Accurate HIV-1 population diversity analysis using deep sequencing with Primer ID approach

Aleksandra Vasileva[1], Ilya Bizin[1], Oleg Talalov[1], Andrei Kozlov[2,3], Alexey Masharsky[2], Dmitrij Frishman[5,4,1]

[1]*Laboratory of Bioinformatics, RASA Research Center, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, 195251, Russia, aleksandra.vasileva.spbpu@gmail.com*

[2]*Laboratory of Molecular Virology and Oncology, CAS, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, 195251, Russia*

[3]*The Biomedical Center, St. Petersburg, 194044, Russia*

[4]*Institute for Bioinformatics and Systems Biology, HMGU German Research Center for Environmental Health, Neuherberg, Germany*

[5]*Department of Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Freising, Germany, d.frishman@wzw.tum.de*

In 2010 the group of Prof. A. P. Kozlov began studying the transmission of HIV-1 variants in longitudinal blood samples of injection drug users (IDUs) [1]. In order to shed light on the viral dynamics at acute and chronic infection phases the blood samples were analyzed using the single genome amplification (SGA) approach. Full-length *env* gene sequences of HIV-1 were obtained for each blood sample. Phylogenetic analysis of these sequences revealed the transmission of a single viral variant in 70% of cases. This phenomenon, called genetic bottleneck, implies that there is a strict selection against transmitted viruses. Previously, this phenomenon has been extensively studied for the sexual transmission route of the virus [2]. In this case, the genetic bottleneck can be explained by a physiological barrier as viral transmission through the mucosa has low efficiency. However, there are no physiological barriers associated with the parenteral HIV transmission and the selective factors causing genetic bottleneck along this route are currently unknown.

SGA is labor-intensive low-throughput approach, which may miss minor virus variants, resulting in genetic uniformity at the onset of the infection. We therefore resorted to deep sequencing to confirm the HIV genetic bottleneck phenomenon for IDUs. Due to its high resolution deep sequencing allows for a more accurate detection of minor variants, but it

suffers from high rate of sequencing errors, which can lead to an artificial increase in population diversity.

A common approach to alleviating the effect of sequencing errors is the use of specific molecular barcodes, the so called Primer IDs [3], to tag each viral cDNA. The Primer ID sequence is usually 8-9 nucleotides in length. Following cDNA amplification and sequencing, the reads with the same Primer IDs are considered to be copies of the same original cDNA and are used to create a consensus sequence in order to reconstruct the original cDNA sequence while eliminating sequencing errors, as illustrated in Figure 1.

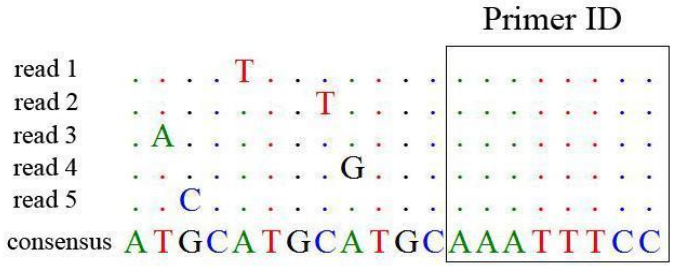

Fig. 1. Consensus creation for reads tagged with the same Primer ID sequence. All PCR biases and sequencing errors introduced during amplification and sequencing of viral templates are removed in the consensus sequence.

The Primer ID approach, while efficient, also has certain drawbacks. For example, sequencing errors may affect Primer ID sequences themselves. To solve this problem reads are subjected to stringent filtering, with all reads not beginning with the exact primer sequence being removed [4]. This procedure can result in the loss of a large number of otherwise high-quality reads, leading to a biased assessment of the viral population.

In order to address this challenge, we have developed an improved computational pipeline to support the application of the Primer ID approach in virology while avoiding loss of data. The pipeline relies on a combination of publicly available tools and several in-house algorithms developed in the Python programming language, with some functions of the Biopython package [5] additionally used. The pipeline involves the following processing steps:

1. Each read is aligned against the reference sequence by the BWA [6] algorithm, or dropped if it cannot be mapped;
2. Primer ID sequences are identified in each read;
3. Reads with errors (insertion/deletion) in Primer ID sequences are removed;
4. Reads get split into Primer ID groups, such that each group contains reads with identical Primer ID sequences;
5. Multiple alignment of reads within each Primer ID group is calculated by mafft [7];
6. Forward and reverse reads are combined in one long marker gene fragment and a consensus cDNA sequence is derived for each set of aligned reads with the same Primer ID;
7. Consensus sequences with bad quality positions are excluded from consideration;
8. Consensus sequences that are hybrids of several different cDNA sequences are removed. Such hybrids are caused by a Primer ID sequence attaching to several different cDNAs;
9. Multiple alignment of all consensus sequences from a given sequencing sample with additional reordering according to the sequence similarity is computed by mafft;
10. The identical consensus sequences get split into groups and a representative sequence from each group is picked for downstream analysis. These representative sequences are used for visualization of the results.

The main advantage of our approach is that it allows to keep as many reads as reasonably possible. For example, instead of analyzing the entire sequence segment in the beginning of the read (including all additional primers for cDNA synthesis and PCR amplification), we only identify the part of the read that corresponds to the Primer ID according to the alignment with the reference sequence. The reference for this step of analysis consists of additional primers, a universal Primer ID designation (e.g. ‘NNN NNN NN’ for Primer IDs consisting of 8 random nucleotides), and a marker gene. The candidate Primer ID sequence is required to have exactly the same length as the utilized Primer ID, and be free of insertions/deletions.

Work is in progress to enhance the pipeline described above based on a mathematical model of the Primer ID approach, which will allow to benchmark our results against synthetic sequencing data.

This work is supported by Russian Science Foundation under grant 15-14-00026.

1. A.E.Masharsky et al. (2010) A substantial transmission bottleneck among newly and recently HIV-1-infected injection drug users in St Petersburg, Russia. T*he Journal of Infectious Diseases,* **201:** 1697–1702.

2. B.F.Keele et al. (2008) Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences USA*, **105:** 7552–7557.

3. C.B.Jabara et al. (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences USA*, **108:** 20166–20171.

4. S.Zhou et al. (2015) Primer ID Validates Template Sampling Depth and Greatly Reduces the Error Rate of Next-Generation Sequencing of HIV-1 Genomic RNA Populations. *Journal of Virology,* **89:** 8540–8555.

5. P.A.Cock et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25:** 1422–1423.

6. H.Li, R.Durbin (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics,* **26:** 589–95.

7. K.Katoh et al. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research,* **30:** 3059–3066.