HIV-1 tropism prediction by the XGboost and HMM methods

Xiang Chen¹, Zhi-Xin Wang¹ & Xian-Ming Pan²

Human Immunodeficiency Virus 1 (HIV-1) co-receptor usage, called tropism, is associated with disease progression towards AIDS. Furthermore, the recently developed and developing drugs against co-receptors CCR5 or CXCR4 open a new thought for HIV-1 therapy. Thus, knowledge about tropism is critical for illness diagnosis and regimen prescription. To improve tropism prediction accuracy, we developed two novel methods, the extreme gradient boosting based XGBpred and the hidden Markov model based HMMPred. Both XGBpred and HMMPred achieved higher specificities (72.56% and 72.09%) than the state-of-the-art methods Geno2pheno (61.6%) and G2p_str (68.60%) in a 10-fold cross validation test at the same sensitivity of 93.73%. Moreover, XGBpred had more outstanding performances (with AUCs 0.9483, 0.9464) than HMMPred (0.8829, 0.8774) on the Hivcopred and Newdb (created in this work) datasets containing larger proportions of hard-to-predict dual tropic samples in the X4-using tropic samples. Therefore, we recommend the use of our novel method XGBpred to predict tropism. The two methods and datasets are available via http://spg.med.tsinghua.edu.cn:23334/XGBpred/. In addition, our models identified that positions 5, 11, 13, 18, 22, 24, and 25 were correlated with HIV-1 tropism.
while CM constructs gapped PSSMs and takes the 11/25 rule and net charge into consideration. Recently, many genotypic methods based on machine learning have also been published. The method Geno2pheno combines two machine learning approaches, support vector machine (SVM) and decision trees, and uses clinical information such as viral loads and CD4-cell counts if available. Another method from the same laboratory, G2p_str, combines SVM and Lasso regression and uses the amino acid structure feature. Hivcopred is based on SVMlight with the split amino acid composition feature. T-CUP employs random forests (RFs) with the structural information of hydrophobicity and electrostatic potential. Currently, Geno2pheno is the most widely used method and the only genotypic method recommended for usage in clinical routines by the European Consensus Group.

Genotypic methods can predict R5 viruses (~90%) accurately, but are inaccurate in the prediction of X4-using viruses (~50–70%). Thus, more accurate tropism prediction methods are required. Here, we present two methods, XGBpred and HMMpred. We analyzed the HIV-1 tropism prediction ability of our methods and compared them with the Geno2pheno, G2p_str, Hivcopred, CM and WebPSSM methods. The results show that XGBpred is robust with the hard-to-predict dual tropic sequences.

### Methods

#### Datasets.

To construct the Newdb dataset, we extracted 6790 R5 tropic, 590 X4 tropic and 1125 dual tropic sequences from the Los Alamos HIV sequence database (http://www.hiv.lanl.gov/, last update: 10 Sep 2017). The tropisms of the sequences from the Los Alamos HIV sequence database have been phenotypically determined, none of them have been inferred from sequences. Then we removed sequences containing non-canonical residues, reserved sequences with lengths between 31 and 39, and dislodged duplicated sequences to guarantee the high quality of genotype/phenotype pairs. This process finally generated 2335 R5 and 663 X4-using (245 X4 and 418 dual) tropic sequences. The distribution of the six major subtypes in the Newdb dataset is shown in Table 1. To compare our methods with the Geno2pheno, G2p_str, Hivcopred, CM and WebPSSM methods, we used the datasets constructed in these studies, respectively. These datasets can be accessed in Supplementary Spreadsheet S1. The distributions of tropisms in different datasets are shown in Table 2.

### Tables

**Table 1.** Distribution of the six major subtypes in the Newdb dataset. Notes: *The number of R5 tropic sequences. *The number of X4 tropic sequences. *The number of dual tropic sequences.

| Subtype | Number (R5, X4, D) | Percentage |
|---------|--------------------|------------|
| B       | 1503 (1209, 93, 201) | 50.13%     |
| C       | 511 (460, 26, 25)   | 17.04%     |
| D       | 233 (120, 52, 61)   | 7.77%      |
| 01_AE   | 213 (149, 45, 19)   | 7.10%      |
| A       | 155 (140, 5, 10)    | 5.17%      |
| 02_AG   | 124 (50, 3, 71)     | 4.14%      |

**Table 2.** Distribution of tropisms in the different datasets. Notes: *Removed 31 duplicated sequences from the original Hivcopred dataset which are marked as not only R5 tropism but also X4-using tropism. *Training set. *Validation set.

| Dataset   | R5 | X4-using | Sum |
|-----------|----|----------|-----|
|           |    |          |     |
|           |    |          |     |
|           |    |          |     |
|           |    |          |     |
|           |    |          |     |

### Machine learning method: XGBpred.

Extreme gradient boosting (XGBoost), like RFs used by T-CUP, is an ensemble algorithm of decision trees. The ensemble works by combining a set of weaker machine learning algorithms to get an improved machine learning algorithm in overall. The main difference between XGBoost and RFs is the way of sampling. RFs are based on uniform sampling with return. Instead, XGBoost gives higher weights to the wrongly predicted samples in the current weaker learner, and then these samples will be paid more attention when training the next weaker learner. In addition, XGBoost adds regularization to avoid overfitting. Therefore, XGBoost is a more complicated algorithm than RFs, and thus always outperforms.

Because XGBoost is designed for vectors, it is necessary to convert V3 loop string sequences of different lengths to numerical vectors. For this task, we used many kinds of features to describe the characteristics of protein sequences, such as split amino acid composition, dipeptide composition, and net charge or hydrophobicity. We also proposed an additional set of features: the alignment score. The 35-dimensional alignment scores were generated by scoring alignments using the block substitution matrices BLOSUM62, BLOSUM90 or BLOSUM100, and the alignments were generated by aligning sequences to the consensus sequence with...
35 residues by the means of Needleman-Wunsch (Version: EMBOSS: 6.6.0)\textsuperscript{31}. For the XGBpred method, we tested these different features and their combinations to find the optimal model to discriminate R5 and X4-using sequences.

**Statistics method: HMMpred.** Hidden Markov model (HMM) is a finite model applied in time series and linear sequences. Just as the PSSM profile, HMM also can be used to describe protein families. The HMM profile described by state-transition and symbol-emission probabilities performs better than PSSM in terms of sequence alignment and homology recognition because it can deal with gaps in protein families better by hidden state chains\textsuperscript{32}.

**HMM profile construction.** We used the maximum likelihood estimation method to establish R5 and X4-using specific HMM profiles from R5 and X4-using tropic multiple sequence alignments generated by Clustalo\textsuperscript{33}, respectively. In addition, we simply assigned columns that had more than half gap characters as insertion states. The structure of HMM that we used was no transition allowed from $D_{i}$ to $I_{i}$ or from $I_{i}$ to $D_{i+1}$ (This kind of structure performed better than the full structure, as shown in Supplementary Table S1). M, D, and I denote match, deletion and insertion states, respectively.

\[
\hat{a}_{kl} = \frac{A_{kl} + 1}{\sum_{l'}A_{l'l}} + \frac{1}{3}
\]  

\[
\hat{e}_{k}(a) = \frac{E_{k}(a) + 1}{\sum_{a'}E_{k}(a')} + \frac{1}{21}
\]

Where $i$, $k$, and $l$ are indices over states M, D, or I; $a$ is an amino acid symbol or gap; $\hat{a}_{kl}$ means the estimated probability of transiting from state $k$ to state $l$; $\hat{e}_{k}(a)$ means the estimated probability of emitting residue $a$ at state $k$, and $A_{kl}$ and $E_{k}(a)$ are the corresponding frequencies. In order to avoid the zero probability which represents it cannot happen in the future, we applied the Laplace's pseudo-count rule that added one to each frequency.

**Sequence-profile alignment.** We employed Viterbi algorithm\textsuperscript{34}, a dynamic programing algorithm, to get two alignment scores $S_{R5}$ and $S_{non-R5}$. Those alignment scores represent the optimal state pathway scores from the R5 and X4-using HMM profiles, respectively. The final score was defined as:

\[
S = S_{R5} - S_{non-R5}
\]

Then the given sequence would be classified as R5 tropic if the final score $S$ is higher than a threshold, otherwise it would be classified as X4-using tropic.

**Ten-fold cross validation.** The widely-used 10-fold cross validation was used to evaluate the performance of our methods in this study, where the sequences were divided into 10 subsets randomly, one subset was used as the testing set, and the others were used as the training set. After ten repetitions, the final performance was average of the performances of those ten subsets.

**Evaluation parameters.** For evaluation, we used sensitivity, specificity, accuracy and Matthew's correlation coefficient (MCC). In particular, MCC is robust even when the size of classes varies widely\textsuperscript{35}. An MCC value ‘0’ corresponds to a completely random prediction, while ‘1’ corresponds to a perfect perdition. These parameters were calculated using the following equations:

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]  

\[
\text{Specificity} = \frac{TN}{FP + TN}
\]  

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}
\]  

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

where $TP$ is the number of true positives, $FP$ false positives, $TN$ true negatives and $FN$ false negatives. We regarded R5 tropic samples as positives in this study.

In contrast to the four threshold-dependent parameters, the receiver operating characteristic (ROC) curve, a threshold-independent parameter, illustrates the trade-off between sensitivity and specificity at various threshold settings. In this study, we used the area under the curve (AUC) to measure a predictive power, where 0.5 means a random method, and 1 means a perfect method\textsuperscript{36}.
Results

Performance on the Newdb dataset. The feature set and the model that gave the strongest predictive power for the XGBpred and HMMpred methods were found, respectively (Supplementary Tables S1 and S2). The performances of the two methods on the Newdb dataset in a same 10-fold cross validation test are shown in Fig. 1A and Table 3. XGBpred had a higher specificity, accuracy, MCC and AUC than HMMpred when having the same sensitivity. Furthermore, the specificity of XGBpred was higher than 80% (84.62%) at the sensitivity of 91.78%. Results from the two methods were highly consistent: they predicted same tropisms for 87.96% of total samples, and achieved 96.70% sensitivity, 83.39% specificity and 93.93% accuracy.

Considering the poorer performance of HMMpred, the score distributions of the two methods were plotted (Fig. 1B). As depicted, the scores of dual tropic sequences mostly placed in the middle of the scores of X4 and R5 tropic sequences. Furthermore, HMMpred generated higher scores for a considerable number of dual tropic samples than XGBpred. This phenomenon illustrates that it is hard for dual tropic sequences to be correctly classified, especially by HMMpred.

The performances of the two methods for the six major subtypes (subtypes B, C, D, 01_AE, A and 02_AG) in the Newdb dataset were analyzed due to the sequence divergence among different subtypes and the different number of sequences in each subtype (Fig. 1C). HMMpred for subtypes B and D showed much lower AUCs (0.8942 and 0.8839) than for subtypes C and 01_AE (0.9369 and 0.9486). The reason was that subtypes B and D contained more hard-to-predict dual tropic sequences (Table 1). This also resulted in a low AUC (0.5887) for subtype 02_AG, and a higher AUC (0.9029) for subtype A than for subtype D (0.8839) generated by HMMpred. In contrast, the performance of XGBpred was not so deeply influenced by dual tropic sequences. XGBpred had higher AUCs for the top four most common subtypes (subtypes B, C, D and 01_AE) than for subtypes A and

Figure 1. Performance of the XGBpred and HMMpred methods on the Newdb dataset. (A) ROC curves on the Newdb dataset in a same 10-fold cross validation test. The legend lists AUCs and specificities at the sensitivity of 91.78% which is plotted as the dashed black line. (B) Distribution of V3 loop sequence scores calculated from XGBpred and HMMpred on the Newdb dataset. The score distribution of the R5 tropic sequences is shown in blue, that of X4 is carmine and that of dual is yellow. (C) ROC curves of XGBpred and HMMpred for the six major subtypes. The legend lists AUCs and mAPs.
with the WebPSSM method, although the WebPSSM dataset is small, XGBpred had a higher AUC (0.9043) than the positions that we identified as correlated with HIV-1 tropism are exactly in accordance with the results from a previous study.

In this section, to evaluate our methods, we compared with the previously published methods Geno2pheno, G2p_str, Hivcopred, CM, and WebPSSM by implementing our methods in a 10-fold cross validation test on the datasets used in these published methods, respectively. The exception was WebPSSM where we used the training set from WebPSSM to model our methods in a 10-fold cross validation test and used the validation set from WebPSSM to test (Table 3).

| Dataset | Method | Specificity | Accuracy | MCC | AUC  |
|---------|--------|-------------|----------|-----|------|
| Newdb   | XGBpred | 84.62%      | 90.19%   | 0.7310 | 0.9465 |
|         | HMMpred | 70.59%      | 87.09%   | 0.6247 | 0.8774 |
| G2p_str | Geno2pheno | 61.6%       | —        | —    | 0.860 |
|         | G2p_str  | 68.6%       | —        | —    | 0.892 |
|         | XGBpred  | 72.56%      | 89.90%   | 0.6605 | 0.8952 |
|         | HMMpred  | 72.09%      | 89.81%   | 0.6570 | 0.9002 |
| Hivcopred | Hivcopred | 81.44%      | 87.07%   | 0.67  | 0.904 |
|         | XGBpred  | 87.13%      | 88.52%   | 0.7154 | 0.9483 |
|         | HMMpred  | 71.08%      | 84.63%   | 0.5899 | 0.8829 |
| CM      | CM      | 92.92%      | 95.21%   | 0.885 | 0.97  |
|         | XGBpred  | 93.85%      | 95.33%   | 0.8106 | 0.9809 |
|         | HMMpred  | 89.54%      | 94.81%   | 0.7826 | 0.9635 |
| WebPSSM | WebPSSM | 83.3%       | —        | —    | 0.881 |
|         | XGBpred  | 83.33%      | 83.10%   | 0.6419 | 0.9043 |
|         | HMMpred  | 75.00%      | 80.28%   | 0.5693 | 0.8678 |

Table 3. Performance of the XGBpred and HMMpred methods on the different datasets. Performance of XGBpred and HMMpred on the Newdb, G2p_str, Hivcopred, CM and WebPSSM datasets at the sensitivities of 91.78%, 93.73%, 89.99%, 95.54% and 82.98%, respectively.

In this section, to evaluate our methods, we compared with the previously published methods Geno2pheno, G2p_str, Hivcopred, CM, and WebPSSM by implementing our methods in a 10-fold cross validation test on the datasets used in these published methods, respectively. The exception was WebPSSM where we used the training set from WebPSSM to model our methods in a 10-fold cross validation test and used the validation set from WebPSSM to test (Table 3).

First when comparing with the Geno2pheno and G2p_str methods, XGBpred and HMMpred achieved AUCs of 0.8952 and 0.9002, respectively. Our methods had higher AUCs than Geno2pheno (0.860) and G2p_str (0.892). In addition, XGBpred and HMMpred achieved specificities of 72.56% and 72.09% at the sensitivity of 93.73%. The specificities were obviously higher than the specificities of Geno2pheno (61.6%) and G2p_str (68.6%) at the same sensitivity. Second, when comparing with the Hivcopred method, XGBpred had a higher AUC (0.9483) than Hivcopred (0.904), but HMMpred had a low AUC (0.8829) as on the Newdb dataset. Third, when comparing with the CM method, our methods were as accurate as the CM method on the CM dataset which only contains a small amount of hard-to-predict dual tropic samples (Table 2). Finally, when comparing with the WebPSSM method, although the WebPSSM dataset is small, XGBpred had a higher AUC (0.9043) than WebPSSM (0.881), and HMMpred presented a similar AUC (0.8678) with WebPSSM.

In this section, to evaluate our methods, we compared with the previously published methods Geno2pheno, G2p_str, Hivcopred, CM, and WebPSSM by implementing our methods in a 10-fold cross validation test on the datasets used in these published methods, respectively. The exception was WebPSSM where we used the training set from WebPSSM to model our methods in a 10-fold cross validation test and used the validation set from WebPSSM to test (Table 3).

First when comparing with the Geno2pheno and G2p_str methods, XGBpred and HMMpred achieved AUCs of 0.8952 and 0.9002, respectively. Our methods had higher AUCs than Geno2pheno (0.860) and G2p_str (0.892). In addition, XGBpred and HMMpred achieved specificities of 72.56% and 72.09% at the sensitivity of 93.73%. The specificities were obviously higher than the specificities of Geno2pheno (61.6%) and G2p_str (68.6%) at the same sensitivity. Second, when comparing with the Hivcopred method, XGBpred had a higher AUC (0.9483) than Hivcopred (0.904), but HMMpred had a low AUC (0.8829) as on the Newdb dataset. Third, when comparing with the CM method, our methods were as accurate as the CM method on the CM dataset which only contains a small amount of hard-to-predict dual tropic samples (Table 2). Finally, when comparing with the WebPSSM method, although the WebPSSM dataset is small, XGBpred had a higher AUC (0.9043) than WebPSSM (0.881), and HMMpred presented a similar AUC (0.8678) with WebPSSM.

In this section, to evaluate our methods, we compared with the previously published methods Geno2pheno, G2p_str, Hivcopred, CM, and WebPSSM by implementing our methods in a 10-fold cross validation test on the datasets used in these published methods, respectively. The exception was WebPSSM where we used the training set from WebPSSM to model our methods in a 10-fold cross validation test and used the validation set from WebPSSM to test (Table 3).

First when comparing with the Geno2pheno and G2p_str methods, XGBpred and HMMpred achieved AUCs of 0.8952 and 0.9002, respectively. Our methods had higher AUCs than Geno2pheno (0.860) and G2p_str (0.892). In addition, XGBpred and HMMpred achieved specificities of 72.56% and 72.09% at the sensitivity of 93.73%. The specificities were obviously higher than the specificities of Geno2pheno (61.6%) and G2p_str (68.6%) at the same sensitivity. Second, when comparing with the Hivcopred method, XGBpred had a higher AUC (0.9483) than Hivcopred (0.904), but HMMpred had a low AUC (0.8829) as on the Newdb dataset. Third, when comparing with the CM method, our methods were as accurate as the CM method on the CM dataset which only contains a small amount of hard-to-predict dual tropic samples (Table 2). Finally, when comparing with the WebPSSM method, although the WebPSSM dataset is small, XGBpred had a higher AUC (0.9043) than WebPSSM (0.881), and HMMpred presented a similar AUC (0.8678) with WebPSSM.

Feature importance analysis. Given the high performance of XGBpred presented in the previous subsections, we discussed which features XGBpred provided with its predictive power (Fig. 2). We did not analyze the feature importance on the WebPSSM dataset as it contains few training samples (Table 2). In the XGBpred method, the feature alignment score in the 5th position of the V3 loop appeared in the top three most important features on all datasets. Interestingly, amino acid Tyr in position 5 appeared more frequently in X4-using tropic than in R5 tropic sequences (Supplementary Fig. S1). Currently, X4-using tropism can be predicted by the 11/25 rule19. However, since position 5 was as important as positions 11 and 25, the pragmatic 11/25/5 rule was proposed to predict a virus as X4-using tropic by the presence of a positively charged amino acid in positions 11 or 25, or by the presence of amino acid Tyr in position 5 of its V3 loop. Compared with the 11/25 rule, the 11/25/5 rule reduced specificities by 1.29%, 1.03%, 1.14% and 1.19% on the Newdb, G2p_str, Hivcopred and CM datasets, respectively. The 11/25/5 rule also had higher accuracies and MCCs than the 11/25 rule at the sensitivities of 91.78%, 93.73%, 89.99%, 95.54% and 82.98%, respectively.
co-receptor usage or not. By XGBpred is a feasible method to judge whether a new discovered association pattern is of importance to tropism. Furthermore, the feature importance distribution generated by XGBpred is based on machine learning, and HMMpred is based on statistics. XGBpred performed best on the datasets containing larger proportions of hard-to-predict dual tropic samples in the X4-using samples, while HMMpred performed worst. In contrast, the predictive powers of the two methods were similar on the smaller G2p_str and CM datasets containing fewer dual tropic samples (Table 3). The poor ability of HMMpred to predict tropism stemmed from the high probability that HMMpred incorrectly predicted dual tropic samples as R5 tropic (Fig. 1B and Supplementary Fig. S2). The profiles used in HMMpred may not be meticulous enough. Several reasons may account for this phenomenon. Firstly, the two sequence families are highly similar since even one amino acid substitution may change their tropisms42,43. Secondly, the characteristics of dual tropic sequences may be overwhelmed by R5 and X4 tropic sequences. Finally, the unavailability of X4-using tropic samples makes it uncertain to learn its accurate HMM profile. Moreover, as the number of samples increased, the gap of predictive powers between XGBpred and HMMpred became large (Tables 2 and 3). This corresponds to the fact that the machine learning based Geno2pheno method is more widely used than the statistics based 11/25 rule and WebPSSM. As a result, a machine learning based method, in particular XGBpred, is recommended to predict co-receptor usage as the number of samples continues to expand.

In an effort to further increase the predictive power, we also generated three meta methods by the means of stacking44. The scores generated by XGBpred, Hivcopred (SVM light) and HMMpred were added as additional features to the new stacking based XGBpred models. Compared with the original XGBpred method, the new stacking-based XGBpred methods had slightly higher AUCs on the G2p_str dataset but lower AUCs on the other datasets (Supplementary Table S4). The poor performances of the meta methods may due to the poorer predictive abilities of Hivcopred and HMMpred than the original XGBpred method, and/or the dependence of the results generated by XGBpred, Hivcopred and HMMpred (Supplementary Table S5). This may stem from the fact that the V3 loop is not the sole determinant of viral tropism. Moreover, V1, V2, C4 and the bridge sheet regions of gp120 also have an impact on co-receptor usage45,46. To predict tropism, several methods gain a higher accuracy by employing other information in addition to the V3 loop, such as clinical information47, V2 loop sequences48 and structure information23,25,41. Therefore, the stacking based method can be constructed to improve its predictive power by combining methods with different kinds of information.

In summary, the two methods we developed performed comparably on the datasets containing less hard-to-predict dual tropic sequences, but XGBpred performed much better on the datasets with more dual tropic sequences. This means XGBpred is more robust to predict dual tropic sequences than other methods. Thus, we strongly recommend to use XGBpred to predict viral tropism. Our two methods have been implemented as a freely available webserver under http://spg.med.tsinghua.edu.cn:23334/XGBpred/.

References
1. Hladik, F. et al. Initial events in establishing vaginal entry and infection by human immunodeficiency virus type-1. Immunity 26, 257–270, https://doi.org/10.1016/j.immuni.2007.01.007 (2007).
2. Wilen, C. B., Tilton, J. C. & Doms, R. W. HIV: cell binding and entry. Cold Spring Harb Perspect Med 2, https://doi.org/10.1101/cshperspect.a006866 (2012).
3. Koning, F., van Ruij, R. & Schuitemaker, H. Biological and Molecular Aspects of HIV1 Coreceptor Usage. (2019).
4. Lengauer, T., Sander, O., Sierra, S., Thielen, A. & Kaiser, R. Bioinformatics prediction of HIV coreceptor usage. Nature biotechnology 25, 1407–1410, https://doi.org/10.1038/nbt1371 (2007).
5. Vandekerckhove, L. P. R. et al. European guidelines on the clinical management of HIV-1 tropism testing. The Lancet Infectious Diseases 11, 394–407, https://doi.org/10.1016/S1473-3099(11)70319-4 (2011).
6. Berger, E. A. et al. A new classification for HIV-1. Nature 491, 240, https://doi.org/10.1038/nature134571 (1998).
7. Hoffmann, C. The epidemiology of HIV coreceptor tropism. European journal of medical research 12, 385–390 (2007).
8. Panos, G. & Watson, D. HIV and AIDS CCR-5 Virus. (2007).
9. Su, Z. et al. Response to vicriviroc in treatment-experienced subjects, as determined by an enhanced sensitivity coreceptor tropism assay: reanalysis of AIDS clinical trials group A5211. The Journal of infectious diseases 200, 1724–1728, https://doi.org/10.1093/infdis/jih494 (2009).
10. Obermeier, M., Symons, J. & Wensing, A. M. Infectious disease reports 5, e6, https://doi.org/10.4081/ird.2013.s1.e6 (2013).
11. Koning, F., van Rij, R. & Schuitemaker, H. Improved Coreceptor Usage Prediction and Genotypic Monitoring of R5-to-X4 Transition by Motif Analysis of Human Immunodeficiency Virus Type 1 env V3 Loop Sequences. Journal of virology 77, 13376–13388, https://doi.org/10.1128/jvi.00387-14 (2013).
12. Agwu, A. L. & Heider, D. A simple structure-based model for the prediction of HIV-1 co-receptor tropism. Bioinformatics (Oxford, England) 27, 22–31, https://doi.org/10.1093/bioinformatics/btr132 (2011).
13. Nelson, M. & Panos, G. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. Journal of virology 66, 3183–3187 (1992).
14. Heider, D., Dybowsky, J. N., Wilms, C. & Hoffmann, D. A simple structure-based model for the prediction of HIV-1 coreceptor tropism. A review of the National Academy of Sciences of the United States of America 89, 10915–10919 (1992).
15. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology 48, 443–453 (1970).
16. Eddy, S. R. Profile hidden Markov models. Bioinformatics (Oxford, England) 14, 755–763 (1998).
17. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology 7, 539, https://doi.org/10.1038/msb.2011.75 (2011).
18. N respiratory virus infection. The Pediatric Infectious Disease Journal 35, 777–781, https://doi.org/10.1097/inf.0000000000001158 (2016).
19. Boeke, J. D. & Janssen, W. P. Population genotypic tropism testing and its clinical significance. Emergence of CXCR4-using human immunodeficiency virus type 1 (HIV-1) variants in a minority of HIV-1-infected patients following treatment with the CCR5 antagonist maraviroc is from a pretreatment CXCR4-using virus reservoir. Journal of virology 80, 4909–4920, https://doi.org/10.1128/jvi.01825-09 (2006).
20. van Rij, R. & Schuitemaker, H. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. Journal of virology 66, 3183–3187 (1992).
21. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology 48, 443–453 (1970).
22. Eddy, S. R. Profile hidden Markov models. Bioinformatics (Oxford, England) 14, 755–763 (1998).
23. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology 7, 539, https://doi.org/10.1038/msb.2011.75 (2011).
24. David Forney, G. Jr. The Viterbi Algorithm: A Personal History. (2005).
25. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phase lysozyme. Biochimica et biophysica acta 405, 442–451 (1975).
26. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36, https://doi.org/10.1148/radiology.143.1.76063747 (1982).
27. Rao, C. T. & Nelson, M. & Panos, G. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. Journal of virology 66, 3183–3187 (1992).
28. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology 48, 443–453 (1970).
29. Eddy, S. R. Profile hidden Markov models. Bioinformatics (Oxford, England) 14, 755–763 (1998).
30. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology 7, 539, https://doi.org/10.1038/msb.2011.75 (2011).
31. David Forney, G. Jr. The Viterbi Algorithm: A Personal History. (2005).
32. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phase lysozyme. Biochimica et biophysica acta 405, 442–451 (1975).
33. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36, https://doi.org/10.1148/radiology.143.1.76063747 (1982).
34. Gao, F. et al. The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. Journal of virology 70, 7013–7029 (1996).
35. Carr, J. K. et al. Full genome sequences of human immunodeficiency virus type 1 subtypes G and A/G intersubtype recombinants. Virology 247, 22–31, https://doi.org/10.1006/viro.1998.9211 (1998).
36. Riemenschneider, M. et al. Genotypic Prediction of Co-receptor Tropism of HIV-1 Subtypes A and C. Sci Rep 6, 24883, https://doi.org/10.1038/srep24883 (2016).
37. Lochel, H. F., Riemenschneider, M., Frishman, D. & Heider, D. SCOTCH: subtype A coreceptor tropism classification in HIV-1. Bioinformatics (Oxford, England) 34, 2575–2580, https://doi.org/10.1093/bioinformatics/bty170 (2018).
38. Sander, O. et al. Structural describers of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. PLoS Comput Biol 3, e58, https://doi.org/10.1371/journal.pcbi.0030058 (2007).
39. Shimizu, N. et al. Changes in and discrepancies between cell tropics and coreceptor uses of human immunodeficiency virus type 1 induced by salt point mutations at the V3 tip of the env protein. Virology 259, 324–333, https://doi.org/10.1006/viro.1999.9764 (1999).
40. De Jong, J. J. De Ronde, A., Keulen, W., Termette, M. & Goudsmitt, J. Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution. Journal of virology 66, 6777–6780 (1992).
44. Wolpert, D. H. Stacked generalization. *Neural Networks* 5, 241–259, https://doi.org/10.1016/S0893-6080(05)80023-1 (1992).
45. Monno, L. *et al.* Impact of mutations outside the V3 region on coreceptor tropism phenotypically assessed in patients infected with HIV-1 subtype B. *Antimicrobial agents and chemotherapy* 55, 5078–5084, https://doi.org/10.1128/aac.00743-11 (2011).
46. Dimonte, S. *et al.* Selected amino acid mutations in HIV-1 B subtype gp41 are associated with specific gp120V3 signatures in the regulation of co-receptor usage. *Retrovirology* 8, 33, https://doi.org/10.1186/1742-4690-8-33 (2011).
47. Brumme, Z. *et al.* Molecular and clinical epidemiology of CXCR4-using HIV-1 in a large population of antiretroviral-naive individuals. *The Journal of infectious diseases* 192, 466–474, https://doi.org/10.1086/431519 (2005).
48. Thileen, A. *et al.* Improved prediction of HIV-1 coreceptor usage with sequence information from the second hypervariable loop of gp120. *The Journal of infectious diseases* 202, 1435–1443, https://doi.org/10.1086/656600 (2010).

Acknowledgements
This work was partly supported by Grant 31770841 from the National Natural Science Foundation of China, and 2016YFA0502114 from the Ministry of Science and Technology of China. The funding source had no role in the design or conduct of the study, or in the collection, analysis, or interpretation of the data.

Author Contributions
X.C. designed the study, developed the method, implemented the data analysis and wrote the manuscript. Z.-X.W and X.-M.P. participated in study design and revision of the manuscript. All the authors read and approved the final manuscript.

Additional Information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-46420-4.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019