


Commentary

Chemistry and Bioinformatics Considerations in Using Next-Generation Sequencing Technologies to Inferring HIV Proviral DNA Genome-Intactness

Guinevere Q. Lee 

Division of Infectious Diseases, Department of Medicine, Weill Cornell Medicine, New York, NY 10021, USA; gul4001@med.cornell.edu

Abstract: HIV persists via integration of the viral DNA into the human genome. The HIV DNA pool within an infected individual is a complex population that comprises both intact and defective viral genomes, each with a distinct integration site, in addition to a unique repertoire of viral quasi-species. Obtaining an accurate profile of the viral DNA pool is critical to understanding viral persistence and resolving interhost differences. Recent advances in next-generation deep sequencing (NGS) technologies have enabled the development of two sequencing assays to capture viral near-full-genome sequences at single molecule resolution (FLIP-seq) or to co-capture full-length viral genome sequences in conjunction with its associated viral integration site (MIP-seq). This commentary aims to provide an overview on both FLIP-seq and MIP-seq, discuss their strengths and limitations, and outline specific chemistry and bioinformatics concerns when using these assays to study HIV persistence.



Citation: Lee, G.Q. Chemistry and Bioinformatics Considerations in Using Next-Generation Sequencing Technologies to Inferring HIV Proviral DNA Genome-Intactness. *Viruses* **2021**, *13*, 1874. <https://doi.org/10.3390/v13091874>

Academic Editor: Francesco Andrea Procopio

Received: 21 July 2021

Accepted: 6 September 2021

Published: 19 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: HIV genomes; HIV persistence; deep sequencing

1. Introduction

HIV-1 infection leads to lifelong viral persistence. Upon infection, the viral RNA genome is reverse transcribed into viral cDNA, which is followed by an irreversible integration into the human genome [1]. This results in the establishment of a viral DNA reservoir that fuels subsequent viral replication cycles when treatment is stopped [2–5]. Antiretroviral treatment is effective in suppressing ongoing viral replications but does not eliminate integrated HIV genomes. As such, HIV-infected individuals rely on lifelong treatment to suppress ongoing viral replications. In the absence of treatment, viral replication resumes, new CD4⁺ cells are infected, and infected individuals progress to develop acquired immunodeficiency syndrome (AIDS) [6].

The viral DNA reservoir that sustains HIV persistence is extremely stable in size and has been estimated to have a half-life of approximately 44 months in individuals receiving long term antiretroviral treatment [5], suggesting the general inability of the immune system to naturally clear the viral reservoir despite suppression of active viral replication. Recent studies have revealed that viral persistence is maintained, at least partially, by clonal expansion of infected cells [7–15], which is driven by mechanisms such as antigenic stimulation [9,16], homeostatic proliferation [15,17], and proliferation associated with host genomic locations of viral integration [10,18]. The relative abundances of these clonally expanded infected cells wax and wane over time [15]. In other words, the viral DNA pool within an individual is not a static population and has a relatively slow half-life.

The viral DNA pool is also genetically diverse: even during hyperacute heterosexually-transmitted HIV infection, when a single founder virus is presumed [19], each HIV-DNA genome contains at least one single-base nucleotide substitution mutation [20], presumably attributable to the high error rate of the HIV reverse transcriptase. Genetically similar but non-identical viral sequences within an infected individual are called “viral quasi-species”.

In addition, a close examination of the genotypic compositions of HIV DNA in chronically infected individuals revealed that over 90% of the viral DNA genomes are heavily truncated, have deleterious insertions/deletions, have been excessively hypermutated, and/or have single-base substitution mutations that would yield premature stop codons in essential viral genes [21,22]. Viral genomes that contained such decapacitating alterations are incapable of fueling virologic rebound in the absence of treatment and have been termed “defective HIV-DNA genomes” as opposed to “genome-intact HIV-DNA genomes,” which lack obvious defects. Furthermore, each HIV-DNA genome is also integrated into distinct locations within the host chromosome, creating another factor that contributes to reservoir population diversity. Recent studies have suggested that integration sites of genome-intact HIV proviruses into transcriptionally less-active human chromosomal regions may be associated with a decreased likelihood of viral transcription activation [23,24].

In summary, the HIV reservoir population structure within a single infected individual is complex, changes over time, and contains viral genomes that are either intact or defective, while each viral genome is associated with unique viral integration sites that may impact their likelihoods of transcription activation. To study viral persistence and its longitudinal dynamics and to identify future targets for HIV cure research, it is therefore crucial to accurately characterize “genome-intact” HIV-DNA genomes. In this commentary, the author will discuss technical considerations and limitations of two assays, FLIP-seq and MIP-seq, both of which are single-copy, next-generation deep sequencing techniques for the study of HIV DNA genomes and reservoirs.

2. Traditional Assays and the Subsequent Development of FLIP-Seq and MIP-Seq

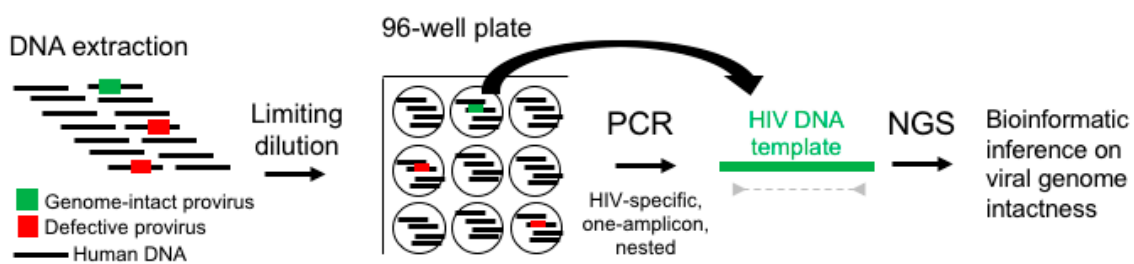
Traditional techniques for the study of HIV reservoirs include cell-culture-based quantitative viral outgrowth assay (qVOA), quantitative short-amplicon PCR (qPCR), and droplet digital PCR (ddPCR). Each of these assays has merits and limitations: qVOA relies on the stimulation of infected cells *ex vivo* and detection of viral RNA production in culture supernatant and measures true replication-competent and genome-intact proviruses but is labor-intensive, and it has been shown that a single round of activation is insufficient to reactivate all replication-competent genome-intact proviruses [13]. Both qPCR- and ddPCR-based total HIV-DNA reservoir sizes quantification approaches amplify and quantify short regions of the viral DNA genome and are relatively inexpensive and scalable but do not distinguish between intact versus defective HIV-DNA genomes [25,26].

In 2013, using HIV near-full-genome Sanger sequencing, a paradigm-shifting study by Ho et al. [21] showed that the vast majority of viral DNA in infected individuals are genome-defective, prompting the HIV persistence and cure research community to shift focus onto identification of cells infected with genome-intact proviruses. In 2017, two groups independently developed and published next-generation deep sequencing versions (as opposed to Sanger sequencing [21]) for near-full-genome HIV-DNA sequencing [7,8], nowadays known as FLIPS or FLIP-seq. Since then, multiple research groups also have, based on these existing proviral full-genome sequence data, developed new qPCR or ddPCR methods, such as the ddPCR-based Intact Proviral DNA Assay (IPDA) [27], a hybrid qPCR/sequencing method Q4PCR [28], and a ddPCR method by Levy et al. [29], all of which use multiple probes and multiplexing to infer and quantify intact versus defective proviral genome status. This article will focus on evaluating single genome amplification and sequencing methods and will use the term FLIP-seq to refer to the assay as published by [7], but the biochemical and technical considerations discussed below may be applied to any single genome amplification and sequencing assays.

Briefly, similar to the 2013 Sanger sequencing method [21], FLIP-seq (Figure 1a) starts with a DNA extraction of an infected cell population (further discussed in Section 2), followed by a rough quantification of total HIV-DNA copies by either qPCR/ddPCR or serial dilution to estimate the copy numbers of total HIV-DNA genomes concentration per extraction volume within the sample. Using this concentration estimate, limiting dilution of the DNA extract is performed by diluting the extract to one HIV-DNA template-positive per

three PCR reactions according to Poisson distribution (Section 3). Then, using HIV-specific primers validated for subtype B [7], C [20], and D [30] HIV-1, each reaction is subjected to PCR (Section 4) to amplify near-full-genome HIV DNA. Since each PCR-positive reaction contains the amplification products originated from approximately a single HIV-DNA molecule, this set up is termed “single genome amplification” (SGA). Resulting amplicons are each subjected to next-generation sequencing library preparation and tagging using unique molecular indexes, then pooled and deep sequenced (Section 4). Resulting short reads are demultiplexed, de-novo assembled, and subjected to bioinformatics inferences on genome-intactness (Section 5).

(a) FLIP-seq



(b) MIP-seq

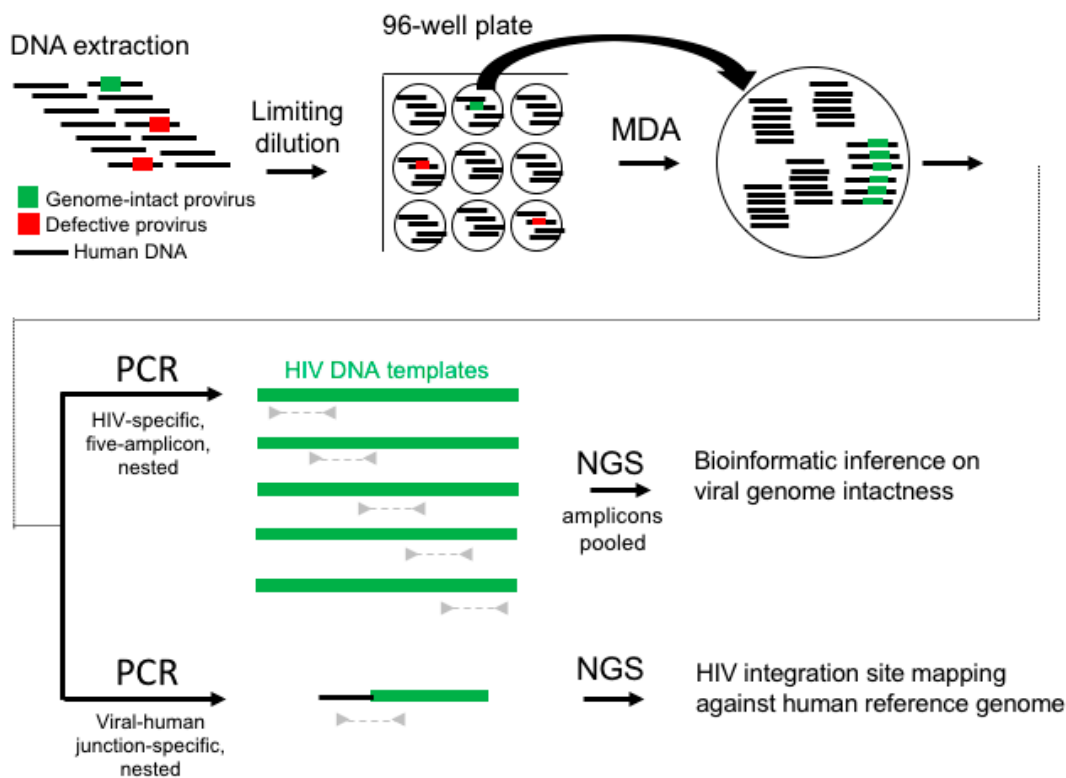


Figure 1. FLIP-seq and MIP-seq workflow. (a) FLIP-seq begins with DNA extraction (further discussed in Section 2), followed by limiting dilution to a single HIV-DNA template per subsequent PCR reaction (Section 3), near-full-genome single-amplicon HIV-DNA PCR amplification (Section 4), and finally next-generation deep sequencing (NGS) and bioinformatic

inference on viral-genome intactness (Section 5). (b) Similar to FLIP-seq, MIP-seq begins with DNA extraction (Section 2), followed by limiting dilution to achieve single HIV-DNA template per subsequent reaction (Section 3), then multiple displacement amplification (MDA) by random primers. Resulting reaction is split for near-full-genome five-overlapping-amplicon HIV-DNA PCR amplification and viral-human junction amplification (Section 4), then subjected to NGS and bioinformatics inference on viral-genome intactness and mapping of viral-human DNA junctions against the human reference genome (Section 5). Note that MIP-seq would not yield full-length sequences of defective viral genomes that do not contain any of the primer binding sites targeted by the 20 primers used in the five-amplicon nested-PCR approach and was designed specifically to capture near-full-length HIV DNA that are approximately >8000 base pairs in length. A single-amplicon, near-full-genome PCR approach was not used in MIP-seq because it was markedly less sensitive due to the average amplification product lengths at the MDA step.

FLIP-seq yields high-resolution HIV-DNA genome sequences; however, to study the integration site of genome-intact proviruses, another technological breakthrough was needed: traditional techniques to examine HIV integration sites involves Sanger or deep sequencing of the viral-host junctions [31]. This approach is scalable, but targeting a short genomic region around the viral-host junction did not allow for the discrimination of integration sites associated with genome-intact versus defect HIV DNA. In 2019, two groups independently published deep sequencing methods to co-sequence full-genome HIV DNA and viral integration sites. These assays were named MIP-seq [32] and MDA-SGS (Multiple Displacement Amplification Single Genome Sequencing) [33], respectively. This article specifically focuses on evaluating biochemical and technical considerations of MIP-seq [32], but the considerations discussed below may be applied to both assays.

Similar to FLIP-seq, MIP-seq (Figure 1b) starts with DNA extraction of an infected cell population (Section 2) and limiting dilution (Section 3). Each single viral genome dilution aliquot is then subjected to multiple displacement amplification (MDA) in order to unbiasedly amplify all DNA genetic materials within the aliquot. This reaction is then split into two portions: one of which is subjected to a five-overlapping-amplicon HIV genome PCR amplification (Section 4), and the other portion is subjected to viral-human DNA junction amplification. All resulting amplicons are deep sequenced (Section 4), followed by bioinformatic inferences of viral genome intactness and the identification of viral integration sites coordinates within the human genome (Section 5).

3. DNA Extraction

One of main purposes of both FLIP-seq and MIP-seq is to capture genome-intact HIV DNA, which is approximately 10,000 base pairs in length [34]. It is therefore crucial that the chosen extraction method does not introduce extensive shearing of DNA templates to below the target capture length. As different extraction methods introduce different DNA shearing profiles [35,36], the choice of extraction method will impact assay sensitivity in terms of full-length viral genome recovery. Another factor that impacts recovery is the extraction mechanism: column-based commercial extraction kits are known to have lower overall DNA recovery compared to magnetic bead-based methods [37]. Other factors, such as incubation method, time, and temperature, also impact percentage shearing and recovery [38].

To monitor DNA shearing and assay recovery and to ensure assay reproducibility, it is therefore necessary to implement quality control protocols. Two methods will be discussed below: the first is using Agilent Bioanalyzer systems or similar technologies. For example, the Agilent 2200 TapeStation is a chip-based capillary electrophoresis system that will analyze the DNA fragment-size distribution in a given sample [39]. Nucleic acid extractions prepared for FLIP-seq and MIP-seq processing could be analyzed via similar platforms to ensure the presence of fragments around 10,000 base pairs to ensure maximal recovery of genome-intact proviral genomes.

The second method is complementary and involves the use of a positive control with assay-specific primers. The positive control can be any known HIV-DNA material that has known clonal full-length viral genomes. One example is a cell line called 8E5/LAV (NIH

AIDS Reagent Program Catalog #95 [40]), which has roughly a single copy of integrated full-length HIV genome per cell. After nucleic acid extraction, the sample is split into two aliquots: one is subjected to limiting dilution and SGA short-amplicon HIV-specific PCR amplification (e.g., *pol*), whereas the other aliquot will be subjected to the same limiting dilution factor identical to the short-amplicon reactions but will be amplified for near-full-length viral genomes in the case of FLIP-seq or subjected to the five amplicon PCR approach in the case of MIP-seq. The ratio between the recovery in the full-genome amplification approaches relative to the short amplification approach would reveal the assay sensitivity against the shorter amplification region. Note, this quality control method measures comprehensive assay sensitivity that includes both DNA shearing and PCR DNA polymerase efficiency, which will be discussed in Section 4 below.

Finally, in light of variabilities in recovery and extent of template shearing depending on extraction methods, it is important to restrict any FLIP-seq- and/or MIP-seq-based quantitative comparisons across samples and/or cohorts to samples that were processed using identical DNA-extraction methods. It is also important to note that FLIP-seq and MIP-seq are theoretically only semi-quantitative at best, a concept which will be further explored in Section 4.

4. Poisson Distribution and Limiting Dilution

Both FLIP-seq and MIP-seq involve limiting dilutions of the nucleic acid extract to achieve single-genome amplification (SGA) by PCR. There are at least three main reasons why SGA should be strictly enforced: the first and perhaps the most important reason is PCR efficiency [41,42]. If a short, truncated, and defective HIV-DNA genome is present in the same PCR reaction well together with a long, intact HIV genome template, amplification efficiency will be higher for the short relative to the longer genome, resulting in a bias of short genome detection. The second reason is to reduce the likelihood of inter-template recombination, which is a well-described PCR phenomenon [43–45]. The third reason is to resolve viral quasi-species. HIV is genetically diverse due to an error-prone reverse transcriptase, which introduces approximately one error into the viral genome at every viral RNA to DNA conversion step [46]. These mutations accumulate over the course of active viral replication and create a genetically diverse within-host viral quasi-species population that allows for Darwinian selection for drug resistance [47] and/or immune escape [48] variants. Given that every HIV-DNA template is potentially genetically different (with the exception of clonally expanded proviral populations), SGA ensures that even single-base differences would be clearly resolved. Note, resolution also depends on PCR fidelity (further discussed in Section 4).

The rule of thumb in setting up limiting dilutions for both FLIP-seq and MIP-seq is to achieve one PCR-positive reaction in every three reactions, or a “1 in 3” setup, or an SGA ratio of 0.3, to yield a Poisson probability of 85.7% that a given PCR-positive well has originated from a single HIV-DNA molecule. Figure 2 illustrates the theoretical relationship between varying SGA ratios and the probability of single-molecule amplification. Referring to Figure 2, shifting the SGA ratio to a “1 in 2” setup would result in a 77.1%, whereas a “1 in 1” setup would result in a 58.2% Poisson probability of having one template of origin per positive PCR reaction. In contrast, a “1 in 100” setup would result in a 99.5% probability of single-genome amplification. Given these probability values, a researcher setting up SGA reactions should strike a balance between reagent cost and data quality, as increasing the number of amplicon-negative wells dramatically increases PCR reagent costs. The SGA ratio of 0.3 or a Poisson probability of 85.7% is an arbitrary value generally accepted by the research community [49].

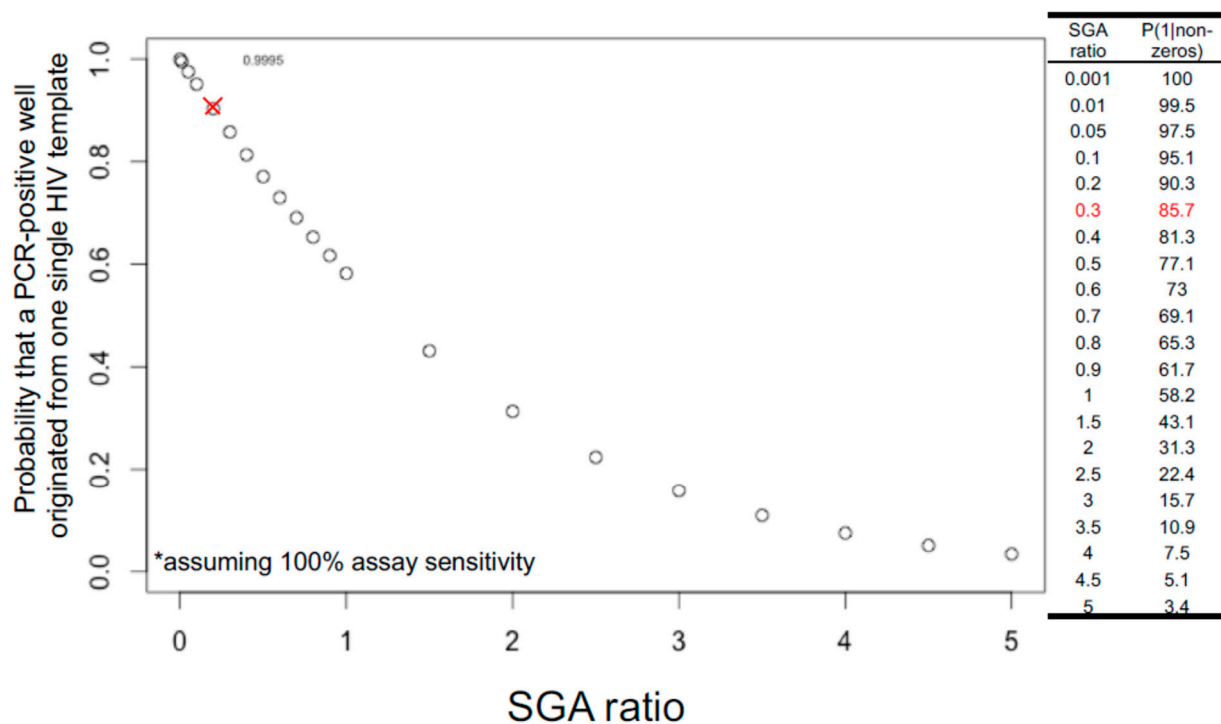


Figure 2. Probability that a PCR-positive reaction originated from one single HIV template follows Poisson probability distribution and decreases as SGA ratio increases. SGA ratio (x-axis) is defined as the number of expected positive reactions divided by the total reactions. For example, if ddPCR short amplicon estimation shows that there are 3 HIV-DNA copies per microliter within a nucleic acid extract, to achieve a limiting dilution at 1:3 (0.3) SGA ratio, one microliter of this extract would be distributed into 9 PCR reactions equally, yielding a probability of 85.7% (y-axis) that a given positive well is derived from one single HIV template (red cross in graph and red highlight in table). This calculation assumes 100% assay sensitivity in the amplification and detection of all input templates (asterisk).

Since a “1 in 3” setup only yields 85.7% probability of single-template amplification, there is a 14.3% probability that a given PCR-positive reaction under this setup could have originated from multiple HIV DNA molecules. Deep sequencing of each PCR-positive reaction allows for post-hoc bioinformatic evaluation of whether there are multiple HIV-DNA species present in the reaction: Applicable to both FLIP-seq and MIP-seq, the presence of base-pair mixtures per genome position at frequencies above the expected sequencing error rate serves as an indicator of the presence of multiple DNA templates. In the case of MIP-seq, presence of multiple HIV integration sites by deep sequencing also marks the potential presence of multiple input templates. Depending on the research question, these multiple-template PCR positive reactions could be removed from the final data analyses to achieve maximal data quality.

Note also that the distribution shown in Figure 2 assumes 100% assay sensitivity; in other words, the probability of 85.7% for single-template amplification is achieved only if every single template input, both long and short, was successfully amplified at 100% PCR efficiency. As discussed above, PCR efficiency varies according to template lengths. Therefore, even PCR-negative wells could have contained a HIV template that was not successfully amplified. This implies that the traditional dilution-factor calculation approach by visually counting the number of PCR-positive reactions by gel electrophoresis after near-full viral genome amplification and then selecting a dilution factor that yields 1 in 3 visually detectable amplicons could possibly lead to under dilution, mainly due to the lower PCR efficiency against longer input HIV-DNA templates. A potential solution to this issue is to calculate the dilution factor for limiting dilution using the concentration of total HIV-DNA genomes derived from a short target region amplification (e.g., via ddPCR

amplification of a short, conserved region in the HIV genome [7]) to achieve a higher PCR efficiency relative to full-genome long template amplification.

Given that each HIV-infected study participant has a distinct profile of HIV-DNA genome lengths [7,8,20–22,50] and given that PCR efficiency is not identical across varying template lengths and that SGA ratios are at best an estimate, plus the fact that PCR-DNA polymerase activity decreases over storage time, both FLIP-seq and MIP-seq should be considered only semi-quantitative with a bias towards detection of shorter viral genomes. In addition, due to the low frequency of productively infected CD4⁺ cells in long-term, antiretroviral-treated, HIV-infected donor samples, typically estimated to be approximately one per million CD4⁺ [5,51], when compounded with imperfect PCR efficiencies, approximately 10 million CD4⁺ cells are typically required in order to detect at least one intact genome per donor. Therefore, despite their high resolution, sample availability can be a major challenge when using FLIP-seq and MIP-seq for genome-intact virus quantification.

5. PCR Fidelity and Sequencing Errors

DNA polymerases, such as Taq, used in PCR reactions and sequencing library preparations can introduce errors in amplification products [52]. PCR fidelity refers to the accuracy of bases incorporated [42]. A high-fidelity DNA polymerase results in a low error amplification profile. Errors can also be introduced by the sequencing process itself: for example, Illumina sequencing is reported to have a baseline sequencing error rate of approximately 0.2% [53]. As discussed in Section 3, each HIV-DNA genome can potentially harbor at least one single-base nucleotide mutation due to the error prone viral reverse transcriptase [46]. Based on this observation, identical HIV-DNA genome sequences obtained from SGA reactions and FLIP-seq are often used as markers for the clonal expansion of infected cells [7,8,50]. The validity of using near-full-genome FLIP-seq sequence-identity to mark clonal expansion has been further supported by later observations from MIP-seq, showing that 100% identical viral sequences also have identical viral-host integration junctions [32]. This implies that PCR and sequencing errors should be strictly monitored for any viral-sequence-based clonal expansion analyses that are not supported by viral integration site data. In addition, PCR and/or sequencing errors can also introduce artificial stop codons into a proviral genome, leading to false classification of genome defectiveness. As such, it is important to optimize both FLIP-seq and MIP-seq to yield the most accurate viral genome sequence data possible.

The first optimization step is to select a DNA polymerase with high fidelity for viral genome amplification: FLIP-seq, as published in [7], uses a third-generation Invitrogen Taq polymerase (catalog number 11304102) at 6X fidelity relative to unmodified regular Platinum Taq [54]. Since no PCR amplification is completely error-free, but errors introduced are relatively random in terms of kinds (base substitutions, deletions, and insertions) and locations [55], it is possible to bioinformatically correct for errors given a deep enough sequencing depth via the generation of consensus sequences (Figure 3). The median sequencing depth (a.k.a. coverage) across the HIV genome for previously published FLIP-seq [56] and MIP-seq [32] data was at approximately 2000 Illumina small reads (150 bp) per base position.

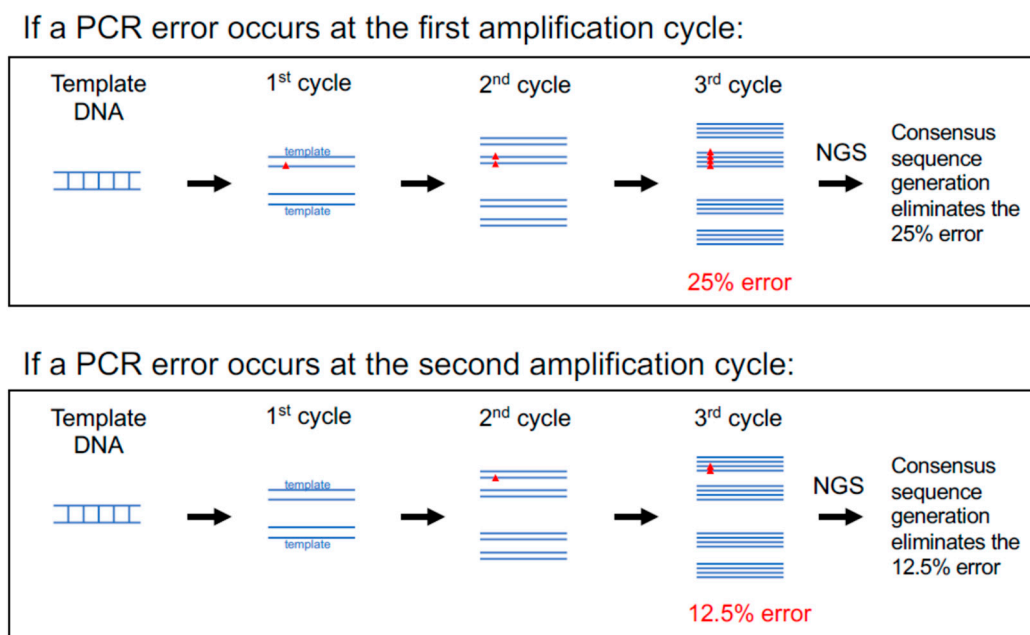


Figure 3. PCR errors are corrected via consensus sequence generation. Errors introduced during the PCR amplification step (red triangles) mainly involve single-base substitution errors at random locations. An error that occurs at an earlier PCR amplification cycle will carry over into a larger proportion within the final amplicon pool (25% if first cycle) relative to an error that occurs at a later amplification cycle (12.5% if second cycle). This figure illustrates that per error introduced, the maximum frequency of representation is 25% within the final amplicon pool, which can be corrected via consensus sequence generation of the deep sequence data.

To bioinformatically measure and/or correct for errors, a consensus viral genome sequence is first generated from deep sequencing reads derived from an SGA reaction, then the distribution and prevalence of non-consensus base pairs across the viral genome is calculated. Note that under the same principal, this consensus-based correction method not only corrects for PCR errors but also serves to correct errors introduced by MDA (in the case of MIP-seq), various sequencing library preparation protocols that are PCR-dependent, as well as the errors introduced during the process of sequencing itself. This error-detection step is an integrated part of the bioinformatics pipeline HIVSeqinR [20] developed for HIV proviral genome-intactness inferences. Another cross-validation for undetectable PCR/sequencing error post-bioinformatics correction, as mentioned above, is through the identification of 100% genetically identical viral DNA genomes in addition to identical MIP-seq-derived integration site coordinates [32].

Note that the above discussion applies only to viral genome sequencing. PCR fidelity is less critical in integration site sequencing and mapping. This is because the viral-host junction sequence data are only used for mapping to the human reference genome for the identification of integration site as opposed to quasi-species differentiation. In one of the MIP-seq algorithms as published in [32], during this step, query fragment lengths in blocks of 20 nucleotides [57] would be evaluated by the mapping algorithm, making the results less susceptible to single-base substitutions, insertion, and deletion errors associated with PCR enzyme fidelity and sequencing errors.

6. Bioinformatics Considerations for Genome-Intactness Inferences

Viral genomes captured by FLIP-seq and MIP-seq are typically subjected to bioinformatics evaluation for genome-intactness. The term “genome intactness” is loosely defined as the lack of any decapacitating mutations that would render a viral genome non-replication competent. However, the exact definitions/criteria vary between research groups and publications [7,8,20,21,27,33,50]. A few common categories of “genome defects” will be discussed below; criteria used in the automated genome-intactness call-

ing computational pipeline HIVSeqinR [20] will be given as examples. A stable release (version 2.7.1 as of date of manuscript preparation) is available in GitHub at <https://github.com/guineverelee/HIVSeqinR> (accessed on 7 September 2021). Another software for HIV-DNA genome-intact inferences, HIVIntact [58], is also publicly available for download and differs from HIVSeqinR in terms of logical order for intact determination as well as specific bioinformatic definitions of “genome defects”. Regardless of the software used, a bioinformatics inference strategy should aim to optimize sensitivity and specificity for the purpose of a specific research question. For example, the software HIVSeqinR [20] was designed to maximize specificity in genome intactness, calling to predict replication competency. A classification algorithm should also be reproducible; in other words, ideally the same inference strategy should be applied to all viral genome sequences in a dataset.

6.1. Large Deletions

A viral genome may be heavily truncated, rendering it non-replication competent. These are genomes that contain “large deletion(s)”. However, “large” is a relative term. In addition, deletion(s) and/or truncation(s) that occurs within an essential gene may directly impacts replication competency. In HIVSeqinR, any near-full-length HIV amplicons less than 8000 bp are automatically categorized as having “large deletions”. In other words, assuming an amplicon spanning HXB2 coordinates 638–9632 [7], any genomes with deletion(s) more than approximately 995 bp relatively to the 8995 bp expected length will be classified as a genome with “large deletion(s)” regardless of the location and frequencies of truncation. Note that this strategy is designed to achieve automation, reproducibility, and to maximize specificity against the detection of a replication-competent virus when used in combination with the other defectiveness categories.

6.2. Internal Inversions

A portion of the viral DNA genome may contain an inversion, rendering it non-replication competent. In HIVSeqinR, inversions are detected by mapping query sequences at an initial block/window size of 11 bp [59]. Adjusting this length can impact the sensitivity of internal inversion detection.

6.3. Hypermutation

Guanosine to adenosine (G-to-A) hypermutations are introduced into viral genomes during the reverse transcription step by a family of host-defense proteins called APOBEC, leading to the occurrences of premature stop codons throughout the genome [60]. A web tool called Hypermut [61] is available in the Los Alamos HIV Sequence Database website to screen whether a given query genome contains APOBEC-associated footprints. This algorithm is reference-sequence dependent: briefly, it counts the occurrences of where Gs are expected based upon the reference genome that the user uploaded. For the most accurate prediction, a donor-matched reference sequence that has been shown to be replication competent experimentally should be used, but this sequence is often not available. In HIVSeqinR’s adaptation of Hypermut [20], HXB2 is used as the universal reference sequence to provide a baseline screen for genomes that have obvious APOBEC footprints; all other genomes that contain a large amount of premature stop codons would be identified as having “premature stop codons” at a later stage in the HIVSeqinR algorithm and will not be classified as intact. In other words, HIVSeqinR compromises on sensitivity for true APOBEC-associated hypermutated genomes in return for automation with a focus on maximizing overall specificity for genome-intact inferences. If the purpose of one’s research is, for example, not to identify intact genomes but to study the impact of APOBEC protein family on HIV DNA reservoirs, then it becomes important to fine-tune this hypermutation inference process as an independent, non-automated step using the most appropriate reference genome available.

6.4. Premature Stop Codons

A viral genome may contain single-base substitution mutation(s) and/or out-of-frame insertion/deletion(s), rendering the genome non-replication competent. Three main considerations should be given when evaluating a specific genome for this category. First, location of a given premature stop codon matters: the HIV genome codes for nine genes (*gag*, *pol*, *vif*, *vpr*, *vpu*, *tat*, *rev*, *env*, and *nef*), while only *gag*, *pol*, and *env* are traditionally considered essential genes [34]. There are known examples of HIV genomes with premature stop codons in *tat* and *nef* that are able to establish infections both *in vitro* (for example, *tat* [62] and Table 1; *nef* [63,64]) and *in vivo* (for example, *tat* [65]; *nef* [66,67]) despite reduced function/replication capacity [63,65,66]. In HIVSeqinR [20], a viral genome is labelled to contain “premature stop codon(s)” only if the stop codon occurs in any one of the essential genes *gag*, *pol*, and/or *env*. Second, “premature” is a relative term: for instance, a premature stop codon that results in the loss of 50% of the expected amino acid length will have a more decapacitating effect relative to a stop codon that results in the loss of 5% amino acid length. In HIVSeqinR [20], a genome will be labelled to contain “premature stop codon(s)” if the stop codon results in an amino acid length of less than 95% relative to HXB2/JR-CSF/NL4-3 in any of the essential genes *gag*, *pol*, and/or *env* (Table 1, expected values). This 95% cutoff value maximizes specificity for genome-intactness inferences. However, these definitions are not absolute and should not be considered 100% predictive of replication competency and should be adapted and evaluated for each scientific question being asked. Finally, it is important to ensure stop codons have not been introduced due to PCR and/or sequencing errors. It is therefore important to perform quality control measures as outlined in Section 4: SGA and sequencing of a clonal population should lead to identical consensus sequences.

Table 1. Amino acid lengths of all HIV gene products and the lengths of the non-coding packaging signal in five commonly used lab/reference strains are summarized below.

Strains	By Lengths					(HIVSeqinR Expected Value Settings)	By Percentages Relative to Expected Values				
	ACH-2	8E5/LAV	HXB2	JR-CSF	NL4-3		ACH-2	8E5/LAV	HXB2	JR-CSF	NL4-3
NIH HIV Reagent Program ID	ARP-349 **	ARP-95 **	NA ***	ARP-394	ARP-114 **		ARP-349	ARP-95	NA	ARP-394	ARP-114
Replication competence	Yes	No	Weak	Yes	Yes		Yes	No	Weak	Yes	Yes
Non-coding (unit, nucleotide length)											
Psi length, HXB2 681-789	112	112	112	111	112	112	100%	100%	100%	99%	100%
Coding (unit, amino acid length)											
Gag	500	500	500	504	500	500	100%	100%	100%	101%	100%
Protease	99	99	99	99	99	99	100%	100%	100%	100%	100%
Reverse transcriptase	440	267	440	440	440	440	100%	61%	100%	100%	100%
RNaseH	120	NA	120	120	120	120	100%	NA	100%	100%	100%
Integrase	288	NA	288	288	288	288	100%	NA	100%	100%	100%
Vif	192	192	192	192	192	192	100%	100%	100%	100%	100%
Vpr	96	37	78	96	96	96 *	100%	39%	81%	100%	100%
Vpu	22	22	82	81	82	82	27%	27%	100%	99%	100%
Env	861	859	856	849	854	856	101%	100%	100%	99%	100%
GP120	486	484	481	474	479	481	101%	101%	100%	99%	100%
GP41	345	345	345	345	345	345	100%	100%	100%	100%	100%
Tat	86	86	86	101	86	101 *	85%	85%	85%	100%	85%
Rev	116	100	116	116	116	116	100%	86%	100%	100%	100%
Nef	206	206	123	216	206	206 *	100%	100%	60%	105%	100%
HIVSeqinR verdict	Intact	PrematureStop	Intact	Intact	Intact		Intact	PrematureStop	Intact	Intact	Intact

* These expected values are based on manually removing mutations associated with defects in HXB2; ** LAV was the parent HIV strain for all of ACH-2, 8E5/LAV, and the 3' end of NL4-3; *** GenBank Accession Number for HXB2 is K03455. Red fonts indicate strain-specific values that are <95% of the expected value settings in HIVSeqinR.

6.5. 5' or Psi (ψ) Defects

The 5' beginning of the HIV genome contains a packaging signal also called ψ (HXB2 coordinates 681–789) [34]. This region has been shown to be essential for viral genome dimerization, nucleocapsid (NC) protein binding, and subsequent viral RNA packaging into viral particles [68]. ψ is non-coding, consists of four stem loops (SL1-4), and depends on the RNA 3D secondary structure to achieve its functions [69–71]. There are currently no algorithms available to accurately predict the RNA 3D structure of a given ψ DNA sequence and to distinguish between functional versus defective ψ . For this reason, HIVSeqinR, for example, imposes a loose definition for ψ defects: given that NL4-3 is replication competent [72] and thus has a functional ψ , 5' defect in HIVSeqinR has been defined as any viral genomes with a ≥ 15 bp insertion and/or deletion in that region relative to NL4-3 ψ , which is identical to HXB2 ψ , which are both 112 base pairs in length. Again, this definition aims to achieve maximal specificities for genome-intactness predictions based on our knowledge of a replication competent viral strain.

6.6. One Verdict per Genome

First, it is important to note that this above list of potential defect-genome categories is not exhaustive: other definitions can also be considered, such as the presence/absence of splice donor 1 (D1) site [50]. Second, it is possible that one genome contains multiple classes of defects: for example, it is not uncommon to observe genomes with large deletions that are also hypermutated [7]. In HIVSeqinR, for reproducibility and downstream statistical purposes, after obtaining a TRUE/FALSE classification of each of the above defective categories described, each viral genome is then given a single verdict in the order of large deletions, internal inversions, hypermutations, premature stop codons, and 5' or psi (ψ) defects. Any genomes without any of the above-mentioned defects would be classified as “genome-intact” by HIVSeqinR. Multiple verdict calling is supported by HIVSeqinR by reviewing the raw per-category TRUE/FALSE output. Note that since the purpose of the HIVSeqinR software was to identify genome-intact proviruses, which is a category derived by elimination, therefore, by definition, it is the only classification category that does not support multiple verdicts.

6.7. Functional Validation

Any bioinformatics-inference algorithms offer only predictions and should be functionally validated. In the case of proviral genome intactness, the corresponding functional data can be one or both of (i) SGA sequence data of full-genome plasma virus assuming that plasma derived sequences are replication competent and/or (ii) SGA sequence data from assays that measure replication competence, such as qVOA. For example, HIVSeqinR was functionally validated to be 100% sensitive in predicting genome-intactness, qVOA-derived outgrowth viral sequences [7]. Finally, it is important to understand that replication competence is a spectrum: mutations in different parts of the viral genome may increase/decrease the replication fitness of the virus to different degrees.

In summary, this section highlights that the term “genome intactness” is a strictly bioinformatic definition for the lack of specific defects in a given HIV DNA genome. Researchers should adapt a definition of genome intactness that best suits their specific research question.

7. Conclusions

Both FLIP-seq and MIP-seq are deep sequencing assays designed to distinguish between intact versus defective HIV proviral DNA genomes. FLIP-seq and similar technologies have been applied to cross-sectionally examine viral reservoir landscapes in various CD4⁺ T-cell subsets [7,8] and to longitudinally examine the evolution of the viral DNA genome populations over time [20,50]. MIP-seq has been applied to compare viral integration sites of intact versus defective genomes [32], reveal unique patterns of genome-intact viral integration sites in HIV elite controllers [24], and has been further developed

by another group of researchers to include co-capturing of T-cell receptor sequences for antigen specificity inferences of the infected cells [73]. Application of these sequencing technologies to various cohorts have resulted in a rich collection of HIV-DNA genome sequences archived in public repositories, such as the HIV Proviral Sequence Databases [73] and the Los Alamos HIV Sequence Database [74], which are used in part to guide the design of relatively low-cost ddPCR-based assays, such as IPDA [27] and a multiplex assay by Levy et al. [29] for the quantification of intact versus defective HIV-DNA genomes. In summary, this commentary highlights that deep sequencing like FLIP-seq and MIP-seq offers advantages, such as high-resolution data quality enabling post-hoc quality control for true single-genome amplification; but in order to take full advantage of these technologies, one has to be mindful to take necessary quality control steps to monitor data quality. The list of chemistry and bioinformatics considerations discussed in this commentary is by no means exhaustive and should be re-evaluated with a given scientific question a researcher sets out to address.

Funding: GQL is supported by NIH grants NIAID UM1 AI164565, R21 AI150398, and R01 AI162221.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The author thanks her previous and current academic mentors: Richard Harri-gan at the University of British Columbia, Mathias Lichterfeld at the Brigham and Women's Hospital and the Ragon Institute of MGH, MIT and Harvard (where FLIP-seq and MIP-seq were developed), and Brad Jones at Weill Cornell Medicine. The author also thanks Massachusetts General Hospital Center for Computational & Integrative Biology DNA Core, specifically Nicole Stange-Thomann, Amy Avery, Kristina Belanger, and Huajun Wang for their continuous input in FLIP-seq and MIP-seq sequence quality assurances. The author thanks Pragya Khadka and Kevin Bernard for their help in proofreading this manuscript.

Conflicts of Interest: GQL receives funding from Merck Co for a non-HIV-related project. The author declares that other conflicts of interest do not exist.

References

1. Flint, S.J.; Enquist, L.W.; Racaniello, V.R.; Skalka, A.M. *Principles of Virology*, 2nd ed.; ASM Press: Washington, DC, USA, 2004.
2. Finzi, D.; Hermankova, M.; Pierson, T.; Carruth, L.M.; Buck, C.; Chaisson, R.E.; Quinn, T.C.; Chadwick, K.; Margolick, J.; Brookmeyer, R.; et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **1997**, *278*, 1295–1300. [[CrossRef](#)]
3. Wong, J.K.; Hezareh, M.; Günthard, H.F.; Havlir, D.V.; Ignacio, C.C.; Spina, C.A.; Richman, D.D. Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **1997**, *278*, 1291–1295. [[CrossRef](#)]
4. Chun, T.-W.W.; Stuyver, L.; Mizell, S.B.; Ehler, L.A.; Mican, J.A.M.; Baseler, M.; Lloyd, A.L.; Nowak, M.A.; Fauci, A.S. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 13193–13197. [[CrossRef](#)]
5. Siliciano, J.D.; Kajdas, J.; Finzi, D.; Quinn, T.C.; Chadwick, K.; Margolick, J.B.; Kovacs, C.; Gange, S.J.; Siliciano, R.F. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nat. Med.* **2003**, *9*, 727–728. [[CrossRef](#)]
6. Walensky, R.P.; Paltiel, A.D.; Losina, E.; Mercincavage, L.M.; Schackman, B.R.; Sax, P.E.; Weinstein, M.C.; Freedberg, K.A. The survival benefits of AIDS treatment in the United States. *J. Infect. Dis.* **2006**, *194*, 11–19. [[CrossRef](#)]
7. Lee, G.Q.; Orlova-Fink, N.; Einkauf, K.; Chowdhury, F.Z.; Sun, X.; Harrington, S.; Kuo, H.-H.; Hua, S.; Chen, H.-R.; Ouyang, Z.; et al. Clonal expansion of genome-intact HIV-1 in functionally-polarized Th1 CD4 T cells. *J. Clin. Investig.* **2017**, *127*, 2689–2696. [[CrossRef](#)]
8. Hiener, B.; Horsburgh, B.A.; Eden, J.-S.S.; Barton, K.; Schlub, T.E.; Lee, E.; von Stockenstrom, S.; Odeval, L.; Milush, J.M.; Liegler, T.; et al. Identification of Genetically Intact HIV-1 Proviruses in Specific CD4+T Cells from Effectively Treated Participants. *Cell Rep.* **2017**, *21*, 813–822. [[CrossRef](#)] [[PubMed](#)]
9. Simonetti, F.R.; Sobolewski, M.D.; Fyne, E.; Shao, W.; Spindler, J.; Hattori, J.; Anderson, E.M.; Watters, S.A.; Hill, S.; Wu, X.; et al. Clonally expanded CD4⁺ T cells can produce infectious HIV-1 in vivo. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 201522675. [[CrossRef](#)] [[PubMed](#)]

10. Maldarelli, F.; Wu, X.; Su, L.; Simonetti, F.R.; Shao, W.; Hill, S.; Spindler, J.; Ferris, A.L.; Mellors, J.W.; Kearney, M.F.; et al. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **2014**, *345*, 179–183. [[CrossRef](#)] [[PubMed](#)]
11. Bui, J.K.; Sobolewski, M.D.; Keele, B.F.; Spindler, J.; Musick, A.; Wiegand, A.; Luke, B.T.; Shao, W.; Hughes, S.H.; Coffin, J.M.; et al. Proviruses with identical sequences comprise a large fraction of the replication-competent HIV reservoir. *PLoS Pathog.* **2017**, *13*, e1006283. [[CrossRef](#)]
12. Bui, J.K.; Halvas, E.K.; Fyne, E.; Sobolewski, M.D.; Koontz, D.; Shao, W.; Luke, B.; Hong, F.F.; Kearney, M.F.; Mellors, J.W. Ex vivo activation of CD4+T-cells from donors on suppressive ART can lead to sustained production of infectious HIV-1 from a subset of infected cells. *PLoS Pathog.* **2017**, *13*, 1–20. [[CrossRef](#)]
13. Hosmane, N.N.; Kwon, K.J.; Bruner, K.M.; Capoferri, A.A.; Beg, S.; Rosenbloom, D.I.S.S.; Keele, B.F.; Ho, Y.-C.C.; Siliciano, J.D.; Siliciano, R.F. Proliferation of latently infected CD4 + T cells carrying replication-competent HIV-1: Potential role in latent reservoir dynamics. *J. Exp. Med.* **2017**, *214*, 1–14. [[CrossRef](#)]
14. Cohn, L.B.; Silva, I.T.; Oliveira, T.Y.; Rosales, R.A.; Parrish, E.H.; Learn, G.H.; Hahn, B.H.; Czartoski, J.L.; McElrath, M.J.; Lehmann, C.; et al. HIV-1 integration landscape during latent and active infection. *Cell* **2015**, *160*, 420–432. [[CrossRef](#)]
15. Wang, Z.; Gurule, E.E.; Brennan, T.P.; Gerold, J.M.; Kwon, K.J.; Hosmane, N.N.; Kumar, M.R.; Beg, S.A.; Capoferri, A.A.; Ray, S.C.; et al. Expanded cellular clones carrying replication-competent HIV-1 persist, wax, and wane. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E2575–E2584. [[CrossRef](#)] [[PubMed](#)]
16. Gantner, P.; Pagliuzza, A.; Pardons, M.; Ramgopal, M.; Routy, J.-P.; Fromentin, R.; Chomont, N. Single-cell TCR sequencing reveals phenotypically diverse clonally expanded cells harboring inducible HIV proviruses during ART. *Nat. Commun.* **2020**, *11*, 4089. [[CrossRef](#)] [[PubMed](#)]
17. Chomont, N.; El-Far, M.; Ancuta, P.; Trautmann, L.; Procopio, F.A.; Yassine-Diab, B.; Boucher, G.; Boulassel, M.R.; Ghattas, G.; Brenchley, J.M.; et al. HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nat. Med.* **2009**, *15*, 893–900. [[CrossRef](#)] [[PubMed](#)]
18. Wagner, T.A.; McLaughlin, S.; Garg, K.; Cheung, C.Y.K.; Larsen, B.B.; Styrchak, S.; Huang, H.C.; Edlefsen, P.T.; Mullins, J.I.; Frenkel, L.M. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **2014**, *345*, 570–573. [[CrossRef](#)]
19. Shaw, G.M.; Hunter, E. HIV Transmission. *Cold Spring Harb. Perspect. Med.* **2012**, *2*, a006965. [[CrossRef](#)]
20. Lee, G.Q.; Reddy, K.; Einkauf, K.B.; Gounder, K.; Chevalier, J.M.; Dong, K.L.; Walker, B.D.; Yu, X.G.; Ndung'u, T.; Lichterfeld, M. HIV-1 DNA Sequence Diversity and Evolution during Acute Subtype C Infection. *Nat. Commun.* **2019**, *10*, 2737. [[CrossRef](#)]
21. Ho, Y.-C.; Shan, L.; Hosmane, N.N.; Wang, J.; Laskey, S.B.; Rosenbloom, D.I.S.; Lai, J.; Blankson, J.N.; Siliciano, J.D.; Siliciano, R.F. Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* **2013**, *155*, 540–551. [[CrossRef](#)]
22. Bruner, K.M.; Murray, A.J.; Pollack, R.A.; Soliman, M.G.; Laskey, S.B.; Capoferri, A.A.; Lai, J.; Strain, M.C.; Lada, S.M.; Hoh, R.; et al. Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nat. Med.* **2016**. [[CrossRef](#)]
23. Einkauf, K.B.; Osborn, M.; Gao, C.; Parsons, E.; Jiang, C.; Lian, X.; Sun, X.; Blackmer, J.E.; Rosenberg, E.S.; Yu, X.; et al. Evolutionary Dynamics of HIV Reservoir Cells via a Novel Single-cell Multiomics Assay. In Proceedings of the Conference on Retroviruses and Opportunistic Infections (CROI 2021), Virtual, 6–10 March 2021.
24. Jiang, C.; Lian, X.; Gao, C.; Sun, X.; Einkauf, K.B.; Chevalier, J.M.; Chen, S.M.Y.; Hua, S.; Rhee, B.; Chang, K.; et al. Distinct viral reservoirs in individuals with spontaneous control of HIV-1. *Nature* **2020**, 1–7. [[CrossRef](#)]
25. Kuo, H.H.; Lichterfeld, M. Recent progress in understanding HIV reservoirs. *Curr. Opin. HIV AIDS* **2018**, *13*, 137–142. [[CrossRef](#)] [[PubMed](#)]
26. Sharaf, R.R.; Li, J.Z. The Alphabet Soup of HIV Reservoir Markers. *Curr. HIV/AIDS Rep.* **2017**. [[CrossRef](#)] [[PubMed](#)]
27. Bruner, K.M.; Wang, Z.; Simonetti, F.R.; Bender, A.M.; Kwon, K.J.; Sengupta, S.; Fray, E.J.; Beg, S.A.; Antar, A.A.R.R.; Jenike, K.M.; et al. A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. *Nature* **2019**, *566*, 120–125. [[CrossRef](#)]
28. C, G.; JCC, L.; TY, O.; L, N.; V, R.; CL, L.; JA, P.; P, M.; M, J.; M, C.; et al. Combination of quadruplex qPCR and next-generation sequencing for qualitative and quantitative analysis of the HIV-1 latent reservoir. *J. Exp. Med.* **2019**, *216*, 2253–2264. [[CrossRef](#)]
29. CN, L.; SM, H.; P, R.; DB, R.; C, A.; H, Z.; ML, H.; Y, W.; ME, B.; NAJ, C.; et al. A highly multiplexed droplet digital PCR assay to measure the intact HIV-1 proviral reservoir. *Cell Rep. Med.* **2021**, *2*. [[CrossRef](#)]
30. Lee, G.Q.; Khadka, P.; Jones, R.B.; Kasule, J.; Kityamuweesi, T.; Buule, P.; Laeyendecker, O.; Reynolds, S.; Quinn, T.; Prodger, J.; et al. Subtype D HIV-1 reservoir levels and viral sequence profiles in Rakai, Uganda. In Proceedings of the International AIDS Society Conference on HIV Science, Virtual, 18–21 July 2021.
31. Bushman, F.D.; Hoffmann, C.; Ronen, K.; Malani, N.; Minkah, N.; Rose, H.M.; Tebas, P.; Wang, G.P. Massively parallel pyrosequencing in HIV research. *AIDS* **2008**, *22*, 1411–1415. [[CrossRef](#)]
32. Einkauf, K.B.K.B.; Lee, G.Q.G.Q.; Gao, C.; Sharaf, R.; Sun, X.; Hua, S.; Chen, S.M.Y.S.M.Y.; Jiang, C.; Lian, X.; Chowdhury, F.Z.F.Z.; et al. Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. *J. Clin. Invest.* **2019**, *129*, 988–998. [[CrossRef](#)]
33. Patro, S.C.; Brandt, L.D.; Bale, M.J.; Halvas, E.K.; Joseph, K.W.; Shao, W.; Wu, X.; Guo, S.; Murrell, B.; Wiegand, A.; et al. Combined HIV-1 sequence and integration site analysis informs viral dynamics and allows reconstruction of replicating viral ancestors. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 25891–25899. [[CrossRef](#)]

34. Foley, B.; Leitner, T.; Apetrei, C.; Hahn, B.; Mizrachi, I.; Mullins, J.; Rambaut, A.; Wolinsky, S.; Korber, B.; Abfalterer, W.; et al. (Eds.) *HIV Sequence Compendium 2018*; Theoretical Biology and Biophysics: Los Alamos, NM, USA, 2018.
35. Ali, N.; Rampazzo, R.D.C.P.; Costa, A.D.T.; Krieger, M.A. Current Nucleic Acid Extraction Methods and Their Implications to Point-of-Care Diagnostics. *Biomed Res. Int.* **2017**. EPUB. [CrossRef]
36. Yuan, S.; Cohen, D.B.; Ravel, J.; Abdo, Z.; Forney, L.J. Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome. *PLoS ONE* **2012**, *7*, e33865. [CrossRef]
37. Bronkhorst, A.J.; Ungerer, V.; Holdenrieder, S. Comparison of methods for the isolation of cell-free DNA from cell culture supernatant. *Tumor Biol.* **2020**, *42*, 101042832091631. [CrossRef] [PubMed]
38. Adamowicz, M.S.; Stasulli, D.M.; Sobestanovich, E.M.; Bille, T.W. Evaluation of Methods to Improve the Extraction and Recovery of DNA from Cotton Swabs for Forensic Analysis. *PLoS ONE* **2014**, *9*, e116351. [CrossRef] [PubMed]
39. Padmanaban, A. End to End Sample Quality Control for Next Generation Sequencing Library Preparation and SureSelect Target Enrichment on the Agilent 2200 TapeStation System. Available online: <https://www.agilent.com/cs/library/applications/5991-3654EN.pdf> (accessed on 18 April 2021).
40. AIDS Reagent Program: 8E5/LAV (Catalog #95). Available online: <https://www.hivreagentprogram.org/Catalog/HRPCellLines/ARP-95.aspx> (accessed on 7 September 2021).
41. Booth, C.S.; Pienaar, E.; Termaat, J.R.; Whitney, S.E.; Louw, T.M.; Viljoen, H.J. Efficiency of the polymerase chain reaction. *Chem. Eng. Sci.* **2010**, *65*, 4996–5006. [CrossRef] [PubMed]
42. Cha, R.S.; Thilly, W.G. Specificity, Efficiency, and Fidelity of PCR. *Genome Res.* **1993**, *3*, S18–S29. [CrossRef] [PubMed]
43. Boltz, V.F.; Rausch, J.; Shao, W.; Hattori, J.; Luke, B.; Maldarelli, F.; Mellors, J.W.; Kearney, M.F.; Coffin, J.M. Ultrasensitive single-genome sequencing: Accurate, targeted, next generation sequencing of hiv-1 rna. *Retrovirology* **2016**, *13*. [CrossRef] [PubMed]
44. Shao, W.; Boltz, V.F.; Spindler, J.E.; Kearney, M.F.; Maldarelli, F.; Mellors, J.W.; Stewart, C.; Volfovsky, N.; Levitsky, A.; Stephens, R.M.; et al. Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of Low-frequency drug resistance mutations in HIV-1 DNA. *Retrovirology* **2013**, *10*, 18. [CrossRef]
45. Zanini, F.; Brodin, J.; Albert, J.; Neher, R.A. Error rates, PCR recombination, and sampling depth in HIV-1 whole genome deep sequencing. *Virus Res.* **2017**, *239*, 106–114. [CrossRef] [PubMed]
46. Griffin, S. Viral Replication. In *Human Immunodeficiency Virus*; Richman, D.D., Ed.; International Medical Press: London, UK, 2003; pp. 313–315.
47. Harrigan, P.R.; Alexander, C.S. Selection of drug-resistant HIV. *Trends Microbiol.* **1999**, *7*, 120–123. [CrossRef]
48. Carlson, J.M.; Le, A.Q.; Shahid, A.; Brumme, Z.L. HIV-1 adaptation to HLA: A window into virus–host immune interactions. *Trends Microbiol.* **2015**, *23*, 212–224. [CrossRef]
49. Butler, D.M.; Pacold, M.E.; Jordan, P.S.; Richman, D.D.; Smith, D.M. The efficiency of single genome amplification and sequencing is improved by quantitation and use of a bioinformatics tool. *J. Virol. Methods* **2009**, *162*, 280–283. [CrossRef] [PubMed]
50. Pinzone, M.R.; VanBelzen, D.J.; Weissman, S.; Bertuccio, M.P.; Cannon, L.; Venanzi-Rullo, E.; Migueles, S.; Jones, R.B.; Mota, T.; Joseph, S.B.; et al. Longitudinal HIV sequencing reveals reservoir expression leading to decay which is obscured by clonal expansion. *Nat. Commun.* **2019**, *10*, 728. [CrossRef] [PubMed]
51. D, F.; J, B.; JD, S.; JB, M.; K, C.; T, P.; K, S.; J, L.; F, L.; C, F.; et al. Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat. Med.* **1999**, *5*, 512–517. [CrossRef]
52. Potapov, V.; Ong, J.L. Examining sources of error in PCR by single-molecule sequencing. *PLoS ONE* **2017**, *12*, e0169774. [CrossRef] [PubMed]
53. Pfeiffer, F.; Gröber, C.; Blank, M.; Händler, K.; Beyer, M.; Schultze, J.L.; Mayer, G. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* **2018**, *8*, 10950. [CrossRef] [PubMed]
54. Invitrogen Platinum TM Taq DNA Polymerase High Fidelity. Available online: <https://www.thermofisher.com/order/catalog/product/11304102#/11304102> (accessed on 22 April 2021).
55. Filges, S.; Yamada, E.; Ståhlberg, A.; Godfrey, T.E. Impact of Polymerase Fidelity on Background Error Rates in Next-Generation Sequencing with Unique Molecular Identifiers/Barcodes. *Sci. Rep.* **2019**, *9*, 1–7. [CrossRef]
56. Lee, G.Q.; Orlova, N.; Serrao, E.; Sun, X.; Chowdhury, N.F.; Rosenberg, E.; Engelman, A.; Yu, X.; Lichterfeld, M. Clonal Expansion of Genome-Intact HIV-1 in Functionally Polarized T-cell Subsets. In Proceedings of the 24th Conference on Retroviruses and Opportunistic Infections (CROI 2017), Seattle, WA, USA, 13–16 February 2017; Abstract #292. Themed Discussion Session 0519.
57. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **2002**, *12*, 656–664. [CrossRef]
58. Wright, I.A.; Bale, M.J.; Shao, W.; Hu, W.-S.; Coffin, J.M.; Van Zyl, G.U.; Kearney, M.F. HIVIntact: A python-based tool for HIV-1 genome intactness inference. *Retrovirology* **2021**, *181*, 1–6. [CrossRef]
59. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinformatics* **2009**, *10*, 1–9. [CrossRef]
60. Harris, R.S.; Liddament, M.T. Retroviral restriction by APOBEC proteins. *Nat. Rev. Immunol.* **2004**, *4*, 868–877. [CrossRef]
61. Rose, P.P.; Korber, B.T. Detecting hypermutations in viral sequences with an emphasis on G-to-A hypermutation. *Bioinformatics* **2000**, *16*, 400–401. [CrossRef]

62. Ho, W.Z.; Tomassini, N.; Cherukuri, R.; Shun-D, G.; Song, L.; Lee, H.R.; Douglas, S.D. Monokine-mediated increase in human immunodeficiency virus type 1 expression in chronically infected promonocyte- and T-cell-derived lines. *Clin. Diagn. Lab. Immunol.* **1994**, *1*, 517–525. [[CrossRef](#)]
63. SU, L.; KANESHIMA, H.; BONYHADI, M.L.; LEE, R.; AUTEN, J.; WOLF, A.; DU, B.; RABIN, L.; HAHN, B.H.; TERWILLIGER, E.; et al. Identification of HIV-1 Determinants for Replication in Vivo. *Virology* **1997**, *227*, 45–52. [[CrossRef](#)] [[PubMed](#)]
64. Chowers, M.Y.; Spina, C.A.; Kwok, T.J.; Fitch, N.J.; Richman, D.D.; Guatelli, J.C. Optimal infectivity in vitro of human immunodeficiency virus type 1 requires an intact nef gene. *J. Virol.* **1994**, *68*, 2906–2914. [[CrossRef](#)]
65. Van der Kuyl, A.C.; Vink, M.; Zorgdrager, F.; Bakker, M.; Wymant, C.; Hall, M.; Gall, A.; Blanquart, F.; Berkhout, B.; Fraser, C.; et al. The evolution of subtype B HIV-1 tat in the Netherlands during 1985–2012. *Virus Res.* **2018**, *250*, 51–64. [[CrossRef](#)] [[PubMed](#)]
66. Rhodes, D.I.; Ashton, L.; Solomon, A.; Carr, A.; Cooper, D.; Kaldor, J.; Deacon, N. Characterization of Three nef-Defective Human Immunodeficiency Virus Type 1 Strains Associated with Long-Term Nonprogression. *J. Virol.* **2000**, *74*, 10581–10588. [[CrossRef](#)] [[PubMed](#)]
67. Learmont, J.C.; Geczy, A.F.; Mills, J.; Ashton, L.J.; Raynes-Greenow, C.H.; Garsia, R.J.; Dyer, W.B.; McIntyre, L.; Oelrichs, R.B.; Rhodes, D.I.; et al. Immunologic and Virologic Status after 14 to 18 Years of Infection with an Attenuated Strain of HIV-1 — A Report from the Sydney Blood Bank Cohort. *N. Engl. J. Med.* **1999**, *340*, 1715–1722. [[CrossRef](#)]
68. Heng, X.; Kharytonchyk, S.; Garcia, E.L.; Lu, K.; Divakaruni, S.S.; LaCotti, C.; Edme, K.; Telesnitsky, A.; Summers, M.F. Identification of a Minimal Region of the HIV-1 5'-Leader Required for RNA Dimerization, NC Binding, and Packaging. *J. Mol. Biol.* **2012**, *417*, 224–239. [[CrossRef](#)]
69. Ding, P.; Kharytonchyk, S.; Waller, A.; Mbaekwe, U.; Basappa, S.; Kuo, N.; Frank, H.M.; Quasney, C.; Kidane, A.; Swanson, C.; et al. Identification of the initial nucleocapsid recognition element in the HIV-1 RNA packaging signal. *Proc. Natl. Acad. Sci.* **2020**, *117*, 17737–17746. [[CrossRef](#)] [[PubMed](#)]
70. Keane, S.C.; Summers, M.F. NMR studies of the structure and function of the HIV-1 5'-leader. *Viruses* **2016**, *8*, 338. [[CrossRef](#)] [[PubMed](#)]
71. Brown, J.D.; Kharytonchyk, S.; Chaudry, I.; Iyer, A.S.; Carter, H.; Becker, G.; Desai, Y.; Glang, L.; Choi, S.H.; Singh, K.; et al. Structural basis for transcriptional start site control of HIV-1 RNA fate. *Science.* **2020**, *368*, 413–417. [[CrossRef](#)]
72. Adachi, A.; Gendelman, H.E.; Koenig, S.; Folks, T.; Willey, R.; Rabson, A.; Martin, M.A. Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. *J. Virol.* **1986**, *59*, 284–291. [[CrossRef](#)] [[PubMed](#)]
73. Cole, B.; Lambrechts, L.; Gantner, P.; Noppe, Y.; Bonine, N.; Witkowski, W.; Chen, L.; Palmer, S.; Mullins, J.; Chomont, N.; et al. In-depth single-cell analysis of translation-competent HIV-1 reservoirs identifies cellular sources of plasma viremia. *Nat. Commun.* **2021**, *12*. [[CrossRef](#)] [[PubMed](#)]
74. Los Alamos National Laboratory Los Alamos HIV Sequence Database. Available online: <http://www.hiv.lanl.gov/> (accessed on 7 September 2021).