Review

Next-Generation Sequencing Methods to Determine the Accuracy of Retroviral Reverse Transcriptases: Advantages and Limitations

Javier Martínez del Río *[ID] **and Luis Menéndez-Arias** *[ID]

Centro de Biología Molecular Severo Ochoa, Consejo Superior de Investigaciones Científicas & Universidad Autónoma de Madrid, c/Nicolás Cabrera 1, 28049 Madrid, Spain

*   Correspondence: javier.martinez@cbm.csic.es (J.M.d.R.); lmenendez@cbm.csic.es (L.M.-A.)

Abstract: Retroviruses, like other RNA viruses, mutate at very high rates and exist as genetically heterogeneous populations. The error-prone activity of viral reverse transcriptase (RT) is largely responsible for the observed variability, most notably in HIV-1. In addition, RTs are widely used in biotechnology to detect RNAs and to clone expressed genes, among many other applications. The fidelity of retroviral RTs has been traditionally analyzed using enzymatic (gel-based) or reporter-based assays. However, these methods are laborious and have important limitations. The development of next-generation sequencing (NGS) technologies opened the possibility of obtaining reverse transcription error rates from a large number of sequences, although appropriate protocols had to be developed. In this review, we summarize the developments in this field that allowed the determination of RNA-dependent DNA synthesis error rates for different RTs (viral and non-viral), including methods such as PRIMER IDs, REP-SEQ, ARC-SEQ, CIR-SEQ, SMRT-SEQ and ROLL-SEQ. Their advantages and limitations are discussed. Complementary DNA (cDNA) synthesis error rates obtained in different studies, using RTs and RNAs of diverse origins, are presented and compared. Future improvements in methodological pipelines will be needed for the precise identification of mutations in the RNA template, including modified bases.

Keywords: reverse transcriptase; next-generation sequencing; fidelity of DNA synthesis; retrovirus; cDNA synthesis

1. Introduction

RNA viruses replicate as complex distributions of closely related variant genomes known as quasispecies [1]. These mutant swarms are subjected to genetic variation, competition and selection, and their diversity is fueled by the high mutation rates of RNA viruses. Genetic variability is one of the hallmarks of human immunodeficiency virus (HIV) infection. The large diversity of HIV is a consequence of its high mutation and recombination rates, the high replication rate of the virus leading to large progenies, and to selective forces that operate during viral dissemination (e.g., immune pressure, host factors, etc.). Estimated mutation rates for HIV type 1 (HIV-1) obtained from studies involving within-host longitudinal samples of DNA sequences could be as high as 10^{-3} nucleotide substitutions per site per year [2–4], although these rates could be different depending on the specific genomic region analyzed [5].

The genome of HIV is composed of two copies of non-covalently linked, unspliced, positive-sense single-stranded RNA (ssRNA). The viral genomic RNA molecules resemble messenger RNA, bearing a 5′ cap (Gppp) and a 3′ poly(A) tail. Each RNA has a length

of approximately 10 kB and contains several open reading frames, including the characteristic retroviral genes *gag*, *pol* and *env*. These genes encode polyproteins Gag, Pol and Env that, after cleavage, produce structural proteins, viral enzymes and envelope glycoprotein, respectively.

A reverse transcriptase (RT) encoded within the *pol* gene is responsible for the conversion of the genomic ssRNA into double-stranded DNA (dsDNA) that integrates into the host's chromosome. This process, known as reverse transcription, is relatively complex (for reviews, see [6,7]) and requires the participation of the RNA- and DNA-dependent DNA polymerase and ribonuclease H (RNase H) activities of the RT. Unlike cellular DNA polymerases, retroviral RTs are devoid of $3'{\rightarrow}5'$ exonucleolytic proofreading activity and are therefore less accurate than prokaryote and eukaryote DNA polymerases while synthesizing DNA (for reviews, see [8–10]). Although retroviral variation is largely due to errors made by the RT during reverse transcription, cellular RNA polymerase II and a series of viral and host factors also contribute to the retroviral mutation rate [11]. Unbalanced nucleotide pools inside the infected cells can facilitate the emergence of specific mutations [12]. More importantly, cytidine deaminases of the APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide) family, such as APOBEC3F and APOBEC3G, are potent inhibitors of HIV-1 infection, while promoting hypermutation G$\rightarrow$A in plus-stranded cDNA during reverse transcription [13]. On the other hand, adenosine deaminases acting on RNA (ADAR) play a dynamic and nuanced role in regulating transcriptomes and are expected to contribute to increase viral diversity, as demonstrated in different RNA viruses, including HIV [14,15].

Studies on the fidelity of RTs have been justified by the importance of HIV-1 RT as a target of antiretroviral therapy [16], as well as for the pivotal role of retroviral enzymes in biotechnological applications, from conventional RT-PCR to transcriptomics and prime editing [17]. The accuracy of DNA synthesis catalyzed by purified RTs has traditionally been determined using enzymological (gel-based) assays or by measuring mutant frequencies in a reporter gene (typically *lacZα* in forward mutation assays) (reviewed in [9,10]). However, these methods are laborious and time-consuming.

The introduction of next-generation sequencing (NGS) technologies has facilitated the analysis of a large number of sequences and the detection of many mutations (as well as minority variants) resulting from single experiments. However, these techniques have some important limitations resulting from their inherent mutational background. Improvements in the methodological workflow, such as, for example, the use of barcodes or unique molecular identifiers (UMIs), can be helpful to extract relevant data on the accuracy of reverse transcription in vitro. In this review, we describe different approaches and NGS-based methods to determine the fidelity of RTs and discuss their advantages and limitations.

2. RT Structure

Retroviral RTs are structurally diverse. The best characterized enzyme is the HIV-1 RT. Currently, the number of crystal structures of this enzyme deposited in the Protein Data Bank is close to 200 and includes complexes of wild-type and mutant RTs with selected inhibitors, as well as binary (RT-DNA/DNA and RT-RNA/DNA) and ternary complexes (e.g., RT-DNA/DNA-dNTP) containing different types of nucleic acids (for a recent review, see [18]). The HIV-1 RT is an asymmetric heterodimer composed of subunits of 66 and 51 kDa, termed as p66 and p51, respectively (Figure 1). The p66 subunit contains a DNA polymerase domain spanning residues 1–440 and an RNase H domain including residues 441–560. Catalytic sites of the DNA polymerase and RNase H activities are located in p66. Active DNA polymerase residues are Asp110, Asp185 and Asp186, while

the catalytic residues responsible for the cleavage of RNA in RNA/DNA complexes are Asp444, Glu478, Asp498 and Asp549. Both catalytic sites coordinate divalent cations (Mg^{2+} in physiological conditions, although Mn^{2+} binding can be observed in some structures). The p51 subunit has the same amino acid sequence as p66 but lacks the RNase H domain and has a structural role while contributing to enzyme dimerization [19].
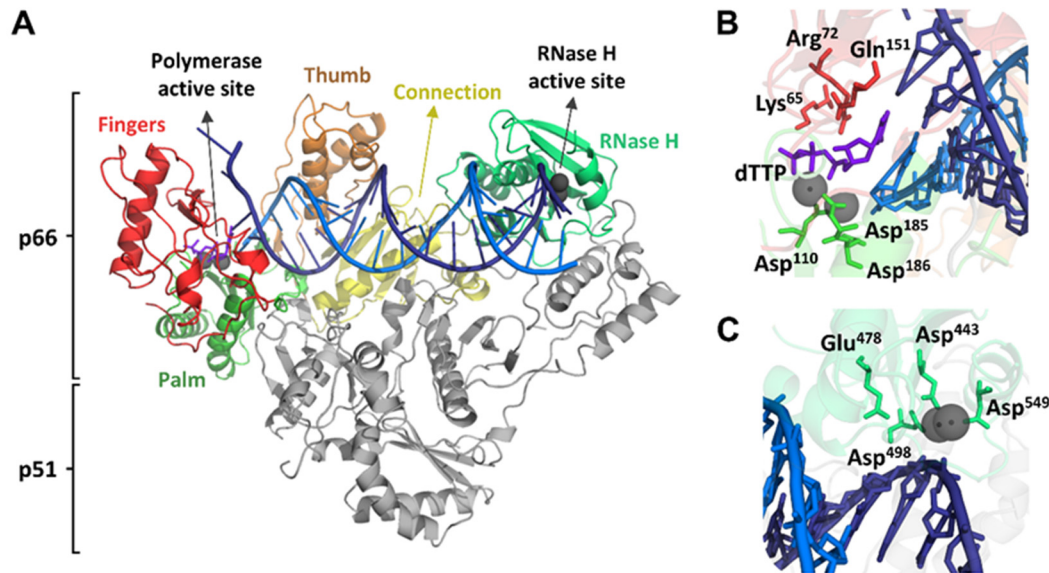

Figure 1. HIV-1 RT structure. (**A**) Crystal structure of HIV-1 RT bound to a DNA primer-template and an incoming deoxythymidine triphosphate (dTTP) (PDB ID: 1RTD). Subdomains of the p66 DNA polymerase domain are shown as red (fingers), green (palm), orange (thumb) and yellow (connection). The RNase H domain is shown as mint green. The p51 subunit is colored gray. The template strand is represented in dark blue, the primer strand in light blue, and the incoming dTTP is shown as purple. Coordinating metal ions (Mg^{2+}) in the DNA polymerase and RNase H active sites are represented as dark gray spheres. (**B**) Detailed view of the DNA polymerase active site, showing the location of the catalytic triad Asp110, Asp185 and Asp186, along with key residues Lys65, Arg72 and Gln151, the incoming dTTP, the template-primer and the two metal ions. (**C**) Detailed view of the RNase H active site, showing the location of Asp443, Glu478, Asp498 and Asp549 (DEDD motif) interacting with the metal ions and the template-primer.

As expected, similar crystal structures have been determined for closely related RTs, such as those of HIV-2 [20] or feline immunodeficiency virus (FIV) [21]. Similar to lentiviral RTs, Baldwin et al. [22] recently reported the crystal structure of a ternary complex containing the human endogenous retrovirus-K (HERV-K) RT bound to double-stranded DNA and an incoming dNTP. Its structure shows a striking similarity to the HIV-1 RT asymmetric heterodimer.

Other examples of heterodimeric retroviral RTs are those of alpharetroviruses, such as Rous sarcoma virus (RSV), avian leucosis virus (ALV) or avian myeloblastosis virus (AMV). AMV RT is a heterodimer consisting of a 63-kDa α subunit and a 95-kDa β subunit. The α subunit contains the DNA polymerase and RNase H domains of RT, while the β subunit has an additional C-terminal integrase domain [23].

In contrast, murine leukemia virus (MLV) RT is a monomer of 80 kDa. Wild-type and recombinant variants of MLV RT, particularly those showing high thermal stability, are probably the most frequently used enzymes in applications requiring the synthesis of complementary DNA (cDNA) [17]. The crystal structure of the closely related xenotropic murine leukemia virus-related virus (XMRV) RT complexed with double-stranded DNA has been reported [24]. Interestingly, studies carried out with marmoset foamy virus (MFV) RT have revealed an unexpected ability of this enzyme to adopt monomeric or

dimeric configurations, depending on the type of bound substrate [25]. Thus, MFV can adopt both a monomeric configuration in the presence of RNA/DNA hybrids and an asymmetric dimer arrangement in the presence of dsDNA. Foamy virus RTs are enzymes of around 750 residues with all the enzymatic functions of the viral protease and RT. The enzyme has an N-terminal protease domain bound to the RT and is monomeric in solution [26,27].

Recently, interest has been focused on prokaryotic RTs, which are extraordinarily diverse and are key components of different lines of defense against phages and other mobile elements (for a recent review, see [28,29]). All these enzymes have enormous biotechnological potential. Among them, thermostable intron II RTs, such as TGIRT®-III RT, MarathonRT or Induro® RT, are commercialized and used in many transcriptomic applications. TGIRT®-III RT is a group II intron RT purified from *Geobacillus stearothermophilus* [30], while MarathonRT was obtained from *Eubacterium rectale* [31]. These enzymes are all thermostable and show better performance than engineered thermostable MLV RTs while copying transcripts longer than 10 kb. The crystal structure of a full-length thermostable group II intron RT (TGIRT®-III RT) in complex with an RNA/DNA hybrid and incoming dNTP has been obtained at 3.0-Å resolution [32]. The structure is similar to those of retroviral RTs, but unlike HIV-1 RT, the group II intron RT binds the template-primer substrate as a monomer.

3. Methods to Determine the Fidelity of DNA Synthesis Catalyzed by Purified RTs

Soon after the discovery of retroviral RTs in 1970, and taking advantage of research focused on the effects of potential carcinogens in promoting mutagenesis in DNA polymerization, researchers developed biochemical assays measuring the simultaneous incorporation of complementary and non-complementary nucleotides during DNA synthesis catalyzed by purified RTs. These classical assays included measurements of the incorporation of dCMP in place of dTMP in cell-free DNA synthesis with polyadenylic acid as a template or competition assays between complementary dGTP and noncomplementary dATP, using poly(rC)-oligo(dG) as a template-primer, and allowed estimates of fidelity for AMV RT [33,34]. Further sophistication of the assays led to the development of enzymological or "gel-based" assays that provided accurate measurements of the nucleotide selectivity of the polymerase at a given position, although their analysis was limited to a relatively small number of incorporation sites. Later on, researchers developed genetic assays using reporter genes that allowed the exploration of a wide variety of sequence contexts and helped with the identification of relevant hot spots related to RT errors. These methods are briefly summarized in the following sections.

3.1. Enzymological Assays

Enzymological assays include dNTP exclusion assays, misinsertion assays and mismatch extension assays. The dNTP exclusion assays, also known as "minus" sequencing gel assays or primer extension assays, provide a rough estimate of enzyme fidelity [35]. In these assays, only three dNTPs are supplied to extend an oligonucleotide primer annealed to a template. A faithful polymerase is expected to halt synthesis at the position where the missing dNTP would normally be incorporated. Therefore, the presence of full-extension bands indicates the insertion of incorrect nucleotides. The fidelity of RTs can be compared by examining the pattern of bands produced by each enzyme.

Quantitative characterization of nucleotide incorporation kinetics by RTs and other DNA polymerases has been used to analyze nucleotide selectivity in different sequence contexts and with many different wild-type and mutant polymerases (for reviews, see [8,36]). In misinser-

tion fidelity assays, the ability to extend a primer is evaluated in the presence of correct or incorrect nucleotides using a template-primer. The efficiency of primer elongation is measured through quantitative gel electrophoresis, and the resulting data are analyzed using the polymerase kinetic parameters k_{pol} and K_D for each nucleotide [37,38]. Misinsertion efficiency (f_{ins}) is calculated as the ratio of k_{pol}/K_D for the incorrect nucleotide to that for the correct nucleotide:

$$f_{ins} = k_{pol}/K_D \text{ (incorrect nucleotide)}/k_{pol}/K_D \text{ (correct nucleotide)}$$

On the other hand, mismatch extension assays evaluate DNA polymerase accuracy by comparing the kinetics of extending template-primer complexes with a mismatched 3′-end primer (e.g., C-C) and with a correctly matched template-primer hybrid (e.g., G-C). The mismatch extension ratio (f_{ext}) is calculated by dividing the catalytic efficiency (k_{pol}/K_D) of the mismatched primer by that of the correctly matched primer:

$$f_{ext} = k_{pol}/K_D \text{ (mismatched primer)}/k_{pol}/K_D \text{ (matched primer)}$$

Although k_{pol} and K_D are the preferred constants, their measurement requires rapid transient kinetics and sophisticated equipment for their correct assessment. Alternatively, equivalent parameters (i.e., k_{cat} and K_M) can be obtained under steady-state conditions, although in this case, the obtained values are strongly influenced by the slow rate of dissociation of the template-primer from the enzyme (k_{off}). In general, enzymological assays offer precise measurements of nucleotide selectivity at specific positions. However, they are labor-intensive, limited to a small number of incorporation sites and, for precise determinations, sophisticated equipment (quench-flow apparatus) might be needed. These methods provide good estimates of nucleotide selectivity, but the obtained results cannot be interpreted in terms of an overall error rate [8].

3.2. Genetic Assays

Genetic assays estimate fidelity by measuring mutant frequencies in a reporter gene (e.g., *lacZα*, *gfp*, *htk*) (for reviews see [8,9]). The assays can be designed to detect the reversion of an inactivating mutation introduced in the reporter gene (codon reversion assays) or to detect the inactivation of the reporter gene (forward mutation assays). Codon reversion assays measure base substitution rates of DNA polymerases using single-stranded phage DNA templates with nonsense codons, such as amber (TAG) in phage ΦX174 or opal (TGA) in the *lacZ* gene of bacteriophage M13mp2 (reviewed in [9]). Reversion of these codons to coding triplets is detected as revertant plaques after bacterial transformation or as dark blue M13 plaques on X-Gal indicator plates. These assays typically focus on specific base substitution errors, particularly G→A transitions. However, the limited scope of target sites may not accurately reflect the overall mutation rate.

Forward mutation assays assess the fidelity of RTs by detecting nucleotide changes across many sites within a reporter gene [39]. This method typically involves the use of a gapped dsDNA template from phage M13mp2, where the *lacZα* gene in one of the two strands is missing. The gapped DNA serves as template for DNA-dependent DNA synthesis to obtain the complementary strand of the *lacZα* gene (the antisense strand of the gene). The resulting DNA is then used to transform *Escherichia coli* cells, which are then plated on X-Gal indicator plates. Mutations can be identified by altered plaque color, such as light blue or colorless, indicating a loss of β-galactosidase activity because of errors introduced by the RT. If desired, DNA from mutant plaques can be purified and sequenced by Sanger sequencing to determine the specific mutations introduced. The method can also be adapted to study RNA dependent DNA polymerase fidelity (Figure 2).
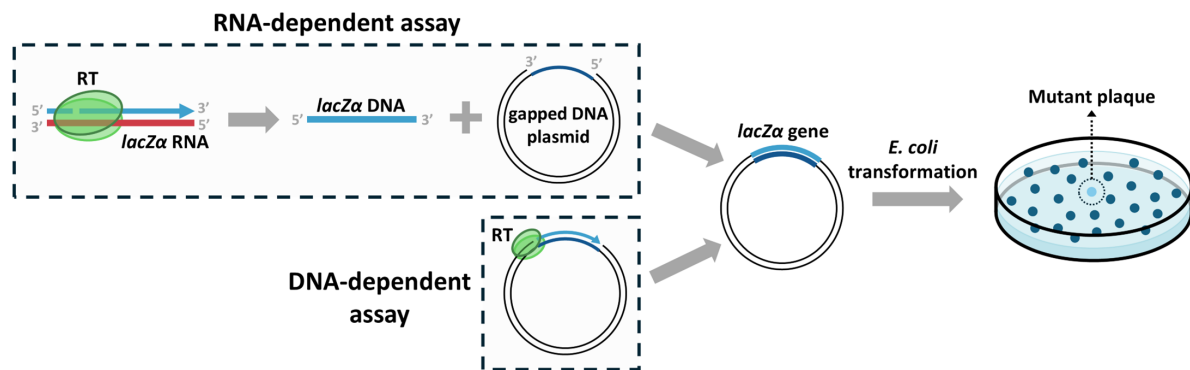
Figure 2. Overview of forward mutation assays with the *lacZα* gene. The initial steps (outlined with dashed lines) differ depending on the desired measurement. For RNA-dependent DNA polymerase fidelity assessment, the RT uses an RNA molecule with the *lacZα* sequence as a template, and the obtained DNA is hybridized with a gapped DNA plasmid carrying the complementary sequence of the *lacZα* gene. In the case of DNA-dependent DNA polymerase fidelity, the gapped DNA plasmid serves as template for the RT to produce the complementary *lacZα* gene strand. Following the initial steps, *E. coli* is transformed with the resulting plasmids, and errors are identified as light blue or colorless colonies.

Forward mutation assays offer several advantages, such as the ability to detect a wide range of mutations, including base substitutions and frameshifts, across many sites, making them valuable for identifying mutational hotspots. However, they have limitations (for reviews, see Refs. [8,9,36]). First, *Escherichia coli* mismatch repair machinery can potentially correct some errors. Secondly, even if the errors are not corrected, not all mutations result in detectable phenotypic changes, especially synonymous mutations that do not alter the protein sequence. A third limitation is that the observed mutation rates represent the sum of the RT mutation rate and errors already present in the template, especially relevant in the RNA-dependent assays. Additionally, the method can be labor-intensive and subjective, as plaque color changes may be subtle.

4. NGS-Based Methods for RT Fidelity Measurements

Over the past two decades, the advent of NGS technologies has revolutionized the ability to assess error rates in a relatively short time and using vast amounts of data [40,41]. However, short-read NGS platforms (e.g., Illumina), exhibit background error rates around 10^{-2} mutations per site and 10^{-3} to 10^{-4} mutations per site for stringently filtered data [42–45], too high for a reliable determination of DNA polymerase error rates. On the other hand, long-read NGS technologies (also referred to as third-generation sequencing technologies), such as PacBio or Oxford Nanopore, show even higher error rates, reaching up to 10^{-1} errors per base [46].

High error rates in sequencing create background noise that hinders the measurement of RT errors, which are typically below 10^{-4} in forward mutation assays [8,9]. Additionally, certain steps during library preparation (e.g., PCR amplification or gel electrophoresis) contribute to background errors. Consequently, accurate determinations of RT error rates require strategies to minimize and distinguish those artifacts.

The most widely used strategy for artifact correction is "consensus sequencing". This approach identifies and eliminates artifacts introduced by sequencing and during library preparation steps. It is valuable not only for determining RT error rates but also for other applications, such as detecting low-frequency mutations in RNA and DNA samples (reviewed in [47]). Suitable methods based on consensus sequencing were described for the first time around 15 years ago [48–50].

Consensus sequencing can be approached using different strategies, all sharing the common principle of obtaining multiple sequences that derive from an original DNA or RNA molecule (Figure 3). To analyze errors introduced by RTs, multiple copies of the same original cDNA molecule are typically examined. Mutations consistently observed at the same position across all molecules are considered true errors introduced during reverse transcription, whereas mutations appearing only in some molecules are regarded as sequencing errors or other artifacts.
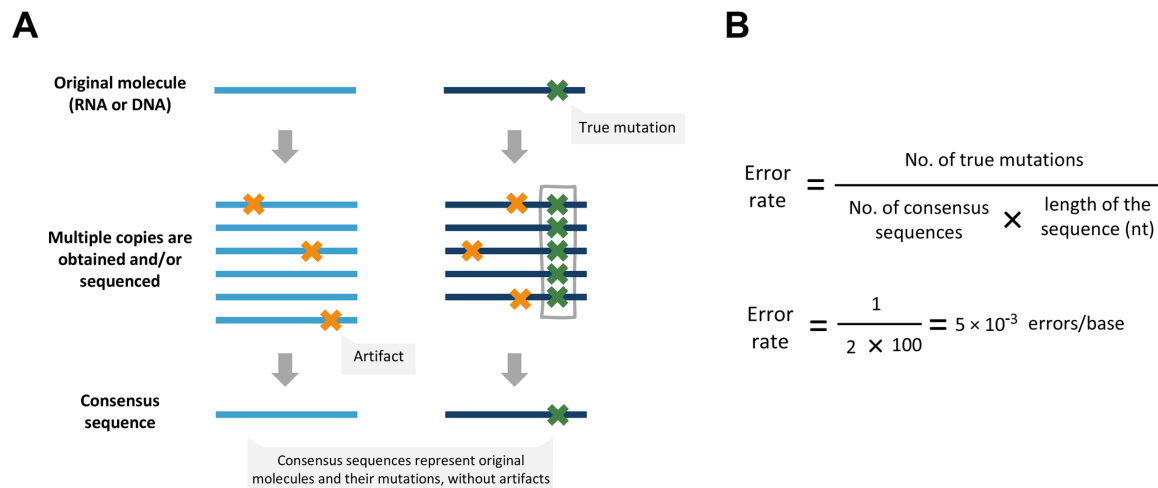

Figure 3. Rationale of consensus sequencing. (**A**) Error correction of consensus sequencing. Depending on the method, multiple copies of an original molecule (DNA or RNA) are obtained and sequenced, or the molecule is sequenced several times. Sequences that derive from the same molecule are aligned, and a consensus is established. True mutations from the original molecule, which are expected to be present in all copies, are added to the consensus. On the other hand, artifacts, which are expected to be inconsistent across the copies, are discarded. (**B**) Formula to calculate error rates. Once consensus sequences are obtained, an accurate error rate can be calculated. If we consider that the consensus sequences shown on the left have a length of 100 nt, the error rate of the original molecules would be 5×10^{-3} errors/base.

A number of consensus generation methods have been used (or could potentially be used) to measure RT errors. These methods can be classified depending on the use of barcodes.

4.1. Methods Based on the Use of Barcodes

A widely adopted approach for generating consensus sequences in short-read sequencing is the use of barcodes or UMIs, which are sequences of random degenerate nucleotides (typically > 8 nucleotides in length). These barcodes can be added to RNA molecules used as templates for reverse transcription, cDNA molecules obtained or both types of molecules.

Barcoding of cDNA molecules is a fast, effective and widely used strategy, initially described as the "PRIMER IDs" method by Jabara et al. [49]. This technique was first applied to identify minority variants found in viral genomic RNA from HIV-infected patients. Later, this approach was used by several labs to identify RT errors and measure RT error rates [31,51–56]. The procedure is summarized in Figure 4A. Reverse transcription is performed to generate cDNAs using primers that incorporate a unique barcode at their 5' end. This process renders cDNAs tagged with different barcodes and ensures that each cDNA molecule is uniquely identified. The barcoded cDNAs are then amplified via PCR, followed by sequencing. Reads sharing the same UMI are grouped together, allowing for the identification of mutations that are consistently observed within each group. These

mutations are considered to be original errors present in the cDNA molecule, as they are likely to have arisen during the transcription or reverse transcription processes.
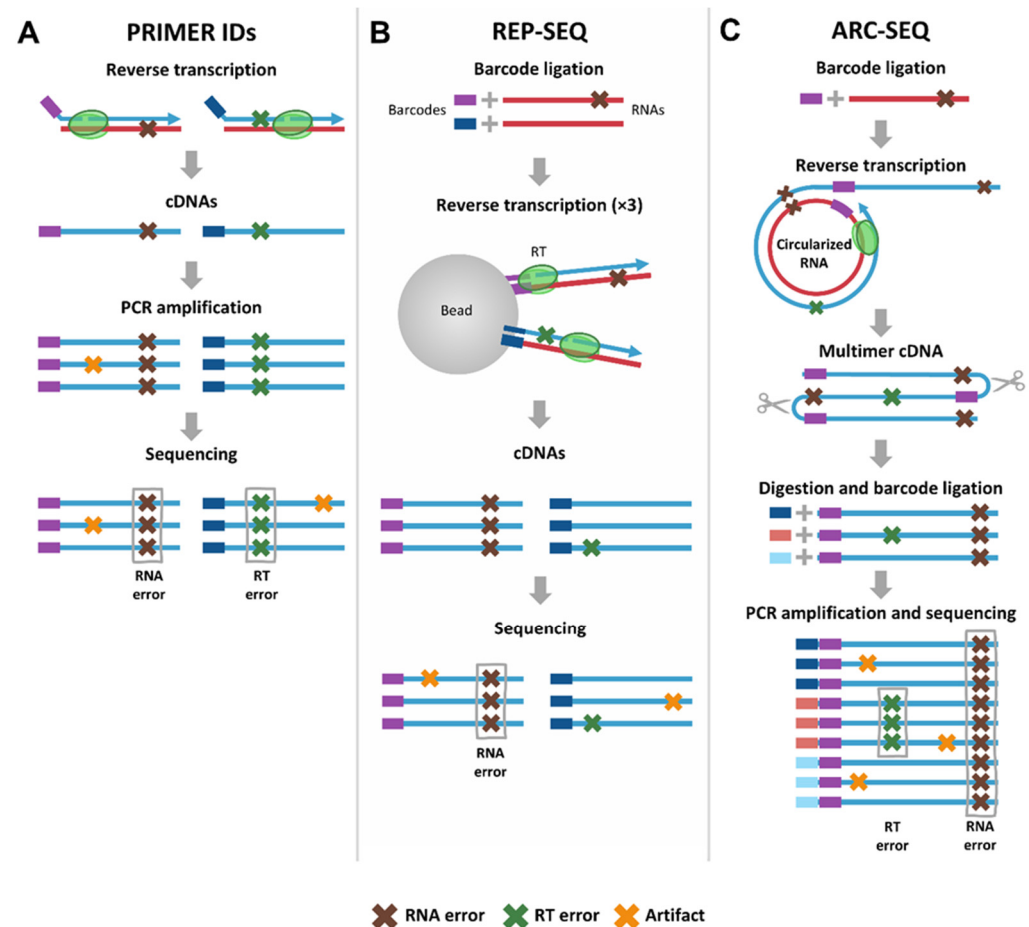

Figure 4. Barcode-based methods for reverse transcription error rate measurements. Brown crosses represent RNA errors, green crosses depict RT errors and yellow crosses represent artifacts (i.e., other errors, such as sequencing or PCR errors). (**A**) PRIMER IDS [49]. Reverse transcription is performed using primers with barcodes at their 5′ ends. The labeled cDNAs are amplified by PCR and sequenced. Combined RNA and reverse transcription errors are expected to be in all sequences sharing a barcode. (**B**) Replicated sequencing (REP-SEQ) [57]. RNA molecules are tagged with barcodes and attached to beads. Reverse transcription is performed, and cDNAs are washed away multiple times. The resulting cDNAs are then sequenced. Transcription errors are expected to be present in all reads sharing a barcode and are distinguished from RT errors and artifacts. (**C**) Accurate RNA consensus sequencing (ARC-SEQ) [58]. Barcoded RNA molecules are circularized and subjected to rolling circle reverse transcription, yielding tandem-repeated cDNA copies. These multimeric cDNAs are restricted to cDNA monomers, which are then tagged with barcodes, amplified through PCR and sequenced. Transcription errors are expected to be present in all sequences sharing an RNA barcode, while reverse transcription errors should be shared only among sequences with the same cDNA barcode.

Another approach similar to the PRIMER IDs method is known as duplex sequencing, or DUP-SEQ [59]. Unlike the PRIMER IDs method, which generates a consensus sequence from a single DNA strand, DUP-SEQ creates consensus sequences for each strand of a dsDNA molecule. In this method, following reverse transcription, the complementary strand of the cDNA is synthesized, and special adapters containing barcodes are added to both ends of the dsDNA. These adapters individually label each strand and enable their independent amplification due to the presence of non-complementary segments in their structure. Finally, the amplified copies are sequenced and analyzed in silico. This

process allows for the generation of separate consensus sequences for both the original cDNA strand and its complementary strand. True mutations are expected to be found in the same position in both sequences, while mutations present in only one strand would be considered artifacts not discarded during the single-strand consensus construction. While duplex sequencing methods have been widely used to study ultra-rare mutations in DNA samples [59–61], they have not yet been applied to measure RT fidelity. Nevertheless, this approach could prove valuable for studying high-fidelity RTs or identifying errors during second-strand DNA synthesis.

Both the PRIMER IDs and DUP-SEQ methods provide a robust framework for distinguishing true mutations from errors introduced during sequencing or PCR amplification. However, these mutations reflect not only errors introduced by the RT used but also mutations present in the RNA template. Consequently, in these setups, it is not possible to distinguish between RNA-derived mutations and reverse transcription errors.

The application of barcodes to RNA molecules allows the identification of RNA mutations. For example, Gout et al. [57] identified RNA mutations with a method called REP-SEQ (Figure 4B). In REP-SEQ procedures, RNA molecules are tagged with barcodes and bound to beads. Reverse transcription is carried out, followed by repeated washing of the cDNAs. The resulting cDNAs are then sequenced. Transcription errors are expected to appear in all reads sharing the same barcode. RT errors can then be estimated as those mutations not attributable to the RNA, but this estimation may also include artifacts introduced during library preparation.

The best way to discard both RNA mutations and sequencing artifacts would be the labeling of both RNA and cDNA, as in the ARC-SEQ method [58] (Figure 4C). This dual-barcoding approach enables researchers to distinguish between RNA-derived mutations and errors introduced during reverse transcription and, at the same time, allows the identification and elimination of sequencing artifacts. RNA molecules are first tagged with barcodes, circularized and subjected to rolling circle reverse transcription, a process that generates tandem repeats of cDNA copies. These multimeric cDNAs are then processed into individual cDNA monomers, each of which is then tagged with a unique barcode. The barcoded cDNAs are amplified through PCR and subsequently sequenced. In this method, transcription errors are expected to be present across all sequences that share the same RNA barcode, as these errors occur during the transcription of the RNA itself. On the other hand, reverse transcription errors should only appear in sequences that share the same cDNA barcode. This method has been exclusively applied to analyze RNA errors in published studies, but it could be adapted to focus on RT errors as well.

4.2. Methods Avoiding the Use of Barcodes

While the use of RNA barcodes in conjunction with cDNA barcodes could effectively identify reverse transcription errors, adding barcodes to RNA presents significant challenges. Indeed, Traverse and Ochman [62] reported difficulties in reproducing results obtained with the REP-SEQ method. In view of these problems, most studies assessing the fidelity of RTs have either employed the "PRIMER IDs" approach or resorted to barcode-less alternatives.

A straightforward example of a barcode-less method is CIR-SEQ [63] (Figure 5A), which utilizes rolling circle amplification. This process is conceptually similar to the rolling circle reverse transcription used in the ARC-SEQ method. In CIR-SEQ, DNA templates (which may be generated after reverse transcription of RNA molecules) are circularized. These circularized templates are then amplified multiple times in tandem using a rolling circle polymerase, producing long, repetitive sequences. These tandem-repeat molecules are subsequently sequenced, and the resulting sequences are analyzed to derive a consensus

sequence. To date, this method has primarily been used to measure mutations in DNA, but it could also be adapted for the determination of RT error rates.
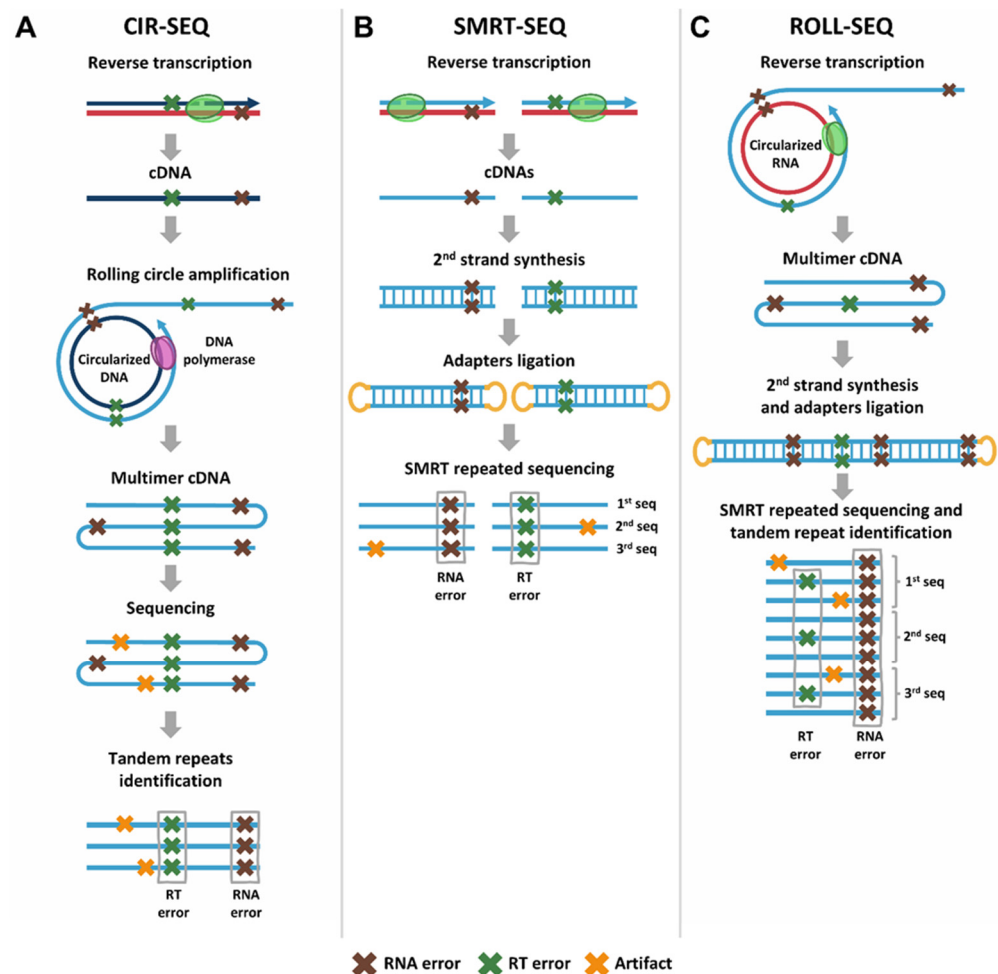

Figure 5. Barcode-free methods for reverse transcription error rate measurements. Brown crosses represent RNA errors, green crosses depict RT errors and yellow crosses represent artifacts (i.e., other errors, such as sequencing or PCR errors). (**A**) Circular sequencing (CIR-SEQ) [63]. Circularized DNA fragments (for example, cDNAs obtained after reverse transcription) are obtained and used as a template for rolling circle amplification reactions to obtain multimer DNAs. After sequencing, the repeated units of each multimer cDNA are identified and aligned. Transcription errors are expected to be found in all the aligned sequences from the same multimer cDNA molecule. (**B**) Single molecule real-time sequencing (SMRT-SEQ) [64]. RNA is used as template for reverse transcription. The obtained cDNAs are used as templates for second-strand DNA synthesis. PacBio SMRT adapters are attached to the dsDNA and subjected to SMRT repeated sequencing. Each read of the original cDNA strand is aligned. Combined RNA transcription and reverse transcription errors are expected to be in all the aligned sequences. Additionally, second-strand DNA synthesis errors could be identified by aligning the reads of the complementary strand. (**C**) Rolling circle sequencing (ROLL-SEQ) [65]. Rolling circle reverse transcription reactions are conducted to obtain multimeric cDNAs. The complementary DNA strand is then synthesized, and PacBio SMRT adapters are attached. SMRT repeated sequencing is performed, and the repeated monomers of the cDNA strand are identified for each read. RNA errors are expected to be in all the monomers of the same read and in all reads, while RT errors are expected to be only in the aligned reads of the same monomer.

However, as found in the case of the "PRIMER IDs" method, error rates obtained with CIR-SEQ would reflect the cumulative errors introduced during the reverse transcription process. The CIR-SEQ method has also been adapted to study RNA mutations [66–69]. In those adaptations (also named CIRC-SEQ), tandem-repeat cDNAs are directly obtained

from RNA molecules via rolling circle reverse transcriptions. These multimeric cDNAs are sequenced and are then analyzed the same way as described above, but mutations identified in consensus sequences will only represent RNA errors.

Another barcode-free consensus generation method is SMRT-SEQ [70], a long-read sequencing technology that allows repeated sequencing of the same DNA molecule. This method was used by Potapov et al. [64] to measure the error rates of several RTs (Figure 5B). In these analyses, RNA serves as a template for reverse transcription, generating cDNA molecules. The resulting cDNAs are then used as templates for second-strand DNA synthesis, producing double-stranded DNA. PacBio SMRT adapters are attached to the dsDNA, which is then subjected to repeated sequencing using SMRT technology. Each read corresponds to the original cDNA strand, and these reads are aligned to generate a consensus sequence. Like cDNA barcoding methods, SMRT-seq captures both reverse transcription errors and RNA-derived errors, making it challenging to distinguish between the two types of errors.

A recently published method, ROLL-SEQ [65] (Figure 5C), combines elements of CIR-SEQ and SMRT-SEQ: it incorporates rolling circle reverse transcription while allowing differentiation of RNA errors. This method provides a more accurate measurement of RT errors by distinguishing them from the RNA-specific errors. In this approach, rolling circle reverse transcription reactions are performed to generate multimeric cDNA molecules. Once these multimeric cDNAs are synthesized, the complementary DNA strand is created, and PacBio SMRT adapters are attached to the dsDNA. The samples undergo SMRT repeated sequencing, which captures long reads and allows for the identification of repeated monomers of the cDNA strand in each sequencing read. RNA-derived errors are expected to appear in all monomers of the same read, as well as in all reads that share the same RNA template. In contrast, RT errors are expected to appear only in the aligned reads of the same cDNA monomer. Both methods based on SMRT sequencing (i.e., SMRT-SEQ and ROLL-SEQ) offer the potential to study second-strand DNA synthesis errors if desired. As in the case of duplex sequencing, it is possible to generate independent consensus sequences for each strand: the one corresponding to the cDNA and its complementary strand.

In addition, we can also find barcode-less error correction strategies applied to Oxford Nanopore sequencing. For example, INC-SEQ [71] combines rolling circle amplification, followed by nanopore sequencing. Unlike the ROLL-SEQ method, it only allows the sequencing of a single strand. This method has been used to detect low-abundance mutations, but it could also be used to measure errors in cDNA molecules.

5. Limitations of NGS-Based Fidelity Measurements

Measuring and comparing the error rates of RTs is often challenging. Even when the same type of experiment and the same RT enzyme are used, numerous factors can influence fidelity. Since RTs can synthesize DNA using two types of templates, it is important to distinguish between fidelity when using RNA or DNA as a template. Other factors that can affect the measured fidelity include the sequence of the template, the source of the template (e.g., chemically synthesized, in vitro synthesized by polymerases, extracted from biological samples), the presence of nucleic acid modifications, mutations in the RT that may affect fidelity and conditions during reverse transcription (e.g., temperature, magnesium concentration, pH, nucleotide concentration) [55,56,64,72], which are reviewed in Refs. [9,73].

5.1. NGS Challenges

In contrast to other methods, NGS assays enable the rapid and efficient determination of error rates across a large number of sequences and can identify all types of errors, includ-

ing silent mutations. However, there is a great diversity of techniques and bioinformatics pipelines to analyze the results. The parameters employed in bioinformatics analyses vary between studies and can significantly affect error rate calculations. As a result, a consistent standard for measuring RTs' fidelity with NGS platforms is still lacking. This also includes the bioinformatics pipelines used to analyze the NGS results and their parameters. Open and clearly documented pipelines are required to allow for the easy reproduction of results while ensuring transparency in the analysis of NGS data.

5.2. Barcode Challenges: Consensus Sequences and Cutoff and Threshold Values

A common feature of all NGS methods is the need to generate consensus sequences in order to identify true mutations and discard artifacts. Two key parameters are critical: the minimum number of sequences required to build a consensus sequence (cutoff) and the threshold used to identify mutations, i.e., the minimum percentage of sequences in which a mutation must be present (at a given position) to be considered a true mutation. For example, a 100% threshold requires the mutation to appear in all sequences to be included in the consensus, whereas a 50% threshold allows the mutation to be included if it appears in at least half of the sequences.

The choice of cutoff and threshold values can significantly affect error measurements. However, there is no consensus on the optimal parameters for accurate mutation selection, and a wide variety of parameter choices exist (discussed below). Reid-Bayliss and Loeb [58] reported that, for a fixed 70% threshold, error rates reach a plateau when using a cutoff of three. On the other hand, in the study by Martínez del Río et al. [56], errors were analyzed with different cutoffs and thresholds, as well. For instance, with a fixed cutoff of at least four sequences per consensus, the authors found a difference of one order of magnitude in the HIV-1 RT error rate when using a 0% threshold as compared to 100% threshold (1.6×10^{-4} and 1.6×10^{-5} errors/base, respectively). The study also included computational simulations to assess threshold selection, revealing a trade-off between sensitivity (also known as recall or true positive rate) in detecting true mutations and specificity (true negative rate) in discarding artifacts. Strict 100% thresholds provide high specificity at the cost of reduced sensitivity, while less stringent thresholds (0–50%) achieve 100% sensitivity but include more artifacts. Figure 6 illustrates two scenarios where sensitivity can be compromised if a high threshold is applied.

5.3. Strategies to Avoid Sequencing Artifacts

As mentioned before, another issue is the inclusion of artifacts as true errors (low specificity), especially when a low threshold is used to reach a high sensitivity. The repeated occurrence of an artifact in the same position of sequences could be explained by several factors. For example, if the method includes PCR amplification, errors introduced in the first steps are expected to be present in a high number of DNA molecules that are sequenced. However, sequencing pipelines involving the elimination of sequences neighboring the target sequence have been designed to reduce errors introduced in the first amplification steps, reporting a 100-fold reduction over standard Illumina error rates [74]. On the other hand, the occurrence of sequencing error hotspots could be a source of repeated artifacts. Deakin et al. [75] performed Illumina sequencing with Sanger-sequenced barcodes and reported an uneven distribution of NGS errors along the sequences, with pronounced sequencing error hotspots.
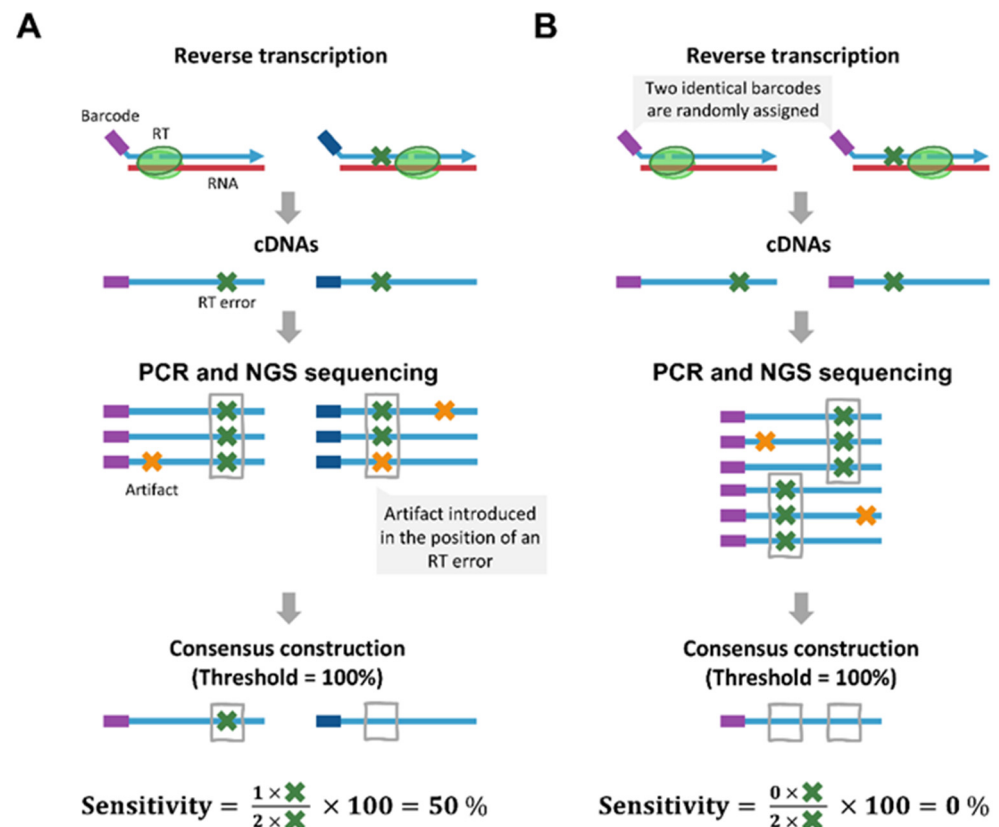
Figure 6. Sensitivity in cDNA barcoding methods. (**A**) shows an example where sensitivity is affected by the introduction of an artifact during PCR or sequencing at a position already containing an RT mutation. (**B**) illustrates a scenario where sensitivity is impacted by assigning the same barcode to sequences originating from different cDNAs. This can occur if barcodes with identical sequences are used by chance during reverse transcription. Alternatively, different barcodes may initially be assigned, but a mutation introduced during library preparation can generate a new barcode sequence that matches another barcode already present in the library pool. Sensitivity is calculated as the percentage of original mutations found in the consensus sequences relative to all mutations present in the original cDNAs that underwent PCR amplification and sequencing.

The use of accurate sequencers, stringent quality filtering and strategies like paired-end sequencing can undoubtedly reduce artifacts due to sequencing errors [76] but it does not offer a perfect solution to discard artifacts, especially considering there may also be other sources of artifacts. Including controls in experiments can help estimate background errors that may be quantified despite the consensus-building strategy. However, the use of controls is often limited. For example, Wang et al. [55], who used the cDNA barcoding method to measure HIV RT error rates, aimed to estimate background error rates by using a DNA control obtained by an asymmetric PCR (in substitution of cDNA obtained by reverse transcription). The highest error rate obtained in one control (9.2×10^{-5}) was not far from the lowest error rate reported (3.4×10^{-4}), which may indicate a high background error in their method. Furthermore, they found a large number of C→A and G→T substitutions in their controls that were not considered in their RT error rate determinations.

In the study by Martinez del Río et al. [56], the authors also included a DNA control in their experiments of cDNA barcoding fidelity measurements. The DNA control was produced through a single-cycle PCR using the high-fidelity Q5 DNA polymerase and a plasmid-derived DNA as a template. This DNA control underwent the same preparation and sequencing processes as the cDNAs. When using a threshold of 100%, their control exhibited a very low error rate of 3.67×10^{-7}, close to a previous estimate of the Q5

DNA polymerase error rate [77]. However, when using a 0% threshold, the error rate was significantly higher (1.36×10^{-5}), highlighting the significant impact of threshold values on specificity.

Additionally, methods like DUP-SEQ, compared to PRIMER IDs, offer improved specificity due to the creation of two independent consensus sequences: one from the cDNA strand and one from its complementary strand. This dual-consensus approach enhances specificity by identifying and discarding errors that appear in only one strand, as they are treated as artifacts. Schmitt et al. [59] reported that DUP-SEQ achieves a theoretical background error rate as low as 10^{-9}. However, this increased specificity comes at the cost of greater methodological complexity. Moreover, the need to amplify and sequence both strands for each cDNA molecule reduces the overall yield of consensus sequences obtained compared to single-strand consensus sequencing methods. This lower efficiency can compromise the sensitivity of the approach, although bioinformatics analyses have been proposed to enhance sensitivity while maintaining the specificity characteristic of duplex sequencing methods [78].

5.4. Artifacts Affecting the Mutational Spectra and Hotspot Distribution

Another challenge is the non-uniform distribution of reverse transcription errors along the sequence being studied. For example, both *lacZ* forward mutation assays [79,80] and NGS-based assays [56] have reported a remarkable accumulation of errors at the first positions that are reverse transcribed by the viral polymerase. The inclusion of these positions in the error rate calculations could have a large effect on the obtained values, especially when the consensus sequences are relatively short. This elevated error rate at the initial positions has also been observed with other DNA polymerases [81]. Therefore, if the NGS method includes a PCR (or a rolling circle) amplification step, there could be an accumulation of errors that could contribute to a higher background error rate. Furthermore, analyses of sequences obtained with the Illumina platform have identified a high frequency of mutations at both ends of the sequenced insert. This pattern is characterized by A→C substitutions at the 5′ end and T→G substitutions at the 3′ end [42,82].

Another issue associated with most methods is the inability to exclude errors already present in molecules used as a template for reverse transcription. Since RNA cannot be synthesized with high-fidelity polymerases, this problem becomes more significant when RNA is used as a template instead of DNA. Consequently, when using RNA as a template, the error rates measured after the consensus construction represent the cumulative error of reverse transcription and RNA transcription errors. This phenomenon, designated as "transcriptional threshold", may attenuate RNA-dependent DNA synthesis fidelity when using high-fidelity RTs [80]. Methods that involve labeling RNA molecules can address this problem, but they come with increased technical difficulty and higher costs. On the other hand, methods like ROLL-SEQ avoid the use of RNA barcodes and allow the discrimination of RNA errors from RT errors. However, RTs with elevated rolling-circle activity are needed to generate long concatemeric cDNA. This might be possible with thermostable intron II RTs that appear to be particularly efficient in copying large transcripts [30–32].

5.5. Artifacts Due to the Generation of "Offspring Barcodes" and Barcode Duplications

The generation of "offspring barcodes" or "false families" (i.e., the appearance of artifacts in sequenced barcodes) can be another difficulty associated with the use of barcode-based methods, such as PRIMER IDs, REP-SEQ or ARC-SEQ (Figure 7). This can lead to situations where distinct consensus sequences actually originate from the same cDNA molecule, potentially quantifying the same error twice, or it can lead to a loss of sensitivity if the offspring barcode shares its sequence with another barcode (Figure 6B). Zhou et al. [83]

addressed this issue, and based on calculations of maximum frequencies of offspring families, they provided cutoff values that should be used to minimize the counting of offsprings. This value is directly correlated to the number of different barcodes observed after sequencing. However, these calculations were obtained assuming that sequencing and other artifacts have a cumulative error rate of 2% in the sequences. In addition, the election of high cutoffs to avoid offsprings may diminish the robustness of error rate calculations, since the majority of barcodes are often repeated only a couple of times. In fact, it is difficult to obtain barcodes that are repeated. The majority of barcodes in experiments are repeated only once [53,56], which does not allow the construction of consensus sequences. Figure 7C shows the typical distribution of barcode frequencies.

A problem affecting barcode-based methods is the occurrence of repeated UMIs (collisions or tag clashes). This can be attributed to random assignments or because an offspring sequence generates a barcode that already exists in the barcode pool. Collisions may result in underestimation or overestimation of mutation frequency, although, on average, underestimations are more probable according to informatics simulations [84]. To reduce the number of collisions, it is ideal to use a small number of input molecules and a much larger number of barcode combinations. This number is directly proportional to the barcode length. Nevertheless, detecting low error rates requires a high number of input molecules, which can sometimes exceed the available barcode combinations, thereby increasing the likelihood of collisions. In the study by Jabara et al. [49], the authors used barcodes that were only eight nucleotides long, thereby increasing the likelihood of repetition, as pointed out by Sheward et al. [85], after applying formulae related to the well-known birthday paradox and demonstrating the high probability of having repeated barcodes. However, longer barcodes also increase the likelihood of obtaining offspring barcodes and potentially affect error rate measurements [86]. It is also worth noting that even in cases where the probability of having at least one barcode is high, the expected number of collisions may be very small to be able to have an impact in error rate calculations [87]. In Table 1, we show the expected number of collisions for different numbers of barcodes used and for different barcode lengths.
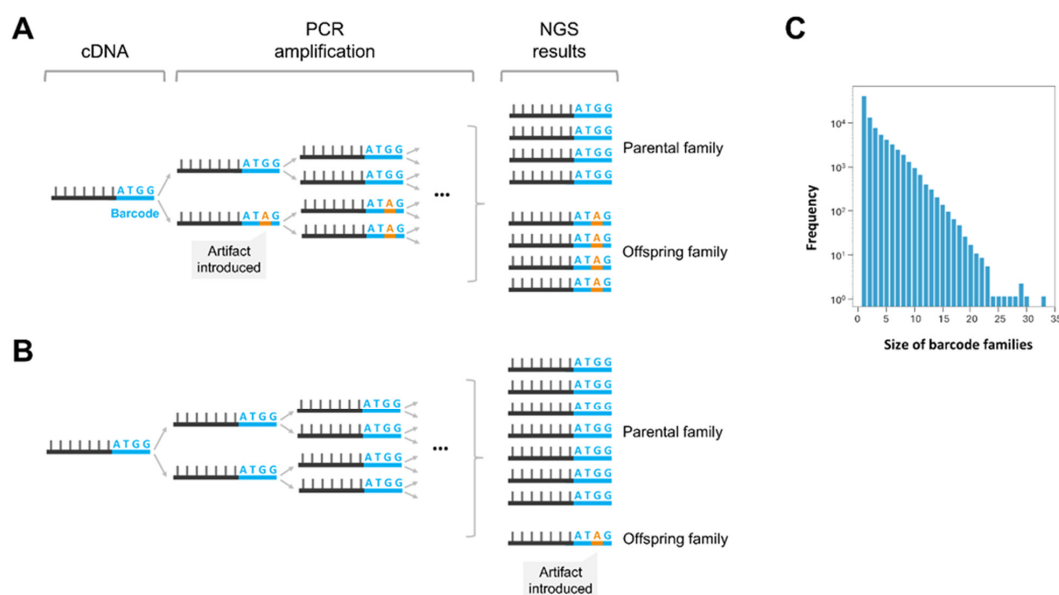

Figure 7. Effect of mutations in barcodes on the generation of offspring barcodes. (**A**) Impact of introducing an artifact in a barcode during the early PCR cycles. (**B**) Impact of introducing an artifact in the later PCR cycles or during sequencing. (**C**) Barcode frequency distribution in a typical NGS experiment. Ordinate values are represented on a logarithmic scale. Adapted from [56].

Table 1. Expected collisions for degenerated barcodes of 8-, 10-, 12-, 14- and 16-nt. Calculations were determined using the "birthday problem" formula indicated below, assuming different numbers of barcodes [1].

Number of Barcodes (*n*)	Expected Number of Collisions				
	8-nt UMI	10-nt UMI	12-nt UMI	14-nt UMI	16-nt UMI
1000	15.13	0.95	<0.01	<0.01	<0.01
2000	60.08	3.81	0.24	0.02	<0.01
5000	367.21	23.78	1.49	0.09	<0.01
10,000	1415.05	94.91	5.96	0.37	0.02
20,000	5259.93	377.84	23.83	1.49	0.09
50,000	26,685.09	2328.19	148.79	9.31	0.58
100,000	78,256.93	9096.03	594.27	37.25	2.33
500,000	499,757.03	189,627.97	14,681.28	930.45	58.20
1,000,000	999,999.76	614,677.19	57,863.01	3718.36	232.80
2,000,000	2,000,000.00	1,703,052.94	224,755.88	14,845.78	931.11
10,000,000	10,000,000.00	9,999,278.49	4,490,142.91	365,675.47	23,255.98

[1] Calculations shown in the table were carried out using the classical "birthday problem" formulation. Instead of individuals, we used barcodes (*n*), and instead of the 365 possible birthdays, we considered the different combinations for a barcode of a certain length (c). The expected number of barcode "collisions", i.e., the number of barcodes sharing the same sequence in a given experiment (or data collection), is given by the following equation: E (number of barcodes with shared sequences) = $n\,(1 - (1 - (1/c))^{n-1})$.

6. The Intrinsic Fidelity to Retroviral RTs Calculated by NGS Methods

Forward mutation assays using the M13mp2 *lacZα* reporter gene [39] have been widely used to determine the DNA-dependent DNA synthesis fidelity of many different RTs. Calculated mutation rates range from 5.8×10^{-6} for feline leukemia virus RT to 1.7×10^{-4} for prototype primate foamy virus RT [55,80,88–97]. As discussed above, these are underestimates of the error rate since *lacZα*-based assays are unable to detect silent mutations, as well as those that do not produce a detectable phenotypic change. In addition, variability in the reported error rates has been observed for the widely studied HIV-1 RT, with values in the range of 6.3×10^{-5} to 1.9×10^{-4} (reviewed in [9]). The high error rates of HIV-1 RT (around 10 times higher than those of AMV or MLV RTs) have been attributed to the relatively high Mg^{2+} concentration used in the assays. Achuthan et al. [98] showed that the accuracy of purified HIV-1 RT could be increased up to 5 to 7 times in the presence of free Mg^{2+} concentrations similar to those found in lymphocytes (~0.25 mM).

The comparison of error rates obtained with HIV-1 RT in reporter gene-based assays failed to provide consistent results when comparing RNA vs. DNA templates. These assays are technically more difficult to perform if RNA is used as template. However, error rates of 1.9×10^{-4} for DNA-dependent DNA synthesis and 2.0×10^{-4} for RNA-dependent DNA synthesis were obtained from studies carried out with HIV-1 RT using M13 and *env* V1 as the vector and reporter gene, respectively [99]. The differences were also small for similar assays carried out with the pBluescript plasmid and *lacZ* (1.7×10^{-4} and 1.4×10^{-4}, for DNA and RNA templates, respectively) [100]. However, other studies showed that RNA-dependent DNA synthesis was around three- to five-fold more accurate than DNA-dependent DNA synthesis in reactions catalyzed by HIV-1 RT, in the presence of high concentrations of Mg^{2+} [79,80]. Using forward mutation assays and *lacZα* as a reporter gene, Sebastián Martín et al. obtained error rates of 2.5×10^{-5} to 3.5×10^{-5} for RNA-dependent DNA synthesis with different retroviral RTs, including those of HIV-1, MLV and AMV [80].

As discussed earlier, these values are underestimates and include the transcriptional error rate of phage T7 polymerase used in the synthesis of the template RNA. However, these cumulative error rates (resulting from errors in the template RNA, plus errors in the cDNA synthesized by the RT, and any potential changes occurring in the bacteria) were broadly consistent with error rates calculated ex vivo using *lacZ* as reporter gene and

estimated at around 3.4×10^{-5} per nucleotide and replication cycle in HeLa and CEM cells [101,102] and 1.4×10^{-5} in HOS cells [103].

One of the most popular applications of NGS in retrovirology focuses on the detection of minority variants in the viral quasispecies, particularly for the early identification of drug-resistant HIV variants that could hamper the efficacy of antiretroviral therapies (for recent reviews, see Refs. [104,105]). NGS platforms are error-prone, and the use of barcodes or UMIs (originally called Primer IDs) during cDNA synthesis opens the possibility of detecting minority variants present at less than 20% abundance [49,83], as reviewed in Ref. [106]. In these assays, total HIV RNA was obtained from infected patients, and the frequency of mutations was determined with NGS and a cDNA barcoding method. These assays allowed the calculation of error rates for reverse transcription catalyzed by SuperScript III RT while copying two different regions of the HIV-1 *env* gene (corresponding to V1/V2 and C2/V3 regions) and the protease-coding region (Table 2). The calculated overall error rate was around 1.1×10^{-4}, although the base substitution error rates showed striking variability and ranged from 3.0×10^{-5} for env V1/V2 to 1.2×10^{-4} for the protease-coding region [83].

SuperScript RTs are engineered MLV RTs that show high thermostability and are frequently used in RT-PCR and other applications [17]. As shown in Table 1, error rates obtained with engineered MLV RTs (e.g., SuperScript II, III and IV; MM4; and ProtoScript II) were in the range of 5.0×10^{-5} to 1.8×10^{-4} and were not largely affected by the origin of the RNA used in the assays (i.e., natural RNA vs. RNA synthesized with phage RNA polymerase). The differences between wild-type and mutant MLV RTs were relatively small in experiments where enzymes were assayed under the same conditions [54,64]. Potapov et al. [64] also showed that the error rates of AMV and MLV RTs were very small (7.5×10^{-5} and 6.3×10^{-5}, respectively), and both enzymes showed a preference for T→C as the most frequent substitution introduced in the reverse transcription process (Table 2). The substitution T→C was also the most frequent mutation introduced by group-II intron-encoded RTs (e.g., MarathonRT, TGIRT and Induro RT) [31,64]. These enzymes showed cumulative error rates slightly reduced in comparison with SuperScript IV [31].

In contrast, the calculated HIV-1 RT error rates were higher, in agreement with previous reports based on the results provided by forward mutation assays and the M13 *lacZα* reporter gene. Side-by-side comparisons of HIV-1 RT and a thermostable variant of MLV RT showed a 2.6-fold increase in the cumulative error rate of the HIV-1 enzyme, measured on the *cesD* gene of *Bacillus cereus* [52]. The calculated HIV-1 RT error rate (around 3×10^{-4}) was obtained from cDNA synthesis reactions carried out in the presence of 5 mM Mg^{2+} and was reduced to 1×10^{-4} in 1 mM Mg^{2+} [53]. Similar effects, caused by the reduction in the Mg^{2+} concentration, were also reported by Wang et al. and Martínez del Río et al. using the M13 *lacZα* and the HIV-1 protease-coding sequences, respectively, as targets [55,56]. The lower error rates reported by Martínez del Río et al. (Table 2) can be attributed to the consensus and threshold values used in their experiments. These conditions were introduced to improve the specificity of the assay and eliminate potential artifacts not discarded during consensus construction. On the other hand, the error background reported by Wang et al. [55] was very high, as determined in their negative controls. Error rate estimates of HIV-1 RT have been generally obtained using in vitro transcribed RNA [52,53,55,56]. Errors made by phage polymerases during transcription can be a source of errors. However, Martínez del Río et al. [56] have shown that the RNA source has little impact on the overall error rate, although the most prevalent mutations differed between both experiments. A detailed analysis of the locations and mutation types obtained in both assays revealed a likely contribution of adenosine deaminases (e.g., ADAR1) in the observed variability, as suggested in other studies [103,107].

Table 2. RT error rates calculated by NGS for RNA-dependent DNA synthesis, as determined in different studies [1].

Study (Ref.)	RNA Source (Target Sequence)	RTs	RT Error Rates	Most Common SNP	Method	Consensus Cutoff	Threshold (%)
Jabara et al. [49]	Human plasma samples HIV-infected subject (HIV-1 PR-coding sequence)	**SuperScript III**	7.05×10^{-5}	-	cDNA barcoding	3	0
Zhou et al. [83]	HIV-1 RNA from 8E5/LAV infected cells (HIV-1 *env* C2/V3 and V1/V2, HIV-1 PR)	**SuperScript III**	1.1×10^{-4}	A→G	cDNA barcoding	Variable (to avoid offsprings)	0
Gout et al. [57]	*Caenorhabditis elegans*	**SuperScript II**	1.14×10^{-4}	G→A	REP-SEQ	2–3	100
Ellefson et al. [51]	Human heart total RNA (*HSPCB* gene)	**MMLV** RTx RTx (exo-)	1.1×10^{-4} 3.7×10^{-5} 1.0×10^{-4}	G→A G→T A→G	cDNA barcoding	3	66.6
Potapov et al. [64]	In vitro transcription using T7 RNA polymerase (artificial sequence)	**MMLV** **ProtoScript II** **AMV** Bst	6.3×10^{-5} 5.6×10^{-5} 7.5×10^{-5} 1.8×10^{-4}	T→C T→C T→C T→C	SMRT-SEQ	-	-
Zhao et al. [31]	RepA D3 RNA (a 1.6-kb mouse lncRNA)	**SuperScript IV** MarathonRT TGIRT	1.8×10^{-4} 9.9×10^{-5} 1.3×10^{-4}	- T→C T→C	cDNA barcoding	3	100
Potapov et al. [65]	In vitro transcription using T7 RNA polymerase (artificial sequence)	Induro RT L1-RT Fbu	1.1×10^{-4} 2.8×10^{-4} 1.8×10^{-4}	A→G T→C T→C	ROLL-SEQ	3	100
Houlihan et al. [54]	Human total RNA (GAPDH mRNA)	**MMLV** **SuperScript III** **ProtoScript II** RTx RT521K RT-TR RT-H11	9.77×10^{-5} 5.01×10^{-5} 7.66×10^{-5} 9.60×10^{-4} 2.51×10^{-3} 1.22×10^{-5} 1.08×10^{-5}	C→T C→T T→A A→G C→A G→A G→A	cDNA barcoding	3	66.6
Yasukawa et al. [52]	In vitro transcription using T7 RNA poly-merase (*cesD*, *B. cereus*)	**HIV-1** **MM4** RTx	2.6×10^{-4} 1.0×10^{-4} 7.5×10^{-5}	A→G C→A C→A	cDNA barcoding	5	100

Table 2. *Cont.*

Study (Ref.)	RNA Source (Target Sequence)	RTs	RT Error Rates	Most Common SNP	Method	Consensus Cutoff	Threshold (%)
Okano et al. [53]	In vitro transcription using T7 RNA poly-merase (*cesD, B. cereus*)	**HIV-1**	1×10^{-4} (1 mM Mg^{2+}) 3×10^{-4} (5 mM Mg^{2+})	-	cDNA barcoding	5	100
Wang et al. [55]	In vitro transcription using T3 RNA polymerase (*lacZα*)	**HIV-1**	3.05×10^{-4} (0.5 mM Mg^{2+}) 1.12×10^{-3} (6 mM Mg^{2+})	C→T and G→A	cDNA barcoding	5	100
Martínez del Río et al. [56]	In vitro transcription using T7 RNA polymerase (HIV-1 PR-coding sequence)	**HIV-1**	9.0×10^{-6} - (0.5 mM Mg^{2+}) 7.3×10^{-5} (3 mM Mg^{2+})	A→G G→A	cDNA barcoding	4	75
	HIV RNA from 8E5/LAV infected cell line (HIV-1 PR-coding sequence)		1.8×10^{-5} (0.5 mM Mg^{2+}) 7.4×10^{-5} (3 mM Mg^{2+})	C→T C→T			

[1] RT sources are AMV, avian myeloblastosis virus; Bst, Bst 2.0 and Bst 3.0 DNA polymerases; MMLV, Moloney murine leukemia virus; MM4 is a thermostable MLV RT containing amino acid substitutions E286R/E302K/L435R/D524A; RT521K is a replicative DNA polymerase from *Thermococcus gorgonarius* that possesses RT activity; RT-TR and RT-H11 are high-fidelity variants of RT521K that contain amino acid substitutions P410T/S411R and R406P/L408A/P410T/S411R/I412F, respectively; RTx is a DNA polymerase from *Thermococcus kodakarensis*; RTx(exo) is the proofreading-deficient *T. kodakarensis* DNA polymerase; Superscript II, III and IV are proprietary thermostable MLV RT variants; MarathonRT and TGIRT are group-II intron RTs from *Eubacterium rectale* and *Geobacillus stearothermophilus*, respectively; Induro RT is a commercial group-II intron-encoded RT marketed by New England Biolabs; L1-RT is a human LINE-1 RT (a non-LTR RT) (LINE: long interspersed nuclear element); and FBu is an RT encoded by the intestinal fluke *Fasciolopsis buski* R2 non-LTR retrotransposon. RT error rates reported in Houlihan et al. [54] correspond to median values. Retroviral RTs are indicated in bold.

Most of the studies described in Table 2 provide cumulative error rates (i.e., those resulting from the combination of errors present in the RNA source and those made by the RT during cDNA synthesis). In some studies, the RNA source was heterogeneous in nature and reflected the variability of the viral quasispecies [49,83]. Methods rendering cumulative error rates are not suitable to identify high-fidelity RTs since errors present in the RNA constitute a transcriptional threshold that represents a basal error rate that blurs any improvement in RT's accuracy. Therefore, a challenge for these methods is the distinction between transcription and cDNA synthesis errors. Theoretically, this could be achieved using different sets of barcodes for RNA and cDNA molecules, as described by Reid-Bayliss and Loeb, using the ARC-SEQ method [58]. In their study, the authors used in vitro-transcribed RNA from phage m13mp18, as well as total yeast RNA as templates. The estimated error rate for the T7 RNA polymerase was around 3×10^{-5}, while the error rates derived from transcription in the yeasts were estimated at 4.21×10^{-5}. ProtoScript RT was used for cDNA synthesis in these experiments, but the reverse transcription error rate was not reported.

Using the REP-SEQ methodology, Gout et al. [57] estimated a transcription error rate of 4.1×10^{-6}, well below their estimated cDNA synthesis error rate, which was 1.14×10^{-4} (for SuperScript II RT). Interestingly, the most common mutation introduced in reverse transcription was G→A, while C→T was the most common single nucleotide polymorphism (SNP) found in the RNA. The transcriptional error rate was calculated for the total RNA obtained from *Caenorhabditis elegans*. Potapov et al. recently reported separate error rates for transcription and cDNA synthesis, using a ROLL-SEQ protocol [65]. The RNA template represents an artificial sequence and was obtained by in vitro transcription with T7 RNA polymerase, and the results were analyzed using Pacific Biosciences Single Molecule Real-Time sequencing. The transcriptional threshold was estimated at 5.3×10^{-5}, similar to the cDNA synthesis error rate of the Induro RT (5.7×10^{-5}) (Table 2).

7. Conclusions and Future Developments

Understanding the fidelity of RTs, along with other sources of mutations, is essential to understand viral dynamics and plays an important role in the development of new tools for biotechnological applications. The accuracy of these enzymes has an important impact on genetic diversity, plays a critical role in the development of drug resistance and represents a critical challenge in vaccine development [108]. Accuracy is important for single-cell transcriptomics, where the detection and precise characterization of low-abundance RNAs are critical [17,109]. RTs with increased accuracy could also be valuable in synthetic biology for long-term nucleic acid data storage that uses naturally nuclease-resistant, chemically modified nucleic acids [110].

Finally, there are biotechnological applications of RTs where low fidelity is sought. One example is the detection of modified ribonucleotides, where having a high error rate at those positions is a desirable feature [111–113]. The use of NGS for the detection of RNA modifications has opened the field of epitranscriptomics [114], where the RT error rates could be exploited for the development of new methods to identify modified bases. NGS-based methods using appropriate RTs could avoid the use of antibodies directed against modified bases that often produce inaccurate results. Typically, the "PRIMER IDs" method (Figure 4A) is used to measure errors introduced by the RT when reverse transcribing an RNA template. Modifications are detected as RT signatures (e.g., mismatches, indels or RT stops) that arise when Watson-Crick pairing is disrupted, such as with N^1-methyladenosine (m1A) [115]. However, certain modifications such as N^6-methyladenosine (m6A), are categorized as "RT-silent", and thus pose a greater detection challenge. For this type of modifications, the Watson–Crick pairing is usually not affected and the information

about the base modification is erased during the reverse transcription step [116]. However, strategies exist to enhance the appearance of RT signatures, either by engineering the RT [111] or by altering the modifications to induce distortions [117].

Still, the possibility of obtaining more faithful RTs represents an important challenge that cannot be properly addressed without reducing the transcriptional threshold. Obtaining more faithful RNA polymerases could be important to manufacture high-quality mRNA, particularly relevant in vaccine development.

On the other hand, NGS is increasingly used for HIV drug resistance genotyping due to its enhanced sensitivity compared to classical Sanger sequencing methods. However, much of the observed variability in NGS sequences may reflect PCR error rates, rather than authentic mutations. Empirical thresholds for reporting drug resistance mutations detected by NGS are usually between 0.5 and 5% [118,119]. However, Tzou et al. [120] have shown that single genome sequencing (involving the use of barcodes and eliminating most of the PCR errors) could detect highly unusual mutations in the protease-, RT- and integrase-coding regions in proportions ranging from 3.9 to 6.5%, suggesting that the thresholds to be used for HIV drug resistance genotyping should be higher. In line with this proposal, comparative studies carried out in different laboratories have shown that the best agreement between NGS and Sanger sequencing methods was obtained when minority variants represented more than 20% of the NGS variation observed at one position [121], although other studies suggested that lower thresholds could be valid for specific resistance mutations (e.g., for those associated with resistance to non-nucleoside RT inhibitors) [122]. Still, the detection of minority variants is limited by errors made by the RT used in cDNA synthesis, prior to PCR amplification, and the bioinformatics pipelines used in the analysis.

Taken together, measuring transcriptional error rates, as well as designing reliable methods to reduce cDNA synthesis errors is still challenging, requiring laborious and expensive methodologies. Further developments in NGS affecting accuracy, reliability and cost will be necessary to address these issues.

References

1. Domingo, E.; Perales, C. Viral quasispecies. *PLoS Genet.* **2019**, *15*, e1008271. [CrossRef] [PubMed]
2. Suzuki, Y.; Yamaguchi-Kabata, Y.; Gojobori, T. Nucleotide substitution rates of HIV-1. *AIDS Rev.* **2000**, *2*, 39–47.
3. Fu, Y.X. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. *Mol. Biol. Evol.* **2001**, *18*, 620–626. [CrossRef] [PubMed]
4. Cuevas, J.M.; Geller, R.; Garijo, R.; López-Aldeguer, J.; Sanjuán, R. Extremely high mutation rate of HIV-1 in vivo. *PLoS Biol.* **2015**, *13*, e1002251. [CrossRef]

5. Geller, R.; Domingo-Calap, P.; Cuevas, J.M.; Rossolillo, P.; Negroni, M.; Sanjuán, R. The external domains of the HIV-1 envelope are a mutational cold spot. *Nat. Commun.* **2015**, *6*, 8571. [CrossRef]

6. Hughes, S.H. Reverse transcription of retroviruses and LTR retrotransposons. *Microbiol. Spectr.* **2015**, *3*, MDNA3-0027-2014. [CrossRef]

7. Menéndez-Arias, L.; Sebastián-Martín, A.; Álvarez, M. Viral reverse transcriptases. *Virus Res.* **2017**, *234*, 153–176. [CrossRef]

8. Svarovskaia, E.S.; Cheslock, S.R.; Zhang, W.H.; Hu, W.S.; Pathak, V.K. Retroviral mutation rates and reverse transcriptase fidelity. *Front Biosci.* **2003**, *8*, d117-34. [CrossRef] [PubMed]

9. Menéndez-Arias, L. Mutation rates and intrinsic fidelity of retroviral reverse transcriptases. *Viruses* **2009**, *1*, 1137–1165. [CrossRef]

10. Menéndez-Arias, L. HIV reverse transcriptase fidelity, clade diversity, and acquisition of drug resistance. In *Human Immunodeficiency Virus Reverse Transcriptase*; LeGrice, S., Götte, M., Eds.; Springer: New York, NY, USA, 2013; pp. 225–252. [CrossRef]

11. O'Neil, P.K.; Sun, G.; Yu, H.; Ron, Y.; Dougherty, J.P.; Preston, B.D. Mutational analysis of HIV-1 long terminal repeats to explore the relative contribution of reverse transcriptase and RNA polymerase II to viral mutagenesis. *J. Biol. Chem.* **2002**, *277*, 38053–38061. [CrossRef] [PubMed]

12. Meyerhans, A.; Vartanian, J.P.; Hultgren, C.; Plikat, U.; Karlsson, A.; Wang, L.; Eriksson, S.; Wain-Hobson, S. Restriction and enhancement of human immunodeficiency virus type 1 replication by modulation of intracellular deoxynucleoside triphosphate pools. *J. Virol.* **1994**, *68*, 535–540. [CrossRef] [PubMed]

13. Malim, M.H. APOBEC proteins and intrinsic resistance to HIV-1 infection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2009**, *364*, 675–687. [CrossRef]

14. Pfaller, C.K.; George, C.X.; Samuel, C.E. Adenosine deaminases acting on RNA (ADARs) and viral infections. *Annu. Rev. Virol.* **2021**, *8*, 239–264. [CrossRef] [PubMed]

15. Piontkivska, H.; Wales-McGrath, B.; Miyamoto, M.; Wayne, M.L. ADAR editing in viruses: An evolutionary force to reckon with. *Genome Biol. Evol.* **2021**, *13*, evab240. [CrossRef] [PubMed]

16. Menéndez-Arias, L.; Delgado, R. Update and latest advances in antiretroviral therapy. *Trends Pharmacol. Sci.* **2022**, *43*, 16–29. [CrossRef] [PubMed]

17. Martín-Alonso, S.; Frutos-Beltrán, E.; Menéndez-Arias, L. Reverse transcriptase: From transcriptomics to genome editing. *Trends Biotechnol.* **2021**, *39*, 194–210. [CrossRef] [PubMed]

18. Xavier Ruiz, F.; Arnold, E. Evolving understanding of HIV-1 reverse transcriptase structure, function, inhibition, and resistance. *Curr. Opin. Struct. Biol.* **2020**, *61*, 113–123. [CrossRef] [PubMed]

19. Menéndez-Arias, L.; Abraha, A.; Quiñones-Mateu, M.E.; Mas, A.; Camarasa, M.J.; Arts, E.J. Functional characterization of chimeric reverse transcriptases with polypeptide subunits of highly divergent HIV-1 group M and O strains. *J. Biol. Chem.* **2001**, *276*, 27470–27479. [CrossRef] [PubMed]

20. Ren, J.; Bird, L.E.; Chamberlain, P.P.; Stewart-Jones, G.B.; Stuart, D.I.; Stammers, D.K. Structure of HIV-2 reverse transcriptase at 2.35-A resolution and the mechanism of resistance to non-nucleoside inhibitors. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 14410–14415. [CrossRef] [PubMed]

21. Galilee, M.; Alian, A. The structure of FIV reverse transcriptase and its implications for non-nucleoside inhibitor resistance. *PLoS Pathog.* **2018**, *14*, e1006849. [CrossRef] [PubMed]

22. Baldwin, E.T.; Götte, M.; Tchesnokov, E.P.; Arnold, E.; Hagel, M.; Nichols, C.; Dossang, P.; Lamers, M.; Wan, P.; Steinbacher, S.; et al. Human endogenous retrovirus-K (HERV-K) reverse transcriptase (RT) structure and biochemistry reveals remarkable similarities to HIV-1 RT and opportunities for HERV-K-specific inhibition. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2200260119. [CrossRef] [PubMed]

23. Leis, J.; Duyk, G.; Johnson, S.; Longiaru, M.; Skalka, A. Mechanism of action of the endonuclease associated with the alpha beta and beta beta forms of avian RNA tumor virus reverse transcriptase. *J. Virol.* **1983**, *45*, 727–739. [CrossRef] [PubMed]

24. Nowak, E.; Potrzebowski, W.; Konarev, P.V.; Rausch, J.W.; Bona, M.K.; Svergun, D.I.; Bujnicki, J.M.; Le Grice, S.F.; Nowotny, M. Structural analysis of monomeric retroviral reverse transcriptase in complex with an RNA/DNA hybrid. *Nucleic Acids Res.* **2013**, *41*, 3874–3887. [CrossRef] [PubMed]

25. Nowacka, M.; Nowak, E.; Czarnocki-Cieciura, M.; Jackiewicz, J.; Skowronek, K.; Szczepanowski, R.H.; Wöhrl, B.M.; Nowotny, M. Structures of substrate complexes of foamy viral protease-reverse transcriptase. *J. Virol.* **2021**, *95*, e0084821. [CrossRef] [PubMed]

26. Wöhrl, B.M. Structural and functional aspects of foamy virus protease-reverse transcriptase. *Viruses* **2019**, *11*, 598. [CrossRef]

27. Harrison, J.J.E.K.; Tuske, S.; Das, K.; Ruiz, F.X.; Bauman, J.D.; Boyer, P.L.; DeStefano, J.J.; Hughes, S.H.; Arnold, E. Crystal structure of a retroviral polyprotein: Prototype foamy virus protease-reverse transcriptase (PR-RT). *Viruses* **2021**, *13*, 1495. [CrossRef] [PubMed]

28. González-Delgado, A.; Mestre, M.R.; Martínez-Abarca, F.; Toro, N. Prokaryotic reverse transcriptases: From retroelements to specialized defense systems. *FEMS Microbiol. Rev.* **2021**, *45*, fuab025. [CrossRef]

29. Mestre, M.R.; Gao, L.A.; Shah, S.A.; López-Beltrán, A.; González-Delgado, A.; Martínez-Abarca, F.; Iranzo, J.; Redrejo-Rodríguez, M.; Zhang, F.; Toro, N. UG/Abi: A highly diverse family of prokaryotic reverse transcriptases associated with defense functions. *Nucleic Acids Res.* **2022**, *50*, 6084–6101. [CrossRef]

30. Mohr, S.; Ghanem, E.; Smith, W.; Sheeter, D.; Qin, Y.; King, O.; Polioudakis, D.; Iyer, V.R.; Hunicke-Smith, S.; Swamy, S.; et al. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* **2013**, *19*, 958–970. [CrossRef]

31. Zhao, C.; Liu, F.; Pyle, A.M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* **2018**, *24*, 183–195. [CrossRef]

32. Stamos, J.L.; Lentzsch, A.M.; Lambowitz, A.M. Structure of a thermostable group II intron reverse transcriptase with template-primer and its functional and evolutionary implications. *Mol. Cell* **2017**, *68*, 926–939.e4. [CrossRef]

33. Battula, N.; Dube, D.; Loeb, L.A. Avian myeloblastosis virus DNA polymerase. Kinetic studies on the incorporation of noncomplementary nucleotides. *J. Biol. Chem.* **1975**, *250*, 8404–8408. [CrossRef] [PubMed]

34. Sirover, M.A.; Loeb, L.A. On the fidelity of DNA replication. Effect of metal activators during synthesis with avian myeloblastosis virus DNA polymerase. *J. Biol. Chem.* **1977**, *252*, 3605–3610. [CrossRef] [PubMed]

35. Preston, B.D.; Poiesz, B.J.; Loeb, L.A. Fidelity of HIV-1 reverse transcriptase. *Science* **1988**, *242*, 1168–1171. [CrossRef] [PubMed]

36. Menéndez-Arias, L. Molecular basis of fidelity of DNA synthesis and nucleotide specificity of retroviral reverse transcriptases. *Prog. Nucleic Acid Res. Mol. Biol.* **2002**, *71*, 91–147. [CrossRef] [PubMed]

37. Echols, H.; Goodman, M.F. Fidelity mechanisms in DNA replication. *Annu. Rev. Biochem.* **1991**, *60*, 477–511. [CrossRef]

38. Johnson, K.A. Conformational coupling in DNA polymerase fidelity. *Annu. Rev. Biochem.* **1993**, *62*, 685–713. [CrossRef] [PubMed]

39. Bebenek, K.; Kunkel, T.A. Analyzing fidelity of DNA polymerases. *Methods Enzymol.* **1995**, *262*, 217–232. [CrossRef]

40. Goodwin, S.; McPherson, J.D.; McCombie, W.R. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **2016**, *17*, 333–351. [CrossRef]

41. Satam, H.; Joshi, K.; Mangrolia, U.; Waghoo, S.; Zaidi, G.; Rawool, S.; Thakare, R.P.; Banday, S.; Mishra, A.K.; Das, G.; et al. Next-generation sequencing technology: Current trends and advancements. *Biology* **2023**, *12*, 997. [CrossRef]

42. Dohm, J.C.; Lottaz, C.; Borodina, T.; Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **2008**, *36*, e105. [CrossRef] [PubMed]

43. Nakamura, K.; Oshima, T.; Morimoto, T.; Ikeda, S.; Yoshikawa, H.; Shiwa, Y.; Ishikawa, S.; Linak, M.C.; Hirai, A.; Takahashi, H.; et al. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **2011**, *39*, e90. [CrossRef]

44. Quail, M.; Smith, M.E.; Coupland, P.; Otto, T.D.; Harris, S.R.; Connor, T.R.; Bertoni, A.; Swerdlow, H.P.; Gu, Y. A tale of three next generation sequencing platforms: Comparison of Ion torrent, pacific biosciences and Illumina MiSeq sequencers. *BMC Genom.* **2012**, *13*, 341. [CrossRef]

45. Schirmer, M.; Ijaz, U.Z.; D'Amore, R.; Hall, N.; Sloan, W.T.; Quince, C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **2015**, *43*, e37. [CrossRef] [PubMed]

46. Logsdon, G.A.; Vollger, M.R.; Eichler, E.E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **2020**, *21*, 597–614. [CrossRef] [PubMed]

47. Salk, J.J.; Schmitt, M.W.; Loeb, L.A. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat. Rev. Genet.* **2018**, *19*, 269–285. [CrossRef]

48. Hiatt, J.B.; Patwardhan, R.P.; Turner, E.H.; Lee, C.; Shendure, J. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods* **2010**, *7*, 119–122. [CrossRef]

49. Jabara, C.B.; Jones, C.D.; Roach, J.; Anderson, J.A.; Swanstrom, R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 20166–20171. [CrossRef] [PubMed]

50. Kinde, I.; Wu, J.; Papadopoulos, N.; Kinzler, K.W.; Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 9530–9535. [CrossRef]

51. Ellefson, J.W.; Gollihar, J.; Shroff, R.; Shivram, H.; Iyer, V.R.; Ellington, A.D. Synthetic evolutionary origin of a proofreading reverse transcriptase. *Science* **2016**, *352*, 1590–1593. [CrossRef] [PubMed]

52. Yasukawa, K.; Iida, K.; Okano, H.; Hidese, R.; Baba, M.; Yanagihara, I.; Kojima, K.; Takita, T.; Fujiwara, S. Next-generation sequencing-based analysis of reverse transcriptase fidelity. *Biochem. Biophys. Res. Commun.* **2017**, *492*, 147–153. [CrossRef] [PubMed]

53. Okano, H.; Baba, M.; Ryota, H.; Iida, K.; Li, T.; Kojima, K.; Takita, T.; Yanagihara, I.; Fujiwara, S.; Yasukawa, K. Accurate fidelity analysis of the reverse transcriptase by a modified next-generation sequencing. *Enzyme Microb. Technol.* **2018**, *115*, 81–85. [CrossRef] [PubMed]

54. Houlihan, G.; Arangundy-Franklin, S.; Porebski, B.T.; Subramanian, N.; Taylor, A.I.; Holliger, P. Discovery and evolution of RNA and XNA reverse transcriptase function and fidelity. *Nat. Chem.* **2020**, *12*, 683–690. [CrossRef]

55. Wang, R.; Belew, A.T.; Achuthan, V.; El Sayed, N.; DeStefano, J.J. Physiological magnesium concentrations increase fidelity of diverse reverse transcriptases from HIV-1, HIV-2, and foamy virus, but not MuLV or AMV. *J. Gen. Virol.* **2021**, *102*, 001708. [CrossRef]

56. Martínez del Río, J.; Frutos-Beltrán, E.; Sebastián-Martín, A.; Lasala, F.; Yasukawa, K.; Delgado, R.; Menéndez-Arias, L. HIV-1 reverse transcriptase error rates and transcriptional thresholds based on single-strand consensus sequencing of target RNA derived from in vitro-transcription and HIV-infected cells. *J. Mol. Biol.* **2024**, *436*, 168815. [CrossRef]

57. Gout, J.F.; Thomas, W.K.; Smith, Z.; Okamoto, K.; Lynch, M. Large-scale detection of in vivo transcription errors. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 18584–18589. [CrossRef]

58. Reid-Bayliss, K.S.; Loeb, L.A. Accurate RNA consensus sequencing for high-fidelity detection of transcriptional mutagenesis-induced epimutations. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 9415–9420. [CrossRef] [PubMed]

59. Schmitt, M.W.; Kennedy, S.R.; Salk, J.J.; Fox, E.J.; Hiatt, J.B.; Loeb, L.A. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 14508–14513. [CrossRef] [PubMed]

60. Kennedy, S.R.; Schmitt, M.W.; Fox, E.J.; Kohrn, B.F.; Salk, J.J.; Ahn, E.H.; Prindle, M.J.; Kuong, K.J.; Shen, J.C.; Risques, R.A.; et al. Detecting ultralow-frequency mutations by duplex sequencing. *Nat. Protoc.* **2014**, *9*, 2586–2606, Erratum in *Nat. Protoc.* **2014**, *9*, 2903. [CrossRef] [PubMed]

61. Newman, A.M.; Lovejoy, A.F.; Klass, D.M.; Kurtz, D.M.; Chabon, J.J.; Scherer, F.; Stehr, H.; Liu, C.L.; Bratman, S.V.; Say, C.; et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat. Biotechnol.* **2016**, *34*, 547–555. [CrossRef]

62. Traverse, C.C.; Ochman, H. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 3311–3316. [CrossRef]

63. Lou, D.I.; Hussmann, J.A.; McBee, R.M.; Acevedo, A.; Andino, R.; Press, W.H.; Sawyer, S.L. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 19872–19877. [CrossRef]

64. Potapov, V.; Fu, X.; Dai, N.; Corrêa, I.R.; Tanner, N.A.; Ong, J.L. Base modifications affecting RNA polymerase and reverse transcriptase fidelity. *Nucleic Acids Res.* **2018**, *46*, 5753–5763. [CrossRef]

65. Potapov, V.; Krudup, S.; Maguire, S.; Unlu, I.; Buss, J.A.; Guan, S.; Smail, B.; Van Euwen, T.; Taylor, M.S.; Burns, K.H.; et al. Discrete measurements of RNA polymerase and reverse transcriptase fidelity reveal evolutionary tuning. *RNA* **2024**, *30*, 1246–1258. [CrossRef] [PubMed]

66. Acevedo, A.; Brodsky, L.; Andino, R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **2014**, *505*, 686–690. [CrossRef] [PubMed]

67. Gout, J.-F.; Li, W.; Fritsch, C.; Li, A.; Haroon, S.; Singh, L.; Hua, D.; Fazelinia, H.; Smith, Z.; Seeholzer, S.; et al. The landscape of transcription errors in eukaryotic cells. *Sci. Adv.* **2017**, *3*, e1701484. [CrossRef] [PubMed]

68. Traverse, C.C.; Ochman, H. Genome-wide spectra of transcription insertions and deletions reveal that slippage depends on RNA:DNA hybrid complementarity. *mBio* **2017**, *8*, e01230. [CrossRef]

69. Chung, C.; Verheijen, B.M.; Navapanich, Z.; McGann, E.G.; Shemtov, S.; Lai, G.-J.; Arora, P.; Towheed, A.; Haroon, S.; Holczbauer, A.; et al. Evolutionary conservation of the fidelity of transcription. *Nat. Commun.* **2023**, *14*, 1547. [CrossRef]

70. Eid, J.; Fehr, A.; Gray, J.; Luong, K.; Lyle, J.; Otto, G.; Peluso, P.; Rank, D.; Baybayan, P.; Bettman, B.; et al. Real-time DNA sequencing from single polymerase molecules. *Science* **2009**, *323*, 133–138. [CrossRef]

71. Li, C.; Chng, K.R.; Boey, E.J.H.; Ng, A.H.Q.; Wilm, A.; Nagarajan, N. INC-Seq: Accurate single molecule reads using nanopore sequencing. *GigaScience* **2016**, *5*, s13742-016-0140-7. [CrossRef] [PubMed]

72. Álvarez, M.; Menéndez-Arias, L. Temperature effects on the fidelity of a thermostable HIV-1 reverse transcriptase. *FEBS J.* **2014**, *281*, 342–351. [CrossRef]

73. Yeo, J.Y.; Goh, G.R.; Su, C.T.; Gan, S.K. The determination of HIV-1 RT mutation rate, its possible allosteric effects, and its implications on drug resistance. *Viruses* **2020**, *12*, 297. [CrossRef] [PubMed]

74. Meissner, M.E.; Julik, E.J.; Badalamenti, J.P.; Arndt, W.G.; Mills, L.J.; Mansky, L.M. Development of a user-friendly pipeline for mutational analyses of HIV using ultra-accurate maximum-depth sequencing. *Viruses* **2021**, *13*, 1338. [CrossRef] [PubMed]

75. Deakin, C.T.; Deakin, J.J.; Ginn, S.L.; Young, P.; Humphreys, D.; Suter, C.M.; Alexander, I.E.; Hallwirth, C.V. Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence. *Nucleic Acids Res.* **2014**, *42*, e129. [CrossRef]

76. Zhang, T.H.; Wu, N.C.; Sun, R. A benchmark study on error-correction by read-pairing and tag-clustering in amplicon-based deep sequencing. *BMC Genom.* **2016**, *17*, 108. [CrossRef] [PubMed]

77. Potapov, V.; Ong, J.L. Examining sources of error in PCR by single-molecule sequencing. *PLoS ONE* **2017**, *12*, e0169774. [CrossRef]

78. Wang, T.T.; Abelson, S.; Zou, J.; Li, T.; Zhao, Z.; Dick, J.E.; Shlush, L.I.; Pugh, T.J.; Bratman, S.V. High efficiency error suppression for accurate detection of low-frequency variants. *Nucleic Acids Res.* **2019**, *47*, e87. [CrossRef] [PubMed]

79. Boyer, J.C.; Bębenek, K.; Kunkel, T.A. Unequal human immunodeficiency virus type 1 reverse transcriptase error rates with RNA and DNA templates. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 6919–6923. [CrossRef]

80. Sebastián-Martín, A.; Barrioluengo, V.; Menéndez-Arias, L. Transcriptional inaccuracy threshold attenuates differences in RNA-dependent DNA synthesis fidelity between retroviral reverse transcriptases. *Sci. Rep.* **2018**, *8*, 627. [CrossRef]

81. Shagin, D.A.; Shagina, I.A.; Zaretsky, A.R.; Barsova, E.V.; Kelmanson, I.V.; Lukyanov, S.; Chudakov, D.M.; Shugay, M. A high-throughput assay for quantitative measurement of PCR errors. *Sci. Rep.* **2017**, *7*, 2718. [CrossRef] [PubMed]

82. Kircher, M.; Heyn, P.; Kelso, J. Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genom.* **2011**, *12*, 382. [CrossRef] [PubMed]

83. Zhou, S.; Jones, C.; Mieczkowski, P.; Swanstrom, R. Primer ID validates template sampling depth and greatly reduces the error rate of next-generation sequencing of HIV-1 genomic RNA populations. *J. Virol.* **2015**, *89*, 8540–8555. [CrossRef]

84. Canzoniero, J.V.; Cravero, K.; Park, B.H. The impact of collisions on the ability to detect rare mutant alleles using barcode-type next-generation sequencing techniques. *Cancer Inform.* **2017**, *16*, 117693511771923. [CrossRef]

85. Sheward, D.J.; Murrell, B.; Williamson, C. Degenerate primer IDs and the birthday problem. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E1330; author reply E1331. [CrossRef]

86. Smith, T.; Heger, A.; Sudbery, I. UMI-tools: Modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.* **2017**, *27*, 491–499. [CrossRef] [PubMed]

87. Liang, R.H.; Mo, T.; Dong, W.; Lee, G.Q.; Swenson, L.C.; McCloskey, R.M.; Woods, C.K.; Brumme, C.J.; Ho, C.K.Y.; Schinkel, J.; et al. Theoretical and experimental assessment of degenerate primer tagging in ultra-deep applications of next-generation sequencing. *Nucleic Acids Res.* **2014**, *42*, e98. [CrossRef] [PubMed]

88. Roberts, J.D.; Bebenek, K.; Kunkel, T.A. The accuracy of reverse transcriptase from HIV-1. *Science* **1988**, *242*, 1171–1173. [CrossRef] [PubMed]

89. Roberts, J.D.; Preston, B.D.; Johnston, L.A.; Soni, A.; Loeb, L.A.; Kunkel, T.A. Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis in vitro. *Mol. Cell. Biol.* **1989**, *9*, 469–476. [CrossRef]

90. Stuke, A.W.; Ahmad-Omar, O.; Hoefer, K.; Hunsmann, G.; Jentsch, K.D. Mutations in the SIV env and the M13 lacZα gene generated in vitro by reverse transcriptases and DNA polymerases. *Arch. Virol.* **1997**, *142*, 1139–1154. [CrossRef]

91. Diamond, T.L.; Souroullas, G.; Weiss, K.K.; Lee, K.Y.; Bambara, R.A.; Dewhurst, S.; Kim, B. Mechanistic understanding of an altered fidelity simian immunodeficiency virus reverse transcriptase mutation, V148I, identified in a pig-tailed macaque. *J. Biol. Chem.* **2003**, *278*, 29913–29924. [CrossRef]

92. Operario, D.J.; Reynolds, H.M.; Kim, B. Comparison of DNA polymerase activities between recombinant feline immunodeficiency and leukemia virus reverse transcriptases. *Virology* **2005**, *335*, 106–121. [CrossRef] [PubMed]

93. Boyer, P.L.; Stenbak, C.R.; Hoberman, D.; Linial, M.L.; Hughes, S.H. In vitro fidelity of the prototype primate foamy virus (PFV) RT compared to HIV-1 RT. *Virology* **2007**, *367*, 253–264. [CrossRef] [PubMed]

94. Álvarez, M.; Matamoros, T.; Menéndez-Arias, L. Increased thermostability and fidelity of DNA synthesis of wild-type and mutant HIV-1 group O reverse transcriptases. *J. Mol. Biol.* **2009**, *392*, 872–884. [CrossRef] [PubMed]

95. Jamburuthugoda, V.K.; Eickbush, T.H. The reverse transcriptase encoded by the non-LTR retrotransposon R2 is as error-prone as that encoded by HIV-1. *J. Mol. Biol.* **2011**, *407*, 661–672. [CrossRef] [PubMed]

96. Barrioluengo, V.; Wang, Y.; Le Grice, S.F.J.; Menéndez-Arias, L. Intrinsic DNA synthesis fidelity of xenotropic murine leukemia virus-related virus reverse transcriptase. *FEBS J.* **2012**, *279*, 1433–1444. [CrossRef]

97. Álvarez, M.; Sebastián-Martín, A.; García-Marquina, G.; Menéndez-Arias, L. Fidelity of classwide-resistant HIV-2 reverse transcriptase and differential contribution of K65R to the accuracy of HIV-1 and HIV-2 reverse transcriptases. *Sci. Rep.* **2017**, *7*, 44834. [CrossRef] [PubMed]

98. Achuthan, V.; Keith, B.J.; Connolly, B.A.; DeStefano, J.J. Human immunodeficiency virus reverse transcriptase displays dramatically higher fidelity under physiological magnesium conditions in vitro. *J. Virol.* **2014**, *88*, 8514–8527, Erratum in *J. Virol.* **2015**, *89*, 1944. [CrossRef]

99. Ji, J.; Loeb, L.A. Fidelity of HIV-1 reverse transcriptase copying a hypervariable region of the HIV-1 env gene. *Virology* **1994**, *199*, 323–330. [CrossRef]

100. Ji, J.; Loeb, L.A. Fidelity of HIV-1 reverse transcriptase copying RNA in vitro. *Biochemistry* **1992**, *31*, 954–958. [CrossRef] [PubMed]

101. Mansky, L.M.; Temin, H.M. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J. Virol.* **1995**, *69*, 5087–5094. [CrossRef]

102. Mansky, L.M. Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. *AIDS Res. Hum. Retroviruses* **1996**, *12*, 307–314. [CrossRef] [PubMed]

103. Abram, M.E.; Ferris, A.L.; Shao, W.; Alvord, W.G.; Hughes, S.H. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *J. Virol.* **2010**, *84*, 9864–9878. [CrossRef]

104. Ávila-Ríos, S.; Parkin, N.; Swanstrom, R.; Paredes, R.; Shafer, R.; Ji, H.; Kantor, R. Next-generation sequencing for HIV drug resistance testing: Laboratory, clinical, and implementation considerations. *Viruses* **2020**, *12*, 617. [CrossRef]

105. Metzner, K.J. Technologies for HIV-1 drug resistance testing: Inventory and needs. *Curr. Opin. HIV AIDS* **2022**, *17*, 222–228. [CrossRef]

106. Zhou, S.; Swanstrom, R. Fact and fiction about 1%: Next generation sequencing and the detection of minor drug resistant variants in HIV-1 populations with and without unique molecular identifiers. *Viruses* **2020**, *12*, 850. [CrossRef] [PubMed]

107. Rawson, J.M.O.; Gohl, D.M.; Landman, S.R.; Roth, M.E.; Meissner, M.E.; Peterson, T.S.; Hodges, J.S.; Beckman, K.B.; Mansky, L.M. Single-strand consensus sequencing reveals that HIV type but not subtype significantly impacts viral mutation frequencies and spectra. *J. Mol. Biol.* **2017**, *429*, 2290–2307. [CrossRef] [PubMed]

108. Megens, S.; Laethem, K.V. HIV-1 genetic variation and drug resistance development. *Expert Rev. Anti-Infect. Ther.* **2013**, *11*, 1159–1178. [CrossRef]

109. Huber, L.B.; Betz, K.; Marx, A. Reverse transcriptases: From discovery and applications to xenobiology. *ChemBioChem* **2023**, *24*, e202200521. [CrossRef] [PubMed]

110. Shroff, R.; Ellefson, J.W.; Wang, S.S.; Boulgakov, A.A.; Hughes, R.A.; Ellington, A.D. Recovery of information stored in modified DNA with an evolved polymerase. *ACS Synth. Biol.* **2022**, *11*, 554–561. [CrossRef] [PubMed]

111. Aschenbrenner, J.; Werner, S.; Marchand, V.; Adam, M.; Motorin, Y.; Helm, M.; Marx, A. Engineering of a DNA polymerase for direct m6A sequencing. *Angew. Chem.* **2017**, *57*, 417–421. [CrossRef] [PubMed]

112. Zhou, H.; Rauch, S.; Dai, Q.; Cui, X.; Zhang, Z.; Nachtergaele, S.; Sepich, C.; He, C.; Dickinson, B.C. Evolution of a reverse transcriptase to map N1-methyladenosine in human messenger RNA. *Nat. Methods* **2019**, *16*, 1281–1288. [CrossRef]

113. Huber, L.B.; Kaur, N.; Henkel, M.; Marchand, V.; Motorin, Y.; Ehrenhofer-Murray, A.E.; Marx, A. A dual-purpose polymerase engineered for direct sequencing of pseudouridine and queuosine. *Nucleic Acids Res.* **2023**, *51*, 3971–3987. [CrossRef]

114. Motorin, Y.; Helm, M. General principles and limitations for detection of RNA modifications by sequencing. *Acc. Chem. Res.* **2024**, *57*, 275–288. [CrossRef]

115. Hauenschild, R.; Tserovski, L.; Schmid, K.; Thüring, K.; Winz, M.-L.; Sharma, S.; Entian, K.-D.; Wacheul, L.; Lafontaine, D.L.J.; Anderson, J.M.; et al. The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent. *Nucleic Acids Res.* **2015**, *43*, 9950–9964. [CrossRef] [PubMed]

116. Vilfan, I.D.; Tsai, Y.-C.; Clark, T.A.; Wegener, J.; Dai, Q.; Yi, C.; Pan, T.; Turner, S.W.; Korlach, J. Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription. *J. Nanobiotechnol.* **2013**, *11*, 8. [CrossRef] [PubMed]

117. Wang, Y.; Xiao, Y.; Dong, S.; Yu, Q.; Jia, G. Antibody-free enzyme-assisted chemical approach for detection of N6-methyladenosine. *Nat. Chem. Biol.* **2020**, *16*, 896–903. [CrossRef] [PubMed]

118. Ji, H.; Enns, E.; Brumme, C.J.; Parkin, N.; Howison, M.; Lee, E.R.; Capina, R.; Marinier, E.; Avila-Rios, S.; Sandstrom, P.; et al. Bioinformatic data processing pipelines in support of next-generation sequencing-based HIV drug resistance testing: The Winnipeg Consensus. *J. Int. AIDS Soc.* **2018**, *21*, e25193. [CrossRef]

119. Taylor, T.; Lee, E.R.; Nykoluk, M.; Enns, E.; Liang, B.; Capina, R.; Gauthier, M.K.; Domselaar, G.V.; Sandstrom, P.; Brooks, J.; et al. A MiSeq-HyDRA platform for enhanced HIV drug resistance genotyping and surveillance. *Sci. Rep.* **2019**, *9*, 8970. [CrossRef]

120. Tzou, P.L.; Rhee, S.Y.; Shafer, R.W. Amino acid prevalence of HIV-1 pol mutations by direct polymerase chain reaction and single genome sequencing. *AIDS Res. Hum. Retroviruses* **2019**, *35*, 924–929. [CrossRef] [PubMed]

121. Parkin, N.T.; Avila-Rios, S.; Bibby, D.F.; Brumme, C.J.; Eshleman, S.H.; Harrigan, P.R.; Howison, M.; Hunt, G.; Ji, H.; Kantor, R.; et al. Multi-laboratory comparison of next-generation to Sanger-based sequencing for HIV-1 drug resistance genotyping. *Viruses* **2020**, *12*, 694. [CrossRef]

122. Ouyang, F.; Yuan, D.; Zhai, W.; Liu, S.; Zhou, Y.; Yang, H. HIV-1 drug resistance detected by next-generation sequencing among ART-naïve individuals: A systematic review and meta-analysis. *Viruses* **2024**, *16*, 239. [CrossRef] [PubMed]