# Chapter 3

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

### 3.1 Introduction

High throughput sequencing (HTS) 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 sequence viral quasispecies and enables the characterization of low frequency variants from highly heterogeneous viral population samples (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, high rate of sequencing errors are incorporated and accumulated at 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, 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. 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 (Cas 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.

**3.2 Methods and Materials**

**3.2.1 Primer ID raw sequence reads.**

The structure of Primer ID raw sequence reads is determined before cDNA generation from viral RNA genome. A universal primer sequence and a customary spacer sequence of any length is prepared to which a MID is attached to the 3’ end. The MID sequence varies between the samples. A different spacer sequence is then attached to the 3’ end of the MID sequence. For every sample, a set of Primer IDs 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 Primer ID from a set for every sample is then attached to the 3’ end of the sequence prepared above. A cDNA primer is then attached to the 3’ end of the Primer ID in the prepared sequence. Because, theoretically, each Primer ID is a unique sequence, it can be used as a tag sequence of the cDNA primer. The cDNA primer binds to viral RNA and extends from 3’ end to generate a cDNA (**Figure 3.1 A**).

PCR amplification of the cDNA follows the cDNA production step. The PCR primer and the amplicon forward primer bind to the cDNA to produce millions of sequences. The forward primer may or may not extend to the end of PCR primer region covering the tags – Primer ID and MID (**Figure 3.1 B**) while PCR primer usually extend to cover the tags Primer ID and MID including the reverse primer (**Figure 3.1 C**). The reverse primer is similar to the cDNA primer used in cDNA production step.

**3.2.1 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 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:

**3.2.1 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 below the primer tolerance, the sequence read amplicon is identified with the aligned forward primer and the search for reverse primer is skipped. If none of the forward primers are identified in the read the algorithm searches the read for the presence of each of the reverse primers. Similarly to forward primer, a subsequence of length equal to reverse primer is obtained from is the reverse primer region and is pair-wise aligned with all reverse primers one at a time. The 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 MID is searched using the supplied known format of the Primer ID, spacers, MID and PCR Priming site. If the nucleotide pattern occurs as known format in a sequence, the MID sequence is obtained and pair-wise aligned with list of supplied MID sequences one at a time. The MID that matches the obtained MID with less or equal to defined tolerance in the pair-wise alignment is used to identify the sequence read sample or discarded if none of the supplied MID matches 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’ 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.

**3.2.2 Selection of sequences with threshold length**

In the raw sequence data generated with Primer ID technology, the sequence with forward primer in the 5’ end should be full length to cover the known pattern of information blocks at the 3’ end. On the other hand, the sequences that start with information blocks at 5’ end, followed by reverse primer can end anywhere and, thus, the user has the option to filter the reverse sequence with defined sequence length supplied in gene file.

The sequences are, then, filtered with their length. Forward sequences lower than full length are discarded whereas reverse sequences lower than user-defined length are discarded (**Figure 3.3** green text).

**3.2.3 Selection of Primer IDs with minimum number of sequences**

PIDA process all the sequences under analysis to get the number of sequences per Primer ID per amplicon per sample. The Primer ID representing below the threshold number of sequences is discarded along with the sequences that it represents. The default threshold number of sequences is 3 and that is the recommended (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).

**3.2.4 Quality trimming**

Quality trimming is an optional step in the algorithm. If the user selects the trimming option, the algorithm uses QTrim 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).

**3.2.5 Generating a Consensus Sequence**

Sequences are binned together in a file with the same primer, MID and Primer ID tags ( 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 nucleotide that occurs the most at each position in the alignment **(Figure 3.3** Sky blue text). In 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 number of ties, users can choose an odd number as a minimum number of sequences. This is a reason that the default minimum number of sequences for consensus sequence generation is three.

**3.2.6 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 plate for each dataset.

### 3.3 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. This resulted in the loss of 257 (0.2%) and 37 (0.1%) reads for Run1 and Run2 respectively (**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 primer ID data and checking of sequence length**

Data was quality trimmed, which resulted in loss of 2950 (2.34%) and 14 (0.034%) sequences from Run1 and Run2 respectively. Then, the every resulting high-quality read was assessed for length. In the forward reads, the MID and Primer ID had to be at the 3’ end and therefore, the amplicon had to be sequenced completely. However, in reverse reads, the tags were at the 5’ region and therefore, these sequences had to have read length that covers the interested region (**Table 3.2**).

Furthermore, forward and reverse sequence read lengths were checked independently and those sequences that were shorter than the read length of interest were removed from the analysis, which resulted in the loss of 32 (0.03%) and 23 (0.06%) sequences for Runs1 and Run2 respectively (**Table 3.3**).

For all subsequent processing and analysis, sequence reads from each amplicon and patient were binned together (**Table 3.4**) and each group of sequences was analyzed independently of all others.

**Independent analysis of every amplicon in each patient – trimmed data.**

Sequences in each group were further binned by Primer ID tag. The total number of groups, each represented by a Primer ID, was inconsistent between amplicons and patients with wide range in both Run1 and Run2 (**Table 3.5**)

Primer IDs with representing similar number of sequences were grouped and were counted together. It showed that large number of the Primer IDs had very few sequences in the most amplicons while very few Primer IDs had large number of sequences in few amplicons in both Run1 and Run2 (**Figure 3.4 – 3.7**).

Each Primer ID representing group was further split into two groups: one group contains the number of Primer IDs representing less than three sequences and the other group containing the number of Primer IDs representing three or more sequences. The result showed that the percentage of Primer IDs representing less than three sequences was on average three times higher than the percentage of Primer IDs representing three or more sequences (**Table 3.6**).

The average number of reads in each group of Primer IDs with less than three reads and three or more reads were obtained. The Primer IDs in the group with less than three sequences had in average between one and two sequences, which it should be while the Primer IDs in the group with three or more sequences had in wide range of average sequences from four to 229 reads per Primer ID (**Table 3.7**).

**Generation of consensus sequences**

All the sequence reads containing a Primer ID that represent less than three sequences were not considered for further analysis as a consensus sequence could not be generated from the Primer IDs. The removal of these sequences from further analysis resulted into lost of large number of Primer IDs (**Table 3.6**). The other group of Primer IDs representing three or more sequences was used for generating a consensus sequence per Primer ID.

A consensus sequence per Primer ID was generated representing the most frequent nucleotide in the given base position of the sequences with same Primer ID or the ambiguous base of the nucleotides if there was tie between them.

Generation of the consensus sequences showed that few Primer IDs could represent a large number of sequences (**Figure 3.8**). For example, Patient A *env* amplicon in Run1 (**Figure 3.8 A**) has the highest number of reads (19700 reads) represented by Primer IDs with three or more sequences but the number of Primer IDs that represent those sequences was only 241. On the other hand, Patient B *env* amplicon has lower number of reads than Patient A for same amplicon but has the number of Primer IDs over six times.

The average number of reads per consensus sequences was obtained, which showed that there was inconsistently wide distribution of sequences between the consensus sequences (**Table 3.7**). Furthermore, from the average number of reads per consensus sequence, the next average was obtained across the amplicon in every patient and across patients in every amplicon (**Table 3.7**). This showed that the number of consensus sequences generated was inconsistent across the amplicons as well as across the patients as shown by the wide range of average number of sequences per Primer ID (**Table 3.7**).

### 3.4 Discussion

Primer ID is a novel technology to tag every cDNA primer that generates a cDNA from a viral RNA and to accurately quantify the viral variants in the HIV quasispecies. *PCR and sequencing errors in the sequences can be resolved using multiple reads over a given base* (Jabara et al., 2011). As sequences with the same Primer ID in a sample originate from the same template viral RNA sequence, an error in one sequence relative to all others from the same template can be attributed to PCR or sequencing error and will be removed by the generation of a consensus sequence from all sequences with the same Primer ID. Therefore, the technology can be anticipated to be widely accepted in future HTS involving highly heterogeneous population sample (Sheward et al., 2012). Because the technology developers and publishers have no intention of releasing the code for public use, the PIDA algorithm has been developed.

The PIDA algorithm was tested on two real Primer ID raw sequence data sets. The algorithm was able to demultiplex the reads with Primer ID along with MID and amplicon primers. However, about 9% and 5% of the sequence reads were discarded in Run1 and Run2 respectively. Most of the sequences were discarded for the reason that MID was not found. MID appears at the proximity of 3’ end of the forward sequences and the quality of the sequences drops at the 3’ end. Therefore, the sequence reads discarded was mostly attributed to sequencing error. Other sequences that were discarded for the reason that a primer was not found (primers appear at 5’ end) were attributed to PCR artifacts.

The MID and Primer ID tags appear at 3’ end of forward sequences and 5’ end of reverse sequences. Forward sequences, and not reverse sequences, had to be full amplicon in order to cover Primer ID and MID tags. Reverse sequence length had to be greater or equal to the read length that had to cover the region of interest in the amplicon. Very few sequences were discarded in both Run1 and Run2 data sets; all were reverse sequences and none were forward sequences. The reason that none were forward sequences was because forward sequences are mostly discarded due to missing MID tags. If MID was present, the forward sequence would cover full amplicon and not be discarded for being short in length.

However, a large number of quality-trimmed reads (18336 reads in Run1 and 7411 reads in Run2) were discarded in both data sets as their Primer ID represented the number of reads less than three (**Figure 3.8**). In Run1, 7542 Primer IDs represented only one sequence and 3030 Primer IDs represented two sequences after demultiplexing step (**Figure 3.8**). The percentage of those Primer IDs representing one or two sequences was on average three times the percentage of Primer IDs representing three or more sequences (**Table 3.6**). Similar result was obtained for Run2. This observation correlates with the result publication by (Jabara et al., 2011) who showed that about 30% and 8% (total 38%) of sequences in population were lost as large number of Primer IDs represented only one sequence and two sequences respectively. The lost of such a huge number of sequences and/or Primer IDs would mean that the final analysis under represent the true diversity of viral variants (Sheward et al., 2012). However, it was not the Primer ID limitation and Jabara et al were not sure about the reason why the higher number of Primer IDs representing single reads were obtained but it could be the PCR bias in amplification of viral variants in abundance (Storhoff et al., 2004).

Consensus sequence generation required at least two sequences therefore, the Primer IDs that represent one sequence cannot generate a consensus sequences while Primer IDs that represent two sequences might generate a consensus sequence with large number of ambiguous bases in it as an ambiguous base is added if there was a tie between two bases at a base position. This was the reason that the minimum number of sequences for consensus generation should be three.

The consensus sequences generated from Primer IDs representing three or more sequences showed that the highest number of sequences did not produce the highest number of consensus sequences. It was because there was a wide difference in the number of sequences represented by each Primer ID (**Figure 3.4 – 3.7**). Our observation showed that for higher average number of sequences per consensus sequence in an amplicon per patient, lesser the consensus sequences were obtained (**Figure 3.8**).

The wide differences in the average number of sequences per consensus sequence across patients and amplicons showed that the amplification of sequences in PCR per Primer ID appears to be random.

The ability of Primer ID technology depended on multiple sequences from same source sequence. Therefore, although, the technology could resolve PCR and sequencing errors, it was unable to resolve the nucleotide misincorporation errors during cDNA synthesis and first cycle of PCR as they had not been amplified. These errors go undetected during the analysis. It was, therefore, cannot be assured that a consensus sequences generated from a Primer ID representing three or more sequences was 100% exactly an original viral variant RNA sequence.

A limitation of Primer ID technology could be the duplication of Primer ID sequences during generation of a set of Primer IDs at random. A mathematical proof of birthday problem (McKinley, 1966) showed that the number of people required such that the chance of two people having same birthday by at least 50% is 23. It is therefore very clear that there was a high chance of duplication of many Primer IDs when generating a set of 65536 Primer IDs (for Primer ID of length 8 nucleotide) randomly (Sheward et al., 2012). Analysis of the two real data sets also showed that there was up to 1000-fold duplication of Primer IDs (**Table 3.8**). If two or more duplicated Primer IDs tagged the same amplicon region in same sample but from different viral variants, the variant represented by the most sequences would get represented in the generated consensus sequence losing the other variants. Therefore, the duplication of Primer IDs could also be under representing the true viral diversity (Sheward et al., 2012)**.**

### 3.5 Conclusion

Primer ID is a novel technology for correcting sequencing errors in HTS data. 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 set and the comparative results were observed in both the datasets. PIDA algorithm shows that the use of Primer ID can be under representing the true variants in the heterogeneous population.

Bibliography

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.

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.

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.

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.

McKinley, EH (1966) Generalized birthday problem. *The American Mathematical Monthly* **73**: 385 - 387.

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.

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.

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.

Storhoff, JJ, Lucas, AD, Garimella, V, Bao, YP, Muller, UR (2004) Homogeneous detection of unamplified genomic DNA sequences based on colorimetric scatter of gold nanoparticle probes. *Nat Biotechnol* **22**: 883-887.

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.