Identification of different sequence properties between HIV-1 DNA and RNA across subtypes using the *k*-mer-based approach

Janusz Wiśniewski[1], Karol Serwin[2], Miłosz Parczewski[2], Anna Kula-Păcurar[3], Pavel Skums[4], Alexander Kirpich[5], Sergiy Yakovlev[6,7], and Heng-Chang Chen[1,*,§]

[1]Quantitative Virology Research Group, Population Diagnostics Center, Łukasiewicz Research Network – PORT Polish Center for Technology Development, Stablowicka 147, 54-066 Wrocław, Poland

[2]Department of Infectious, Tropical Diseases and Immune Deficiency, Pomeranian Medical University in Szczecin, Szczecin, Poland

[3]Molecular Virology, Małopolska Centre of Biotechnology, Jagiellonian University, Gronostajowa 7A str, 30-387 Kraków, Poland

[4]School of Computing, University of Connecticut, 371 Fairfield Way, Storrs, CT, 06269, USA

[5]Department of Population Health Sciences, School of Public Health, Georgia State University, Atlanta, GA, USA

[6]Institute of Mathematics, Lodz University of Technology, 90-924 Lodz, Poland

[7]Institute of Computer Science and Artificial Intelligence, V.N. Karazin Kharkiv National University, 61022 Kharkiv, Ukraine

[§] Equal contributions

[*] Correspondence: heng-chang.chen@port.lukasiewicz.gov.pl

Abstract

Advanced analytical tools that enable mining of the masked features hidden in intricate datasets and strengthening the biological interpretation of multigenomic outputs hold paramount importance. In this study, we present an updated version of a *k*-mer-based approach, PORT-EK-v2, allowing for a comparison of multiple genomic datasets and identification of over-represented genomic regions, *k*-mers, related to specific organisms. Using PORT-EK-v2, we exemplified that most likely DNA and RNA sequence properties are distinct across HIV-1 subtypes. Furthermore, we showcased that "isolate *k*-mer count" could serve as a default choice in classifying the DNA versus RNA sequence property. Lastly, results based on Markov chain Monte Carlo modeling unveiled a discontinuous nature of the sequence property in terms of *k*-mer frequencies across HIV-1 subtypes. Altogether, we propose that the sequence property (DNA versus RNA) is distinct across HIV-1 subtypes and has a consequential impact on identifying new and emerging subtypes in the future.

Introduction

The current human immunodeficiency type 1 (HIV-1) pandemic is phylogenetically classified into M, N, O, and P four groups[1–3] based on RNA sequences. Among them, group M is the most prevalent worldwide and is responsible for pandemics in humans. This high genomic diversity is one of the main characteristics of the HIV-1 genomes due to its high mutation rate and a lack of proofreading activity of reverse transcriptase[4,5] and viral recombination-prone nature[6–8]. At the epidemiological level, At the molecular level, the apparent sequence diversity of HIV-1 RNA mainly resulting from a lack of proofreading of its replicated RNA genomes[9] and alternative splicing[10,11,12]. HIV-1 RNA is thus crucial for tracking the transmission dynamics of primary HIV-1 infections, monitoring the circulation and evolution of HIV-1 subtypes and variants, and detecting emerging drug-resistance mutations in HIV-1.

In contrast, in cases where the viral load is too low for successful HIV-1 RNA genotyping, HIV-1 DNA may serve as an alternative to identify drug resistance mutations within patients subjected to ART, as evidenced by several studies[13-17]. Nevertheless, the feasibility of the proviral DNA-based HIV resistance testing remains debated, given that this strategy often captures apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC)-mediated G-to-A hypermutation and genetically defective and replication-incompetent viruses[18-20]. Indeed, more than 90% of proviral sequences are genetically defective[21], particularly falling in the region of the HIV *env* gene[22,23], further resulting in a heterogeneity of provirus genomic sequences in reservoir cells. Although diversities between HIV-1 DNA and RNA have been characterized at the molecular and epidemiological level, appropriate HIV-1 subtyping remains an unsolved issue. A genomic sequence property can be decoded at different tiered levels, ranging from nucleotides, sequence motifs, individual genes, to a cluster of genes involved in the same regulatory network. In this work, we investigate the HIV DNA and RNA genomic sequence diversity at the level of enriched *k*-mers, short genomic sequences (i.e., 13 base pairs (bp), 15 bp, and 17 bp), and hypothesize that the difference in the sequence property between HIV-1 DNA and RNA exists and influences the classification of HIV-1 genomic diversity. More comprehensive insights into the difference between them may enhance the accuracy and the granularity of genome surveillance of HIV-1[24,25].

Traditional methods for HIV-1 taxonomies are mainly based on gene-based phylogenetic analysis embeded with alignment-based methods. Individual HIV-1 gene sequences (often from the *env* gene[26] and *gag* region[27]) are aligned with curated subtype reference sequences, followed by a comparison of homologous nucleotide patterns or motifs to create evolutionary trees[28,29]. Such methods can be computationally expensive and time-consuming, given that similarity statistics over a sliding window are computed[30,31]. To conquer this limitation, various *k*-mer-based alignment-free methods, including K_{AMERIS}[32], a multi-label learning algorithm[33], KINN[34], KEVOLVE[35], and a *k*-mer-based eXtreme Gradient Boosting model[36], have been proposed for HIV-1 subtyping and phylogenetic analysis[34,37,38], and HIV-1-resistent drug detection[39]. Overall, the feasibility of *k*-mer-based methods for the classification of HIV-1 subtypes may mitigate this challenge.

In this work, we showcased the enhanced effectiveness and accuracy of the strengthened PORT-EK[40] pipeline, a *k*-mer-based approach, which facilitates the comparison of various multi-genomic datasets from diverse hosts and the detection of over-represented genomic regions (i.e., *k*-mers) linked to specific sequences; this was applied as PORT-EK version 2 (PORT-EK-v2) to pinpoint *k*-mers abundant in HIV-1 DNA or RNA sequences across group M HIV-1 subtypes. PORT-EK-v2 calculates enriched *k*-mer frequencies rather than just detecting the presence of a single genomic sequence between two species; this capability facilitates the recapture of more biologically meaningful patterns embedded in sequence property and architecture. Technically, PORT-EK-v2 is a streamlined and efficient pipeline designed to make *k*-mer frequency enrichment statistics available in a readable table format. Using PORT-EK-v2, we identified distinct HIV-1 enriched DNA and RNA *k*-mer frequencies through the viral genome across subtypes.

Results

An overview of PORT-EK-v2

The critical steps of PORT-EK-v2 include (i) *k*-mers matrix preparation, (ii) *k*-mers filtering and selection of enriched *k*-mers, and (iii) mapping of enriched *k*-mers to a reference genome. The predictive accuracy using *k*-mer-related features was tested independently from the PORT-EK-v2 pipeline. Technical advances in PORT-EK-v2 include (i) the establishment of a streamlined computing pipeline, allowing for reducing computing cost (computational resources requirement) (**Table 1**) and offering minimal computing time (**Table 1**), and leveraging its effectiveness, and (ii) the improvement of mapping algorithms, thereby providing a superior performance of retrieving enriched *k*-mers. Detailed rationale and background of PORT-EK-v2 and its optimized parameterization in comparison to PORT-EK[40] are provided in **online Methods**. The systematic comparison of key parameters between PORT-EK[40] and PORT-EK-v2 is summarized in **Table S1**. In addition, we performed pipeline runs from input HIV-1 RNA sequences and two subsets of randomly bootstrapped HIV-1 RNA sequences of different sizes to retrieve enriched RNA *k*-mers using PORT-EK[40] as a benchmark. We benchmarked PORT-EK-v2 against PORT-EK[40], revealing a remarkable reduction in computing speed and peak memory footprint of the enrichment of *k*-mer frequencies (**Table 1**). The *k*-mer counting speed of PORT-EK-v2 was found to be comparable to that achieved by Jellyfish[41] and Kmer-bd[42,43] (**Table 2**).

Distinct molecular properties unveiled by enriched *k*-mers between HIV-1 DNA and RNA sequences across subtypes

A total of 10,015 HIV-1 DNA and 5,494 RNA sequences were downloaded from the Los Alamos National Laboratory HIV Databases (https://www.hiv.lanl.gov/). After the examination of the quality of genomic sequences, the subset of 10,013 and 5,490 HIV-1 DNA and RNA sequences were subject to PORT-EK-v2 and downstream analyses. Analyzed sequences and corresponding exact IDs are archived at GitHub (see **Methods**). To minimize a bias resulting from the imbalance of the sample size across each subtype, we combined genomes from rare subtypes for which a small number of sequences were retrieved as one group named "rare subtypes" for the following analysis. Eventually, five groups of HIV-1 subtypes—the subtypes A, B, C, D, and the rare subtypes, including subtypes F1, F2, G, H, J, K, and L—were utilized for the downstream analyses. *K*-mers enriched in corresponding groups of HIV-1 subtypes were determined by the average count of individual *k*-mers.

Using PROT-EK-v2, we first observed distinct distributions of enriched DNA and RNA *k*-mers with a length of 13 bp [hereinafter, total enriched DNA *k*-mers (13 bp), **Fig. 1a**; hereinafter, total enriched RNA *k*-mers (13 bp), **Fig. 1c**] and 15 bp [hereinafter, total enriched DNA *k*-mers (15 bp), **Fig. 1b**; hereinafter, total enriched RNA *k*-mers (15 bp), **Fig. 1d**]. A linear distribution of enriched DNA *k*-mers, aligning with subtypes C, A, the rare subtypes, D, and B, reflects a linear similarity of DNA sequence in terms of enriched *k*-mer frequencies across isolates in different subtypes: DNA sequences retrieved from subtypes A, D, and the rare subtypes shared a relatively high sequence similarity compared with sequences retrieved from subtypes B and C (**Fig. 1a** and **1b**). Of note, a small fraction of enriched DNA *k*-mers in subtype B was

observed. In contrast, an equivalent radial distribution of enriched RNA *k*-mers inferred that quantitatively, the RNA sequence diversity of isolates across subtypes A, B, and C is indistinguishable (**Fig. 1c** and **1d**); the similarity of RNA sequences shared between subtypes B and D, as well as subtype D and the rare subtypes. Similar distribution patterns also appeared in enriched DNA *k*-mers (17 bp) (**Extended data Fig. S1a**) and enriched RNA *k*-mers (17 bp) (**Extended data Fig. S1b**). Given their similar patterns, in the following content, we systematically present results using *k*-mers of length 13 bp and 15 bp to verify our findings with one another.

A clear separation across different groups of HIV-1 subtypes was also confirmed by polygenetic trees, in which the distance is computed based on enriched *k*-mer count (**Fig. 1e-1h**). Intriguingly, a small subset appearing in subtypes A and D inferred the presence of a higher diversity in enriched DNA *k*-mer count in these two groups compared with others (**Fig. 1e** and **1f**); whereas an independent subset was detected in the rare subtypes when enriched RNA *k*-mers were processed (**Fig. 1g** and **1h**). No significant difference in the proportion of enriched *k*-mers that are exclusively present in corresponding groups can be detected (**Extended data Fig. S1c**).

We further computed three additional features of enriched *k*-mers: (1) "*k*-mer weight"—representing the composition of each enriched *k*-mer at a nucleotide level, (2) "subtype *k*-mer count"—the sum of individual *k*-mers normalized by different groups of HIV-1 subtypes, and (3) "isolate *k*-mer count"—the appearance of individual *k*-mers across different isolates (**Fig. 1i**). The definition of each feature is detailed in **online Methods**. We observed a modest and significant enrichment of enriched DNA and RNA *k*-mers in subtype C compared with *k*-mers enriched in other groups of HIV-1 subtypes while comparing "*k*-mer weight" across each group (**Extended data Fig. S1d**). No difference in "*k*-mer weight" between *k*-mers in a length of 13 bp and 15 bp, except for a minor enrichment in enriched DNA *k*-mers (13 bp) in subtypes B (enriched DNA *k*-mers: 7.564, enriched RNA *k*-mers: 7.501) and C (enriched DNA *k*-mers: 7.463, enriched RNA *k*-mers: 7.416) compared with enriched RNA *k*-mers (**Extended data Fig. S1e**). Although a bias was observed when different ordinal encoding was assigned to every nucleotide in an enriched *k*-mers (**Extended data Fig. S2a-S2c**), the overall patterns of "*k*-mer weight" across different groups of HIV-1 subtypes were reproducible (**Extended data Fig. S2d-S2f**).

A more abundant "subtype *k*-mer count" appeared in subtype D, whereas a relatively few counts were in subtype B, irrespective of enriched DNA or RNA *k*-mers and their lengths (**Extended data Fig. S1f**). A diverse and varied "subtype *k*-mer count" between enriched DNA and RNA *k*-mers across different groups of HIV-1 subtypes was also observed (**Extended data Fig. S1g**).

While computing "isolate *k*-mer count", a superior enrichment of enriched DNA *k*-mers appeared in the group containing the rare subtypes, followed by subtype D and A (**Extended data Fig. S1h**), whereas the majority of enriched RNA *k*-mers accumulated in the HIV-1 subtype D, followed by the rare subtypes; a relatively minor number of *k*-mers were enriched in isolates across the subtypes A, B, and C (**Extended data**

Fig. S1h). A higher "isolate *k*-mer count" in enriched DNA *k*-mers than in enriched RNA *k*-mers was observed in subtypes A, C, and the rare subtypes (**Extended data Fig. S1i**); whereas an opposite pattern was detected in subtypes B and D (**Extended data Fig. S1i**).

To visualize the coverage of enriched DNA and RNA *k*-mers, we summarized the HIV-1 genome by 20 different quantile values (**Fig. 1j**, *k*-mer = 13 bp; **Fig. 1k**, *k*-mer = 15 bp) and observed that a higher enriched *k*-mer coverage appears in subtypes A and D, followed by subtype C and the rare subtypes, whereas the coverage of enriched RNA *k*-mers across subtypes A, C, D, and the rare subtypes was comparable (**Fig. 1j** and **1k**). Identical patterns were observed in the circumstances that the HIV-1 genome was summarized by 10 (**Extended data Fig. S1j**, *k*-mer = 13 bp; **Extended data Fig. S1k**, *k*-mer = 15 bp) and 7 (**Extended data Fig. S1l**, *k*-mer = 13 bp; **Extended data Fig. S1m**, *k*-mer = 15 bp) different quantile values. The overall coverage was illustrated in **Fig. 1l** (*k*-mer = 13 bp) and **1m** (*k*-mer = 15 bp). The coverage of enriched RNA *k*-mers in length of 15 bp in subtype D was slightly higher than in other groups of HIV-1 subtypes. In all scenarios, a low *k*-mer coverage was shown in subtype B. Overall, these observations suggested that most likely the sequence space varies between HIV-1 DNA and RNA sequences and differs across HIV-1 subtypes in the context of genomic sequence composition unveiled by enriched *k*-mers.

Distinct sequence properties between HIV-1 DNA and RNA were revealed at the sequence level rather than at a genic level

Next, we retrieved enriched *k*-mers that are exclusively present in either DNA (hereinafter, unique enriched DNA *k*-mers) or RNA (hereinafter, unique enriched RNA *k*-mers) enriched *k*-mers, separated by distinct groups of HIV-1 subtypes (**Extended data Fig. S3a**). We computed that a range between 15% to 40% of unique enriched DNA *k*-mers and 15% to 39% of unique enriched RNA *k*-mers were present in corresponding groups of HIV-1 subtypes (**Extended data Fig. S3a**). Distinct enrichment patterns of "*k*-mer RMSE" between unique enriched DNA and RNA *k*-mers suggest a strong correlation between unique enriched *k*-mers and distinct sequence properties across groups of HIV-1 subtypes (**Extended data Fig. S3b**).

As we computed "*k*-mer weight", only minor differences across different groups of HIV-1 subtypes were observed (**Extended data Fig. S3c**). Notably, different from the pattern shown in total enriched DNA *k*-mers (**Extended data Fig. S1f**), the increased abundance of unique enriched DNA *k*-mers accumulated in subtype B, highlighting the heterogeneity of subgenomic locations throughout the DNA sequence (unique versus common *k*-mers), as "subtype *k*-mer count" was examined (**Extended data Fig. S3d**). In contrast, a similar pattern between total and unique enriched RNA *k*-mers was observed (**Extended data Fig. S1f** and **Fig. S3d**).

To target unique enriched *k*-mers with significant enrichment, we retrieved those with the fold change in RMSE greater than 0.5 (arbitrary units) (**Fig. 2a**, *k*-mer = 13 bp; **Fig. 2c**, *k*-mer = 15 bp) and 1 (**Fig. 2b**, *k*-mer = 13 bp; **Fig. 2d**, *k*-mer = 15 bp) in a logarithmic scale ($\log_2$). The majority of unique enriched DNA *k*-

mers appeared in subtype A, whereas abundant unique enriched RNA *k*-mers accumulated in subtype D (**Fig. 2e**), suggesting that the enrichment of unique *k*-mers differs between DNA and RNA sequences. Similar patterns were observed in both cutoff thresholds. We also depicted the landscapes of unique enriched DNA and RNA *k*-mers throughout the HIV-1 genome, verifying that distinct loci overlapping unique *k*-mers enriched in DNA and RNA sequences across different groups of HIV-1 subtypes were once again confirmed (**Extended data Fig. S3e**, *k*-mer = 13 bp; **Extended data Fig. S3f**, *k*-mer = 15 bp). A superior frequency of unique enriched *k*-mers overlapped the HIV-1 po*l* gene, followed by the *env* and *gag* genes (**Fig. 2f**, *k*-mer = 13 bp; **Fig. 2g**, *k*-mer = 15 bp). Such a bias was, however, indistinguishable either between DNA and RNA sequences or across different groups of HIV-1 subtypes (**Fig. 2f** and **2g**).

With respect to identical *k*-mers that are enriched in both DNA and RNA sequences, namely common enriched *k*-mers, separated by different groups of HIV-1 subtypes, a positive correlation in "*k*-mer RMSE" (**Extended data Fig. S3g**, *k*-mer = 13 bp, R^2 = 0.63; **Extended data Fig. S3h**, *k*-mer = 15 bp, R^2 = 0.6), "*k*-mer weight" (**Extended data Fig. S3i**, *k*-mer = 13 bp; **Extended data Fig. S3j**, *k*-mer = 15 bp), and "subtype *k*-mer count" (**Extended data Fig. S3g**, *k*-mer = 13 bp, R^2 = 0.89; **Extended data Fig. S3h**, *k*-mer = 15 bp, R^2 = 0.88) were detected. These findings suggest that the sequence property of common enriched *k*-mers is independent of DNA and RNA sequences. No difference in the frequency of enriched unique (**Fig. 2f** and **2g**) and common (**Fig. 2h** and **2i**) *k*-mers that overlap HIV-1 genes was observed.

"Isolate *k*-mer count" enables the classification and prediction of different groups of HIV-1 subtypes

To evaluate the biological importance of five features associated with enriched *k*-mers, we constructed classifiers based on three approaches: logistic regression (**Fig. 3a**), multinomial logistic regression (**Fig. 3b**), and the neural network (**Fig. 3c**). We observed that a model constructed by "isolate *k*-mer count" surpassed others in the prediction of the origin of enriched *k*-mers (DNA versus RNA) (**Fig. 3a**). Notably, "isolate *k*-mer count" enabled the prediction of the likelihood of different groups of HIV-1 subtypes, with diverse prediction power between enriched DNA and RNA *k*-mers across distinct groups of HIV-1 subtypes (**Fig. 3b**). Such distinct predictive performance may imply the involvement of additional determinants (e.g., inherent genetic and biological variations, and sequence recombinations) in addition to the cause of the difference in the readthrough between DNA and RNA sequences. Prediction performance of "isolate *k*-mer count" in the context of HIV-1 subtypes was also confirmed by a neural network-based model (**Fig. 3c**). The values of AUC (**Fig. 3a**), probability (**Fig. 3b**), and accuracy (**Fig. 3c**) are summarized in **Supplementary Tables S10**, **S11**, and **S12**, respectively. All these findings signify the importance of the feature "isolate *k*-mer count". We further computed the Euclidean distance between two isolates based on "isolate *k*-mer count" and observed a clear separation in isolates across subtypes A, B, and C, irrespective of enriched DNA and RNA *k*-mers (**Fig. 3d-3g**). In contrast, isolates in subtypes D and the remaining subtypes were indistinguishable (**Fig. 3d-3g**).

Intrinsic barriers lie between sequence properties across HIV-1 subtypes

We further constructed a pentapartite graph (see **online Methods** for the definition), representing a network consisting of the five groups of HIV-1 subtypes, and used this graph to illustrate spatial interactions between isolates (**Extended data Fig. S4a-S4d**) or enriched *k*-mers (**Extended data Fig. S4f-S4i**) across different groups of HIV-1 subtypes. Overall, *k*-mers that are enriched in the same groups of HIV-1 subtypes tended to connect. In the former scenario, a relatively high assortativity degree was observed in subtypes A, B, and C, whereas a relatively low assortativity degree was present in subtype D and the rare subtypes (**Extended data Fig. S4e**). In the latter scenario, although a strong tendency between enriched *k*-mers in different groups of HIV-1 subtypes was structured, subsets of enriched *k*-mers bridged by enriched *k*-mers assigned to the group of the rare subtypes (**Extended data Fig. S4f-S4i**). A relatively high assortativity degree was observed in subtype B, whereas other subtypes, especially subtype D and the rare subtypes, demonstrated the lowest assortativity degree (**Extended data Fig. S4j**).

Lastly, we applied the Markov chain Monte Carlo (MCMC) method[44] and simulated 10 paths of random walks up to 10,000 steps (**Fig. 4a**) in networks constructed by either "isolate *k*-mer count" (**Fig. 4b-4f**) or "subtype *k*-mer count" (real, **Fig. 4g-4k**; labeling mismatch, **Extended data Fig. S5a-S5e**), respectively. We inspected five different assumptions, in which a random walk initiates in an isolate or an enriched *k*-mer assigned to different groups of HIV-1 subtypes and computed the probability that isolates or enriched *k*-mers in each group of HIV-1 subtypes are involved in a path of a random walk (**Fig. 4a**; see **online Method**).

In a network constructed by "isolate *k*-mer count", a superior subtype-specific correlation between initiation of isolates and the highest probability that isolates categorized into the identical group upon the initiation step in each assumption was observed (**Fig. 4b-4f**), resonating with disconnected graphs (**Extended data Fig. S4a-S4d**). A similar pattern appeared in networks constructed by "subtype *k*-mer count" (**Fig. 4g-4k**), with more significant discrepancy demonstrated in paths initiated from an enriched *k*-mer assigned to subtype B (**Fig. 4h**) and C (**Fig. 4i**). No difference in labeling mismatch networks across each tested assumption was observed (**Extended data Fig. S5**). Altogether, these findings verified the presence of profound barriers between sequence spaces across different groups of HIV-1 subtypes.

Discussion

PORT-EK-v2 has achieved major performance improvements, offering minimal computing time (PORT-EK[40], 01:01:27; PORT-EK-v2, 00:06:50 [hh:mm:ss]), efficient computing costs, and enhanced analytical capabilities for multi-genomic datasets when contrasted with PORT-EK[40]. Overall, PORT-EK-v2 minimizes manual intervention while maintaining high performance accuracy and is readily adaptable to other multi-genomic datasets. Remarkably, PORT-EK-v2 is an efficient pipeline designed to make *k*-mer frequency enrichment statistics available in a readable form. Moreover, PORT-EK-v2 features *k*-mer alignment functionality, which operates without additional supporting metrics. Using PORT-EK-v2, we analyzed multiple HIV-1 DNA and RNA sequences and provided several fresh insights based on findings presented in this work. Firstly, distinct behaviours between unique and common enriched *k*-mers (**Fig. 2** and **Extended data Fig. S3**) suggest that perhaps subgenomic locations are present in the sequence property: unique enriched *k*-mer frequencies are most likely the main determinants governing intrinsic differences across different HIV-1 subtypes. Secondly, indistinguish pattern between enriched DNA and RNA *k*-mers mapped to HIV-1 genes may infer insufficiency of resolution for the classification of the sequence property at the genic level (**Fig. 2f-2i**). This also suggests that future taxonomies of HIV-1 subtypes should be manifested in a higher tiered order at the level of enriched *k*-mer frequencies to discover more profound diversifications of HIV-1 subtypes and recombinant forms. In addition, one should also seek informative features from non-coding regions, given that the majority of variants identified by genome-wide association studies of complex phenotypes are non-coding[45]. Thirdly, the consistent gap present (ranging from 8.3 kb to 8.6 kb) across all HIV-1 subtypes in both enriched DNA and RNA *k*-mer profiles (**Extended data Fig. S3e** and **S3f**) indicates that nucleotide compositions in this gap are either highly varied or conserved across HIV-1 subtypes. Biological consequences of this consensus genomic region will require further investigation. Fourthly, our findings that relatively high "isolate *k*-mer count" in rare subtypes (**Extended data Fig. S1h**), an independent subset in the group of rare subtypes in phylogenetic trees constructed using enriched RNA *k*-mers (**Fig. 3j** and **3k**) unveil the genomic heterogeneity within rare HIV-1 subtypes, agreeing with the fact that many of these rare subtypes are more frequently found in recombinant forms[28,46–50]. Although subtypes C and B are the most prevalent globally, rare subtypes, including F, H, J, K, and L, which are individually responsible for less than 1% of HIV-1 infections worldwide, have been reported to play relevant roles in viral evolution, global persistence, recombination, and drug resistance as well[51,52]. In this circumstance, a more sophisticated and sensitive method that enables the detection of the genomic discrepancy at the molecular level will be essential to tackle ongoing viral outbreaks, pandemic preparedness, and to respond to ongoing and emerging threats.

Limitations of this study

One of the major limitations is that recombinant subtypes have not been included in this study, like the circulating recombinant form (CRF) 01_AE is one of the most prevalent HIV-1 subtypes together with subtypes C, A, and B[53]. In addition, mutations emerging on enriched *k*-mers have not been investigated yet. Future research focus should be placed on closer investigation of the functional roles (e.g., identification of drug-resistant mutations across distinct HIV-1 subtypes) of *k*-mers between DNA and RNA sequences in

clinical management, enabling a better application in the context of HIV-1 molecular epidemiology associated with precise genotypic HIV-1 drug resistance testing[54,55], and genotypic therapeutics. To this extent, PORT-EK-v2 may have the potential to serve as one of the novel molecular epidemiology approaches applied to the future investigation of strategies for curbing HIV transmission, preventing resistance, and improving the understanding of HIV pathogenesis and HIV-1 genomic surveillance.

Methods

Data acquisition and quality control of retrieved sequences. A total of 10,015 and 5,494 HIV-1 DNA and RNA genomic sequences across different subtypes within the group M HIV-1 were retrieved from the Los Alamos HIV Sequence Database (https://www.hiv.lanl.gov/). Only complete genomic sequences were included in this study. Analyzed sequences and corresponding exact IDs are archived at GitHub (https://github.com/Quantitative-Virology-Research-Group/PORT-EK-version-2/tree/main/Input.HIV.sequence). 2 and 4 HIV-1 DNA and RNA sequences, which cannot be aligned to the reference genome (HIV-1 HXB2, complete genome K03455.1, curated in the National Center for Biotechnology Information database), provided in the Los Alamos HIV Sequence Database were discarded.

Conception, bioinformatic analysis, modeling, data visualization, and data availability. The rationale of PORT-EK-v2 and definitions of each feature computed in this work are detailed in **online Methods**. The final output of enriched DNA and RNA *k*-mers can be referred to **Supplementary Table S2-S5**. Enriched DNA and RNA *k*-mer counts at the coordinate of isolates can be referred to **Supplementary Table S6-S9** (archived at Zenodo DOI:10.5281/zenodo.18757144). The PORT-EK-v2 code, the instruction of installation and execution of the PORT-EK-v2 pipeline, and all codes for downstream analyses performed in this work are available at GitHub (https://github.com/Quantitative-Virology-Research-Group/PORT-EK-version-2).

References

1. Charneau, P. *et al.* Isolation and envelope sequence of a highly divergent HIV-1 isolate: definition of a new HIV-1 group. *Virology* **205**, 247–253 (1994).
2. Simon, F. *et al.* Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nat Med* **4**, 1032–1037 (1998).
3. Plantier, J.-C. *et al.* A new human immunodeficiency virus derived from gorillas. *Nat Med* **15**, 871–872 (2009).
4. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral mutation rates. *J Virol* **84**, 9733–9748 (2010).
5. Redd, A. D., Quinn, T. C. & Tobian, A. A. R. Frequency and implications of HIV superinfection. *Lancet Infect Dis* **13**, 622–628 (2013).
6. Simmonds, P. *et al.* Discontinuous sequence change of human immunodeficiency virus (HIV) type 1 env sequences in plasma viral and lymphocyte-associated proviral populations in vivo: implications for models of HIV pathogenesis. *J Virol* **65**, 6266–6276 (1991).
7. Sabino, E. C. *et al.* Identification of human immunodeficiency virus type 1 envelope genes recombinant between subtypes B and F in two epidemiologically linked individuals from Brazil. *J Virol* **68**, 6340–6346 (1994).
8. Zhang, M. *et al.* The role of recombination in the emergence of a complex and dynamic HIV epidemic. *Retrovirology* **7**, 25 (2010).
9. Preston, B. D., Poiesz, B. J. & Loeb, L. A. Fidelity of HIV-1 reverse transcriptase. *Science* **242**, 1168–1171 (1988).
10. Purcell, D. F. & Martin, M. A. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. *J Virol* **67**, 6365–6378 (1993).
11. Ocwieja, K. E. *et al.* Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing. *Nucleic Acids Res* **40**, 10345–10355 (2012).
12. Nguyen Quang, N. *et al.* Dynamic nanopore long-read sequencing analysis of HIV-1 splicing events during the early steps of infection. *Retrovirology* **17**, 25 (2020).
13. Sarmati, L. *et al.* Drug-associated resistance mutations in plasma and peripheral blood mononuclear cells of human immunodeficiency virus type 1-infected patients for whom highly active antiretroviral therapy is failing. *J Clin Microbiol* **41**, 1760–1762 (2003).
14. Chew, C. B. *et al.* Assessment of drug resistance mutations in plasma and peripheral blood mononuclear cells at different plasma viral loads in patients receiving HAART. *J Clin Virol* **33**, 206–216 (2005).
15. Turriziani, O. *et al.* Genotypic resistance of archived and circulating viral strains in the blood of treated HIV-infected individuals. *J Acquir Immune Defic Syndr* **44**, 518–524 (2007).
16. Vicenti, I., Razzolini, F., Saladini, F., Romano, L. & Zazzi, M. Use of peripheral blood DNA for genotype antiretroviral resistance testing in drug-naive HIV-infected subjects. *Clin Infect Dis* **44**, 1657–1661 (2007).
17. Geretti, A. M. *et al.* HIV DNA Sequencing to Detect Archived Antiretroviral Drug Resistance. *Infect*

Dis Ther **11**, 1793–1803 (2022).

18. Bruner, K. M. *et al.* Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nat Med* **22**, 1043–1049 (2016).
19. Sengupta, S. & Siliciano, R. F. Targeting the Latent Reservoir for HIV-1. *Immunity* **48**, 872–895 (2018).
20. Simonetti, F. R. *et al.* Intact proviral DNA assay analysis of large cohorts of people with HIV provides a benchmark for the frequency and composition of persistent proviral DNA. *Proc Natl Acad Sci U S A* **117**, 18692–18700 (2020).
21. Ho, Y.-C. *et al.* Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* **155**, 540–551 (2013).
22. Hiener, B. *et al.* Identification of Genetically Intact HIV-1 Proviruses in Specific CD4 T Cells from Effectively Treated Participants. *Cell Rep* **21**, 813–822 (2017).
23. Bruner, K. M. *et al.* A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. *Nature* **566**, 120–125 (2019).
24. Dauwe, K., Staelens, D., Vancoillie, L., Mortier, V. & Verhofstede, C. Deep sequencing of HIV-1 RNA and DNA in newly diagnosed patients with baseline drug resistance showed no indications for hidden resistance and is biased by strong interference of hypermutation. *J. Clin. Microbiol.* **54**, 1605–1615 (2016).
25. Lee, G. Q. *et al.* HIV-1 DNA sequence diversity and evolution during acute subtype C infection. *Nat Commun* **10**, 2737 (2019).
26. Myers, G., MacInnes, K. & Korber, B. The emergence of simian/human immunodeficiency viruses. *AIDS Res Hum Retroviruses* **8**, 373–386 (1992).
27. Louwagie, J. *et al.* Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. *AIDS* **7**, 769–780 (1993).
28. Triques, K. *et al.* Near-full-length genome sequencing of divergent African HIV type 1 subtype F viruses leads to the identification of a new HIV type 1 subtype designated K. *AIDS Res Hum Retroviruses* **16**, 139–151 (2000).
29. Foley, B. *et al. HIV Sequence Compendium 2018.* https://www.osti.gov/servlets/purl/1458915/ (2018) doi:10.2172/1458915.
30. Rozanov, M., Plikat, U., Chappey, C., Kochergin, A. & Tatusova, T. A web-based genotyping resource for viral sequences. *Nucleic Acids Res* **32**, W654–9 (2004).
31. Pineda-Peña, A.-C. *et al.* 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 (2013).
32. Solis-Reyes, S., Avino, M., Poon, A. & Kari, L. An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. *PLoS One* **13**, e0206409 (2018).
33. Tang, R. *et al.* Genetic source completeness of HIV-1 circulating recombinant forms (CRFs) predicted by multi-label learning. *Bioinformatics* **37**, 750–758 (2021).
34. Tang, R., Yu, Z. & Li, J. KINN: An alignment-free accurate phylogeny reconstruction method based on

inner distance distributions of k-mer pairs in biological sequences. *Mol Phylogenet Evol* **179**, 107662 (2023).

35. Lebatteux, D., Soudeyns, H., Boucoiran, I., Gantt, S. & Diallo, A. B. Machine learning-based approach KEVOLVE efficiently identifies SARS-CoV-2 variant-specific genomic signatures. *PLoS One* **19**, e0296627 (2024).
36. Wade, K. E., Chen, L., Deng, C., Zhou, G. & Hu, P. Investigating alignment-free machine learning methods for HIV-1 subtype classification. *Bioinform Adv* **4**, vbae108 (2024).
37. Ma, Y. *et al.* Phylogenetic analysis of HIV-1 genomes based on the position-Weighted K-mers method. *Entropy (Basel)* **22**, 255 (2020).
38. Wang, T., Yu, Z.-G. & Li, J. CGRWDL: alignment-free phylogeny reconstruction method for viruses based on chaos game representation weighted by dynamical language model. *Front Microbiol* **15**, 1339156 (2024).
39. Paremskaia, A. I. *et al.* Web Service for HIV Drug Resistance Prediction Based on Analysis of Amino Acid Substitutions in Main Drug Targets. *Viruses* **15**, (2023).
40. Wiśniewski, J. & Chen, H.-C. Identification of potential SARS-CoV-2 genomic regions representing hallmarks for adaptation to different hosts. *iMetaOmics* **2**, (2025).
41. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
42. Deorowicz, S., Gudys, A., Dlugosz, M., Kokot, M. & Danek, A. Kmer-db: instant evolutionary distance estimation. *Bioinformatics* **35**, 133–136 (2019).
43. Zielezinski, A. *et al.* Ultrafast and accurate sequence alignment and clustering of viral genomes. *Nat Methods* **22**, 1191–1194 (2025).
44. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
45. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
46. Gao, F. *et al.* An isolate of human immunodeficiency virus type 1 originally classified as subtype I represents a complex mosaic comprising three different group M subtypes (A, G, and I). *J Virol* **72**, 10234–10241 (1998).
47. Oelrichs, R. B., Workman, C., Laukkanen, T., McCutchan, F. E. & Deacon, N. J. A novel subtype A/G/J recombinant full-length HIV type 1 genome from Burkina Faso. *AIDS Res Hum Retroviruses* **14**, 1495–1500 (1998).
48. Vidal, N., Bazepeo, S. E., Mulanga, C., Delaporte, E. & Peeters, M. Genetic characterization of eight full-length HIV type 1 genomes from the Democratic Republic of Congo (DRC) reveal a new subsubtype, A5, in the A radiation that predominates in the recombinant structure of CRF26_A5U. *AIDS Res Hum Retroviruses* **25**, 823–832 (2009).
49. Yebra, G. *et al.* A high HIV-1 strain variability in London, UK, revealed by full-genome analysis: Results from the ICONIC project. *PLoS One* **13**, e0192081 (2018).
50. Lan, Y. *et al.* Genetic characteristics of HIV-1 CRF06_cpx and CRF56_cpx strains isolated in

Guangzhou, China. *Arch Virol* **166**, 2853–2857 (2021).

51. Hemelaar, J. *et al.* Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and trend analysis. *Lancet Infect Dis* **19**, 143–155 (2019).
52. Indihar, D. F. *et al.* Rare but relevant: a comprehensive narrative review of rare HIV-1 group M subtypes. *mBio* **16**, e0126425 (2025).
53. Serwin, K. *et al.* Molecular epidemiology and HIV-1 variant evolution in Poland between 2015 and 2019. *Sci Rep* **11**, 16609 (2021).
54. Günthard, H. F. *et al.* Human Immunodeficiency Virus Drug Resistance: 2018 Recommendations of the International Antiviral Society-USA Panel. *Clin Infect Dis* **68**, 177–187 (2019).
55. Günthard, H. F., Kusejko, K. & Kouyos, R. D. Phylogenetics and molecular evolution to understand and curb the HIV pandemic. *Nat Rev Microbiol* (2025) doi:10.1038/s41579-025-01202-w.

Supplementary Data statement

Supplementary Tables S1-S12.

Tables S2-S9. Excel files containing additional data too large to fit in a PDF and are archived at Zenodo (DOI:10.5281/zenodo.18757144).

Acknowledgements

The study was partially supported by the IMPRESS-U program, funded by the Narodowe Centrum Nauki (NCN, Poland; grant UMO-2023/05/Y/ST6/00263) and the National Science Foundation (NSF, USA; grant OISE-2412914).

Author Contributions Statement

Conceptualization: HCC

Methodology: JW and HCC

Software: JW and HCC

Formal analysis: JW and HCC

Investigation, JW, SK, and HCC

Resources: HCC

Data curation: JW and HCC

Writing of original draft manuscript: JW and HCC

Writing, manuscript review and editing: JW, SK, MP, AKP, PS, AK, SY, and HCC

Visualization: JW and HCC

Supervision: HCC

Project administration: AK, SY, and HCC

Funding acquisition: AK, SY, and HCC

Conflict-of-interest declarations

The authors declare no conflicts of interest.

Figure Legends

Fig. 1. Distinct enriched DNA and RNA *k*-mers across different groups of HIV-1 subtypes. (**a**-**d**) Two dimensional principal component analysis (PCA) revealing the discrepancy of enriched DNA (**a**, *k*-mer = 13 bp; **b**, *k*-mer = 13 bp) and RNA *k*-mers (**c**, *k*-mer = 13 bp; **d**, *k*-mer = 13 bp). Dots marked green, orange, red, purple, and dark blue represent HIV-1 subtype A, B, C, D, and rare subtypes, respectively. (**e**-**h**) Phylogenetic trees representing the evolutionary distance across isolates assigned to different groups of HIV-1 subtypes. Trees were constructed by "*k*-mer average count" computed from enriched DNA (**e**, *k*-mer = 13 bp; **f**, *k*-mer = 15 bp) and RNA (**g**, *k*-mer = 13 bp; **h**, *k*-mer = 15 bp) *k*-mers across different groups of HIV-1 subtypes. 250 enriched *k*-mers were bootstrapped from each pool of enriched *k*-mers to illustrate a tree plot. The color code annotation denotes the group of HIV-1 subtypes: green, orange, red, purple, and dark blue represent HIV-1 subtype A, B, C, D, and rare subtypes, respectively. (**i**) Schematic representation of the definitions of enriched *k*-mer-related features, including "*k*-mer weight", "subtype *k*-mer count", and "isolate *k*-mer count". Corresponding mathematical equations were provided along with respective features. (**j**, **k**) Quantile plots representing the coverage of enriched *k*-mers in a length of 13 (**j**) or 15 bp (**k**) throughout the complete HIV-1 genome at the quantile of 0.5 kb. Facets on the y-axis separate enriched *k*-mers computed by the DNA or RNA sequence. Dots and dashed lines marked green, orange, red, purple, and dark blue represent HIV-1 subtype A, B, C, D, and rare subtypes, respectively. (**l**, **m**) Line plots representing the coverage of enriched *k*-mers in a length of 13 (**l**) or 15 bp (**m**) throughout the complete HIV-1 genome. Facets at the x-axis separate different groups of HIV-1 subtypes; facets at the y-axis separate enriched *k*-mers computed by DNA or RNA sequence. Lines marked green, orange, red, purple, and dark blue represent HIV-1 subtype A, B, C, D, and rare subtypes, respectively.

Fig. 2. Unique enriched *k*-mers account for the discrepancy of the sequence properties across HIV-1 subtypes. (**a**-**d**) Scatter plots representing the selection of unique enriched DNA and RNA *k*-mers in a length of 13 bp (**a** and **b**) and 15 bp (**c** and **d**). Two distinct cutoffs (**a** and **c**, cutoff = 0.5; **b** and **d**, cutoff = 1) based on the ratio of "*k*-mer RMSE" represented on a logarithmic scale were applied. Dots marked in orange, red, and light blue represent enriched *k*-mers, which are exclusively present in a pool of enriched DNA and enriched RNA *k*-mers, respectively; dots marked in grey represent enriched *k*-mers that do not pass the selection threshold. (**e**) Bar charts representing the percentage of unique DNA and RNA *k*-mers in two distinct cutoffs (cutoff = 0.5 versus cutoff = 1.0) across different groups of HIV-1 subtypes. Facets at the x-axis separate enriched *k*-mers in a length of 13 bp or 15 bp; facets at the y-axis separate *k*-mers enriched by DNA or RNA sequence. Squares marked in light and dark green represent the cutoff threshold at 0.5 and 1.0, respectively. (**f**, **g**) Clustering heatmaps representing the frequency of unique enriched DNA (written in red beneath the heatmap) and RNA (written in blue beneath the heatmap) enriched *k*-mers aligned to HIV-1 genes (row) in a *k*-mer length of 13 bp (**f**) and 15 bp (**g**) across different groups of HIV-1 subtypes. The color scale depicts the frequency of unique enriched *k*-mers appearing in HIV-1 genes. (**h**, **i**) Clustering heatmaps representing the frequency of common enriched k-mers in a length of 13 bp (**h**) and 15 bp (**i**) across different groups of HIV-1 subtypes. The color scale depicts the frequency of unique enriched *k*-mers appearing at HIV-1 genes.

Fig. 3. Isolate k-mer count enables the classification of different groups of HIV-1 subtypes. (**a**) Box plots representing the area under the curve (AUC) of logistic regression models, constructed by distinct enriched *k*-mer-related features, including "isolate *k*-mer count", "subtype *k*-mer count", "*k*-mer average count", "*k*-mer RMSE", and "*k*-mer weight" (from left to right at the x-axis), to classify the origin of enriched *k*-mers retrieved between DNA and RNA sequences. Each classifier was constructed by 1,000 enriched *k*-mers bootstrapped from the pool of total enriched DNA or RNA *k*-mers, separated between a *k*-mer length equal to 13 bp (the left-hand side panel) and 15 bp (the right-hand side panel). Each classification iteration was repeated 1,000 times for statistical significance. Significance levels are denoted as follows: ns for no significance, **p* 0.05, ***p* 0.01, ****p* 0.001, *****p* 0.0001. (**b**) Box plots representing the probability of multinomial logistic regression models, constructed by distinct enriched *k*-mer-related features, including "isolate *k*-mer count", "subtype *k*-mer count", "*k*-mer average count", "*k*-mer RMSE", and "*k*-mer weight", to classify the origin of enriched *k*-mers retrieved across different groups of HIV-1 subtypes. Each classifier was constructed by 1,000 enriched *k*-mers bootstrapped from the pool of total enriched DNA or RNA *k*-mers, separated between a *k*-mer length equal to 13 bp (top panels) and 15 bp (bottom panels). Each classification iteration was repeated 1,000 times for statistical significance. Facets at the x-axis separate different enriched *k*-mer-related features; facets at the y-axis separate a *k*-mer length between 13 bp and 15 bp. Boxes marked in orange red and sky blue represent enriched DNA and RNA *k*-mers, respectively. (**c**) Box plots representing the accuracy of simple neural network-based models, constructed by distinct enriched *k*-mer-related features, including "isolate *k*-mer count", "subtype *k*-mer count", "*k*-mer average count", "*k*-mer RMSE", and "*k*-mer weight" (from left to right at the x-axis), to verify the predictive performance contributed by individual enriched *k*-mer-related features in the classification of the origin of enriched *k*-mers retrieved across different groups of HIV-1 subtypes, shown in (**b**). Each classifier was constructed by 1,000 enriched *k*-mers bootstrapped from the pool of total enriched DNA or RNA *k*-mers, separated between a *k*-mer length equal to 13 bp (left-hand side panels) and 15 bp (right-hand side panels). Each classification iteration was repeated 10 times for statistical significance. Significance levels are denoted as follows: ns for no significance, **p* 0.05, ***p* 0.01, ****p* 0.001, *****p* 0.0001. (**d**-**g**) Clustering heatmaps representing the spatial correlation between two adjacent isolates presenting a topological network constructed by "isolate *k*-mer count" computed from enriched DNA (**d**, *k*-mer = 13 bp; **e**, *k*-mer = 15 bp) and RNA (**f**, *k*-mer = 13 bp; **g**, *k*-mer = 15 bp) *k*-mers across different groups of HIV-1 subtypes. The color scale represents the Euclidean distance between two adjacent isolates. The color code annotation beneath and at the right-hand side of the heatmap denotes the group of HIV-1 subtypes: green, orange, red, purple, and dark blue represent HIV-1 subtype A, B, C, D, and rare subtypes, respectively.

Fig. 4. MCMC modeling the intrinsic barrier in sequence properties between HIV-1 DNA and RNA sequences. (**a**) Schematic representation of the strategy of subsampling of enriched k-mers, followed by the construction of a pentapartite graph of a topological network. Five scenarios under a framework of Markov assumption were set up, in which 10 paths of random walks up to 10,000 steps were simulated across distinct network properties. Lastly, the probability of appearing vertices assigned to different groups of HIV-1

subtypes within a path in each scenario was computed. (**b**-**f**) Box plots representing the probability that vertices assigned to different groups of HIV-1 subtypes appear in a path in each scenario, whereby a random walk initiates a vertex assigned to HIV-1 subtype A (**b**), B (**c**), C (**d**), D (**e**), and rare subtypes (**f**). Networks were constructed by "isolate *k*-mer count", whereby each vertex represents an individual HIV-1 isolate. Facets at the x-axis separate enriched DNA and RNA *k*-mers; facets at the y-axis separate a *k*-mer length between 13 bp and 15 bp. Boxes marked in green, orange, red, purple, and dark blue represent HIV-1 subtype A, B, C, D, and rare subtypes, respectively. Significance levels are denoted as follows: ****p 0.0001. (**g**-**k**) Box plots representing the probability that vertices assigned to different groups of HIV-1 subtypes appear in a path in each scenario, whereby a random walk initiates a vertex assigned to HIV-1 subtype A (**g**), B (**h**), C (**i**), D (**j**), and rare subtypes (**k**). Networks were constructed by "subtype *k*-mer count", whereby each vertex represents an individual enriched *k*-mer. Facets at the x-axis separate enriched DNA and RNA *k*-mers; facets at the y-axis separate a *k*-mer length between 13 bp and 15 bp. Boxes marked in green, orange, red, purple, and dark blue represent HIV-1 subtype A, B, C, D, and rare subtypes, respectively. Significance levels are denoted as follows: ns for no significance, *p 0.05, **p 0.01, ***p 0.001.

Table 1. Benchmarking results of PORT-EK-v2 vs PORT-EK.

Dataset (size)	PORT-EK[40] (*K*-mer = 15 bp)		PORT-EK-v2 (*K*-mer = 15 bp)	
	Total wall time [hh:mm:ss]	**Peak memory usage (RSS) [GB]**	**Total wall time [hh:mm:ss]**	**Peak memory usage (RSS) [GB]**
HIV-1 RNA (5,490 genomes)	01:01:27	26.4	00:06:50	4.7
Downsampled HIV-1 RNA set 1 (2,000 genomes)	00:22:43	7.8	00:04:23	2.0
Downsampled HIV-1 RNA set 2 (1,000 genomes)	00:14:29	2.5	00:03:36	1.3

Table 2. Benchmarking results of *k*-mer counting speed using PORT-EK-v2, Jellyfish, and Kmer-db based on HIV-1 RNA sequences (5,490 genomes).

Software	**PORT-EK[40]**	**PORT-EK-v2**	**Jellyfish[41]**	**Kmer-db[42,43]**
***K*-mer counting time [mm:ss]**				
***K*-mer = 13 bp**	00:53	00:41	00:43	00:036
***K*-mer = 15 bp**	00:53	00:42	00:045	00:034

Figure 1

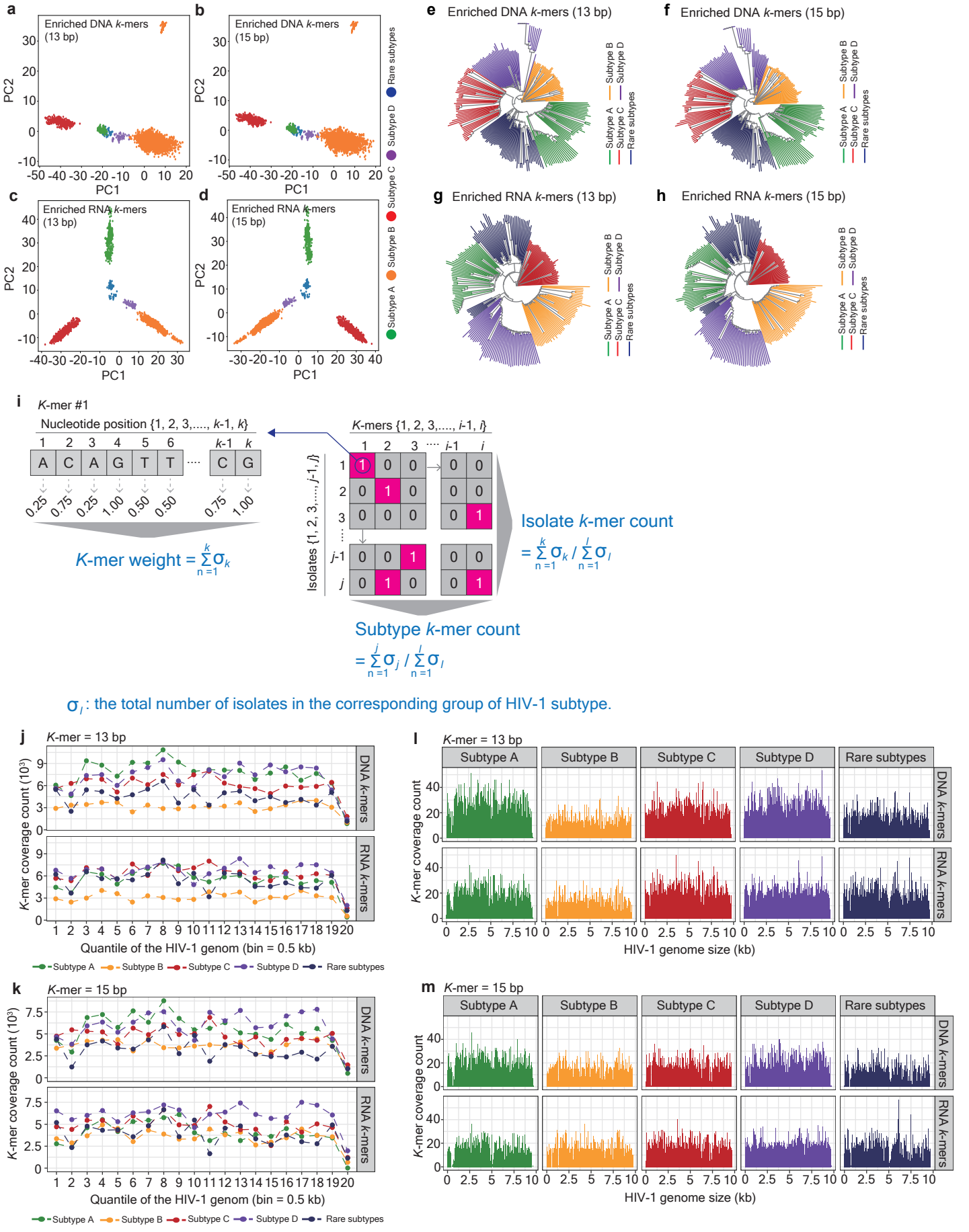

a
Enriched DNA k-mers (13 bp)
b
Enriched DNA k-mers (15 bp)
c
Enriched RNA k-mers (13 bp)
d
Enriched RNA k-mers (15 bp)
PC1
PC2
Rare subtypes
Subtype D
Subtype C
Subtype B
Subtype A
e
Enriched DNA k-mers (13 bp)
f
Enriched DNA k-mers (15 bp)
g
Enriched RNA k-mers (13 bp)
h
Enriched RNA k-mers (15 bp)
i
K-mer #1
Nucleotide position {1, 2, 3,...., k-1, k}
A C A G T T .... C G
0.25 0.75 0.25 1.00 0.50 0.50 0.75 1.00
K-mer weight = $\sum_{n=1}^{k} \sigma_k$
K-mers {1, 2, 3,...., i-1, i}
Isolates {1, 2, 3,...., j-1, j}
Isolate k-mer count
$= \sum_{n=1}^{k} \sigma_k / \sum_{n=1}^{l} \sigma_l$
Subtype k-mer count
$= \sum_{n=1}^{j} \sigma_j / \sum_{n=1}^{l} \sigma_l$
σ_l: the total number of isolates in the corresponding group of HIV-1 subtype.
j
K-mer = 13 bp
K-mer coverage count (10^3)
DNA k-mers
RNA k-mers
Quantile of the HIV-1 genom (bin = 0.5 kb)
Subtype A
Subtype B
Subtype C
Subtype D
Rare subtypes
l
K-mer = 13 bp
K-mer coverage count
HIV-1 genome size (kb)
k
K-mer = 15 bp
m
K-mer = 15 bp

Figure 2

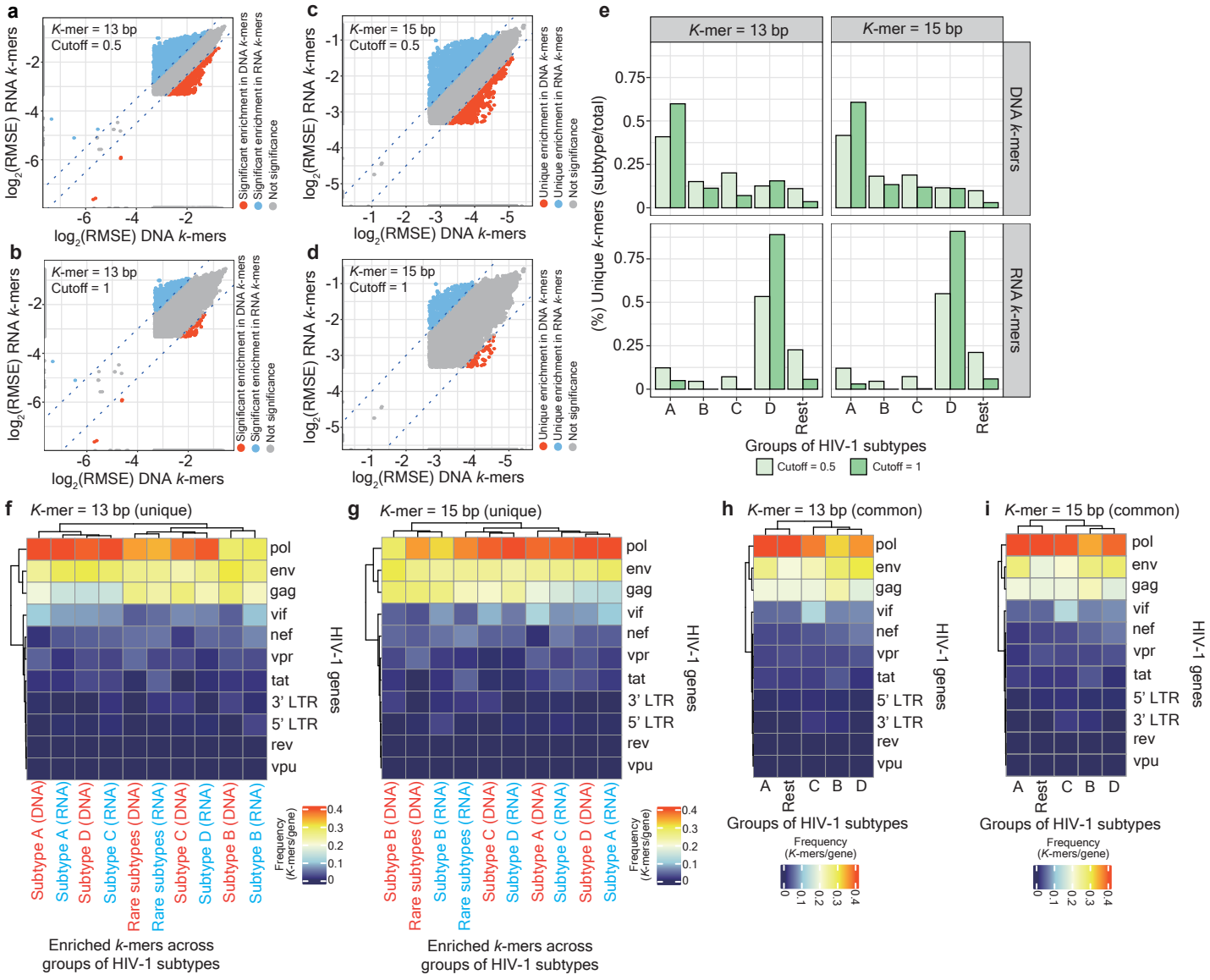

a
K-mer = 13 bp
Cutoff = 0.5
log2(RMSE) RNA k-mers
log2(RMSE) DNA k-mers
Significant enrichment in DNA k-mers
Significant enrichment in RNA k-mers
Not significance
b
K-mer = 13 bp
Cutoff = 1
c
K-mer = 15 bp
Cutoff = 0.5
Unique enrichment in DNA k-mers
Unique enrichment in RNA k-mers
d
K-mer = 15 bp
Cutoff = 1
e
K-mer = 13 bp
K-mer = 15 bp
DNA k-mers
RNA k-mers
(%) Unique k-mers (subtype/total)
A B C D Rest
Groups of HIV-1 subtypes
Cutoff = 0.5
Cutoff = 1
f
K-mer = 13 bp (unique)
g
K-mer = 15 bp (unique)
h
K-mer = 13 bp (common)
i
K-mer = 15 bp (common)
pol
env
gag
vif
nef
vpr
tat
3' LTR
5' LTR
rev
vpu
HIV-1 genes
Subtype A (DNA)
Subtype A (RNA)
Subtype D (DNA)
Subtype C (RNA)
Rare subtypes (DNA)
Rare subtypes (RNA)
Subtype C (DNA)
Subtype D (RNA)
Subtype B (DNA)
Subtype B (RNA)
Frequency (K-mers/gene)
Enriched k-mers across groups of HIV-1 subtypes

Figure 3

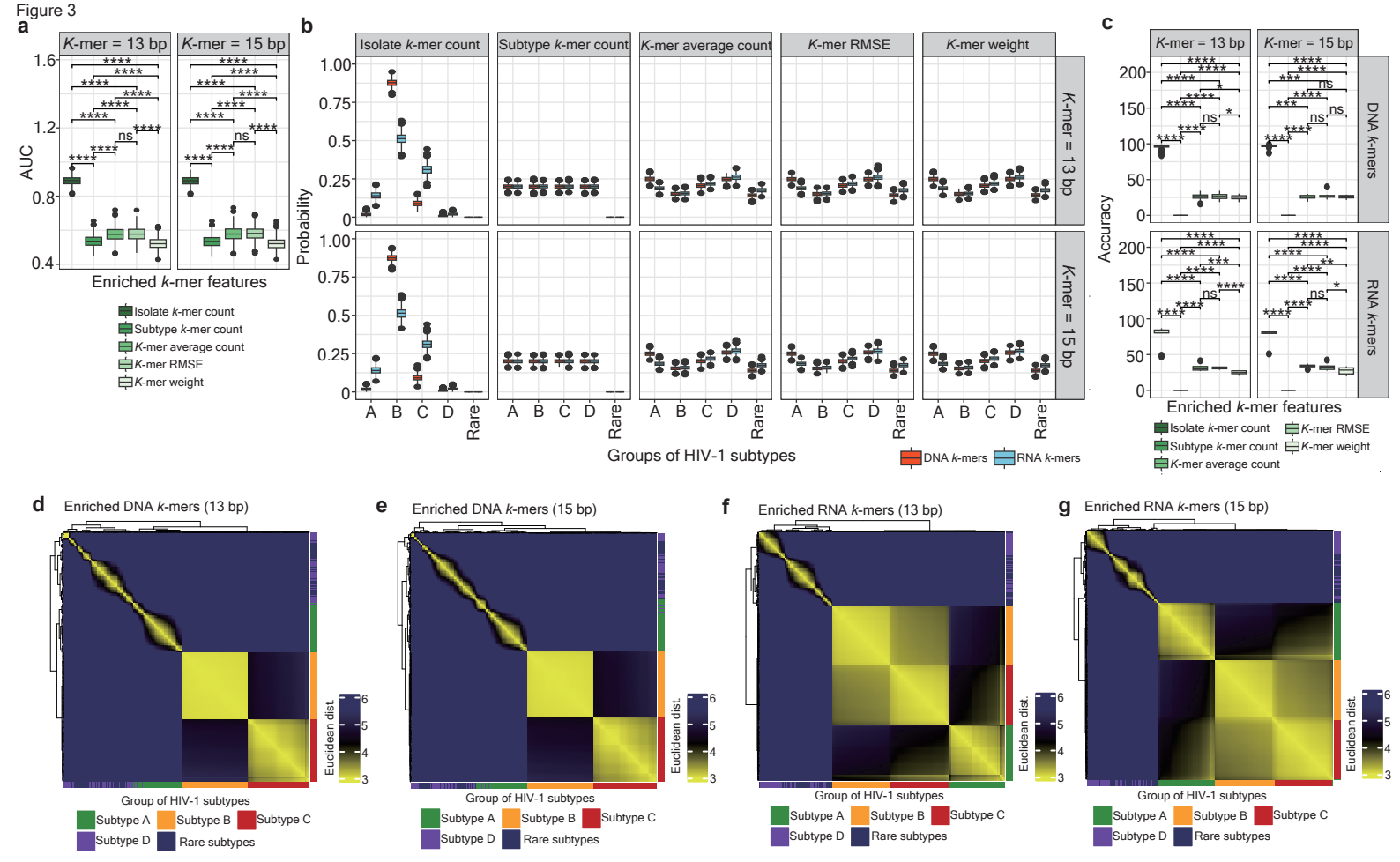
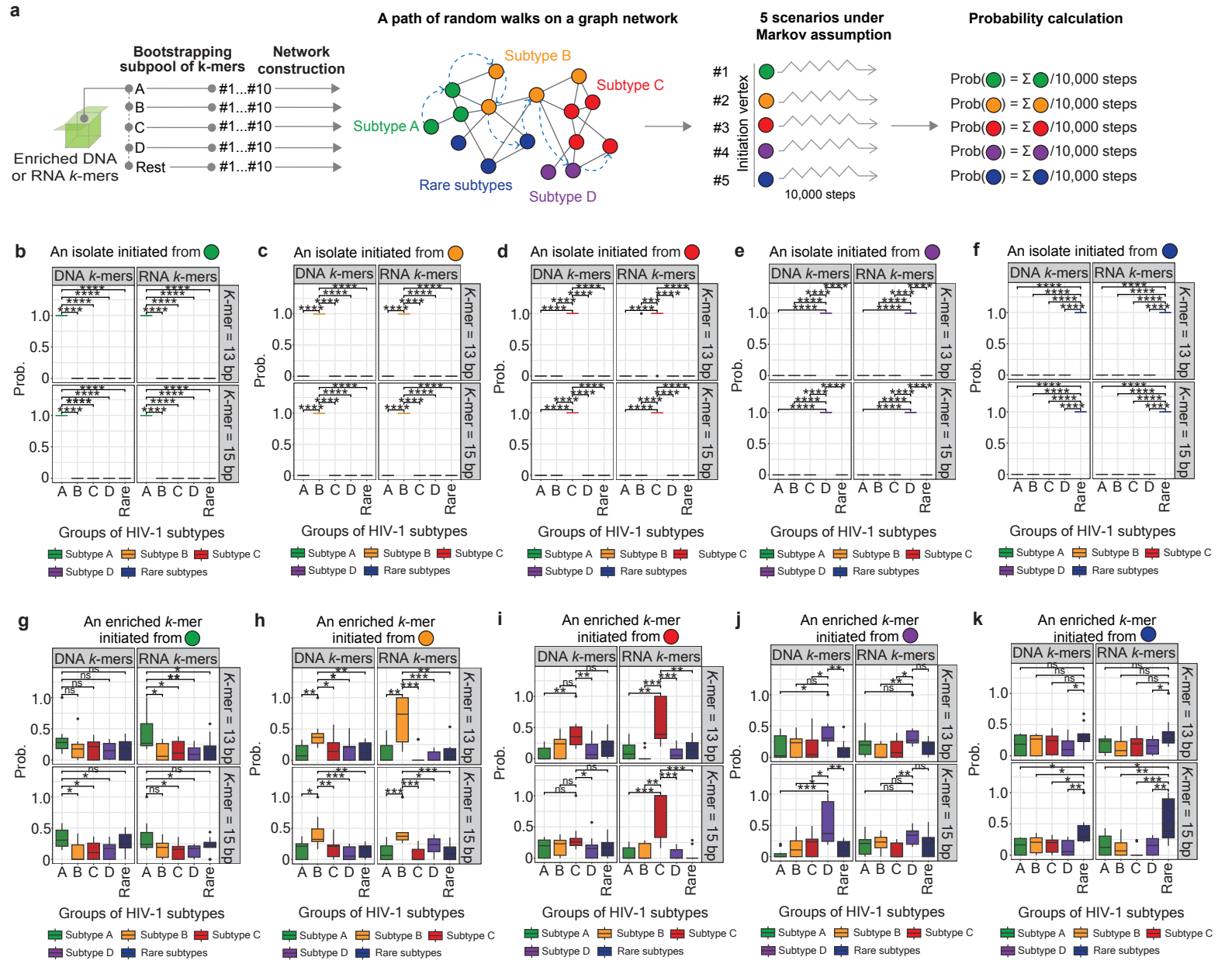

a
K-mer = 13 bp
K-mer = 15 bp
AUC
Enriched k-mer features
Isolate k-mer count
Subtype k-mer count
K-mer average count
K-mer RMSE
K-mer weight
b
Isolate k-mer count
Subtype k-mer count
K-mer average count
K-mer RMSE
K-mer weight
Probability
K-mer = 13 bp
K-mer = 15 bp
A B C D Rare
Groups of HIV-1 subtypes
DNA k-mers
RNA k-mers
c
K-mer = 13 bp
K-mer = 15 bp
DNA k-mers
RNA k-mers
Accuracy
Enriched k-mer features
Isolate k-mer count
Subtype k-mer count
K-mer average count
K-mer RMSE
K-mer weight
d
Enriched DNA k-mers (13 bp)
e
Enriched DNA k-mers (15 bp)
f
Enriched RNA k-mers (13 bp)
g
Enriched RNA k-mers (15 bp)
Euclidean dist.
Group of HIV-1 subtypes
Subtype A
Subtype B
Subtype C
Subtype D
Rare subtypes

Figure 4


a
Bootstrapping subpool of k-mers
Network construction
Enriched DNA or RNA k-mers
A
B
C
D
Rest
#1...#10
A path of random walks on a graph network
Subtype A
Subtype B
Subtype C
Subtype D
Rare subtypes
5 scenarios under Markov assumption
Initiation vertex
#1
#2
#3
#4
#5
10,000 steps
Probability calculation
Prob( ) = Σ /10,000 steps
b An isolate initiated from
c An isolate initiated from
d An isolate initiated from
e An isolate initiated from
f An isolate initiated from
g An enriched k-mer initiated from
h An enriched k-mer initiated from
i An enriched k-mer initiated from
j An enriched k-mer initiated from
k An enriched k-mer initiated from
DNA k-mers
RNA k-mers
K-mer = 13 bp
K-mer = 15 bp
Prob.
A B C D Rare
Groups of HIV-1 subtypes
Subtype A
Subtype B
Subtype C
Subtype D
Rare subtypes