Machine learning identified genetic features associated with HIV sequences in the monocytes

Xiaorong Peng, Biao Zhu

State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, National Clinical Research Center for Infectious Diseases, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang 310003, China.

To the Editor: Human immunodeficiency virus (HIV) DNA has been detected in circulating monocytes isolated from HIV-infected individuals with or without antiretroviral therapy (ART).[1] Infrequent and low levels of HIV DNA were detected in circulating monocytes.[2] Monocytes harbor replication-competent, non-latent HIV-1 in patients on ART. Earlier studies on circulating monocytes have revealed distinct viral populations and genetic characteristics compared to T cells.[3,4] However, there is currently no easy way to clearly distinguish virus in monocyte from virus in T-cell.

Monocytes persist in the blood for only a few days before dying or migrating into tissues and differentiating into macrophages. HIV-infected monocytes could provide an important mechanism to disseminate the virus to sites, like the central nervous system (CNS) and male genital system, where evidences for compartmentalization of viral populations had been identified.[5] One possibility is that monocytes are being infected by macrophage-tropic viruses replicating in tissue macrophage.[6] Defining genetic determinants of viruses in monocytes could enhance our overall understanding of macrophage-tropic viruses in different anatomic sites and have implications for the design of curative strategies for HIV.

In most studies of macrophage-tropic viruses, the technique used was to clone the *env* sequence into pseudoviruses, and tropism was defined as entry and replication. These studies showed that the ability of viruses to gain entry into target macrophages using a low density of CD4 and coreceptor is a key feature.[7] Mutations within and adjacent to the envelope (*env*)-CD4 binding site have been identified as determining macrophage tropism, including the N283 substitution in the C2 region, which enhances the binding affinity of gp120 to CD4, and the loss of the N-linked glycosylation site N386 in V4. Another important region that has been identified as contributing to the macrophage tropism phenotype is the V3 loop, which is linked to co-receptor binding. Specifically, V3 loop substitutions S306R and I326 have been suggested to contribute to macrophage tropism.

Machine learning (ML) may provide an effective tool for the prediction of HIV provirus in the macrophage reservoir. This highly promising technique can build a potent model by exploring a vastly large set of parameters. ML tools have been applied to discover patterns in noisy biological datasets.[8] ML methods trained on HIV sequences have accurately predicted biologically relevant outcomes such as coreceptor usage, immune epitopes, and drug resistance mutations, and identified functional groupings of amino acid positions within protein classes. In this study, we have applied ML methods to attempt to distinguish viral genomes in monocyte from viral genomes in T cells based on *env* sequences.

All available subtype B *env* C2V3C3 segments isolated from subjects with paired monocytes and T cells were obtained from the Los Alamos HIV database (http://www.hiv.lanl.gov/) on May 21, 2022. Multiple alignments of codon and protein sequences were produced using Gene Cutter (https://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html). According to the previous publications,[3,4] $CD4^+$ T cells and $CD14^+$ monocytes were isolated from peripheral blood mononuclear cells (PBMC) using column purification and magnetic antibody beads. The purity of $CD14^+$ monocytes and $CD4^+$ T cells were about 99%. The accession numbers for these sequences in the Los Alamos HIV database are summarized in Supplementary Table 1, http://links.lww.com/CM9/B836.

Access this article online

Quick Response Code:	Website: www.cmj.org
	DOI: 10.1097/CM9.0000000000002932

Correspondence to: Biao Zhu, State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, National Clinical Research Center for Infectious Diseases, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang 310003, China
E-Mail: zhubiao1207@zju.edu.cn

Chinese Medical Journal 2023;136(24)
Received: 26-07-2023; **Online:** 28-11-2023 **Edited by:** Yanjie Yin

The approximate maximum likelihood phylogenetic trees for *env* segments were estimated in Molecular Evolutionary Genetics Analysis (MEGA X) using the general time reversible (GTR) + Γ + I nucleotide substitution model (www.megasoftware.net). The genetic compartmentalization between monocytes and T cells was calculated in the HyPhy (v2.24, http://hyphy.org) by using the tree-based Slatkin–Maddison test (SM test) and the distance-based Hudson test (Snn test). The criteria for genetic compartmentalization were $P \leq 0.01$.

All computational analyses were conducted in R ver. 4.0.2 (https://www.r-project.org/) as previously described in R package HANDPrediction on GitHub (https://github.com/masato-ogishi/HANDPrediction).[8]

All sequences were split into the training and testing group by 4:1. Each amino acid was converted to five quantitative measures of biophysicochemical properties: "Hydro," "Charge," "Polar," "Distribution," and "Flexi," by AAIndex metrics (http://www.genome.jp/aaindex/). Seventy-six features were used to represent the biophysicochemical property of a certain amino acid at a certain position. In the training group, columns with zero variance and highly correlating columns were deleted using the preProcess function. After the filtration, 4880 unique features were left. Support vector machine (SVM), random forest (RF), gradient boosting machine (GBM), extreme gradient boosting with linear booster (XGBL), and extreme gradient boosting with tree booster (XGBT) were applied to compare their accuracy in predicting viruses for monocyte/macrophage.[8] In addition, extreme gradient boosting with tree booster was used to "stack" of these classifiers. Ten-fold repeated three-fold cross-validation was conducted in the training phase to improve the generalizability of the classifiers.

The varImp function in the caret package was applied to estimate model-special feature importance. The 20 most important genetic feature differences between the monocyte and T-cell data subsets were identified in each algorithm. Welch's *t*-test was used to test the distribution of the feature values predicted by more than two algorithms among the monocyte and T-cell groups. The most important features were defined as the features with adjusted *P*-values of lower than 0.05.

Five hundred and four *env* C2V3C3 sequences data for paired T cell and monocyte specimens were collected from eight patients. A total of 266 partial HIV *env* sequences were collected from T cell specimens (range, 14–58 sequences for each individual), 238 from monocytes specimens (range, 18–42 sequences for each individual) [Supplementary Table 2, http://links.lww.com/CM9/B813]. Using previously developed tree-based (SM test) and distance-based (Snn test) tests of compartmentalization, we found evidence for compartmentalization of viral populations from eight individuals between T cells and monocytes [Supplementary Table 2, http://links.lww.com/CM9/B813].

Five distinct algorithms with 10 different random seeds were compared for their accuracy in correctly predicting viruses in monocytes. Stacking of the five classifiers was also attempted. The accuracy of the ML model predicting virus in monocytes in each of the 10 random-split seeds of the dataset is displayed in Supplementary Figure 1, http://links.lww.com/CM9/B813. The mean and best accuracy of the best classifier (XGBL) were 79.0% and 86.5%, respectively. The accuracy of XGBL classifier for the whole dataset was 88.3% (95% confidence interval [CI]: 0.86, 0.91), specificity is 0.94.

The model identified five C2V3C3 *env* features most significant in distinguishing between proviruses in monocytes. The identified feature positions are 297, 326, 335, 355, and 395. The distributions of the most important features were significantly different between viruses in monocytes and T cells. The AAIndex ARGP820101 of position 297, 326, and 335 represents a hydrophobicity index. The difference between the virus in monocyte and T cell was explained by the increase in the frequency of 297I, 326D, and 335N in the monocyte group. Meanwhile, the AAIndex PONP800105 of position 355 is related to the surrounding hydrophobicity in the beta-sheet. The AAIndex KUMS000101 of position 395 represents the distribution of amino acid residues in the 18 non-redundant families of thermophilic proteins [Supplementary Table 3, http://links.lww.com/CM9/B813]. Variants enriched in the monocyte group, such as 355G and 395W/T, contributed to the different feature distributions [Figure 1]. Based on these five features, the mean and best accuracy of the best classifier (XGBL) were 74.3% and 82.5%, respectively. The accuracy of this classifier for the whole dataset was 73.4% (95% CI: 0.69, 0.77). Specificity is 0.78.

Previous studies showed several important features were identified in virus in macrophages. Some CD4 binding site substitutions have been confirmed to enhance the binding affinity of gp120 to CD4. The gp120 protein is comprised of five variable (V1–V5) and five conserved constant (C1–C5) domains. V3 loop sequence variations were associated with viral phenotype of viral entry and co-receptor usage.[7] Positions 297 and 326 were located at V3 loop and may be associated with co-receptor usage. The presence of I326 in the gp120 V3 loop stem is located at the gp120-coreceptor interface and predicted to interact with the CXCR4 N terminus. I326 was found to be critical for efficient C-X-C chemokine receptor type 4 (CXCR4)-mediated monocyte-derived macrophages (MDM) entry of divergent CXCR4-using Envs. Another study found that all V3 loop sequences from HIV-1 long-term non-progressors contained a proline at position 326.

Positions 335 and 355 are located in the C3 region. The detailed role of these positions played in the interaction needs further investigation. Previously, 335R was one of the seven positively selected sites at the surface positions of the alpha-helix (positions 335–347 in the C3 region) in the opposite face for CD4 binding. Position 355, to the best of our knowledge, has not previously been reported within the context of cell tropism.

HIV can infect cells in different phases of the process from monocytes to macrophages.[5] Monocyte progenitors may be infected with HIV in the bone marrow. Intermediate

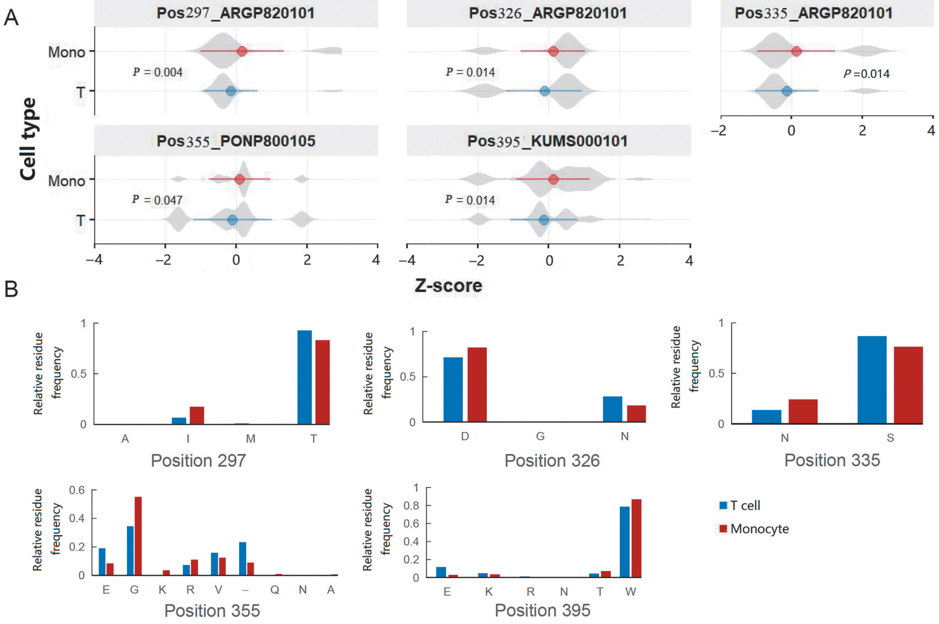

Figure 1: The most important different features between the virus in monocyte and T cell. (A) Distributions of detected features among monocyte and T cell groups. The values of each feature were converted to *Z*-score for visualization purposes. (B) Relative residue frequencies in sequence sets derived from virus in monocyte and T cell. The alignment position numbers correspond to the positions in the HIV-1 reference sequence (HXB2; NCBI accession: K03455). The X-axis represents different amino acid residue. A: Alanine; D: Aspartate; N:Asparagine; E: Glutamate; G: Glycine; I: Isoleucine; K: Lysine; M: Methionine; N: Asparaginate; Q: Glutamine; R: Arginine; S: Serine; T: Threonine; V: Valine; W: Tryptophan; –: None. HIV-1: Human immunodeficiency virus-1.

monocytes may be infected with HIV in the tissues such as the spleen. Circulating monocytes may also be infected with HIV in the blood.[5] Circulating HIV-infected monocytes can enter different anatomical sanctuary sites such as the brain, differentiate into macrophages and thus seed as tissue reservoirs. Macrophages may also be infected with HIV in these tissues by cell-to-cell transmission. Macrophages infected by HIV could promote local inflammation and be the origin of viral rebound. The difference of viral population in monocytes and T cells may explain viral compartmentalization in several anatomical sites. Due to the difficulty in obtaining samples, our overall understanding of macrophage-tropic viruses in different anatomic sites is limited. Defining genetic determinants of virus in monocytes could enhance the understanding about the size and characteristics of the myeloid (monocyte/macrophage) reservoir.

In summary, the five important features and several variants were identified in the virus in the monocyte group. This study could help us understand the phenotype of the virus in macrophages and the reservoir component, guiding a new study direction to eradicate the HIV reservoir.

Conflicts of interest

None.

Funding

This study was supported by grants from the National Natural Science Foundation of China (No. 82202522) and the Independent Task of State Key Laboratory for Diagnosis and Treatment of Infectious Diseases (No. 2019ZZ18).

References

1. Veenhuis RT, Abreu CM, Costa PAG, Ferreira EA, Ratliff J, Pohlenz L, *et al.* Monocyte-derived macrophages contain persistent latent HIV reservoirs. Nat Microbiol 2023;8:833–844. doi: 10.1038/s41564-023-01349-3.
2. Massanella M, Bakeman W, Sithinamsuwan P, Fletcher JLK, Chomchey N, Tipsuk S, *et al.* Infrequent HIV infection of circulating monocytes during antiretroviral therapy. J Virol 2019;94:e1174–e1119. doi: 10.1128/JVI.01174-19.
3. Llewellyn N, Zioni R, Zhu H, Andrus T, Xu Y, Corey L, *et al.* Continued evolution of HIV-1 circulating in blood monocytes with antiretroviral therapy: Genetic analysis of HIV-1 in monocytes and CD4+ T cells of patients with discontinued therapy. J Leukoc Biol 2006;80:1118–1126. doi: 10.1189/jlb.0306144.
4. Fulcher JA, Hwangbo Y, Zioni R, Nickle D, Lin X, Heath L, *et al.* Compartmentalization of human immunodeficiency virus type 1 between blood monocytes and CD4+ T cells during infection. J Virol 2004;78:7883–7893. doi: 10.1128/JVI.78.15.7883-7893.2004.
5. Chitrakar A, Sanz M, Maggirwar SB, Soriano-Sarabia N. HIV latency in myeloid cells: Challenges for a cure. Pathogens 2022;11:611. doi: 10.3390/pathogens11060611.
6. Zhao JC, Deng K. Heterogeneity of HIV-1 latent reservoirs. Chin Med J 2020;133:2867–2873. doi: 10.1097/cm9.0000000000001085.
7. Duncan CJ, Sattentau QJ. Viral determinants of HIV-1 macrophage tropism. Viruses 2011;3:2255–2279. doi: 10.3390/v3112255.
8. Ogishi M, Yotsuyanagi H. Prediction of HIV-associated neurocognitive disorder (HAND) from three genetic features of envelope gp120 glycoprotein. Retrovirology 2018;15:12. doi: 10.1186/s12977-018-0401-x.

How to cite this article: Peng XR, Zhu B. Machine learning identified genetic features associated with HIV sequences in the monocytes. Chin Med J 2023;136:3002–3004. doi: 10.1097/CM9.0000000000002932