Chapter 3

Primer ID Algorithm PIDA – Algorithm for processing Ultra-Deep High Throughput Sequence Data generated using Primer ID technology

# Introduction

Ultra Deep PyroSequencing (UDPS) platforms are capable of generating as much as millions of sequence reads from DNA fragments at low cost and in less time than other sequencing approaches. This capability enables the potential to fully characterize viral quasispecies including the low frequency variants (Fischer et al., 2010; Hoffmann et al., 2007; Mitsuya et al., 2008; Rozera et al., 2009; Varghese et al., 2009; Wang et al., 2007) However, a high rate of sequencing errors are incorporated and accumulated at the PCR amplification step (Kanagawa, 2003) and by instrumental/hardware error and sequencing errors like nucleotide insertion and deletion errors (reviewed in (Metzker, 2009)). These errors inflate and confound the real genetic diversity in the population (Kunin et al., 2009; Zagordi et al.). The errors generated at PCR step are: **1)** incorporation of wrong nucleotide by polymerase enzyme during many cycles of amplification (Hughes and Totten, 2003; Kanagawa, 2003) **2)** recombination of two DNA fragments producing a new chimeric DNA (Judo et al., 1998; Meyerhans et al., 1990; Yang et al., 1996) **3)** differential amplification of DNA fragments change the ratio before and after PCR step, obscuring true original sample diversity (Liu et al., 1996; Polz and Cavanaugh, 1998). In order to avoid modification and over diversification of the original sample and to ensure downstream results are truly reflective of the actual viral diversity, it is essential to correct those inevitable errors.

The accurate quantification of low abundance drug resistant HIV viruses, in particular, may be substantially improved by the implementation of the primer ID approach as described by Jabara and colleagues (Jabara et al., 2011). Primer ID is a unique identifier that is annealed to a viral cDNA during the reverse transcription step from RNA to DNA to enable identification of the viral template from which each individual sequence reads originated. A number of studies have already used this approach for such purposes (Beerenwinkel et al., 2012; Eisele and Siliciano, 2012; Jabara et al., 2011; Schmitt et al., 2012). While the original publication describes the development of an algorithm to analyze the complex data output from primer ID-based sequencing (Jabara et al., 2011), this code has not been made available to the public and is unlikely to be made so in the near future (Cassandra B. Jabara, personal communication). Thus, to facilitate the analysis of HIV drug resistance sequence data generated using the primer ID approach in the Seq2Res resistance testing computational pipeline, this chapter describes the development and application of a such a tool.

# Methods and Materials

## Raw sequence reads containing primer ID

The structure of a raw sequence read containing a primer ID is shown in Figure 3. A set of primer ID sequences of custom length (usually 8) is generated randomly. The number of primer IDs depends on the length of primer ID. A set of primer ID of length 8 has 65536 (48) unique nucleotide combinations. A single primer ID is embedded within the primer used in the cDNA synthesis. The set of primer ID and primers creates random library of sequences. A cDNA primer binds to a viral RNA and extends from 3’ end to generate a cDNA that now contains a primer ID tag. A unique multiplex identifier (MID) sequence per sample with a spacer sequence at both ends of the MID is prepared. The cDNA with primer ID at 5’ end is then attached to the 3’ end of the MID with spacer sequences. A PCR priming site sequence is then added to the 5’ end.

The cDNA sequence with primer ID, MID and PCR priming site is then PCR amplified to produce millions of sequences. The primer ID is copied through the PCR steps thus enabling tracking of the viral template from which each sequence derived. During PCR amplification a forward read is extended through the target sequence followed by the primer ID, MID and PCR priming site. In reverse sequences, however, the primer is extended the target sequence is the last part to be extended following the primer ID, MID and PCR priming site. Correct assignment of forward sequences requires that the full fragment has been amplified and sequenced in order to cover the primer ID sequence at the end of the read (**Figure 3.1 B**) however reverse sequences do not need to be full length in order to assign the read to a particular template on the basis of the primer ID (**Figure 3.1 C**).

## Processing primer ID data using PIDA

A novel algorithm - Primer ID algorithm (PIDA), was developed for integration into the Seq2Res pipeline to facilitate fast and accurate processing of sequence reads generated using the primer ID approach. The algorithm requires the raw sequence reads input in FASTQ format while other required files contain information about the primers used, the multiplex identifiers used (if present) and the minimum allowed read lengths.

The primer file is a five column tab delimited file containing the amplicon name in the first column followed by the forward and reverse primer sequences in the second and third column while the fourth and fifth columns contain start and end nucleotide positions set by first nucleotide position of forward primer and last nucleotide position of reverse primer relative to the standard HIV *pol* reference sequence. (**Figure 3.2 A**).

In instances when multiple samples have been sequenced together on the same sequencing plate, each sample is tagged with a unique MID sequence. In order to interpret these, the user must provide a tab-delimited file with the MID name in the first column and a unique patient identifier in the second one (**Figure 3.2 B**). When the standard Roche MIDs are not used then the MID number can be replaced with the MID sequence.

In some instances the end-user may only be interested in subsequent analysis of a short amplicon fragment located within an amplicon thereby enabling non full-length sequences to be analysed. Thus, we allow the user to define the minimum read length required for both the forward and reverse sequences for each amplicon. The gene file details these lengths with the amplicon name in the first column followed by the forward and reverse sequence minimum read lengths in columns two and three respectively (**Figure 3.2 C**).

The other information that the end-users are required to supply are the universal PCR primer sequence and the format of the sequence containing primer ID, spacers, MID and PCR primer that was prepared for cDNA production. For example, a user may input the format as primerid8.cg.mid5.tga.primingsite, which indicates a primer ID of length 8 nucleotides, a spacer sequences ‘cg’, MID sequence of length 5 nucleotides, another spacer sequence ‘tga’ and followed by the word ‘primingsite’. Users also have options to choose:

1. Threshold number of sequences required to generate consensus sequence
2. Maximum mismatches allowed between a user supplied primer and primer region in a sequence read. This is defined as primer tolerance.
3. Maximum mismatches allowed between a user-supplied MID and the sequence in the MID region of a sequence read. This is defined as MID tolerance.

The steps of processing the raw data into consensus sequences in the algorithm are discussed in detail below:

### Sequence Demultiplex using tag sequences

For each sequence read, the information provided in primer file and MID file is used to identify the amplicon and sample to which the read belongs (**Figure 3.3**, red text). The algorithm begins with a search for the forward primer at 5’ end of the read. A subsequence of length equal to forward primer is obtained from forward primer region at 5’ end, which is then pair-wise aligned with all forward primers one at a time. If the number of mismatches in pair-wise alignment is less or equal to, the primer tolerance, the sequence read amplicon is designated as being identified with the aligned forward primer and the search for reverse primer is skipped. If none of the forward primers are identified within the read the algorithm searches for the presence of each of the reverse primers. A subsequence of length equal to the reverse primer is obtained from the reverse primer region and is pair-wise aligned with every reverse primer one at a time. If a reverse primer that is aligned with the number of mismatches less or equal to primer tolerance, the sequence read amplicon is identified with the aligned reverse primer. All sequences in which a reverse primer is identified are reverse complemented to ensure all subsequent analysis is performed on sequences in the same strand orientation. If neither forward nor reverse primers are found, sequence reads are discarded.

Following identification of a sequence read’s source amplicon, PIDA identifies the MID associated with that read and bins all reads with the same primer and MID together for downstream analysis.

A subsequence from MID region of a sequence read is extracted and pair-wise aligned with list of supplied MID sequences one at a time. If the number of mismatches between a MID and the subsequence is less or equal to defined tolerance, the sample for the associated MID is identified for the sequence read or discarded if none of the MIDs match with the obtained MID from the sequence.

Once a matching MID is found, a sequence of ‘n’ nucleotides is obtained as an primer ID sequence, where ‘n’ is the length of primer ID, from the sequence read region at the 5’ end of MID and spacer sequence. The read is discarded if there is a presence of an ambiguous base in the obtained ‘n’ nucleotides. The tags - Primer, MID and primer ID - are then added at sequence id for further downstream processing.

### Selection of sequences with threshold length

With sequence reads containing the primer ID sequence information at the 5’ end (Figure 3.1C) the entire target sequence plus the downstream information must be sequenced in order for the important MID and primer ID information to be identified and the sequence read retained for further analysis. Conversely, when the primer ID information is contained at the 3’ end, the entire target sequence does not need to be sequenced to identify the read based upon the primer ID and MID motifs (Figure 3.1B). Therefore, if the entire query sequence is not required for downstream analysis, the user can set a parameter to define the minimum length of query sequence that must sequenced to retain a non full-length reverse read. (**Figure 3.3** green text).

### Selection of Primer IDs with minimum number of sequences

The default number of representative sequences required for a single primer ID to generate a consensus sequence is three as recommended in the original publication to avoid ambiguous bases (Jabara et al., 2011). However, end-users have the option to set this value as required. Only the primer IDs with a number of representative sequences greater than, or equal to, the defined threshold number are passed through for subsequent analysis (**Figure 3.3** blue text).

### Quality trimming

Quality trimming is an optional step in the algorithm. If the user selects the trimming option, the algorithm uses QTrim (Shrestha et al., 2014) to quality trim the non-discarded sequence reads. As quality trimming can change the sequence length once it is complete the PIDA algorithm repeats the previous two steps of the process to ensure that the quality trimmed reads are of a sufficient length and quantity for consensus sequence generation (**Figure 3.3** black text).

### Generating a Consensus Sequence

For each primer ID the representative sequences are binned together (**Figure 3.3** purple text) and are aligned to each other using MAFFT (Katoh et al., 2005; Katoh et al., 2002; Katoh and Toh, 2008, 2010). A consensus sequence is generated from the resulting alignment by calling the most common nucleotide that at each position in the alignment **(Figure 3.3** Sky blue text). In the case of ties in frequency between two or more bases at a position, an ambiguous base representing the bases is added to the consensus sequence. To avoid large numbers of ambiguous bases in a consensus sequence users can define an odd number as the minimum number of sequences.

### Test datasets

Two datasets (described here as Run1 and Run2) were generated by our collaborators (Prof Carolyn Williamson’s research group, University of Cape Town) using the Primer ID approach and were used here to evaluate PIDA. Each dataset comprised one sequencing run (Roche/454 Junior plate) containing data from four HIV infected patients from a study to analyze vaccine response in HIV subtype C (the results from this study do not comprise part of this thesis and will be published elsewhere). For each patient, four amplicons covering one region in the envelope gene (*env)*, two regions of gag (gag54, *gag*472) and one region in the nef gene (*nef*23) were amplified using subtype C specific primers with each primer used in the cDNA generation step associated with a unique primer ID. Amplicons for each patient were tagged using a unique MID and sequence data was generated using one Roche/454 Junior system for each dataset.

## Results

### Initial demultiplexing

Runs 1 and Run2 were analyzed independently of each other using PIDA. Before any processing of the data was undertaken the total number of raw sequence reads for each dataset was 125,865 and 40,544 for Runs 1 and 2 respectively.

The initial demultiplexing step involved identifying sequence reads that had incomplete/missing primer or MID motifs and discarding them from subsequent analysis. For Run 1 the primer was not found in 2962 (2.4%) reads while 7557 (6%) had no MID present. Similarly, for Run 2 the primer was not found in 751 (1.9%) of reads while 1109 (2.74%) had no MID present (**Table 3.1**).

For the remaining reads that contained complete primers and MIDs, we examined the sequence of the primer ID and retained reads that did not contain ambiguous bases at primer ID region. There were 257 (0.2%) and 37 (0.1%) reads with ambiguous bases in primer ID in Run1 and Run2 respectively and were discarded (**Table 3.1**)

Thus, following initial demultiplexing, 91% of sequences from Run1 were retained, while 95% of the sequences from Run2 were passed through for subsequent analysis.

### Quality trimming of sequence data and sequence length evaluation

All sequence reads were quality trimmed using QTrim (Shrestha et al., 2014) with a mean quality score of 20, resulting in a loss of 2950 (2.34%) and 14 (0.034%) sequences from Run1 and Run2 respectively as a result of poor quality. For forward reads the MID and primer ID sequences are located at the 3’ end and, thus, the sequenced amplicons must be full length in order to extract all necessary information for downstream analysis. For reverse reads, however, the required information is located at the 5’ end and, thus, full-length sequence reads are not always necessary as the region of interest may be close to the 5’ end. Therefore, the required read lengths to define reads that could be passed through to the next stage of analysis (i.e. they cover the region of interest in the query sequence) were different for forward and reverse sequence reads (Table 3.2)

Using these parameters forward and reverse sequence read lengths were analyzed independently of each other and those sequences that were shorter than the required read length were removed from the analysis. 32 (0.03%) and 23 (0.06%) sequences from Runs1 and Run2 respectively were discarded as short reads in the trimming step (**Table 3.3**).

For all subsequent processing and analysis, sequence reads from each amplicon and patient were binned together resulting in 16 unique datasets generated from each sequencing run (**Table 3.4**). Each of these datasets was subsequently analyzed independently of all others.

### Characterization of primer IDs

For each dataset, the number of unique primer ID tags contained in the data was counted. A wide range of unique primer ID tags was observed between the various datasets ranging from 110 for the gag472 amplicon in patient E to 4193 for the *env* amplicon in patient B (**Table 3.5**)

For each dataset the number of sequence reads tagged with each unique primer ID in that dataset were grouped together and counted. We observed that the number of representative sequences for each primer ID tag ranged from one to 4144 (**Figure 3.4-3.7**). The minimum number of sequences required to generate a consensus sequence representing a primer ID tag is three and, thus, for each dataset we separated primer ID tags with two or less sequences from those with three or more representative sequences. We observed that the percentage of primer ID tags with less than three representative sequences was, on average, three times higher than those with three or more representative sequences (**Table 3.6**).

While only three sequences are required to generate a consensus sequence, we find that for each dataset the average number of sequence reads for each primer ID is significantly greater than three (**Table 3.7**). Patient G contained an average of six sequences per primer ID for each amplicon with the remainder of patients averaging between 18 and 88 representative sequences per primer ID.

### Generation of consensus sequences

Consensus sequences were only generated for those primer ID tags with three or more representative sequences (**Figure 3.8**). Generation of consensus sequences showed that while there may be a large number of sequences representing a particular amplicon for a patient (**Figure 3.8 A and B**), this could comprise data representing a small number of primer IDs meaning that the resulting number of consensus sequences was, in fact, quite low (**Figure 3.8 C and D**). For example, Patient A’s *env* amplicon in Run1 had the highest total number of sequence reads (19700, **Figure 3.8A**) but the number of primer IDs that they represent was only 241 (**Figure 3.8C**). On the other hand, Patient B’s *env* amplicon had a lower number of reads (**Figure 3.8A**) but had more than six times the number of primer ID tags than patient A had for the same amplicon (**Figure 3.8C**).

## Discussion

Since the introduction of Roche/454 and other ultra deep pyrosequencing technologies, the sequencing of the full spectrum of the diversity of the HIV quasispecies within a HIV infected individual has become possible (Abbate et al., 2011; Beerenwinkel and Zagordi, 2011; Mild et al., 2011; Rozera et al., 2009; Zagordi et al., 2010). However, there are a number of potential biases that must be considered in the experimental protocol and subsequent data analysis.

Current pyrosequencing technologies require a high volume of input DNA and therefore the sample DNA need to be PCR amplified prior to sequencing. The PCR amplification step can introduce various errors into the amplified data including 1) in-vitro recombination of the viral DNA (Judo et al., 1998; Meyerhans et al., 1990; Yang et al., 1996), 2) misincorporation of a base at one or more positions in the growing strand (Hughes and Totten, 2003; Kanagawa, 2003), 3) differential amplification of two different viral templates thereby artificially inflating the prevalence of one viral variant relative to the others (Liu et al., 1996; Polz and Cavanaugh, 1998), and 4) resampling of the DNA sample (Eckert and Kunkel, 1991; Kanagawa, 2003; Liu et al., 1996). These PCR errors introduce allelic skewing, artificial homogeneity, artificial diversity, and inflate genetic variation (Kanagawa, 2003).

Jabara et al (Jabara et al., 2011) introduced a novel strategy that tags each cDNA-generating primer with a unique primer identifier (primer ID) sequence thereby allowing the identification of sequence reads that originate from the same viral template. As sequences with the same primer ID indicate that they originate from the same template viral RNA sequence, any difference in one sequence relative to all others from the same template can be attributed to PCR or sequencing error and can be removed by the generation of a consensus sequence from all of the sequences representing that primer ID. The technology can be anticipated to be widely accepted in future UDPS involving highly heterogeneous population sample (Sheward et al., 2012). Because the technology developers and publishers have not made the code available for public use, the PIDA algorithm has been developed to facilitate the analysis of HIV drug resistance genotyping data generated using the primer ID approach.

Primer ID is an elegant strategy to reduce errors in the final data for analysis. The primer ID is added to the cDNA primer for reverse transcription process, not to the viral RNA. Therefore primer ID tag cannot address the errors introduced in to the cDNA in the reverse transcription (Boyer et al., 1992). The primer ID that is added to the cDNA is copied through the PCR amplification and ultra deep sequencing protocols. The PCR amplification may introduce errors in every cycle of DNA synthesis. Although primer ID strategy can reduce PCR errors, the errors introduced in the first cycle of PCR amplification cannot be addressed due to lack of enough amplification of the template sequences (Horton, 1995). On total, the mutations from reverse transcription and first cycle of PCR amplification can range from 0.01 to 0.1%. This would mean that there are unavoidable errors in the experimental analysis that include reverse transcription and PCR in the experimental protocols. The primer ID strategy also fails to address these errors. Despite this, a number of studies have shown that using the primer ID approach significantly reduces the artificial generated in a sequence dataset through PCR and sequencing induced errors. With and colleagues showingFurther

We counted the number of primer IDs by pooling the sequences with same primer ID in to a bin. There were 19558 and 8694 primer IDs in Run1 and Run2 respectively. The number of representative sequences for each primer ID ranged from one to 4144 in Run1 and one to 1119 in Run2, while Jabara et al reported the small range of one to 96 representative sequences per primer ID (Jabara et al., 2011). The variation in number of sequences reads in each primer ID bin was over thousand folds (Figure 3.4 – 3.7), while Jabara et al reported only 100 folds (Jabara et al., 2011). The median sequence reads per primer ID in Run1 and Run2 was three, while Jabara et al (Jabara et al., 2011) reported the median of six. The wide range in the number of representative sequence per primer ID is clearly the poor feature of PCR as indicated by Jabara et al (Jabara et al., 2011). Polz and Cavanough also indicated PCR has amplification bias (Polz and Cavanaugh, 1998).

### Lost of HIV variants due to PCR artifacts

In our study, primers IDs with one or two representative sequence reads were discarded as the number of sequence reads is below the threshold required for consensus sequence generation. This resulted in a total of 73% and 93% of primer IDs in Run1 and Run2 respectively being discarded. This loss was much higher than 38% that was observed by Jabara et al (Jabara et al., 2011). A primer ID tags a single viral cDNA and the removal of primer IDs with one or two representative sequences would mean likely removal of true HIV variants from the dataset. This would mean that the sequence final dataset probably under represents the true diversity of the viral variants present in the viral quasispecies of that HIV-infected individual.. Therefore, while the primer ID approach undoubtedly reduces artificial diversity by removing many of the PCR and sequencing-induced errors from a dataset, it appears to be too stringent resulting in an underrepresentation of the true viral diversity. Thus the primer ID approach should be further optimized and extended to to improve its accuracy.

For example, Schmitt and colleagues have developed an approach to further correct any form of sequencing errors including the errors in reverse transcription and early PCR cycle errors by tagging both strands of duplex DNA (Schmitt et al., 2012). This approach compares one strand with the other strand of the DNA for error correction (Schmitt et al., 2012). The authors showed that the error could be reduced to less than one per billion bases sequenced. However, the approach has not been applied at UDPS of HIV quasispecies with primer ID strategy. While this doesn’t fix the problems in relation to underrepresentation of the data as a result of low numbers of sequences for some primer IDs, it is certainly a step in the right direction.

The large numbers of primer IDs with one or two representative sequence reads perhaps indicate the poor feature of PCR protocol. A study by Polz and Cavanough (Polz and Cavanaugh, 1998) also showed that it is the PCR artifact to considerably and reproducibly over amplify specific template DNA. Their study also showed that GC rich priming site amplified with higher efficiency indicating that primer binding energies might play role in over amplification (Polz and Cavanaugh, 1998). These might be the possible reasons for the huge difference in the amplification of the viral templates. Christopherson et al (Christopherson et al., 1997) suggested that the amplification efficiency could be improved up to 10 fold by adjusting the PCR annealing temperature and gradually increasing the temperature during cDNA synthesis. All these suggest that the PCR protocol needs to be improved for better amplification of DNA templates to avoid the lost of viral depth.

### Loss of HIV variants due to Primer ID collision

Sheward et al (Sheward et al., 2012) suggested that the generation of truly unique primer IDs is extremely unlikely. The generation of only 10,000 primer IDs of 8 bases length, could generate ~726 duplicated primer IDs called primer ID collision. The result of primer ID collision could tag multiple templates with the same random primer ID. Of the two HIV variants tagged with same primer ID, the variant with low representative sequences gets lost during consensus sequence generation. The authors calculated that 30 of 2000 consensus sequences would be the result of primer ID collision. We also observed over 1000 primer IDs collision in our study (Table 3.8). This could lose viral variants depth in the experimental sample.

## Conclusion

Primer ID is a novel technology for correcting PCR and sequencing errors in UDPS data. Primer ID error correction requires at least three representative sequences for a primer ID to generate consensus sequence. Therefore the PCR protocol needs to be optimized for templates amplification. The improved PCR protocol for the high number of template amplification would increase the generation of a consensus sequence per primer ID. The high number of consensus sequences increases the depth of HIV variants in the analysis.

We have developed an algorithm called PIDA for the analysis of raw datasets generated using this technology. We have tested the algorithm in two real primer ID raw sequence read data sets and the comparative results were observed in both the datasets. PIDA algorithm showed that primer ID technology has ability to reduce errors through consensus sequence generation.

Bibliography

.

Abbate, I, Vlassi, C, Rozera, G, Bruselles, A, Bartolini, B, Giombini, E, Corpolongo, A, D'Offizi, G, Narciso, P, Desideri, A, Ippolito, G, Capobianchi, MR (2011) Detection of quasispecies variants predicted to use CXCR4 by ultra-deep pyrosequencing during early HIV infection. *AIDS* **25**: 611-617.

Beerenwinkel, N, Gunthard, HF, Roth, V, Metzner, KJ (2012) Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front Microbiol* **3**: 329.

Beerenwinkel, N, Zagordi, O (2011) Ultra-deep sequencing for the analysis of viral populations. *Curr Opin Virol* **1**: 413-418.

Boyer, JC, Bebenek, K, Kunkel, TA (1992) Unequal human immunodeficiency virus type 1 reverse transcriptase error rates with RNA and DNA templates. *Proc Natl Acad Sci U S A* **89**: 6919-6923.

Christopherson, C, Sninsky, J, Kwok, S (1997) The effects of internal primer-template mismatches on RT-PCR: HIV-1 model studies. *Nucleic Acids Res* **25**: 654-658.

Eckert, KA, Kunkel, TA (1991) DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl* **1**: 17-24.

Eisele, E, Siliciano, RF (2012) Redefining the viral reservoirs that prevent HIV-1 eradication. *Immunity* **37**: 377-388.

Fischer, W, Ganusov, VV, Giorgi, EE, Hraber, PT, Keele, BF, Leitner, T, Han, CS, Gleasner, CD, Green, L, Lo, CC, Nag, A, Wallstrom, TC, Wang, S, McMichael, AJ, Haynes, BF, Hahn, BH, Perelson, AS, Borrow, P, Shaw, GM, Bhattacharya, T, Korber, BT (2010) Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS One* **5**: e12303.

Hoffmann, C, Minkah, N, Leipzig, J, Wang, G, Arens, MQ, Tebas, P, Bushman, FD (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* **35**: e91.

Horton, RM (1995) PCR-mediated recombination and mutagenesis. SOEing together tailor-made genes. *Mol Biotechnol* **3**: 93-99.

Hughes, JP, Totten, P (2003) Estimating the accuracy of polymerase chain reaction-based tests using endpoint dilution. *Biometrics* **59**: 505-511.

Jabara, CB, Jones, CD, Roach, J, Anderson, JA, Swanstrom, R (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc Natl Acad Sci U S A* **108**: 20166-20171.

Judo, MS, Wedel, AB, Wilson, C (1998) Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Res* **26**: 1819-1825.

Kanagawa, T (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J Biosci Bioeng* **96**: 317-323.

Katoh, K, Kuma, K, Toh, H, Miyata, T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**: 511-518.

Katoh, K, Misawa, K, Kuma, K, Miyata, T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059-3066.

Katoh, K, Toh, H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* **9**: 286-298.

Katoh, K, Toh, H (2010) Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* **26**: 1899-1900.

Kinde, I, Wu, J, Papadopoulos, N, Kinzler, KW, Vogelstein, B (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A* **108**: 9530-9535.

Kunin, V, Engelbrektson, A, Ochman, H, Hugenholtz, P (2009) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **12**: 118-123.

Liu, SL, Rodrigo, AG, Shankarappa, R, Learn, GH, Hsu, L, Davidov, O, Zhao, LP, Mullins, JI (1996) HIV quasispecies and resampling. *Science* **273**: 415-416.

Metzker, ML (2009) Sequencing technologies — the next generation. *Nature Reviews Genetics* **11**: 31-46.

Meyerhans, A, Vartanian, JP, Wain-Hobson, S (1990) DNA recombination during PCR. *Nucleic Acids Res* **18**: 1687-1691.

Mild, M, Hedskog, C, Jernberg, J, Albert, J (2011) Performance of ultra-deep pyrosequencing in analysis of HIV-1 pol gene variation. *PLoS One* **6**: e22741.

Mitsuya, Y, Varghese, V, Wang, C, Liu, TF, Holmes, SP, Jayakumar, P, Gharizadeh, B, Ronaghi, M, Klein, D, Fessel, WJ, Shafer, RW (2008) Minority human immunodeficiency virus type 1 variants in antiretroviral-naive persons with reverse transcriptase codon 215 revertant mutations. *J Virol* **82**: 10747-10755.

Polz, MF, Cavanaugh, CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol* **64**: 3724-3730.

Preston, BD, Poiesz, BJ, Loeb, LA (1988) Fidelity of HIV-1 reverse transcriptase. *Science* **242**: 1168-1171.

Roberts, JD, Bebenek, K, Kunkel, TA (1988) The accuracy of reverse transcriptase from HIV-1. *Science* **242**: 1171-1173.

Rozera, G, Abbate, I, Bruselles, A, Vlassi, C, D'Offizi, G, Narciso, P, Chillemi, G, Prosperi, M, Ippolito, G, Capobianchi, MR (2009) Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations. *Retrovirology* **6**: 15.

Schmitt, MW, Kennedy, SR, Salk, JJ, Fox, EJ, Hiatt, JB, Loeb, LA (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* **109**: 14508-14513.

Sheward, DJ, Murrell, B, Williamson, C (2012) Degenerate Primer IDs and the birthday problem. *Proc Natl Acad Sci U S A* **109**: E1330; author reply E1331.

Shrestha, RK, Lubinsky, B, Bansode, VB, Moinz, MB, McCormack, GP, Travers, SA (2014) QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform. *BMC Bioinformatics* **15**: 33.

Varghese, V, Shahriar, R, Rhee, SY, Liu, T, Simen, BB, Egholm, M, Hanczaruk, B, Blake, LA, Gharizadeh, B, Babrzadeh, F, Bachmann, MH, Fessel, WJ, Shafer, RW (2009) Minority variants associated with transmitted and acquired HIV-1 nonnucleoside reverse transcriptase inhibitor resistance: implications for the use of second-generation nonnucleoside reverse transcriptase inhibitors. *J Acquir Immune Defic Syndr* **52**: 309-315.

Wang, C, Mitsuya, Y, Gharizadeh, B, Ronaghi, M, Shafer, RW (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res* **17**: 1195-1201.

Yang, YL, Wang, G, Dorman, K, Kaplan, AH (1996) Long polymerase chain reaction amplification of heterogeneous HIV type 1 templates produces recombination at a relatively high frequency. *AIDS Res Hum Retroviruses* **12**: 303-306.

Zagordi, O, Klein, R, Daumer, M, Beerenwinkel, N (2010) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res* **38**: 7400-7409.