Supplementary Text 1 – analysis of primer IDs.

Background. In this study, we use uniquely barcoded primers during the reverse transcription process to uniquely tag each viral genome that underwent reverse transcription. This method, also known as the "primer ID" method [1], allows for several improvements in deep sequencing analysis. First, it allows to quantify the number of sequenced templates [2]. Second, if coverage is deep enough, it allows for correcting process errors using consensus sequences obtained with each unique barcode. Third, it allows for true normalization of observed variability and reduction of PCR amplification bias.

AccuNGS primer ID approach. AccuNGS uses the primer ID method only for quantifying the amount of actually sequenced templates. This is partly because error correction is not required when the process fidelity is initially very high. Moreover, using primer ID on clinical samples for reducing the process errors requires splitting the sample into many aliquots, in a way that reduces the overall yield in the common low-biomass clinical scenario.

Primer ID design. Two factors limit the primer ID inference. One is the number of different templates in the sample, which may be low in the clinical sample use case. In the acute HIV sample, this was estimated to be 300,000 viruses using an external viral load assay. The second limiting factor is the depth of the sequencing, or how many reads with primer ID were outputted by the sequencer. Good primer ID length selection should take the minimum factor of both into account, and provide sufficient length to avoid having two similar primer IDs attached to two different templates ("primer ID collision") [3]. As previously indicated, the collision problem can modelled by the "birthday problem" [4]. We therefore created a birthday-problem-based calculator where one can calculate the chances for collision, given a desired primer ID length and an estimate on the limiting factor (actual templates count or sequencing depth). In the acute HIV sample, we estimated that 30,000 viruses will be sequenced and therefore designed a 15-nt long barcodes.

Quantifying the amount of templates. A common problem when relying on barcodes in the sequence is that the barcodes themselves are not error-prone and accumulate sequencing errors as well. When the number of sequenced templates is much lower than the sequencing depth, this may manifest as some primer IDs that are observed only a few times, and may be at a short "edit distance" from their source barcode. Such barcodes are often named "offsprings". Previous study has mapped the relationship between the most abundant barcode and the reliable depth a barcode is required to be sequenced in order to be considered as non-offspring [5]. Notably, this relationship was inferred using an experimental protocol that is more error-prone than AccuNGS. Currently, no standard exists for analyzing the primer IDs and downstream analysis. Zhou et al. provided a computational script that identifies primer ID by finding sequenced reads that has a perfect match to the known flanking sequences of the degenerate barcoded region. However, as sequencing errors may perturb the flanking regions this may result in incomplete barcode recovery.

We therefore developed our own primer ID recovery algorithm using a local alignment of the immediate flanking regions of the sequenced reads. The number of unique primer IDs identified is therefore an upper bound on the number of actually-sequenced templates, which in our sample was 22,971. The distribution of primer ID abundances is provided (Fig. S6), a figure that is similar to Fig. 7 in Zhou et al [5]. The most abundant primer ID in our study was sequenced 104 times. According to the Zhou et al. calibrated model, primer IDs sequenced one or two times should be considered offsprings. This will result in a lower-bound on the number of actually-sequenced templates at 8,617.

Offspring analysis. As most sequenced unique barcodes were sequenced once or twice and suspected to be offsprings of some paternal barcode, we tested the Hamming distance between all pairs of barcodes. Our analysis revealed that 15,880 barcodes had their closest barcode at a hamming distance of at least 3, suggesting that in AccuNGS most sequenced barcodes are real, and low barcode abundance may occur due to overall lower coverage and not due to sequencing errors of a more abundant barcode. We therefore conclude that we sequenced at least 15,880 different viruses.

References

1. 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.
2. Kivioja, T., et al., *Counting absolute numbers of molecules using unique molecular identifiers.* Nature methods, 2012. **9**(1): p. 72.
3. Liang, R.H., et al., *Theoretical and experimental assessment of degenerate primer tagging in ultra-deep applications of next-generation sequencing.* Nucleic Acids Research, 2014. **42**(12): p. e98-e98.
4. Sheward, D.J., B. Murrell, and C. Williamson, *Degenerate Primer IDs and the birthday problem.* Proceedings of the National Academy of Sciences, 2012. **109**(21): p. E1330-E1330.
5. 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.