Chapter 3

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

# Introduction

Ultra Deep PyroSequencing (UDPS) platforms are capable of generating as much as millions of sequence reads from DNA fragments at low cost and in less time than other sequencing approaches. This capability enables the potential to fully characterize viral quasispecies including the low frequency variants (Fischer et al., 2010; Hoffmann et al., 2007; Mitsuya et al., 2008; Rozera et al., 2009; Varghese et al., 2009; Wang et al., 2007) However, 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.

# Methods and Materials

## 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 Multiplex Identifier (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.

## 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:

### 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.

### 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).

### 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 three 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).

### 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.

### 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.

## Results

### Initial demultiplexing

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

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

For the remaining reads that contained complete primers and MIDs, we examined the sequence of the Primer ID and retained reads that did not contain ambiguous bases at Primer ID region. 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 sequence data and sequence length evaluation

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

Using these parameters forward and reverse sequence read lengths were analyzed independently of each other and those sequences that were shorter than the required read length were removed from the analysis, 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 resulting in 16 unique datasets generated from each sequencing run (**Table 3.4**). Each of these datasets was subsequently analyzed independently of all others.

### Characterization of primer IDs

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

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

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

### Generation of consensus sequences

Consensus sequences were only generated for those Primer ID tags with three or more representative sequences (Figure 3.8). Generation of consensus sequences showed that few Primer IDs could be represented by a large number of sequences (**Figure 3.8**). For example, Patient A’s *env* amplicon in Run1 had the highest total number of sequence reads (19700, Figure 3.8A) but the number of Primer IDs that they represent was only 241 (Figure 3.8C). On the other hand, Patient B’s *env* amplicon had a lower number of reads (Figure 3.8A) but had more than six times the number of Primer ID tags than patient A had for the same amplicon (Figure 3.8C).

## Discussion

HIV has turnover of 1010 new viruses per replication cycle and generates errors at the rate of 10-5 per base per replication cycle due to its error prone reverse transcriptase (Ho et al., 1995; Perelson et al., 1996). Therefore, HIV creates a complex and highly diverse population within the host (Drake and Holland, 1999; Duffy et al., 2008; Onafuwa-Nuga and Telesnitsky, 2009). UDPS technology has enabled sampling of the highly diverse population (Bentley et al., 2008; Eid et al., 2009; Fischer et al., 2010; Margulies et al., 2005; Metzner et al., 2009; Shafer, 2009). However, PCR amplification of the DNA sample precedes the actual UDPS, and the PCR step may incorporates errors like recombination of DNA (Judo et al., 1998; Meyerhans et al., 1990; Yang et al., 1996), misincorporation of a base at new growing strand (Hughes and Totten, 2003; Kanagawa, 2003), differential amplification of two different DNA creating new sample with different ratio of DNA than the original sample (Liu et al., 1996; Polz and Cavanaugh, 1998). These PCR errors create false viral diversity in the viral population, which needs to be corrected. Recently a technology called Primer ID (Jabara et al., 2011) has been introduced as an effort to corrected PCR and sequencing related errors. Kindle et al also showed that errors were reduced by approximately 20 fold using the Primer ID technology (Kinde et al., 2011).

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 can be removed by the generation of a consensus sequence from sequences represented by a Primer ID. Therefore, the technology can be anticipated to be widely accepted in future UDPS involving highly heterogeneous population sample (Sheward et al., 2012). Because the technology developers and publishers have not made the code available for public use, the PIDA algorithm has been developed.

### Demultiplexing of sequence reads

The PIDA algorithm was tested on two real Primer ID raw sequence datasets. The algorithm was able to demultiplex the sequence reads with Primer ID along with MID and amplicon primers. However, 9% and 5% of the sequence reads were discarded at demultiplex step in Run1 and Run2 respectively.

The reasons PIDA discarded sequence reads were unable to locate either MID sequence, primer or the presence of ambiguous base in Primer ID sequence.

The MID is located at the proximity of 3’ end of the forward sequences and close to 5’ end of reverse sequences (Figure 3.1). We set a criterion of zero tolerance in spacer and MID region and a tolerance of five in priming site region. Most of the sequences are discarded for the reason that PIDA is unable to find exact MID in forward sequence reads. As the base quality drops towards 3’ end of sequence reads in UDPS (Cock et al., 2010), there may be error in bases close to the end and thus the algorithm cannot find the exact matching MID sequence. A subsequence that is exactly similar to MID might occur anywhere within a sequence read but a sample specific MID must be present in between the spacer sequences close to either 5’ or 3’ end (Figure 3.1). An error in single base either in spacer sequence or in MID sequence disables identification of a MID for the sequence resulting to discard of the sequence read.

The presence of a gene specific primer either forward or reverse was checked if a sample specific MID was found in a sequence read. PIDA used local pairwise alignment to search primers and we allowed a maximum tolerance of three bases while searching for a primer in a sequence reads. A mismatch of over three bases in the primer region discarded a sequence read. PIDA discarded 2962 and 751 sequence reads in Run1 and Run2 respectively for not finding a primer.

Finally, after the search of a MID and a primer are successful, PIDA copies ‘n’ bases as a Primer ID (where ‘n’ is length of Primer ID) of the sequence from the known Primer ID region. A presence of any ambiguous base in the noted Primer ID means it is not possible to identify the original template cDNA sequence thus the sequence reads are discarded. PIDA discarded 257 and 37 sequence reads from Run1 and Run2 respectively for the presence of an ambiguous base in the Primer ID. Almost all the sequence reads discarded in this category were forward orientation. Primer IDs are present towards 3’ end of forward sequence reads and as the quality drops towards the end, an ambiguous base might be incorporated during sequencing.

### Forward and reverse sequence read length

In forward sequence reads, the MID and Primer ID tags are present at the 3’ end of forward sequences. In order to demultiplex by MID and Primer ID tags, the forward sequence reads have to be full-length amplicon to cover the tags. We observed high number of sequence reads was discarded for the reason that no MID were found (Table 3.1). No forward sequence reads were discarded for the reason of low read length (Figure 3.3) and the reason was that PIDA searched for MID and Primer ID before the sequence read length.

In reverse sequence reads, the MID and Primer ID tags are present at the 5’ end. Therefore, all reverse sequence reads cover MID and Primer ID tags. However, minimum read length for reverse sequence reads were defined such that they cover the region of interest in the amplicon. Only 32 and 23 sequence reads were discarded for not covering the region of interest in Run1 and Run2 respectively.

### Analysis of Primer IDs by their representative number of sequences

For each dataset binned together by sample and primer, the sequence reads were further sub-binned by Primer ID. There were 19558 and 8694 Primer IDs in Run1 and Run2 respectively. The number of representative sequences in the Primer IDs ranged from one to 4144 in Run1 and one to 1119 in Run2. Jabara et al reported the range of one to 96 representative sequences per Primer ID (Jabara et al., 2011). The variation in number of sequences reads in each sub-bin was over thousand folds (Figure 3.4 – 3.7), which was much higher than 100 folds as observed by Jabara and colleagues (Jabara et al., 2011). The median sequence reads per Primer ID in Run1 and Run2 was three while Jabara et al (Jabara et al., 2011) reported the median of six. We could not explain the difference in range of representative sequence reads and median sequence reads per Primer ID between Jabara et al’s observations and ours.

Primer IDs with one or two representative sequence reads were discarded, as they could not generate consensus sequence. The total of 73% and 93% of Primer IDs in Run1 and Run2 respectively was discarded for representing less than minimum number of sequence reads required to generate a consensus sequence. This loss was much higher than 38% that was observed by Jabara et al (Jabara et al., 2011). Because each Primer ID tags a unique cDNA of a HIV virus in a population, the removal of Primer IDs with one or two representative sequences would mean that the final analysis under represents the true diversity of viral variants in the obtained sample (Sheward et al., 2012).

### Representative sequences per Primer ID and consensus generation

The minimum number of representative sequence reads per Primer ID to generate a consensus sequence is three. Primer IDs with two representative sequence reads can be used to generate a consensus sequence but they can have large number of ambiguous bases in it as an ambiguous base is added if there is a tie between two bases at many base positions. Therefore the ideal minimum number of representative sequence reads per Primer ID for consensus generation should be three.

In theory, an amplicon with the highest number of sequences should have high number of unique Primer IDs. Each unique Primer ID generates a consensus sequence, therefore, the amplicon with the highest number of sequences should also generate the highest number of consensus sequences. But we observed that the amplicon with the highest number of sequences did not produce the highest number of consensus sequences. It was because there was a wide range in the number of representative sequences per Primer ID per amplicon (**Figure 3.4 – 3.7**).

### Limitations of Primer ID technology

Although, the PCR technology was developed to resolve PCR and sequencing errors, it was unable to resolve the nucleotide misincorporation errors during cDNA synthesis (Boyer et al., 1992) and initial cycles of PCR (Horton, 1995). These initial errors are propagated to the new DNA sequences during amplification in subsequent cycles (Horton, 1995). Thus they are difficult to detect during downstream analysis. Therefore, we were not assured that a consensus sequence generated from a Primer ID with three or more representative sequence reads represents an error free viral sequence from a sampled viral population. Schmitt et al has developed an approach to further correct any form of sequencing errors by tagging both strands of duplex DNA (Schmitt et al., 2012). This approach compares one strand with the other strand of the DNA for error correction (Schmitt et al., 2012). However, the approach has not been applied at UDPS of HIV quasispecies

Another 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 Primer IDs when generating a set of 65536 Primer IDs (for Primer ID of length 8 nucleotide) randomly (Sheward et al., 2012). In correlation to this, we observed that Run1 dataset patient B had 1103 Primer IDs while in Run2 patient G had 252 Primer IDs that were repeated between two or more amplicons of the same patient (**Table 3.8**). If two or more duplicated Primer IDs tagged an 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)**.**

## Conclusion

Primer ID is a novel technology for correcting sequencing errors in UDPS 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 showed that Primer ID technology has ability to reduce errors but was unable to represent the true variants spectrum in the viral quasispecies due to technical artifacts.

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.

Bentley, DR, Balasubramanian, S, Swerdlow, HP, Smith, GP, Milton, J, Brown, CG, Hall, KP, Evers, DJ, Barnes, CL, Bignell, HR, Boutell, JM, Bryant, J, Carter, RJ, Keira Cheetham, R, Cox, AJ, Ellis, DJ, Flatbush, MR, Gormley, NA, Humphray, SJ, Irving, LJ, Karbelashvili, MS, Kirk, SM, Li, H, Liu, X, Maisinger, KS, Murray, LJ, Obradovic, B, Ost, T, Parkinson, ML, Pratt, MR, Rasolonjatovo, IM, Reed, MT, Rigatti, R, Rodighiero, C, Ross, MT, Sabot, A, Sankar, SV, Scally, A, Schroth, GP, Smith, ME, Smith, VP, Spiridou, A, Torrance, PE, Tzonev, SS, Vermaas, EH, Walter, K, Wu, X, Zhang, L, Alam, MD, Anastasi, C, Aniebo, IC, Bailey, DM, Bancarz, IR, Banerjee, S, Barbour, SG, Baybayan, PA, Benoit, VA, Benson, KF, Bevis, C, Black, PJ, Boodhun, A, Brennan, JS, Bridgham, JA, Brown, RC, Brown, AA, Buermann, DH, Bundu, AA, Burrows, JC, Carter, NP, Castillo, N, Chiara, ECM, Chang, S, Neil Cooley, R, Crake, NR, Dada, OO, Diakoumakos, KD, Dominguez-Fernandez, B, Earnshaw, DJ, Egbujor, UC, Elmore, DW, Etchin, SS, Ewan, MR, Fedurco, M, Fraser, LJ, Fuentes Fajardo, KV, Scott Furey, W, George, D, Gietzen, KJ, Goddard, CP, Golda, GS, Granieri, PA, Green, DE, Gustafson, DL, Hansen, NF, Harnish, K, Haudenschild, CD, Heyer, NI, Hims, MM, Ho, JT, Horgan, AM, Hoschler, K, Hurwitz, S, Ivanov, DV, Johnson, MQ, James, T, Huw Jones, TA, Kang, GD, Kerelska, TH, Kersey, AD, Khrebtukova, I, Kindwall, AP, Kingsbury, Z, Kokko-Gonzales, PI, Kumar, A, Laurent, MA, Lawley, CT, Lee, SE, Lee, X, Liao, AK, Loch, JA, Lok, M, Luo, S, Mammen, RM, Martin, JW, McCauley, PG, McNitt, P, Mehta, P, Moon, KW, Mullens, JW, Newington, T, Ning, Z, Ling Ng, B, Novo, SM, O'Neill, MJ, Osborne, MA, Osnowski, A, Ostadan, O, Paraschos, LL, Pickering, L, Pike, AC, Chris Pinkard, D, Pliskin, DP, Podhasky, J, Quijano, VJ, Raczy, C, Rae, VH, Rawlings, SR, Chiva Rodriguez, A, Roe, PM, Rogers, J, Rogert Bacigalupo, MC, Romanov, N, Romieu, A, Roth, RK, Rourke, NJ, Ruediger, ST, Rusman, E, Sanches-Kuiper, RM, Schenker, MR, Seoane, JM, Shaw, RJ, Shiver, MK, Short, SW, Sizto, NL, Sluis, JP, Smith, MA, Ernest Sohna Sohna, J, Spence, EJ, Stevens, K, Sutton, N, Szajkowski, L, Tregidgo, CL, Turcatti, G, Vandevondele, S, Verhovsky, Y, Virk, SM, Wakelin, S, Walcott, GC, Wang, J, Worsley, GJ, Yan, J, Yau, L, Zuerlein, M, Mullikin, JC, Hurles, ME, McCooke, NJ, West, JS, Oaks, FL, Lundberg, PL, Klenerman, D, Durbin, R, Smith, AJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59.

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

Cock, PJA, Fields, CJ, Goto, N, Heuer, ML, Rice, PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* **38**: 1767-1771.

Drake, JW, Holland, JJ (1999) Mutation rates among RNA viruses. *Proc Natl Acad Sci U S A* **96**: 13910-13913.

Duffy, S, Shackelton, LA, Holmes, EC (2008) Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* **9**: 267-276.

Eid, J, Fehr, A, Gray, J, Luong, K, Lyle, J, Otto, G, Peluso, P, Rank, D, Baybayan, P, Bettman, B (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133–138.

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.

Ho, DD, Neumann, AU, Perelson, AS, Chen, W, Leonard, JM, Markowitz, M (1995) Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**: 123-126.

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

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

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

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

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

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

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

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

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

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

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

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

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

Margulies, M, Egholm, M, Altman, WE, Attiya, S, Bader, JS, Bemben, LA, Berka, J, Braverman, MS, Chen, Y-J, Chen, Z, Dewell, SB, Du, L, Fierro, JM, Gomes, XV, Godwin, BC, He, W, Helgesen, S, Ho, CH, Irzyk, GP, Jando, SC, Alenquer, MLI, Jarvie, TP, Jirage, KB, Kim, J-B, Knight, JR, Lanza, JR, Leamon, JH, Lefkowitz, SM, Lei, M, Li, J, Lohman, KL, Lu, H, Makhijani, VB, McDade, KE, McKenna, MP, Myers, EW, Nickerson, E, Nobile, JR, Plant, R, Puc, BP, Ronan, MT, Roth, GT, Sarkis, GJ, Simons, JF, Simpson, JW, Srinivasan, M, Tartaro, KR, Tomasz, A, Vogt, KA, Volkmer, GA, Wang, SH, Wang, Y, Weiner, MP, Yu, P, Begley, RF, Rothberg, JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.

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.

Metzner, KJ, Giulieri, SG, Knoepfel, SA, Rauch, P, Burgisser, P, Yerly, S, Gunthard, HF, Cavassini, M (2009) Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and -adherent patients. *Clin Infect Dis* **48**: 239-247.

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.

Onafuwa-Nuga, A, Telesnitsky, A (2009) The remarkable frequency of human immunodeficiency virus type 1 genetic recombination. *Microbiol Mol Biol Rev* **73**: 451-480, Table of Contents.

Perelson, AS, Neumann, AU, Markowitz, M, Leonard, JM, Ho, DD (1996) HIV-1 Dynamics in Vivo: Virion Clearance Rate, Infected Cell Life-Span, and Viral Generation Time. *Science* **271**: 1582-1586.

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.

Shafer, RW (2009) Low-abundance drug-resistant HIV-1 variants: finding significance in an era of abundant diagnostic and therapeutic options. *J Infect Dis* **199**: 610-612.

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.

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.