Analysis of HIV-1 quasispecies sequences generated by High-throughput Sequencing (HTS) using HIVE

Naila Gulzar1, Bhavna Hora2, Konstantinos Karagiannis1, Krista Smith1, Feng Gao2, Raja Mazumder1
1Department of Biochemistry and Molecular Medicine, The George Washington University, Washington, DC, 20037
2Duke Human Vaccine Institute and Departments of Medicine, Duke University Medical Center, Durham, NC 27715, USA

ABSTRACT

The high level of genetic variability of Human Immunodeficiency Virus type 1 (HIV-1) is caused by the low fidelity of its replication machinery. This leads to evolution of swarm-like viral populations often described as quasispecies. High-throughput sequencing (HTS) technology provides higher resolution over Sanger sequencing, enabling detection of low frequency variant genomes. However, quasispecies analysis is still a challenge due to the systematic noise, introduced by HTS technology. This leads to the increase in type I errors (also known as false positives) and the underestimation of genetic diversity, which can lead to overestimation of type I errors (also known as false negatives). We have developed a pipeline using the tools in the High-performance Integrated Virtual Environment (HIVE), an HTS tool designed for big data analysis and management, to analyze viral sequence data from each sample and identify the sequence patterns and recombination patterns of recombinants. RNA was extracted from 70 plasma samples of chronic HIV-1 infected patients. The 3′ half-genome of HIV-1 were amplified using RT-PCR and PCR products were sequenced using Illumina MiSeq. The paired-end reads for each sample were assembled using Genium software and analyzed for presence of HIV-1 quasispecies using HIVE tools. Subtype analysis of 70 samples using Genium software identified 17 A, 4 B, 30 C, 1 D, 0 CRF02_AG, and 12 unique recombinant forms (URFs). Additionally, we found up to 178 ambiguous sites in the consensus sequences from 41 viral samples (58.6%), suggesting the presence of viral subpopulations. However, Genium could not determine the major quasispecies populations in each sample. We analyzed the same HTS reads using the HIV-1 quasispecies analysis pipeline and found one predominant population in 11 samples (15.7%), two to ten distinct populations in 45 samples (64.3%), 11-20 in 13 samples (18.6%), and 21 in one sample. Interestingly, two equally major viral populations that were not detected by Genium were identified in five samples (7.1%) by HIVE. The HIV-1 quasispecies analysis pipeline is reliable and more sensitive in its ability to identify distinct viral populations and the recombination patterns not identified by the Genium software.

METHOD

RNA was extracted from 70 plasma samples of HIV-1 infected patients. The 3′ half-genome of HIV-1 were amplified using RT-PCR and PCR products were sequenced using Illumina MiSeq. The paired-end reads for each sample were assembled using Genium software and analyzed for presence of HIV-1 quasispecies using HIVE tools. The quality control analysis of the raw reads was done in HIVE and then the reads were mapped to the consensus sequence HIVE hexagon. The nearest neighbors were selected from the alignment results and then raw reads were mapped to the nearest neighbors. Clonal discovery tool, Hexagon discovery was run on the alignment using mutual frame of the nearest neighbors for co-linear system.

HIV-1 ANALYSIS PIPELINE

HTS raw reads
Step 1: Upload files to HIVE
Step 2: Automatic QC analysis
Step 3: Raw reads alignment
20 reference strains
Iterative loop (Repeated within different parameters)
Step 6B: Recombinant mosaic discovery
Step 6A: Clonal discovery
Hexagon tool
Step 7: Report dominant subtype/recombinant
Step 4: Selection of Nearest Neighbors and mutual alignment of Nearest Neighbors using MAFFT
Step 5: Align samples to Nearest Neighbors

RESULTS

Figure 2. Sankey plots generated in HIVE using hexagon tool depicting subtype, quasispecies and recombinant patterns. Sample PK006 (A) depicts sample PK006 having two predominant populations of subtype A1. (B) two subtypes, A1 and C observed in sample PK016. (C) recombination between A1 and CRF02_AG seen in sample PK006 and (D) single predominant A1 population in sample PK015. Width of the sankey plot is reference to the coverage by reads in the area. For each sample different colors represent different subtypes.

Figure 3. Identification of two major and five minor quasispecies populations, recombination patterns and genetic distances between the clones. Two predominant clones, clone 1 and 2 were verified as representative of subtype A1 and C using REGA subtyping tool. Phylogenetic trees were constructed for each minor clone with the overlapping clones in the region using Mega software (D-H) to estimate the genetic distances between major and minor clonelines.

Figure 4. Occurrence of quasispecies populations in 70 samples. Percent samples with minor clonal populations are shown.

Table 1. Analysis of genetic populations in 70 chronic infection samples

<table>
<thead>
<tr>
<th>Sample</th>
<th>VL</th>
<th>Subtype</th>
<th>Predominant populations</th>
<th>Genetic distances</th>
</tr>
</thead>
</table>
| SampleA | 70 | A1 | A1 | 100%
| SampleB | 40 | A1 | A1 | 100%
| SampleC | 30 | A1 | A1 | 100%
| SampleD | 20 | A1 | A1 | 100%

ACKNOWLEDGEMENT


SUMMARY

- In chronic HIV-1 infection samples multiple viral populations (1-26) were detected.
- Out of 70 samples, five (7.1%) samples were identified with two predominant viral populations by HIVE, not identified by Genium.
- On comparing the subtyping and recombination analysis results of the predominant populations obtained from HIVE with the consensus sequences obtained by Genium, the results were found comparable.

CONCLUSIONS

The HIV-1 quasispecies analysis pipeline is reliable and more sensitive in its ability to identify distinct major and minor virus quasispecies populations not identified by Genium software and determine the recombination patterns between different subtypes within the viral quasispecies populations in each sample.