# 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 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).

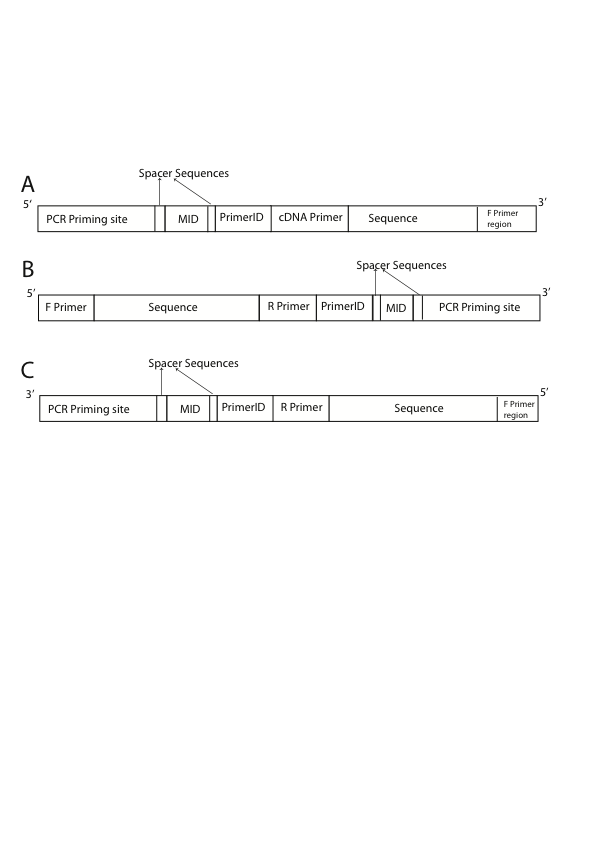


Figure 3.1: The structure of Primer ID raw sequence data. A) cDNA structure with locations of PCR priming site, spacer sequences, MID and Primer ID. B) A forward strand Primer ID raw sequence read produced in PCR step. C) A reverse sequence produced in the PCR step. The primers tags the amplicon region which is being amplified, MID tags the specific individual from whom the sample is obtained and Primer ID tags the template RNA sequence. Two spacer sequences are used to separate the MID sequence from Primer ID and PCR priming site. The nucleotide sequence and length of spacers are customary and may vary between the projects.

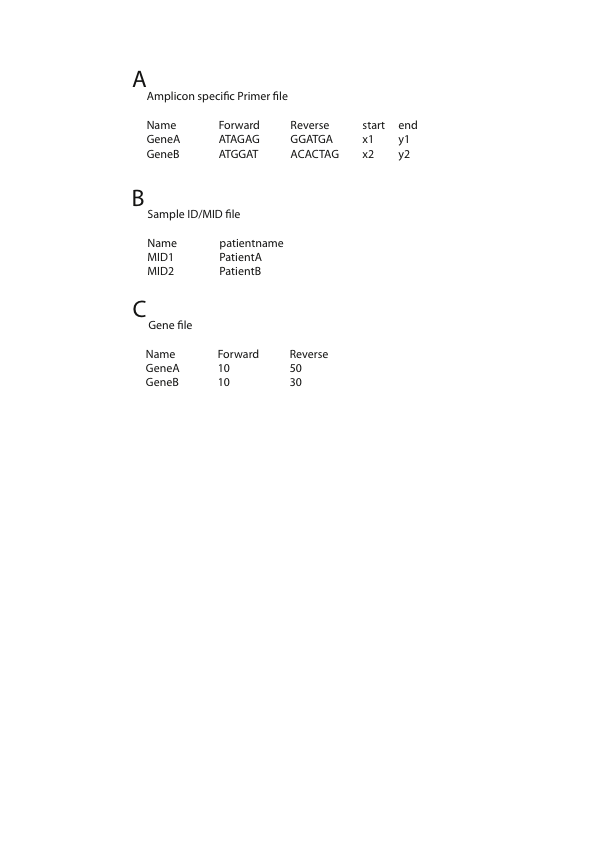


Figure 3.2:Different files required as input for the PIDA algorithm. Each column in all files has to be separated by a single tab. A) The amplicon specific primer file contains gene names that are amplified; the forward and reverse primers used for each amplicon; and the start and end nucleotide positions of the targeted amplicon region relative to standard HIV *pol* reference sequence. B) The MID file contains names of MID sequence used to tag the samples and sample names for identification. C) The gene file contains the amplicon names and minimum required sequence length for forward and reverse sequences. The information displayed in the files are just for the purpose and can be changed as required.

Note: x1, y1, x2 and y2 can be replaced with numbers that represent the amplicon start and end positions.

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.

Seq2Res_primerid_pipeline_flow.pdf

Figure 3.3: The logical flow of the PIDA algorithm to process raw sequence data generated with Primer ID technology into generation of consensus sequence. The colors denote PIDA processing steps. Red: sequence demultiplexing, Green: Sequence filter by length, Blue: sequence filter by number of sequence represent by each Primer ID, Black: Quality trimming, Purple: Sequencing binning, Sky blue: Consensus sequence generation

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 3.

**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

Runs 1 and Run2 were analyzed independently of each other using PIDA. The raw Primer ID sequence reads in Run1 and Run2 are 125,865 and 40,544 respectively 115057 and 38624 raw reads were demultiplexed with the obtained primer, MID and Primer ID information from each read for Run1 and Run2 respectively, as the other raw reads were discarded for reasons like – no primer was found, no MID was found, an ambiguous bases in the Primer ID sequence and the reads were shorter than threshold length (Table 3.1).

When the demultiplexed reads were grouped by unique Primer ID per amplicon per sample, 18011 and 7397 reads were discarded as their Primer ID represented below the threshold number of sequences (Table 3.2). The rest reads were quality trimmed using QTrim, which discarded very few sequences (2950 and 14 reads in Run1 and Run2 respectively) and that showed that the reads in downstream analysis were of high quality (Table 3.2).

Discarding the reads in quality trimming lowered the number of sequence reads represented by Primer IDs per amplicon per sample. After the quality trimmed reads were grouped by Primer IDs per amplicon per sample, only 365 reads for Run 1 and two reads for Run2 were discarded as their Primer ID represents below the threshold number of sequences (Table 3.2).

The final number of sequence reads that were remained before the generation of consensus sequences was 93731 (74.47%) for Run1 and 31321 (77.25%) for Run2 of

Table 3.1: Analysis of raw sequence data before accounting Primer ID for downstream analysis.

|  |  |  |
| --- | --- | --- |
|  | Run1 | Run2 |
| Total raw reads | 125865 | 40544 |
| Total read discarded for no primer | 2962 | 751 |
| Total reads discarded for no MID | 7557 | 1109 |
| Total reads discarded for ambiguous base in Primer ID | 257 | 37 |
| Total reads discarded for short length | 32 | 23 |
| Total reads after demultiplexing | 115057 | 38624 |

Table 3.2: Analysis of the number of demultiplexed sequence reads with Primer ID into generation of consensus sequences

|  |  |  |
| --- | --- | --- |
|  | Run1 | Run2 |
| Total reads discarded because their Primer ID had less than 3 representative sequences | 18011 | 7397 |
| Total reads after discarding Primer ID with less than 3 representative sequences | 97046 | 31337 |
| Total reads after quality trimming | 94096 | 31323 |
| Total reads after filtering by sequence length and Primer ID with less than 3 representative sequences | 93731 | 31321 |
| Total consensus sequence generated | 4704 | 2751 |

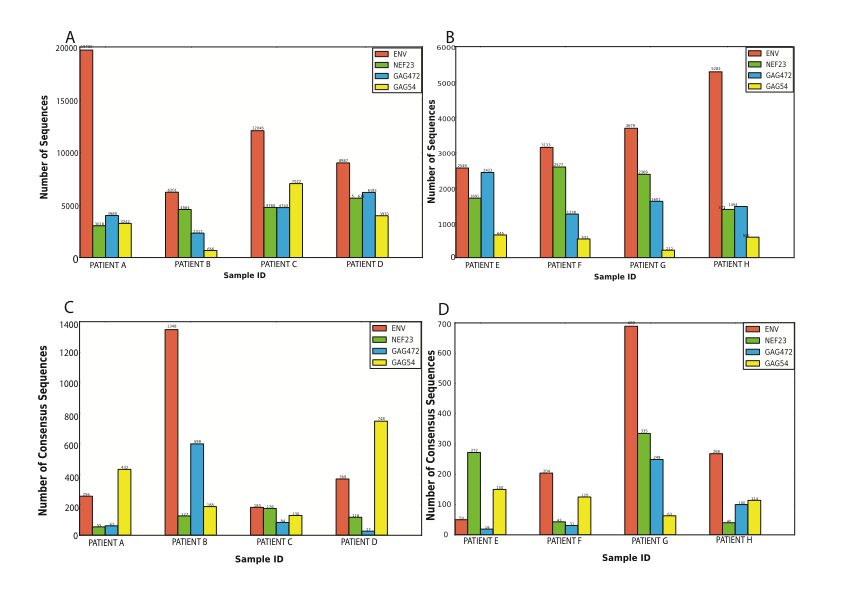
****

Figure 3.4: The distribution of final total number of sequences, each represented by Primer ID with greater or equal to threshold number of sequence, per amplicon per sample before consensus sequence generation in (A) Run1 and (B) Run2. The total number of consensus sequences generated per amplicon per sample in (C) Run1 and (D) Run2. Each consensus sequence is generated by collapsing the sequences represented by a unique Primer ID in an amplicon of a sample.

the total raw reads in Run1 and Run2 respectively. The distribution of these sequences reads per amplicon per sample shows that there was a huge variation in the number of sequence reads per amplicon per sample in both data sets, ranging from 666 -19700 reads between the amplicon in Run1 and 212 to 5283 reads between the amplicon in Run2 (Figure 3.4 A and B).

The sequence reads per Primer ID per amplicon per sample were collapsed to a single consensus sequence and the distribution of consensus sequences per amplicon per sample showed that the number of consensus sequences were not proportional to the number of sequences per amplicon per sample before consensus sequences generation in both Run1 and Run2 data sets (Figure 3.4 C and D). An amplicon with high number sequences may produce low number of consensus sequences whereas an amplicon with low number of sequences may produce high number of consensus sequences. For example, Run1 patient A *env*  amplicon had the highest number of sequences (19700) (Figure 3.4 A) that produced only 256 consensus sequences (Figure 3.4 C) while patient B *env*  amplicon had around one-third sequences but produces over five times the consensus sequences than patient A *env* amplicon. The range in number of consensus sequences generated in Run1 was 27 – 1348 and in Run2 was 19 – 609 (Figure 3.4 C and D).

### 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, a large number of demultiplexed 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 3.5). In Run1, 7542 Primer IDs represented only one sequence and 3030 Primer IDs represented two sequences after demultiplexing step (Figure 3.5). 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 their Primer ID represented only one sequence and two sequences respectively. The lost of such a huge number of sequences 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 I believe it was the PCR bias in amplification of viral variants in abundance (Storhoff et al., 2004). This characteristic of PCR changed the ratio of viral variants in original sample and Primer

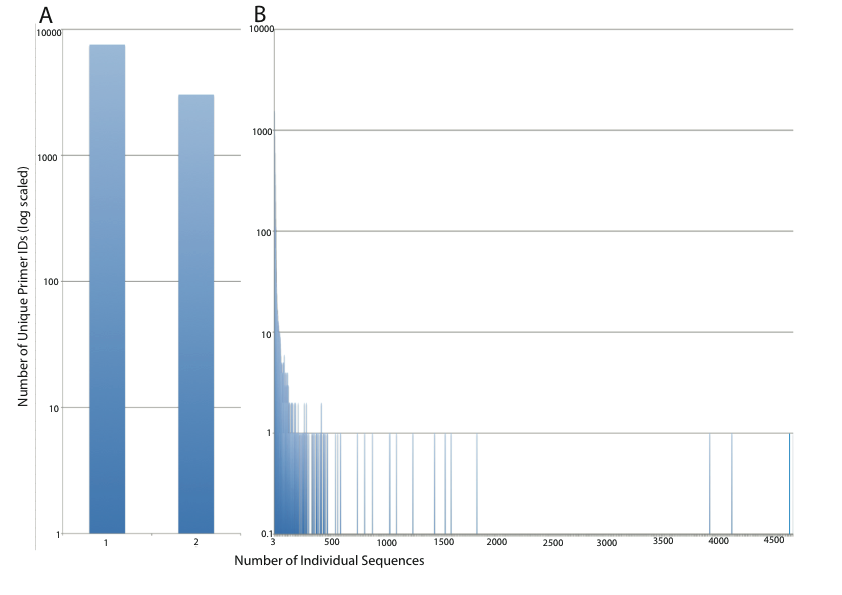


Figure 3.5: Distribution of number of unique Primer IDs by number of individual sequences represented in Run1 data set. A) Number of unique Primer IDs that represented one or two individual sequences that were discarded as they could not generate consensus sequences. B) Number of unique Primer IDs that represented three or greater number of individual sequences. The highest number of individual sequences represented by a single Primer ID was 4705.

ID technology was able to quantify the ratio by the number of sequences that a Primer ID represented.

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 3.

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 randomly. 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 randomly (Sheward et al., 2012). Analysis of the two real

Table 3.3: The number of duplicated Primer IDs in samples before and after filtering the Primer IDs representing less than three sequences from Run1.

|  |  |  |
| --- | --- | --- |
| Sample | Number of Duplicated Primer IDs before filtering by threshold number of sequences | Number of Duplicated Primer IDs after filtering by threshold number of sequences |
| Patient A | 77 | 8 |
| Patient B | 1103 | 64 |
| Patient C | 44 | 5 |
| Patient D | 75 | 7 |

data sets also showed that there was up to 1000 fold duplication of Primer IDs (Table 3.3). 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. The PIDA algorithm can be incorporated in computational pipelines.

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.