Characterization of highly diverse viral populations by fast reference selection and accurate read mapping

Gennady Fedonin

Federal Budget Institution of Science "Central Research Institute for Epidemiology", Moscow, Russia,

Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia, gennady.fedonin@gmail.com

Alexey Neverov

Federal Budget Institution of Science "Central Research Institute for Epidemiology", Moscow, Russia,

Department of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University,

Moscow, Russia, neva_2000@mail.ru

Ability to characterize genetic composition of viral populations is critical to the selection of effective treatment of various viral infections. Previous studies [1-9] have shown that many viruses including Dengue, HCV, HIV, Influenza, Polio and West Nile all maintain diverse populations within a single host. Genome wide deep sequencing can be used to detect low frequency variants resistant to various drug treatments in such heterogeneous viral population. The sequencing data can be aligned to a reference for variant detection, but the use of a reference genome that is too genetically distant from the sample population may yield inaccurate read alignments and substantial data loss; both factors decrease the ability to detect biological variants [10,11]. Because viral consensus can vary substantially between patients, using of any fixed reference may be problematic, and it may be difficult to choose reference from existing ones that allows good alignment of reads [2,12]. One solution is to start the analysis by *de novo* assembly of each patient sample, allowing use of the patient consensus as the reference for variant detection [2]. Standard *de novo* assemblers are highly optimized for long homogeneous genomes and show poor performance in comparison to specially designed ones, like VICUNA [13]. VICUNA can produce good assemblies, but preliminary read filtering is needed for best performance, which involves computing of similarity of all reads to reference MSA. Still, contamination, presence of chimerical reads and unevenness of genome coverage due to PCR amplification of target DNA needed for clinical samples may significantly decrease performance of *de novo* assembly. VICUNA also failed to assemble viral metagenomes: reads from different genomes should be separated by mapping to appropriate reference sequences [18].

In our research we propose alternative solution mainly based on read mapping to selected reference, but also performing local *de novo* assembly in highly variable regions. First, one or more references, closest to sample population, are selected from given reference MSA using fast read localization algorithm and overlap-layout-consensus like contig assembly. All reads are assigned to selected references allowing one read to be assigned to multiple references. On second step we use read mapping to built whole-genome consensus sequences and reassemble regions with poor alignment. When final consensus sequences are ready all reads are remapped and resulting alignments can be used for variant calling and further analysis.

Read mapping. Most existing aligners utilize hashing algorithms [14] or the Burrows-Wheeler transform [15,16] to search exact matches as their first step. We use hashing approach: to map read we first quickly localize it on genome using hash-table storing all short k-mers of genome with their genome positions and then use Smith-Waterman algorithm to precisely align read to previously found location. We've extended this approach to MSA, allowing fast read localization on reference MSA without slow alignment step.

Reference selection. Sample viral population may contain two or more distant subpopulations. Thus, we search for a minimal set of references, which pairwise identity is less than given threshold and all reads in the sample have identity above the same threshold to at least one reference from this set. To select references we first localize all reads on reference MSA, then cluster closely localized reads with corresponding genome fragments using Uclust [17] with given similarity threshold, build consensus sequences in each cluster (short contigs) and merge them to longer contigs by traversing overlap graph built on short contigs.

Genome-wide viral population consensus building. Selected references may still be too far from sample viral subpopulations. To build closer reference, all reads are mapped to corresponding references and consensus sequences are built for each reference. This procedure can be repeated multiple times, but we found two iterations to be enough for convergence. Then we cut regions with low quality read alignment and reassemble them by greedy algorithm, which join one nucleotide at a time to both edges of the cut using reads already mapped to the edges but having unmapped parts hanging down.

Performance on clinical samples. We tested our method on 51 samples obtained using

Illumina MiSeq from clinical samples of HIV-1 infected patients. These data were provided by the Retroviridae Lentivirus group at the Wellcome Trust Sanger Institute and can be obtained from http://www.ncbi.nlm.nih.gov/sra (submission number ERA012006). Consensus sequences were built using our algorithm for all samples using HXB2 as initial reference and compared with *de novo* assembled VICUNA contigs. Our assemblies have very high identity with vicuna contigs (more than 98%) and usually cover larger part of reference genome in case of low total coverage (Fig. 1).

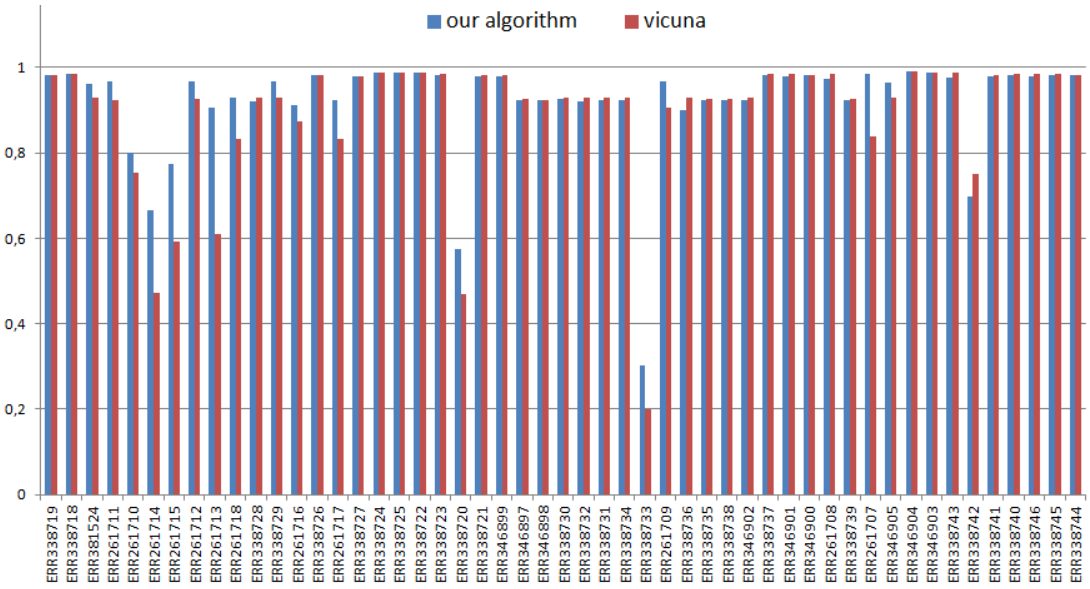

Fig.1 Fraction of reference genome covered by assembly for 51 clinical samples

We also tested our reference selection method using the same data: optimal references were selected from Viral Genotyping Tool HIV-1 2009 Reference Set. From 51 samples being tested for 1 sample with extremely high contamination reference was not selected at all, for 41 samples one sequence was selected, for remaining 9 – two sequences. Viral genotypes were identified from selected references and compared to genotypes of references closest to the assemblies discussed earlier in the same reference set. For 38 samples genotypes are the same, in 4 cases genotypes are different, in 8 cases genotype of one of two selected references coincide with assembly genotype. According to Viral Genotyping Tool these 12 samples have complex recombinant genotype, making it possible to assume these samples to be mixes of different genotypes.

1. Vignuzzi M, Stone JK, Andino R: **Ribavirin and lethal mutagenesis of poliovirus: molecular mechanisms, resistance and biological implications,** *Virus res* 2005, **107**(2)**:**173-181.
2. Henn MR, *et al*.: **Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection,** *PLoS Pathogens* 2012, **8**(3)**:**e1002529.
3. Herbeck JT, *et al*: **Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes,** *J Virol* 2011, **85**(15)**:**7523-7534.
4. Lauck M, *et al*: **Analysis of Hepatitis C Virus Intrahost Diversity across the Coding Region by Ultradeep Pyrosequencing,** *J Virol* 2012, **86**(7)**:**3952-3960.
5. Jerzak G, Bernard KA, Kramer LD, Ebel GD: **Genetic variation in West Nile virus from naturally infected mosquitoes and birds suggests quasispecies structure and strong purifying selection,** *J Gen Virol* 2005, **86**(Pt 8)**:**2175-2183.
6. Murcia PR, *et al*: **Intra- and interhost evolutionary dynamics of equine influenza virus,** *J Virol* 2010, **84**(14)**:**6943-6954.
7. Vignuzzi M, *et al*: **Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population,** *Nature* 2006, **439**(7074)**:**344-348.
8. Lin SR, *et al*: **Study of sequence variation of dengue type 3 virus in naturally infected mosquitoes and human hosts: implications for transmission and evolution,** *J Virol* 2004, **78**(22)**:**12717-12721.
9. Thai KT, *et al*: **High-resolution analysis of intrahost genetic diversity in dengue virus serotype 1 infection identifies mixed infections,** *J Virol* 2012, **86**(2)**:**835-843.
10. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G: **De novo assembly and genotyping of variants using colored de Bruijn graphs,** *Nat Genet* 2012, **44**(2)**:**226-232. (13)
11. Archer J, Rambaut A, Taillon BE, Harrigan PR, Lewis M, Robertson DL: **The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time–an ultra-deep approach,** *PLoS Comput Biol* 2010, **6**(12)**:**e1001022. (14)
12. Willerth SM, Pedro HA, Pachter L, Humeau LM, Arkin AP, Schaffer DV: **Development of a low bias method for characterizing viral populations using next generation sequencing technology,** *PloS one* 2010, **5**(10)**:**e13564. (15)
13. Xiao Yang, *et al*: **De novo assembly of highly diverse viral populations,** *BMC Genomics* 2012, **13**:475.
14. Wan-Ping Lee, *et al*: **MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping**, *PloS one* 2014, **9**(3): e90581.
15. Li H, Durbin R: **Fast and accurate short read alignment with BurrowsWheeler transform**, Bioinformatics 2009, 25: 1754–1760.
16. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2,** Nat Methods 2012, 9: 357–360.
17. Edgar, R.C. **Search and clustering orders of magnitude faster than BLAST**, Bioinformatics 2010, 26(19): 2460-2461.
18. Jorge F Vázquez-Castellanos, *et al***: Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut**, *BMC Genomics* 2014, **15**:37