Infection, Genetics and Evolution 77 (2020) 104080

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research Paper

A novel alignment-free method for HIV-1 subtype classification

Lily He[a,1], Rui Dong[a,1], Rong Lucy He[b], Stephen S.-T. Yau[a,*]

[a] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, PR China
[b] Department of Biological Sciences, Chicago State University, Chicago, United States of America

ARTICLE INFO

Keywords:
SNV
Alignment-free
HIV-1
Classification

ABSTRACT

HIV-1 is the most common and pathogenic strain of human immunodeficiency virus consisting of many subtypes. To study the difference among HIV-1 subtypes in infection, diagnosis and drug design, it is important to identify HIV-1 subtypes from clinical HIV-1 samples. In this work, we propose an effective numeric representation called Subsequence Natural Vector (SNV) to encode HIV-1 sequences. Using the representation, we introduce an improved linear discriminant analysis method to classify HIV-1 viruses correctly. SNV is based on distribution of nucleotides in HIV-1 viral sequences. It not only computes the number of nucleotides, but also describes the position and variance of nucleotides in viruses. To validate our alignment-free method, 6902 complete genomes and 11,668 pol gene sequences of HIV-1 subtypes were collected from the up-to-date Los Alamos HIV database. SNV outperforms the three popular methods, Kameris, Comet and REGA, with almost 100% Sensitivity and Specificity, also with much less time. Our subtyping algorithm especially works better for circulating recombinant forms (CRFs) consisting of a few sequences. Our approach is also powerful to separate unique recombinant forms (URFs) from other subtypes with 100% Sensitivity and Specificity. Moreover, phylogenetic trees based on SNV representation are constructed using full-length HIV-1 genomes and pol genes respectively, where viruses from the same subtype are clustered together correctly.

1. Introduction

Human Immunodeficiency Viruses (HIV) infected about 36.9 million people as of 2017. HIV-1 is more infective than HIV-2 and causes the great majority of HIV infections worldwide. Without treatment, average survival time after infection with HIV-1 varies from 9 to 11 years according to the HIV-1 subtype, host genetics, route of infection, and psychological state, ect. The differences among HIV-1 subtypes also have important effect on disease progression, therapeutic reaction and vaccine research and development. Based on previous research, HIV-1 has been divided into a main group M and three small groups: N, O and P (Nuno et al., 2014). The M group is further divided into nine pure subtypes (A-D, F−H, J and K subtypes) and 97 circulating recombinant forms (CRFs) and several unique recombinant forms (URFs) (Los Alamos National Lab (LANL) database, accessed February 2019). Inter-subtype recombinant genomes are common, but many of them are found only in the one dually-infected (or multiply-infected) individual patient in which they arose. Such recombinant forms are called URFs. If an inter-subtype recombinant virus is transmitted to many people, it becomes one of the circulating strains in the HIV epidemic and thus should be classified as a CRF. Correct subtype classification of clinical HIV-1 samples is a challenging and significant problem in HIV research, due to the frequent recombination of HIV-1 and the importance of subtype difference for epidemiological studies.

In previous work, a quantity of computational techniques have been proposed to categorize HIV-1 strains into subtypes. A popular kind is the widely used alignment-based methods which identify HIV-1 subtypes either by similarity search such as BLAST (Altschul et al., 1990), USEARCH (Edgar, 2010) and KCLUST (Hauser et al., 2013) etc. or by pairwise distance such as PASC (Bao et al., 2014) and DEmARC (Lauber and Gorbalenya, 2012) etc. Another kind of HIV-1 subtype methods is based on phylogenetic analysis such as REGA (Pineda-Pea et al., 2013; De Oliveira T. et al., 2005), SCUEAL (Kosakovsky Pond et al., 2009), and Pplacer (Matsen et al., 2010) etc. Using these methods based on phylogenetic analysis, a query sequence's class can be predicted by its neighbor's information in phylogenetic trees. However, these alignment-based and phylogeny-based approaches are difficult to deal with large data because of time consumption. Thus the alignment-free methods become more and more popular with the increasing amount of data. Among these methods, Comet (Struck et al., 2014) adapted from partial matching compression and Kameris (Solis-Reyes et al., 2018) based on k-mer frequency achieves high accuracy.

* Corresponding author.
E-mail address: yau@uic.edu (S.S.-T. Yau).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.meegid.2019.104080
Received 3 July 2019; Received in revised form 8 October 2019; Accepted 20 October 2019
Available online 01 November 2019

Table 1
Classification Sensitivity (Sens), Specificity (Spec) and AUC are reported for the set of 1718 whole HIV-1 genomes from the LANL database by using SNV (L = 65), Kameris (Kam.) and Comet (Com.). "U": Unique recombinant form (URF).

	Subtype	Sens	Sens	Sens	Spec	Spec	Spec	AUC	AUC	AUC
		SNV	Kam.	Com.	SNV	Kam.	Com.	SNV	Kam.	Com.
		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
CRFs	CRF01_AE	100	100	98.31	100	100	95.79	100	100	99.15
	CRF02_AG	100	100	89.66	100	99.88	96.26	100	99.94	94.83
	CRF06_cpx	100	0	75	100	100	96.20	100	50	87.50
	CRF07_BC	100	100	70	100	100	96.31	100	100	85
	CRF08_BC	100	100	50	100	99.94	96.31	100	99.97	75
	CRF11_cpx	100	100	25	100	99.88	96.32	100	99.94	62.50
	CRF13_cpx	100	0	100	100	100	96.14	100	50	100
	CRF14_BG	100	0	66.67	100	100	96.20	100	50	83.33
	CRF22_01A1	100	0	0	100	100	96.32	100	50	50
	CRF35_AD	100	0	83.33	100	100	96.19	100	50	91.67
	CRF42_BF	100	0	50	100	100	96.26	100	50	75
	CRF63_02A	100	0	0	100	100	96.26	100	50	50
	CRF64_BC	100	0	0	100	100	96.20	100	50	50
	CRF71_BF1	100	0	100	100	100	96.14	100	50	100
	CRF85_BC	100	0	100	100	100	96.14	100	50	100
Pure (M)	A	100	100	95.74	100	99.36	96.17	100	99.68	97.87
	B	100	100	97.75	100	99.14	94.50	100	99.57	98.87
	C	100	100	97.65	100	99.77	95.76	100	99.89	98.82
	D	100	100	82.35	100	99.94	96.29	100	99.97	91.18
	F1	100	100	78.57	100	99.94	96.30	100	99.97	89.29
	G	100	100	100	100	99.57	96.10	100	99.79	100
	H	100	0	100	100	100	96.14	100	50	100
	N	100	0	100	100	100	96.14	100	50	100
	O	100	100	100	100	100	96.11	100	100	100
URF	**"U"**	100	0	100	100	100	96.14	100	50	98.13
Average	Total	100	48	74.40	100	99.90	96.11	100	73.95	87.13

1 Pure includes A-D, F-H, J and K subtypes; CRFs:circulating recombinant forms; URF: unique recombinant form.

In this paper, we propose an alignment-free method called Subsequence Natural Vector (SNV) to identify HIV-1 subtypes. Unlike k-mer methods, SNV not only utilizes the frequency of subsequences of nucleotides but also includes the position and variance information of nucleotides in viral sequences. Two HIV-1 datasets consisting of complete genomes and pol genes respectively retrieved from Los Alamos HIV database are used to evaluate our method. Comparison in running time is also made among SNV, Comet, Kameris and REGA on each dataset, and SNV enjoys a higher time-efficiency than others.

2. Materials and methods

2.1. Natural Vector (NV)

Let $\Phi = \{A, C, G, T\}$ be the set of 4 types of nucleotides, and a sequence ($S = s_1, s_2, \cdots, s_N$, $s_i \in \Phi$, $i = 1, 2\cdots, N$), where N is the length of the DNA sequence. For each $\alpha \in \Phi$, we define a function $\omega_\alpha(\cdot) : \Phi \to \{0, 1\}$, i.e.,

$$w_\alpha(s_i) = \begin{cases} 1, & s_i = \alpha \\ 0, & s_i \neq \alpha \end{cases} \quad i = 1, 2, \cdots, N \tag{1}$$

1. Let $n_\alpha = \Sigma_{i=1}{}^{n} \omega_\alpha(s_i)$ describe the number of nucleotide α in S.
2. Let $\mu_\alpha = \sum_{i=1}^{n} \frac{\omega_\alpha(s_i) * i}{n_\alpha}$ be the mean position of nucleotide α.
3. Let $D_\alpha = \sum_{i=1}^{n} \frac{(i-\mu_\alpha)^2 \omega_\alpha(s_i)}{n_\alpha n}$ be the normalized 2-nd central moment of position of nucleotide α.

Then natural vector of the DNA sequence S is defined as (Deng et al., 2011):

$$(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_A, D_C, D_G, D_T)$$

2.2. Subsequence Natural Vector (SNV)

To capture the local distribution of nucleotides in DNA sequences, we propose a new feature vector based on the previous natural vector method. Given a DNA sequence $S = s_1, s_2, \cdots, s_N$, we divide the sequence into L non-overlapping segments. To normalize all subsequences with equal length, we apply the method proposed in (Zhao et al., 2011). We define q as the quotient and r as the remainder when dividing N by L, i.e.,

$$q = \left[\frac{N}{L}\right], r = N - L * q. \ (0 \leq r < L) \tag{2}$$

The first r segments ($SubStr_1, SubStr_2, \cdots, SubStr_r$) all consist of $q + 1$ nucleotides and the remaining $L - r$ segments ($SubStr_{r+1}, SubStr_{r+2}, \cdots, SubStr_L$) all consist of q nucleotides. L is a preset integer ($L \ll N$) which can be adjusted from datasets. Then we compute the natural vector of each subsequence to obtain L natural vectors. Subsequently those L natural vectors are concatenated into our proposed vector with $12 * L$ dimension.

We call the vector Subsequence Natural Vector (SNV). Specifically, the SNV of the DNA sequence S is defined as:

$$(n_A^{SubStr_1}, n_C^{SubStr_1}, n_G^{SubStr_1}, n_T^{SubStr_1}, \mu_A^{SubStr_1}, \mu_C^{SubStr_1}, \mu_G^{SubStr_1}, \mu_T^{SubStr_1}, D_A^{SubStr_1}, D_C^{SubStr_1}, D_G^{SubStr_1}, D_T^{SubStr_1}, \ldots, n_A^{SubStr_L}, n_C^{sL}, n_G^{SubStr_L}, n_T^{SubStr_L}, \mu_A^{SubStr_L}, \mu_C^{SubStr_L}, \mu_G^{SubStr_L}, \mu_T^{SubStr_L}, D_A^{SubStr_L}, D_C^{SubStr_L}, D_G^{SubStr_L}, D_T^{SubStr_L})$$

In this paper, we choose the L in the following way:

$$L = [M/(12 * log(M))]. \tag{3}$$

where M is the number of viruses in the dataset and [] means rounding down to integer.

Table 2
Classification Sensitivity (Sens), Specificity (Spec) and AUC are reported for the set of 2928 pol genes from the LANL database by using SNV (L = 103), Kameris (Kam.) and Comet (Com.).

	Subtype	Sens	Sens	Sens	Spec	Spec	Spec	AUC	AUC	AUC
		SNV	Kam.	Com.	SNV	Kam.	Com.	SNV	Kam.	Com.
		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
CRFs	CRF01_AE	99.83	100	97.97	100	99.87	98.44	99.92	99.94	98.98
	CRF02_AG	100	100	97.67	100	99.90	98.36	100	99.95	98.84
	CRF04_cpx	100	0	33.33	100	100	98.42	100	50	66.67
	CRF06_cpx	100	0	100	100	100	98.34	100	50	100
	CRF07_BC	100	100	90.91	100	100	98.40	100	100	95.45
	CRF08_BC	100	100	80	100	100	98.48	100	100	90
	CRF11_cpx	100	100	100	100	100	98.34	100	100	100
	CRF13_cpx	100	0	50	100	100	98.38	100	50	75
	CRF14_BG	100	0	0	100	100	98.48	100	50	50
	CRF22_01A1	100	0	0	100	100	98.52	100	50	50
	CRF24_BG	100	0	100	100	100	98.35	100	50	100
	CRF35_AD	100	0	100	100	100	98.34	100	50	100
	CRF42_BF	100	0	100	100	100	98.35	100	50	100
	CRF59_01B	100	0	100	100	100	98.35	100	50	100
	CRF63_02A	100	0	100	100	100	98.35	100	50	100
	CRF64_BC	100	0	33.33	100	100	98.42	100	50	66.67
	CRF71_BF1	100	0	100	100	100	98.35	100	50	100
	CRF83_cpx	100	0	100	100	100	98.35	100	50	100
	CRF85_BC	100	0	0	100	100	98.45	100	50	50
	CRF90_BF1	100	0	100	100	100	98.35	100	50	100
Pure (M)	A	100	100	100	99.96	99.61	98.28	99.98	99.80	99.91
	B	100	100	99.77	100	99.37	97.17	100	99.69	99.89
	C	100	100	98.93	100	99.60	98.18	100	99.80	99.47
	D	100	100	100	100	100	98.34	100	100	100
	F1	100	100	100	100	99.90	98.34	100	99.95	100
	F2	100	0	66.67	100	100	98.38	100	50	83.33
	G	100	100	95.45	100	99.79	98.37	100	99.90	97.73
	H	100	0	100	100	100	98.35	100	50	100
	N	100	0	100	100	100	98.35	100	50	100
	O	100	100	100	100	100	98.34	100	100	100
URF	**"U"**	100	100	100	100	99.38	98.35	100	99.69	99.55
Total	Average	99.99	41.94	82.07	99.99	99.92	98.33	99.99	70.93	91.02

1 Pure includes A-D, F−H, J and K subtypes; CRFs:circulating recombinant forms; URF: unique recombinant form.

Example:

Given a dataset with 400 ($M = 400$) sequences. One of the DNA sequence is $S =$ "*ACGGACTCGTACGGATCGATACGAATCC*", the length of S is 28, i.e., $N = 28$. According to 3, we choose $L = 5$, then: $q = 5$, $r = 3$,

Subsequence	Length
$SubStr_1 =$ "*ACGGAC*"	6
$SubStr_2 =$ "*TCGTAC*"	6
$SubStr_3 =$ "*GGATCG*"	6
$SubStr_4 =$ "*ATACG*"	5
$SubStr_5 =$ "*AATCC*"	5

In order to create reliable datasets, we download full set of HIV-1 complete genomes and pol genes from the well-known Los Alamos HIV database last updated on February 222,019, accessible at (https://www.hiv.lanl.gov/components/sequence/HIV/search/search.html). The pol region is a necessary region for the process of HIV-1 replication which includes protease (PR), reverse transcriptase (RT) and integrase (IN). Genetic variations among subtypes are about 7–20% for gag gene, 20–30% for env gene, and 10% for pol gene (Gao et al., 1998). For the dataset of whole genomes, the query parameters virus: HIV-1, genomic region: complete genome, excluding problematic are applied to test the performance of the proposed method. For the dataset of pol genes, we test the following query parameters virus: HIV-1, genomic region: Pol CDS, excluding problematic. In these datasets, some CRFs classes have few samples, which has an unfair effect on model training. Thus small classes with samples less than 9 are removed. Then the final complete genome dataset includes a total of 6902 HIV-1 full length genomes with an average length of 8955 bp. It contains 25 subtypes, CRFs and URFs. The pol gene dataset includes 11,668 sequences with an average length of 3010 bp. This collection consists of 31 subtypes and CRFs and URFs. The details of subtypes complete genomes and pol genes are available in Tables 1 and 2 respectively.

To validate the performance of our method, for each dataset 75% samples are randomly selected to form a training dataset and the rest 25% samples form test dataset. The training dataset is utilized to build models and the test dataset to evaluate model performance. Note that test datasets do not involve in training process. For subtype classification based on complete genomes, the training dataset comprises 5184 genomes and the test dataset comprises 1718 genomes. For subtype classification based on pol genes, the training dataset comprises 8740 pol gene sequences and the test dataset comprises 2928.

All computations in this paper are done on a Dell PowerEdge R730 equipped with Intel Xeon E5–2667 v3 Processor under Linux Home Premium with 384 GB RAM.

2.3. Linear discriminant analysis

Linear discriminant analysis (LDA) is a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification. We review the Linear discriminant analysis method as follows

Table 3
Classification Sensitivity (Sens), Specificity (Spec) and AUC are reported for the set of 2928 pol genes from the LANL database by using REGA method.

	Subtype	Sens	Spec	AUC
		REGA	REGA	REGA
CRFs	CRF01_AE	98.98%	99.91%	99.45%
	CRF02_AG	95.35%	100%	97.67%
	CRF04_cpx	100%	100%	100%
	CRF06_cpx	100%	100%	100%
	CRF07_BC	100%	100%	100%
	CRF08_BC	75%	100%	87.50%
	CRF11_cpx	100%	100%	100%
	CRF13_cpx	100%	100%	100%
	CRF14_BG	75%	100%	87.50%
	CRF22_01A1	0%	100%	50%
	CRF24_BG	100%	100%	100%
	CRF35_AD	100%	100%	100%
	CRF42_BF	100%	100%	100%
	CRF59_01B	0%	100%	50%
	CRF63_02A	0%	100%	50%
	CRF64_BC	0%	100%	50%
	CRF71_BF1	0%	100%	50%
	CRF83_cpx	0%	100%	50%
	CRF85_BC	0%	100%	50%
	CRF90_BF1	0%	100%	50%
Pure (M)	A	100%	99.61%	99.80%
	B	98.56%	100%	99.28%
	C	99.70%	99.51%	99.60%
	D	100%	100%	100%
	F1	100%	100%	100%
	F2	100%	100%	100%
	G	100%	99.97%	99.98%
	H	100%	100%	100%
	N	100%	100%	100%
	O	100%	99.97%	99.98%
URF	"U"	100%	98.86%	99.43%
Total	Average	72.34%	99.93%	86.14%

1 Pure includes A-D, F—H, J and K subtypes; CRFs:circulating recombinant forms; URF: unique recombinant form.

Table 4
Running time on the datasets of 1718 complete genomes and 2928 pol genes by SNV, Kameris and Comet.

Complete genomes			
	Training-time (second)	Testing-time (second)	Total-time (second)
SNV(L = 65)	5.83	0.037	5.867
Kameris	162.66	2.5	165.16
Comet	–	153.92	153.92
Pol genes			
	Training-time (second)	Testing-time (second)	Total-time (second)
SNV (L = 103)	19.754	0.103	19.857
Kameris	194.04	3.89	197.93
Comet	–	98.6	98.6
REGA	–	About 3 days	About 3 days

(Balakrishnama and Ganapathiraju, 1998). Suppose that we have a set of K classes in a dataset and for each sample we know its class label. Let C_k be indices of the n_k samples in class k, where n_k denotes the number of samples belonging to class k, $n = n_1 + \cdots + n_K$. The centroid for class k is defined as $\bar{x}_k = \sum_{i \in C_k} x_i / n_k$, and the centroid for all classes is defined as $\bar{x} = \sum_i^n x_i / n$. Let W denote the within-class covariance matrix, that is the covariance matrix of the variables centred on the class centroid:

$$W = \frac{\sum_{k=1}^{K} \sum_{j \in C_k} (x_j - \bar{x}_k)(x_j - \bar{x}_k)^T}{n - K} \tag{4}$$

Let B denote the between-class covariance matrix:

$$B = \frac{\sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T}{K - 1} \tag{5}$$

Fisher linear discriminant analysis (LDA) projects high dimension data x into low dimension space to find linear combinations $\omega^T x$ such that between-class variance is maximized relative to the within-class variance, that is, maximizing the ratio:

$$P = \frac{\omega_1^T B \omega_1}{\omega_1^T W \omega_1} \tag{6}$$

This is a generalized eigenvalue problem, with ω_1 being the eigenvector corresponding to the largest eigenvalue of $W^{-1}B$. Similarly the next direction ω_2 can be found by maximizing ${\omega_2}^T B \omega_2 / {\omega_2}^T W \omega_2$ such that ω_2 is orthogonal in W to ω_1. These ω_k are called discriminant coordinates or canonical coordinates. Therefore at most K-1 such directions will be given by $K - 1$ positive eigenvalues.

We might project each sample x onto all directions by ${\omega_k}^T x$, $k = 1, 2, \ldots, K - 1$. Thus each sample is projected onto the subspace spanned by above directions $\omega_1, \ldots, \omega_{K-1}$. There exists a fundamental dimension reduction in LDA, namely, that we only need to only consider samples in a subspace of dimension at most $K - 1$, which is far less than dimension of the sample which is SNV here. For a new sample, its distance to each class centroid is computed in this low dimensional space. We classify the sample to the class with smallest distance.

2.4. Classification

Given a training set of n sequences including K classes, we should first represent each sequence with a $12 * L$ dimensional SNV. We train an LDA model with the training set. We obtain the $K - 1$ directions which span the $K - 1$ dimensional projection space and the projected vector for each SNV. Note that all SNVs become $K - 1$ dimension on the projection space. We also get K centroid denoted as $\mu_1, \mu_2, \cdots, \mu_K$ and each centroid is a $K - 1$ dimensional vector.

Given a HIV-1 sequence S in test dataset represented by SNV, for prediction, we first project it into the projection space and denote the projected vector as P_S. Then we calculate the distance between P_S and all class centroid μ_i, $i = 1, 2, \cdots, K$ as follows:

$$d_i = \|P_S - \mu_i\|_2, \; i = 1, 2, \cdots, K \tag{7}$$

where $\|\cdot\|$ is Euclidean distance. We denote the class μ_{01} with smallest distance d_{01} from P_S to the class centroid. According to the traditional LDA method, this virus will be classified into class μ_{01}. However, the classification may be incorrect when the virus belongs to a new class that does not appear in training dataset.

To overcome the misclassification may also due to the high sequence similarity between classes of traditional LDA, we propose an improved LDA using a ratio to remove unreliable prediction. Suppose the distance from P_S to centroid of class μ_{02} is the second smallest distance among d_i, $i = 1, 2, \cdots, K$ and the distance is denoted d_{02}. Then we compute the ratio of d_{01} and d_{02}: $Ratio = d_{01}/d_{02}$. If $Ratio < 0.9$, we predict virus S belongs to class μ_{01}. Otherwise, we determine virus S as" Unassigned".

Finally, the widely used Sensitivity, Specificity and AUC are chosen to evaluate classification performance. The definition of these measures are as follows:

$$Sensitivity = TP/(TP + FN) \tag{8}$$

and

$$Specificity = TN/(FP + TN) \tag{9}$$

where TP, TN, FP, and FN are the number of true positive, true negative, false positive and false negative predictions respectively. ROC curve is created by plotting the Sensitivity against 1-Specificity at various threshold settings. AUC is the area under the ROC curve.

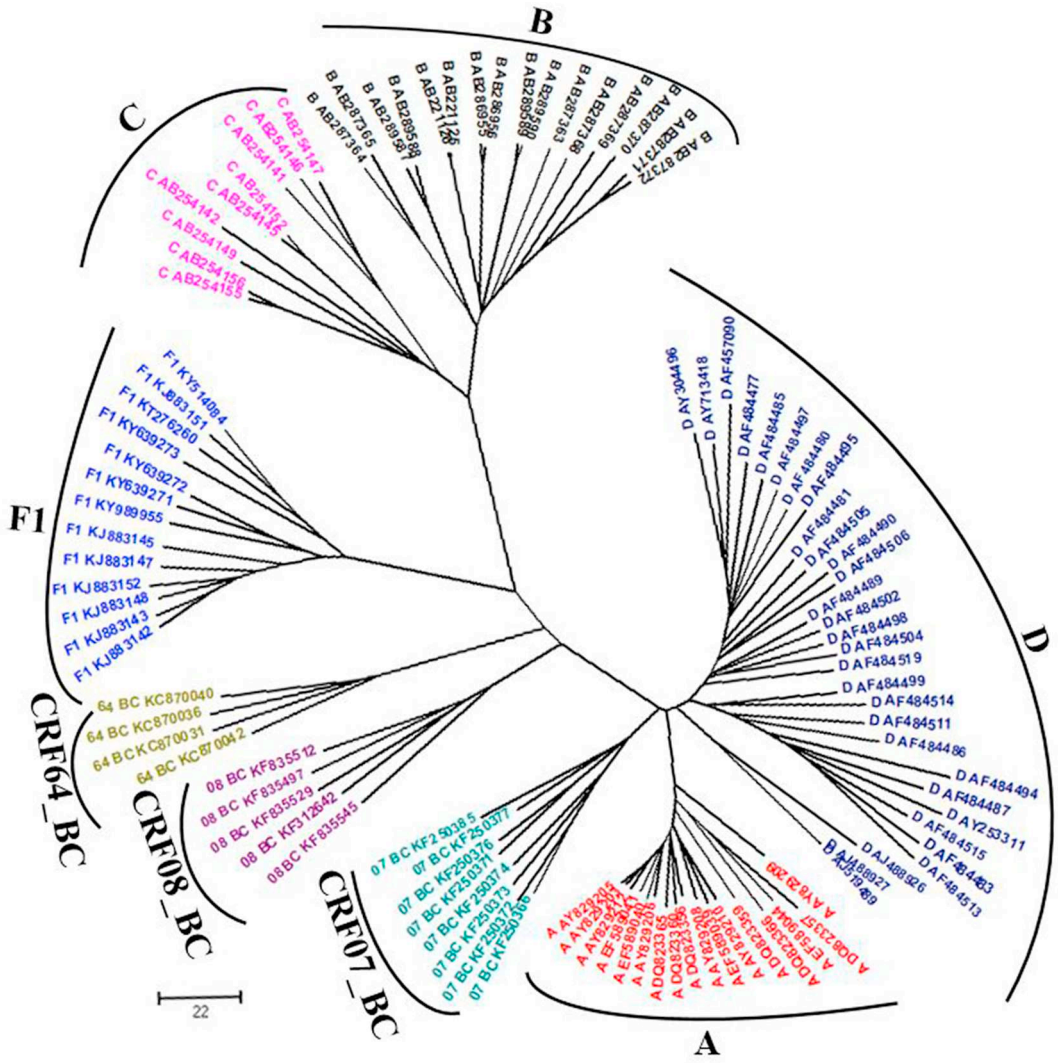

Fig. 1. Unrooted neighbor-joining tree depicting the genetic relationship among the 102 complete genomes including 8 subtypes and CRFs by SNV method. Each group is highlighted with a different color.

3. Results

In our work, the classification of HIV-1 viruses is performed on complete genomes and the pol gene datasets respectively. Although different URF has different recombinant form, we view URFs as a group. Based on our method we can capture their common features and determine whether they are URF or not. To make comparison, the Kameris, Comet and REGA methods are also applied to the HIV-1 subtype classification.

3.1. Results based on complete genome

For HIV-1 complete genome datasets consisting of both training and test datasets ($M = 6902$), L is chosen to be 65 according to Eq. (3). Traditional LDA is then trained on the whole complete genome training set of 5184 viruses. Our improved LDA is used to classify the 1718 complete genomes in the test set. The performance is displayed in Table 1. We also provide the numbers of training, unassigned viruses for each subtype in Table S2 in supporting files. As shown in Table 1, for each subtype, Sensitivity, Specificity and AUC of our SNV based method are all 100%. Especially for HIV-1 strains in subtype URF, we can predict whether they are URFs or not correctly. Its average Sensitivity and AUC are only 48% and 73.95% respectively, due to the fact that Kmaeris can't deal with subtypes including small samples. For Comet method, the average Sensitivity, Specificity and AUC are 74.40%, 96.11% and 87.13% respectively. Since the Sensitivity of Comet is 100%, the predicted viruses in URF class truly belong to URF. However, the method misclassifies some strains into URF which results in the Specifcity not 100%. Although *CRF64_BC*, *CRF85_BC* and *H* have only nine sequences, our approach can achieve 100% Sensitivity, Specificity and AUC. The length and the number of virus is too large for REGA, we can't apply REGA to the complete dataset. This proves that our approach is able to predict HIV-1 classes not only on class with large samples but also on class with few samples. In summary, our SNV method works much better than Kameris and Comet in Sensitivity, Specificity and AUC.

To compare with REGA subtyping tool for complete genomes, we randomly select 90 complete HIV genomes from the test data consisting of 1718 complete genomes and make sure each subtype in the test data occurs in the 90 genomes. It takes about 3 h to classify the 90 viruses using online REGA subtyping software. The result shows that 67 viruses are not applicable as the software produces errors during analysis paupfragment. For the rest 23 viruses, 4 viruses are incorrectly predicted. However, our method is able to classify all the 1718 viruses correctly as shown in Table 1. Accession numbers of the 90 sequences are shown in supplementary files.

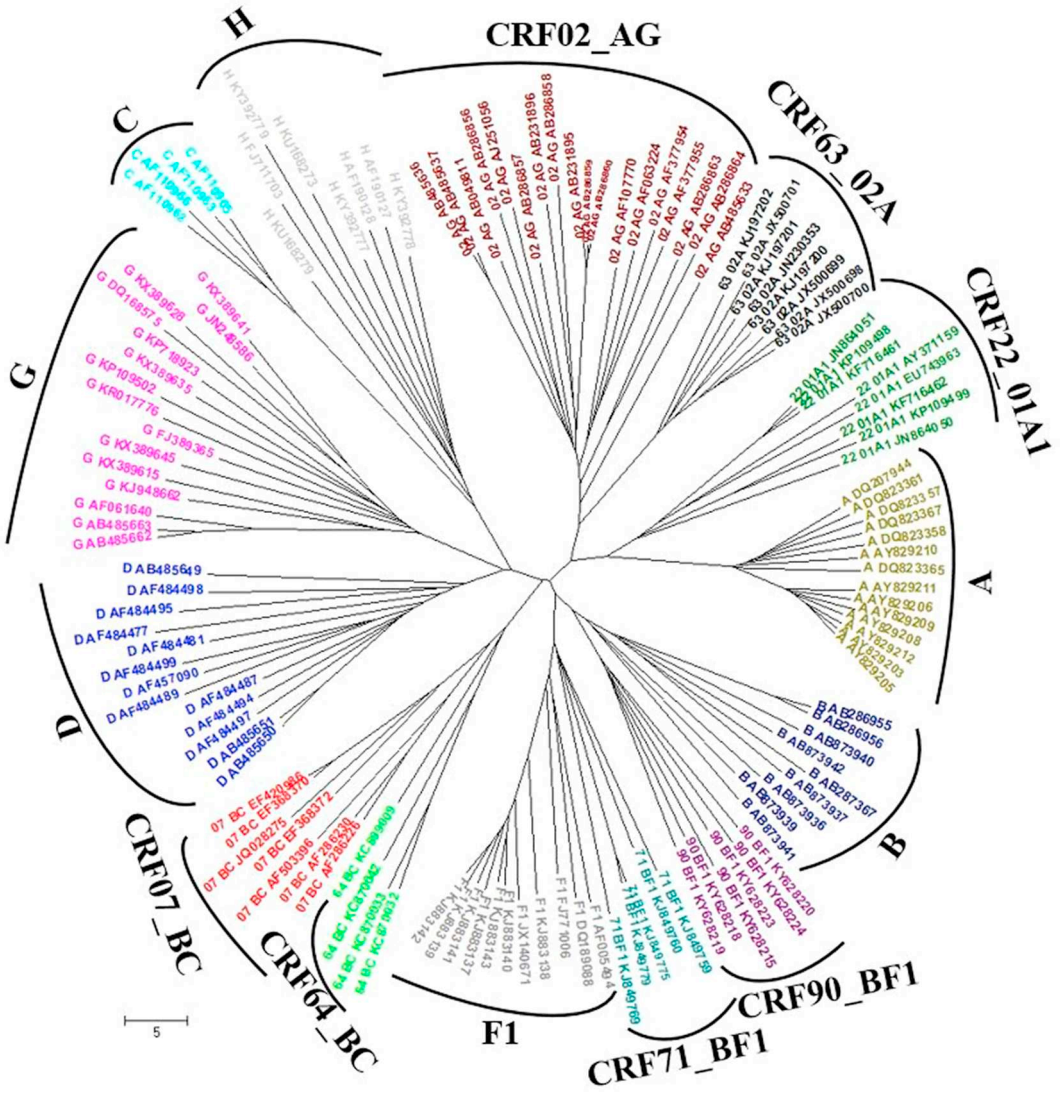

Fig. 2. Unrooted neighbor-joining tree depicting the genetic relationship among the 130 pol genes including 14 subtypes and CRFs by SNV method. Each group is highlighted with a different color.

3.2. Results based on pol genes

For HIV-1 pol gene data consisting of both training and test datasets, the value of $L = 103$ is chosen according to Eq. (3). Traditional LDA method is trained on the whole pol gene training set of 8740 viruses. Our improved LDA method is applied to predict the class label for 2928 viruses in pol gene test set. The results are reported in Table 2. We also provide the numbers of training, unassigned viruses for each subtype in Table S3 in supporting files. As shown in Table 2, Sensitivity of our SNV based method reaches 100% except CRF01_AE class. The sensitivity of CRF01_AE class is 99.83%. Thus the average Sensitivity of our method is 99.99%. Except pure A subtype with 99.96% Specificity, the Specificity of all other subtypes is 100%. Thus the average Specificity of our method is as high as 99.99%. The average AUC of our method obtains more than 99.99%. For Kameris, the average of Sensitivity and AUC are just 41.94% and 70.93% respectively. For Comet, the average of Sensitivity and AUC are just 82.07% and 91.02% respectively. For URF class, the tree methods all can make correct prediction for true URF. But Kameris and Comet may incorrectly classify viruses from other subtypes into URF. Only our method can identify URF correctly. Note that some classes especially CRFs only contain a few viruses. For these subtypes, our approach achieves almost 100% Sensitivity while Kameris always obtains zero Sensitivity and Comet obtains low Sensitivity even zero. We also use REGA method based on alignment of HIV-1 viruses. The results for this method is shown in Table 3. This method gets low Sensitivity and AUC on CRFs such as CRF22_01A1, CRF59_01B and CRF63_02A. Our SNV method outperforms Kameris, Comet and REGA in both Specificity and Sensitivity for small subtypes.

3.3. Results of running time

MATLAB R2016a is utilized for calculation of all SNV and R 3.5.0 for model training and testing. The parameter k for Kameris is chosen as k = 6. Since the online server of Comet has been trained and the time for its training is unknown, we only compute its running time for HIV-1 subtype classification. The time consumption for SNV, Kameris,Comet and REGA is shown in Table 4. For HIV-1 identification based on complete genomes, the total running time of SNV is only about 5.8 s which is much less than Kameris (165 s) and Comet (> 153 seconds). Based on pol genes, the total running time of SNV is only about 19.8 s, while the time of Kameris and Comet are about 197 s and at least 98 s respectively. Since REGA is based on multiple sequence alignment, the time of REGA is about 3 days which is too long to accept.

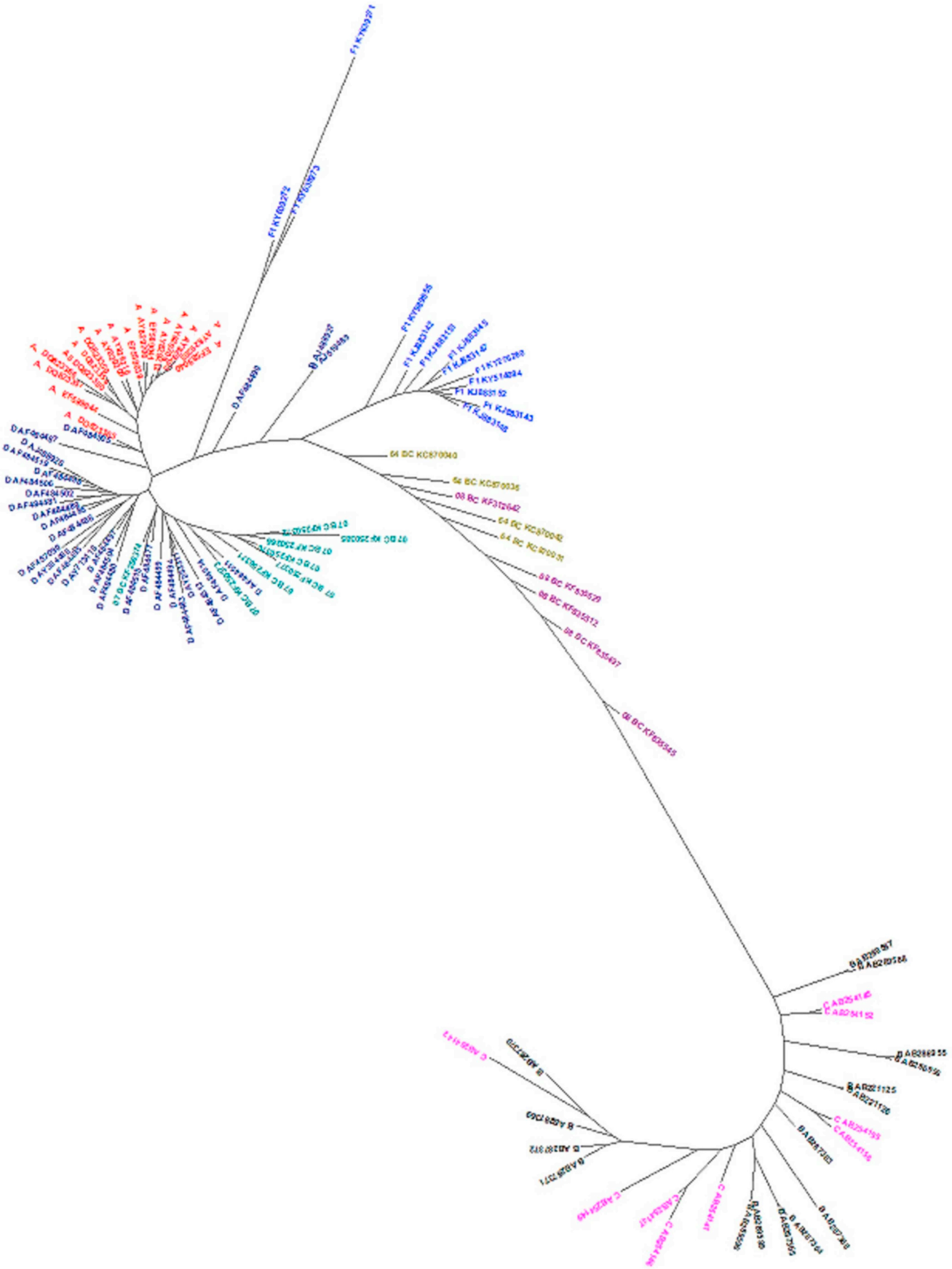

Fig. 3. The neighbor-joining tree depicting the genetic relationship among the 102 complete genomes including 8 subtypes and CRFs by NV method. Each group is highlighted with a different color.

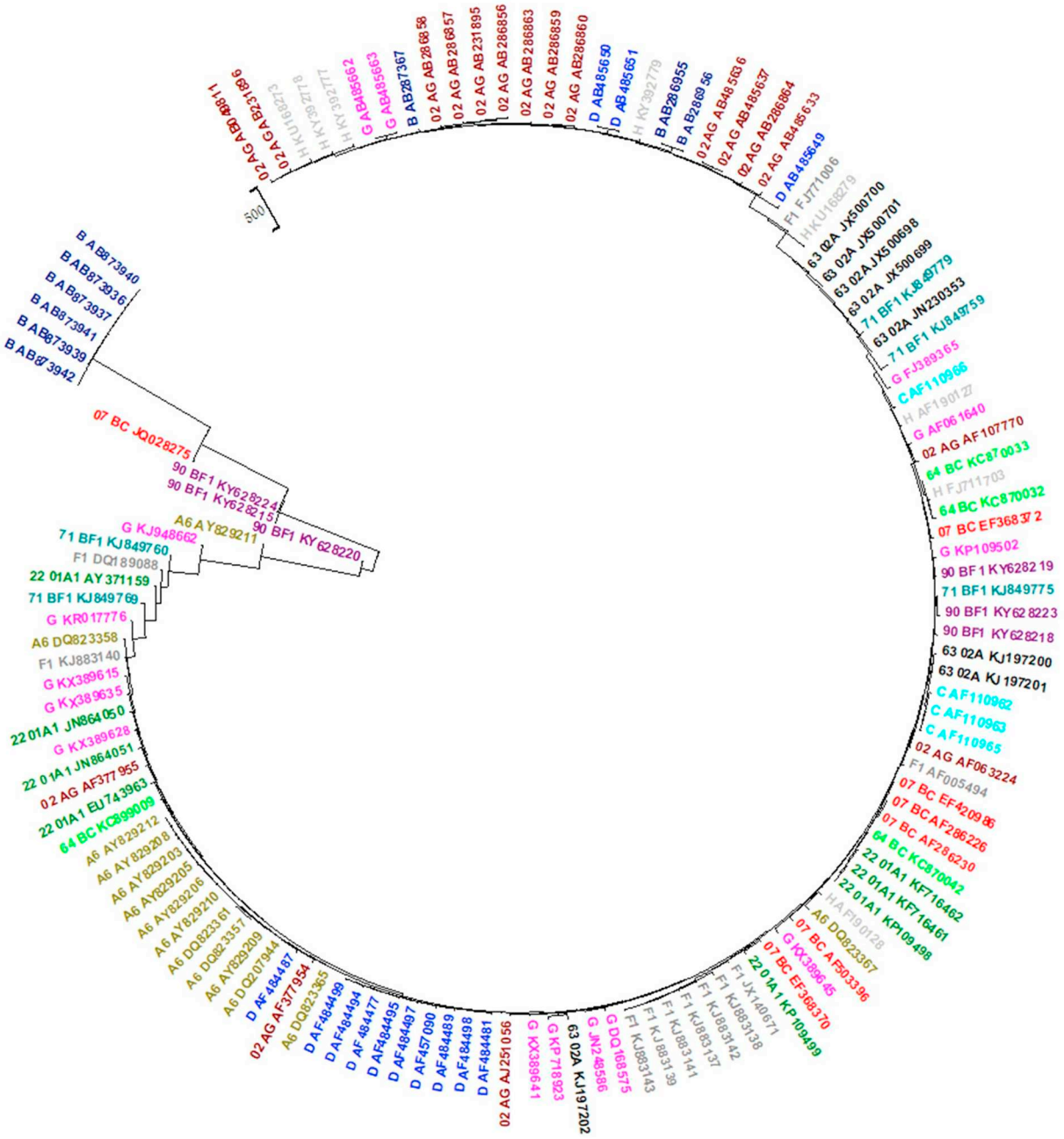

Fig. 4. Phylogenetic tree depicting the genetic relationship among the 130 pol genes including 14 subtypes and CRFs by NV method. Each group is highlighted with a different color.

3.4. Results of phylogenetic trees

To further validate the advantage of our SNV representation for HIV-1 viruses, we construct phylogenetic trees of HIV-1 based on complete genomes and pol genes respectively. We choose some pure subtypes and CRFs recombined by these pure subtypes. These strains are close in evolution and hard to separate from each other. Thus we randomly select 102 complete genomes consisting of some subtypes such as B, C and their recombinants such as CRF64_BC and CRF07_BC. Based on the SNV representation of complete genomes, an unrooted neighbor-joining tree is constructed by Mega 7 (Kumar et al., 2016). Each group is highlighted with a different color in the phylogenetic tree. As shown in Fig. 1, the 102 complete genomes are divided into 8 subtypes and CRFs: A, B, C, D, F_1, CRF07_BC, CRF08_BC, and CRF64_BC. Although CRF07_BC, CRF08_BC and CRF64_BC are recombinant from B and C subtype, our method can effectively cluster each group together. HIV-1 strains from the same pure subtype are also clustered together.

We randomly select 130 pol genes including subtypes such as B, F1 and their recombinants such as CRF71_BF1 and CRF90_BF1. Then we compute the SNV representation of each sequence. An unrooted neighbor-joining tree of these genes is constructed by Mega 7. As shown in Fig. 2, these pol genes consist of 14 subtypes and CRFs. HIV-1 strains from the same subtype are clustered together. Moreover, CRF02_AG, CRF63_02A, CRF22_01A1 and A groups are cluster together correctly. CRF71_BF1, CRF90_BF1, B and F1 form a lineage. CRF07_BC and CRF64_BC become a clade. From the two phylogenetic trees, we can see that SNV representation is useful to build HIV-1 evolutionary relationship.

To compare our SNV with previous NV method, we build the neighbor-joining tree for the 102 complete genomes. As shown in Fig. 3,

Table 5
Classification Sensitivity (Sens), Specificity (Spec) and AUC are reported for p51RT from the LANL database by using SNV and Comet (Com.). U: Unique recombinant form (URF). Ave. means average on the related column. Pure includes A-D, F−H, J and K subtypes; CRFs: circulating recombinant forms.

	Subtype	Test	Sens	Sens	Spec SNV	Spec	SNV	AUC
		Number	SNV	Com.	SNV	Com.	SNV	Com.
CRFs	CRF01_AE	812	99.88%	99.38%	100.00%	100.00%	99.94%	99.69%
	CRF02_AG	106	100.00%	82.08%	100.00%	100.00%	100.00%	91.04%
	CRF04_cpx	3	100.00%	33.33%	100.00%	100.00%	100.00%	66.67%
	CRF06_cpx	17	100.00%	76.47%	99.98%	100.00%	99.99%	88.24%
	CRF07_BC	185	98.92%	85.41%	99.97%	100.00%	99.44%	92.70%
	CRF08_BC	109	100.00%	88.07%	99.95%	99.97%	99.97%	94.02%
	CRF11_cpx	7	100.00%	57.14%	99.98%	100.00%	99.99%	78.57%
	CRF12_BF	3	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CRF13_cpx	3	100.00%	0.00%	100.00%	100.00%	100.00%	50.00%
	CRF14_BG	4	100.00%	0.00%	99.92%	100.00%	99.96%	50.00%
	CRF22_01A1	5	100.00%	0.00%	100.00%	100.00%	100.00%	50.00%
	CRF24_BG	3	100.00%	66.67%	100.00%	100.00%	100.00%	83.33%
	CRF35_AD	6	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CRF42_BF	5	100.00%	0.00%	100.00%	100.00%	100.00%	50.00%
	CRF55_01B	3	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CRF59_01B	3	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CRF63_02A	4	100.00%	0.00%	100.00%	100.00%	100.00%	50.00%
	CRF64_BC	3	100.00%	0.00%	99.97%	100.00%	99.98%	50.00%
	CRF65_cpx	4	100.00%	75.00%	100.00%	100.00%	100.00%	87.50%
	CRF83_cpx	3	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	CRF85_BC	3	100.00%	0.00%	100.00%	100.00%	100.00%	50.00%
	CRF90_BF1	3	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Pure (M)	A1	152	100.00%	100.00%	99.98%	98.99%	99.99%	99.49%
	A2	20	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	A3	4	100.00%	0.00%	100.00%	100.00%	100.00%	50.00%
	A6	29	100.00%	0.00%	100.00%	100.00%	100.00%	50.00%
	B	2744	99.96%	99.74%	99.91%	99.76%	99.94%	99.75%
	C	1609	99.69%	99.63%	100.00%	99.20%	99.84%	99.41%
	D	146	97.95%	98.63%	100.00%	99.95%	98.97%	99.29%
	F1	19	100.00%	100.00%	100.00%	99.97%	100.00%	99.98%
	F2	4	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	G	38	86.84%	97.37%	100.00%	99.88%	93.42%	98.63%
	N	4	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	O	16	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
URF	U	5	80.00%	60.00%	100.00%	99.52%	90.00%	79.76%
Ave.	Total		98.95%	66.25%	99.99%	99.92%	99.47%	83.09%

1 Pure includes A-D, F-H, J and K subtypes; CRFs:circulating recombinant forms; URF: unique recombinant form.

the subtypes A, D and CRF_07BC are incorrectly clustered together. The subtypes B and C are incorrectly mixed together. The CRF_08BC and CRF_64BC should not be placed together. Using NV method, we also construct the neighbor-joining tree for the 130 pol genes. As shown in Fig. 4, the result is poor because all 14 subtypes and CRFs are incorrectly placed. Therefore, our SNV performs much better than NV.

3.5. Results of partial genomic sequences

To test the performance of our classifier for partial genomic sequences, we download 24,576 p51RT fragments of pol genes from the LANL database. The p51RT gene is the major fragment of a pol gene and has about 1320 nucleotides. We repeat the training and test process for this dataset. L is chosen as 10 by experience. Our method is also compared with Comet method. The results are reported in Table 5. As shown in Table 5, our method achieves 98.95% sensitivity, 99.99% specificity and 99.47% AUC on average. Comet method achieves 66.25% sensitivity, 99.92% specificity and 83.09% AUC on average. Thus our method performs much better than Comet specifically in sensitivity. Except pure D and G subtype, our method achieves same or better AUC than Comet. For pure D and G subtypes, although the sensitivity of our method is lower than that of Comet, the specificity of our method is better than that of Comet. For 23 of 35 kinds of subtypes, our method achieves better than Comet with respect to AUC.

4. Discussion and conclusions

In this paper, we propose an efficient Subsequence Natural Vector (SNV) to encode HIV-1 viruses and improve traditional linear discriminant analysis (LDA) to subtype viruses. For each virus, we first use SNV representation to encode HIV-1 sequences into numerical vectors. By traditional LDA classification method, each SNV is projected into a low dimensional space whose dimension is the number of classes. Thus HIV-1 viruses are identified in the reduced space by their distance to class centroid. HIV-1 viruses are split into training data and test data. The training data is used to estimate the parameters such as discriminant coordinates in the LDA method. These discriminant coordinates vectors form the projection space on which each SNV is represented by a much lower dimensional vector. For each HIV-1 virus in test set, the distance to all class centroid in the projection space is computed. Unlike the traditional LDA, we introduce a ratio of the smallest distance to the second smallest distance to predict the class label of viruses reliably. For a virus, we predict it should belong to the class where the distance is smallest if the ratio is less than 0.9. Otherwise, we do not make prediction for this virus. Our improved LDA is able to identify URFs and to reduce the prediction error when classes have high similarity.

SNV representation is based on distribution of each nucleotide in viral sequences. According to those results for HIV-1 subtype classification based on complete genomes and pol genes, our improved LDA

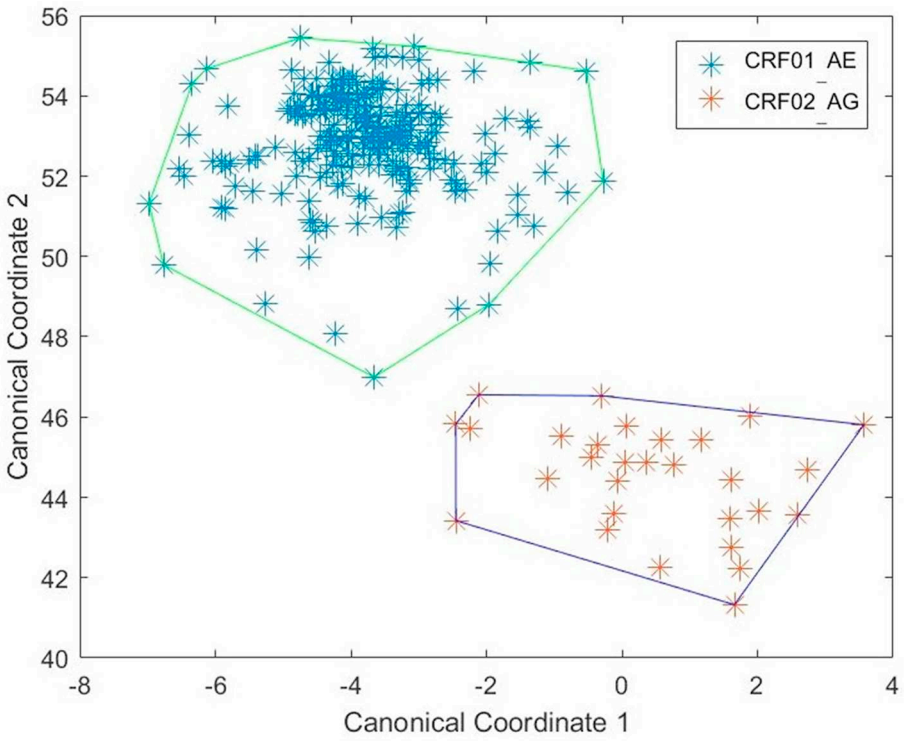

Fig. 5. Two dimensional projection for 236 CRF01_AE and 29 CRF02_AG complete genomes. The blue and orange convex polygon are formed by CRF01_AE and CRF02_AG viruses respectively. The two subtypes are clearly separate from each other. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

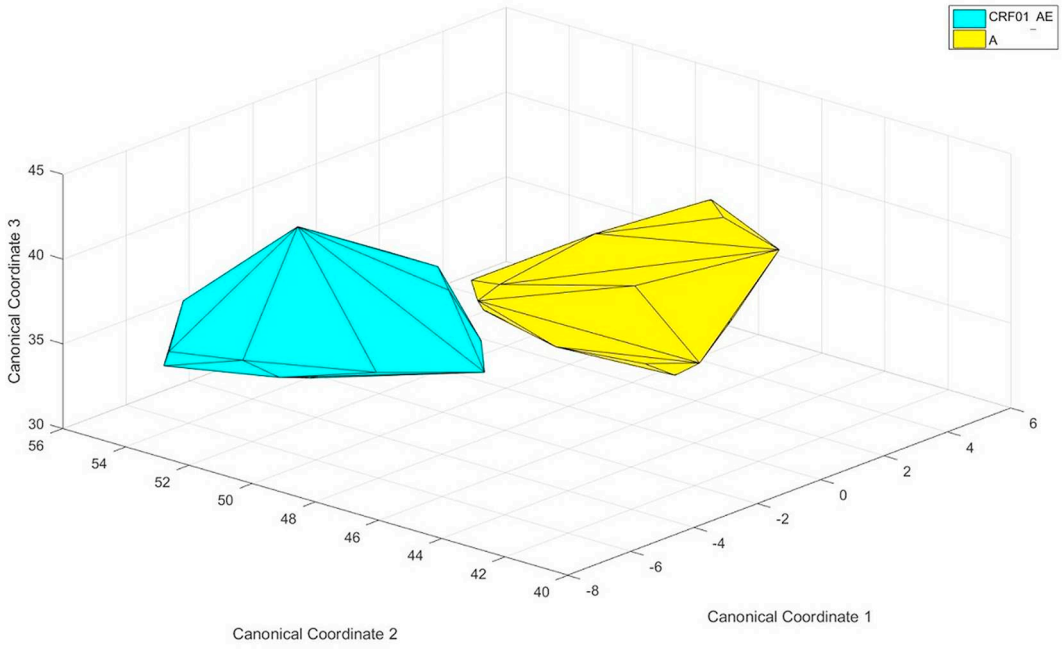

Fig. 6. Three dimensional projection for 236 CRF01_AE and 95 A complete genomes. The cyan and yellow convex polygon are formed by CRF01_AE and A viruses respectively. The two subtypes are clearly separate from each other. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

approach based on SNV representation can achieve almost 100% Sensitivity and Specificity on both datasets. Moreover, our method performs better than Kameris, Comet and REGA especially for these subtype with few samples. Compared with Comet, our approach obtains higher Sensitivity and Specificity for both complete genome and pol gene datasets, with a very robust characteristic. Compared with REGA, our method not only runs far more fast but also shows better performance for all subtypes. For both datasets, we can get close to 100% Sensitivity for classes with a few sequences while Kameris can't work. From the perspective of dimension reduction, our method projects each sequence onto a $12 * L$ dimensional space while Kameris projects each sequence into a $4^6 = 4096$ dimensional space.

In this paper, L was chosen by experience. For Eq. (3), according to some numerical simulation, when the size of dataset M changes in a certain range, L keeps unchanged. For example, for 6902 complete genomes used in our paper, L is always 65 if $6894 \leq M \leq 7013$. Thus L

is unchanged when some new unknown viruses are predicted by our trained classifier. If a lot of new viruses are imported, we indeed need to train our model again.

As comparison, we also choose L = [d/12*log(d)],where d is the average length of sequences. The 6902 whole genomes are used to test this new rule. Since the average length of this dataset is 8956, the value of L is 82. Using this L, we repeat the model training and test process. The classification results are listed in Table S1. From Table S1 we can see, 5 viruses from subtype B and C, CRFs11_cpx, CRF42_BF and CRF71_BF1 are incorrectly classified while no virus is incorrectly classified using our original L.

LDA method is able to represent HIV-1 viruses in low dimensional space using our SNV representation. Moreover, LDA can even separate some different subtypes in 2 or 3 dimensional space which is clear for visualization. To illustrate the point clearly, we choose all CRF01_AE and CRF02_AG from the complete test dataset removing the unassigned viruses. Therefore, there are 236 CRF01_AE viruses and 29 CRF02_AG viruses are chosen. These viruses are then represented by SNV vectors. Using LDA method, we project the SNV vectors into 2 dimensional plane constructed by the first two canonical coordinate vector. As show in Fig. 5, the CRF01_AE viruses and CRF02_AG viruses form two different classes respectively. The boundary of each class forms a convex polygon in plane. Note that the two classes are clearly separate for each other. In addition, we collect all 236 subtype CRF01_AE viruses and 95 pure A viruses from test dataset and project them into three dimensions. As shown in Fig. 6, CRF01_AE and A viruses form two classes and the boundary of each class forms a convex polyhedron. The two classes CRF01_AE and A are separate from each other. From the two figures we can see, LDA method using our new SNV representation may separate some subtypes in 2 or 3 space which is clear for visualization.

Acknowledgements

This study is supported by the National Natural Science Foundation of China (91746119), Tsinghua University Education Foundation fund (042202008) and Tsinghua University start-up fund. Stephen S.-T. Yau is grateful to National Center for Theoretical Sciences (NCTS) for providing excellent research environment while part of this research was done.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2019.104080.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Balakrishnama, S., Ganapathiraju, A., 1998. Linear discriminant analysis-a brief tutorial. Inst. Sig. Inf. Process. 18, 1–8.
Bao, Y., Chetvernin, V., Tatusova, T., 2014. Improvements to pairwise sequence comparison (pasc): a genome-based web tool for virus classification. Arch. Virol. 159, 3293–3304.
De Oliveira, T., Deforche, K., Cassol, S., et al., 2005. An automated genotyping system for analysis of hiv-1 and other microbial sequences. Bioinformatics 21, 3797–3800.
Deng, M., Yu, C., Liang, Q., He, R.L., Yau, S.S.T., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS One 6, e17293.
Edgar, R.C., 2010. Search and clustering orders of magnitude faster than blast. Bioinformatics 26, 2460–2461.
Gao, F., Robertson, D.L., Carruthers, C.D., et al., 1998. A comprehensive panel of near-full-length clones and reference sequences for non-subtype b isolates of human immunodeficiency virus type 1. J. Virol. 72, 5680–5698.
Hauser, M., Mayer, C.E., Sding, J., 2013. kclust: fast and sensitive clustering of large protein sequence databases. BMC Bioinforma. 14, 248.
Kosakovsky Pond, S.L., Posada, D., Stawiski, E., et al., 2009. An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in hiv-1. PLoS Comput. Biol. 5, e1000581.
Kumar, S., Stecher, G., Tamura, K., 2016. Mega7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874.
Lauber, C., Gorbalenya, A.E., 2012. Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses. J. Virol. 86, 3890–3904.
Matsen, F.A., Kodner, R.B., Armbrust, E.V., 2010. pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinforma. 11, 538.
Nuno, R.F., Andrew, R., Marc, A.S., Guy, B., Trevor, B., et al., 2014. The early spread and epidemic ignition of hiv-1 in human populations. Science 346, 56–61.
Pineda-Pea, A.C., Faria, N.R., Imbrechts, S., et al., 2013. Automated subtyping of hiv-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new rega version 3 and seven other tools. Infect. Genet. Evol. 19, 337–348.
Solis-Reyes, S., Avino, M., Poon, A., Kari, L., 2018. An open-source k-mer based machine learning tool for fast and accurate subtyping of hiv-1 genomes. PLoS One 13, e0206409.
Struck, D., Lawyer, G., Ternes, A.M., Schmit, J.C., Bercoff, D.P., 2014. Comet: adaptive context-based modeling for ultrafast hiv-1 subtype identification. Nucleic Acids Res. 42, e144.
Zhao, B., He, R.L., Yau, S.S.T., 2011. A new distribution vector and its application in genome clustering. Mol. Phylogenet. Evol. 59, 438–443.