# Chapter 3

## PIDA – Algorithm for Accurate HIV -1 Quasispecies Sequencing and Sampling of Ultra Deep High Throughput Sequence Data using Primer ID

### 3.1 Introduction

Roche/454 ultra-deep high throughput next generation sequencing technology is capable of generating millions of short sequence reads from DNA fragments at low cost and less time. For more detail on NGS technologies, see Chapter 1, section 1.11. The ultra deep high throughput capability has enabled sequencing of low frequent variants from highly heterogeneous viral population sample (Fischer et al., 2010; Hoffmann et al., 2007; Mitsuya et al., 2008; Rozera et al., 2009; Varghese et al., 2009; Wang et al., 2007). This characteristic of ultra-deep high throughput sequencing has allowed unprecedented details of HIV diversity, evolution and dynamic (Salazar-Gonzalez et al., 2009). Therefore, a prominent use of ultra-deep sequencing is the characterization of drug resistant minor variants in a sample of viral quasispecies from a HIV infected individual. However, high rate of sequencing errors are incorporated and accumulated at PCR amplification step (Kanagawa, 2003) and by instrumental/hardware error and sequencing error (reviewed in (Metzker, 2009)). These errors inflate and confound the real viral population genetic diversity (Kunin et al., 2009; Zagordi et al., 2010). 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 (Liu et al., 1996; Polz and Cavanaugh, 1998). An error in Roche/454 sequencing that is highly related in HIV-1 drug resistance genotyping is the homopolymer sequencing error (Gilles et al., 2011), which is due to the disproportion in the light intensity detection and number of base incorporations in a particular cycle of base flow during sequencing. Due to the presence of drug resistant mutations at homopolymer regions in HIV-1 protease and reverse transcriptase genes, accurate genotyping is highly essential. For more details, see Chapter 1 section 1.11 and 1.12. It is essential to correct the cinch errors that modifies and over diversifies the original sample while analyzing complex and diverse HIV-1 sample.

Some sequencing error correction methods based on statistical analysis have been developed for estimation of diversity in HIV population (Zagordi et al., 2011; Zagordi et al., 2010). However, those methods could not remove the errors completely and were tested to correct only the substitution errors but Roche/454 generates insertion/deletion errors. Therefore, the methods are not suitable for HIV-1 drug resistance genotyping.

Recently, a novel technology, based on bar coding the original template viral RNA 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 (Figure 3.1 A). For a Primer ID of length 8 bases, 65536 (48) unique Primer IDs can be produced. In the subsequent PCR cycles, the cDNA with its associated Primer ID is amplified. Sequences with same Primer ID are, theoretically derived from the same viral RNA template sequence. Same Primer ID sequences can be used to create a consensus sequence that theoretically represents the template viral RNA sequence (Figure 3.1 B).

Figure 3.1: Application of Primer ID to identify the original viral 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).

This approach removes all PCR related errors and sequencing error.

### 3.2 Methods and Materials

We have developed a novel algorithm called Primer ID algorithm (PIDA) that processes raw sequence reads generated using Primer ID technology into generation of consensus sequences from reads with similar information tags that are primer, Sample ID or Multiplex IDentifier (MID) and Primer ID. The algorithm requires sequence input file in FASTQ format; other required files are tab-delimited primer file, sample id file and gene file in a defined format as shown in Figure 3.2. Users also require to input sequence of blocks of information and the length of each information block. For example, a user may input blocks of information as primerid8.cg.mid5.tga.pcr primer sequence, in 5’ to 3’ order, which indicates a Primer ID of length 8 nucleotides, two spacer sequences ‘cg’ and ‘tga’, MID sequence of length 5 nucleotides and a universal PCR primer sequence.

**3.2.1 Information search and demultiplexing**

The algorithm searches amplicon specific primers (both forward and reverse primers) and sample specific MID using pairwise sequence alignment allowing user defined maximum threshold of mismatches. The sequence read is discarded if any information tag is not found. If the sequence is not discarded, a sequence of ‘n’ nucleotides is extracted as an Primer ID sequence, where ‘n’ is the length of Primer ID, from the read after the sequence of known blocks of information that are PCR primer, spacer, MID and spacer. 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 to the sequence id for further downstream processing. All non-discarded sequences are saved in a single file.

**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 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.

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.

**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 information tags 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 information tags is required. User has an option to select a minimum number of sequences that have same combinations of information 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.

**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. For more details on quality trimming using QTrim, see Chapter 2. Quality trimming may change the sequence length. Therefore, the algorithm repeats **sequence filtration by length** and **Primer ID filtration** with minimum number of sequences.

**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 the information tags.

**3.2.6 Generating a Consensus Sequence**

Sequences binned together in a file with the same information tags are aligned using a multiple sequence alignment tool called MAFFT. A consensus sequence is generated adding the bases that occur the most at a given position in the alignment. 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.

The logical flow of the algorithm is shown in figure 3.2.

### 3.3 Result

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.

Datasets Run1 and Run2 are analyzed using PIDA. Dataset Run1 contains 1,25,865 raw sequence reads. A total of 10808 (8.59%) sequences are discarded in the demultiplex step. The number of sequences demultiplexed by using information tags MID and amplicon primer only is shown in figure 3.3 (A) and the final number of consensus sequences after binning further using Primer ID tag is shown in figure 3.3 (B).

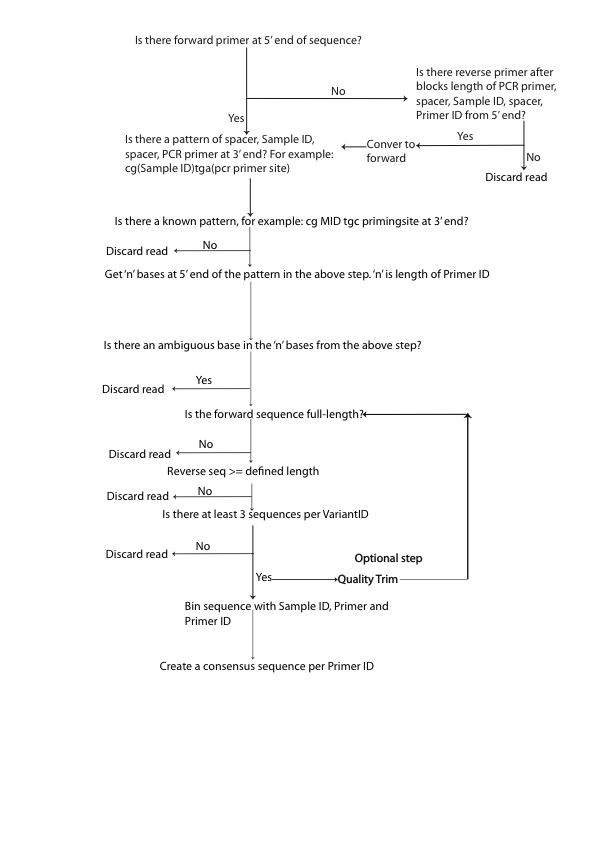


Figure 3.2: The logical flow of the PIDA algorithm to process raw sequence data generated with Primer ID technology into generation of consensus sequence.

### PrimerID_figure.pdf

Figure 3.3: Differences in data analysis using sequences generated with and without the use of Primer ID. A) Sequence distribution after demultiplexing by sample and genomic region. This data would be used in downstream analysis, if no Primer ID were added in reverse transcription. Sequences with same Primer ID are multiple aligned and a consensus sequence is generated. B) Consensus sequence distribution by sample and genomic region generated using Primer ID information.

**3.3.1 Loss of HIV-1 Variants using Primer ID**

The PIDA algorithm generated a total of 4704 consensus sequences for dataset Run1. However, the number of sequences with unique combinations of information tags is much higher than the number of final consensus sequences. A total of 7542 unique combination of information tags, each representing one sequence and a total of 3030 unique combination of information tags, each representing two sequences are discarded as the number of sequence they represent is below the minimum number of sequence required for consensus sequence generation. As each unique combination of tags represent a unique HIV-1 variant, a total of 10572 (7542 + 3030) viral RNA sequences are lost in the analysis as these RNA template sequences were not amplified at the PCR step before ultra-deep sequencing.

**3.3.2 Disproportion between number of sequences and consensus sequences**

We observe that patient A has the highest number of individual sequences (19700) in *env* gene. Consensus sequence generation from these sequences showed it sampled only 256 viral RNA templates. On the other hand, patient B *env* amplicon has only 6201 individual sequences, significantly less than patient A’s *env* amplicon but it sampled 1348 viral RNA templates. Similar disproportion between number of individual sequences and consensus sequences were observed in other amplicons between different patients and within a patient. This shows that higher the number of sequences does not sample out higher number of template viral RNAs.

**3.3.3 Differential amplification**

Tagging of viral RNA template with Primer ID has enabled to detect differential amplification of original DNA templates in PCR. As discussed in section 3.3.1, huge number of sequences was rejected because they were not amplified to the threshold number of sequences required to generate consensus sequences. However, there were unique combinations of information tags that represent huge number of sequences, which means those viral RNA template sequences were largely amplified (Figure 3.4).

**3.3.4 Duplication of Primer IDs**

Primer IDs of 8 nucleotides, generated at random, may have a chance that some Primer IDs are generated multiple times (Sheward et al., 2012).We observed that large number of Primer IDs is duplicated within a sample across the amplicons as pointed by Sheward (Sheward et al., 2012) (Table 3.1). It is observed that huge number of duplicated Primer IDs in a sample, across the amplicons, represent individual sequence that are not amplified. Therefore, while filtering the sequences with the minimum number of sequences represent by a Primer ID in a particular sample, large number of Primer IDs was discarded (Figure 3.4).

Comparative results were observed for Run2 dataset when analyzed with PIDA for the four point discussed above. (data not shown).

### 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 related to choosing the Primer ID sequence length as pointed by (Sheward et al., 2012). A set of Primer ID of different combinations depends on length of degenerate bases in the ID. It is very rare to select a subset of unique IDs from a set of 65536 possibilities (Sheward et al., 2012) leading to duplication of the IDs, which was also observed in our analysis . Ultra-deep high throughput sequencing requires PCR amplification of the template viral cDNA sequences.

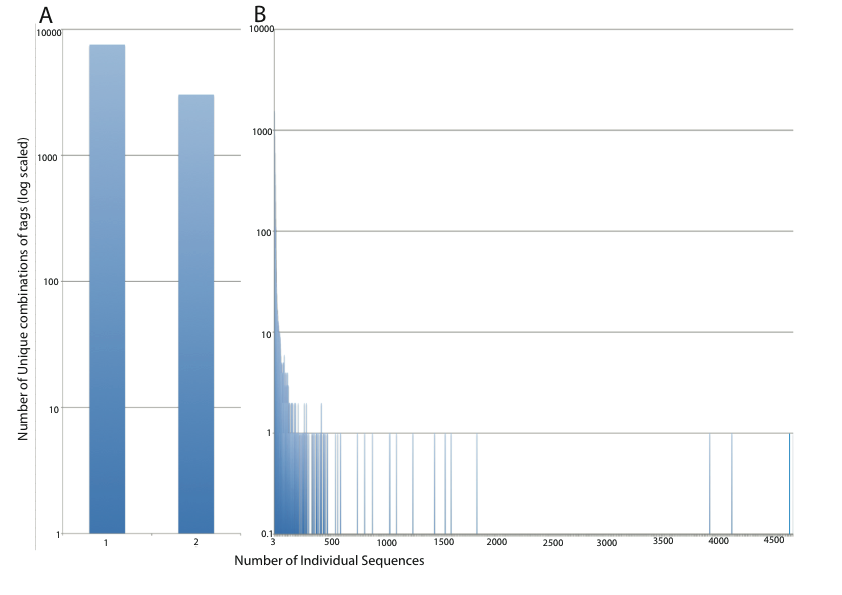


Figure 3.4: Distribution of number of unique combinations of tags by number of individual sequences represented. A) Number of unique combinations of tags that represent one or two individual sequences. B) Number of unique combinations of tags that represent three or greater number of individual sequences. The highest number of individual sequences represented by a combination of tags is 4705 individual sequences.

|  |  |  |
| --- | --- | --- |
| 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 |

Table 3.1: The number of duplicated Primer IDs in samples before and after filtering the Primer IDs representing less than three sequences.

However, the primary concern in PCR is that large number of templates does not get amplified at all (Storhoff et al., 2004). 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). Large number of single reads of Primer IDs was observed in the datasets (7542 reads in Run1 and 3600 reads in Run2). Our result correlates with the analysis in publication by Jabara et al (Jabara et al., 2011). Although Primer ID technology cannot rule out PCR dependencies causing no amplification of large number of templates, the technology enables to quantify the viral variants in HIV quasispecies. 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). But this effect is minimal (Jabara et al., 2011).

### 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 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.

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.

Gilles, A, Meglecz, E, Pech, N, Ferreira, S, Malausa, T, Martin, JF (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* **12**: 245.

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.

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, Bhattacharya, A, Eriksson, N, Beerenwinkel, N (2011) ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* **12**: 119.

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.