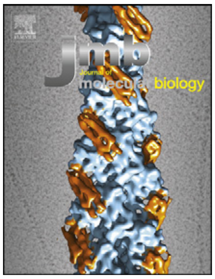
A Comprehensive Analysis of Primer IDs to Study Heterogeneous HIV-1 Populations



A Comprehensive Analysis of Primer IDs to Study Heterogeneous HIV-1 Populations

David Seifert[1,2], Francesca Di Giallonardo[3], Armin Töpfer[1,2], Jochen Singer[1,2], Stefan Schmutz[3,4], Huldrych F. Günthard[3,4], Niko Beerenwinkel[1,2] and Karin J. Metzner[3,4]

1 - ***Department of Biosystems Sciences and Engineering,*** *ETH Zurich, 4058 Basel, Switzerland*
2 - ***SIB Swiss Institute of Bioinformatics,*** *4058 Basel, Switzerland*
3 - ***Division of Infectious Diseases and Hospital Epidemiology,*** *University Hospital Zurich, University of Zurich, 8091 Zurich, Switzerland*
4 - ***Institute of Medical Virology,*** *University of Zurich, 8091 Zurich, Switzerland*

Correspondence to Niko Beerenwinkel and Karin J. Metzner: *N. Beerenwinkel is to be contacted at: Department of Biosystems Sciences and Engineering, ETH Zurich, 4058 Basel, Switzerland; K. J. Metzner, Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, University of Zurich, 8091 Zurich, Switzerland. niko.beerenwinkel@bsse.ethz.ch; karin.metzner@usz.ch*

http://dx.doi.org/10.1016/j.jmb.2015.12.012
Edited by M. Sternberg

Abstract

Determining the composition of viral populations is becoming increasingly important in the field of medical virology. While recently developed computational tools for viral haplotype analysis allow for correcting sequencing errors, they do not always allow for the removal of errors occurring in the upstream experimental protocol, such as PCR errors. Primer IDs (pIDs) are one method to address this problem by harnessing redundant template resampling for error correction. By using a reference mixture of five HIV-1 strains, we show how pIDs can be useful for estimating key experimental parameters, such as the substitution rate of the PCR process and the reverse transcription (RT) error rate. In addition, we introduce a hidden Markov model for determining the recombination rate of the RT PCR process. We found no strong sequence-specific bias in pID abundances (the same RT efficiencies as compared to commonly used short, specific RT primers) and no effects of pIDs on the estimated distribution of the references viruses.

Introduction

Knowing the composition of the intrahost viral population of patients afflicted with HIV, hepatitis virus, and other viral diseases is considered crucial for the personalized treatment of these diseases [1]. Previously, determining this composition required laborious biological assays, such as limiting dilution assays followed by single-genome sequencing, to obtain a small sample of variants of the virus population under study [2]. Such assays do not lend themselves to large-scale clinical applications, where diagnostics of a potentially large group of people need to be undertaken in order to devise individualized treatment plans.

The advent of next-generation sequencing (NGS) is likely to facilitate personalized viral diagnostics, as the entire pipeline from patients' samples to virus population structures is feasible today [3]. In addition, NGS provides data of viral populations that are much deeper, such that minor variants can be observed, which can be important for treatment outcome [4]. Lastly, due to the high yield of an NGS experiment, researchers are not constrained to study only certain parts of viruses, but they can use this technology with a focus on determining the whole-genome composition of a viral population [5], a feat that is impracticable with limited dilution assays and subsequent bulk sequencing.

While NGS is poised to revolutionize personalized medicine, the technology is not without its drawbacks. Of the many technology platforms currently available, such as Ion Torrent, Illumina, and Pacific Biosciences, all have drawbacks that require downstream computational and statistical handling. The Ion Torrent technology, for instance, is prone to make errors in homopolymeric stretches of DNA and Pacific Biosciences' Single Molecule Real Time Sequencing (SMRT®) technology is marred by a high insertion–deletion (indel) rate [6]. Even the currently most commonly used technology by Illumina contains idiosyncratic error patterns [7].

Statistical models accounting for sequencing errors aim to infer the true viral strains or haplotypes in order to make this error-prone NGS data useful. A popular class of methods involves nonparametric Bayesian clustering approaches, including ShoRAH [8] and PredictHaplo [9]. After correcting locally for errors, we can extend these methods, for instance, with graph-based algorithms or constrained extension of the local Bayesian clustering solution in order to determine global haplotypes, which are substantially longer than the average read length. The graph-based method HaploClique [10] involves finding sufficient overlaps between reads (or read pairs) and extending them to phase longer haplotype sequences. We use the term haplotype as synonym for genotype, which implies a different meaning than in human genetics. In HIV, a viral haplotype can consist of more than two alleles at any given locus.

These statistical tools can potentially correct for a majority of sequencing errors; however, misincorporations or recombinations in one of the initial cycles of the PCR cannot be corrected for, if such PCR mutants show up in frequencies above the sequencing error threshold. This effect is exacerbated by a fluctuating number of PCR duplicates and by the necessity of reverse transcription (RT) prior to PCR.

A different approach to this problem has been devised by experimental means. Instead of PCR duplicates possibly skewing the frequencies in the results, redundancy is taken advantage of. Tags are used to uniquely label cDNA molecules during RT, such that all observed heterogeneity of sequences after PCR and NGS are necessarily the result of substitutions in these two steps. Taking the consensus sequence of all sequences with the same identifier or tag by majority vote allows for removing most, if not all, PCR and sequencing errors. This procedure allows for deriving a near-perfect picture of the pool of cDNAs after RT.

Primer IDs (pIDs), barcodes, or tags are used in diverse fields. Immunologists employ barcodes in order to tackle the immense diversity of IgG haplotypes [11,12], an endeavor that is likely intractable with the aforementioned statistical tools, due to the low frequency of each haplotype in the population. pIDs have also seen application in cancer to distinguish preexisting from *de novo* resistance mutations [12,13]. The idea of using IDs to tag cDNA uniquely during RT has been advanced by Jabara *et al.* in the field of HIV-1 [14].

Given that pIDs are short, random DNA sequences of length k, or k-mers, with finite diversity, experiments need to be designed carefully in order to avoid the same k-mer being tagged to two different cDNA molecules, a statistical problem known as the Birthday paradox [15]. Liang *et al.* have analyzed the pID protocol in light of these collisions and showed how different minimum numbers of PCR replicates influence the fraction of data lost during analysis [16]. Brodin *et al.* have determined similar efficiency challenges [17] with input template numbers analogous to Liang *et al.*

Here, we have applied the pID protocol to estimate essential experimental parameters with respect to a validated reference set of five viruses. The well-defined mixture of these five viruses serves as the ground truth under which we can evaluate the pID protocol with respect to bias, efficiency, and comparing it to standard haplotype reconstruction tools. This mixture has been used in other studies, where its known composition allowed for stronger conclusions than using patient samples of unknown compositions [5,18]. We performed amplicon sequencing, where a specific locus is selectively amplified, because the pID protocol is not applicable to whole-genome sequencing.

Results

The RT efficiency is not impaired by using pIDs

To test the potential impact of long oligonucleotides used for the pID approach on cDNA synthesis, we designed a variety of primers differing in total length (14–67 nt), length of primer binding site (i.e., HIV-1-specific region), and tails added to the HIV-1-specific

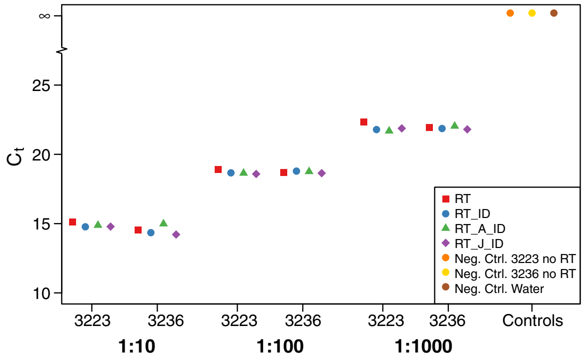


Fig. 1. RT efficiency as a function of the primer length. HIV-1 RNA was isolated from a dilution series of the HIV-1_{NL4-3} virus stock (1:10, 1:100, and 1:1000), reverse transcribed using primers of increasing length (Table S1), and quantified by qPCR. Depicted are the means of threshold cycles (C_t) of duplicates. Water and no RT samples served as negative controls.

Table 1. NGS sequence reads obtained from the six independent experiments

Run	Raw reads	After preprocessing	Aligned reads (valid, properly paired)	Pairs after filtering (paired reads)	Fraction of retained reads (%)
3223a	1,327,580	1,191,694	1,037,266	430,506 (861,012)	64.9
3223b	1,156,640	848,104	440,300	166,684 (333,368)	28.8
3223c	1,368,880	1,220,926	1,070,384	446,435 (892,870)	65.2
3236a	1,487,022	1,360,358	1,237,278	515,358 (1,030,716)	69.3
3236b	1,508,126	1,334,142	1,193,852	522,257 (1,044,514)	69.3
3236c	1,533,840	1,370,464	1,244,358	533,962 (1,067,924)	69.6

regions (Table S1). The total performance of eight RT primers were identical, as determined by quantitative PCR (qPCR) (Fig. 1).

NGS statistics

In all but one sequencing experiment, roughly 65–70% of the reads (Table 1 and Table S4) were used for the final analysis after pID quality checking and collision removal. Sequencing run 3223b suffered from degraded performance as can be observed in the FastQC tile qualities (supplementary information, section 1d, Fig. S4). Furthermore, we observed that the reverse reads of all experiments include more masked nucleotides than the forward reads due to the general decrease of quality of the reverse reads and the drop in quality for later sequencing cycles (Figs. S3–S8). We required the full length of the 23-nt PCR primer region to align, as we observed that reads with this region truncated are of very low quality. More comprehensive read statistics can be found in Tables S2–S4.

The distribution of pID lengths is peaked at 10-mers, as is expected, and the remaining lengths account for ~3% of all pID lengths (Fig. S10), with shorter pIDs being more common, that is, a general preference for deletions over insertions. For the distribution of the number of reads per pID, we found that a large fraction of pIDs is represented by only one read (Table S5), and beyond 10 reads per pID collection, the remaining collections display an approximately exponential decline in their frequency

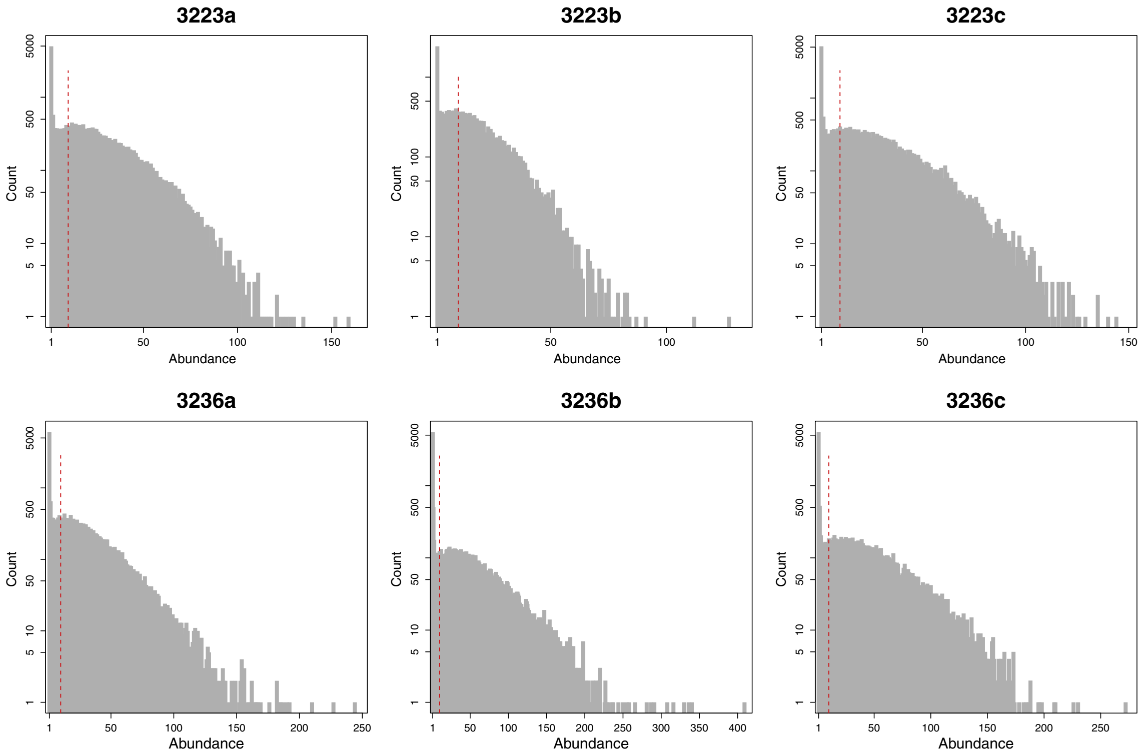


Fig. 2. Histograms of the abundance of reads per pID. The top row shows the triplicates of the shorter primer RT_J_ID pol 3223 and the bottom row shows the abundances of the longer primer RT_J_ID pol 3236. The spike at 1 depicts the number of unique pIDs, that is, singletons (Table S8). The red broken line indicates our required size of a pID collection, which is 10. All pIDs with less than 10 reads were discarded.

as a function of their abundance (Fig. 2), supporting the notion of a neutral amplification process.

Enzymatic error rates

We estimated the PCR substitution rate to be 9.68×10^{-5} [95% confidence interval (CI): 9.31×10^{-5}, 1.01×10^{-4}] and the RT substitution rate to be 6.02×10^{-4} per base per cycle (95% CI: 5.92×10^{-4}, 6.12×10^{-4}) after correcting for multiple sources of errors that would show up as conserved mutations. In order to correct for such possibly inflating sources of error to the estimated RT substitution rate, we devised a model to arrive at estimators of the substitution rate (supplementary information, section 2b). In practice, bacterial mutations do not affect this estimate. The PCR substitution rate is comparable to that of the standard *Taq* DNA polymerase reported in the literature [19] but is higher than expected for the high-fidelity variant of the enzyme. The RT substitution rate is higher than the expected rate of 3.4×10^{-5} given in the manufacturer's specification [20]. We found no indication that PCR-mediated recombination is a problem (supplementary information, section 1g). Of all filtered raw reads, at least 96.5% of raw reads could be explained by zero, one, or two substitutions in the heterozygous loci of one of the five clones. We estimated the combined RT PCR recombination rate to be 3.40×10^{-6} per base per cycle (95% CI: 2.56×10^{-6}, 4.40×10^{-6}), which is lower than that for the RT [21] alone. The CI values in this case are larger relative to the parameter estimate than for the RT substitution rate since the latent Markov chain introduces more uncertainty, which results in a higher variance of the estimator. We tested for changes in these parameters when collisions were not removed, and we found relative differences of less than 5%.

pID biases are not biologically relevant

The marginal position-wise frequencies of the four bases within the pID are independent of the position within the pID (Fig. 3). The position-wise distributions further show that unique pIDs or pIDs weighted by their PCR multiplicities have the same position-wise base distribution within the pID. Furthermore, the dinucleotide distributions of adjacent positions in the pID show no signs of coupling biases (supplementary information, section 3a, Fig. S18).

The intersections of the 500 most abundant pIDs per experiment are almost empty, indicating a very large class of equally efficient pIDs (Fig. 4a and b). Only experiments 3223a, 3236a, 3236b, and 3236c

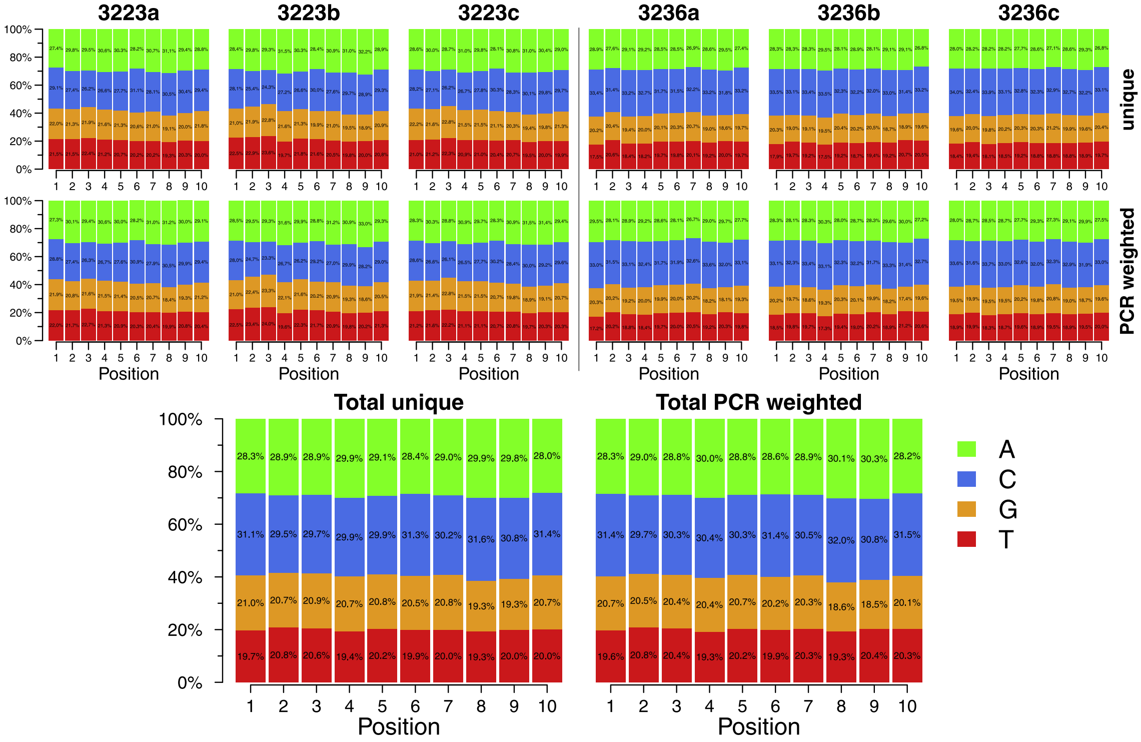


Fig. 3. Barplots of the position-wise nucleotide distributions within pIDs. The six graphs in the first row are based on uniquely observed pIDs, whereas the second row depicts nucleotide distributions on the basis of counts, weighted by their PCR multiplicities. The last two graphs at the bottom depict the pooled results from all experiments.

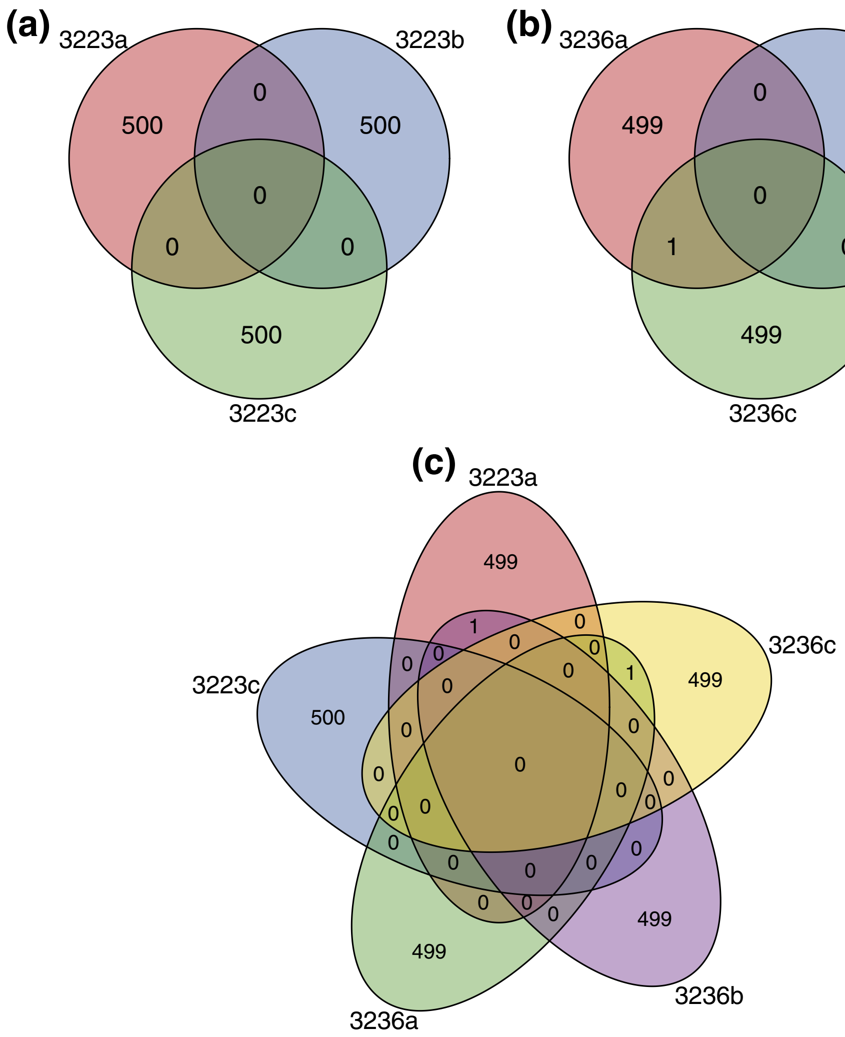


Fig. 4. Venn diagrams of the top 500 most abundant pIDs in all experiments. (a and b) Intersection for the shorter primer RT_J_ID pol 3223 and the longer primer RT_J_ID pol 3236. The union of all experiments is shown in (c), where run 3223b was omitted due to its low quality.

show any overlap with any other experiment in their top 500 pIDs. The top 500 pIDs of experiments 3223b and 3223c have an empty intersection with all other experiments (Fig. 4c).

We have fitted the PCR bottleneck model, which assumes equal efficiency of all pIDs, to the rank-abundance curves. Deviations from this model are indicative of pID selection. In the case of 3223a, 3223b, 3223c, 3236a, and 3236c, we have excellent fits for our neutral stochastic model to the frequency abundance curves (Fig. 5). For 3236b, the fits of our model are less optimal. This is mainly due to the large dynamic range (273) of the abundances in the first 100 ranks of this experiment. All estimates for the number of RNAs tagged vary between 7000 and 16,000 (Table 2), confirming the target bottleneck of around 10,000 initial molecules for the second PCR (Fig. S2).

pIDs neither bias nor significantly improve precision of population estimators

With the devised statistical Dirichlet test (supplementary information, section 4), we tested for differences in the bias and variance of the three derived population frequency estimators (Table 3). In constructing the five-virus mix, the target of 1:5 ratios of each clone in the mix was not achieved due to noisy RT qPCR quantification. As a result, relative frequencies of the clones deviate from the theoretical 20% and vary between 6% and 38% in practice. While the HaploClique frequencies appeared to be slightly further apart from the raw and pID frequencies, the difference was not statistically significant ($p > 0.05$; Fig. S21). Similarly, we found no evidence of any estimator having less variance than another ($p > 0.05$; Fig. S22). The HaploClique-based

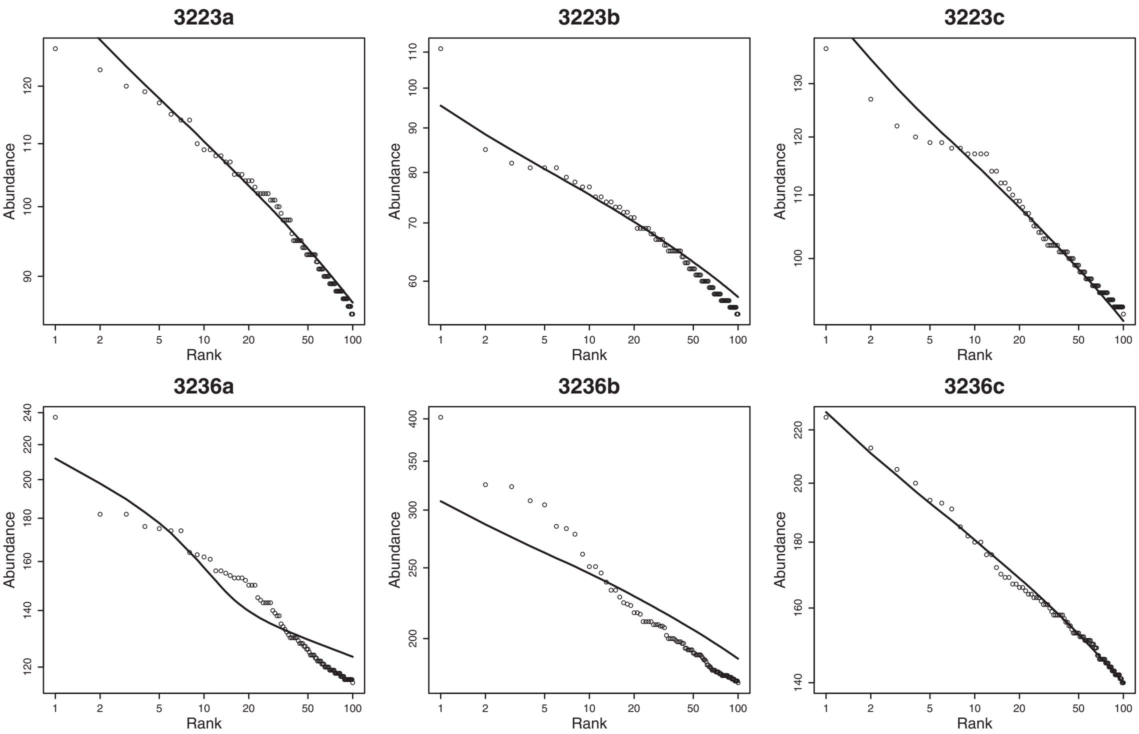


Fig. 5. Rank-abundance plots for the experiments consisting of two primers with three replicates. The dots depict the abundance of reads per pID for the first 100 ranks. The smooth continuous lines are the best-determined fits from the neutral stochastic RT PCR model.

frequencies had the highest variance, but the difference was not statistically significant. Similarly, the pID-based estimators had the lowest variance, but this difference was again not statistically significant. In addition, given the precision values of the raw and pID estimators, the difference does not appear to be biologically relevant, as the relative differences between these values are small. This is also supported by the standard deviation summary statistic of each clone across experiments (Table 3), where standard deviations of raw read, pID frequency, and HaploClique estimators are on the same order.

Table 2. Estimated size of bottleneck for the six independent experiments

Run	Rank 1 abundance	Rank 100 abundance	Dynamic range	Inferred bottleneck size
3223a	127	68	59	16,391
3223b	111	42	69	9686
3223c	137	72	65	14,808
3236a	237	85	152	7012
3236b	402	129	273	8128
3236c	225	108	117	11,921

The dynamic range is defined to be the difference in abundance between the pID at rank 1 and the pID at rank 100.

Discussion

We analyzed a pID protocol on the basis of an validated reference set of five viruses. Previous studies of this innovative protocol have always used patient samples, which does not allow for strong conclusions regarding the performance of this approach due to lack of a ground truth. Without such ground truth, the strengths and limitations of this method cannot be fully understood.

We found that the RT efficiency does not depend on the length of the RT primer for the primer lengths we studied. This indicates that the rate-determining step of RT annealing/ligation is dominated by the efficiency of hybridization within the first few bases of the primer region, such that the following bases, including the random pID, have a negligible impact on efficiency. The number of bases crucial for this step necessarily has to be less than 14 nt, given that the longer primer has the same efficiency. This has the positive effect that the pID length can likely be increased in order to decrease the probability of collisions of different RNA molecules.

The loss of reads due to an incorrect length of the pID is marginal and around 3% for all experiments in our case. We observed deletions to be favored over insertions within the pID, which is not surprising,

Table 3. Frequency estimates from the six independent experiments by the three different frequency estimators

Method	Clones	3223a (%)	3223b (%)	3223c (%)	3236a (%)	3236b (%)	3236c (%)	σ (%)
Raw	89.6	11.03	9.70	8.83	11.20	11.24	8.40	1.27
	HXB2	13.99	15.29	14.39	13.08	15.33	14.96	0.87
	JR-CSF	35.88	37.01	38.90	36.92	35.78	38.24	1.25
	NL4-3	23.86	21.56	21.87	23.79	24.22	22.93	1.11
	YU2	15.23	16.44	16.00	15.00	13.43	15.47	1.04
pID	89.6	10.80	9.28	8.27	11.14	11.02	8.26	1.36
	HXB2	14.06	15.02	14.75	13.32	15.38	15.45	0.83
	JR-CSF	36.64	37.16	39.51	37.30	36.26	38.99	1.31
	NL4-3	23.55	22.09	21.40	23.36	23.71	22.30	0.94
	YU2	14.94	16.45	16.06	14.88	13.63	15.00	1.00
HC	89.6	10.16	8.06	7.78	10.08	10.15	6.87	1.46
	HXB2	14.12	14.79	14.37	12.57	14.90	14.58	0.86
	JR-CSF	36.16	37.84	39.84	37.88	36.53	39.03	1.41
	NL4-3	24.48	23.32	22.26	24.78	25.59	24.12	1.17
	YU2	15.07	15.99	15.76	14.68	12.83	15.41	1.14

Raw frequencies and pID frequencies are determined by assigning raw reads and pID consensus sequences to the five clones without allowing for mismatches in the heterozygous loci. HaploClique (HC) determines the frequencies of haplotypes by normalizing the sizes of the haplotypes' respective maximal cliques in the course of its algorithm. The standard deviation of the frequencies of each clone over the six experiments is denoted by σ.

given the tendency of the MiSeq® sequencer to preferentially call deletions [22] and a coupling efficiency of below 100% during primer synthesis. While the primers ordered were PAGE purified, a certain fraction will always be shorter. The rate of collisions is between 1% and 2%, which is acceptable given a pID diversity of 1,024,768, considering that our collision-calling algorithm is somewhat conservative and will likely remove pIDs that have not been collisions but rather multiple PCR substitutions in the early cycles of a template.

We estimated enzymatic parameters of the RT and PCR steps. The estimated PCR substitution rate of roughly 10^{-4} is compatible with current known estimates of the *Taq* DNA polymerase, which is unexpected, as the employed Platinum *Taq* DNA polymerase High Fidelity has an advertised fidelity rate of 6 times the reference *Taq*. While our estimate is higher than the advertised rate of the manufacturer, this is likely due to experimental conditions that are unlike the conditions set forth by the manufacturer. However, we believe that our estimate more likely reflects the commonly experienced PCR substitution rate because, first, we optimized the PCR conditions to reduce PCR errors [18] and, second, our estimate is likely a lower bound, as it is based on substitutions occurring during the first cycles of the PCR, that is, when PCR conditions are still optimal. Some factors are especially detrimental to the PCR error rate in later PCR cycles when, for instance, unbalanced nucleotide concentrations can occur [23].

We estimated an RT substitution rate that is also higher than expected from the data given by the manufacturer. We excluded bacterial mutations by showing the negligible contribution deriving from the Luria-Delbrück distribution in the bacterial amplification, given even the highest of currently known mutation rates. We corrected the RT substitution rate estimate by accounting for other sources of errors that show up as conserved mutations, such as RNA polymerase II substitutions and substitutions in the first PCR that fixated during subsampling for the second PCR.

The RT PCR recombination rate of approximately 3×10^{-6} is lower than what is known from the literature. This has multiple reasons, one of which is likely the design of the reference five-virus mix, as the locus under study possesses little diversity. Due to the resulting small number of heterozygous loci, observing patterns of recombination is difficult and a number of recombination events will be missed due to the inherent similarity of some templates. In addition, the recombination rate critically depends on the amount of input material [24], which we have purposely kept low. The only way to improve these estimates would be in selecting different templates such that recombination becomes directly observable and recombination events can be counted, making models such as the hidden Markov model used here redundant. On the other hand, clones have to retain a high degree of homology; otherwise, recombination is unlikely to occur between templates [25]. The fact that 96.5% of all raw reads can be explained with a maximum of two mismatches in the heterozygous loci of one of the five clones, given the Hamming distances between clones (Table S6), provides strong indication that recombination is not a pervasive issue. This is in addition to the low recombination rate, which further supports the previously optimized experimental protocol for minimizing artificial recombination [18]. Given that the artificially constructed diversity of our sample is unexpected in patient samples, recombination in practice could be higher than what we inferred. Nonetheless, we believe such recombination effects not to be orders of magnitude above what we have observed, as the strains in our study still share a high degree of homology. Substitution and recombination rate estimates are always dependent on experimental conditions and hence cannot be generalized across different experimental methodologies [26].

We have observed negligible sequence biases in the form of a nonequiprobable position-wise marginal nucleotide distribution in the pID. These marginal effects are likely due to unequal nucleoside concentrations or intrinsic biases between nucleosides in the oligonucleotide solid-phase synthesis of the primers and not due to any systematic biases in the pID protocol. Given that the position-wide distribution of bases for unique pIDs and PCR-weighted counts is practically the same, this further corroborates our suggestion that pID efficiency is not a function of the pID sequence, such that an interaction between the pID and the adjacent sequence of the template seems unlikely.

In addition, pIDs showed practically no overlap between our experiments. This requires, at a minimum, a sufficiently large class of pIDs that are all equally selected for during the RT and/or PCR or a general preclusion of strong selective advantages of all pIDs. Any small number of pIDs that have supposedly very high efficiencies would be reproducibly represented in the top 500 of pIDs between experiments, unlike what we observed. Lastly, the rank-abundance spectra support at least in part a generally neutral RT PCR model, with only experiment 3236b that cannot be fitted to a purely neutral model. Given the steep gradient of observed abundance decline with rank in experiment 3236b, we believe that this might be due to multiple bottlenecks of the same magnitude, which our model cannot reproduce, because it only includes one strong bottleneck. Multiple bottlenecks of equal strength will leave different signatures in the rank-abundance graphs. Our model does not capture these bottlenecks and can hence not reproduce these steep gradients.

We see the pID protocol as an attractive tool with many advantages over conventional sequencing. When one step fails in the preparatory upstream protocol, conventional NGS will reveal only a homogenous population due to only a select number of RNA/cDNA being captured in the process. The pID protocol would have revealed an extremely small number of pIDs and can therefore serve as important quality control tool. In addition, the pID protocol shines when it comes to low-frequency variants [27,28]. Whereas NGS generally does not allow for detecting SNVs below the average sequencing error rate of around ~1% (unless mutations are cooccurring within reads [29]), the pID protocol pushes this detection threshold down to the RT error rate, which is more than 1 order of magnitude lower than the average sequencing error rate. Furthermore, the pID protocol does not require sophisticated (and imperfect) haplotype reconstructions algorithms [30]. Finally, the pID protocol allows for determining the largest bottleneck of the experimental protocol. In the case of a nonnested PCR for amplification of the input material, this bottleneck will likely be the RT, such that the number of tagged RNA molecules can be estimated. We could not estimate the number of RNA molecules because we used a nested PCR, where the strongest bottleneck is the subsampling step between the two stages of the PCR.

While we see the pID protocol as an attractive tool for studying heterogeneous HIV and other viral intrahost populations, the protocol is not without its associated drawbacks. First, whole-genome sequencing beyond average amplicon lengths is not suitable and will require different protocols for shotgun sequencing, such as BAsE-Seq [31], with their associated strengths and weaknesses. While the pID protocol obviates the requirement for haplotype reconstruction tools for single-amplicon sequencing, phasing multiple amplicons will require some computational algorithms, as the pID protocol cannot remove RT errors and some PCR errors in nested PCR setups. Second, the pID protocol has the disadvantage that pooling/multiplexing of multiple samples per sequencing experiment is problematic due to the required high PCR resampling redundancy [16]. The inherent sensitivity–redundancy tradeoff of the pID protocol requires considerable redundancy should detection sensitivity of low-frequency variants be desired. This might be problematic in diagnostic setups, when pooling of patient samples for economic reasons requires a strong reduction in the number of reads per patient, with an equal or overproportionate reduction in the number of input RNA molecules [16], which is contrary to the desired sensitivity of the protocol. Third, using the very sensitive HaploClique to infer the haplotypes above 1% and then piling up the raw reads on these will deliver unbiased estimators with comparable variance, avoiding the pID protocol altogether. Given that we observe an overall neutral process, this is not surprising, as pID consensus sequences then only represent a finite subsampling step without replacement, which in itself will not improve estimator variances [32].

In conclusion, we found the pID protocol to be a powerful quality control tool that can aid in determining whether the library preparation step has failed. It should be noted that our conclusions are based only on the five-virus mix and its frequency ranges from 6% to 38%. Analyses of more realistic virus populations are required to generalize, refine, or falsify our findings. We believe the pID protocol to be a valuable tool for inferring SNVs and haplotypes below the sequencing error rate when a short single fragment (amplicon) is of interest.

Materials and Methods

RT, amplification, and NGS of HIV-1 RNA using oligonucleotides containing pIDs

A well-characterized five-virus mix (Fig. S1) comprising five different HIV-1 virus strains was used [5,18] and ~ 10^6 HIV-1 RNA copies of it were isolated using the NucleoSpin® RNA Virus Kit (Macherey-Nagel) according to the manufacturer's protocol, including a DNase treatment on the column with 30 U DNase (DNase I recombinant, RNase-free; Roche). RNA was eluted in 25 μl water. RT was performed with 10 μl of RNA (approximately 400,000 copies) and 1 μM of the primer RT_J_ID pol 3236 or RT_J_ID pol 3223 (Table S1), with the two primers differing only in the HIV-1-specific sequences. RNA plus oligonucleotides were incubated at 80 °C for 5 min followed by cooling at 4 °C for 2 min. cDNA synthesis was performed using SuperScript® III RT (Invitrogen) according to the manufacturer's protocol. The cDNA was treated with RNase H (New England Biolabs, Bioconcept) following the manufacturer's instructions and then purified with the NucleoSpin® Gel and PCR Clean-Up Kit (Macherey-Nagel) according to the manufacturer's description for single-stranded DNA clean-up. One-twelfth of purified cDNA (approximately 32,000 copies) were used for the subsequent PCR. PCRs were performed in a total volume of 20 μl: 94 °C-2′, 30 cycles of 94 °C-15″, 55 °C-30″, 68 °C-60″ containing 0.4 mM dNTPs (Fermentas), 1.5 U Platinum® *Taq* DNA Polymerase High Fidelity (Invitrogen), and 0.2 μM of each forward oligonucleotide pol 2316 (5′-GCTCTATTA GATACAGGAGCAG-3′; nucleotides 2316–2337 based on HIV-1_{HXB2}, GenBank accession number K03455) and reverse oligonucleotide 5′-GCCTTGCCAGCACGCT-CAGGC-3′, of which the latter is similar to the 5′ part of the oligonucleotides used for RT. PCR products were purified with the NucleoSpin® Gel and PCR Clean-Up Kit (Macherey-Nagel) according to the manufacturer's description and quantified by qPCR. Approximately 10,000 DNA copies were used for a second PCR with oligonucleotides specific for Illumina amplicon sequencing. Pools of four oligonucleotides per forward and reverse direction were used, forward 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG$N_{0\text{-}3}$ TACAATACTCCAGTATTTGCC-3′ and reverse 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA G$N_{0\text{-}3}$CCAGCACGCTCAGGCCTTGCA-3′ creating amplicons of variable length (±0–6 bp) to guarantee an even distribution of all four bases in subsequent MiSeq® sequencing cycles. PCR was performed as described above, except that 35 cycles were performed. Each setup was performed in three independent experiments: the samples were tagged with oligonucleotides harboring different indices and Illumina-specific sequences. PCR conditions were chosen according to the manufacturer's description (Illumina), and the samples were sequenced using a MiSeq® Benchtop Sequencer generating paired-end reads of 2 × 250 bp length (v2 kit). The complete protocol is shown in Fig. S2.

RT qPCRs to test efficiencies of oligonucleotides containing pIDs

A dilution of the HIV-$1_{NL4\text{-}3}$ virus stock was used to isolate HIV-1 RNA from approximately 8.5×10^7, 8.5×10^6, and 8.5×10^5 HIV-1 RNA copies. The RNA was treated three times with DNase (DNase I recombinant, RNase-free; Roche), that is, prior, during (on column), and after RNA isolation to minimize the amount of remaining plasmid used for transfection of 293T cells to generate the virus stock. Isolated RNA was diluted 1:10 and reverse transcribed using the SuperScript III RT (Invitrogen) according to the manufacturer's protocol (input per RT reaction: approximately 4.25×10^6, 4.25×10^5, and 4.25×10^4 HIV-1 RNA copies, respectively). RT was performed using different oligonucleotides (Table S1) with or without pID sequence. The cDNA was quantified by qPCR as previously described [18], and the oligonucleotide pair is given in Table S1.

NGS preprocessing

Data from all NGS runs were analyzed with FastQC in order to determine the overall quality of the lane[†]. PRINSEQ [33] was used to clip bases with a Phred score below 30 in a sliding window of size 5 from both 5′ and 3′ ends of the reads. Only reads that had a minimum length of 230 after clipping were retained. Finally, reads with a missing mate after preprocessing were also discarded (Fig. S9).

Alignment of NGS data

To prevent erroneous, low-quality pIDs from contaminating the subsequent downstream analysis, we replaced internal bases of reads with Phred scores below 30 by the ambiguous base “N” using fastq_masker from the FASTX Toolkit[‡]. For alignment of the NGS data, a custom alignment software, named pIDalign, on the basis of the SeqAn [34] library was developed. The global-to-local, or glocal, alignment algorithm is based on the Needleman-Wunsch algorithm but does not penalize overhangs. For performance reasons, a consensus sequence with ambiguous bases was constructed from the reference five-virus mix. Alignment of the six datasets was stored in the SAM sequence alignment format [35]. Aligned reads were removed if their alignment scores dropped below 210, or included a consecutive deletion of length 20 or more, or the template length was less than 520 or the read is a mate of a read with a failed alignment. Template length is defined as per the SAM specification, that is, the distance on the reference genome of the leftmost mapped base of the left read to the rightmost mapped base of the right read. The minimum alignment score of 210 ensures a sufficient number of matches between the read and the reference sequence.

Preprocessing

All analysis was performed with a custom analysis pipeline referred to as pIDalyse[§]. First, the homozygous and heterozygous loci of the RT reading frame under study of the five reference HIV-1 strains were determined. By a homozygous locus, we mean a locus where all five clones have the same base, and by a heterozygous locus, we mean a locus where at least two clones have different bases (Fig. S11). Determining these homozygous loci and heterozygous loci is required for discerning recombination between clones later on.

Second, aligned read mates of one fragment were paired by their read identifiers for all six SAM alignments. Reads that have a missing mate were discarded; otherwise, they were merged into one complete read with a gap between the mates. From all properly paired reads, the distribution of pID lengths was determined. Third, reads were filtered to retain only those that have a pID of exactly 10 nt. In addition, reads were only retained if the length of the non-HIV-specific overhang beyond the pID was at least 23 nt long, that is, if the read included the full 3′ PCR primer.

Removing collisions

In order to determine the most likely pool of sequences after the RT step, we proceeded to call the consensus sequence. Determining the consensus sequence was performed by a majority vote of a minimum 80% of bases at one position of an ensemble of PCR replicates of one pID. In order to not call spurious consensus sequences, at least 10 reads per pID collection were required, where a pID collection is defined as all the PCR molecules produced from one original cDNA template. Additionally, we used the following two-step heuristic to remove collisions from doubly tagged RNA molecules in the RT step with the same pIDs.

In order to classify sequences from one pID collection, we used a k-means clustering with k = 2. We defined the first cluster to be composed of all those reads originating from the cDNA of the RT. The second cluster contained all those sequences that were assigned to this pID by either mutations in the pID from a different one via PCR substitutions or errors in the NGS step. In order to initialize the cluster centers, which are defined to be the consensus sequences by majority vote, we defined a graph in which nodes represent reads of one pID collection. We defined an edge between two nodes (or reads) if the number of mismatches between their heterozygous loci was at most one. We determined the connected components of the graph and initialized the first cluster center with the consensus sequence of the largest connected component. The second cluster center was initialized with the consensus sequence of the second largest connected component. We then performed the k-means clustering algorithm for five iterations, assigning reads to the cluster center with the least number of mismatches in the heterozygous loci and calling the cluster center by majority vote. Five rounds provided ample convergence to the final clustering. A collision was called if the first and largest cluster contained less than 80% of the reads of a pID collection. An indecisive case exists when the largest cluster contained more than 80% of the reads, but the absolute number of reads in this cluster was less than 10. In both cases, the pID with its reads was discarded; otherwise, the first cluster center was taken as the final consensus sequence. Reads from the second cluster were discarded as likely not originating from the original cDNA. Consensus sequences were assigned to a clone if there were no mismatches between heterozygous loci. If no error-free assignment to any of the heterozygous loci of the clones was possible, the pID consensus sequence was not used for frequency estimation. This pileup of consensus sequences to clones eventually allows for estimating the relative proportion, or frequency, of each clone in the experiment. Unless otherwise stated, all further analysis involves the collision-removed data.

Estimating the PCR substitution rate

The substitution rate of the *Taq* DNA polymerase was estimated by determining the fraction of mutants that likely occurred in the first cycle of the PCR branching process (Fig. S12). Focusing only on homozygous loci, substitutions occurring in the first cycle of the PCR will show up with a frequency of roughly 50%. A substitution in the second cycle will result in a frequency of about 25%, yet the probability of a substitution occurring in the second cycle is twice as likely as in the first cycle, given that one and only one substitution occurred in the PCR branching process. In all of these calculations, we assumed the PCR process to be deterministic in the starting cycles; that is, the number of molecules doubles exactly between successive cycles. We probabilistically counted the number of prospective PCR substitutions having occurred in the first cycle in all homozygous loci of all pID collections and added these up to arrive at the total number of substitutions distributed approximately according to a binomial distribution. Dividing this number by the total number of inspected homozygous loci over all pIDs provides an estimator of the PCR substitution rate (see supplementary information, section 2a).

Estimating the RT substitution rate

In order to estimate the substitution rate of the reverse transcriptase, we concentrated on the homozygous loci of the five virus clones. Assuming the amplification of the plasmids in the bacteria to be error free, all substitutions at the homozygous loci of the consensus sequence of one pID collection are likely due to either substitutions by the reverse transcriptase or substitutions by the *Taq* DNA polymerase in the first PCR of the nested PCR (Fig. S14). Due to the small number of molecules sampled in the first PCR for the second round, any PCR error will likely show up in all reads, that is, as a conserved substitution, at a homozygous locus. We counted the number of substitutions and the total number of bases in all homozygous loci of collision-free consensus pID sequences. The fraction of mutant counts divided by the total number of homozygous bases yielded the maximum-likelihood estimate for the RT substitution rate. This estimate was then corrected by subtracting various factors that could inflate the error rate, such as the aforementioned PCR substitution rate, the RNA polymerase II substitution rate of the 293T cells, and the inherent bacterial mutation rate (supplementary information, section 2b).

Estimating the combined RT PCR recombination rate

We first assessed whether considerable PCR recombination artifacts are present in the data, by assigning filtered raw reads to any of the five clones with a minimum of mismatches in the heterozygous loci (supplementary information, section 1g). An essential impediment in estimating the recombination rate of the RT and PCR is the unobserved process of the reverse transcriptase and the *Taq* DNA polymerase. Due to the large number of homozygous loci, the exact base at which either enzyme switches the template or prematurely terminates cannot be determined exactly. As such, we require a latent model that explicitly averages over all recombination paths. We have developed a hidden Markov model, where the (unknown) Markov chain models the movement of the RT and the first PCR along the RNA/DNA molecule with the possibility of switching templates between or pausing after every base.

We determined the combined RT PCR recombination rate by maximum-likelihood estimation with numerical methods. For the substitution rate of this process, that is, the emission probabilities, we used the estimated parameter from the previous subsection. CI values were determined by suitably inverting the log-likelihood ratio test (supplementary information, section 2c).

Assessing potential biases of pIDs

To estimate whether a strong bias exists in the distribution of pIDs, we devised a number of summary statistics. After the consensus calling procedure, we stored each pID and its associated number of PCR replicates in order to determine the position-wise nucleotide distribution.

We analyzed position-wise nucleotide frequencies of the sequenced pIDs to determine whether there exist any biases. In addition, to exclude higher-order bias effects, we also analyzed the dinucleotide distributions of neighboring positions in the pIDs. Beyond position-wise dinucleotide analysis, we analyzed the data to find evidence for extreme selective effects of single pIDs. For each dataset, we picked the top 500 pID sequences of the most abundant pID collections. For both primers RT_J_ID pol 3223 and RT_J_ID pol 3236, we determined the intersection between replicates and generated Venn diagrams. Finally, we combined all experiments in one Venn diagram for the union of all pID experiments.

As a goodness-of-fit test for neutrality of all pIDs, we sorted the abundances of the pIDs by the size of their PCR collection and plotted them as a function of their rank. We determined the goodness of fit by simulating a model of unbiased (neutral) random pID sampling in the RT and stochastic PCR amplification followed by a sequencing sample without replacement from the amplified PCR pool that we termed the PCR bottleneck model. With this model, we determined the size of the strongest bottleneck and estimated the efficiency of the PCR procedure (supplementary information, section 3b).

Bias and variance of population estimators

We assessed whether the pID protocol significantly improves the frequency estimators of the clonal proportions in comparison to HaploClique [10], a tool for correcting spurious errors from NGS data to infer viral haplotypes and their frequencies. To this end, we aligned the data from the NGS preprocessing step using pIDalign. After alignment, we employed AmpliconClipper[||] to remove the PCR primers, including the pID. The clipped alignment was transformed back to raw paired-end FastQ data using Picard tools[¶]. These emulated raw data, without the pID and primer segments, were aligned using bwa [36]. Unpaired reads and improperly paired reads were removed using SAMtools [35] and the remaining aligned reads were saved to a bam file. The pipeline leading to the final alignment represents a typical haplotype reconstruction workflow (left branch; Fig. S9). Finally, this alignment was used as input to HaploClique, which we ran for 30 iterative cycles to produce final error-corrected haplotypes. To exclude spurious errors introduced in the RT and early cycles of the PCR, we filtered out reconstructed haplotypes with a frequency of less than 1% and renormalized the remaining fractions. Other population frequency estimators using first raw reads assigned to the five clones and then using the pID consensus sequences were calculated by normalizing their counts as determined by pIDalyse[§].

To compare bias and variance of the haplotype frequency estimator based on the pID protocol to the other population estimators, we devised a statistical test referred to as the likelihood ratio Dirichlet test (supplementary information, section 4). We first tested whether a bias between any of the methods exists. In a second step, we tested whether the statistical difference in estimator variances is significant.

Availability

The dataset consisting of six lanes of the Illumina 2× 250-bp protocol has been uploaded to the National Center for Biotechnology Information sequence read archive (accession number SRP060688). All software and scripts for producing the statistics and plots are available with pIDalyse[§].

Acknowledgments

The HIV-1 full-length plasmids were obtained through the National Institutes of Health AIDS Research and Reference Reagent Program, Division of AIDS, National Institute of Allergy and Infectious Diseases, National Institutes of Health: pYK-JRCSF from Irvin S. Y. Chen and Yoshio Koyanagi, pNL4–3 from Malcolm Martin, pYU2 from Beatrice Hahn and George M. Shaw, and p89.6 from Ronald G. Collman. pHXB2 was kindly provided by Marek Fischer. This work was supported by ETH research grant ETH-33 13-1 (to N.B.) and by the Swiss National Science Foundation under grant CR32I2_146331 (to N.B., K.J.M., and H.F.G.).

Conflict of Interest: H.F.G. has been an adviser and/or consultant for the following companies: GlaxoSmithKline, Abbott, Gilead, Novartis, Boehringer Ingelheim, Roche, Tibotec, Pfizer, and Bristol-Myers Squibb and has received unrestricted research and educational grants from Roche, Abbott, Bristol-Myers Squibb, Gilead, Astra-Zeneca, GlaxoSmithKline, and Merck Sharp & Dohme (all money went to institution). K.J.M. received travel grants and honoraria from Gilead Sciences, Roche Diagnostics, Tibotec, Bristol-Myers Squibb, and Abbott; the University of Zurich has received research grants from Gilead, Roche, and Merck Sharp & Dohme for studies that K.J.M. serves as principal investigator and advisory board honoraria from Gilead Sciences. All other authors declare no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2015.12.012.

Received 7 August 2015;
Received in revised form 25 November 2015;
Accepted 16 December 2015
Available online 19 December 2015

Keywords:
HIV-1;
next-generation sequencing;
primer ID;
PCR;
RT

This work has been partly presented at the 22nd Conference on Retroviruses and Opportunistic Infections, February 23–26, 2015, Seattle, USA (abstract #255).

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

†http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

‡http://hannonlab.cshl.edu/fastx_toolkit/.

§https://github.com/cbg-ethz/PrimerID.

||https://github.com/SoapZA/AmpliconClipper.

¶http://picard.sourceforge.net.

Abbreviations used:
pID, primer ID; RT, reverse transcription; NGS, next-generation sequencing; CI, confidence interval; qPCR, quantitative PCR.

References

[1] T. Lengauer, N. Pfeifer, R. Kaiser, Personalized HIV therapy to control drug resistance, Drug Discov. Today Technol. 11 (2014) 57–64.

[2] P. Rieder, B. Joos, A.U. Scherrer, H. Kuster, D. Braun, C. Grube, et al., Characterization of human immunodeficiency virus type 1 (HIV-1) diversity and tropism in 145 patients with primary HIV-1 infection, Clin. Infect. Dis. 53 (2011) 1271–1279.

[3] N. Beerenwinkel, H.F. Gunthard, V. Roth, K.J. Metzner, Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data, Front. Microbiol. 3 (2012) 329.

[4] A. Cozzi-Lepri, M. Noguera-Julian, F. Di Giallonardo, R. Schuurman, M. Daumer, S. Aitken, et al., Low-frequency drug-resistant HIV-1 and risk of virological failure to first-line NNRTI-based ART: A multicohort European case-control study using centralized ultrasensitive 454 pyrosequencing, J. Antimicrob. Chemother. 70 (2015) 930–940.

[5] F.D. Giallonardo, A. Töpfer, M. Rey, S. Prabhakaran, Y. Duport, C. Leemann, et al., Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations, Nucleic Acids Res. 42 (2014) e115.

[6] M.A. Quail, M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, et al., A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, BMC Genomics 13 (2012) 341.

[7] M. Allhoff, A. Schonhuth, M. Martin, I.G. Costa, S. Rahmann, T. Marschall, Discovering motifs that induce sequencing errors, BMC Bioinformatics 14 (2013) S1.

[8] O. Zagordi, A. Bhattacharya, N. Eriksson, N. Beerenwinkel, ShoRAH: Estimating the genetic diversity of a mixed sample from next-generation sequencing data, BMC bioinformatics 12 (2011) 119.

[9] S. Prabhakaran, M. Rey, O. Zagordi, N. Beerenwinkel, V. Roth, HIV haplotype inference using a propagating Dirichlet process mixture model, IEEE/ACM Trans. Comput. Biol. Bioinform. 11 (2014) 182–191.

[10] A. Töpfer, T. Marschall, R.A. Bull, F. Luciani, A. Schonhuth, N. Beerenwinkel, Viral quasispecies assembly via maximal clique enumeration, PLoS Comput. Biol. 10 (2014) e1003515.

[11] M. Shugay, O.V. Britanova, E.M. Merzlyak, M.A. Turchaninova, I.Z. Mamedov, T.R. Tuganbaev, et al., Towards error-free profiling of immune repertoires, Nat. Methods 11 (2014) 653–655.

[12] I. Kinde, J. Wu, N. Papadopoulos, K.W. Kinzler, B. Vogelstein, Detection and quantification of rare mutations with massively parallel sequencing, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 9530–9535.

[13] H.E. Bhang, D.A. Ruddy, V. Krishnamurthy Radhakrishna, J.X. Caushi, R. Zhao, M.M. Hims, et al., Studying clonal dynamics in response to cancer therapy using high-complexity barcoding, Nat. Med. 21 (2015) 440–448.

[14] C.B. Jabara, C.D. Jones, J. Roach, J.A. Anderson, R. Swanstrom, Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 20166–20171.

[15] D.J. Sheward, B. Murrell, C. Williamson, Degenerate primer IDs and the birthday problem, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) E1330.

[16] R.H. Liang, T. Mo, W. Dong, G.Q. Lee, L.C. Swenson, R.M. McCloskey, et al., Theoretical and experimental assessment of degenerate primer tagging in ultra-deep applications of next-generation sequencing, Nucleic Acids Res. 42 (2014) e98.

[17] J. Brodin, C. Hedskog, A. Heddini, E. Benard, R.A. Neher, M. Mild, et al., Challenges with using primer IDs to improve accuracy of next generation sequencing, PLoS One 10 (2015) e0119123.

[18] F. Di Giallonardo, O. Zagordi, Y. Duport, C. Leemann, B. Joos, M. Kunzli-Gontarczyk, et al., Next-generation sequencing of HIV-1 RNA genomes: Determination of error rates and minimizing artificial recombination, PLoS One 8 (2013) e74249.

[19] K.R. Tindall, T.A. Kunkel, Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase, Biochemistry 27 (1988) 6008–6013.

[20] J. Potter, W. Zheng, J. Lee, Thermal stability and cDNA synthesis capability of SuperScript III reverse transcriptase, Focus 25 (2003) 19–24.

[21] M. Negroni, M. Ricchetti, P. Nouvel, H. Buc, Homologous recombination promoted by reverse transcriptase during copying of two distinct RNA templates, Proc. Natl. Acad. Sci. U. S. A. 92 (1995) 6971–6975.

[22] M.G. Ross, C. Russ, M. Costello, A. Hollinger, N.J. Lennon, R. Hegarty, et al., Characterizing and measuring bias in sequence data, Genome Biol. 14 (2013) R51.

[23] R.C. Cadwell, G.F. Joyce, Randomization of genes by PCR mutagenesis, PCR Methods Appl. 2 (1992) 28–33.

[24] D.J. Lahr, L.A. Katz, Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase, Biotechniques 47 (2009) 857–866.

[25] J. Liu, H. Song, D. Liu, T. Zuo, F. Lu, H. Zhuang, et al., Extensive recombination due to heteroduplexes generates large amounts of artificial gene fragments during PCR, PLoS One 9 (2014) e106658.

[26] P. McInerney, P. Adams, M.Z. Hadi, Error rate comparison during polymerase chain reaction by DNA polymerase, Mol. Biol. Int. 2014 (2014) 287430.

[27] J.R. Keys, S. Zhou, J.A. Anderson, J.J. Eron Jr., L.A. Rackoff, C. Jabara, et al., Primer ID informs next-generation sequencing platforms and reveals preexisting drug resistance mutations in the HIV-1 reverse transcriptase coding domain, AIDS Res. Hum. Retrovir. 31 (2015) 658–668.

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

[29] K. McElroy, O. Zagordi, R. Bull, F. Luciani, N. Beerenwinkel, Accurate single nucleotide variant detection in viral populations by combining probabilistic clustering with a statistical test of strand bias, BMC Genomics 14 (2013) 501.

[30] M.C. Prosperi, L. Yin, D.J. Nolan, A.D. Lowe, M.M. Goodenow, M. Salemi, Empirical validation of viral quasispecies assembly algorithms: State-of-the-art and challenges, Sci. Rep. 3 (2013) 2837.

[31] L.Z. Hong, S. Hong, H.T. Wong, P.P. Aw, Y. Cheng, A. Wilm, et al., BAsE-Seq: A method for obtaining long viral haplotypes from short sequence reads, Genome Biol. 15 (2014) 517.

[32] E. Venrick, The statistics of subsampling, Limnol. Oceanogr. 16 (1971) 811–818.

[33] R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets, Bioinformatics 27 (2011) 863–864.

[34] A. Doring, D. Weese, T. Rausch, K. Reinert, SeqAn an efficient, generic C++ library for sequence analysis, BMC Bioinformatics 9 (2008) 11.

[35] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al., The Sequence Alignment/Map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.

[36] H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform, Bioinformatics 26 (2010) 589–595.