AccuNGS: a sequencing protocol for detection of ultra-rare variants reveals extensive variation in HIV during first days of infection

Maoz Gelbart[1*], Sheri Harari[1*], Ya'ara Ben-Ari[1], Talia Kustin[1], Moran Meir[1], Danielle Miller[1], Orna Mor[2,3] and Adi Stern[1]

[1]School of Molecular Cell Biology and Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

[2]Central Virology Laboratory, Ministry of Health, Sheba Medical Center, Ramat-Gan, Israel.
[3]Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel

[*]Authors contributed equally

Correspondence to AS: sternadi@tau.ac.il

ABSTRACT

Next generation sequencing (NGS) is widely used to identify genetic diversity within a sample. This includes basic scientific interrogations such as quantifying standing genetic variation and estimating mutation rates, as well as clinical applications such as surveillance of drug-resistance mutations and assessment of a pre-tumorigenic state. While several novel sequencing methods were established for accurately measuring rare genetic variants (prevalence below 1:100), the low yield of these methods and their relative complexity limits their utility for clinical samples. Here we describe a simple NGS protocol with high template recovery that enables detection of variants at frequencies as rare as 1:10,000. We thoroughly explored possible contributors to the background error level of the protocol by performing differential sequencing that targeted virtually every step of the protocol. Our results reveal that unlike most protocols in use, errors introduced during sample amplification are not predominant. We highlight that as opposed to what is commonly assumed, Illumina MiSeq paired-end reads are not independent from each other, and biochemical processes inside the sequencer itself likely contribute most to the background errors. Finally, we applied our method to an RNA sample from a seronegative HIV-1 patient during acute infection, i.e., with an infection duration of approximately twenty days. Our results revealed transition mutations in around 50% of the positions sequenced with elevated rates of G>A mutations, possibly driven by host deaminases. These results highlight the early rich accumulation of genetic diversity in early HIV infection at depths previously unseen.

INTRODUCTION

Recent advances in high-throughput nucleic acid sequencing have revolutionized our ability to identify the prevalence of minor traits in a heterogeneous sample. Identification of rare single nucleotide variants (SNVs) is important in diverse disciplines spanning post-transcriptional modifications, cancer genetics, non-invasive prenatal diagnoses and microbiology [1-4]. SNV identification in virus populations is currently at the heart of many studies monitoring drug resistance, estimating mutation rates, quantifying standing genetic variation and predicting the fitness costs of single mutations [4-10]. Accountable quantification of such variants present in clinical specimens requires high template recovery, sufficient sequencing depth, and discrimination of real minor variants from the background errors of the sequencing process [11, 12]. However, using the standard next generation sequencing (NGS) protocols may result in significant background error rates. In fact, following typical post-processing of NGS data, mutations that are at frequencies lower than 1-5% are discarded, drastically limiting rare SNV identification [9, 13, 14]. This limitation of NGS has been recently pointed as one of the major gaps in using genotyping to survey resistance mutations and understand HIV treatment failure [15, 16].

In the past few years, several innovative experimental approaches were suggested to reduce the background error rates of the NGS process: rolling-circle-based redundant coding of the amplified fragments [1, 17-19]; consensus sequencing of barcoded genomic fragments [19-24]; error reduction by overlapping paired reads in paired-end sequencing [25-27]; and usage of improved polymerases [28]. Complementary to the library preparation methods, several computational methods were created to facilitate discrimination of true variants from process errors, based on systematic background error modeling [24, 29-33]. Apart from the usage of overlapping read pairs (ORP), most experimental methods described above are designed for samples with high biomass and are inapplicable for sequencing of viruses from clinical samples, where the biomass of the viruses may be extremely low. Furthermore, these experimental protocols may provide accuracy at the cost of increased technical complexity, and may introduce their own artifacts to the sequencing process [34]. On the variant calling side, most variant callers model strand-specific sequencing bias but do not properly incorporate the information from the paired region that may have different error characteristics [35]. Moreover, it has been suggested that these well-established variant callers do not perform well on clinical virus samples [36].

We therefore sought to create a simple and highly accurate sequencing protocol that will be suitable for sequencing of clinical samples, with a special focus on low-biomass samples of RNA viruses. The motivation was based on the fact that many of these viruses replicate at huge census population sizes with high mutation rates of ~10^{-4}-10^{-5} mutations/base/replication, so viral populations from clinical samples are expected to contain very high levels of heterogeneity [5, 37,

38]. We set out to perform a step-by-step optimization of NGS protocols with an emphasis in each step on high accuracy and high yield. The resulting optimized protocol termed “AccuNGS” is a simple and rapid sequencing method and includes an associated variant caller, which if combined can reliably detect ultra-rare variants at frequencies of 1:10,000 or lower. Using homogenous DNA and RNA samples, we characterized the typical error landscape of the AccuNGS protocol, pinpointed the potential sources that generate errors in our protocol and suggest potential solutions. We applied our method to an RNA sample taken from a patient recently infected with HIV-1 (acute stage, seronegative) to examine the breadth of accumulation of mutations in the viral population in this critical period of the infection.

MATERIALS AND METHODS

Ethics declaration

The study was approved by the local institutional review board of the Sheba Medical Center (approval number SMC 1765-14) and of Tel-Aviv University. Written informed consent for retention and testing of residual plasma samples was provided by the patients.

Preparation of plasmids

In order to maintain the plasmid stock as homogenous as possible, plasmids were transformed to a chemically competent bacteria cells [DH5alpha (BioLab, Israel) or TG1 [A kind gift by Itai Benhar (Tel Aviv University, Tel Aviv, Israel)]] by a standard heat-shock protocol. Based on the fact that *Escherichia coli* doubling time is 20 mins in average using rich growing medium [39], a single colony was selected and grown to a maximum of 100 generations. Plasmids were column purified (HiYield™ Plasmid Mini Kit, RBC Bioscience) and stored at -20°C until use.

Construction of baseline control amplicon

Baseline control amplicon was based on clonal amplification and sequencing of the pLAI.2 plasmid, which contains a full-length HIV-1_{LAI} proviral clone [40] (obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: pLAI.2 from Dr. Keith Peden, courtesy of the MRC AIDS Directed Program). The Integrase region of pLAI.2 was amplified using primers: KLV70 - 5’TTC RGG ATY AGA AGT AAA YAT AGT AAC AG and KLV84 - 5’TCC TGT ATG CAR ACC CCA ATA TG [41]. Polymerase chain reaction (PCR) amplification was conducted using the high-fidelity Platinum™ SuperFi™ DNA Polymerase (Invitrogen) in a 50µl reaction using 20-40 ng of the plasmid as input and according to the manufacturer instructions. Amplification in a thermal cycler was performed as follows: initial denaturation for 3min at 98°C, followed by 40 cycles of denaturation for 20sec at 98°C, annealing for 30sec at 60°C and extension for 1min at 72°C, and final extension for 2min at 72°C. An alternative high-fidelity DNA polymerase used was Q5 High-Fidelity DNA Polymerase (New England Biolabs, NEB). PCR cycles were set according to each manufacturer’s instructions using the above described PCR program. The Integrase amplicon was gel purified (Wizard® SV Gel and PCR Clean-Up System, Promega) and the concentration

determined by Qubit fluorometer (Invitrogen) according to each manufacturer instructions. The purified product was further used for library construction.

For the AmpR sample, the conserved *AmpR* gene was amplified from PLAI.2 plasmid using primers: AmpR FW - 5'AAA GTT CTG CTA TGT GGC GC and AmpR RV - 5'GGT CTG ACA GTT ACC AAT GC. PCR amplification was carried out as described above, except for extension duration of 30sec instead of 1min. Similarly, the conserved *RpoB* gene was amplified from the bacteria genome using the following primers: RpoB FW 5'- ATG GTT TAC TCC TAT ACC GA and RpoB RV 5'- GTG ATC CAG ATC GTT GGT G and the following PCR program: initial denaturation for 3min at 98°C, followed by 40 cycles of denaturation for 10sec at 98°C, annealing for 10sec at 60°C and extension for 4sec at 72°C, and final extension for 2min at 72°C.

Construction of alternative purification amplicons

The agarose gel purification step of the amplified integrase gene was replaced with other purification methods; (1) For the gel-free sample, the amplified integrase gene was purified using 25µl of AMPure XP beads (0.5X ratio, Beckman Coulter) according to the manufacturer instructions; And (2) For the ExoSap sample, 10µl of the amplified integrase gene were mixed with 4µl of ExoSap (ExoSAP-IT™ PCR Product Cleanup Reagent, Applied Biosystems) and incubated according to the manufacturer instructions. No other changes in the generation of amplicon protocol were made.

Construction of a PCR free control amplicon

For the PCR-free sample, 10ug of PLAI.2 plasmid was digested using the restriction enzymes: NheI, StuI and XcmI (NEB) according to the manufacturer instructions. A ~1500bp fragment containing the integrase gene was gel purified and concentration was determined by Qubit. The purified product was further used for library construction.

Construction of an RNA control amplicon

A plasmid containing the full cDNA of Coxsackie virus B3 (CVB3) under a T7 promoter was a kind gift from Marco Vignuzzi (Institut Pasteur, Paris, France). The plasmid was used to generate an RNA control pool. Ten micrograms of this plasmid were linearized using SalI (NEB), purified by AMPure XP beads (0.5X ratio), and then *in-vitro* transcribed using T7 RNA polymerase (NEB) according to the manufacturer instructions. The transcribed RNA was purified using AMPure XP beads (0.5X ratio) and reverse transcribed with random hexamers using SuperScript III Reverse Transcriptase (Invitrogen) according to the manufacturer instructions. Four microliters of the reverse transcription reaction were used as template for a PCR reaction using primers: CVB FW - 5'GGA GAG AAG GTC AAC TCT ATG GAA GC and CVB RV - 5'TAC CAC CCT GTA GTT CCC CA, which amplify a ~1500bp fragment within the CVB genome. PCR reaction (50µl total) was set and amplified using Platinum™ SuperFi™ as follows: initial denaturation for 3min at 98°C, followed by 40 cycles of denaturation for 20sec at 98°C, annealing for 30sec at 60°C and

extension for 15sec at 72°C, and final extension for 2min at 72°C. The CVB amplicon was gel purified and the concentration measured by Qubit. The purified product was further used for library construction.

Construction of clinical HIV-1 amplicon with primer-ID

Plasma sample from a recently diagnosed HIV-1 patient (clinical sample, ID 83530) with >$1x10^7$ c/ml HIV-1 viral load was provided by the National HIV Reference Laboratory, Chaim Sheba Medical Center, Ramat-Gan, Israel. The mode of HIV-1 transmission for this patient was MSM, men who have sex with men. HIV-1 viral load was determined with Xpert HIV-1 viral load assay on GeneXpert (Cepheid Inc., Sunnyvale, CA), according to the manufacturer instructions [42]. RNA was extracted from 0.5 mL plasma by NucliSENS Easy MAG (Biomerieux, Marcy l'Etoile, France) according to the manufacturer's protocol, eluted in a final volume of 55 µl and stored in -80°C until use. A primer specific to the Gag gene of HIV-1 was designed with a 15 N-bases unique barcode followed by a linker sequence for subsequent PCR, Gag ID RT - 5'TAC CCA TAC GAT GTT CCA GAT TAC GNN NNN NNN NNN NNN NAC TGT ATC ATC TGC TCC TG TRT CT. Based on the measured viral load and sample concentration, 4 µl (containing the genomes of roughly 300,000 viruses) were taken for reverse transcription reaction. RT was performed using SuperScript IV Reverse Transcriptase (Invitrogen) according to the manufacturer instructions with the following adjustments; (1) In order to maximize the primer annealing to the viral RNA, the sample was allowed to cool down gradually from 65°C to room temperature for 10 minutes before it was transferred to ice for 2min; And (2) The reaction was incubated for 30min at 55°C to increase the overall reaction yield. To remove excess primers, the resulting cDNA was purified using AMPure XP beads (0.5X ratio) and eluted with 35µl nuclease-free water. To avoid loss of barcoded primers due to coverage drop at the ends of a read as a result of the tagmenation process, the PCR forward primer was designed with a 60bp overhang so the barcode ("primer ID") is far from read end, Gag ID FW - 5'AAG CGA GGA GCT GTT CAC TGC CAT CCT GGT CGA GCT ACC CAT ACG ATG TTC CAG ATT ACG and Gag ID RV **-** 5'CTC AAT AAA GCT TGC CTT GAG TGC. PCR amplification was accomplished using Platinum™ SuperFi™ in a 50µl reaction with 33.5µl of the purified cDNA as input using the following conditions: initial denaturation for 3min at 98°C, followed by 40 cycles of denaturation for 20sec at 98°C, annealing for 30sec at 60°C and extension for 1min at 72°C, and final extension for 2min at 72°C. The Gag amplicon was gel purified and the concentration determined by Qubit. The purified product was further used for library construction.

Libraries construction

PCR fragmentation and indexing of samples for sequencing was performed using the Nextera XT DNA Library Prep Kit (Illumina) with the following adjustments to the manufacturer instructions; (1) In order to get a short insert size of ~250bp, 0.85 ng of input DNA was used for tagmentation; (2) No neutralization of the tagmentation buffer was done, as described previously [43]; (3) For

library amplification of the tagmented DNA, the Nextera XT DNA library prep PCR reagents were replaced with high-fidelity DNA polymerase reagents (the same DNA polymerase that was used for the amplicon generation). The PCR reaction (50µl total) was set as depicted. Directly to the tagmented DNA, index 1 (i5, illumina, 5µl), index 2 (i7, illumina, 5µl), buffer (10µl), high-fidelity DNA polymerase (0.5µl), dNTPs (10mM, 1µl) and nuclease-free water (8.5µl) were added; (4) Amplification was performed with annealing temperature set to 63°C instead of 55°C, as introduced previously [43] and final extension for 2min; (5) Post-amplification clean-up was achieved using AMPure XP beads in a double size-selection manner [44] to remove larger fragments as well as smaller fragments, in order to obtain a narrower size-selection that will maximize the fraction of fully overlapping read pairs. For the first size-selection, 32.5µl of beads (0.65X ratio) were added to bind the large fragments. These beads were separated and discarded. For the second-size selection, 10µl of beads (0.2X ratio) were added to the supernatant to allow binding of intermediate fragments, and the supernatant containing the small fragments was discarded. The intermediate fragments were eluted and their size was determined using a high-sensitivity DNA tape in Tapestation 4200 (Agilent). A mean size of ~370bp, corresponding to the desired insert size of ~250bp, was achieved; And (6) Normalization and pooling was performed manually.

Alternative library purification methods

For the AMPure XP beads-free sample, post-amplification clean-up by double size-selection was replaced with an agarose gel purification of a ~370bp fragment, with no other changes in the library construction protocol.

Alternative tagmentation sample

For the alternative tagmentation sample, a 250bp amplicon within the integrase region was designed, using specific primers with an overhang corresponding to the sequence inserted during the tagmentation step of the NexteraXT DNA library prep kit, NexteraXT free FW - 5'TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG ACT TGT CCA TGC ATG GCT TCT C and NexteraXT free RV - 5'GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GTC TAT CTG GCA TGG GTA CCA GCA. PCR reaction was set up using Platinum™ SuperFi™ and carried out as follows: initial denaturation for 3min at 98°C, followed by 40 cycles of denaturation for 20sec at 98°C, annealing for 30sec at 62°C and extension for 15sec at 72°C, and final denaturation for 2min at 72°C. The PCR product was gel purified and the concentration was measured by Qubit. The purified product was indexed by a succeeding PCR amplification using primers corresponding to i5 and i7 NexteraXT primers (IDT) as mentioned previously [43] at a final concentration of 1uM. The PCR reaction was set up using Platinum™ SuperFi™ and amplified as detailed: initial denaturation for 3min at 98°C, followed by 12 cycles of denaturation for 20sec at 98°C, annealing for 30sec at 63°C and extension for 30sec at 72°C, and final extension for 2min at 72C. Size selection was achieved by gel purification of ~370bp fragments.

NextSeq libraries construction

Illumina NextSeq supports shorter reads than MiSeq. The longest NextSeq read length is 150bp, therefore we selected a shorter insert size of 270bp, compared to the 370bp insert size for the MiSeq platform. The first size selection of the post-NexteraXT amplification cleanup was performed using 42.5µl of AMPure XP beads (0.85X ratio) [44].

Sequencing

Sequencing of all samples (except for the NextSeq sample) was performed on the Illumina MiSeq platform using MiSeq Reagent Kit v2 (500-cycles) [Illumina]. Sequencing of the NextSeq samples was performed on the Illumina NextSeq 500 platform using NextSeq 500/550 High Output Kit (300-cycles) [Illumina].

Reads processing and base calling

The paired-end reads from each control library were aligned against the reference sequence of that control using an in-house script that relies on BLAST command-line tool [45-47]. The paired-end reads from the clinical HIV-1 sample were aligned against HIV-1 subtype B HXB2 reference sequence (GenBank accession number K03455.1) and then realigned against the consensus sequence obtained. Bases were called using an in-house script only if the forward and reverse reads agreed and their average Q-score was above an input threshold (30 or 38). At each position, for each alternative base, we calculate mutation frequencies by dividing the number of reads bearing the mutation by sequence coverage. Positions were retained for analysis only if sequenced to a depth of at least 100,000 reads. In order to analyze the errors in the sequencing process we used Python (Anaconda distribution) with the following packages: pandas, seaborn, numpy and stats. Distributions of errors on control plasmids were compared using t-test or Mann-Whitney U test.

Variant calling

In order to facilitate discrimination of true variants from AccuNGS process artifacts, we created a variant caller based on two principles: (i) positions that exhibit relatively high level of error on a control sample are error-prone for the clinical sample as well; and (ii) process errors on a control sample follow a gamma distribution. A gamma distribution was fitted for each mutation type in the control sequence. In order to detect and remove outliers from the fitting process we used the “three-sigma-rule”, and positions that showed error higher than three standard deviations from the mean of the fitted distribution were removed. For these rare loci a base was called only if the mutation was more prevalent in the sample by an order of magnitude. For G>A transition mutations, four distinct gamma distributions were fitted, corresponding for all four G>A combinations with preceding nucleotide. Accordingly, for C>T transition mutations four gamma distributions were fitted as well, on the four C>T reverse complement mutations of the G>A mutations. Variants were called on the input RNA sample only if a mutation was in the extreme 1% of the corresponding gamma distribution fitted on the DNA control.

Standard sequencing control sample

Standard control sample of pLAI.2 was taken from a previous study [41]. For obtaining mutation frequencies we used the same pipeline as for the AccuNGS samples, but without correcting overlapping paired reads. Positions in this sample were analyzed only if sequenced to a depth of at least 2,000 bases.

RESULTS

Protocol overview

AccuNGS protocol was designed to enable targeted sequencing of a desired genomic region from clinical samples with maximal yield, high accuracy and rapid implementation. It combines several concepts, some of which were previously reported separately. The principles of our protocol are (i) use of high-fidelity polymerases during reverse transcription (RT) and amplification. In particular we chose to use either the SuperScript III or SuperScript IV RT enzymes (reported mean error rate near $5x10^{-5}$ [29, 48]) and the Platinum SuperFi DNA polymerase (error rate of near $5x10^{-7}$, [49] and manufacturer at https://www.thermofisher.com/); (ii) significant reduction of sequencer errors by overlapping paired-end sequencing; and (iii) a specialized base calling bioinformatics pipeline that incorporates mutation-specific and locus-specific distribution of background errors (see Methods and Fig. 1). Based on the above information we were able to calculate *theoretical* means of $1.78*10^{-5}$ and $6.78*10^{-5}$ errors per base introduced in our AccuNGS protocol for DNA and RNA samples, respectively (Table 1). An auxiliary part of our protocol allows for quantification of the actual number of RNA genomes sequenced using uniquely barcoded primers introduced early in the RT step (the "primer-ID" method) [2, 20, 21, 50, 51]. This is a critical measure when analyzing the levels of diversity present in clinical samples, since low genetic diversity observed in a sample may stem from a small number of sequenced templates rather than from real reduced diversity in the sample. We note that this manuscript focuses on two measures: the mean error rate, reported when we characterize the method, and a cutoff error rate (based on the gamma distribution of errors), which we report as a measure to be used when performing base-calling. Naturally the cutoff value is higher than the mean error.

Table 1. Mean expected error rates in AccuNGS.

Step	Error rate	Number of rounds	Expected error
Polymerase Chain Reaction (PCR)	$6.47*10^{-7}$ [ThermoFisher]	40 + 12	$1.68*10^{-5}$
Sequencing	10^{-6} (Q30 filtering & overlapping paired reads)	1	$1*10^{-6}$

Total (DNA starting material)			**$1.78*10^{-5}$**
Reverse Transcriptase (RT)	$3.1*10^{-5}$ – $6.5*10^{-5}$ [29, 48]	1	$4.8*10^{-5}$
Total (RNA starting material)			**$6.58*10^{-5}$**

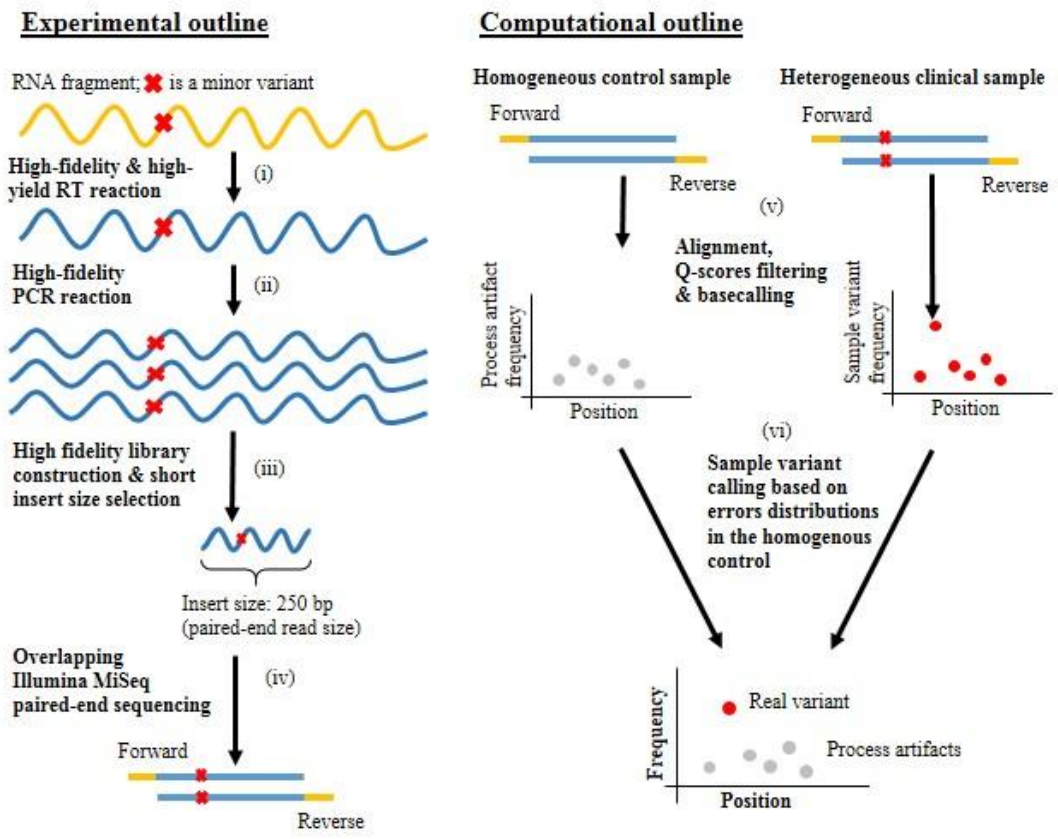

Fig. 1. AccuNGS protocol principles. (i) High fidelity and high-yield RT reaction. The RT primers may be designed with a unique N-base barcode (e.g. "primer-ID") to allow template quantification downstream (ii) High-fidelity PCR reaction (iii) Library construction with size selection for insert size as short as a single paired-end read (iv) Paired-end sequencing where each base in the insert is sequenced twice, once in the forward read and again in the reverse read (v) Alignment of reads, Q-score filtering on both reads and basecalling of both the sample and a homogeneous control (vi) Sample variant calling based on fitted distributions of process errors and position-specific error propensity.

AccuNGS error sources analysis at the DNA level

We began by evaluating AccuNGS when a DNA plasmid was used as starting material. Our underlying assumption throughout our working process was that our DNA starting material is homogenous with respect to the theoretical error rate we calculated. This assumption is based on

the fact that we used low-copy plasmids that were grown in *E. coli*, and only a single colony was subsequently sequenced. The mutation rate of *E. coli* is in the order of $1x10^{-10}$ errors/base/replication [23], and sequencing of a single colony ensures only a limited number of replication cycles. Accordingly, error rates in the purified plasmids are expected to be much lower than the expected protocol mean error of ~$1x10^{-5}$. Thus, errors observed when comparing the results of the sequencing to our known reference sequence reflect errors created by the library preparation or by the sequencing process itself. All samples underwent basecalling using our specialized bioinformatics pipeline and positions were considered for analysis only if their coverage exceeded 100,000 bases per position (see Methods). Table S1 provides statistics about the number of reads and the distribution of miscalls in each sample.

We thus set out to sequence the HIV-1 pLAI.2 plasmid [40] using AccuNGS at its baseline conditions, including 40 cycles of PCR amplification of a target region with a high-fidelity polymerase, followed by Nextera XT library preparation with a high-fidelity polymerase and size selection of a 250bp insert (see Methods and Table 2). We then compared the baseline AccuNGS error rate to the results of a protocol typically used in clinical (and other) settings, where less focus is put on the fidelity of the process [52]. Fig. 2 compares the proportion of errors observed on our plasmid sequence under AccuNGS and under standard sequencing, broken down according to type of mutation (base-to-base). Reassuringly, AccuNGS showed a significant improvement of one to two orders of magnitude over the standard sequencing protocol. For example, the mean A>G error rate went down from $2.6x10^{-3}$ down to around $9.2x10^{-5}$. While this improvement was large, perplexingly it was still an order of magnitude higher than the theoretical error rate we had expected of ~$1x10^{-5}$. We thus set out to optimize the AccuNGS protocol and try to pinpoint the unexpected source introducing errors into the process.

Sources of process errors based on differential sequencing. We next performed a set of sequencing trials, whereby at each trial we tested if removing or changing a specific stage of the protocol alleviates some of the observed errors and improves the fidelity of AccuNGS. Our immediate suspect was the PCR amplification step. Due to the exponential nature of PCR, errors introduced at early stages of the amplification will be carried over and have been reported to create a high background error for ultra-deep sequencing [53, 54]. Accordingly, we were worried that any misspecification of the error rate of the SuperFi DNA Polymerase we use (Table 1) would lead to an inflation in PCR errors. To test this hypothesis we created a sample with no PCR amplification, by harvesting larger quantities of the plasmid from the bacteria. This led to only a slight decrease in the mean error rate of transition mutations (Fig. 2). We hence concluded that the forty PCR cycles that take place in the PCR amplification prior to library preparation do not explain most of the errors of the AccuNGS protocol.

Table 2. Differential DNA control samples sequenced in this study.

Sample name	PCR enzyme	PCR cycles	PCR purification method	Library purification method	Cloning bacteria	Tagmentation method	Target region
Baseline	SuperFi (Thermofisher)	40+12	Gel	Beads	*E.coli* (DH5α)	Nextera XT	Integrase (pLAI)
Q5	Q5 (NEB)	.	.	.	.	.	.
PCR free	.	12	.	.	.	.	Integrase (extended)
Alt. PCR purification (beads)	.	.	Beads	.	.	.	.
Alt. PCR purification (Exosap)	.	.	Exosap	.	.	.	.
Alt. library purification (gel)	.	.	.	Gel	.	.	.
TG1	.	.	.	.	*E.coli* (TG1)	.	.
Alt. tagmentation	.	.	.	.	.	PCR	.
AmpR	.	.	.	.	.	.	*AmpR* (pLAI)
RpoB	.	.	.	.	.	.	*RpoB* (DH5α)
NextSeq (Illumina)	.	.	.	.	.	.	.

"." in a cell indicates a condition that is similar to the baseline; Alt=alternative. All samples besides the NextSeq sample were sequenced on the Illumina MiSeq sequencing platform.

We next focused on various so-called “chemical” processes that take place in AccuNGS library preparation. Mainly we were concerned that using gel extraction for DNA size selection, particularly UV light exposure, may introduce mutations. Indeed, when replacing gel extraction for the PCR products with magnetic beads extraction, we observed a slight reduction of the mean transition error rates (Fig. 2). On the other hand, this sample showed elevated levels of C:G>A:T errors. These errors are often signatures of oxidative stress and are discussed below. Alternative PCR purification using the Exosap cleanup reagent did not show elevated C:G>A:T errors compared to their levels in the baseline protocol, however it showed error levels comparable to the baseline protocol.

We next tested if the source plasmid itself was the major source of observed errors, focusing on the conditions whereby we grew the plasmid. First we tested if the mutations were accumulated naturally due to lack of selection on the HIV genes by sequencing the antibiotic resistance marker *AmpR* on pLAI.2 that is presumably under strong selection against mutations. Next we tested if

the plasmid was the source of errors by sequenced the highly conserved *RpoB* gene from *E. coli* itself. Finally we tested if the errors were introduced by the bacteria during plasmid replication by growing the plasmid in an alternative strain of *E. coli* (TG1) with a presumably lower mutation rate [55]. However, we observed no change in the error rate distribution in any of these conditions, suggesting that the DNA input was not the major source of minor variants.

We next hypothesized that the tagmentation process in the NexteraXT DNA library preparation kit (Illumina) might be the cause of artifacts. We resorted to a home-made tagmentation protocol based on introduction of NexteraXT-compatible adapters and indices via PCR amplification of a 250bp fragment of pLAI.2. Yet again, we observed no significant change in the distribution of errors. Replacing the Illumina MiSeq sequencer with the Illumina NextSeq, which is based on a two-channel sequencing process rather than a four-channel process and a different flow cell, resulted in a small increase in error levels. This suggests that the Illumina MiSeq and the higher throughput HiSeq (which employs the same detection method as MiSeq) are slightly more suitable for AccuNGS sequencing.

After having ruled out all sources of error we could conceive of testing, we were still left with the enigma of what causes the errors observed consistently and reproducibly across all the samples we sequenced. We were left with one condition that we could not alter directly: the sequencing step itself, as discussed next.

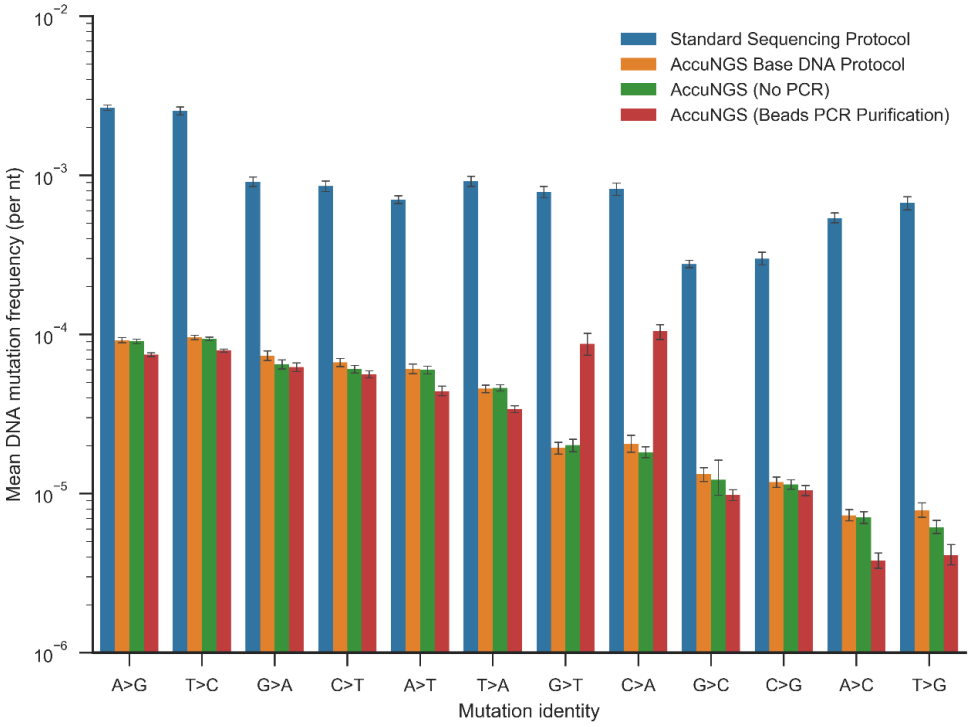

Fig. 2. Mean background error rates of different sequencing protocols at the DNA level.

Sequencing quality effect. Each base reported by Illumina sequencers is assigned with a probability of that base being wrong, termed the Q-score. The range of Q-scores reported from the Illumina MiSeq is 0 to 40, which translates to a probability of an erroneous call between 1 and 0.0001, respectively. In the AccuNGS base calling scheme, we consider only sites where the two overlapping reads reported the same base with an average Q-score of 30 or higher, as in [25]. Our original interpretation of overlapping reads, in line with previous works [56, 57] was that the base called jointly on both reads has a corrected Q-score equals approximately to the sum of the Q-scores from both reads. Accordingly, this means that if we filter for bases with a corrected Q-score of at least 60, this translates to an error probability $<=1x10^{-6}$ per base called, far below our theoretical threshold of $1.78x10^{-5}$. We set out to test if this is indeed true. First, we determined whether the independent Q-scores indeed reflect what they are supposed to. When inspecting errors observed on one read only with a Q-score of 30 or higher, we found that their maximal frequency was indeed around 10^{-3} (Fig. S3). Thus, it seems that the individual Q-scores on each read are reflective of the error rates of the process. However, we suspected that the joint Q-scores that we calculate are incorrect – mainly, that each reported Q-score is not independent of the Q-score on the mate read. To test this, we examined whether using a more stringent filter criterion improves AccuNGS results. We applied a very stringent quality filtering of Q38 on the AmpR sample that was sequenced to extreme depth of 1,500,000 bases per position (Table S1). The very high coverage and the good quality of the sequencing allowed us to still retain most sites at a coverage of above 100,000 reads per base (Table S1). We expected that this filtering will improve the results by the difference between twice Q30 (Q60, error probability of $1x10^{-6}$) and twice Q38 (Q76, error probability of $\sim 2x10^{-8}$). This difference translates to an improvement which is far below our observed error rate and hence we did not expect to see any improvement. Surprisingly, we observed a dramatic reduction in the rates of errors for A:T>G:C miscalls, and a modest reduction for C:G>T:A miscalls (Fig. 3, Table 3 and Table S2). We hence concluded that the assumption that the Q-scores of overlapping reads are independent is an incorrect assumption, and that the sequencer itself is the major source of errors in AccuNGS.

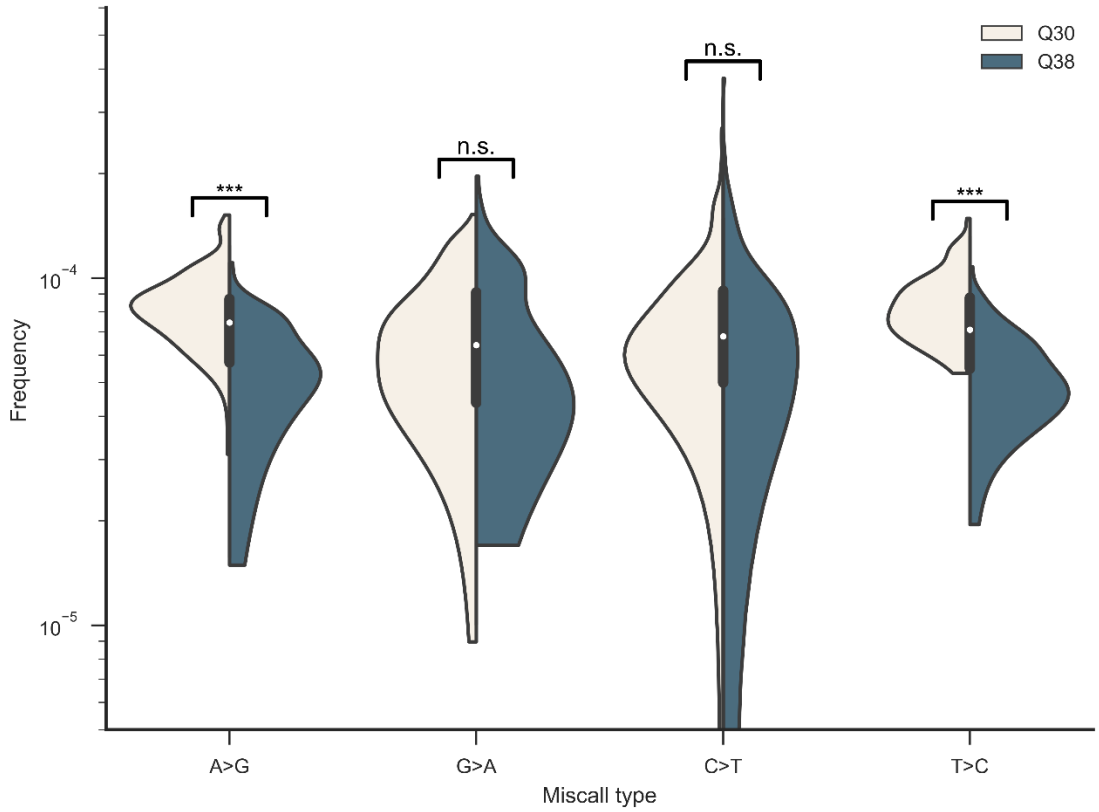

Fig. 3. The effect of increasing the Q-score filtering threshold on AccuNGS error rates. Error distributions for Q30 and Q38 filtering for the AmpR amplicon, presented for each type of transition error.

n.s., not significant; ***p<0.0001

Table 3. Fitted gamma distributions for different Q-scores cutoffs and substitution types, on the AmpR amplicon.

Q-score Cutoff	Substitution type	Shape (κ)	Scale (θ)	0.95% of fitted gamma	Mean error rate	Median error rate
Q30	A->G	8.855585	9.31E-06	1.32E-04	8.24E-05	8.30E-05
	C->T	4.511632	1.69E-05	1.42E-04	7.61E-05	6.49E-05
	G->A	5.32	1.33E-05	1.27E-04	7.05E-05	6.69E-05
	T->C	12.8402	6.57E-06	1.26E-04	8.43E-05	8.02E-05
Q38	A->G	5.863207	8.51E-06	8.79E-05	4.99E-05	4.89E-05

C->T	2.773564	2.49E-05	1.48E-04	6.92E-05	5.90E-05
G->A	2.879898	2.17E-05	1.32E-04	6.25E-05	5.36E-05
T->C	7.09643	7.01E-06	8.39E-05	4.98E-05	4.77E-05

Effects of surrounding nucleotides on error rates. Previous studies have indicated that the nucleotides surrounding a called base may influence its propensity to be miscalled [13, 28]. Indeed, we found that the identity of the surrounding bases sometimes affected the error rate observed with AccuNGS: when focusing on G>A mutations, we observed a higher error when the G was preceded by a C (CpG) and a lower error when the G was preceded by an A (ApG, Fig. S1). The exact same pattern was observed for the reverse complement C>T mutations (data now shown). These phenomena were found in all sequenced samples, suggesting that they aren't affected by any of the differential conditions we tested. When analyzing transversion artifacts, we observed a higher error rate for G:C>T:A mutations compared to all other types of transversions. This enrichment was more prominent in some samples than in the others and is typically associated with oxidative damage [13, 58-61]. When characterizing the nucleotide context of the C:G>A:T errors, we found that G>T errors occurred more frequently when the mutated G base was followed by A or another G (GpA\G). As in the G>A transitions, the reverse complement C>A mutations were more frequent when C was preceded by C or T (C\TpC; Fig. S2).

AccuNGS error analysis at the RNA level

Since one of the ultimate goals of the development of AccuNGS was the sequencing of RNA viruses, we set out to characterize how the protocol fares for RNA. In order to obtain a homogenous RNA sample we performed *in-vitro* transcription of a homogeneous plasmid using T7 polymerase, whose error rate has been approximated in the order of 10^{-6} [62], an order of magnitude lower than the error rate observed for DNA with AccuNGS. The RNA was then used as input for reverse transcription reaction with random hexamers using SuperScript III, whose mean error rate has been approximated to be between 3.1×10^{-5} and 6.5×10^{-5} [29, 48]. We then proceeded with the AccuNGS protocol for DNA as previously described. In this RNA control sample, we expected the observed errors to be the union of those introduced by the DNA part, those introduced during *in-vitro* transcription and those introduced during reverse transcription. With Q30 filtering, the RNA control sample indeed showed a higher mean transition error rate of 9.52×10^{-5} compared to the mean transition error rate of 8.49×10^{-5} in the DNA control sample (Table S1 and Fig. 4). The difference of 1.03×10^{-5} is indeed in line with most additional errors in the RNA control sample stemming from the RT step. By using the difference between the medians of these two control samples, we were able to calculate upper bounds on the base-by-base error rates of the RT used in the process, which were found to be lower for some mutation types than previously reported (Table S3).

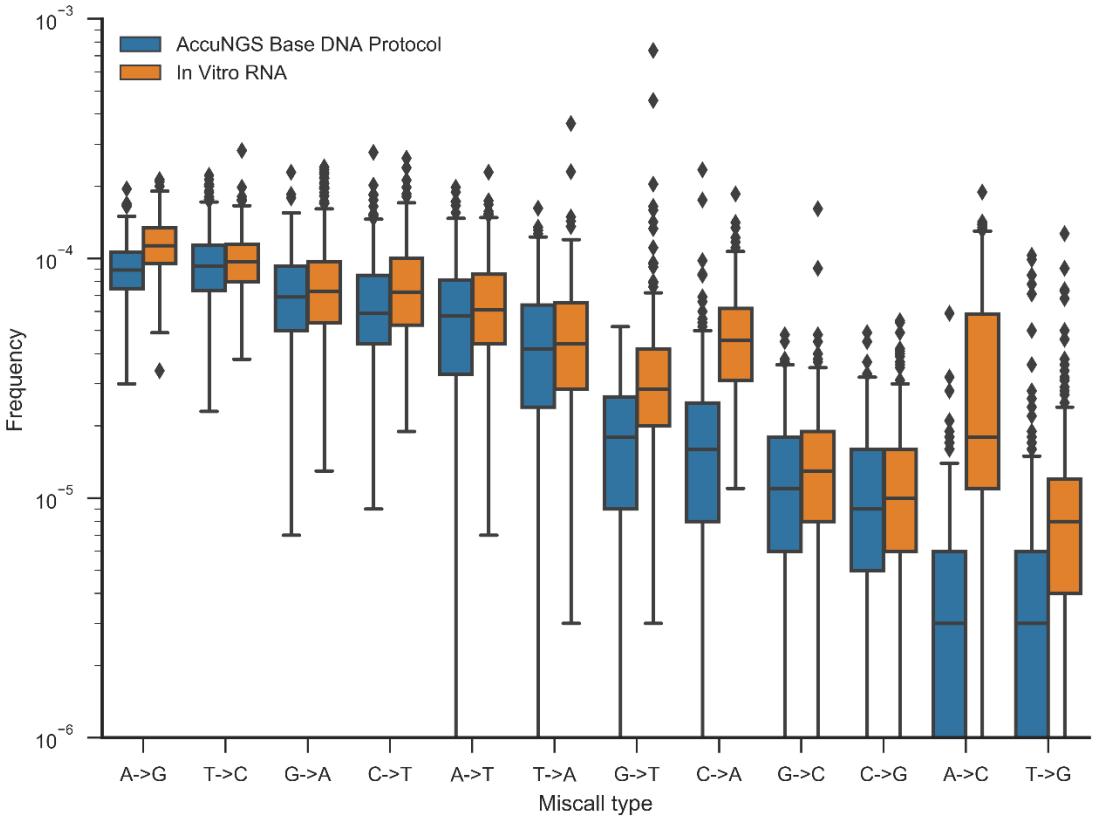

Fig. 4. Error rates of the DNA versus RNA control samples. Boxplots of errors per type of base changes are shown. Raw read bases were filtered when their average Q-score was less than Q30.

Clinical sample sequencing – acute HIV-1 infection

We next went on to test our method on direct sequencing of a clinical HIV-1 sample. HIV-1 infections typically begin with one to few viruses, indicating that the virus diversity at the population level at the time of infection is very limited [63]. Large population sizes coupled with high mutation rates in the order of 10^{-5} mutations/base/replication cycle allow the virus to obtain mutations shortly after infection [38, 64]. However, HIV-1 populations sequenced shortly after infection, while the patient is at acute infection, have shown very limited diversity [65-67]. At this time point, most variants in HIV are expected to be at frequencies below 10^{-3}, thus mostly obscured by the common clinical sequencing protocols' error rates. We obtained a plasma sample from a recently infected HIV-1 patient with laboratory confirmed seroconversion (a negative HIV-1 confirmatory assay followed by a positive test, 2 weeks apart), indicating this patient was likely 15-20 days after infection (so called acute HIV infection [68]). We chose to sequence the *gag* region of the virus (nearly 1800 bases) as it is mostly under purifying selection, but is also targeted by the HLA component of the adaptive immune system and possibly HLA

escape mutations will be seen at this early stage of infection [69]. We prepared each RT primer with a unique barcode ("primer ID"), 15 nucleotides long that allowed us to quantify the amount of viruses we have actually sequenced (see Methods and Suppl. File 1). We aimed to sequence ~30,000 viruses, which is roughly the inverse of the error rate we obtain with AccuNGS (which is around 1 in 10,000). To this end, we started the protocol with roughly 300,000 viruses, as estimated from the viral load of the sample. Based on RT processivity, we estimated that 10% of the viruses would hence be sequenced [70].

For background errors control, we amplified and sequenced a clonal pLAI DNA in parallel. The sequenced sample and the control had a median coverage of nearly 400,000 bases per position at Q30, allowing us to filter called bases for a minimum of Q38 for each base in the Forward and Reverse reads. As expected, background error rates in the control plasmid were similar to those obtained by our previous controls. Primer-ID analysis revealed that nearly 16,000 viruses were actually sequenced, so variants identified at frequencies of 10^{-4} are likely to represent true diversity (see Suppl. File 1). We applied our variant caller on the sample using the pLAI control to serve as the background error distribution. Using a p-value cutoff of 0.01 for each variant called (i.e., this variant is at or above the 99th percentile of the gamma distribution for this mutation type), between 40% and 50% of all sequenced positions were identified as containing true transition mutations (Fig. 5A and Table 4). Using the standard sequencing, only *four* transition mutations would have been identified; AccuNGS revealed that several hundreds of transition variants exist at this time of infection at low frequencies of 10^{-4}-10^{-3} (Table 4). When analyzing the type of mutations observed at low frequencies, synonymous mutations were the most prevalent, then non-synonymous mutations and then nonsense mutations (Fig. 5B). This suggests that signals of purifying selection can be already captured at this early time-point of infection and also strongly suggests that the variation observed is true. Interestingly, G>A minor variants were more prevalent than all other mutation classes (Fig. 5). Specifically, G>A variants preceded by A (GpA) were the most prevalent among all G>A variants, followed by G>A variants preceded by G (GpG, Fig. S4). This is possibly evidence for cytosine deamination activity by host APOBEC enzymes on the minus strand during reverse transcription, reflected as excess G>A mutations in the genome of virus [71]. This is also in line with the favored editing context of APOBEC3F [71]. However, G>A mutations are also the most common replication error of HIV-RT [64], and we cannot rule out that this drives higher frequencies of G>A as observed here.

Table 4. Variants identified with AccuNGS in the acute HIV sample. P-values were calculated with AccuNGS variant caller with the specified control sample as control. Sites were considered if their coverage exceeded 25,000.

Control	#Sites with sufficient coverage	Significance level	Number of variants (Ts+Tv)	Number of transition variants	% Positions with	Average transition variant frequency

					transition variants	
pLAI (DNA)	1469	5%	1535	799	54%	8.2E-04
pLAI (DNA)	1469	1%	1297	727	49.4%	8.98E-04
In vitro RNA	1469	5%	1105	649	44.1%	9.9E-04
In vitro RNA	1469	1%	984	594	40.4%	1.06E-03

Ts, Transitions; Tv, Transversions.

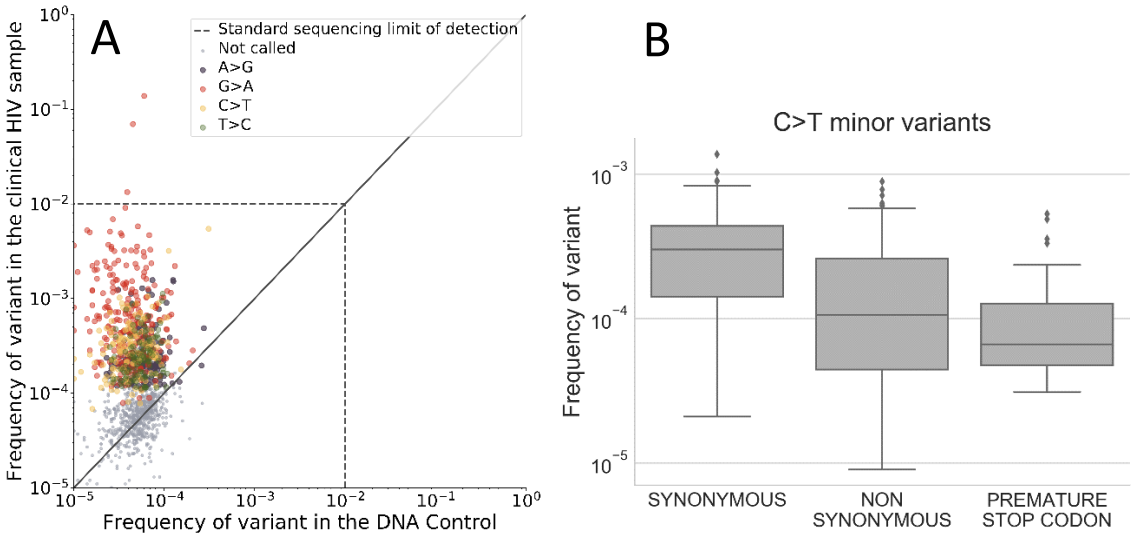

Fig. 5. Transition minor variants identified in the acute HIV-1 sample. (A) Called variants plotted against their respective frequency on the DNA control. Variants are colored if they reside in the top 1% of the fitted error distributions based on the DNA control. (B) Minor C>T transition variants by type. Synonymous mutations are more prevalent than non-synonymous and stop-codon-forming mutations.

DISCUSSION

Application of next generation sequencing on clinical samples is still limited by the ability to reliably capture minor variants [15, 16, 36]. Here we describe AccuNGS, a simple, rapid and optimized experimental protocol and associated computational pipeline for detecting ultra-rare variants from low-biomass clinical RNA and DNA samples. AccuNGS aims to accurately detect minor variants present in a population of genomes at frequencies of 1:10,000 or lower, close to the mutation rate of RNA viruses [5]. By performing differential sequencing we demonstrate that as opposed to many sequencing protocols, PCR is not a major source of errors in AccuNGS [53, 72], and conversely we suggest that the sequencer is a major source for errors, even when correcting the strand-bias using overlapping read pairs. We show that the mean transition error rate of the protocol is around 7.83×10^{-5} when filtering for Q30, and 5.80×10^{-5} when filtering for

Q38. These error rates translate to a cutoff of $1.35x10^{-4}$ and $1.16x10^{-4}$ based on the 95 percentile of a fitted gamma distribution of all transition errors. Notably when focusing on specific types of transitions (mostly T>C and A>G) the cutoff drops to below $1x10^{-4}$.

AccuNGS excels especially when the input is low-biomass heterogeneous RNA. Comparable methods for accurate sequencing such as rolling-circle-based methods typically require extremely high-biomass input, making them irrelevant for clinical virus sequencing. Furthermore we and others have observed that such protocols exhibited a relatively high C>T rate (exceeding 10^{-4} [1, 17, 18, 73], unpublished results). Such error levels were not recapitulated using AccuNGS, suggesting that these may have been artifacts of the rolling circle approach. A possible alternative for clinical sequencing would be the use of barcoded primers (also known as primer-IDs) during RT reaction, to generate consensus sequences that will correct errors inserted during amplification and sequencing [2]. The advantage of the barcoded approaches is that they can also correct for unequal PCR sampling. However, the downside of these approaches is that they require splitting the input sample into numerous reactions, since a barcode has to be attached to each sequencing read (typically spanning 500-600 bases). When the number of initial viruses in the sample is not huge (as typical for limited clinical samples), this is problematic since only a relatively small number of viruses will be sequenced in each reaction. AccuNGS is only limited by the capacity of the RT and PCR reactions (i.e., the length of the targeted sequence that undergoes one RT or PCR reaction), which may span several thousand bases. We also noted that error rates in barcoding-based protocols are comparable to AccuNGS [35, 74]. We do note that in the AccuNGS approach we recommend primer barcoding only on one end of the amplicon, but this is aimed to understand how many RNA templates were actually sequenced rather than for error correction.

The overlapping read pairs (ORP) concept to reduce sequencing errors was first reported by Chen-Harris et al. [25] and further used by PELE-Seq [26]. We note that our approach is novel in that it hinges on the combined use of high fidelity enzymes, ORP and a bioinformatics pipeline that compares variant frequencies to the fitted gamma distribution of process errors ([24, 27, 75] and BioRxiv: https://doi.org/10.1101/281659)). Indeed, AccuNGS improves over Chen-Harris et al., and we further show that the use of a high fidelity polymerase is key to bringing down the mean transition error rate by approximately 30% (Fig. S5). We further provide here a step-by-step dissection and optimization of the sequencing process. This has allowed us to refute the commonly assumed notion that the forward and reverse reads are independent and can be considered technical replicas (e.g. [56, 57, 76] and BioRxiv: https://doi.org/10.1101/276576). The observed improvement in error rates of A:T>G:C transitions when using a more stringent Q-score filtering threshold strongly indicate that the Illumina chemistry, involving low-fidelity polymerases [77], is a major source of sequencer errors that are only partially reflected in the Q-scores, and

are usually not modelled. This is supported by the observation that Taq DNA polymerase, which operates during cluster generation in the sequencer [77], tends to make more errors on A:T>G:C [54, 72]. The lack of improvement in the reciprocal G:C>A:T process errors in spite of more stringent filtering suggests these may stem from Cytosine deamination during thermal cycling, rather than from polymerization during cluster generation [78]. Our results also demonstrate that MiSeq to some extent better suits AccuNGS than the newer NextSeq platform, although significant improvement of error rates was also seen on the latter platform. We suggest that the observed error rate of the Q38 samples approaches the limit of detection that can be achieved using any traditional Illumina sequencing protocol. Hopefully in the future sequencing vendors will create a "high-fidelity" sequencing program that incorporates high fidelity enzymes that could minimize process artifacts.

We further note that the modeling process errors has several advantages over position-specific error models [25, 30]. First, position-specific error models do not perform well when the consensus base in the sample differs from the consensus base in the background homogeneous control. Second, the stochastic nature of process errors at a given position may result in significant differences between technical replicates, highlighting the fragility of an error estimation based on a single base observation. When possible, we strongly recommend using a homogenous control that is as similar as possible to the samples at hand, since this directly allows detecting loci that are highly error prone. In the absence of such a control, any homogenous control (e.g., a plasmid) is useful to control for process errors.

Performing our benchmarking on relatively long genomic regions allowed us to find that some errors tend to occur more in specific contexts than in others. We find that C:G>T:A mutations, that often arise from spontaneous cytosine deamination during thermal cycling [61, 72, 78], tend to occur when in CpG context (for both C and G), whereas they are less frequent when in CpT (for C>T) and ApG (for G>A) contexts. Similarly, we find that C:G>A:T mutations, which often stem from the formation of 8-oxo-Guanine under oxidative stress [61], are more prevalent in specific contexts. AccuNGS incorporates this observed bias into its variant calling method, based on more-specific background errors distributions.

We used AccuNGS to characterize HIV-1 diversity shortly after infection. To the best of our knowledge, the immediate evolution of HIV-1 following a new infection has been rarely observed [65-67], mainly due to the lack of resolution associated with the common sequencing protocols. We demonstrate that AccuNGS captures minor transition mutations that mostly segregate at frequencies between 1:100 and 1:10,000, and match the expected properties of real biological variants such as different rates for silent, missense and nonsense mutations. We found that in this sample, G>A mutations were three times more prevalent than C>U mutations (Fig. 5B). In spite of being the most abundant type of mutation induced by the viral reverse transcriptase [6],

the high level of observed minor frequencies exceeded our predictions, given the early time point. This may be partially explained by activity of APOBEC3 editing enzymes acting on the antisense of the HIV reverse transcribed genome. The most common APOBEC3 signature observed in our clinical sample (characterized by GpA>ApA mutations) was that of APOBEC3A/D/F/H, unlike a previous study that highlighted the contribution of APOBEC3G to the observed G>A variants in HIV-1 proviral DNA (characterized by GpG>ApG, [79]). It is possible that in our patient there are variants of APOBEC3 that are more active [80], or that in this patient, the viral Vif protein (that encounters host APOBEC3 proteins) lost the activity against some APOBEC3 enzymes [81]. Future studies spanning more patients may elucidate this issue.

Summary

To summarize, we anticipate that using AccuNGS will be highly useful in detecting previously uncharacterized genetic diversity in biological samples. The ease of use of this approach should make it highly amenable for many different studies.

COMPUTATIONAL RESOURCES

We have developed the following computational resources that complement the AccuNGS sequencing protocol:

(a) Base coverage calculator. AccuNGS relies on overlapping read pairs and high Q-scores for both reads of a pair. The calculator receives as input the length of the target regions and the desired coverage, and outputs the recommended number of reads required for sequencing each sample. It is available at [lab page, or excel suppl. file, or some DOI-based resource].

(b) Computational pipeline for computing the number of unique RNA molecules sequenced, based on primer-ID barcodes (see Suppl. File 1).

(c) Computational pipeline for base-calling and inferring site by site base frequencies.

All resources are freely available at https://github.com/SternLabTAU/AccuNGS.

DATA AVAILABILITY

The datasets generated and reported in the current study were deposited in the Sequencing Read Archive (SRA, available at https://www.ncbi.nlm.nih.gov/sra), under BioProject PRJNA476431.

SUPPLEMENTARY FILES

Supplementary File 1- clinical sample primer ID analysis.

ACKNOWLEDGEMENTS

We thank the members of the Stern lab for helpful comments and Roy Moscona for helpful discussions and support.

FUNDING

This work was supported by the Israeli Science Foundation [1333/16] to AS; by the German Israeli Foundation [I-1096-411.8-2015] to AS; by the United-States-Israel Binational Science Foundation [2016555] to AS; by the Edmond J. Safra center for bioinformatics in Tel Aviv University to MG, TK and DM; and by the Constantiner Institute for Molecular Genetics in Tel Aviv University to MG.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

REFERENCES

1. Reid-Bayliss, K.S. and L.A. Loeb, *Accurate RNA consensus sequencing for high-fidelity detection of transcriptional mutagenesis-induced epimutations.* Proceedings of the National Academy of Sciences, 2017: p. 201709166.
2. Salk, J.J., M.W. Schmitt, and L.A. Loeb, *Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations.* Nature Reviews Genetics, 2018.
3. Achermann, J.C., et al., *Disorders of sex development: effect of molecular diagnostics.* Nature Reviews Endocrinology, 2015. **11**: p. 478.
4. Beerenwinkel, N., et al., *Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data.* Frontiers in Microbiology, 2012. **3**(329).
5. Sanjuan, R., et al., *Viral mutation rates.* Journal of virology, 2010. **84**(19): p. 9733-9748.
6. Zanini, F., et al., *In vivo mutation rates and the landscape of fitness costs of HIV-1.* Virus Evolution, 2017. **3**(1).
7. Brumme, C.J. and A.F. Poon, *Promises and pitfalls of Illumina sequencing for HIV resistance genotyping.* Virus research, 2017. **239**: p. 97-105.
8. McElroy, K., T. Thomas, and F. Luciani, *Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions.* Microbial Informatics and Experimentation, 2014. **4**(1): p. 1.
9. Casadella, M. and R. Paredes, *Deep sequencing for HIV-1 clinical management.* Virus Res, 2017. **239**: p. 69-81.
10. Hartl, M., et al., *Within-patient HIV mutation frequencies reveal fitness costs of CpG dinucleotides, drastic amino acid changes and G→ A mutations.* bioRxiv, 2016: p. 057026.
11. Sims, D., et al., *Sequencing depth and coverage: key considerations in genomic analyses.* Nature Reviews Genetics, 2014. **15**(2): p. 121.
12. Gallet, R., et al., *The number of target molecules of the amplification step limits accuracy and sensitivity in ultradeep-sequencing viral population studies.* Journal of virology, 2017. **91**(16): p. e00561-17.
13. Meacham, F., et al., *Identification and correction of systematic error in high-throughput sequence data.* BMC bioinformatics, 2011. **12**(1): p. 451.
14. Huber, M., et al., *MinVar: A rapid and versatile tool for HIV-1 drug resistance genotyping by deep sequencing.* Journal of Virological Methods, 2017. **240**: p. 7-13.

15. Boucher, C.A., et al., *State of the Art in HIV Drug Resistance: Science and Technology Knowledge Gap.* AIDS reviews, 2018. **20**(1): p. 27-42.
16. Clutter, D.S., et al., *HIV-1 drug resistance and resistance testing.* Infection, Genetics and Evolution, 2016. **46**: p. 292-307.
17. Acevedo, A., L. Brodsky, and R. Andino, *Mutational and fitness landscapes of an RNA virus revealed through population sequencing.* Nature, 2014. **505**(7485): p. 686.
18. Lou, D.I., et al., *High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing.* Proceedings of the National Academy of Sciences, 2013. **110**(49): p. 19872-19877.
19. Wang, K., et al., *Ultrasensitive and high-efficiency screen of de novo low-frequency mutations by o2n-seq.* Nature Communications, 2017. **8**: p. 15335.
20. Jabara, C.B., et al., *Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID.* Proceedings of the National Academy of Sciences, 2011. **108**(50): p. 20166-20171.
21. Zhou, S., et al., *Primer ID validates template sampling depth and greatly reduces the error rate of next-generation sequencing of HIV-1 genomic RNA populations.* Journal of virology, 2015. **89**(16): p. 8540-8555.
22. Kennedy, S.R., et al., *Detecting ultralow-frequency mutations by Duplex Sequencing.* Nature protocols, 2014. **9**(11): p. 2586-2606.
23. Jee, J., et al., *Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing.* Nature, 2016. **534**(7609): p. 693-696.
24. Newman, A.M., et al., *Integrated digital error suppression for improved detection of circulating tumor DNA.* Nature Biotechnology, 2016. **34**: p. 547.
25. Chen-Harris, H., et al., *Ultra-deep mutant spectrum profiling: improving sequencing accuracy using overlapping read pairs.* BMC genomics, 2013. **14**(1): p. 96.
26. Preston, J.L., et al., *High-specificity detection of rare alleles with paired-end low error sequencing (pele-seq).* BMC genomics, 2016. **17**(1): p. 464.
27. Schirmer, M., et al., *Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform.* Nucleic Acids Research, 2015. **43**(6): p. e37-e37.
28. Imashimizu, M., et al., *Direct assessment of transcription fidelity by high-resolution RNA sequencing.* Nucleic Acids Research, 2013. **41**(19): p. 9090-9104.
29. Orton, R.J., et al., *Distinguishing low frequency mutations from RT-PCR and sequence errors in viral deep sequencing data.* BMC genomics, 2015. **16**(1): p. 229.
30. Gerstung, M., et al., *Reliable detection of subclonal single-nucleotide variants in tumour cell populations.* Nature communications, 2012. **3**: p. 811.
31. Wilm, A., et al., *LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets.* Nucleic acids research, 2012. **40**(22): p. 11189-11201.
32. Van der Borght, K., et al., *QQ-SNV: single nucleotide variant detection at low frequency by comparing the quality quantiles.* BMC bioinformatics, 2015. **16**(1): p. 379.
33. Yang, X., et al., *V-Phaser 2: variant inference for viral populations.* BMC Genomics, 2013. **14**(1): p. 674.
34. Brodin, J., et al., *Challenges with Using Primer IDs to Improve Accuracy of Next Generation Sequencing.* PLOS ONE, 2015. **10**(3): p. e0119123.
35. Zhang, T.-H., N.C. Wu, and R. Sun, *A benchmark study on error-correction by read-pairing and tag-clustering in amplicon-based deep sequencing.* BMC Genomics, 2016. **17**(1): p. 108.

36. McCrone, J.T. and A.S. Lauring, *Measurements of intrahost viral diversity are extremely sensitive to systematic errors in variant calling.* Journal of virology, 2016. **90**(15): p. 6884-6895.
37. Roberts, J.D., K. Bebenek, and T.A. Kunkel, *The accuracy of reverse transcriptase from HIV-1.* Science, 1988. **242**(4882): p. 1171-1173.
38. Preston, B.D., B.J. Poiesz, and L.A. Loeb, *Fidelity of HIV-1 reverse transcriptase.* Science, 1988. **242**(4882): p. 1168-1171.
39. Sezonov, G., D. Joseleau-Petit, and R. D'Ari, *Escherichia coli physiology in Luria-Bertani broth.* Journal of bacteriology, 2007. **189**(23): p. 8746-8749.
40. Peden, K., M. Emerman, and L. Montagnier, *Changes in growth properties on passage in tissue culture of viruses derived from infectious molecular clones of HIV-1LAI, HIV-1MAL, and HIV-1ELI.* Virology, 1991. **185**(2): p. 661-672.
41. Moscona, R., et al., *Comparison between next-generation and Sanger-based sequencing for the detection of transmitted drug-resistance mutations among recently infected HIV-1 patients in Israel, 2000–2014.* Journal of the International AIDS Society, 2017. **20**(1).
42. Mor, O., et al., *Evaluation of the RealTime HIV-1, Xpert HIV-1, and Aptima HIV-1 quant Dx assays in comparison to the NucliSens EasyQ HIV-1 v2. 0 assay for quantification of HIV-1 viral load.* Journal of clinical microbiology, 2015. **53**(11): p. 3458-3465.
43. Baym, M., et al., *Inexpensive multiplexed library preparation for megabase-sized genomes.* PLoS One, 2015. **10**(5): p. e0128036.
44. Bronner, I.F., et al., *Improved protocols for illumina sequencing.* Current protocols in human genetics, 2014: p. 18.2. 1-18.2. 42.
45. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.
46. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Research, 1997. **25**(17): p. 3389-3402.
47. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.
48. Potter, J., W. Zheng, and J. Lee, *Thermal stability and cDNA synthesis capability of SuperScript III reverse transcriptase.* Focus, 2003. **25**(1): p. 19-24.
49. Barnes, W.M., *The fidelity of Taq polymerase catalyzing PCR is improved by an N-terminal deletion.* Gene, 1992. **112**(1): p. 29-35.
50. Seifert, D., et al., *A comprehensive analysis of primer IDs to study heterogeneous HIV-1 populations.* Journal of molecular biology, 2016. **428**(1): p. 238-250.
51. Kivioja, T., et al., *Counting absolute numbers of molecules using unique molecular identifiers.* Nature methods, 2012. **9**(1): p. 72.
52. Van Laethem, K., et al., *A genotypic assay for the amplification and sequencing of integrase from diverse HIV-1 group M subtypes.* Journal of Virological Methods, 2008. **153**(2): p. 176-181.
53. Brodin, J., et al., *PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data.* PloS one, 2013. **8**(7): p. e70388.
54. Zanini, F., et al., *Error rates, PCR recombination, and sampling depth in HIV-1 whole genome deep sequencing.* Virus research, 2016.
55. Casali, N., *Escherichia coli host strains*, in *E. coli Plasmid Vectors*. 2003, Springer. p. 27-48.
56. Edgar, R.C. and H. Flyvbjerg, *Error filtering, pair assembly and error correction for next-generation sequencing reads.* Bioinformatics, 2015. **31**(21): p. 3476-3482.

57. Zhang, J., et al., *PEAR: a fast and accurate Illumina Paired-End reAd mergeR.* Bioinformatics, 2013. **30**(5): p. 614-620.
58. Costello, M., et al., *Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation.* Nucleic acids research, 2013. **41**(6): p. e67-e67.
59. Nishimura, S., *8-Hydroxyguanine: a base for discovery.* DNA repair, 2011. **10**(11): p. 1078-1083.
60. Park, G., et al., *Characterization of background noise in capture-based targeted sequencing data.* Genome biology, 2017. **18**(1): p. 136.
61. Arbeithuber, B., K.D. Makova, and I. Tiemann-Boege, *Artifactual mutations resulting from DNA lesions limit detection levels in ultrasensitive sequencing applications.* DNA Research, 2016. **23**(6): p. 547-559.
62. Sultana, S., et al., *Transcriptional fidelities of human mitochondrial POLRMT, yeast mitochondrial Rpo41, and Phage T7 single-subunit RNA polymerases.* Journal of Biological Chemistry, 2017. **292**(44): p. 18145-18160.
63. Keele, B.F., et al., *Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection.* Proceedings of the National Academy of Sciences, 2008. **105**(21): p. 7552-7557.
64. Abram, M.E., et al., *Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication.* Journal of virology, 2010. **84**(19): p. 9864-9878.
65. Zanini, F., et al., *Population genomics of intrapatient HIV-1 evolution.* Elife, 2015. **4**: p. e11282.
66. Kijak, G.H., et al., *Rare HIV-1 transmitted/founder lineages identified by deep viral sequencing contribute to rapid shifts in dominant quasispecies during acute and early infection.* PLoS pathogens, 2017. **13**(7): p. e1006510.
67. Henn, M.R., et al., *Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection.* PLOS Pathogens, 2012. **8**(3): p. e1002529.
68. Fiebig, E.W., et al., *Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection.* Aids, 2003. **17**(13): p. 1871-1879.
69. Markowitz, M., et al., *A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo.* Journal of virology, 2003. **77**(8): p. 5037-5038.
70. Levesque-Sergerie, J.-P., et al., *Detection limits of several commercial reverse transcriptase enzymes: impact on the low-and high-abundance transcript levels assessed by quantitative RT-PCR.* BMC molecular biology, 2007. **8**(1): p. 93.
71. Bishop, K.N., et al., *Cytidine Deamination of Retroviral DNA by Diverse APOBEC Proteins.* Current Biology, 2004. **14**(15): p. 1392-1396.
72. Potapov, V. and J.L. Ong, *Examining sources of error in PCR by single-molecule sequencing.* PloS one, 2017. **12**(1): p. e0169774.
73. Stern, A., et al., *The Evolutionary Pathway to Virulence of an RNA Virus.* Cell, 2017. **169**(1): p. 35-46.e19.
74. Kou, R., et al., *Benefits and challenges with applying unique molecular identifiers in next generation sequencing to detect low frequency mutations.* PloS one, 2016. **11**(1): p. e0146638.
75. Hao, Y., et al., *RareVar: A Framework for Detecting Low-Frequency Single-Nucleotide Variants.* Journal of Computational Biology, 2017. **24**(7): p. 637-646.

76. Masella, A.P., et al., *PANDAseq: paired-end assembler for illumina sequences.* BMC bioinformatics, 2012. **13**(1): p. 31.
77. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**: p. 53.
78. Chen, G., et al., *Cytosine deamination is a major cause of baseline noise in next-generation sequencing.* Molecular diagnosis & therapy, 2014. **18**(5): p. 587-593.
79. Cuevas, J.M., et al., *Extremely High Mutation Rate of HIV-1 In Vivo.* PLOS Biology, 2015. **13**(9): p. e1002251.
80. Harari, A., et al., *Polymorphisms and splice variants influence the antiretroviral activity of human APOBEC3H.* Journal of virology, 2009. **83**(1): p. 295-303.
81. Simon, V., et al., *Natural Variation in Vif: Differential Impact on APOBEC3G/3F and a Potential Role in HIV-1 Diversification.* PLOS Pathogens, 2005. **1**(1): p. e6.

SUPPLEMENTARY MATERIAL

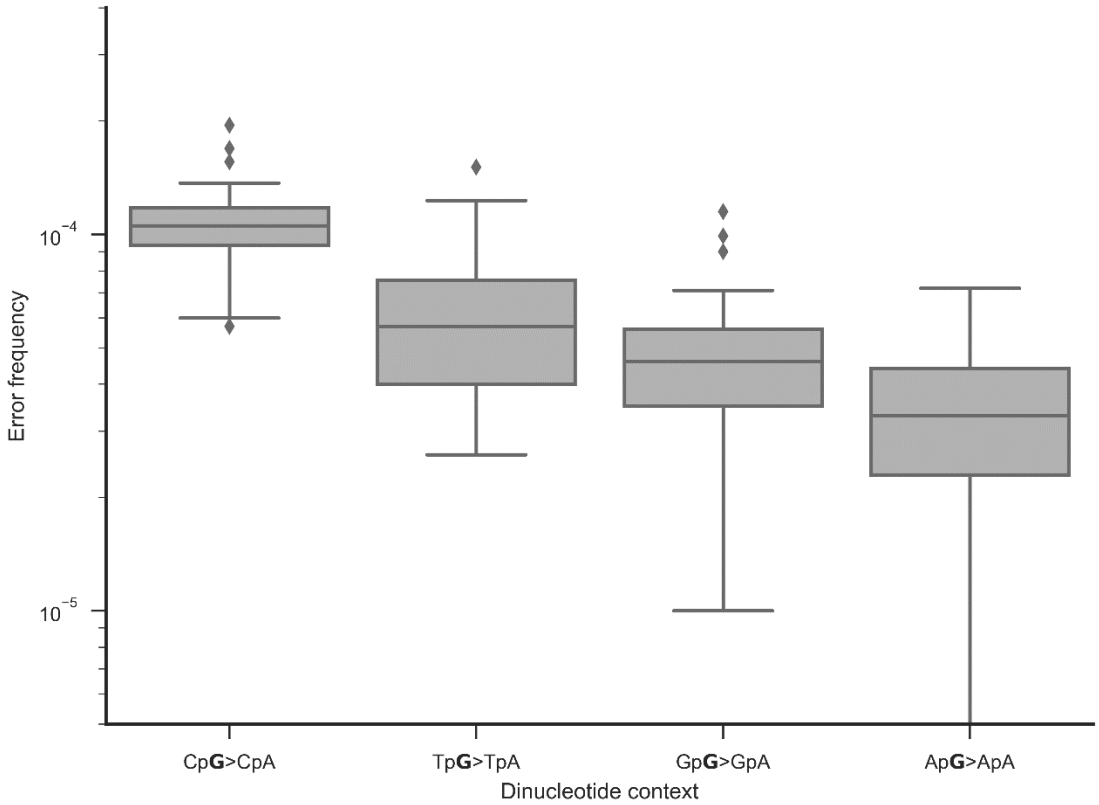

Fig. S1. Effect of sequence context on G>A error frequencies. Data shown on the AmpR sample (with Q38 filtering).

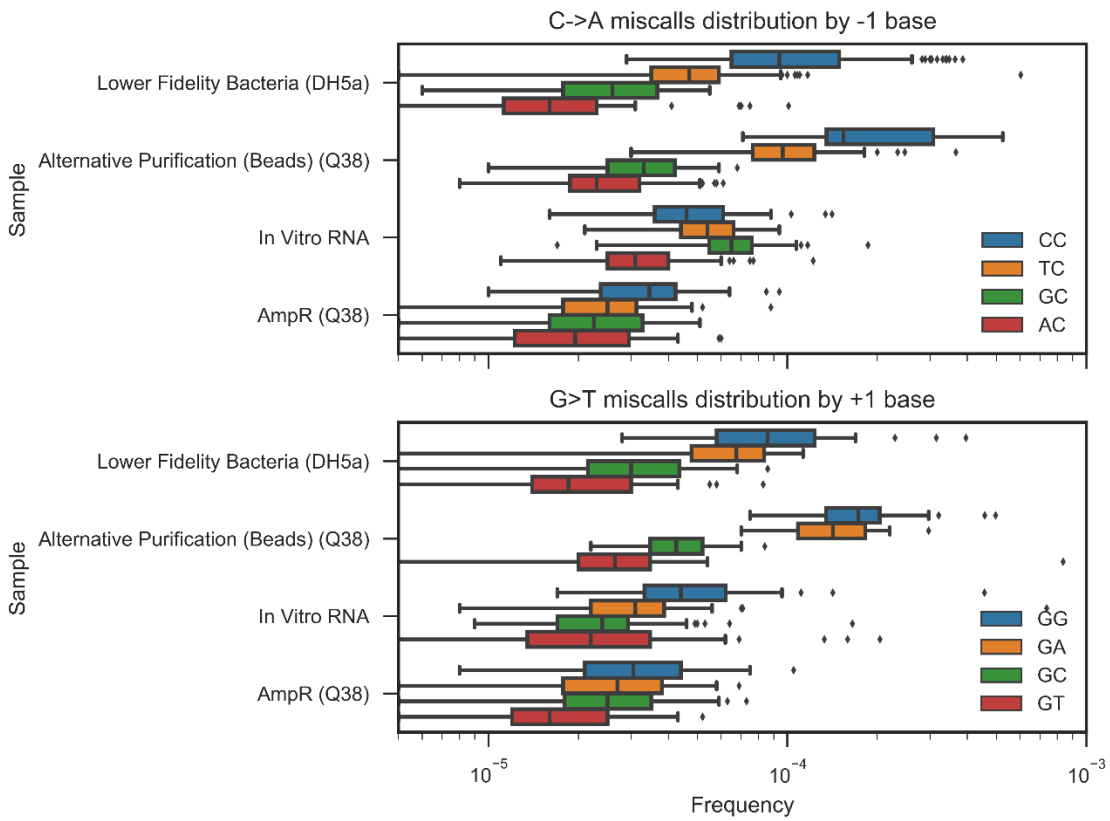

Fig. S2. Distributions of process errors associated with oxidative damage (G:C>T:A). Notably in the *In Vitro* RNA sample, the C>A errors pattern is different from the pattern observed in other samples due to the single-stranded origin of this sample.

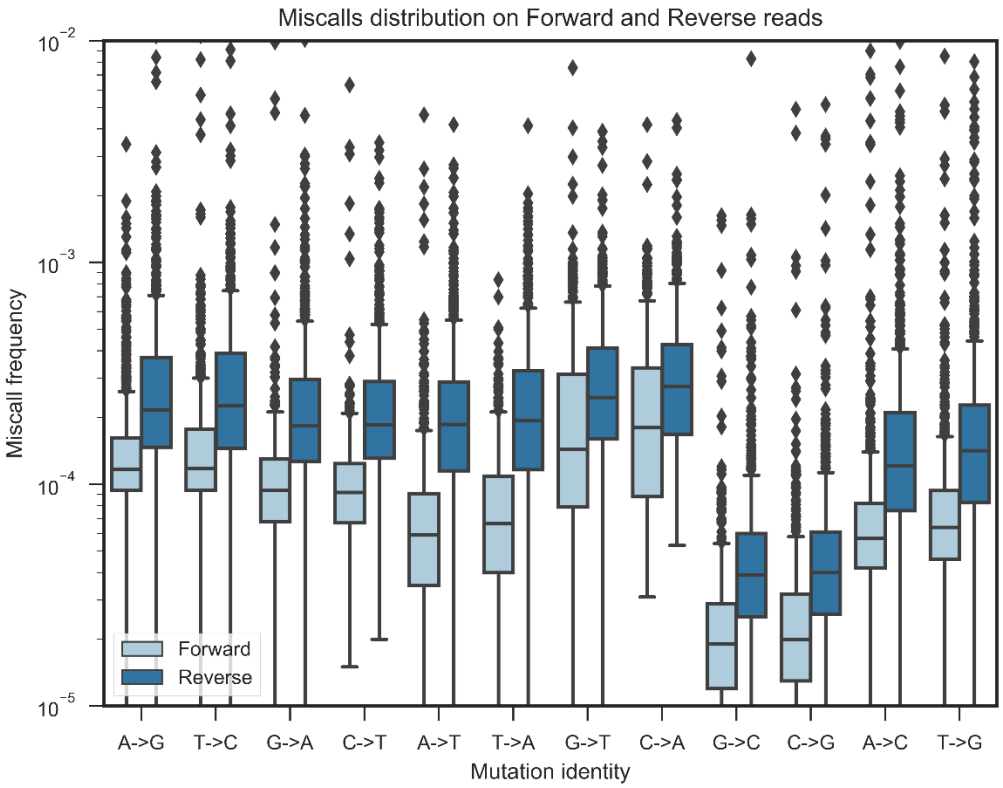

Fig. S3. Error distribution of forward and reverse reads in the No PCR sample. Bases were included if their quality was at least Q30.

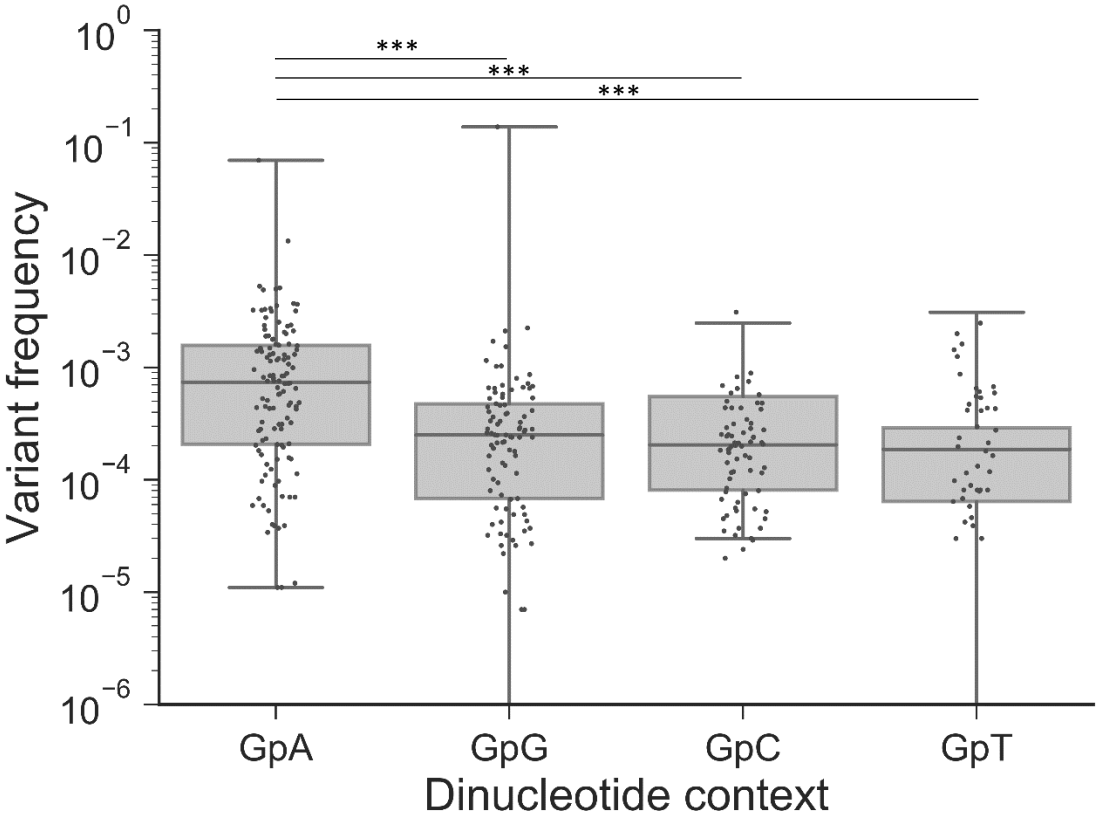

Fig. S4. Distribution of G>A minor variants in the HIV-1 sample, broken down by downstream nucleotide. G>A mutations in the context of A in the +1 position (GpA) were the most prevalent minor variants in the sample. Mann-Whitney U Test was used to test for different distributions of the various contexts.

n.s., not significant; ***p<0.0001

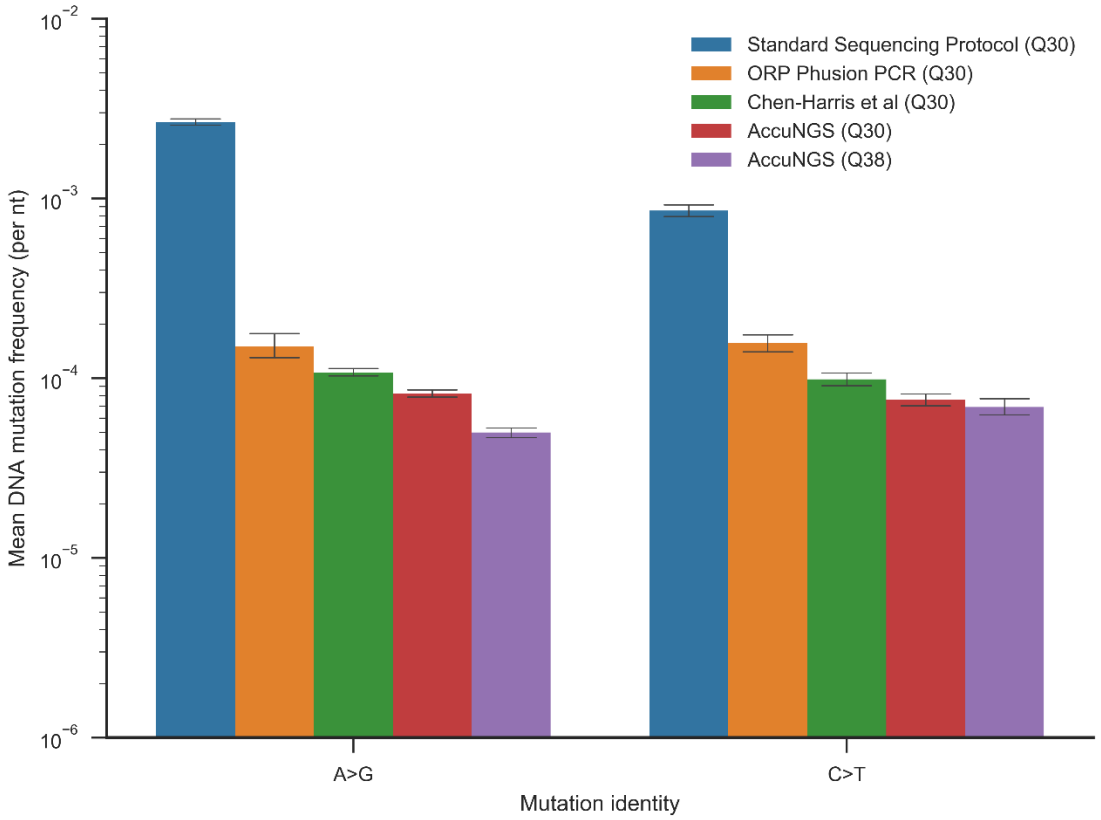

Fig. S5. A>G and C>T error frequencies in various DNA controls. Base quality filtering level is indicated in parentheses. The ORP Phusion sample is a sample prepared with AccuNGS with the Phusion High-Fidelity DNA Polymerase (Thermo Scientific) as a PCR enzyme (unpublished data). Chen-Harris et al is a control taken from [25], processed with the AccuNGS computational pipeline (i.e., no filtering beside the Q30 filtering).

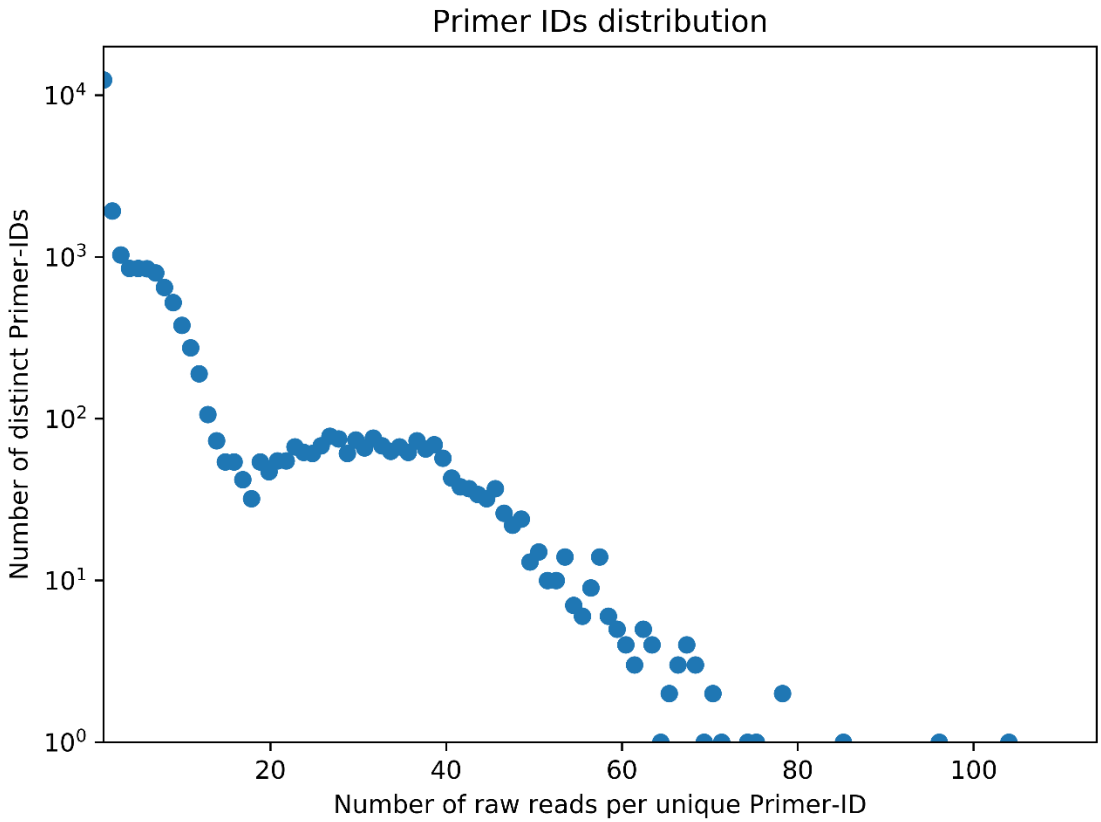

Fig S6. Primer-ID's abundance in the acute HIV sample.

Table S1. Sequencing statistics.

Sample	Target length	#positions called	#Reads	#Bases	Median coverage per position	Gamma 95%	Mean transition rate
Baseline	1,253	1,230	2,879,486	232,774,036	170,441	1.52e-4	8.49e-5
Baseline (biological replicate)	1,253	1,228	3,020,054	200,478,326	154,049	1.65e-4	8.90e-5
Alt. PCR enzyme (Q5)	1,253	1,228	4,633,464	346,227,696	270,777	1.47e-4	8.29e-5
PCR free	1,518	1,500	4,686,406	360,292,348	238,172	1.46e-4	8.11e-5
TG1	1,253	1,228	3,614,432	230,325,625	181,449	1.68e-4	9.19e-5
Alt. PCR purification (beads)	1,253	1,236	14,478,306	963,991,509	722,334	1.17e-4	6.94e-5
Alt. PCR purification (Exosap)	1,253	1,228	3,497,002	237,147,581	183,652	1.37e-4	7.71e-5
Alt. tagmentation (PCR)	250	188	337,378	28,775,994	154,609	1.72e-4	9.12e-5
Alt. library purification (Gel)	1,253	1,228	2,410,520	185,585,883	139,380	2.53e-4	1.06e-4
RpoB	1,031	1,009	2,759,920	368,214,150	180,053	1.53e-4	8.37e-5
In Vitro RNA	1,437	1,420	6,485,480	439,433,015	318,968	1.66e-4	9.51e-5
AmpR (Q30)	646	646	24,574,758	963,968,760	1,543,706	1.35e-4	7.53e-5
AmpR (Q38)	646	635	24,574,758	244,802,727	395,120	1.16e-4	5.03e-5
AmpR (Nextseq, Q30)	646	644	6,101,462	389,420,916	585,861	1.77e-4	9.80e-5
pLAI gag control (Q38)	1,860	1,720	12,824,998	673,387,667	358,723	1.02e-4	5.33e-5
In-vivo HIV gag (Q38)	1,860	1,743	7,289,860	802,204,580	494,427	n.a.	4.60e-4

Table S2. Relative contribution of identified factors to AccuNGS transition errors. Data were calculated based on the relative improvement in error rate in each DNA control sample, compared to the baseline protocol. Relative improvement for each control was calculated by comparing the median error frequency for each transition type.

Transition	PCR effect	Gel effect	Q-score effect	Unexplained
A>G	1%	18%	41%	40%
C>T	9%	12%	9%	70%
G>A	14%	17%	20%	50%
T>C	1%	17%	40%	42%

Table S3. SuperScript III median error rates estimations. Estimations are based on the difference in the medians of each mutation category between the In Vitro RNA control library and the AccuNGS baseline DNA protocol (pLAI). The RNA control library errors includes both the T7 RNA polymerase errors and SuperScript III errors, and therefore the provided estimates are expected to be upper bounds.

From\To	**A**	**C**	**G**	**T**	**Total**
A		1.14E-05	2.30E-05	1.31E-05	**3.83E-05**
C	2.99E-05		1.15E-06	4.47E-06	**4.41E-05**
G	3.91E-06	1.66E-06		1.02E-05	**1.57E-05**
U	2.36E-06	4.47E-06	1.51E-06		**8.34E-06**