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

Naila Gulzar[1], Bhavna Hora[2], Konstantinos Karagiannis[1], Krista Smith[1], Feng Gao[2], Raja Mazumder[1]

•[1]Department of Biochemistry and Molecular Medicine, The George Washington University, Washington, DC, 20037

•[2]Duke Human Vaccine Human Institute and Departments of Medicine, Duke University Medical Center, Durham, NC 27710, 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 underlying genetic diversity, which can lead to mathematically insolvable type II errors (also known as false negatives). We have developed a pipeline using the tools in the High-performance Integrated Virtual Environment (HIVE), an HTS platform designed for big data analysis and management, to analyze viral populations within each sample and identify their subtype classification and recombination patterns of recombinants. RNA was extracted from 70 plasma samples of chronic HIV-1 infected patients. The 3' half genomes 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 Geneious software and analyzed for presence of HIV-1 quasispecies using HIVE tools. Subtype analysis of 70 samples using Geneious software identified 17 A1s, 4 Bs, 30 Cs, 1 D, 6 CRF02_AG, and 12 unique recombinant forms (URFs). Additionally, we found up to 178 ambiguous bases in the consensus sequences from 41 viral samples (58.6%), suggesting the presence of viral subpopulations. However, Geneious 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 (18.16%), and 26 in one sample (1.4 %). Interestingly, two equally major viral populations that were not detected by Geneious 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 Geneious software.

BACKGROUND

High-throughput sequencing (HTS) technology has the ability to characterize HIV-1 genome sequences with higher resolution over Sanger sequencing. However, quasispecies analysis of half or whole genome sequence is still a challenge due to the systematic noise introduced by the HTS technology. We have developed a pipeline to identify the distinct viral populations, subtypes and recombination breakpoints in each sample using the High-performance Integrated Virtual Environment (HIVE).

AIM

The aim of this study is to design a workflow which is reliable and more sensitive in identifying the major and minor viral quasispecies in the HIV-1 infected samples using the HTS data.

METHOD

RNA was extracted from 70 plasma samples of chronic HIV-1 infected patients. The 3' half genomes 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 Geneious 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 20 reference set using HIVE-hexagon. The nearest neighbors were selected from the alignment results and then raw reads were mapped to the nearest neighbors. Clonal discovery tool, Hexahedron was run on the alignment using mutual frame of the nearest neighbors for co-ordinate system.

HIV-1 ANALYISIS PIPELINE

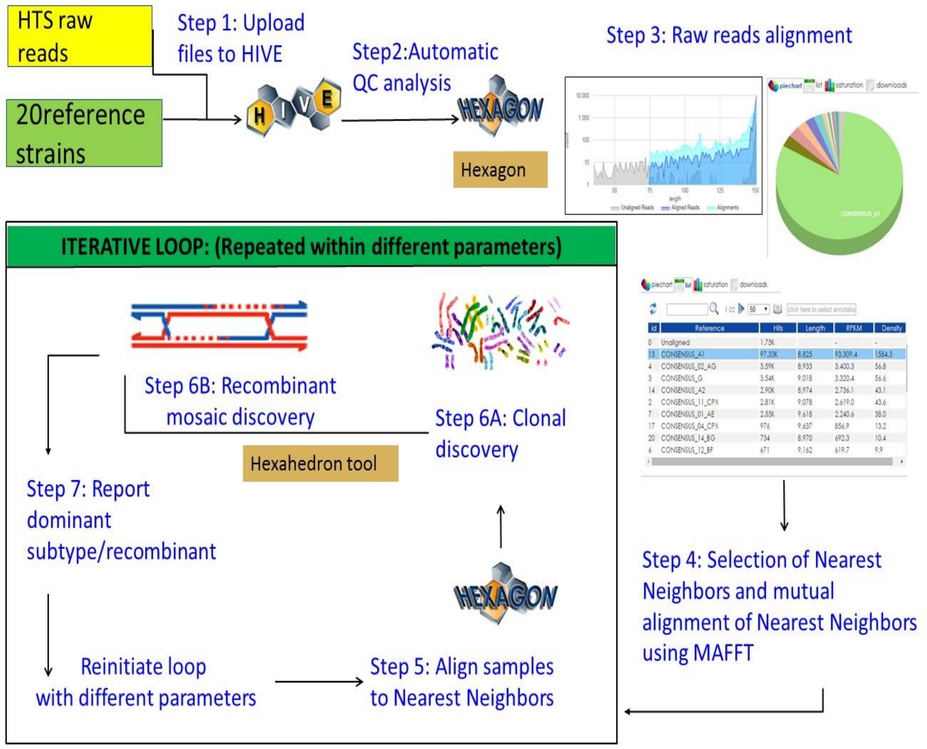

Figure 1. The workflow for the identification of quasispecies, subtypes and recombination breakpoints in HIV-1 infected patient samples in HIVE

RESULTS

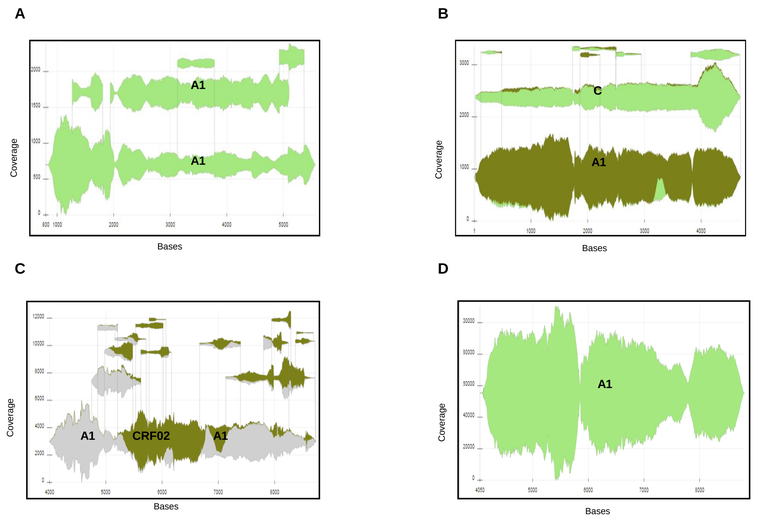

Figure 2. Sankey plots generated in HIVE using hexahedron tool depicting subtype, quasispecies and recombination pattern. Sankey plot (A) depicts sample PK040 having two predominant populations of subtype A1, (B) two subtypes, A1 and C observed in sample 707010038, (C) recombination between A1 and CRF02 seen in sample PK006 and (D) single predominant A1 population in sample PK016. Width of the sankey plot is a reference to the coverage by reads in the area. For each sample different colors represent different subtypes.

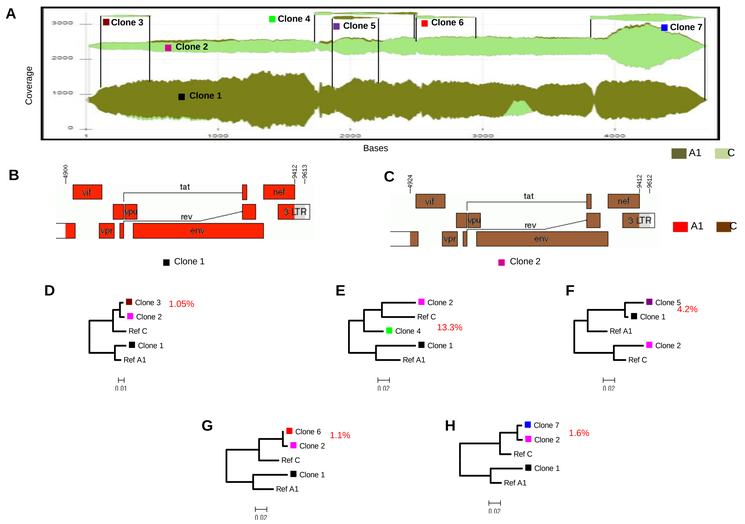

Figure 3. Identification of two major and five minor quasispecies populations, recombination patterns and genetic distances between the clones. Two predominant clones: clone 1 and 2 were verified as representative of subtype A1 and C using REGA subtyping tool. Phylogenetic trees were constructed for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

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

			HIVE		Geneious
Sample	VL	Subtype	Predominant populations	No. of clonal populations	Ambiguous bases
703010957	95,554	C	1	1	0
704010566	26,700	C	1	1	0
707010225	2,417	D	1	1	1
PK005	610	CRF02	1	1	2
PK009	173,000	C	1	1	0
PK011	8,750	URF	1	1	2
PK016	13,000	A1	1	1	0
PK026	5,350	A1	1	1	26
PK032	6,400	CRF02	1	1	44
PK034	4,665	A1	1	1	28
PK036	96,500	A1	1	1	20
706010391	271,000	C	1	2	0
PK021	93,500	A1	1	2	42
PK025	26,900	URF	1	2	17
703010234	183,452	C	1	3	0
703010619	29,059	C	1	3	0
703010835	32,974	C	1	3	3
703011871	150,182	C	1	3	0
705010614	98,700	C	1	3	0
707010277	97,122	URF	1	3	0
PK004	464,000	A1	1	3	4
PK040	111,000	A1	2	3	43
702010675	146,346	C	1	4	5
703010523	88,878	C	1	4	0
705010474	40,497	C	1	4	0
707010169	28,598	A1	1	4	0
707010789	22,880	A1	1	4	5
PK039	60,000	CRF02	1	4	11
700010516	10,310	B	1	5	0
705010366	15,329	C	1	5	0
PK031	225,500	A1	1	5	0
PK035	461,500	CRF02	1	5	3
702010118	32,081	C	1	6	2
705010303	26,833	C	1	6	0
705010381	21,352	C	1	6	0

			HIVE		Geneious
Sample	VL	Subtype	Predominant populations	No. of clonal populations	Ambiguous bases
705010661	31,081	C	1	6	0
705010801	4,704	C	1	6	0
706010375	119,000	C	1	6	1
PK007	106,000	A1	1	6	2
PK038	456,000	CRF02	1	6	5
700010501	31,875	B	1	7	0
704010715	319,000	C	1	7	0
PK015	300,000	URF	1	7	6
707010038	17,213	A1 and C	2	7	30
702010322	57,901	C	1	8	4
706010413	78,000	C	1	8	0
PK023	65,500	URF	1	8	28
PK003	13,600	URF	2	8	101
702010350	85,381	C	1	9	0
704010486	253,000	C	1	9	1
PK017	119,000	A1	1	9	178
PK027	111,500	A1	1	9	46
703010632	43,600	C	1	10	0
705010699	46,057	C	1	10	4
PK008	625,000	URF	1	10	129
PK033	425,500	URF	1	10	34
700010329	11,014	B	1	11	0
707010134	15,208	URF	2	11	0
PK018	88,500	A1	1	11	79
707010585	79,367	URF	2	14	60
PK006	710,000	URF	1	14	9
PK019	760,000	CRF02	1	15	134
PK002	488,000	A1	1	16	4
700010260	74,588	B	1	18	0
PK013	193,000	A1	1	18	24
705010645	425,760	C	1	19	1
PK020	208,000	A1	1	19	131
702010133	591,344	C	1	20	2
703010539	97,857	C	1	20	0
PK010	1,250	C	1	26	NA

Samples highlighted in red ink contain more than one predominant population.

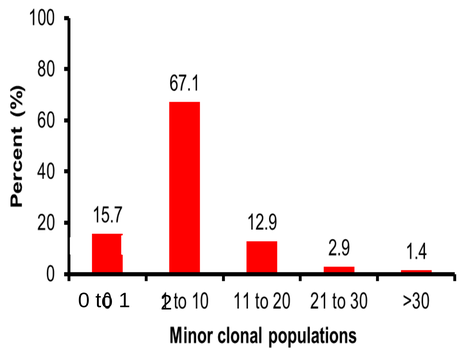

Figure 4. Occurrence of quasispecies populations in 70 samples. Percent samples with minor clonal populations are shown.

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 Geneious.
- On comparing the subtyping and recombination analysis results of the predominant populations obtained from HIVE with the consensus sequences obtained by Geneious, 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 viral quasispecies populations not identified by Geneious software and determine the recombination patterns between different subtypes within the viral quasispecies populations in each sample.

ACKNOWLEDGEMENT

Vahan Simonyan, Yue Chen, Fangping Cai, Chang Su, Sharaf Ali Shah, Manzoor Ahmed, Ana M. Sanchez, Mars Stone, Myron S. Cohen, Barton F. Haynes, Michael Busch, Thomas N. Denny