Identification of different intrinsic sequence patterns 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. At present, HIV-1 subtyping remains a challenging task in a great part due to analytical tool discordance. To tackle this issue, in this study, we present an updated version of a *k*-mer-based approach, PORT-EK-v2, a streamlined bioinformatic pipeline, allowing for a comparison of multiple genomic datasets and identification of over-represented genomic regions, *k*-mers, related to specific origins of datasets. Using PORT-EK-v2, we exemplified that intrinsic sequence patterns between HIV-1 DNA and RNA are distinct across group M HIV-1 subtypes. Furthermore, we showcased that "isolate *k*-mer count", a predictive variable computed in this work, could serve as a default choice in classifying the HIV-1 DNA versus RNA sequences across subtypes. Lastly, results based on network-based analyses and Markov chain Monte Carlo modeling unveiled a clear discontinuation of a random walk throughout the network properties corresponding to each tested group of HIV-1 subtypes, confirming the specificity of enriched *k*-mer retrieved by PORT-EK-v2 and the genomic diversity across group M HIV-1 subtypes. Source code for PORT-EK-v2 is at https://github.com/Quantitative-Virology-Research-Group/PORT-EK-version-2 and is freely available.

Keywords

K-mers, machine learning, HIV-1 subtypes, group M HIV-1, sequence property

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, a lack of proofreading activity of reverse transcriptase[4,5], viral recombination-prone nature[6–8], a lack of proofreading of its replicated RNA genomes[9], and alternative splicing[10,11,12]. Nevertheless, HIV-1 RNA is 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 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 antiretroviral therapy (ART), as evidenced by several studies[13-17]. Although the feasibility of the proviral DNA-based HIV-1 resistance testing remains debated, 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 exacerbating the heterogeneous nature of proviral genomes in reservoir cells. Although diversities between HIV-1 DNA and RNA have been previously characterized, appropriate HIV-1 subtyping remains an unsolved issue. Given that the composition of a genomic sequence can be viewed at different hierarchical orders, 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-1 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 intrinsic sequence pattern between HIV-1 DNA and RNA exists and may influence the classification of HIV-1 genomic diversity. More comprehensive insights into the difference between HIV-1 DNA and RNA genomic sequences 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 embedded 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]. In addition, presently available reference genomes may not be appropriate for all subspecies or subtypes within certain species, resulting in inaccurate alignment results.

To conquer this limitation, various alignment-free classification methods have been proposed. COMET (COntext-based Modeling for Expeditious Typing)[32] was developed based on adaptive coding and partial string matching[33]. CASTOR[34] was developed based on restriction fragment length polymorphism (RFLP) and has been applied to the subtyping of human papillomavirus (HPV), hepatitis B virus (HBV), and HIV-1. The “Natural Vector representation” approach that classifies single-segmented[35] and multi-segmented[36] viral

genomes, as well as viral proteomes[37]. Phymm, an interpolated Markov-based model, has been applied to the phylogenetic classification of metagenomic samples[38].

In addition to the mentioned methods, the use of *k*-mers is also employed for phylogenetic applications. Various *k*-mer-based alignment-free methods, including KAMERIS[39], a multi-label learning algorithm[40], KINN[41], KEVOLVE[42], and a *k*-mer-based eXtreme Gradient Boosting model[43], have been proposed for HIV-1 subtyping and phylogenetic analysis[41,44,45], and HIV-1-resistent drug detection[46]. Although the feasibility of *k*-mer-based methods may mitigate the difficulty in the classification of HIV-1 subtypes, limitations exist. For example, while KAMERIS[39] achieves high overall accuracy, recall across minority classes seems to be insufficient. Although KEVOLVE[42] achieves good classification performance, it tends to mistakenly classify recombinant subtypes as pure subtypes[43]. The *k*-mer and position-based vectorization method, in conjunction with multi-class k-nearest neighbours (KNN) algorithm, was developed using only a single gene, which may risk failing to capture the complexity of the entire HIV-1 genome[40].

In this work, we showcased the enhanced effectiveness and accuracy of the streamlined pipeline of the Pathogen Origin Recognition Tool using Enriched *K*-mers (PORT-EK)[47] version 2, namely PORT-EK-v2. PORT-EK-v2 is a *k*-mer-based approach, which facilitates the comparison of various multigenomic datasets of HIV-1 subtypes and the detection of over-represented genomic regions (i.e., *k*-mers) linked to specific sequences. Our contributions to PORT-EK-v2 are two fold. First, PORT-EK-v2 computed the relative abundance of enriched *k*-mers (i.e., enriched *k*-mer frequency) between two compared multigenomic datasets, aiming at identifying "quantitative" differences between two species. This feature in PORT-EK-v2 enables unveiling more masked patterns previously hidden in intricate datasets, subsequently gaining a more profound interpretation of results. Second, PORT-EK-v2 automatically computes the optimal length of *k*-mers fitting to the input dataset and generates *k*-mer frequency enrichment statistics in a readable table format (see **Supporting information**). These features allow PORT-EK-v2 to outperform other similar methods. In this work, we applied PORT-EK-v2 to pinpoint *k*-mers abundance in HIV-1 DNA or RNA sequences and showcased distinct sequence patterns between HIV-1 DNA and RNA across group M HIV-1 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. Compared with PORT-EK[47], 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[47] are provided in **Supporting information**. The systematic comparison of key parameters between PORT-EK[47] and PORT-EK-v2 is summarized in **Table S1**. In addition, using input HIV-1 RNA sequences and two subsets of randomly bootstrapped HIV-1 RNA sequences of different sizes, we benchmarked PORT-EK-v2 against PORT-EK[47], revealing a remarkable reduction in computing speed and peak memory footprint of the enrichment of *k*-mer frequencies (**Table 1**). In addition, the *k*-mer counting speed of PORT-EK-v2 was also evaluated to be comparable to that achieved by Jellyfish[48], Kmer-db[49,50], and KMC 3[51] (**Table 2**). It is important to stress that PORT-EK[47] and PORT-EK-v2 are designed and optimized for a large number of small genomes (mainly for viral infectious diseases), as it calculates aggregate statistics concurrently with *k*-mer counting, which may lead to a slow progression for large genomes.

Distinct sequence patterns unveiled by enriched *k*-mers between HIV-1 DNA and RNA 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 subsets 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 the group M HIV-1 subtypes—the subtypes A, B, C, D, and the rare subtypes, including subtypes F1, F2, G, H, J, K, and L—were computed in this work. To identify enriched *k*-mers within a specific HIV-1 subtype, the "*k*-mer average count" (refer to **Supporting information**) has to be substantially greater in one group than in others based on pairwise comparisons (detailed in **Supporting information**).

Using PORT-EK-v2, we first observed distinct distributions between enriched DNA and RNA *k*-mers with a length of 13 bp [hereinafter, total enriched DNA *k*-mers (13 bp), **Fig. 1a**; total enriched RNA *k*-mers (13 bp), **Fig. 1c**] and 15 bp [hereinafter, total enriched DNA *k*-mers (15 bp), **Fig. 1b**; total enriched RNA *k*-mers (15 bp), **Fig. 1d**]. Based on enriched DNA *k*-mers, a bipolar distribution with subtypes B and C was observed (**Fig. 1a** and **1b**). 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, a tripolar and 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 relative position of subtype A genomes is, however, markedly different from subtypes B and C (**Fig. 1c** and **1d**). The similarity of RNA sequences was 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) (**Supplementary Fig. S1a**) and enriched RNA *k*-mers (17 bp) (**Supplementary Fig. S1b**), suggesting the substantial differences in sequence patterns between HIV-1 DNA and RNA.

A clear separation across different groups of HIV-1 subtypes was also confirmed by phylogenetic 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 (**Supplementary 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 at a single isolate level (**Fig. 1i**). The definition of each feature is detailed in **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 (**Supplementary 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 (**Supplementary Fig. S1e**). Although a bias was observed when different ordinal encoding was assigned to every nucleotide in an enriched *k*-mers (**Supplementary Fig. S2a-S2c**), the overall patterns of "*k*-mer weight" across different groups of HIV-1 subtypes were reproducible (**Supplementary Fig. S2d-S2f**).

A more abundant "subtype *k*-mer count" appeared in subtype D, whereas relatively few counts were in subtype B, irrespective of enriched DNA or RNA *k*-mers and their lengths (**Supplementary 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 (**Supplementary Fig. S1g**).

While computing "isolate *k*-mer count", a superior enrichment of enriched DNA *k*-mers appeared in the group of rare subtypes, followed by subtype D and A (**Supplementary Fig. S1h**), whereas the majority of enriched RNA *k*-mers accumulated in the HIV-1 subtype D, followed by the rare subtypes and a relatively minor number of *k*-mers were enriched in isolates across the subtypes A, B, and C (**Supplementary 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 (**Supplementary Fig. S1i**), whereas an opposite pattern was detected in subtypes B and D (**Supplementary 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 (**Supplementary Fig. S1j**, *k*-mer = 13 bp; **Supplementary Fig. S1k**, *k*-mer = 15 bp) and 7 (**Supplementary Fig. S1l**, *k*-mer = 13 bp; **Supplementary 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 of length 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 the sequence pattern varies between HIV-1 DNA and RNA sequences and differs across HIV-1 subtypes, as unveiled by enriched *k*-mers.

Distinct genomic compositions 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 (**Supplementary 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 (**Supplementary Fig. S3a**). Distinct magnitudes of *k*-mer enrichment (measured by "*k*-mer RMSE", see **Supporting information**) verified a strong correlation between unique enriched *k*-mers and distinct genomic composition across different groups of HIV-1 subtypes (**Supplementary Fig. S3b**).

As we computed "*k*-mer weight", only minor differences across different groups of HIV-1 subtypes were observed (**Supplementary Fig. S3c**). Notably, unlike the pattern shown in total enriched DNA *k*-mers (**Supplementary Fig. S1f**), the increased abundance of unique enriched DNA *k*-mers accumulated in subtype B, highlighting the possibility that subgenomic locations throughout the DNA sequence (unique versus common *k*-mers) are present, as "subtype *k*-mer count" was examined (**Supplementary Fig. S3d**). In contrast, a similar pattern between total and unique enriched RNA *k*-mers was observed (**Supplementary 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**). This observation exemplified that the inconsistency between HIV-1 DNA and RNA within the same group of HIV-1 subtypes exists. 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 the

phenotype that distinct loci overlapping unique enriched DNA and RNA *k*-mers across different groups of HIV-1 subtypes were present (**Supplementary Fig. S3e**, *k*-mer = 13 bp; **Supplementary Fig. S3f**, *k*-mer = 15 bp). A superior frequency of unique enriched *k*-mers overlapped the HIV-1 *pol* 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" (**Supplementary Fig. S3g**, *k*-mer = 13 bp, $R^2 = 0.63$; **Supplementary Fig. S3h**, *k*-mer = 15 bp, $R^2 = 0.6$), "*k*-mer weight" (**Supplementary Fig. S3i**, *k*-mer = 13 bp; **Supplementary Fig. S3j**, *k*-mer = 15 bp), and "subtype *k*-mer count" (**Supplementary Fig. S3g**, *k*-mer = 13 bp, $R^2 = 0.89$; **Supplementary Fig. S3h**, *k*-mer = 15 bp, $R^2 = 0.88$) were detected. These findings suggest that the minimal contribution to the distinct sequence pattern results from common enriched *k*-mers. 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 sequence barriers lie across group M HIV-1 subtypes

We further constructed a pentapartite graph (see **Supporting information** for the definition), representing a network consisting of the five groups of HIV-1 subtypes, and used this graph to illustrate the spatial interactions between isolates (**Supplementary Fig. S4a-S4d**) or enriched *k*-mers (**Supplementary Fig. S4f-S4i**). 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 (**Supplementary Fig. S4e**). In the latter scenario, although a strong tendency between enriched *k*-mers in the same group of HIV-1 subtypes was structured, subsets of *k*-mers enriched in the rare subtypes were observed to link distinct enriched *k*-mer

hubs (**Supplementary 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 (**Supplementary Fig. S4j**).

Lastly, we applied the Markov chain Monte Carlo (MCMC) method[52] 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, **Supplementary Fig. S5a-S5e**), respectively. We inspected five different assumptions, in which a random walk initiates in an isolate or an enriched *k*-mer in any HIV-1 subtype group and computed the probability that isolates or enriched *k*-mers in each group of HIV-1 subtypes involve in a path of a random walk (**Fig. 4a**; see **Methods**).

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 was observed (**Fig. 4b-4f**), resonating with disconnected graphs (**Supplementary 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 initiating from an *k*-mer enriched in subtype B (**Fig. 4h**) and C (**Fig. 4i**). No difference in labeling mismatch networks across each tested assumption was observed (**Supplementary Fig. S5**). Altogether, these findings verified the presence of intrinsic barriers between genomic sequences across different groups of HIV-1 subtypes.

Discussion

PORT-EK-v2 has achieved major performance improvements, offering minimal computing time (PORT-EK[47], 01:01:27; PORT-EK-v2, 00:06:50 [hh:mm:ss]), efficient computing costs, and enhanced analytical capabilities for multigenomic datasets when contrasted with PORT-EK[47]. In other words, PORT-EK-v2 minimizes manual intervention while maintaining high performance accuracy and is readily adaptable to different multigenomic datasets. In addition, the improvement of the enriched *k*-mer mapping strategy—pre-calculated reference genome index—has significantly leveraged the performance, resulting in faster execution, the ability to map recurring *k*-mers, and precise indel handling (detailed in **Supporting information** with the pseudocode provided). Overall, PORT-EK-v2 is an efficient and streamlined pipeline designed to immediately provide *k*-mer frequency enrichment statistics in a readable form when dealing with a high volume of small multigenomic datasets. Optimal parameterization in PORT-EK-v2 is detailed in **Supporting information** and **Supplementary Table S1**. 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 **Supplementary Fig. S3**) suggest the presence of subgenomic locations in the HIV-1 genomic composition. Unique enriched *k*-mer frequencies are most likely the main determinants governing intrinsic differences across different HIV-1 subtypes. The phenomenon is remarkably observed in HIV-1 subtype B, when DNA sequence is analyzed (**Supplementary Fig. S3b** and **S3d**). For at least three decades, the epidemic in the Western World has been dominated by subtype B infections[53]. At present, subtype B remains the most disseminated variant and is assumed to be the causative agent in approximately 11% of all cases of HIV worldwide[54]. Relative to other group M HIV-1 subtypes, a higher genomic diversity measured based on enriched *k*-mer frequency in subtype B may be reflected by its global spread, thereby accelerating the evolution of its genomic composition (e.g., accumulation of mutations or deletions, and so on). This assumption could be underpinned by the observation of an outlier of enriched *k*-mers (**Fig. 1a** and **1b**). Further investigation of the evolutionary trajectory of subtype B at the sequence or nucleotide levels will be required to gain a more insightful understanding of its genomic diversity at a larger time scale.

Secondly, an indistinguishable 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 and the phylogenetic analysis should be manifested in a higher-tiered order at the levels of short pieces of genomic sequences or nucleotides to discover more profound diversities of HIV-1 subtypes and recombinant forms. It is important to take into account the inconsistency between HIV-1 DNA and RNA appearing in the same group of HIV-1 subtypes (**Fig. 2e**), which may lead to different biological interpretations. 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[55].

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 (**Supplementary Fig. S3e** and **S3f**) indicates that nucleotide compositions in this gap are highly varied across HIV-1 subtypes. Of note, no conserved *k*-mers were detected in this region.

Intriguingly, this region overlaps with the second exon of *tat-rev* (8,433-8,459 bp) and is in the vicinity of the HIV *rev* response element (RRE, 7,832-7,851 bp)[56-58]. HIV-1 Tat-Rev are essential regulatory proteins that drive HIV-1 gene expression and replication[59] and govern the switch from early-phase to late-phase viral replication, ensuring the delicate balance between early and late infection[60]. Biological consequences of this consensus genomic region will require further investigation.

Fourthly, our findings that relatively high "isolate *k*-mer count" in the rare subtypes (**Supplementary Fig. S1h**) and an independent subset in the group of the rare subtypes in phylogenetic trees constructed using enriched RNA *k*-mers (**Fig. 3j** and **3k**) unveil the genomic heterogeneity within the rare HIV-1 subtypes, supported by the fact that many of these rare subtypes are more frequently found in recombinant forms[28,61–65]. Although subtypes C and B are the most prevalent globally[53,54], the 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[66,67]. 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.

Altogether, in this work, we propose that distinct sequence patterns between HIV-1 DNA and RNA across group M HIV-1 subtypes exist based on enriched *k*-mers retrieved by PORT-EK-v2, and were verified by different downstream analyses. Results from machine learning-based modeling have verified that the enriched *k*-mer-associated feature, "isolate *k*-mer count", enables the classification of HIV-1 DNA and RNA genomic sequences (**Fig. 3a**) and the discrimination of different groups of HIV-1 subtypes (**Fig. 3b** and **3c**). Such intrinsic differences were further confirmed by the calculation of Euclidean distances (**Fig. 3d-3g**). Moving forward, the observations based on the graph-based analysis (**Supplementary Fig. S4**) and MCMC simulations (**Fig. 4**) have suggested that this intrinsic difference can also exist at the global level of network organization. Nevertheless, observations based on these different analyses lead us to hypothesize that the presence of intrinsic differences in HIV-1 DNA and RNA sequences across subtypes can be unveiled based on enriched *k*-mers.

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, which is one of the most prevalent HIV-1 subtypes, together with subtypes C, A, and B[68]. Given the complexity of the genomic composition of CRF01_AE, a putative recombinant between subtypes A and E[69], and the uncertainty of the origin of its parental subtype E[70,71], only the non-recombinant genomic sequences from group M HIV-1 subtypes were applied in this work, enabling a better examination of the feasibility of PORT-EK-v2 and the functionality of enriched *k*-mers. It is also important to note that due to the prevalence of HIV-1 subtype B in Latin America and the Caribbean[72,73], where viruses later spread to North America, as well as in Europe[74], a higher volume of subtype B genomic data is available in the Los Alamos HIV Sequence Database compared to other subtypes. While we cannot currently eliminate the possibility of a bias caused by the large volume of subtype B genomic sequences, which could impact result interpretation, our approach to finding *k*-mer frequency showed no obvious link

between input data size and enriched *k*-mer counts across different groups of HIV-1 subtypes. Nevertheless, a more meticulous review of artefactual *k*-mers (e.g., those resulting from a bias of input data size or sequencing error, and so on) together with other techniques will be expected to increase the reliability of the extracted *k*-mers. 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[75,76], and genotypic therapeutics. At present, PORT-EK-v2 has only been examined using multigenomic sequences from HIV-1 subtypes. In the future, it should also be tested using multigenomic sequences from a diverse array of species to determine its viability for a wider spectrum of organisms. It is also crucial to restate that PORT-EK-v2 was initially tailored for a high volume of small genomes and may not be ideal for large-scale genomic data. Nevertheless, 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, enabling the identification of new and emerging subtypes.

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.

The principle of the Pathogen Origin Recognition Tool using Enriched *K*-mers version 2 (PORT-EK-v2). PORT-EK-v2 is the upgraded version of PORT-EK[47], a pipeline that compares sequences from two different multigenomic datasets and identifies over-represented *k*-mers in respective datasets. The principle of *k*-mer-based approaches is to count the number of distinct substrings of length *k* in a string S[77], or a set of strings, where *k* is a positive integer representing a unique mark in an indicated string S[77]. In this study, we computed five features of enriched *k*-mers, including "*k*-mer average count", "*k*-mer root mean square deviation (RMSE)", "*k*-mer weight", "subtype *k*-mer count", and "isolate *k*-mer count", for better characteristics of enriched *k*-mer. Methodology for the calculation of each feature is detailed in **Supporting information** and the following section, wherever applicable. The Python pipeline and related R scripts for downstream analyses are available at GitHub https://github.com/Quantitative-Virology-Research-Group/PORT-EK-version-2. It is worth noting that we optimized the PORT-EK-v2 parameterization—the number of parameters has been reduced from 6, including k, c, m, min_{RESM}, f_c, and l_{map}, to 4, including k, d (which replaced m), *max_mem,* and *min_freq* (see **Supporting information** for the definition of each parameter). Details of each step involved in PORT-EK-v2 are described in **Supporting information**.

PORT-EK-v2 benchmark versus PORT-EK. Benchmark runs were performed on a desktop PC with AMD Ryzen 7 7700X 8-Core CPU and 32 GB of RAM, running Ubuntu 24.04.01 LTS WSL for Windows 11. Benchmark results are released in **Tables 1** and **2**.

Computation of the feature of enriched *k*-mers. Five features, including "*k*-mer average count", "*k*-mer RMSE", "*k*-mer weight", "subtype *k*-mer count", and "isolate *k*-mer count", were included in this study for a better characterization of enriched *k*-mers. The rationale of computing "*k*-mer average count" and "*k*-mer RMSE" has been detailed in **Supporting information**, together with the rationale of PORT-EK-v2.

K-mer weight: This attribute was designated to unveil the composition of enriched *k*-mers. Here, we applied ordinal encoding, which presents baseline encoding, to represent each of the four nucleotides (A, T, C, and G) as a number between 0 and 1. We first utilized the previous setting of ordinal encoding from the study of Wade et al. (2024)[43], whereby A is encoded as 0.25, T as 0.50, C as 0.75, and G as 1.00, and summed the weights for individual enriched *k*-mers. The "*k*-mer weight" is defined by the Equation below (**Fig. 1i**).

$$Kmer\,weight = \sum_{n=1}^{n} c(w,n) \tag{1}$$

, where n is the index of each nucleotide over a string of an enriched k-mer sequence, and $c(w,n)$ represents the assigned ordinal encoding value w in a specific nucleotide n. The computing code is written in R and available at GitHub (https://github.com/Quantitative-Virology-Research-Group/PORT-EK-version-2/blob/main/Analysis/Main.figures/Fig_1f_Kmer.weight.R). Alterations of ordinal encoding assigned to four nucleotides, as controls, were performed and illustrated in **Supplementary Fig. S2**.

Subtype k-mer count: We aggregated the count of individual k-mers across all isolates present in the same group of the HIV-1 subtype; values were then normalized by the total number of enriched k-mers present in the same group of the HIV-1 subtype. The "subtype k-mer count" is defined by the Equation below (**Fig. 1i**).

$$Subtype\,kmer\,count = 1/K \sum_{i=1}^{i} c(k,i) \tag{2}$$

, where K is the total number of enriched k-mers in an indicated group, i is the index over all isolates in the indicated group, and $c(k,i)$ represents the count of enriched k-mer k in an isolate i. The computing code is written in R and available at GitHub (https://github.com/Quantitative-Virology-Research-Group/PORT-EK-version-2/blob/main/Analysis/Main.figures/Fig_1g_subtype.kmer.count.R).

Isolate k-mer count: We aggregated k-mer counts by HIV-1 isolates present in the same group of HIV-1 subtype. The sum was normalized by the total number of isolates present in the same group of the HIV-1 subtype. The "isolate k-mer count" is defined by the Equation below (**Fig. 1i**).

$$Isolate\,kmer\,count = 1/N \sum_{k=1}^{k} C \tag{3}$$

, where N is the total number of isolates in an indicated group, and C is the summation of all enriched k-mer counts (to the k-th enriched k-mer) in an individual isolate. The computing code is written in R and available at GitHub (https://github.com/Quantitative-Virology-Research-Group/PORT-EK-version-2/blob/main/Analysis/Main.figures/Fig_1h_isolate.kmer.count.R).

Construction of prediction models. *Logistic regression*: To examine the prediction power in the classification of the origin of enriched k-mers (DNA versus RNA k-mers), 1,000 random enriched k-mers from each individual pool of enriched k-mers were bootstrapped with replacement and divided into a training set (80% of the dataset) and a testing set (20% of the dataset) for logistic regression running on R. The logistic regression model was fitted using the function glm() with the argument family specified as "na.exclude" and "binomial" in the R package "stats" (https://www.r-project.org/). Classifiers associated

with the five mentioned features (i.e., "*k*-mer average count", "*k*-mer RMSE", "*k*-mer weight", "subtype *k*-mer count", and "isolate *k*-mer count") were independently constructed; the "response" (i.e., the origin of enriched *k*-mers), coupled with a corresponding *k*-mer feature, was thus included in an object of class "formula", independently. Receiver operating characteristic (ROC) and the area under the curve (AUC) were calculated using the functions multiclass.roc() and auc() in the R package "pROC"[78], respectively. Given that there is no great variety of the input data for modeling, we have chosen ROC-AUC for evaluating a model's overall ranking ability across all thresholds. The whole procedure was repeated 1,000 times for statistical robustness. The computing code is written in R and available at GitHub (https://github.com/Quantitative-Virology-Research-Group/PORT-EK-version-2/blob/main/Analysis/Main.figures/Fig_1g_Kmer.count.isolate.R).

Multinomial logistic regression: To examine the prediction power in the classification of the distinct groups of HIV-1 subtypes based on enriched *k*-mers, 1,000 random enriched *k*-mers from each individual pool of enriched *k*-mers were bootstrapped with replacement and divided into a training set (80% of the dataset) and a testing set (20% of the dataset) for constructing multinomial logistic regression-based models associated with five features, running on R. The multinomial logistic regression model was fitted using the function multinom() in the R package "nnet"[79]. Classifiers associated with the five mentioned features were independently constructed. The predicted probability that enriched *k*-mers falling into distinct groups of HIV-1 subtypes, including subtypes A, B, C, D, and the rare subtypes, coupled with a corresponding *k*-mer feature, was computed using the function predict() with the argument family specified as type = "probs" in the R package "stats" (https://www.r-project.org/). Given that there is no great variety of input data for modeling, we have calculated the probability for evaluating the model's confidence level. Output probabilities were conserved in a dataframe. The whole procedure was repeated 1,000 times for statistical robustness. The computing code is written in R and available at GitHub (https://github.com/Quantitative-Virology-Research-Group/PORT-EK-version-2/blob/main/Analysis/Main.figures/Fig_3b_Kmer.multinomial.LG.subtypes.R).

Simple neural network: We performed simple neural network modeling to verify the prediction result obtained from multinomial logistic regression-based models. 1,000 random enriched *k*-mers from each individual pool of enriched *k*-mers were bootstrapped with replacement and divided into a training set (80% of the dataset) and a testing set (20% of the dataset) for the performance of simple neural network running on R. The model was fitted using the function neuralnet() in the R package "neuralnet"[80]. Classifiers associated with the five mentioned features were independently constructed. The whole procedure was repeated 10 times for statistical robustness. The computing code is written in R and available at GitHub (https://github.com/Quantitative-Virology-Research-Group/PORT-EK-version-2/blob/main/Analysis/Main.figures/Fig_3c_Kmer.NN.R).

It is important to note that the input data representing "isolate *k*-mer count" is a HIV-1 isolate-based matrix. While their balance was first managed during the PORT-EK-v2 processing phase for obtaining enriched *k*-mers, the inherent bias resulting from the unequal distribution of isolates among HIV-1 subtypes, restricted by current data availability, must still be acknowledged.

Calculation of the Euclidean distance between isolates. Euclidean distance is chosen based on the consideration that the magnitude (quantity) of the "isolate *k*-mer count" matters. Given that the length of a *k*-mer has been restricted and the network has been resampled to an identical size of 100 isolates for each group of HIV-1 subtypes, their influence is deemed to be negligible. To reduce computing cost, we bootstrapped 100 enriched *k*-mers and repeated 10 times to create 10 subpools of enriched *k*-mers in each group of HIV-1 subtypes. In each subpool of enriched *k*-mers, an edge weight between two isolates is calculated by the sum of "isolate *k*-mer count" normalized by the total number of isolates in the corresponding group of HIV-1 subtypes. The edge weight is defined by the Equation below.

$$w_{i,i'} = \Sigma(C_i + C_{i'})/N \tag{4}$$

, where $w_{i,i'}$ is the edge weight between isolate i and isolate i', $\Sigma(C_i + C_{i'})$ is the sum of the *k*-mer counts for isolate i and isolate i', respectively, and N is the total number of isolates in an indicated group.

The computed edge weight was assigned to individual isolates and was used to compute a matrix of the Euclidean distance, together with the parameter "isolate *k*-mer count" at an individual isolate level using the function dist() with the argument family specified as method = "euclidean" in the R package "stats" (https://www.r-project.org/). The clustering heatmap was visualized using the R package "ComplexHeatmap"[81,82]. The computing code is written in R and available at GitHub (https://github.com/Quantitative-Virology-Research-Group/PORT-EK-version-2/blob/main/Analysis/Main.figures/Fig_3e_3h_Euclidean.dist.heatmap.R).

Phylogenetic tree analysis. Unrooted phylogenetic trees were constructed from enriched *k*-mer count matrices using a distance-based approach. Optionally, the user can randomly subsample the dataset to a specified number of genomes, either uniformly or with balanced representation of multiple genomes. In this study, all datasets were downsampled to 50 genomes of each HIV-1 subtype, 250 genomes per tree. Pairwise Euclidean distances between all genomes' *k*-mer count vectors were calculated using *pdist* function from *scipy.spatial.distance* library. A distance matrix was then created with *Biopython's Phylo.TreeConstruction.DistanceMatrix* class. Finally, a phylogenetic tree was inferred from this matrix with *Biopython's Phylo.TreeConstruction.DistanceTreeConstructor* class, using the neighbor-joining (nj) method. The resulting tree object was written to Newick file format (.nwk) using the Phylo._io.write function from *Biopython.* The file was then read with the function read.tree() and visualized with the function ggtree() in the R package "ggtree"[83] with the %<+% operator to connect the annotation files corresponding to each phylogenetic tree.

Markov chain Monte Carlo modeling analysis. A concatenation of networks were constructed by either "isolate *k*-mer count" or "subtype *k*-mer count" across different groups of HIV-1 subtypes for simulation (see **Supporting information**). 10 networks constructed by presampling 100 isolates were computed; the readouts of pairwise correlation matrices were saved as a list of data frames. A list of data frames was then expanded by generating all possible pairs of vertices, representing either "isolate *k*-mer count" or "group *k*-mer count" using the function complete() in the R package "tidyr" (https://tidyr.tidyverse.org). Any pair of

vertices with the edge coefficient (i.e., summation of either "isolate *k*-mer count" or "group *k*-mer count" between two adjacent vertices) equal to zero was removed. Moving forward, a list of expanded data frames was reshaped from long to wide format using the function dcast() with the option "fun.aggregate = sum" in the R package "maditr" (https://github.com/gdemin/maditr). We further coordinated each vertex present in a list of expanded data frames with corresponding pairwise correlation matrices, enabling the reassignment of each vertex to respective groups of HIV-1 subtypes.

Five scenarios under the assumption, in which a random walk commences with an isolate or an enriched *k*-mer assigned to different groups of HIV-1 subtypes, were established. Based on the coordinate between each vertex and the corresponding group of HIV-1 subtypes, we forced a random walk beginning with the first appearing vertex in different groups of HIV-1 subtypes and completing within a total of 10,000 steps. In principle, in each step we first sought and listed all neighbor vertex that possess the edge coefficient greater than zero from the current vertex, and randomly selected one neighbor vertex from this list with equal probability using the R built-in function sample(). Lastly, we computed the probability of appearing vertices assigned to different groups of HIV-1 subtypes within a path in each scenario. The computing code is written in R and available at GitHub (https://github.com/Quantitative-Virology-Research-Group/PORT-EK-version-2/blob/main/Analysis/Main.figures/Fig_4b_4f_random.walk.isolate.R and https://github.com/Quantitative-Virology-Research-Group/PORT-EK-version-2/blob/main/Analysis/Main.figures/Fig_4g_4k_random.walk.subtype.R).

Data visualization. All plots were visualized using R (version 4.4.1) with the packages "ggplot2" (https://ggplot2.tidyverse.org), " ComplexHeatmap"[81,82], "ggtree"[83], "treeio"[84], and "igraph"[85].

Statistics. All statistical tests were performed with R with default options. Details are provided where appropriate in the main text.

Data availability. 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).

Code availability. 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. Struck, D., Lawyer, G., Ternes, A.-M., Schmit, J.-C. & Bercoff, D. P. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. *Nucleic Acids Res* **42**, e144 (2014).
33. Cleary, J. & Witten, I. Data compression using adaptive coding and partial string matching. *IEEE Trans. Commun.* **32**, 396–402 (1984).
34. Remita, M. A. *et al.* A machine learning approach for viral genome classification. *BMC Bioinformatics* **18**, 208 (2017).
35. Yu, C. *et al.* Real time classification of viruses in 12 dimensions. *PLoS One* **8**, e64328 (2013).
36. Huang, H.-H. *et al.* Global comparison of multiple-segmented viruses in 12-dimensional genome space. *Mol Phylogenet Evol* **81**, 29–36 (2014).
37. Li, Y., Tian, K., Yin, C., He, R. L. & Yau, S. S.-T. Virus classification in 60-dimensional protein space. *Mol Phylogenet Evol* **99**, 53–62 (2016).
38. Brady, A. & Salzberg, S. L. Phymm and PhymmBL: metagenomic phylogenetic classification with

interpolated Markov models. *Nat Methods* **6**, 673–676 (2009).

39. 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).
40. Tang, R. *et al.* Genetic source completeness of HIV-1 circulating recombinant forms (CRFs) predicted by multi-label learning. *Bioinformatics* **37**, 750–758 (2021).
41. 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).
42. 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).
43. 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).
44. Ma, Y. *et al.* Phylogenetic analysis of HIV-1 genomes based on the position-Weighted K-mers method. *Entropy (Basel)* **22**, 255 (2020).
45. 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).
46. 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).
47. Wiśniewski, J. & Chen, H.-C. Identification of potential SARS-CoV-2 genomic regions representing hallmarks for adaptation to different hosts. *iMetaOmics* **2**, (2025).
48. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
49. Deorowicz, S., Gudys, A., Dlugosz, M., Kokot, M. & Danek, A. Kmer-db: instant evolutionary distance estimation. *Bioinformatics* **35**, 133–136 (2019).
50. Zielezinski, A. *et al.* Ultrafast and accurate sequence alignment and clustering of viral genomes. *Nat Methods* **22**, 1191–1194 (2025).
51. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
52. 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).
53. Magiorkinis, G. *et al.* The global spread of HIV-1 subtype B epidemic. *Infect Genet Evol* **46**, 169–179 (2016).
54. Junqueira, D. M. & Almeida, S. E. de M. HIV-1 subtype B: Traces of a pandemic. *Virology* **495**, 173–184 (2016).
55. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
56. Malim, M. H., Hauber, J., Le, S. Y., Maizel, J. V. & Cullen, B. R. The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature* **338**, 254–257 (1989).
57. Hope, T. J., McDonald, D., Huang, X. J., Low, J. & Parslow, T. G. Mutational analysis of the human

immunodeficiency virus type 1 Rev transactivator: essential residues near the amino terminus. *J Virol* **64**, 5360–5366 (1990).

58. Kjems, J., Frankel, A. D. & Sharp, P. A. Specific regulation of mRNA splicing in vitro by a peptide from HIV-1 Rev. *Cell* **67**, 169–178 (1991).
59. Rana, T. M. & Jeang, K. T. Biochemical and functional interactions between HIV-1 Tat protein and TAR RNA. *Arch Biochem Biophys* **365**, 175–185 (1999).
60. Rosen, C. A. Tat and Rev: positive modulators of human immunodeficiency virus gene expression. *Gene Expr* **1**, 85–90 (1991).
61. 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).
62. 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).
63. 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).
64. 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).
65. 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).
66. 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).
67. Indihar, D. F. *et al.* Rare but relevant: a comprehensive narrative review of rare HIV-1 group M subtypes. *mBio* **16**, e0126425 (2025).
68. Serwin, K. *et al.* Molecular epidemiology and HIV-1 variant evolution in Poland between 2015 and 2019. *Sci Rep* **11**, 16609 (2021).
69. Li, X. *et al.* Tracing the epidemic history of HIV-1 CRF01_AE clusters using near-complete genome sequences. *Sci Rep* **7**, 4024 (2017).
70. Gao, F. *et al.* The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. *J Virol* **70**, 7013–7029 (1996).
71. Anderson, J. P. *et al.* Testing the hypothesis of a recombinant origin of human immunodeficiency virus type 1 subtype E. *J Virol* **74**, 10752–10765 (2000).
72. Osmanov, S. *et al.* Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. *J Acquir Immune Defic Syndr* **29**, 184–190 (2002).
73. Hemelaar, J., Gouws, E., Ghys, P. D., Osmanov, S. & WHO-UNAIDS Network for HIV Isolation and Characterisation. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS* **25**, 679–689 (2011).
74. Abecasis, A. B. *et al.* HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. *Retrovirology* **10**, 7 (2013).
75. 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).

76. 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.
77. Manekar, S. C. & Sathe, S. R. A benchmark study of k-mer counting methods for high-throughput sequencing. *Gigascience* **7**, (2018).
78. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
79. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S.* (Springer, New York, NY, 2010).
80. Fritsch, S., Guenther, F. & Wright, M. N. neuralnet: Training of Neural Networks. *CRAN: Contributed Packages* The R Foundation https://doi.org/10.32614/cran.package.neuralnet (2008).
81. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
82. Gu, Z. Complex heatmap visualization. *Imeta* **1**, e43 (2022).
83. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. Ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
84. Wang, L.-G. *et al.* Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Mol Biol Evol* **37**, 599–603 (2020).
85. Antonov, M. *et al.* Igraph enables fast and robust network analysis across programming languages. *arXiv [cs.SI]* (2023) doi:10.48550/ARXIV.2311.10260.
86. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
87. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
88. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
89. The pandas development team. *Pandas-Dev/pandas: Pandas*. (Zenodo, 2026). doi:10.5281/ZENODO.18328522.
90. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272 (2020).
91. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
92. Lau, B. T. *et al.* Profiling SARS-CoV-2 mutation fingerprints that range from the viral pangenome to individual infection quasispecies. *Genome Med* **13**, 62 (2021).
93. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
94. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a 'kneedle' in a haystack: Detecting knee points in system behavior. in *2011 31st International Conference on Distributed Computing Systems Workshops* (IEEE, 2011). doi:10.1109/icdcsw.2011.20.
95. Grinberg, D. An introduction to graph theory. *arXiv [math.HO]* (2023) doi:10.48550/ARXIV.2308.04512.

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 = 15 bp) and RNA *k*-mers (**c**, *k*-mer = 13 bp; **d**, *k*-mer = 15 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. The definition of each parameter can be found in **Methods** and **Supporting information**. (**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 DNA and RNA *k*-mers. 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 patterns 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 red-orange 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 enriched DNA and RNA *k*-mers. 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 common enriched *k*-mers appearing in 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 0.0001. (**b**) Box plots representing the probability of multinomial logistic regression models, constructed by five enriched *k*-mer-related features 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 red-orange 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 five enriched *k*-mer-related features to verify the predictive performance contributed by individual enriched *k*-mer-related features in the classification of different groups of HIV-1 subtypes, shown in (**b**). Significance levels are denoted as follows: ns for no significance, *p 0.05, ****p 0.0001. (**d-g**) Clustering heatmaps representing the Euclidean distance 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 sequence barrier between HIV-1 DNA and RNA. (**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 assumptions 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 within a path of 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 represent enriched *k*-mers to different groups of HIV-1 subtypes within a path of each scenario, whereby a random walk initiates a vertex assigned to HIV-1 subtype A (**g**), B (**h**), C (**i**), D (**j**), and the rare subtypes (**k**). Networks were constructed by "subtype *k*-mer count". Facets at the x-axis separate

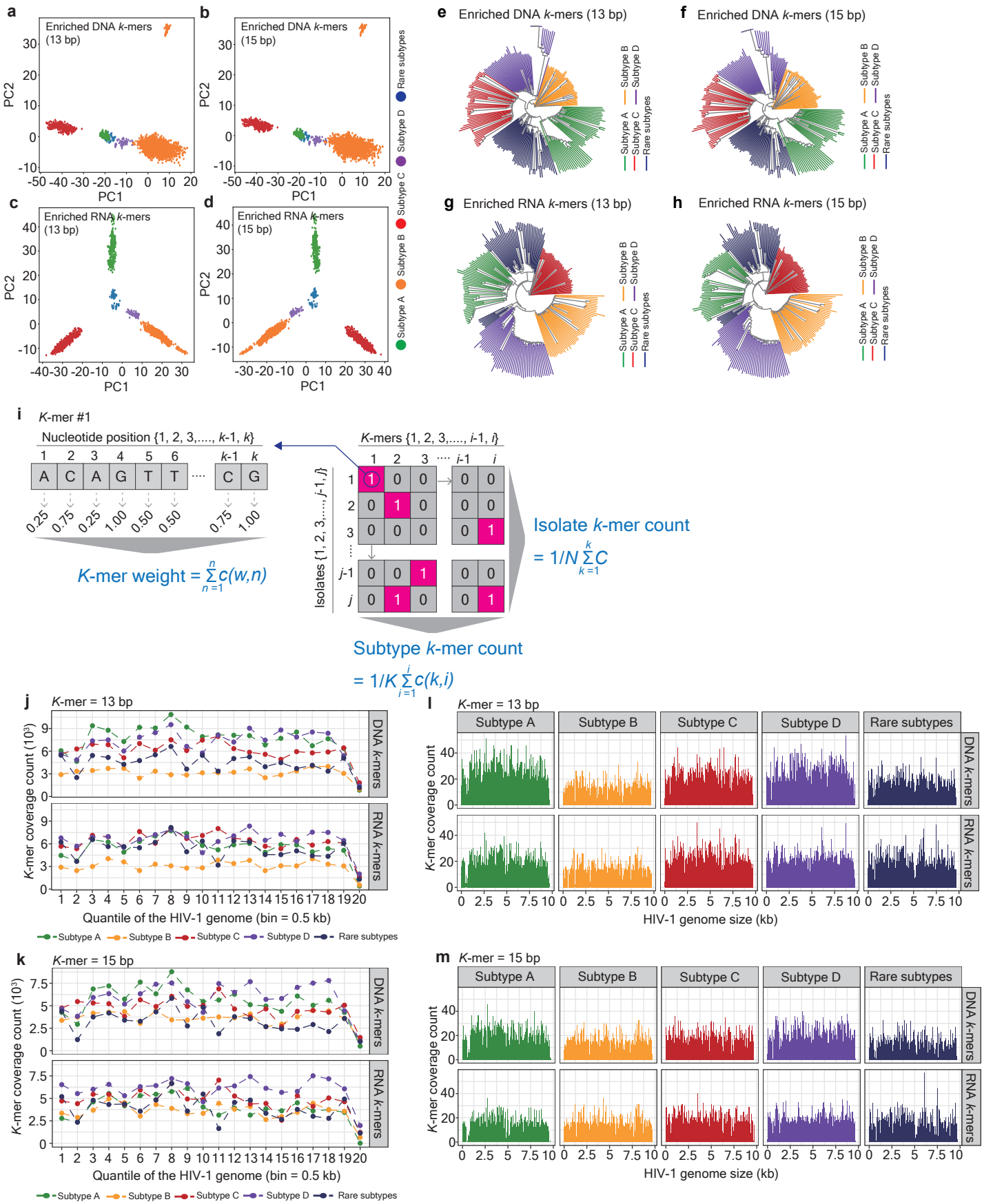
enriched DNA and RNA *k*-mers; facets at the y-axis separate a *k*-mer length between 13 bp and 15 bp. Boxes marked in green, orange, red, purple, and dark blue represent HIV-1 subtype A, B, C, D, and rare subtypes, respectively. Significance levels are denoted as follows: ns for no significance, **p* 0.05, ***p* 0.01, ****p* 0.001.

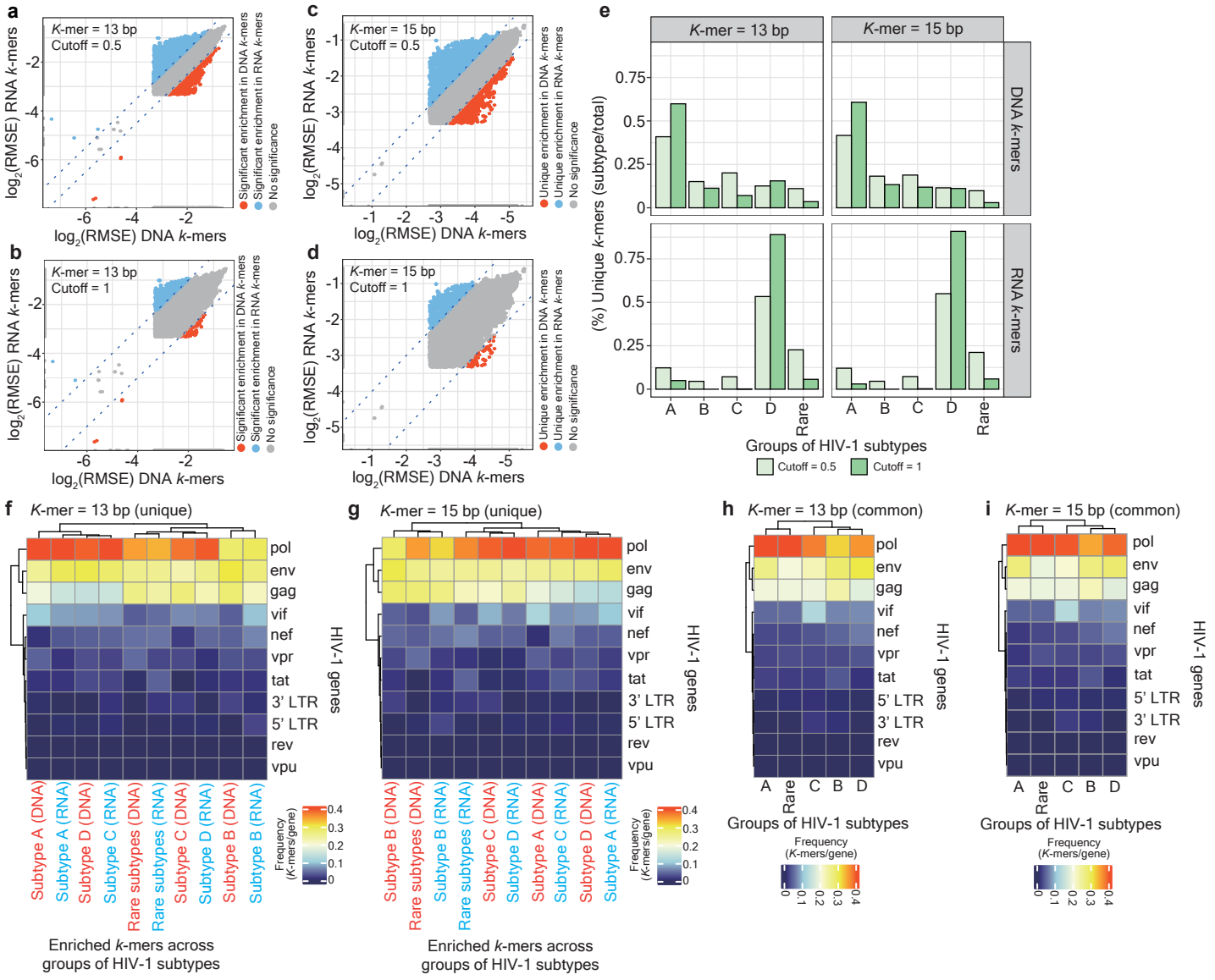
Table 1. Benchmarking results of PORT-EK-v2 versus PORT-EK.

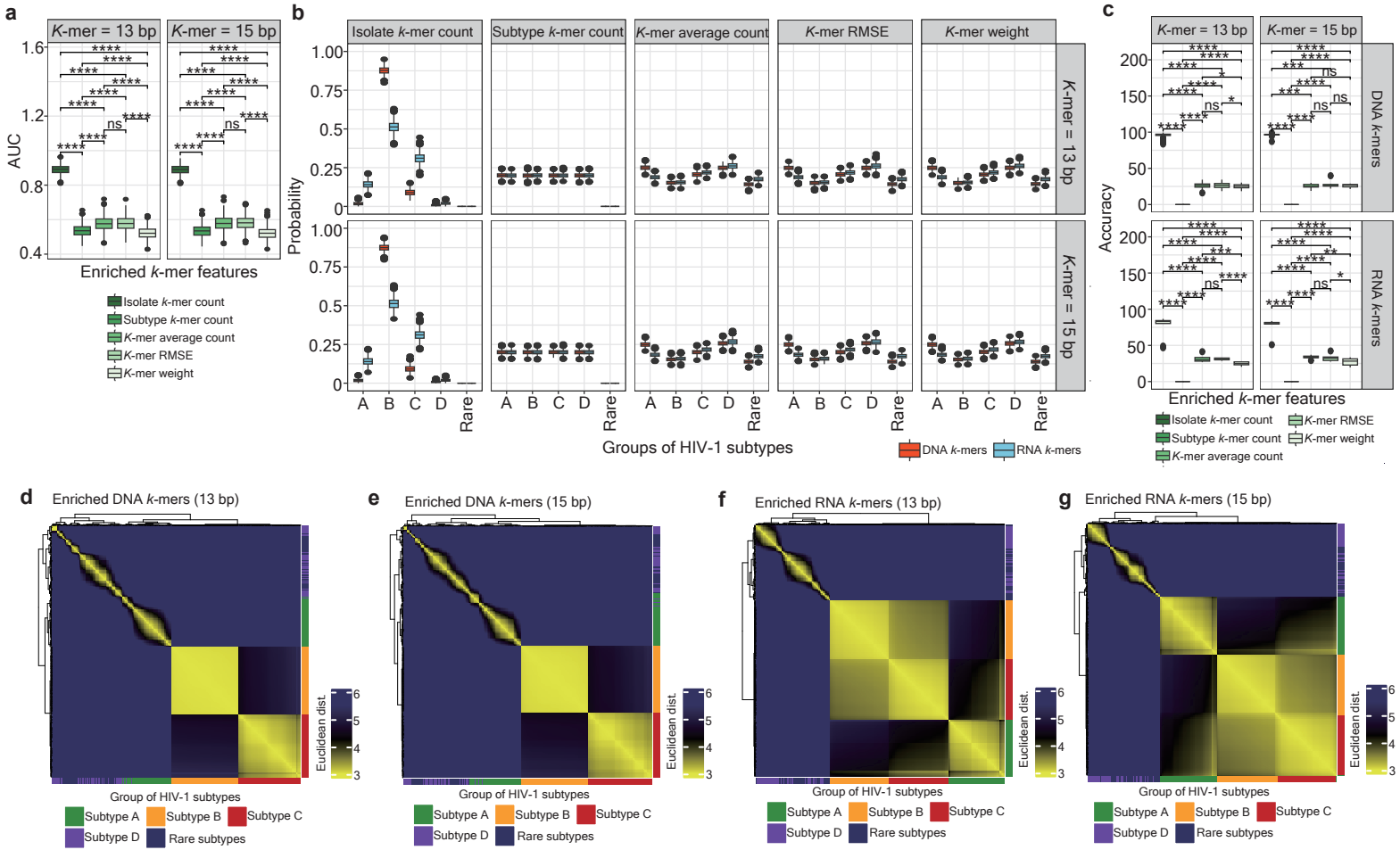
Dataset (size)	PORT-EK[47] (*K*-mer = 15 bp)		PORT-EK-v2 (*K*-mer = 15 bp)	
	Total wall time [hh:mm:ss]	**Peak memory usage (RSS) [GB]**	**Total wall time [hh:mm:ss]**	**Peak memory usage (RSS) [GB]**
HIV-1 RNA (5,490 genomes)	01:01:27	26.4	00:06:50	4.7
Downsampled HIV-1 RNA set 1 (2,000 genomes)	00:22:43	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

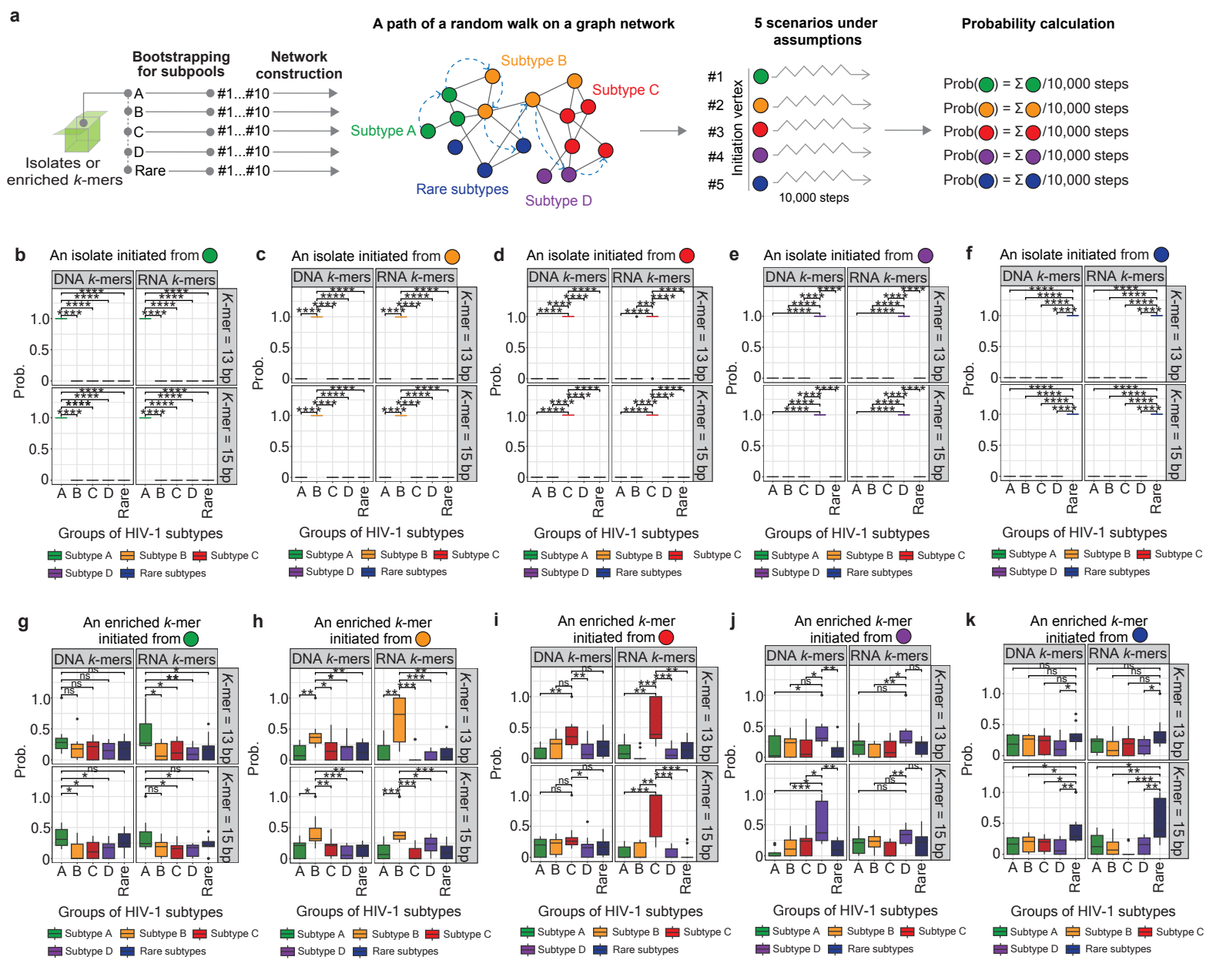
Table 2. Benchmarking results of *k*-mer counting speed using PORT-EK-v2, Jellyfish, Kmer-db, and KMC 3 based on HIV-1 RNA sequences (5,490 genomes).

Software	PORT-EK[47]	PORT-EK-v2	Jellyfish[48]	Kmer-db[49,50]	KMC 3[51]
***K*-mer counting time [mm:ss]**					
***K*-mer = 13 bp**	00:53	00:41	00:43	00:036	00:47
***K*-mer = 15 bp**	00:53	00:42	00:10	00:034	00:46


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 = $\sum_{n=1}^{n} c(w,n)$
K-mers {1, 2, 3,...., i-1, i}
Isolates {1, 2, 3,...., j-1, j}
Isolate k-mer count
$= 1/N \sum_{k=1}^{k} C$
Subtype k-mer count
$= 1/K \sum_{i=1}^{i} c(k,i)$
j
K-mer = 13 bp
K-mer coverage count (10^3)
DNA k-mers
RNA k-mers
Quantile of the HIV-1 genome (bin = 0.5 kb)
Subtype A
Subtype B
Subtype C
Subtype D
Rare subtypes
l
K-mer = 13 bp
Subtype A
Subtype B
Subtype C
Subtype D
Rare subtypes
K-mer coverage count
HIV-1 genome size (kb)
k
K-mer = 15 bp
m
K-mer = 15 bp


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
No 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
Rare
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


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
Accuracy
DNA k-mers
RNA k-mers
Enriched k-mer features
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


a
Bootstrapping for subpools
Network construction
Isolates or enriched k-mers
A
B
C
D
Rare
#1...#10
A path of a random walk on a graph network
Subtype A
Subtype B
Subtype C
Subtype D
Rare subtypes
5 scenarios under assumptions
#1
#2
#3
#4
#5
Initiation vertex
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.
Groups of HIV-1 subtypes
Subtype A
Subtype B
Subtype C
Subtype D
Rare subtypes