Novel Computational Approach for Identification of Highly Mutated Integrated HIV Genomes

Khanipov K., Albayrak L., Golovko G., Pimenova M., Fofanov Y.
Department of Toxicology and Pharmacology
University of Texas Medical Branch
Galveston, TX – United States
kakhanip@utmb.edu

Khanipov K., Pavlidis I.
Department of Computer Science
University of Houston
Houston, TX – United States
kkhanipov@uh.edu

***Abstract*—More than 70 million people have been infected with the human immunodeficiency virus (HIV). There is no cure for HIV and modern treatment is only effective at delaying the onset of acquired immunodeficiency syndrome. Incorporated HIV serves as a reservoir for constant release of virions. Knowing the locations and quantity of HIV in the reservoirs can help guide development of complete treatment. Patient/Organ-specific HIV genome reconstruction allows to significantly improve detection of sequences originating from HIV. In this paper, we present a novel personalized medicine approach based on the reconstruction of patient/organ-specific HIV genome sequences in latent reservoirs.**

HIV; Human immunodeficiency virus; Viral genome reconstruction; Integrated genome reconstruction

I. Introduction

Since the beginning of the HIV epidemic in 1987, more than 70 million people have been infected and about 35 million people have died as a result. The key challenge in the treatment of this infection is due to its life cycle. HIV integrates itself into various locations of the human genome and these “silent copies” can be reactivated months or years later to produce more virions. This results in the presence of latent reservoirs in various body sites called “latent pools” which are established within the first 2 weeks of the infection. Longitudinal studies have shown that the decay rate of a pool of latently infected cells has a half-life of 44 months [1, 2]. Over 70 years of treatment would be required to eradicate the latent reservoirs. Additionally, treatment adherence levels greater than 95% are required to maintain virologic suppression, but actual adherence rates are often far lower; most studies show that 40–60% of patients are less than 90% adherent and adherence also tends to decrease over time [3]. The key to curing the HIV infection is being able to combine highly active retroviral therapy (which eliminates the live virus) and depletion of latent viral pools. Knowing the chromosomal locations and quantity of HIV incorporated in various human organs (e.g., such as brain, liver, or spleen) can provide critical insight to reconstruct the mechanism of infection and design more effective treatment.

To date, latent pool and HIV characterization have been performed using standard molecular approaches such as real-time PCR [4] and in situ hybridization [5] that require a priori knowledge of the viral sequence. Unfortunately, due to extremely high mutation rate: up to 1% per site per year for infected individual for replication-competent virus and even higher for integrated viral sequences makes the efficacy and repeatability of these approaches depend on how close the virus present in the individual is to the sequences used in the PCR and in situ hybridization test design. Considering the duration of the infection (up 35 years) the virus can accumulate up to 30% of mutations which make every 3-rd nucleotide different from the originally infected sequence, the chance that viral sequence will be missed is extremely high [6, 7]. For example, quantitative viral outgrow assays underestimate intact proviral HIV sequences in the 10^2-10^3 orders of magnitude. While, digital PCR on gag gene overestimated intact proviral HIV sequences by 10-10^2 orders of magnitude, meaning that at least 90% of HIV genomes are not intact or fit for viral replication [8].

One of the recent and most effective approaches to estimate the frequency and chromosomal location of incorporated HIV DNA is capture-sequence-map approach where hundreds of HIV-complementary DNA subsequences (probes) are used to capture of HIV-like DNA fragments from the sample followed by whole-genome sequencing of these fragments using High Throughput Sequencing (HTS) instruments [9]. De-novo assembly and reference based mapping are two approaches which can be generally used to reconstruct patient-specific (or even tissue-specific in an individual patient) HIV genome. The high mutation rate, however, remains the significant obstacle in the identification of specific HIV integration sites and overall characterization of the latent reservoir of the infection.

In this manuscript, we present a novel approach for the reconstruction of the patient/tissue-specific HIV-1 genomics sequences based on the iterative use of the combination of reference and de-novo assembly strategies.

II. Iterative Genome Reconstruction Algorithm

The basic idea of the proposed algorithm is to iteratively identify all the sequencing reads which with a high probability belong to the virus and perform de-novo assembly using these reads (Figure 1). In the initial step of the algorithm, sequencing reads are loosely mapped to 2,253 known HIV-1 sequences to create an initial pool of reads which may belong to the viral genome and use them to create the first iteration of the sample/patient-specific reference sequence. Sequentially, these sequences will be used to identify (by mapping) a new (improved) pool of reads to be used in the second assembly and this process will be repeated

2471-7819/17/31.00 ©2017 IEEE
DOI 10.1109/BIBE.2017.00048

until newly assembled sequence will not change anymore (Figure 1).

III. Materials and Methods

A database of all previously publish HIV 1 genomes was downloaded on March 5, 2017, from the Los Alamos National Laboratory HIV database. A total of 2,494 HIV 1 genomes were downloaded. The use of multiple references allowed to increase the number of unique 100 nucleotide long subsequences available for mapping from approximately 1,218 to 1,340,565. The approach has been implemented as a workflow in CLC Genomics workbench (clcbio.com).

IV. Results and Discussion

Patient/Organ-specific HIV genome reconstruction allows to significantly improve the quantity and quality of detection of HIV originating sequences. The proposed approach allowed us to increase the number of detected sequenced reads originating from HIV by 10^2-10^4. In experiments performed using sequencing reads originating from clinical samples, mapping to a whole possible set of HIV sequences increases average number of reads mapped with 90% alignment identity by 50-fold compared to when using a single reference.

In many real cases where the proposed approach has been tested, de-novo assembly from all the reads as well as assembly using reads with human background excluded does not lead to any results. Mapping reads to single reference genome also does not produce enough coverage so the individual reference sequence cannot be identified. However, using thousands of reference genomes to identify reads potentially originated from the target HIV-1 virus could successfully reconstruct the patient specific viral sequence.

It is also important to mention that proposed approach can be used to explore and characterize other aspects of the virus (e.g., heteroplasmy and quasispecies) and opens the opportunity to distinguish between partial and complete (replication competent) incorporations.

References

[1] D. Finzi et al., "Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy," (in eng), Nat Med, vol. 5, no. 5, pp. 512-7, May 1999.

[2] J. D. Siliciano et al., "Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells," (in eng), Nat Med, vol. 9, no. 6, pp. 727-8, Jun 2003.

[3] J. A. Bartlett, "Addressing the challenges of adherence," JAIDS Journal of Acquired Immune Deficiency Syndromes, vol. 29, pp. S2-S10, 2002.

[4] L. Chavez, V. Calvanese, and E. Verdin, "HIV latency is established directly and early in both resting and activated primary CD4 T cells," PLoS pathogens, vol. 11, no. 6, p. e1004955, 2015.

[5] A. Chargin, F. Yin, M. Song, S. Subramaniam, G. Knutson, and B. K. Patterson, "Identification and characterization of HIV-1 latent viral reservoirs in peripheral blood," Journal of clinical microbiology, vol. 53, no. 1, pp. 60-66, 2015.

[6] G. Li et al., "An integrated map of HIV genome-wide variation from a population perspective," Retrovirology, vol. 12, no. 1, p. 18, 2015.

[7] A. Gueler et al., "Life expectancy in HIV-positive persons in Switzerland: matched comparison with general population," AIDS (London, England), vol. 31, no. 3, p. 427, 2017.

[8] K. M. Bruner et al., "Defective proviruses rapidly accumulate during acute HIV-1 infection," Nature medicine, vol. 22, no. 9, pp. 1043-1049, 2016.

[9] P. Miyazato et al., "Application of targeted enrichment to next-generation sequencing of retroviruses integrated into the host human genome," Scientific reports, vol. 6, 2016.

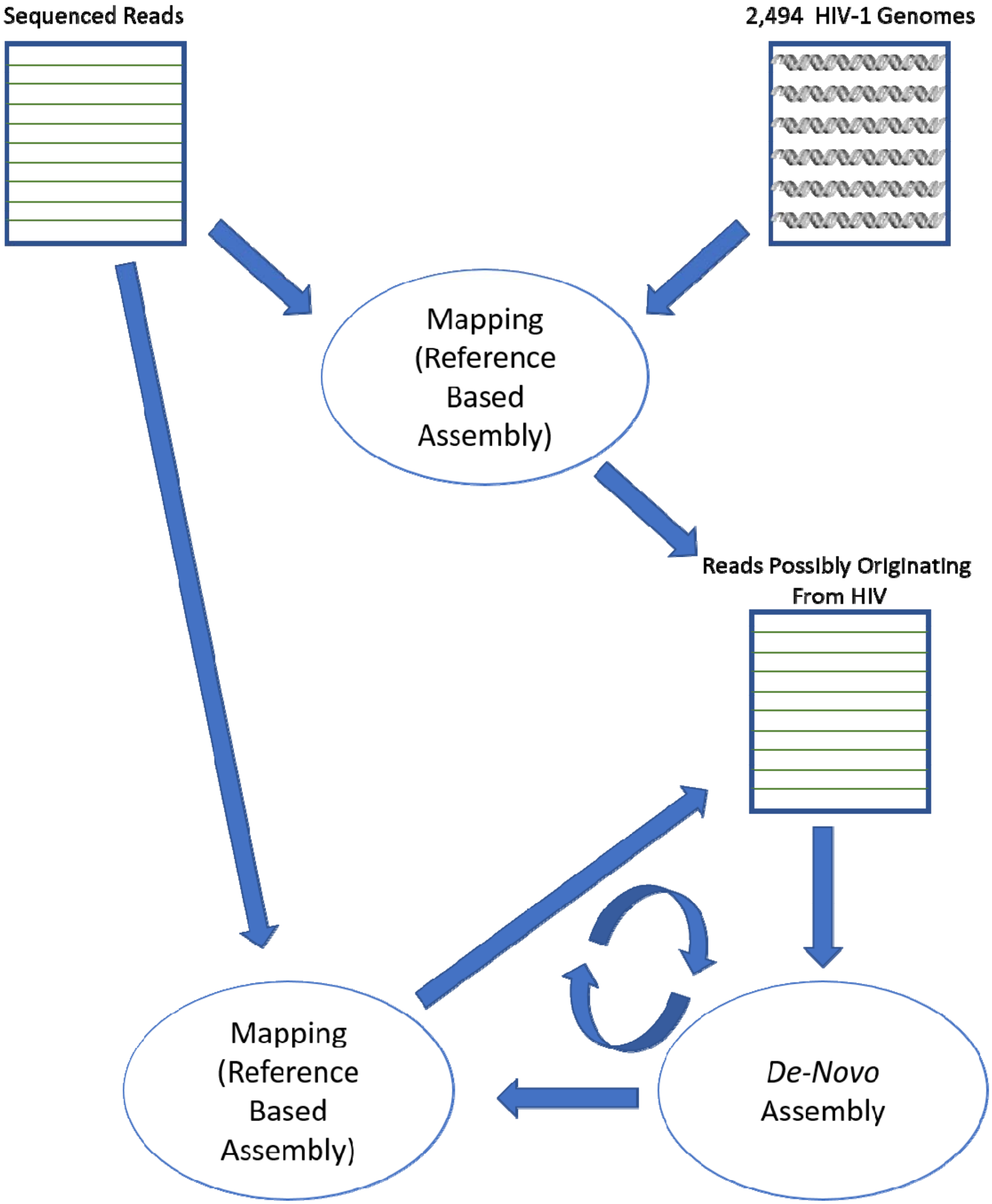
Sequenced Reads
2,494 HIV-1 Genomes
Mapping (Reference Based Assembly)
Reads Possibly Originating From HIV
Mapping (Reference Based Assembly)
De-Novo Assembly

Figure 1. Iterative Genome Reconstruction Workflow