminating small RNA molecules that can initiate RT without an RT primer (Agranovsky 1992; Moison et al. 2011). These molecules are hypothesized to be highly associated with the RTase and are therefore difficult to remove (Moison et al. 2011). Generally, a "no-primer" control identifies said primer-independent RT products, after which they may be excluded from the analysis.

CONCLUSION

The simple linear and perfect conversion of RNA to cDNA is not a reality and should therefore be considered a black-box reaction to date. It relies on multiple factors and depends on close interactions between these different factors. This review highlights how the synthesized cDNA pool might differ from the original RNA repertoire. Where "RT biases" correspond to relatively different cDNA abundances of correctly transcribed molecules, the term "RT artifacts" refers to aberrant cDNA molecules. These can arise from incorrect primer binding or unexpected RTase behavior. We advise detailed transparent reporting and dedicated verification of (surprising) results using RT-independent techniques to reveal the mentioned biases and artifacts. We hope this review can incentivize researchers to be critical of blind trust in the RT reaction.

ACKNOWLEDGMENTS

The authors would like to thank their peers Marieke Vromman and Kathleen Schoofs for proofreading the final manuscript for correctness and clarity.

REFERENCES

Agranovsky AA. 1992. Exogenous primer-independent cDNA synthesis with commercial reverse transcriptase preparations on plant virus RNA templates. *Anal Biochem* **203:** 163–165. doi:10.1016/0003-2697(92)90058-F

Balázs Z, Tombácz D, Csabai Z, Moldován N, Snyder M, Boldogkoi Z. 2019. Template-switching artifacts resemble alternative polyadenylation. *BMC Genomics* **20:** 824. doi:10.1186/s12864-019-6199-7

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Res* **10:** 1001. doi:10.1101/gr.10.7.1001

Begik O, Diensthuber G, Liu H, Delgado-Tejedor A, Kontur C, Niazi AM, Valen E, Giraldez AJ, Beaudoin J-D, Mattick JS, Novoa EM. 2023. Nano3P-seq: transcriptome-wide analysis of gene expression and tail dynamics using end-capture nanopore cDNA sequencing. *Nat Methods* **20:** 75–85. doi:10.1038/s41592-022-01714-w

Bernstein LB, Mount SM, Weiner AM. 1983. Pseudogenes for human small nuclear RNA U3 appear to arise by integration of self-primed reverse transcripts of the RNA into new chromosomal sites. *Cell* **32:** 461–472. doi:10.1016/0092-8674(83)90466-X

Brooks EM, Sheflin LG, Spaulding SW. 1995. Secondary structure in the 3′ UTR of EGF and the choice of reverse transcriptases affect the detection of message diversity by RT-PCR. *BioTechniques* **19:** 806–812.

Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, et al. 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* **55:** 611–622. doi:10.1373/clinchem.2008.112797

Bustin S, Dhillon HS, Kirvell S, Greenwood C, Parker M, Shipley GL, Nolan T. 2015. Variability of the reverse transcription step: practical implications. *Clin Chem* **61:** 202–212. doi:10.1373/clinchem.2014.230615

Cocquet J, Chong A, Zhang G, Veitia RA. 2006. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88:** 127–131. doi:10.1016/j.ygeno.2005.12.013

Das R. 2021. RNA structure: a renaissance begins? *Nat Methods* **18:** 439. doi:10.1038/s41592-021-01132-4

Delviks KA, Pathak VK. 1999. Effect of distance between homologous sequences and 3′ homology on the frequency of retroviral reverse transcriptase template switching. *J Virol* **73:** 7923–7932. doi:10.1128/JVI.73.10.7923-7932.1999

DeStefano JJ, Mallaber LM, Rodriguez-Rodriguez L, Fay PJ, Bambara RA. 1992. Requirements for strand transfer between internal regions of heteropolymer templates by human immunodeficiency virus reverse transcriptase. *J Virol* **66:** 6370–6378. doi:10.1128/jvi.66.11.6370-6378.1992

Fleige S, Pfaffl MW. 2006. RNA integrity and the effect on the real-time qRT-PCR performance. *Mol Aspects Med* **27:** 126–139. doi:10.1016/j.mam.2005.12.003

Garces J, Wittek R. 1991. Reverse-transcriptase-associated RNaseH activity mediates template switching during reverse transcription *in vitro*. *Proc R Soc London Ser B Biol Sci* **243:** 235–239. doi:10.1098/rspb.1991.0037

Geiszt M, Lekstrom K, Leto TL. 2004. Analysis of mRNA transcripts from the NAD(P)H oxidase 1 (*Nox1*) gene: evidence against production of the NAPDH oxidase homolog-1 short (NOH-1S) transcript variant. *J Biol Chem* **279:** 51661–51668. doi:10.1074/jbc.M409325200

Gilboa E, Mitra SW, Goff S, Baltimore D. 1979. A detailed model of reverse transcription and tests of crucial aspects. *Cell* **18:** 93–100. doi:10.1016/0092-8674(79)90357-X

Goldenberger D, Altwegg M. 1995. Eubacterial PCR: contaminating DNA in primer preparations and its elimination by UV light. *J Microbiol Methods* **21:** 27–32. doi:10.1016/0167-7012(94)00024-2

Graber JH, Nazeer FI, Yeh PC, Kuehner JN, Borikar S, Hoskinson D, Moore CL. 2013. DNA damage induces targeted, genome-wide variation of poly(A) sites in budding yeast. *Genome Res* **23:** 1690–1703. doi:10.1101/gr.144964.112

Hansen KD, Brenner SE, Dudoit S. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38:** e131. doi:10.1093/nar/gkq224

Hansen KD, Irizarry RA, Wu Z. 2012. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13:** 204–216. doi:10.1093/biostatistics/kxr054

Houseley J, Tollervey D. 2010. Apparent non-canonical trans-splicing is generated by reverse transcriptase *in vitro*. *PLoS One* **5:** e12271. doi:10.1371/journal.pone.0012271

Khraiwesh B, Arif MA, Seumel GI, Ossowski S, Weigel D, Reski R, Frank W. 2010. Transcriptional control of gene expression by microRNAs. *Cell* **140:** 111–122. doi:10.1016/j.cell.2009.12.023

Kotewicz ML, Sampson CM, D'alessio JM, Gerard GF. 1988. Isolation of cloned Moloney murine leukemia virus reverse transcriptase lacking ribonuclease H activity. *Nucleic Acids Res* **16:** 265–277. doi:10.1093/nar/16.1.265

Kuo KW, Leung MF, Leung WC. 1997. Intrinsic secondary structure of human TNFR-I mRNA influences the determination of gene expression by RT-PCR. *Mol Cell Biochem* **177:** 1–6. doi:10.1023/A:1006862304381

La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A, et al. 2018. RNA velocity of single cells. *Nature* **560:** 494–498. doi:10.1038/s41586-018-0414-6

Lekanne Deprez RH, Fijnvandraat AC, Ruijter JM, Moorman AFM. 2002. Sensitivity and accuracy of quantitative real-time polymerase chain reaction using SYBR green I depends on cDNA synthesis conditions. *Anal Biochem* **307:** 63–69. doi:10.1016/S0003-2697(02)00021-0

Levesque-Sergerie JP, Duquette M, Thibault C, Delbecchi L, Bissonnette N. 2007. Detection limits of several commercial reverse transcriptase enzymes: impact on the low- and high-abundance transcript levels assessed by quantitative RT-PCR. *BMC Mol Biol* **8:** 93. doi:10.1186/1471-2199-8-93

Lewis JB, Doty P. 1970. Derivation of the secondary structure of 5S RNA from its binding of complementary oligonucleotides. *Nature* **225:** 510–512. doi:10.1038/225510a0

Li J, Jiang H, Wong WH. 2010. Modeling non-uniformity in short-read rates in RNA-seq data. *Genome Biol* **11:** R50. doi:10.1186/gb-2010-11-5-r50

Lindén J, Ranta J, Pohjanvirta R. 2012. Bayesian modeling of reproducibility and robustness of RNA reverse transcription and quantitative real-time polymerase chain reaction. *Anal Biochem* **428:** 81–91. doi:10.1016/j.ab.2012.06.010

Luo GX, Taylor J. 1990. Template switching by reverse transcriptase during DNA synthesis. *J Virol* **64:** 4321–4328. doi:10.1128/jvi.64.9.4321-4328.1990

Malboeuf CM, Isaacs SJ, Tran NH, Kim B. 2018. Thermal effects on reverse transcription: improvement of accuracy and processivity in cDNA synthesis. *BioTechniques* **30:** 1074–1084. doi:10.2144/01305rr06

McManus CJ, Duff MO, Eipper-Mains J, Graveley BR. 2010. Global analysis of trans-splicing in *Drosophila*. *Proc Natl Acad Sci* **107:** 12975–12979. doi:10.1073/pnas.1007586107

Minshall N, Git A. 2020. Enzyme- and gene-specific biases in reverse transcription of RNA raise concerns for evaluating gene expression. *Sci Rep* **10:** 8151. doi:10.1038/s41598-020-65005-0

Mir KU, Southern EM. 1999. Determining the influence of structure on hybridization using oligonucleotide arrays. *Nat Biotechnol* **17:** 788–792. doi:10.1038/11732

Mizutani S, Temin HM. 1976. Incorporation of noncomplementary nucleotides at high frequencies by ribodeoxyvirus DNA polymerases and *Escherichia coli* DNA polymerase I. *Biochemistry* **15:** 1510–1516. doi:10.1021/bi00652a023

Moison C, Arimondo PB, Guieysse-Peugeot AL. 2011. Commercial reverse transcriptase as source of false-positive strand-specific RNA detection in human cells. *Biochimie* **93:** 1731–1737. doi:10.1016/j.biochi.2011.06.005

Nam DK, Lee S, Zhou G, Cao X, Wang C, Clark T, Chen J, Rowley JD, Wang SM. 2002. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc Natl Acad Sci* **99:** 6152–6156. doi:10.1073/pnas.092140899

Oshlack A, Wakefield MJ. 2009. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* **4:** 14. doi:10.1186/1745-6150-4-14

Ouhammouch M, Brody EN. 1992. Temperature-dependent template switching during in vitro cDNA synthesis by the AMV-reverse transcriptase. *Nucleic Acids Res* **20:** 5443–5450. doi:10.1093/nar/20.20.5443

Pathak VK, Hu WS. 1997. "Might as well jump!" template switching by retroviral reverse transcriptase, defective genome formation, and recombination. *Semin Virol* **8:** 141–150. doi:10.1006/smvy.1997.0114

Peyrefitte CN, Pastorino B, Bessaud M, Tolou HJ, Couissinier-Paris P. 2003. Evidence for *in vitro* falsely-primed cDNAs that prevent specific detection of virus negative strand RNAs in dengue-infected cells: improvement by tagged RT-PCR. *J Virol Methods* **113:** 19–28. doi:10.1016/S0166-0934(03)00218-0

Piché C, Schernthaner JP. 2018. Background priming during reverse transcription by oligo(dT) carried over from mRNA isolation. *BioTechniques* **34:** 720–724. doi:10.2144/03344bm08

Piskol R, Ramaswami G, Li JB. 2013. Reliable identification of genomic variants from RNA-seq data. *Am J Hum Genet* **93:** 641–651. doi:10.1016/j.ajhg.2013.08.008

Potapov V, Fu X, Dai N, Corrêa IR, Tanner NA, Ong JL. 2018. Base modifications affecting RNA polymerase and reverse transcriptase fidelity. *Nucleic Acids Res* **46:** 5753–5763. doi:10.1093/nar/gky341

Roberts JD, Bebenek K, Kunkel TA, Roberts JD, Bebenek K, Kunkel TA. 1988. The accuracy of reverse transcriptase from HIV-1 published. *Science* **242:** 1171–1173. doi:10.1126/science.2460925

Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. 2011. Improving RNA-seq expression estimates by correcting for fragment bias. *Genome Biol* **12:** R22. doi:10.1186/gb-2011-12-3-r22

Ryvkin P, Leung YY, Silverman IM, Childress M, Valladares O, Dragomir I, Gregory BD, Wang LS. 2013. HAMR: high-throughput annotation of modified ribonucleotides. *RNA* **19:** 1684–1692. doi:10.1261/rna.036806.112

Schulz L, Torres-Diz M, Cortés-López M, Hayer KE, Asnani M, Tasian SK, Barash Y, Sotillo E, Zarnack K, König J, et al. 2021. Direct long-read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts. *Genome Biol* **22:** 190. doi:10.1186/s13059-021-02411-1

Schwartz S, Oren R, Ast G. 2011. Detection and removal of biases in the analysis of next-generation sequencing reads. *PLoS One* **6:** e16685. doi:10.1371/journal.pone.0016685

Sellem CH, Begel O, Sainsard-Chanet A. 2000. Recombinant mitochondrial DNA molecules suggest a template switching ability for group-II-intron reverse transcriptase. *Curr Genet* **37:** 24–28. doi:10.1007/s002940050003

Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-seq. *RNA* **17:** 761–772. doi:10.1261/rna.2581711

Shivram H, Iyer VR. 2018. Identification and removal of sequencing artifacts produced by mispriming during reverse transcription in multiple RNA-seq technologies. *RNA* **24:** 1266–1274. doi:10.1261/rna.066217.118

Sieber MW, Recknagel P, Glaser F, Witte OW, Bauer M, Claus RA, Frahm C. 2010. Substantial performance discrepancies among commercially available kits for reverse transcription quantitative polymerase chain reaction: a systematic comparative investigator-driven approach. *Anal Biochem* **401:** 303–311. doi:10.1016/j.ab.2010.03.007

Smyth RP, Negroni M, Lever AM, Mak J, Kenyon JC. 2018. RNA structure--a neglected puppet master for the evolution of virus and host immunity. *Front Immunol* **9:** 2097. doi:10.3389/fimmu.2018.02097

Srivastava S, Chen L. 2010. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res* **38:** e170. doi:10.1093/nar/gkq670

Ståhlberg A, Kubista M, Pfaffl M. 2004. Comparison of reverse transcriptases in gene expression analysis. *Clin Chem* **50:** 1678–1680. doi:10.1373/clinchem.2004.035469

Svarovskaia SE, Cheslock SR, Zhang W-H, Hu W-S, Pathak VK. 2003. Retroviral mutation rates and reverse transcriptase fidelity. *Front Biosci* **8:** 117–134. doi:10.2741/957

Svoboda M, Frost HR, Bosco G. 2022. Internal oligo(dT) priming introduces systematic bias in bulk and single-cell RNA sequencing count data. *NAR Genom Bioinform* **4:** lqac035. doi:10.1093/nargab/lqac035.

Tan KT, Ding LW, Wu CS, Tenen DG, Yang H. 2021. Repurposing RNA sequencing for discovery of RNA modifications in clinical cohorts. *Sci Adv* **7:** eabd2605. doi:10.1126/sciadv.abd2605

Tian B, Hu J, Zhang H, Lutz CS. 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* **33:** 201–212. doi:10.1093/nar/gki158

Timofeeva AV, Skrypina NA. 2018. Background activity of reverse transcriptases. *BioTechniques* **30:** 22–28. doi:10.2144/01301bm02

Townshend RJL, Eismann S, Watkins AM, Rangan R, Karelina M, Das R, Dror RO. 2021. Geometric deep learning of RNA structure. *Science* **373:** 1047–1051. doi:10.1126/science.abe5650

Tuiskunen A, Leparc-Goffart I, Boubis L, Monteil V, Klingström J, Tolou HJ, Lundkvist A, Plumet S. 2010. Self-priming of reverse transcriptase impairs strand-specific detection of dengue virus RNA. *J Gen Virol* **91:** 1019–1027. doi:10.1099/vir.0.016667-0

Van Gurp TP, McIntyre LM, Verhoeven KJF. 2013. Consistent errors in first strand cDNA due to random hexamer mispriming. *PLoS ONE* **8:** 85583. doi:10.1371/journal.pone.0085583

Werner S, Schmidt L, Marchand V, Kemmer T, Falschlunger C, Sednev MV, Bec G, Ennifar E, Höbartner C, Micura R, et al. 2020. Machine learning of reverse transcription signatures of variegated polymerases allows mapping and discrimination of methylated purines in limited transcriptomes. *Nucleic Acids Res* **48:** 3734–3746. doi:10.1093/nar/gkaa113

Wilkening S, Pelechano V, Järvelin AI, Tekkedil MM, Anders S, Benes V, Steinmetz LM. 2013. An efficient method for genome-wide polyadenylation site mapping and RNA quantification. *Nucleic Acids Res* **41:** e65. doi:10.1093/nar/gks1249

Wrede P, Pongs O, Erdmann VA. 1978. Binding oligonucleotides to *Escherichia coli* and *Bacillus stearothermophilus* 5 S RNA. *J Mol Biol* **120:** 83–96. doi:10.1016/0022-2836(78)90296-6

Xu H, Yao J, Wu DC, Lambowitz AM. 2019. Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. *Sci Rep* **9:** 7953. doi:10.1038/s41598-019-44457-z

Yu CY, Liu HJ, Hung LY, Kuo HC, Chuang TJ. 2014. Is an observed non-co-linear RNA product spliced in *trans*, in *cis* or just *in vitro*? *Nucleic Acids Res* **42:** 9410–9423. doi:10.1093/nar/gku643

Zaphiropoulos PG. 2002. Template switching generated during reverse transcription? *FEBS Lett* **527:** 326. doi:10.1016/S0014-5793(02)03239-8

Zeng XC, Wang SX. 2002. Evidence that BmTXKβ-BmKCT cDNA from Chinese scorpion *Buthus martensii* Karsch is an artifact generated in the reverse transcription process. *FEBS Lett* **520:** 183–184. doi:10.1016/S0014-5793(02)02812-0

Zhang J, Byrne CD. 1999. Differential priming of RNA templates during cDNA synthesis markedly affects both accuracy and reproducibility of quantitative competitive reverse-transcriptase PCR. *Biochem J* **337:** 231–241. doi:10.1042/bj3370231

Zhang H, Hu J, Recce M, Tian B. 2005. PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res* **33:** D116–D120. doi:10.1093/nar/gki055

Zheng W, Chung LM, Zhao H. 2011. Bias detection and correction in RNA-sequencing data. *BMC Bioinformatics* **12:** 290. doi:10.1186/1471-2105-12-290

Zhu S, Li W, Cao Z. 2002. Does MMLV-RT lacking RNase H activity have the capability of switching templates during reverse transcription? *FEBS Lett* **520:** 185. doi:10.1016/S0014-5793(02)02813-2

Zucha D, Androvic P, Kubista M, Valihrach L. 2020. Performance comparison of reverse transcriptases for single-cell studies. *Clin Chem* **66:** 217–228. doi:10.1373/clinchem.2019.307835

RNA

A PUBLICATION OF THE RNA SOCIETY

Artifacts and biases of the reverse transcription reaction in RNA sequencing

Jasper Verwilt, Pieter Mestdagh and Jo Vandesompele

RNA 2023 29: 889-897 originally published online March 29, 2023
Access the most recent version at doi:10.1261/rna.079623.123