The HIVE pipeline to measure HIV variant diversity to predict time since infection

Ilya **Mazo,** Luis V **Santana-Quintero,** Konstantinos **Karagiannis,** Indira **Hewlett, and** Viswanath **Ragupathy**

Center for Biologics Evaluation and Research (CBER), Food and Drug Administration (FDA), Silver Spring, MD, USA

Abstract

Background: The project overarching goal is to develop a PCR/NGS based test to measure the genetic diversity in the HIV env protein from patient samples as a metric to decide whether the infection was recent or not. The rationale is that the longer the virus replicates in the patient's body the more it mutates and therefore the genetic diversity increases. Our objective is to provide the bioinformatics pipeline in the FDA HIVE that given the HIV env targeted NGS data from a patient sample would provide the diagnostic answer: short-term vs long-term infection, and in addition would provide an estimate of the time since infection.

Methods: The virus genome NGS pipeline was developed and optimized using targeted sequencing data from a publicly available dataset of 92 longitudinal samples from 11 patients which included a robust infection time estimate. The optimized pipeline, consisting of HIVE Hexagon, HIVE Heptagon, variant diversity tool, and the pre-trained regression model implemented in Python, was extensively tested to optimize the parameter space and applied to the public validation set.

Results: Our HIV variant diversity pipeline produced variant diversity results and time since infection estimates which are in strong agreement with estimates detected in previous HIV studies. The additional tuning to make the pipeline applicable to the env-specific amplicons is underway.

Conclusion: Using the sequencing pre-processing, variant calling, and ML strategies implemented in the HIVE virus diversity pipeline, prediction of the time since infection estimate detection and analysis is an attractive alternative approach to the antibody level based tests currently in clinical trials.

Introduction

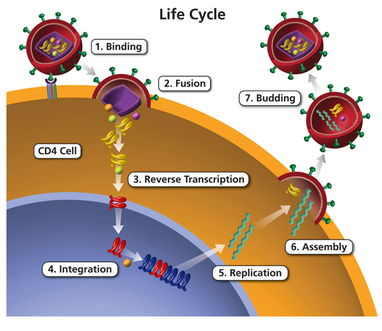

Figure 1. Cycles of HIV reinfection and reverse transcription create mutations and genetic diversity. Longer persistence of the virus is expected to result in more diversity.

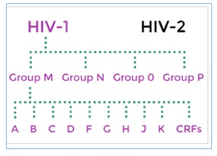

Figure 2. HIV Types and Strains. HIV-1 Group M is the strain responsible for the global HIV epidemic. Group M includes nine subtypes (A b D, F to H, J, and K), five subsubtypes (A1, A2, A3, F1, and F2), and 37 known circulating recombinant forms (CRFs, CRF01 to CRF37).

Materials and Methods

Training Set – multiple datapoints (days since infection) for 11 patients, is described in the works from Neher lab [2,3] who reported successful regression models to predict time since infection from genetic diversity. The FASTQ files were downloaded from [1] and represent multiple HIV amplicons from each of the samples. Our goal was to replicate their results in HIVE.

The pipeline, consisting of (i) HIVE Hexagon [4] and using HBX2 genome as a reference, (ii) HIVE Heptagon [5] using sample-specific consensus as a reference, (iii) variant diversity tool to calculate several diversity metrics (Figure 3) from variant profiles, was applied to every sample, and the data was used (iv) to train a regression model or classifiers to separate recent infection (less than 180 days) from long term infection (>180 days).

Patient	HIV Subtype	No. of Samples	1st Sample [days]	Last Sample [years]
p1	O1_AE	12	122	8.2
p2	B	6	74	5.5
p3	B	10	146	8.4
p4	B	8	78	8.4
p5	B	7	134	5.9
p6	C	7	62	7.0
p7	B	11	1905	15.9
p8	B	7	87	6.0
p9	B	8	106	8.1
p10	B	9	33	6.2
p11	B	7	209	5.6

$$D_A = \frac{1}{L}\sum_{i=1}^{L}\Theta(1-x_i^m-x_c)$$

$$D_H = \frac{1}{L}\sum_{i=1}^{L}\Theta(1-x_i^m-x_c)\left[\sum_{\alpha}x_{i\alpha}(1-x_{i\alpha})\right].$$

$$D_E = -\frac{1}{L}\sum_{i=1}^{L}\Theta(1-x_i^m-x_c)\left[\sum_{\alpha}x_{i\alpha}\log(x_{i\alpha})\right]$$

Figure 3. Genetic Variant Diversity Metrics. $\mathbf{D_A}$ – Average Variant Frequency, $\mathbf{D_H}$ – Diversity Score, $\mathbf{D_E}$ – Shannon Entropy.
α ∈ {A, C, G, T} L – sequence length in nucleotides.
Cutoff xc: Θ(1 − xm − xc) = 1 when 1 − xm > xc and 0 otherwise

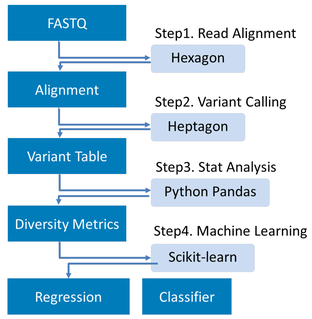

Figure 4. Pipeline to Compute Viral Diversity Metrics.

Results and Discussion

Our HIV variant diversity pipeline produced variant diversity results and time since infection estimates which are in strong agreement with estimates detected in previous HIV studies [1-3].

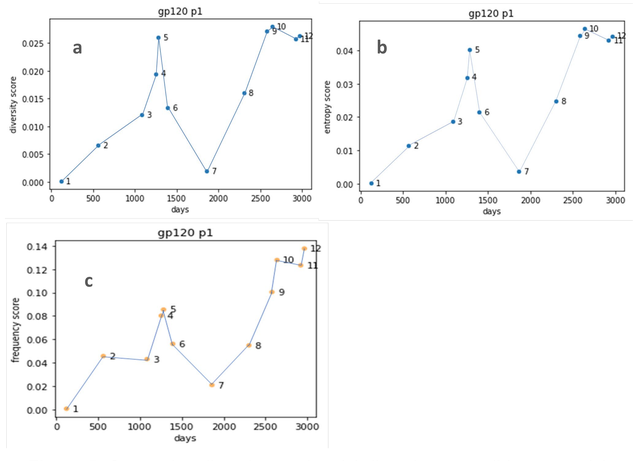

Figure 5. Comparing diversity metrics: (**a**) diversity score, (**b**) entropy, (**c**) frequency score using the 12 datapoints for the patient 1, gp120 amplicons. All metrics show high degree of correlation between them. Low diversity might be due, in addition to biological reasons, to a low RT-PCR efficiency in a specific amplicon [1].

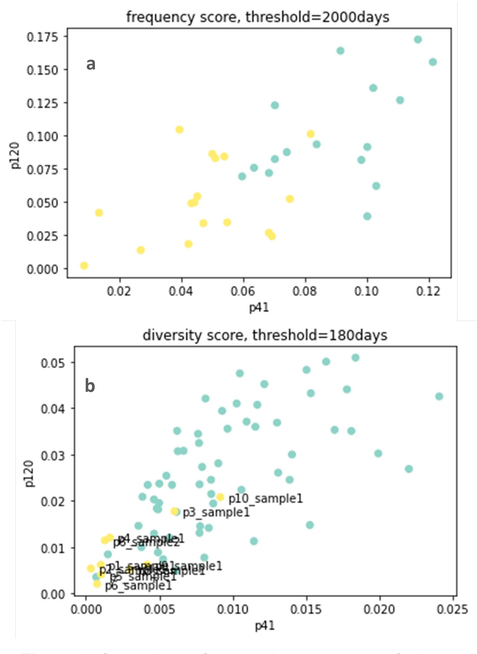

Figure 7. Separation of groups by time since infection.

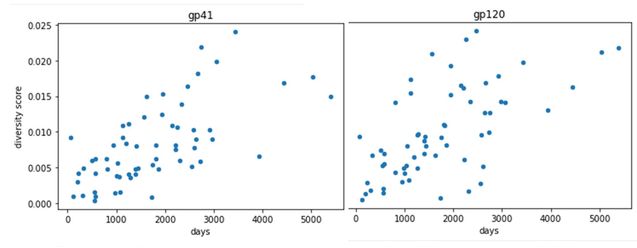

Figure 6. Comparison of diversity scores for different parts of HIV *env* gene. The diversity scores measured based on variants restricted either to gp41 (left panel) or to gp120 (right panel) regions of HIV genome are plotted as a function of time since infection. The data for all 11 patients have been used, each data point treated as a separate sample.gp41 performs better and show less scatter. Importantly, metrics from different regions of *env* can be used as separate parameters in ML classifiers.
X axis - time since infection (in days), Y axis – diversity score.

Figure 7. Separation of groups by time since infection. X axis - diversity score for p41 region, Y axis - diversity score for p120.
(**a**) Yellow - less than 2000 days. Green - more than 2000 days, (**b**) Yellow - less than 180 days. Green - more than 180 days

Conclusion

Using the sequencing pre-processing, variant calling, and ML strategies implemented in the HIVE virus diversity pipeline, prediction of the time since infection estimate detection and analysis is an attractive alternative approach to the antibody level based tests currently in clinical trials. The sources of noise might include low PCR efficiency, the presence of multiple HIV strains, experimental or biological contamination. To develop a clinical grade assay applicable to the broad category of HIV NGS datasets, our efforts will be to include additional noise estimates and internal controls into classifier models.

References

1. https://hiv.biozentrum.unibas.ch/
2. Population genomics of intrapatient HIV-1 evolution. Zanini F, et al. eLife. 2015.
3. Estimating time of HIV-1 infection from next-generation sequence diversity. Puller V, et al. PLoS Comput Biol. 2017.
4. HIVE-hexagon: high-performance, parallelized sequence alignment for next-generation sequencing data analysis. Luis Santana-Quintero et al. PloS One. 2014.
5. HIVE-Heptagon: A Sensible Variant-Calling Algorithm with Post-Alignment Quality Controls. Simonyan V et al. Genomics. 2017.