# Chapter 3

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

### 3.1 Introduction

The Roche/454 high throughput next generation sequencing (NGS) platforms Junior and FLX/FLX+ are capable of generating hundred thousands and a million of short sequence reads respectively from DNA fragments at low cost and less time than other NGS systems. The ultra deep high throughput capability has enabled sequencing of low frequent 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). It is essential to correct the inevitable errors that modifies and over diversifies the original sample while analyzing highly complex and diverse sample.

Recently, a novel technology, based on bar coding an original template RNA sequence in a sample has been introduced (Jabara et al., 2011). The barcode, called Primer ID, is a short unique degenerate nucleotide sequence added to a reverse transcription primer during cDNA production. Before cDNA production, the Primer ID is attached to sample specific ID or Multiplex Identifier (MID) with a spacer in between, which again is attached to a PCR primer with another spacer in between (Figure 3.1 A). In the subsequent PCR cycles, the cDNA with its associated Primer ID is amplified (Figure 3.2 A). Sequences with same Primer ID are, theoretically derived from the same viral RNA template sequence. The sequences tagged with same Primer ID can be used to create a consensus sequence that theoretically represents an error free and original template viral RNA sequence (Figure 3.2 B).

An highly demanding application of Primer ID is ultra-deep high throughput NGS HIV-1 genotyping for the viral drug resistance testing. HIV population in an infected individual is highly heterogeneous. Despite this, ultra-deep sequencing of HIV-1 has enabled to unfold unprecedented viral diversity and understand viral dynamics and evolution (Salazar-Gonzalez et al., 2009). But the errors generated in the process of ultra-deep high throughput sequencing confound the true HIV-1 variants and inflates the diversity (Kunin et al., 2009). With the application of Primer ID technology, the errors can be minimized to zero level and reflect the true variants and diversity.

Jabara et al (Jabara et al., 2011) first introduced the Primer ID technology and analyzed the raw data with the algorithm they have developed but was not released. In the future, this technology is anticipated to be used widely, we have developed an algorithm called Primer ID Algorithm (PIDA,) specifically for raw data generated with Primer ID technology using Roche/454 ultra-deep sequencing technology.

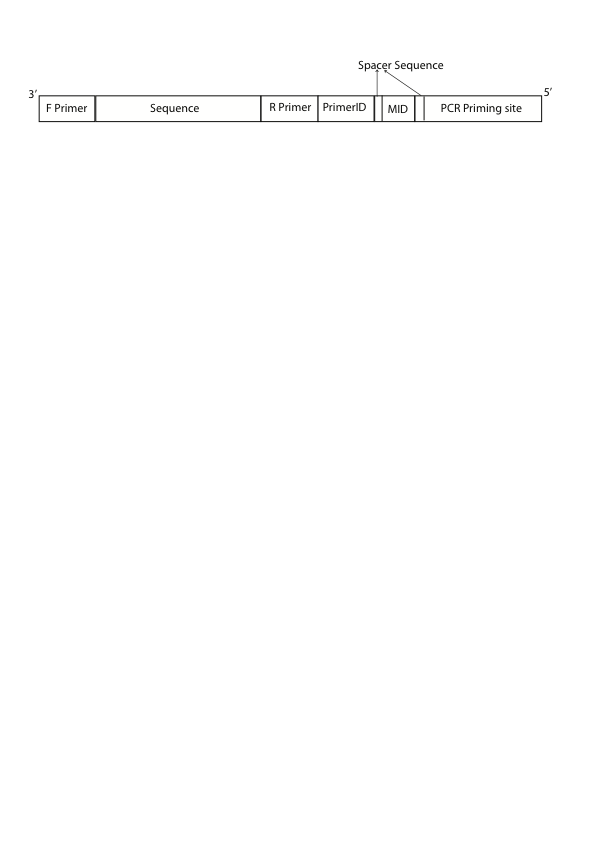


Figure 3.1: Locations of forward and reverse primer, Primer ID, MID and PCR primer in a raw sequence data generated using Primer ID technology. The primers tags the amplicon region which is being amplified, MID tags the specific individual from whom the sample is obtained and Primer ID tags the RNA sequence from where the raw sequence read is originated. A universal PCR primer binds at PCR priming site while PCR amplification. Two spacer sequences are used to separate the MID sequence from Primer ID and PCR priming site. The spacer sequences are completely custom defined.



Figure 3.2: The Primer ID is shown as a series of 8 Ns. Application of Primer ID is to identify the original RNA template sequence. A) Primer IDs are added to cDNA primer during reverse transcription process. B) Sequences with same Primer ID can be used to generate a consensus sequences that is free from sequencing errors, recombination error and PCR misincorporations. Source: Jabara et al 2011 (Jabara et al., 2011).

3.2 Methods and Materials

We have developed a novel algorithm called Primer ID algorithm (PIDA) that processes raw sequence reads generated using the Primer ID approach to generation of consensus sequences from reads with same Primer ID. The algorithm requires sequence input file in FASTQ format.Other required files are primer file, sample ID/MID file and gene file; all in tab delimited format as shown in Figure 3.3 The primer file contains forward and reverse primer that indicates an amplified genomic region. If mutiple samples are sequenced together, each sample is tagged with a unique MID sequence, which needs to be supplied through Sample ID/MID file. The gene file contains user defined minimum forward and reverse sequence lengths. Users are required to input a universal PCR primer sequence and the known design pattern of Primer ID, spacer, MID, another spacer and PCR priming site (Figure 3.1). For example, a user may input the pattern 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 input threshold number of sequences required to generate consensus sequence, maximum mismatches allowed in primer sequence and in MID sequence called primer tolerance and MID tolerance respectively.

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

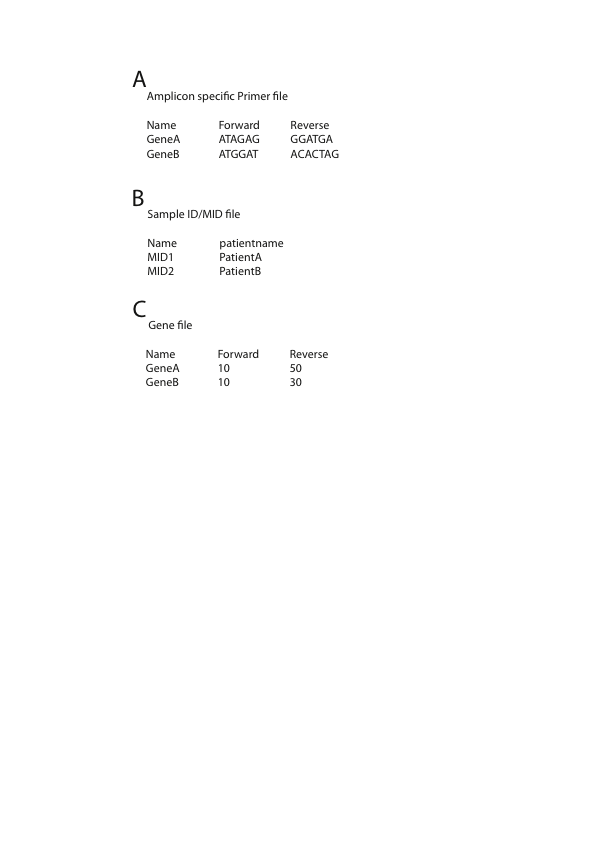


Figure 3.3:Different files required as input for the PIDA algorithm. Each column in all files has to be separated by a single tab. A) The amplicon specific primer file contains gene names that are amplified, the forward primer and reverse used for each amplicon. B) The sample ID/MID file contains names of MID sequence used to tag the samples and sample/patient 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.

**3.2.1 Sequence identification with information tags or Demultiplexing**

The information provided in primer file and MID file is used to identify the gene/amplicon and sample respectively, to which the sequence belong. The step is shown in figure 3.4 in red color text. The algorithm begins with a search for forward primer at 5’ end. The algorithm extracts nucleotides of length equal to the forward primer length at 5’ end and pairwise aligns the extract with forward primers one at a time. If the number of mismatches in pairwise alignment is below the primer tolerance, the sequence is identified with the amplicon/gene name of the forward primer and the search for reverse primer is skipped. If the sequence begins with none of the forward primers, the presence of reverse primer is checked. The reverse primer is located inwards after the length of a PCR priming site, a spacer, a MID, another spacer and a Primer ID. Similar to forward primer, the nucleotides of length equal to reverse primer are extracted from the position where reverse primer should start in the sequence. The extract is pairwise aligned with reverse primers one at a time and a reverse primer that have number of mismatches less or equal to primer tolerance is selected and indentified with its amplicon name. The reverse sequences are reverse complemented to convert into forward sequence. The sequence read is discarded if none of the primers are found.

The algorithm, then, searches for the sample ID/MID. As reverse sequences are converted into forward as well, the MID sequence appears close to 3’ end. The MID is searched using a known nucleotide pattern in a sequence. A known pattern is ordered sequence of spacer, some nucleotide of length equal to MID length, another spacer and PCR primer sequence. If the pattern occurs in a sequence, the sequence between the MID is extracted and pairwise aligned with list of supplied MID sequences one at a time.

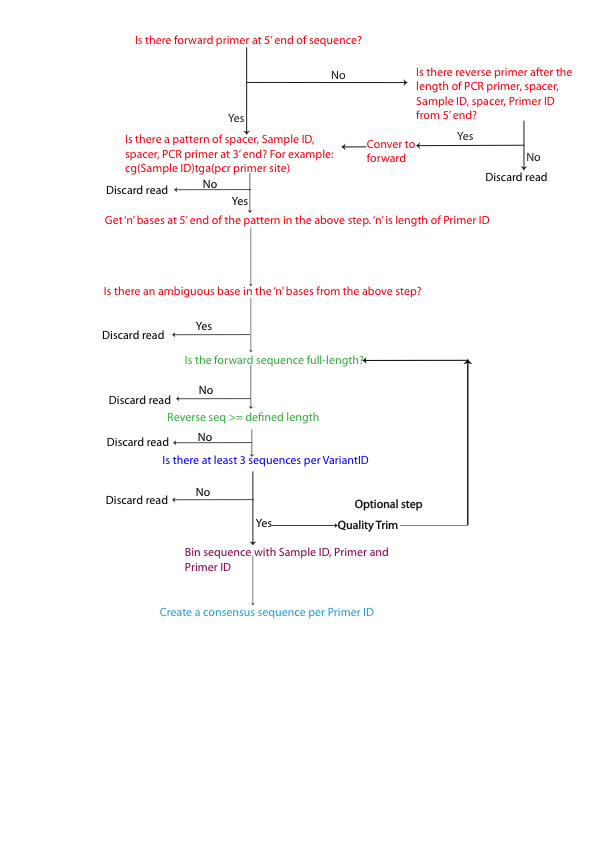


Figure 3.4: The logical flow of the PIDA algorithm to process raw sequence data generated with Primer ID technology into generation of consensus sequence. The colors denote PIDA processing steps. Red: sequence demultiplexing, Green: Sequence filter by length, Blue: sequence filter by number of sequence represent by each Primer ID, Black: Quality trimming, Purple: Sequencing binning, Sky blue: Consensus sequence generation

Once the number of mismatches less or equal to defined tolerance is obtained, the sequence is tagged with the sample name or discarded if none of the MID sequence is obtained. Once a matching MID is found, a sequence of ‘n’ nucleotides is extracted as an Primer ID sequence, where ‘n’ is the length of Primer ID, from the read before the known sequence pattern. The read is discarded if there is a presence of an ambiguous base in the extracted ‘n’ nucleotides. The information tags Primer, MID and Primer ID are added at sequence id for further downstream processing. All non-discarded sequences are saved in a single file in FASTQ format.

**3.2.2 Sequence filtration by 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 user has 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 3.4 in green color).

**3.2.3 Primer ID filtration with minimum number of sequences**

“PCR and sequencing errors in the sequences can be resolved using multiple reads over a given base” (Jabara et al., 2011). As sequences with same primer and Primer ID in a sample originate from same template viral RNA sequence, error in one sequence can be corrected in reference to other sequences. Therefore, more than one sequence with same primer, MID and Primer ID tags is required. User has an option to select a minimum number of sequences that have same primer, MID and Primer ID tags.

The default minimum number of sequences is 3. The algorithm discards all the sequence reads if the number of sequences in the group containing same information tags is less than the user defined minimum sequences (Figure 3.4 in blue color).

**3.2.4 Quality trimming**

Quality trimming is an optional step in the algorithm. If user selects the trimming option, the algorithm calls quality trimming tools to quality trim the non-discarded sequences. Quality trimming may change the sequence length. Therefore, the algorithm repeats section 3.2.2 and 3.2.3 (Figure 3.4 in black color).

**3.2.5 Binning**

The information tags for each sequence are collected from the sequence id. Sequences with same information tags are binned together and saved in a file. This creates a number of files that is equal to the number of unique combinations of primer, MID and Primer ID tags (Figure 3.4 in purple color).

**3.2.6 Generating a Consensus Sequence**

Sequences binned together in a file with the same primer, MID and Primer ID tags are aligned using a multiple sequence alignment tool called MAFFT (Katoh et al., 2005; Katoh et al., 2002; Katoh and Toh, 2008, 2010). A consensus sequence is generated adding the bases that occur the most at a given position in the alignment (Figure 3.4 in Sky blue color). 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.

Two datasets (Run1 and Run2), each containing four HIV infected patients, are obtained from a study to analyze vaccine response in HIV subtype C (The result will be published elsewhere). Four amplicons *env*, gag54, *gag*472, and *nef*23 are amplified using subtype C specified primers associated to a unique Primer ID.

### 3.3 Result

Datasets Run1 and Run2 are analyzed using PIDA. Dataset Run1 and Run2 contains 1,25,865 and 40544 raw sequence reads respectively. The PIDA algorithm discards a number of sequences if none of the user supplied list of primers and MID tags are found in the sequences. Non discared sequences may be further discarded for its short length at downstream processing or if its Primer ID has an ambiguous nucleotide as the Primer ID is not definitive. The total number of sequences discarded at each step of the algorithm before analysis with Primer ID tag is shown in the table 3.1. A total of 10808 (8.59%) and 1920 (4.73%) sequences are discarded in the steps for Run1 and Run2 respectively.

While analyzing the data only with Primer ID, initially, the number of sequences with same Primer ID of a same amplicon belonging to a sample is calculated. For a default threshold number of sequences for consensus sequence generation, if the number of sequences with sample Primer ID in an amplicon of a sample is less than 3, then the sequences are discarded. This discarded 18011 (15.653%) and 7397 (19.151%) of the remaining sequences in Run1 and Run2 respectively (Table 3.2). In the downstream processing, the sequences with Primer ID, spacer sequences, MID and priming sites removed are quality trimmed using QTrim. Quality trimming changes the sequence length and discards low quality sequences. Therefore, once again the sequences are filtered by length and threshold number of sequences to generate consensus sequences.

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 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 |

The final number of sequences for generation of consensus sequences is 74.47% (93731) and 77.25% (31321) of the total raw reads in Run1 and Run2 respectively. The number of sequences distributed per amplicon per sample is shown in figure 3.5. Each sequence has a Primer ID that represents 3 or more sequences required for generation of a consensus sequence. The number of consensus sequences per amplicon per sample is shown in figure 3.6

### 3.4 Discussion

The Primer ID is relatively a new technology to correct PCR and other sequencing errors. The technology can be anticipated to be widely accepted in future high throughput sequencing involving highly heterogeneous population sample (Sheward et al., 2012). However, there are few issues with the technology.

**3.4.1 Error in cDNA production and first cycle of PCR cannot be tracked**

HIV-1 RNA sequences are reverse transcribed into cDNA with reverse transcriptase, which is then amplified with PCR. A misincorporation of a base in the reverse transcription process and in the first cycle of PCR gets amplified in the subsequent PCR cycles, giving a mixture in the base position. They go undetected during analysis.

**3.4.2 Large number of unamplified Primer IDs**

A Primer IDs representing single or two sequences are discarded in the analysis, as they could not generate consensus sequences. A total of 15838 and 7614 unique Primer IDs were observed in Run1 and Run2 respectively in the initial step of demultiplexing.

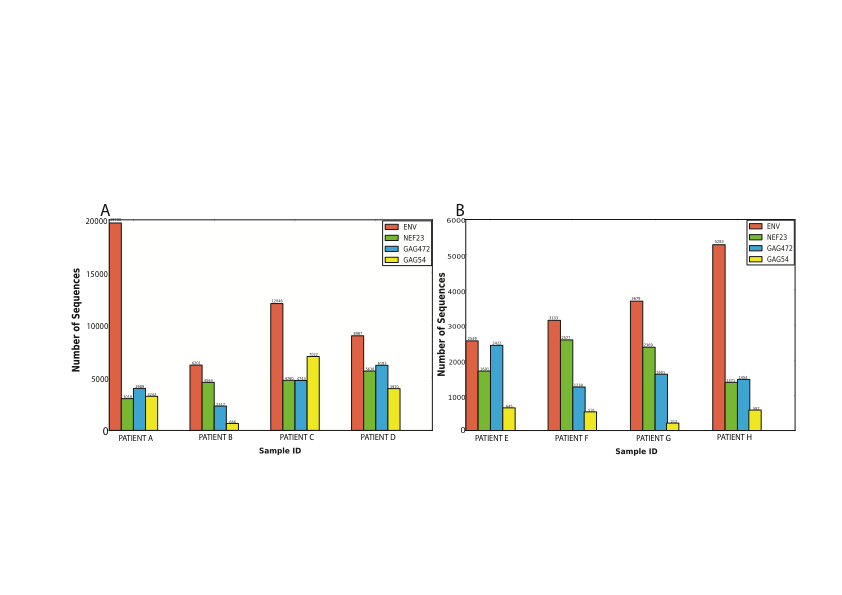
****

Figure 3.5: 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 in (A) Run1 and (B) Run2. Each Primer ID represented sequences in each amplicon of a sample creates a consensus sequence.

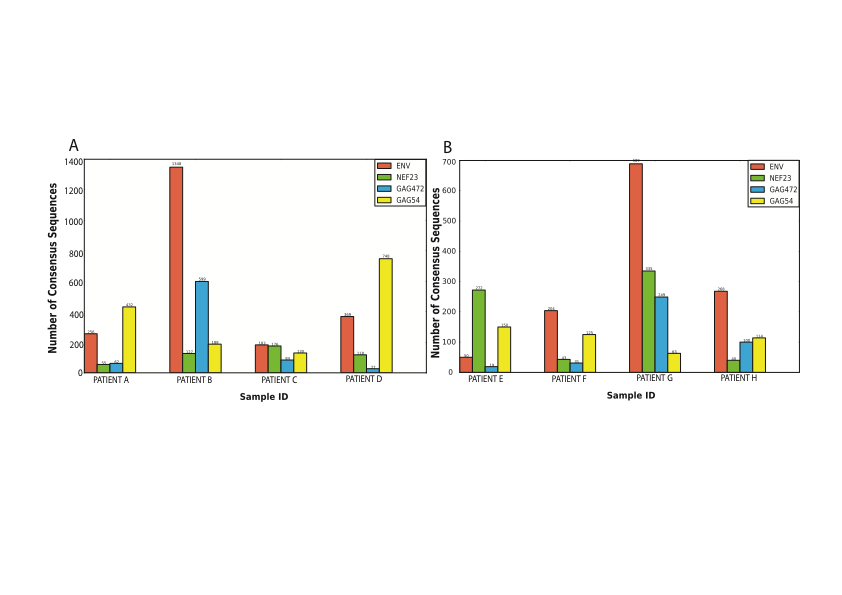


Figure 3.6: The total number of consensus sequences generated per amplicon per sample for (A) Run1 and (B) Run2

However, the final total number of Primer IDs that went on to generate consensus sequences are 4704 (29.7%) and 4981 (65.41%) in Run1 and Run2 respectively, which shows that large number of template sequences do not get amplified in the PCR process (Figure 3.7 A) (Storhoff et al., 2004).

Our result correlates with the analysis in publication by Jabara et al (Jabara et al., 2011). The cause for this effect is attributed to bias in template annealing only with perfect match (Sipos et al., 2007). The reason for the bias might be PCR dependencies like annealing temperature and primer oligonucleotide length (Wu et al., 1991).

Although Primer ID technology cannot rule out no amplification of large number of templates in PCR, the technology enables to quantify the viral variants in HIV quasispecies. However, discarding huge number of unique combination of tags that represent individual sequence below the threshold value might under represent the true viral diversity (Sheward et al., 2012).

**3.5 Allelic skewing**

Primer ID technology enables the number of folds a template sequence is amplified in PCR. We observed the over a 100 fold amplification of sequences than the other, which is consistent with the study by Jabara et al (Jabara et al., 2011). While large number of Primer IDs has only 3 sequences, we observed a Primer ID with up to 4705 sequences (Figure 3.7).

**3.6 Duplication of Primer ID and lost of viral variant**

Primer IDs of 8 nucleotides, generated at random, may have a chance that some Primer IDs are generated multiple times (Sheward et al., 2012). It is observed that large number of Primer IDs is duplicated within a sample across the amplicons as pointed by Sheward (Sheward et al., 2012). I observed over a thousand fold duplication of Primer ID in a single sample (Table 3.3).

Seqcount_primerid.pdf

Figure 3.7: Distribution of number of unique Primer IDs by number of individual sequences represented. A) Number of unique Primer IDs that represent one or two individual sequences that are discarded as they cannot generate consensus sequences. B) Number of unique Primer IDs that represent three or greater number of individual sequences. The highest number of individual sequences represented by a single Primer ID is 4705.

At such a large Primer ID duplication, there is a high chance than two or more Primer IDs tag same genomic region in a sample. It is difficult to differentiate viral variants when two similar Primer IDs tag the same amplicon in same sample. Viral variant might be lost if duplicated Primer IDs tag same amplicon of viral population from a same sample as the variant with less amplification is not represented during consensus sequence generation (Sheward et al., 2012).

### 3.5 Conclusion

Primer ID is a novel technology for correcting sequencing errors in high throughput next generation sequence data. Primer ID technology can be anticipated to be widely adopted in future ultra deep high throughput next generation sequence 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 datasets generated using Primer ID technology and the comparative results were observed in both datasets.

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 |

Bibliography

.

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.

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.

Salazar-Gonzalez, JF, Salazar, MG, Keele, BF, Learn, GH, Giorgi, EE, Li, H, Decker, JM, Wang, S, Baalwa, J, Kraus, MH, Parrish, NF, Shaw, KS, Guffey, MB, Bar, KJ, Davis, KL, Ochsenbauer-Jambor, C, Kappes, JC, Saag, MS, Cohen, MS, Mulenga, J, Derdeyn, CA, Allen, S, Hunter, E, Markowitz, M, Hraber, P, Perelson, AS, Bhattacharya, T, Haynes, BF, Korber, BT, Hahn, BH, Shaw, GM (2009) Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J Exp Med* **206**: 1273-1289.

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.

Sipos, R, Szekely, AJ, Palatinszky, M, Revesz, S, Marialigeti, K, Nikolausz, M (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol Ecol* **60**: 341-350.

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.

Wu, DY, Ugozzoli, L, Pal, BK, Qian, J, Wallace, RB (1991) The effect of temperature and oligonucleotide primer length on the specificity and efficiency of amplification by the polymerase chain reaction. *DNA Cell Biol* **10**: 233-238.

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.