VIROLOGICA SINICA, August 2013, 28 (4): 228-238
DOI 10.1007/s12250-013-3348-z
www.virosin.org
Email: info@virosin.org

Research Article

Inference of Global HIV-1 Sequence Patterns and Preliminary Feature Analysis

Yan Wang[1], Reda Rawi[2], Daniel Hoffmann[2], Binlian Sun[1]✉ and Rongge Yang[1]✉

1. AIDS and HIV Research Group, State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan 430071, China;
2. Research Group for Bioinformatics, Center for Medical Biology, University of Duisburg-Essen, Essen 45141, Germany

The epidemiology of HIV-1 varies in different areas of the world, and it is possible that this complexity may leave unique footprints in the viral genome. Thus, we attempted to find significant patterns in global HIV-1 genome sequences. By applying the rule inference algorithm RIPPER (Repeated Incremental Pruning to Produce Error Reduction) to multiple sequence alignments of Env sequences from four classes of compiled datasets, we generated four sets of signature patterns. We found that these patterns were able to distinguish southeastern Asian from non-southeastern Asian sequences with 97.5% accuracy, Chinese from non-Chinese sequences with 98.3% accuracy, African from non-African sequences with 88.4% accuracy, and southern African from non-southern African sequences with 91.2% accuracy. These patterns showed different associations with subtypes and with amino acid positions. In addition, some signature patterns were characteristic of the geographic area from which the sample was taken. Amino acid features corresponding to the phylogenetic clustering of HIV-1 sequences were consistent with some of the deduced patterns. Using a combination of patterns inferred from subtypes B, C, and all subtypes chimeric with CRF01_AE worldwide, we found that signature patterns of subtype C were extremely common in some sampled countries (for example, Zambia in southern Africa), which may hint at the origin of this HIV-1 subtype and the need to pay special attention to this area of Africa. Signature patterns of subtype B sequences were associated with different countries. Even more, there are distinct patterns at single position 21 with glycine, leucine and isoleucine corresponding to subtype C, B and all possible recombination forms chimeric with CRF01_AE, which also indicate distinct geographic features. Our method widens the scope of inference of signature from geographic, genetic, and genomic viewpoints. These findings may provide a valuable reference for epidemiological research or vaccine design.

Pattern inference, global HIV-1 sequence, Repeated Incremental Pruning to Produce Error Reduction (RIPPER)

INTRODUCTION

During the processes of independent cross-species transmission, different HIV lineages were formed, and these included HIV-1 M, N, O, and P, and HIV-2. The HIV-1 M group has been further subdivided into nine subtypes, A-D, F-H, J, and K, according to the variation in genetic distance of these amino acids. This variation is generally 8-17% and up to 30% within subtypes, whereas between subtypes, it is generally 17-35% and up to 42%, depending on the genomic regions used for subtyping (Hemelaar J, 2012; Sharp P M, et al., 2011). With the increasing sensitivity and range of sequencing techniques, increasing numbers of circulating recombinant forms (CRFs) have been reported.

The globally uneven distribution of the different HIV-1 subtypes and CRFs reflects the molecular epidemiology of the virus. In southern and eastern Africa, the predominant subtype is C, and this makes up 52% of HIV-1 infections worldwide. By contrast, in West and Central Africa, the vast majority of infections are caused by CRF02_AG, while in East Africa, subtypes A and D and their CRFs are the dominant subtypes (Delatorre E O, et al., 2012;

Received: 2013-06-04, Accepted: 2013-07-26
Online First: 2013-07-27

✉ Corresponding authors.
Binlian Sun: Phone: +86-27-87198736, Fax: +86-27-87198736,
Email: sunbl@wh.iov.cn
Rongge Yang: Phone: +86-27-87198736, Fax: +86-27-87198736,
Email: ryang@wh.iov.cn

Hemelaar J, 2012; Kallings L O, 2008; Morris C N, et al., 2006; Njai H F, et al., 2006; Pollakis G, et al., 2003; Shen C, et al., 2011; Tebit D M, et al., 2011; Worobey M, et al., 2008; Zhu T, et al., 1998). Within the homosexual populations in North and South America, Western and Central Europe, Australia, Asia (for example, Hong Kong, Japan, Korea, Taiwan etc.), North Africa, the Middle East, South Africa, and Russia, subtype B is the predominant subtype (Buonaguro L, et al., 2007; Delatorre E O, et al., 2012; Gilbert M T P, et al., 2007; Junqueira D M, et al., 2011; Moran D, et al., 2007; Paraskevis D, et al., 2009). In South America, in addition to the B, C, F, and BF subtypes, recombinant virus subtypes also coexist, and infections caused by the BF recombinant viruses (including CRF12_BF, CRF17_BF, CRF29_BF, and CRF29_BF) accounted for 80% of the HIV-1 infections in Argentina. In Eastern Europe, A1 is the predominant subtype, but subtypes B and CRF03_AB are also common in this region (Bello G, et al., 2007; Masciotra S, et al., 2000; Paraschiv S, et al., 2012; Pérez L, et al., 2006; Sierra M, et al., 2007; Silveira J, et al., 2012; Villanova F E, 2010; Walker P R, et al., 2005).

In contrast to Africa, all subtypes in Asia seem to have originated from different founder events, including the CRF01_AE, B, and C subtypes, as well as the various CRFs derived from these three subtypes. It is worth mentioning that the B subtype in Asia can also be divided into two types; in evolutionary terms, one is closer to the subtype B found in Europe and America, while the other is genetically distant, forming a clear clustering branch in the phylogenetic tree called B′ or Thai B. The coexistence of HIV-1 subtypes in East Asia leads to various CRFs, which are dominant in particular regions such as the BC recombinant epidemic among drug users in Northwestern and Southeastern China, and the various Thai-B and CRF01_AE recombinants found in Thailand and Myanmar (Li Y, et al., 2010; Liao H, et al., 2009; Liu J, et al., 2011; Meng Z, et al., 2012).

The central role that HIV diversity plays in HIV transmission suggests the necessity for global HIV epidemic monitoring and a reasonable sampling strategy. In addition, studies of the association of diversity with spread, viral load, and disease progression may also give crucial clues for the prevention and treatment of HIV (Butler I F, et al., 2007; Fryer H R, et al., 2011; Restif O, 2009; Spira S, 2003; Taylor B S, et al., 2008).

Exploration of the signature patterns in the HIV genome could be the first step toward studying HIV diversity. Data mining of biological sequence requires identifying the rules, extracting features and inferring models from a large but specific biological dataset in order to classify, recognize or predict new data. This usually involves pattern mining and clustering of biological sequences, and these two techniques can usually be used interchangeably (Poonpiriya V, et al., 2008). The performance and effectiveness of the various biological sequence pattern mining and clustering methods differ, depending on the characteristics of the algorithms and the datasets used (Cai Y-D, et al., 2010; Dybowski J N, et al., 2011; Zhao Y, 2011).

Although traditional phylogenetic analysis of HIV sequences supports study of HIV origin, evolution, and dissemination, it is generally unsuitable for application to large samples because of the computational requirements (Blair C, et al., 2011). In the current study, we used an efficient method of data mining known as RIPPER (Repeated Incremental Pruning to Produce Error Reduction). This method is suitable for large-scale sample analysis (Avenue M, et al., 1994) to comprehensively analyze global HIV sequence patterns. We particularly focused on analyzing the Env regions, which cover most of the currently available datasets and include the maximum amount of information (Lynch R M, et al., 2009).

In our study, we compiled four datasets from four HIV-1 pandemic hotspots with different epidemiological and evolutionary features: Southeast Asia, China, Africa, and Southern Africa, and focused in our analysis on answering the following three questions.

1) For the four epidemiological hotspots with different epidemiological features, can we identify signature patterns that are characteristic of HIV-1 sequences from the four geographic classes?

2) Is the performance of the signature pattern inference the same for all four datasets?

3) Can we understand the scope of signature pattern analysis and the application of these patterns?

MATERIALS AND METHODS

The global HIV-1 sequences and associated information were retrieved from the Los Alamos HIV sequence database (http://www.hiv.lanl.gov/).

Dataset compilation

The dataset was downloaded from LANL HIV Sequence Alignments

(http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html) by setting Alignment type as Filter alignment (all complete sequences), Year as 2011, Organism as HIV-1, DNA/Protein as PRO, Region as Env, and Subtype as ALL.

The original downloaded dataset comprised 3,261 sequences.

After removing problematic sequences, especially those with ambiguous amino acids, 2,762 sequences were extracted. The HXB2 Env sequence was used as a reference of amino acid position (with number 1 corresponding to the first Met residue).The combined set was realigned using the software MUSCLE (version 3.5) (Edgar R C, 2004) with default parameters.

The following four datasets were compiled based on the alignment of the aforementioned 2762 sequences. For inference of the signature patterns for Env sequences sampled in Southeast Asia, we obtained an aligned set of 312 Southeast Asian and 2,450 non-Southeast Asian Env sequences by extracting sequences with labels of “TH,” “CN,” “MY,” or “VN” in the sequence headers. For inference of the signature patterns for Env sequences sampled in China, we obtained an aligned set of 162 Chinese and 2,600 non-Chinese Env sequences from the above overall alignment by extracting the sequences with the label of “CN” in the sequence headers. For the inference of signature patterns for Env sequences sampled in Africa, we obtained an aligned set of African Env sequences by extracting the sequences with labels of African countries in the sequence headers. African samples were separated into five regions as follows:

1. Southern (“AO”,“ZM”,“ZW”,“BW”,“ZA”,“NA”);
2. Central (“CF”,“CM”,“CG”,“GA”,“AO”).
3. Western (“ML”,“GH”,“NE”,“NG”,“SN”,“GM”,“BJ”).
4. Eastern (“ET”,“UG”,“KE”,“SO”,“TZ”,“RW”).
5. Northern (“EG”,“SD”,“LY”,“MA”,“TN”).

For the inference of signature patterns of Env sequences sampled in Africa, we obtained an overall aligned set of 1,103 African and 1,659 non-African Env sequences, whereas for the inference of signature patterns for Env sequences sampled in Southern Africa alone, we selected only the Southern African sequence described above (“AO”,”ZM”,”ZW”,”BW”,”ZA”,”NA”), which gave an aligned set of 599 Southern African and 2,163 non-Southern African Env sequences.

Extraction of all labels and the manipulation of characters were performed using the R scripts (Supplementary material 1). The supplementary materials and the four alignment files are available on the website of Virologica Sinica: http://www.virosin.org.

Rule inference

To deduce the signature patterns of the four datasets, we used JRip software (Witten I H, et al., 2011) in RWeka (Hornik K, et al., 2009), which can be used in the R environment (Gentleman R C, et al., 2004; Hornik K, et al., 2009) (http://cran.r-project.org). JRip implements RIPPER, which is an incremental machine learning method. In addition, the rule sets can be inferred directly from the training datasets, thus this method is suitable for the fast inference of rules from large datasets. Further association studies and plotting were performed in the R environment.

Assessment of signature pattern inference

To certify the inference of signature patterns, we tested the classification assessment of the signature patterns. We assessed in detail the performance of signature patterns in the classification of Env sequences of Southeast Asian or non-Southeast Asian samples. We performed a full ‘leave-one-out’ classification run with the same set of 2,762 Env sequences used above; each of the sequences was omitted once from the training data, and a set of signature patterns was learned by RIPPER from the remaining 2,761 sequences and their class labels as described above. This was followed by the classification of the remaining one sequence as either Southeast Asian or non-Southeast Asian, based on this set of signature patterns. Comparison of the 2,761 predicted and true class labels allowed for an assessment of the prediction performance. The same procedure was used for assessment of classification of the other three datasets.

Entropy calculation for pattern positions

In an attempt to explain the positions captured in the pattern inferences from the information theory, R-package bio3d (Grant B J, et al., 2006) was used to manipulate and analyze sequences. Using the “entropy” function, we could compute Shannon entropies S_j for alignment position j based on a 22-letter alphabet, including the conventional amino acid, the gap symbol “-,” and “X” (this letter was last not used here), according to the following formula:

$$S_j = -\log_2 22 \times \sum_{i=1}^{22} p_{ij} \log_{22} p_{ij}$$

with the relative frequency P_{ij} of letter i at alignment position j.

Phylogenetic analysis for pattern positions

Owing to the limitations of phylogenetic analysis, such as computational requirements, we considered in this study only one specific pattern corresponding to subtype B and Thai-B (B′) in Southeast Asian sequences, as this analysis might provide important clues to specific geographic origin corresponding to the Chinese HIV-1 B′ pandemic and help to interpret identified patterns from a phylogenetic viewpoint, which might exclude founder effects.

Making use of the maximum likelihood (ML) method to reconstruct phylogenetic trees, we analyzed a set of 954 global HIV-1 subtype B (B′) Env sequences, with 10 HIV-1 subtype D sequences added as an outgroup. The substitution model we chose was HIVb + I + Gamma, and the heuristic tree searches used Nearest-neighbor interchange (NNI), and branch support estimation used approximate likelihood ratio (aLTR). After obtaining the phylogenetic tree, the association of amino acid pattern with the phylogenetic clustering branches was plotted with the R-package ape (Paradis E, et al., 2004).

RESULTS

Rule inference for Southeast Asian HIV-1 Env sequences

Inspired by our previous findings of Chinese HIV-1 genome signature patterns (Wang Y, et al., 2013), and based on a large body of epidemiological evzidence that Chinese HIV-1 sequences have a close phylogenetic relationship with Southeast Asian sequences, we tested the signature patterns for Southeast Asian Env sequences.

After compiling the first aligned dataset, which comprised 312 Southeast Asian and 2,450 non-Southeast Asian Env sequences (label strategy and methods are shown in Materials and methods) and applying rule inference, we obtained the following 7 rules and "x" reprents the amino acid position.

1. (x219 = T) and (x722 = H) => SE=TRUE (185.0/28.0)
2. (x108 = V) and (x725 = G) and (x63 = T) => SE =TRUE (64.0/8.0)
3. (x553 = R) and (x190 = S) => SE=TRUE (44.0/4.0)
4. (x375 = H) and (x148 = G) and (x820 = I) => SE =TRUE (20.0/0.0)
5. (x219 = T) and (x108 = V) and (x742 = K) => SE =TRUE (12.0/4.0)
6. (x746 = T) and (x317 = L) and (x698 = I) => SE =TRUE (6.0/0.0)
7. => SE=FALSE (2431.0/25.0)

The first rule translates as: "Env sequences that have both a T at amino acid position 219 and a H at position 722 could be considered as Southeast Asian sequences." Sites 219 and 722 were numbered according to the HXB2 reference sequence. This first rule covered 185 sequences, with 28 false positives (that is, sequences not from Southeast Asia). The other rules can be interpreted analogously. The seventh rule "= > Southeast Asian = FALSE" means that if none of the previous six rules has been found in a sequence, it is a non-Southeast Asian sequence. The false-negative rate with this rule is about 1%. The total prediction accuracy for the combination of all 7 rules in distinguishing Southeast Asian from non-Southeast Asian sequences was 97.5%.

Statistical errors for the classification of signature patterns

To address the statistical errors for the classification of signature patterns, we tested the classification performance of signature pattern inference. We performed a full 'leave-one-out' classification as described above, run with the same set of 2,762 Env sequences; each of the sequences was omitted once from the training data, and a set of signature patterns was learned by RIPPER from the remaining 2,761 sequences and their class labels. This was followed by the classification of the omitted sequence in this set of signature patterns as being Southeast Asian or non-Southeast Asian. Comparison of the 2,761 predicted and true class labels allowed assessment of the prediction performance. The receiver operating characteristic (ROC) curve and area under the curve (AUC) indicated good performance of the signature patterns mentioned above (Supplementary material 2).

Analysis of signature patterns found in Southeast Asian HIV-1 sequences

In general, each individual rule covered two or three non-adjacent positions in the above set of patterns. Of the 1,157 alignment sites, only 14 sites were found to occur within the whole rule set made up of the 7 rules. In addition, some of the 14 sites occurred frequently within the whole rule set.

In an attempt to investigate this phenomenon further, we computed the sequence entropy for all alignment sites and plotted it against the frequency of occurrence in the above 6 rules rather than rule 7 (Fig. 1). The figure showed that the most frequently occurring sites within these rules

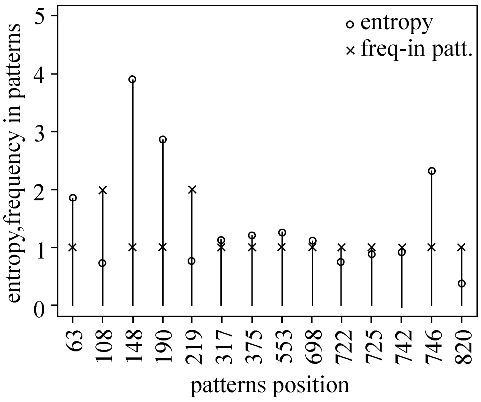

Fig. 1 Sequence entropy and frequency of alignment sites occurring in Southeast Asia signature patterns. The 14 different aligned positions (according to the HXB2 numbering) deduced by the rule set from Southeast Asian HIV-1 Env sequences are listed on the X-axis. Shannon entropy readings are shown as circles and position frequency in patterns is shown as crosses.

tended to be those with higher entropy. However, there were some striking outliers, such as alignment sites 108, 219, and 820, which occurred more frequently but had lower entropy, suggesting that these sites are less variable but more informative in Southeast Asian sequences.

Inferred rules are associated with HIV-1 subtypes

Generally, the classification of HIV into groups and subtypes is based on the variation in genetic distance in HIV sequences. Thus, the HIV-1 subtype itself can be considered as being composed of HIV-1 strains with distinct signature patterns. As for the sequences from Southeast Asia, we would also anticipate that the inferred rules might at least suggest subtype features. We therefore investigated whether such rules existed for the Southeast Asia isolates.

As shown in Fig. 2, some deduced rules were subtype-specific. For example, we found that patterns 1 and 4 were almost exclusively associated with CRF01_AE, whereas pattern 2 reflected features of subtypes C, CRF07_BC, CRF08_BC, and BC recombinants (the latter three recmbinants are chimeric with subtype C in Env), and pattern 3 was characteristic of subtypes B and 01B.

Inferred rules are characteristic of geographic sampling

After investigating associations between rules and subtypes, we next considered the associations between rules and geographical variations. The specific sites 108 and 219, which occurred frequently in all the aforementioned 7 patterns (Fig. 3), are not characteristic of subtype but rather of their origin, that is, Southeast Asia.

Support by phylogenetic analysis for the position combination (x553 = R) and (x190 = S)

To explain these patterns from a phylogenetic viewpoint, we compared the position combination that frequently occurred in patterns, (x553 = R) and (x190 = S), with the evolutionary relationship of this amino acid combination. Again, because the phylogenetic analysis was limited by computational capacity, we only considered the global subtype B/B′ linkage, which is the major pandemic subtype

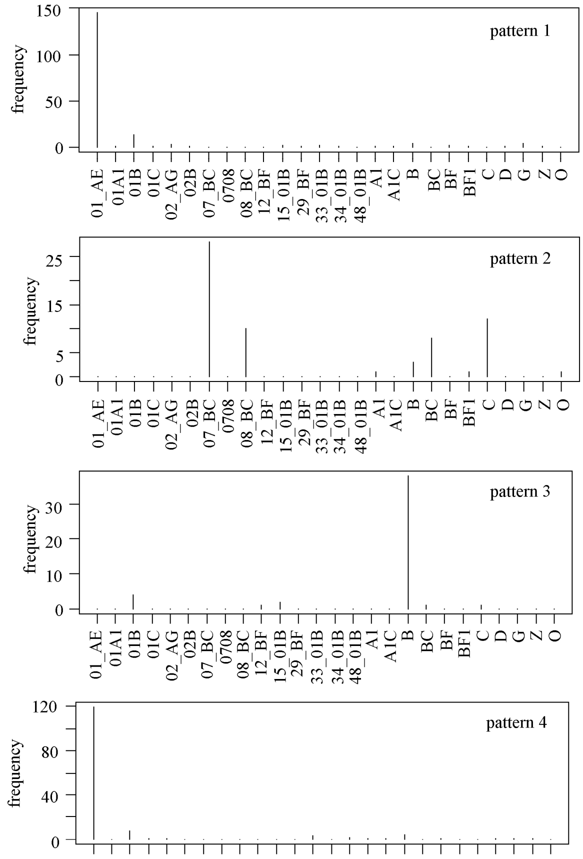

Fig. 2 Signature patterns of Southeast Asian HIV-1 Env sequences and subtypes. The frequency of subtypes is shown for the first four patterns inferred from Southeast Asian HIV-1 Env sequences.

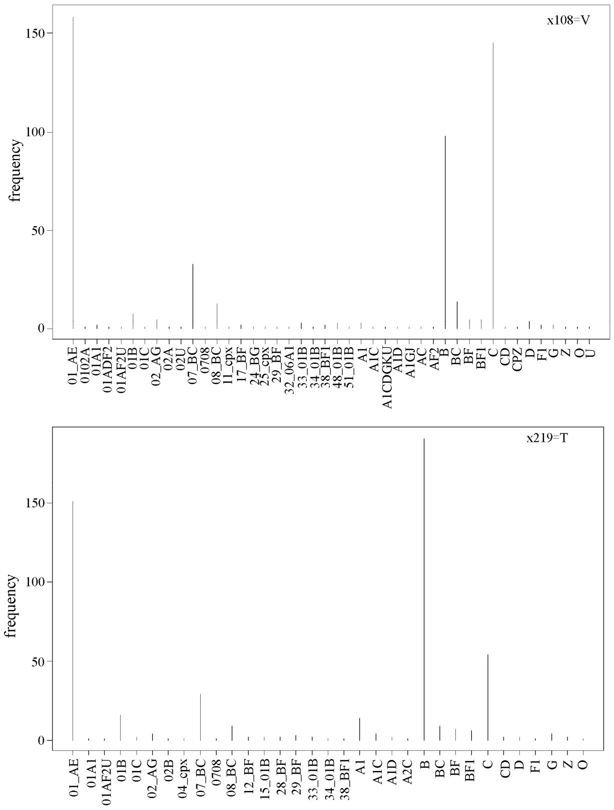

Fig. 3. Relationship of signature patterns of positions 108 and 219 from Southeast Asian HIV-1 sequences with subtypes. The frequency of subtypes is shown for positions 108 and 219 inferred from Southeast Asian HIV-1 Env sequences. These two specific positions, which occur frequently in the whole rule set, are characteristic of having Southeast Asian origin, rather than being subtype-specific.

both in Southeast Asia (B′) and worldwide (B). To this end, we constructed an ML phylogenetic tree for a set of 954 subtype B/B′ amino acid sequences of HIV-1 Env, with 10 subtype D sequences as an outgroup (Fig. 4). Most of the Southeast Asian subtype B sequences were found to lie in a separate cluster distinct from the pandemic global subtype B sequences (Fig. 4, top left: green cluster). Sequences with R553 or S190 only (not the combination of both) were distributed throughout the whole tree (Fig. 4, bottom left and top right: blue and red branches), whereas sequences with both R553 and S190 together were found to lie in similar branches with the Southeast Asian cluster (Fig. 4, bottom right: yellow cluster). In general, this first rule of (x553 = R) and (x190 = S) is consistent with phylogenetic clustering of the Southeast Asian or subtype B/B′ clusters.

Rule inference for Chinese HIV-1 Env sequences

Although we previously completed the rule inference of Chinese HIV-1 genome signature patterns in a combined dataset of 1,047 Chinese and 1,288 non-Chinese sequences (Wang Y, et al., 2013), here we compiled a second aligned dataset, which comprised 162 Chinese and 2,600 Non-Chinese Env sequences (the whole alignment was the same as that for the other three datasets). From this, we obtained 11 rules (Supplementary material 3).

In general, two to five sequence sites were included in each separate rule, and almost all of them were not close to each other. In addition, some sites appeared many times in the whole pattern, such as sites 108 and 219. The most interesting site was site 108, which occurred in five patterns, suggesting a key role for this site in all the Chinese sequences. Although the false-positive rate of the separate pattern was relatively high, the overall false- negative rate was only 0.78%. The whole rule set can be used to distinguish Chinese-specific sequences, with an overall classification accuracy of 98.3%.

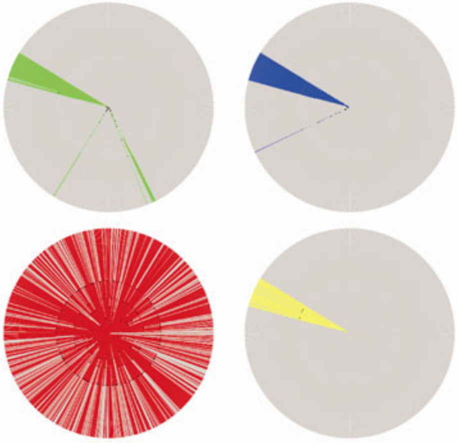

Fig. 4. Maximum likelihood phylogenetic tree of global HIV-1 subtype B/B′ Env sequences. The graph is colored for geographic and amino acid features. Top left: green indicates sequences from Southeast Asia; bottom left: red indicates with sequences with x553 = R; top right: blue indicates sequences with x190 = S; bottom right: yellow indicates sequences carrying the signature pattern (x553 = R) and (x190 = S).

Rule inference for African HIV-1 Env sequences

There are multiple hints of HIV in Africa. Firstly, West and Central Africa were the sites of origin of HIV, which evolved through cross-species transmission of simian immunodeficiency virus in other primates to humans. Secondly, sub-Saharan Africa is the most severely affected region for HIV infection, with a rate of 4.9% in the population, and it accounts for 69% of HIV infections worldwide. Moreover, as outlined in the introduction, the geographical distribution of subtypes and CRF is complex.

Thus, we expected that extending our analysis to the whole African sequences would reveal different patterns specific to these sequences. After compiling the third aligned dataset, which comprised 1,103 African and 1,659 non-African Env sequences and applying rule inference, we obtained 13 rules (Supplementary material 4). We calculated that the combination of all 13 rules would give a total prediction accuracy of 88.4% in distinguishing African from non-African sequences, with a predicted false-negative rate of 7%.

Analysis of signature patterns found in African HIV-1 sequences

Compared to the relatively homologous HIV-1 diversity in Southeast Asia, the highest diversity of whole African HIV-1 sequences may be one of the reseans of less prediction efficiency.

Nevertheless, taking into consideration of 1,167 alignment sites, these rules were still relatively effective at classification. As before, we found that some sites appeared many times in the whole patterns, such as sites 219 (5 times), 315 (4 times), and 720 (3 times). Fig. 5 shows the analysis of the first two patterns specifically. Some deduced rules seemed to be subtype-specific, whereas others were not. For example, rule 1 was found to be associated almost exclusively with subtype C, whereas rule 2 captured features of subtypes C, D, A1D, and CRF02_AG.

Rule inference for Southern Africa HIV-1 Env sequences

The main focus of HIV research is Southern Africa, which has the highest HIV infection rates and most severe HIV pandemics. However, the genetic diversity of HIV in Southern Africa is the lowest.

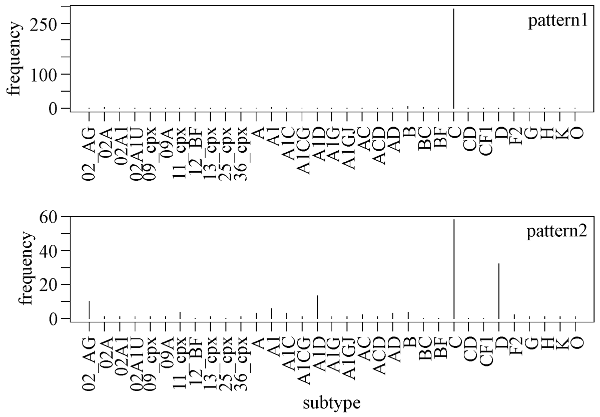

Fig. 5. Signature patterns of African HIV-1 sequences and subtypes. The X-axis shows all the subtypes in the compiling dataset of African origin. The first two patterns were the most representative, and thus subtype frequencies were compared with these patterns.

After compiling the fourth aligned dataset which comprised 599 Southern African and 2163 Non-Southern African Env sequences and applying rule inference, we obtained 6 rules (Supplementary material 5).

The whole signature patterns were extremely succinct, covering 10 different and adjacent sequence sites and a total of 1,167 alignment sites. Although the false-positive rate for separate patterns was relatively high, the overall false-negative rate for rule 6 was 0.35%, and the combination of the six rules (rule set) was able to separate Southern African from non-Southern African sequences with a prediction accuracy of 91.2%. The whole deduced five signature patterns of Southern African HIV-1 sequences were exclusively specific with subtype C. Besides, the informative site 21 was dominant in the whole rule set, suggesting its key role in all the sequences from Southern African. When linking the specific site with the subtypes in Southern African, we also find an extremely dominant pattern with subtype C.

Extensive study on signature patterns for Env sequences of different subtypes sampled worldwide

Since many of the signature patterns inferred above were associated exclusively with specific subtypes and CRFs, we further investigated geographical and evolutionary associations for different subtypes worldwide.

To this end, three sets of patterns were inferred from global subtype C, B and all possible recombination forms chimeric with CRF01_AE Env sequences, including the rules obtained from inference of subtype C Env sequences worldwide, the rules obtained from inference for subtype B Env sequences worldwide and the rules obtained from inference for all possible recombination forms chimeric with CRF01_AE Env sequences worldwide (Supplementary material 6-8).

Comparative analysis of the above three sets of patterns gave an indication of subtypes. As Fig. 6 shows, the signature patterns of HIV-1 global subtype C sequences were extremely common in Zambia (southern Africa), which may also hint at the origin of HIV-1 subtypes and the necessity for paying special attention to this area. Somewhat differently, Fig. 7 shows that signature patterns of HIV-1 global subtype B sequences were associated with some pandemic countries, such as United States, Brazil, Great Britain, Cyprus, and Japan. We also found that site 21 is present in all three rule sets implying some fundamental importance for this site. In detail, there are distinct patterns at single position 21 with glycine, leucine and isoleucine corresponding to subtype C, B and all possible recombination forms chimeric with CRF01_AE, also dominating the above regions separately.

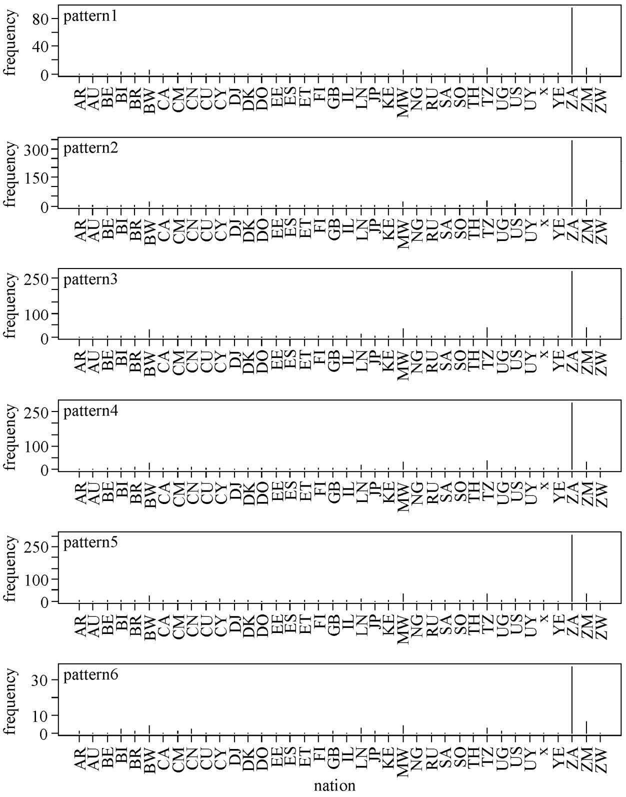

Fig. 6. Signature patterns of HIV-1 global subtype C sequences and countries. The X-axis shows all the countries from which HIV-1 subtype C sequences were sampled. All six patterns are shown, with their frequency distribution in the sampled countries.

DISCUSSION

Scopes of signature pattern analysis

Normally, to infer sequence patterns linked with specific classifications (such as a host group) correctly, it is necessary that the sequences or sequence sites are variable, and that a sufficient number of sequences are available for

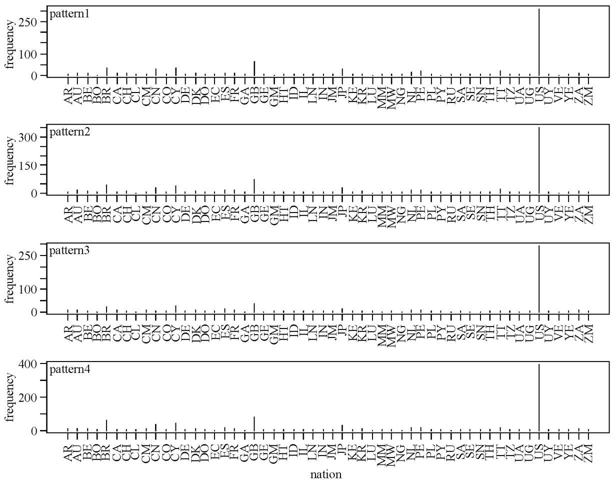

Fig. 7. Signature patterns of HIV-1 global subtype B sequences and countries. The X-axis shows all the countries from which HIV-1 subtype B sequences were sampled. All five patterns are shown, with their frequency distribution in the sampled countries.

statistical analysis. The simple method of rule inference can be used to deduce the patterns of global HIV-1 lineages. However, the effectiveness of this approach varies somewhat among lineages.

Geographically, Southern African HIV-1 sequences, which have the lowest diversity performed best in signature pattern inference, whereas for the whole African sequences, the accuracy of prediction was lower. Compared with the results for rule inference in Chinese and Southeast Asian sequences, these two geographic samplings had no distinct effect on the overall prediction accuracy of signature pattern inference.

We observed that signature patterns of subtype C had extremely high representation in certain of the sampled countries, which may hint at the importance of this region in the early stages of the epidemic. However, it is possible that the sampling bias may have influenced this result. By contrast, the signature patterns of subtype B sequences were associated with different countries. We found that site 21 had clearly distinct patterns among all three subtypes of the global HIV-1 sequence.

When we analyzed the effect of other genome regions (Gag, Pol) on pattern inference, we also found qualitatively similar results (unpublished data).

Mechanisms of these signature patterns

Because there are no specially designed knowledge-mining algorithms, the results are difficult to explain and are unable to meet the requirements of biological research. Thus, appropriate analytical methods are needed to further explain the results of data mining. For example, we can refer to particular features of aligned sequences (WebLogo) (Crooks G E, et al., 2004), immune epitope prediction methods (NetMHC) (Lundegaard C, et al., 2008), mutational modeling methods (pyMOL) (Delano W L, et al., 2004), information theory, and related methods (direct coupling analysis) (Morcos F, et al.) to explain the significance of global HIV-1 sequence patterns from the viewpoints of evolutionary conservation, immune escape, structural stability, and physical contact separately. The preliminary analyses are shown in Fig. 8 and 9.

Application of these signature patterns

The ideal solution for controlling the HIV infection situation is to develop vaccines; however, the diversity of HIV challenges the development of such vaccines.

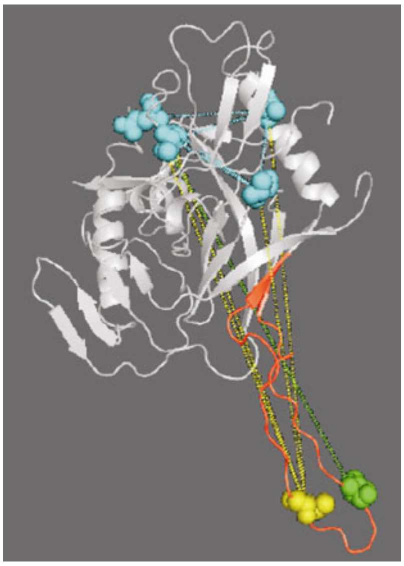

Fig. 8. The deduced physical contacts of the sites found in the Chinese sequence patterns. The background structure shown is gp120, with the V3 region marked in red. The two residues of signature pattern 1 are marked by a green sphere (residue 309) and a yellow sphere (residue 317). The lines connecting these indicate high-DI pairs. Those high-DI amino acid pairs that include residue 309 or residue 317 are shown as blue spheres, and are located in the V2, C2, and V4 regions.

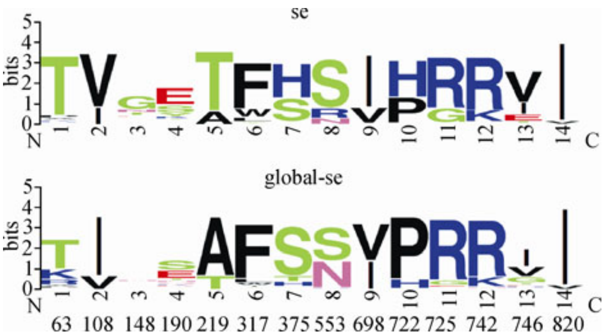

Fig. 9 Results from two datasets: (a) dataset of Southeast Asian sequences and (b) dataset of global sequences covering only the pattern positions inferred from the previous six patterns. Program used was WebLogo version 3.3 (accessed June 3, 2013).

Normally, vaccine strains can be designed either using sequences from contemporary strains (strains within the region), ancestral sequences, artificial consensus sequences, or central sequences in the phylogenetic tree. So far, all Phase III vaccines have been based on the original strain sequence, and as it is considered similar to the contemporary epidemic strain, cross-linking reaction might be induced increasingly. In addition, vaccine design has also focused on the HIV conserved genome region because on the one hand, these regions are more easily recognized by cross-reacting T-cells, while on the other hand, mutations in conserved regions will influence virus fitness. Another development of vaccines is to increase coverage of the epitope so as to increase T-cell response (Fauci A S, et al., 2008; Karlsson Hedestam G B, et al., 2008; Tebit D M, et al., 2011; Walker B D, et al., 2008; Yang O O, 2009). Therefore, our findings should be considered further in the context of vaccine development. For example, patterns characteristic of geographic areas may show a cross- reactivity effect, in contrast to patterns associated with subtype. Sites 108 and 219, which are conserved but informative in the patterns, in the detection of entropy, and in the evolutionary relationship, should also be considered during vaccine design.

Acknowledgements

We gratefully acknowledge the funding by the Chinese Key National Science and Technology Program in the 12th Five-Year Period, grant 2012ZX10001006-002;

And the Deutsche Forschungsgemeinschaft (http://www.dfg.de), grant TRR60/A6; the University of Duisburg-Essen (http://www.uni-due.de)

Author contributions

Yan Wang: Performed the experiments and wrote the article
Reda Rawi: Paticipated in a portion of experiments
Daniel Hoffmann: Designed the project
Binlian Sun: Designed the project and revised the article
Rongge Yang: Designed the project and revised the article

Supplementary materials:

The supplementary materials and the four alignment files are available on the website of Virologica Sinica: http://www.virosin.org.

References

Avenue M, Hill M, Cohen W W, Of C, and Pruning R. 1994. **Fast E ective Rule Induction 2 Previous work 1 in introduction**.

Bello G, Eyer-Silva W a, Couto-Fernandez J C, Guimarães M L, Chequer-Fernandez S L, Teixeira S L M, and Morgado M G. 2007. **Demographic history of HIV-1 subtypes B and F in Brazil**. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases, 7: 263-270.

Blair C, and Murphy R W. 2011. **Recent trends in molecular phylogenetic analysis: where to next?** The Journal of heredity, 102: 130-138.

Buonaguro L, Tagliamonte M, Tornesello M L, and Buonaguro F M. 2007. **Genetic and phylogenetic evolution of HIV-1 in a low subtype heterogeneity epidemic: the Italian example**. Retrovirology, 4: 34-34.

Butler I F, Pandrea I, Marx P a, and Apetrei C. 2007. **HIV genetic diversity: biological and public health consequences**. Current HIV research, 5: 23-45.

Cai Y-D, Lu L, Chen L, and He J-F. 2010. **Predicting subcellular location of proteins using integrated-algorithm method**. Molecular diversity, 14: 551-558.

Crooks G E, Hon G, Chandonia J-m, and Brenner S E. 2004. **WebLogo : A Sequence Logo Generator**. 1188-1190.

Delano W L, and Ph D. 2004. **PyMOL User ' s Guide written by**.

Delatorre E O, and Bello G. 2012. **Phylodynamics of HIV-1 subtype C epidemic in east Africa**. PloS one, 7: e41904-e41904.

Dybowski J N, Riemenschneider M, Hauke S, Pyka M, Verheyen J, Hoffmann D, and Heider D. 2011. **Improved Bevirimat resistance prediction by combination of structural and sequence-based classifiers**. BioData mining, 4: 26-26.

Edgar R C. 2004. **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. Nucleic acids research, 32: 1792-1797.

Fauci A S, Johnston M I, Dieffenbach C W, Burton D R, Hammer S M, Hoxie J a, Martin M, Overbaugh J, Watkins D I, Mahmoud A, and Greene W C. 2008. **HIV vaccine research: the way forward**. Science (New York, N.Y.), 321: 530-532.

Fryer H R, and McLean A R. 2011. **Modelling the spread of HIV immune escape mutants in a vaccinated population**. PLoS computational biology, 7: e1002289-e1002289.

Gentleman R C, Carey V J, Bates D M, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, and Gentry J. 2004. **Bioconductor: open software development for computational biology and bioinformatics**. Genome biology, 5: R80

Gilbert M T P, Rambaut A, Wlasiuk G, Spira T J, Pitchenik A E, and Worobey M. 2007. **The emergence of HIV/AIDS in the Americas and beyond**. Proceedings of the National Academy of Sciences of the United States of America, 104: 18566-18570.

Grant B J, Rodrigues A P C, ElSawy K M, McCammon J A, and Caves L S D. 2006. **Bio3d: an R package for the comparative analysis of protein structures**. Bioinformatics, 22: 2695-2696.

Hemelaar J. 2012. **The origin and diversity of the HIV-1 pandemic**. Trends in Molecular Medicine, 18: 182-192.

Hornik K, Buchta C, and Zeileis A. 2009. **Open-source machine learning: R meets Weka**. Computational Statistics, 24: 225–232.

Junqueira D M, de Medeiros R M, Matte M C C, Araújo L A L, Chies J A B, Ashton-Prolla P, and Almeida S E D M. 2011. **Reviewing the history of HIV-1: spread of subtype B in the Americas**. PloS one, 6: e27489-e27489.

Kallings L O. 2008. **The first postmodern pandemic: 25 years of HIV/ AIDS**. Journal of internal medicine, 263: 218-243.

Karlsson Hedestam G B, Fouchier R a M, Phogat S, Burton D R, Sodroski J, and Wyatt R T. 2008. **The challenges of eliciting neutralizing antibodies to HIV-1 and to influenza virus**. Nature reviews. Microbiology, 6: 143-155.

Li Y, Uenishi R, Hase S, Liao H, Li X-J, Tsuchiura T, Tee K K, Pybus O G, and Takebe Y. 2010. **Explosive HIV-1 subtype B' epidemics in**

Asia driven by geographic and risk group founder events. Virology, 402: 223-227.

Liao H, Tee K K, Hase S, Uenishi R, Li X-J, Kusagawa S, Thang P H, Hien N T, Pybus O G, and Takebe Y. 2009. **Phylodynamic analysis of the dissemination of HIV-1 CRF01_AE in Vietnam**. Virology, 391: 51-56.

Lihana R W. 2012. **Update on HIV-1 Diversity in Africa : A Decade in Review**. 83-100.

Liu J, and Zhang C. 2011. **Phylogeographic analyses reveal a crucial role of Xinjiang in HIV-1 CRF07_BC and HCV 3a transmissions in Asia**. PloS one, 6: e23347-e23347.

Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, and Nielsen M. 2008. **NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11**. Nucleic Acids Research, 36: W509-W512.

Lynch R M, Shen T, Gnanakaran S, and Derdeyn C a. 2009. **Appreciating HIV type 1 diversity: subtype differences in Env**. AIDS research and human retroviruses, 25: 237-248.

Masciotra S, Livellara B, Belloso W, Clara L, Tanuri a, Ramos a C, Baggs J, Lal R, and Pieniazek D. 2000. **Evidence of a high frequency of HIV-1 subtype F infections in a heterosexual population in Buenos Aires, Argentina**. AIDS research and human retroviruses, 16: 1007-1014.

Meng Z, Xin R, Zhong P, Zhang C, Abubakar Y F, Li J, Liu W, Zhang X, and Xu J. 2012. **A new migration map of HIV-1 CRF07_BC in China: analysis of sequences from 12 provinces over a decade**. PloS one, 7: e52373-e52373.

Moran D, and Jordaan J a. 2007. **HIV/AIDS in Russia: determinants of regional prevalence**. International journal of health geographics, 6: 22-22.

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks D S, Sander C, Zecchina R, Onuchic J N, Hwa T, and Weigt M. **Direct-coupling analysis of residue coevolution captures native contacts across many protein families**. Proceedings of the National Academy of Sciences of the United States of America, 108: E1293-E1301.

Morris C N, and Ferguson a G. 2006. **Estimation of the sexual transmission of HIV in Kenya and Uganda on the trans-Africa highway: the continuing role for prevention in high risk groups**. Sexually transmitted infections, 82: 368-371.

Njai H F, Gali Y, Vanham G, Clybergh C, Jennes W, Vidal N, Butel C, Mpoudi-Ngolle E, Peeters M, and Ariën K K. 2006. **The predominance of Human Immunodeficiency Virus type 1 (HIV-1) circulating recombinant form 02 (CRF02_AG) in West Central Africa may be related to its replicative fitness**. Retrovirology, 3: 40-40.

Paradis E, Claude J, and Strimmer K. 2004. **APE: Analyses of Phylogenetics and Evolution in R language**. Bioinformatics, 20: 289-290.

Paraschiv S, Otelea D, Batan I, Baicus C, Magiorkinis G, and Paraskevis D. 2012. **Molecular typing of the recently expanding subtype B HIV-1 epidemic in Romania: evidence for local spread among MSMs in Bucharest area**. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases, 12: 1052-1057.

Paraskevis D, Pybus O, Magiorkinis G, Hatzakis A, Wensing A M, van de Vijver D a, Albert J, Angarano G, Asjö B, Balotta C, Boeri E, Camacho R, Chaix M-L, Coughlan S, Costagliola D, De Luca A, de Mendoza C, Derdelinckx I, Grossman Z, Hamouda O, Hoepelman I, Horban A, Korn K, Kücherer C, Leitner T, Loveday C, Macrae E, Maljkovic-Berry I, Meyer L, Nielsen C, Op de Coul E L, Ormaasen V, Perrin L, Puchhammer-Stöckl E, Ruiz L, Salminen M O, Schmit J-C, Schuurman R, Soriano V, Stanczak J, Stanojevic M, Struck D, Van Laethem K, Violin M, Yerly S, Zazzi M, Boucher C a, and Vandamme A-M. 2009. **Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach**. Retrovirology, 6: 49-49.

Pérez L, Thomson M M, Bleda M J, Aragonés C, González Z, Pérez J, Sierra M, Casado G, Delgado E, and Nájera R. 2006. **HIV Type 1 molecular epidemiology in cuba: high genetic diversity, frequent mosaicism, and recent expansion of BG intersubtype recombinant forms**. AIDS research and human retroviruses, 22: 724-733.

Pollakis G, Abebe A, Kliphuis A, De Wit T F R, Fisseha B, Tegbaru B, Tesfaye G, Negassa H, Mengistu Y, Fontanet A L, Cornelissen M, and Goudsmit J. 2003. **Recombination of HIV type 1C (C'/C") in Ethiopia: possible link of EthHIV-1C' to subtype C sequences from the high-prevalence epidemics in India and Southern Africa**. AIDS research and human retroviruses, 19: 999-1008.

Poonpiriya V, Sungkanuparph S, Leechanachai P, Pasomsub E, Watitpun C, Chunhakan S, and Chantratita W. 2008. **A study of seven rule-based algorithms for the interpretation of HIV-1 genotypic resistance data in Thailand**. Journal of virological methods, 151: 79-86.

Restif O. 2009. **Evolutionary epidemiology 20 years on: challenges and prospects**. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases, 9: 108-123.

Sharp P M, and Hahn B H. 2011. **Origins of HIV and the AIDS Pandemic**. 1-22.

Sharp P M, and Hahn B H. 2011. **Origins of HIV and the AIDS pandemic**. Cold Spring Harbor perspectives in medicine, 1: a006841-a006841.

Shen C, Craigo J, Ding M, Chen Y, and Gupta P. 2011. **Origin and dynamics of HIV-1 subtype C infection in India**. PloS one, 6: e25956-e25956.

Sierra M, Thomson M M, Posada D, Pérez L, Aragonés C, González Z, Pérez J, Casado G, and Nájera R. 2007. **Identification of 3 phylogenetically related HIV-1 BG intersubtype circulating recombinant forms in Cuba**. Journal of acquired immune deficiency syndromes (1999), 45: 151-160.

Silveira J, Santos A F, Martínez A M B, Góes L R, Mendoza-Sassi R, Muniz C P, Tupinambás U, Soares M a, and Greco D B. 2012. **Heterosexual transmission of human immunodeficiency virus type 1 subtype C in southern Brazil**. Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology, 54: 36-41.

Spira S. 2003. **Impact of clade diversity on HIV-1 virulence, antiretroviral drug sensitivity and drug resistance**. Journal of Antimicrobial Chemotherapy, 51: 229-240.

Taylor B S, and Hammer S M. 2008. **The challenge of HIV-1 subtype diversity**. The New England journal of medicine, 359: 1965-1966.

Tebit D M, and Arts E J. 2011. **Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease**. The Lancet Infectious Diseases, 11: 45-56.

Villanova F E. 2010. **Diversity of HIV-1 Subtype B : Implications to the Origin of BF Recombinants**. 5: 1-9.

Walker B D, and Burton D R. 2008. **Toward an AIDS vaccine**. Science (New York, N.Y.), 320: 760-764.

Walker P R, Pybus O G, Rambaut A, and Holmes E C. 2005. **Comparative population dynamics of HIV-1 subtypes B and C:**

subtype-specific differences in patterns of epidemic growth. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases, 5: 199-208.

Wang Y, Rawi R, Wilms C, Heider D, Yang R, and Hoffmann D. 2013. **A small set of succinct signature patterns distinguishes Chinese and non-Chinese HIV-1 genomes**. PloS one, 8: e58804-e58804.

Witten I H, Frank E, and Hall M A. 2011. **Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques**. Elsevier

Worobey M, Gemmel M, Teuwen D E, Haselkorn T, Kunstman K, Bunce M, Muyembe J-j, Kabongo J-m M, Kalengayi R M, Van Marck E, Gilbert M T P, Wolinsky S M, Kalengayi M, and Marck E V. 2008. **Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960**. Nature, 455: 661-664.

Yang O O. 2009. **Candidate vaccine sequences to represent intra- and inter-clade HIV-1 variation**. PloS one, 4: e7388-e7388.

Zhao Y. 2011. **R and Data Mining: Examples and Case Studies 1**.

Zhu T, Korber B T, Nahmias a J, Hooper E, Sharp P M, and Ho D D. 1998. **An African HIV-1 sequence from 1959 and implications for the origin of the epidemic**. Nature, 391: 594-597.