



Article

Identification of HIV Rapid Mutations Using Differences in Nucleotide Distribution over Time

Nan Sun ¹, Jie Yang ²  and Stephen S.-T. Yau ^{1,3,*} 

¹ Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China; sunn19@mails.tsinghua.edu.cn

² Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA; jyang06@uic.edu

³ Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing 101408, China

* Correspondence: yau@uic.edu

Abstract: Mutation is the driving force of species evolution, which may change the genetic information of organisms and obtain selective competitive advantages to adapt to environmental changes. It may change the structure or function of translated proteins, and cause abnormal cell operation, a variety of diseases and even cancer. Therefore, it is particularly important to identify gene regions with high mutations. Mutations will cause changes in nucleotide distribution, which can be characterized by natural vectors globally. Based on natural vectors, we propose a mathematical formula for measuring the difference in nucleotide distribution over time to investigate the mutations of human immunodeficiency virus. The studied dataset is from public databases and includes gene sequences from twenty HIV-infected patients. The results show that the mutation rate of the nine major genes or gene segment regions in the genome exhibits discrepancy during the infected period, and the *Env* gene has the fastest mutation rate. We deduce that the peak of virus mutation has a close temporal relationship with viral divergence and diversity. The mutation study of HIV is of great significance to clinical diagnosis and drug design.



Citation: Sun, N.; Yang, J.; Yau, S.S.-T. Identification of HIV Rapid Mutations Using Differences in Nucleotide Distribution over Time. *Genes* **2022**, *13*, 170. <https://doi.org/10.3390/genes13020170>

Academic Editors: Kenta Nakai and Tun-Wen Pai

Received: 28 November 2021

Accepted: 12 January 2022

Published: 19 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: mutations; natural vector; human immunodeficiency virus; nucleotide distribution difference

1. Introduction

Human immunodeficiency virus (HIV) is a highly variable virus and has a high replication rate in participants, which leads to many quite different genetic variations of the viruses [1]. The HIV-1 polymerase is very error-prone, with the net result to generate escape mutations on newly generated viruses (evolution) undetected by the immune system [2]. The two types of HIV—HIV-1 and HIV-2—will cause acquired immunodeficiency syndrome (AIDS) [3], but HIV-1 is more virulent and infective than HIV-2 [4]. It is the cause of most HIV infections globally, and thousands of infected people appear every year [5]. HIV-1 is composed of two copies of positive-sense single-stranded RNA, and the genome length is about 9.2–9.8 kb (RefSeq accession number in GenBank is NC_001802, sequence length is 9181 bp). There are nine major genes, including three structural genes—*Gag*, *Pol*, and *Env*; two regulatory genes—*Tat* and *Rev*; and four auxiliary genes—*Nef*, *Vpr*, *Vpu*, and *Vif*. The mutation rate of each gene in the HIV-1 genome sequence is different, and the *Env* gene mutation rate is the highest [6]. The mutation rate detection of the gene segment or whole genome has important guiding significance for HIV monitoring, diagnosis, vaccine, and drug treatment.

CD4 count and viral load matter a great deal to measure patients' disease progression. HIV mainly attacks CD4+ T cells [7], macrophages and dendritic cells [8] using the CD4 receptor as a docking site [9]. The infection of HIV results in a gradual change in the number of CD4-expressing T cells [10], which can be measured by CD4 count. Virus load is another important factor to assess the immune system. At the acute infection stage, the

virus rapidly propagates, the virus content in each milliliter of blood can reach millions, and CD4 count will also decrease significantly. After the stage of acute HIV infection, the strong response of the immune system inhibits the activity of the virus and reduces the amount of virus in the blood. The length of incubation period is affected by many factors, ranging from 3 years to 20 years [11,12]. Usually, once the number of CD4 per microliter of blood is less than 200 [13], this means that the immune system is almost compromised and the human body can no longer effectively deal with many common infections. Patients will suffer from AIDS and soon die of cancer [14,15]. Therefore, the relationship of virus mutation, viral load and CD4 count is of great importance.

Mutations may create new HIV quasi-species, so comparison of the mutational nucleotide sequence with the original sequence is necessary. Methods to compare sequences include alignment and alignment-free methods. Alignment methods contain pairwise or multiple sequence alignment [16,17]. Alignment-free methods include power spectrum [18–20], k-mer theory [21], the density-based method [22], and the natural vector method [23], among others [24]. Natural vector (NV) was proposed to classify viral genomic sequences by Yau's team in 2013 [23], and has been successfully applied to many studies [25–27]. Natural vector characterizes the distribution of four nucleotides in the genome, including their counts, mean positions and second central moments of location. Mutations will lead to changes in nucleotide distribution, which inspires us to use the difference in natural vector over time to measure mutations globally.

In this paper, we present a mathematical formula of nucleotide distribution difference over time to investigate the mutations. This is defined as the difference in natural vectors of the two nucleotide sequences divided by the corresponding difference in time points. We use the equation to explore the gene mutations of HIV. The studied dataset includes gene sequences from twenty HIV-1-infected patients. The results show that the mutation rate of the main genes or segment regions in the genome exhibits discrepancy during the infected period. There is a consistent pattern in the temporal sense among the mutation rate of *Env* gene sequences, viral divergence, and viral diversity. This study provides a meaningful and advanced tool to study mutations.

2. Materials and Methods

2.1. The Traditional 12-Dimensional Natural Vector

Natural vector is a powerful method to characterize the statistical features of biological sequences. The definition is as follows. Suppose the genomic sequence is $S = s_1 s_2 s_3 \dots s_n$ with length n . For $k \in L = \{A, C, G, T/U\}$, the indicator function is

$$w_k(s_i) = \begin{cases} 1, & \text{if } s_i = k \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $s_i \in L, i = 1, 2, 3, \dots, n$. Then, the distribution of four nucleotides can be described by a 12-dimensional natural vector:

$$\left(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_2^A, D_2^C, D_2^G, D_2^T \right), \quad (2)$$

n_k denotes the count of nucleotide k within sequence S :

$$n_k = \sum_{i=1}^n w_k(s_i), \quad (3)$$

μ_k specifies the average location of nucleotide k within sequence S :

$$\mu_k = \sum_{i=1}^n i \frac{w_k(s_i)}{n_k}, \quad (4)$$

D_2^k is the second central moment of positions of nucleotide k within sequence S :

$$D_2^k = \sum_{i=1}^n \frac{(i - \mu_k)^2 w_k(s_i)}{n_k n}, \tag{5}$$

Here, we give an example how to calculate the vector. If the sequence is ACTGC-TATGA, the indicator functions are $w_A = 1000001001$, $w_C = 0100100000$, $w_G = 0001000010$, and $w_T = 0010010100$. Each component of the vector is as follows:

- $n_A = 3, n_C = 2, n_G = 2, n_T = 3$
- $\mu_A = \frac{1+7+10}{3} = 6, \mu_C = \frac{2+5}{2} = \frac{7}{2}, \mu_G = \frac{4+9}{2} = \frac{13}{2}, \mu_T = \frac{3+6+8}{3} = \frac{17}{3}$.
- $D_2^A = \frac{(1-6)^2 + (7-6)^2 + (10-6)^2}{3 \cdot 10} = \frac{21}{15}, D_2^C = \frac{(2-\frac{7}{2})^2 + (5-\frac{7}{2})^2}{2 \cdot 10} = \frac{9}{40},$
 $D_2^G = \frac{(4-\frac{13}{2})^2 + (9-\frac{13}{2})^2}{2 \cdot 10} = \frac{5}{8}, D_2^T = \frac{(3-\frac{17}{3})^2 + (6-\frac{17}{3})^2 + (8-\frac{17}{3})^2}{3 \cdot 10} = \frac{19}{45}.$

Then, the 12-dimensional natural vector is:

$$\left(3, 2, 2, 3, 6, \frac{7}{2}, \frac{13}{2}, \frac{17}{3}, \frac{21}{15}, \frac{9}{40}, \frac{5}{8}, \frac{19}{45} \right).$$

2.2. Measure the Changes of Nucleotide Distribution over Time Based on Natural Vector

Sequence mutations may change the genetic information of organisms and obtain selective competitive advantages through natural selection to adapt to environmental changes. This is the driving force of species evolution. All life activities of organisms are related to proteins. Gene mutations may change the structure or function of translated proteins, which may result in abnormal cell operation, a variety of diseases and even cancer [1]. Therefore, it is particularly important to identify gene regions with high mutations.

Counting the number and type is an important method to detect mutations. Mutations will cause changes in nucleotide distribution. Another intuitive idea is to use the change in natural vector over time to measure the mutations globally. Suppose sequence S_1 is examined at time point D_1 ; it mutates over a period of time and becomes sequence S_2 at time point D_2 . Both sequences are transformed into the corresponding natural vectors NV_1 and NV_2 first. Then, the nucleotide distribution difference over time (NDDT) of the two sequences is described as:

$$NDDT = \frac{\|NV_2 - NV_1\|_2}{D_2 - D_1}. \tag{6}$$

where $\|\cdot\|_2$ is the l_2 - norm, which has commonly been used in previous NV studies. We use the formula to measure the HIV mutation rates and will verify the rationality of the definition in the results part.

2.3. Validation Dataset

The first longitudinal sequence dataset covers the whole genomes of eleven HIV-1-infected patients without therapy in Sweden, with long-term follow-up from 1990 to 2003 [28] (<https://hiv.biozentrum.unibas.ch>, accessed on 27 June 2021). These data were sampled with 5–12 time points. Patient 10 was removed from the dataset due to a shortage of time points. Besides the complete genome of each patient, clinical data, including CD4 count, viral load and the position of each gene in the genome sequence, are available. Ten patients' information is presented in supplementary dataset 1.

The second longitudinal sequence dataset includes 1337 sequences of a patient at 17 time points during the first three years of infection [29–31] (https://www.hiv.lanl.gov/content/sequence/HIV/SL_alignments/set10.html, accessed on 25 August 2021). The sequences belong to eight sequence segment regions: *Nef*, p17, p24, RT, *Vpu*, gp41, *Vpr*, and *Tat*. Detailed information can be found in supplementary dataset 2.

The third longitudinal dataset includes 1032 *Env* gene sequences, with an average of 12 sampling time points per person (<https://www.hiv.lanl.gov/content/sequence/>

[HIV/SI_alignments/datasets.html](#), accessed on 27 June 2021). The C2-V5 region of *Env* is used for research because it plays an important role in encoding the target of immune responses and shows a high degree of variation [6,32,33]. The sequences belong to nine HIV-positive participants, who were tracked over a 6~12 years period starting at the time of seroconversion. These participants were all homosexual men enrolled in the Multicenter AIDS Cohort Study (MACS, <http://aidscohortstudy.org>, accessed on 27 June 2021) [34]. The statistical information of the sampling time points and the *Env* gene sequence number of these 1032 sequences are shown in supplementary dataset 3. The virus has been evolving constantly in each infected patient. There might exist different virus variations at the same time point, and more than one sequence was obtained at a given time point. Seven men were treated during this study. Participant 8 did not take any antiretroviral therapy during the whole study. Participant 11 was the slowest progressor and did not receive treatment either. Participant 9 was another typical non-progressor but progressed subsequently [6].

3. Results

3.1. NDDT Comparison of the Nine Major Genes in HIV-1 Genome

The exploration of nucleotide distribution difference over time of the nine main genes of the HIV-1 genome—*Gag*, *Pol*, *Env*, *Tat*, *Rev*, *Nef*, *Vpr*, *Vpu*, and *Vif*—is of great significance. Note that most genes are composed of multiple sequence segments in dataset 1: *Env* includes gp120 and gp41; *Gag* includes P17, P24, P2, P7, P1 and P6; *Pol* includes PR, RT, P15 and IN; *Tat* and *Rev* consist of two segments, respectively; and the remaining four genes—*Vif*, *Vpr*, *Vpu*, and *Nef*—only have one segment, respectively.

We first parse each segment sequence from the genome according to their position records and assemble the segments of each gene together; we then transform each gene sequence into a 12-dimensional natural vector. We calculate the NDDTs of the nine genes for each patient, and show the results in Figures S1–S10. The horizontal axis represents the time from infection, and the red dot represents the nucleotide distribution difference of *Env* gene over time. The figures show that the mutation rate of the *Env* gene is the highest among almost all time periods. To better describe this fact, we take the average of the mutation rates of the nine genes during the whole study period and exhibit the results in Figure 1. Red bars indicate the mutation rate of *Env*, which is the highest compared with those of other genes. *Gag*, *Pol*, *Nef*, and *Vpu* also show high mutation rates.

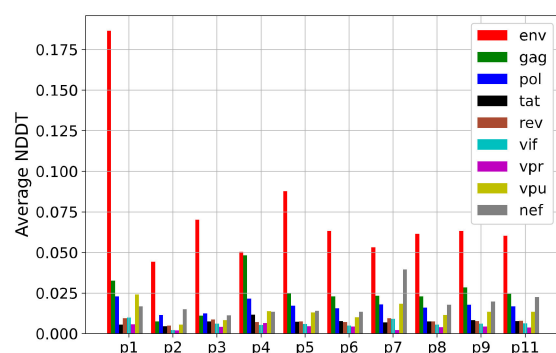


Figure 1. The average nucleotide distribution difference of the nine genes for each patient over all time periods.

To better visualize the mutation distributions and frequencies of virus sequences over time in the same patients, we extend the Mutation Tracker method illustrated in [35,36] to present them. For the data of each patient, multiple gene sequences are aligned by MUSCLE [37,38], and a sequence at the first time point is regarded as the reference sequence. The aligned gene sequences are re-positioned and compared according to reference sequence. Gaps caused by insertion and deletion are also taken into account. Then, the mutation diversity of gene sequences is explored. Here, we assemble the mutation profiles of all genes without considering the noncoding region (the locations of the nine genes in the genomic se-

quence are presented in Figure 2a) and show the results in Figures 2b,c and S11a,b–S19a,b. From the density of the dots in the figures, the changes in nucleotide distribution of the *Env* gene are the largest (the yellow dots are the densest when observed along the horizontal axis of Figures 2b and S11a) and the mutation frequencies are the highest (the yellow bars corresponding to high mutation frequency are higher than others when observed along the longitudinal axis of Figures 2c and S11b–S19b) compared with the reference sequence. This further validates the reliability of the conclusions obtained by NDDT.

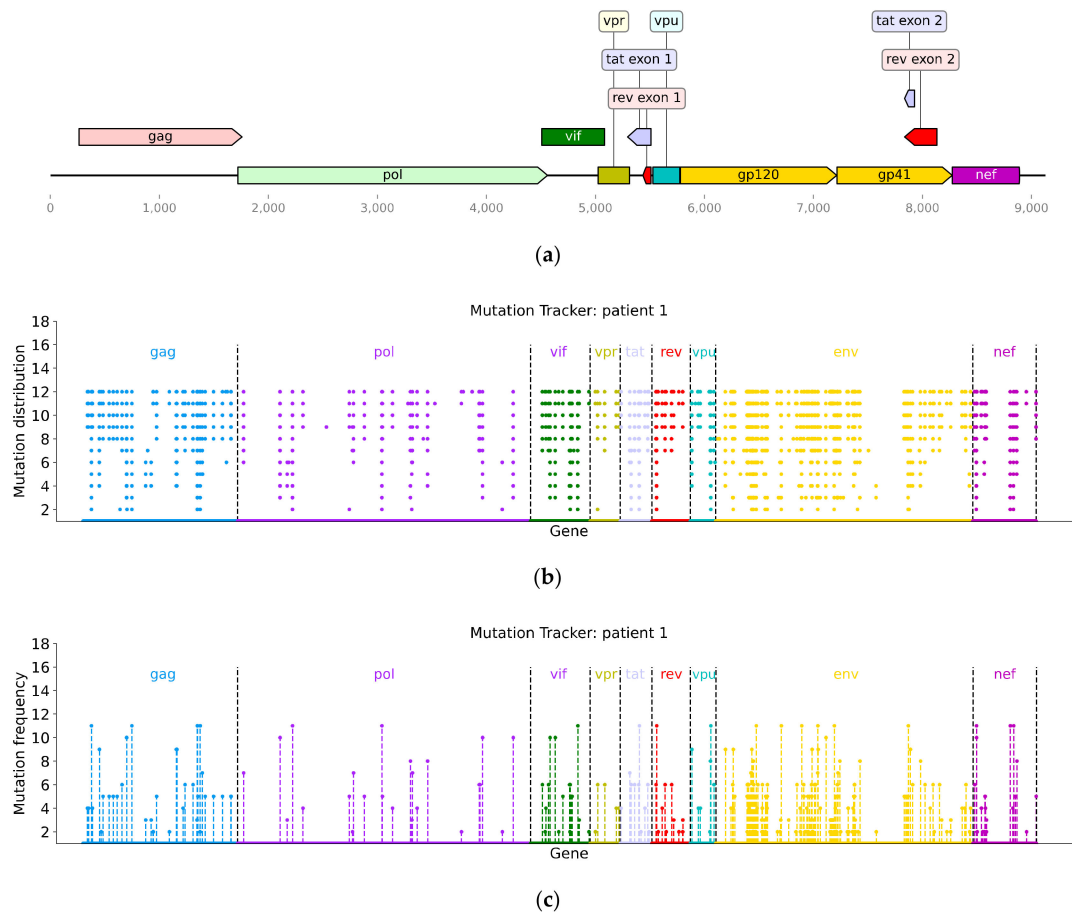


Figure 2. (a) Structure of the RNA genome of HIV-1. The structure is drawn through the DNA viewer (<https://pypi.org/project/dna-features-viewer/>, accessed on 17 April 2021); (b) Single nucleotide mutation distributions of sequences at 12 time points. The horizontal axis represents the gene position, and the vertical axis represents the mutations of the mutant sequence compared with the reference sequence (the sequence at the first time point). For example, the scale “1” means the sequence at the first time point, and the scale “2” means the sequence at the second time point; the scale “12” means the sequence at the 12th time point; (c) Single nucleotide mutation frequencies of sequences at 12 time points. The horizontal axis represents the gene position, and the vertical axis represents the number of sequences mutated at the corresponding gene site. The sequence at the first time point is regarded as the reference sequence, and the mutation profiles of all genes are assembled without considering the noncoding region.

3.2. NDDT Study of Eight Sequence Segment Regions

Furthermore, we select eight sequence segment regions (*Nef*, p17, p24, RT, *Vpu*, gp41, *Vpr*, *Tat*) from the nine major genes to study the mutations. The data contain 1337 sequences provided by dataset 2. In Figure 1, *Gag*, *Pol*, *Nef* and *Vpu* show a higher average mutation rate, and *Vpr* and *Tat* show a lower average mutation rate. This result is also reflected in Figure 3a (in which segments p17 and p24 belong to *Gag* and segment RT belongs to *Pol*). The mutation rate at each time point is shown in Figure S20. Figure 3b,c further verify it by

using mutation distribution and frequency. According to the sampling time points of the dataset, the nucleotide distribution changes of *Gag*, *Pol*, *Nef*, *Vpu*, *Vpr* and *Tat* in the first three years of infection have a similar pattern as those in the whole infection period.

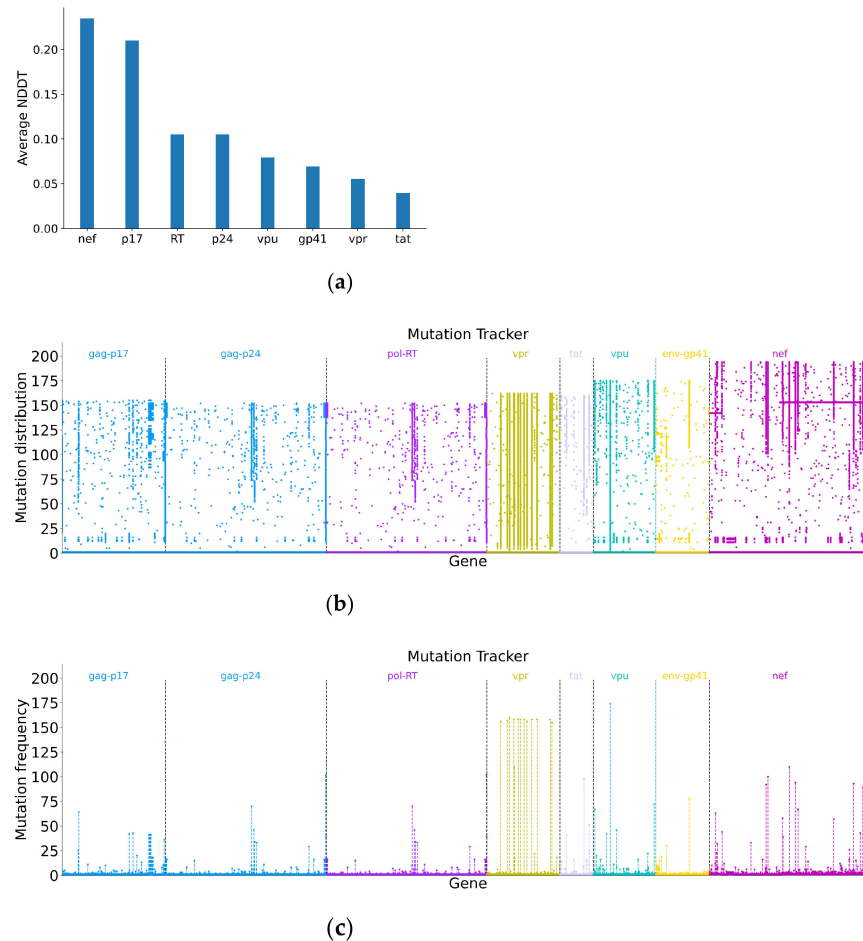


Figure 3. (a) The average nucleotide distribution difference of the eight sequence segments for patients in dataset 2 over all time periods; (b) Single nucleotide mutation distributions of the sequence segments at all time points; (c) Single nucleotide mutation frequencies of the sequence segments at all time points. The sequence at the first time point is regarded as the reference sequence.

In addition, we extract the eight sequence segments of all sequences in dataset 1, calculate the average NDDT of the eight sequence segments for each patient during the whole sampling time period, and analyze the distribution and frequency of the mutation, as shown in Figures S21–S30. In the whole infection stage, gp41 sequences of most patients show the highest mutation rate; *Nef*, *Vpu*, RT, p17 and p24 also show relatively higher average mutation rate, while *Vpr* and *Tat* show lower average mutation rate.

3.3. The Temporal Relationships among Nucleotide Distribution Changes, Viral Divergence, and Diversity

Env is an HIV gene, whose encoded protein forms the viral envelope. The *Env* gene codes for the gp160 protein, and is cut into glycoproteins gp120 and gp41. gp120 binds to the CD4 receptor on the target cell, and make the virus infiltrate the cell to bind to the co-receptor CXCR4 or CXCR5. gp41 provides the second step to let the virus enter the host cell through the target cell membrane. The first targets of HIV vaccine research are gp120 and gp41. It is vital to have a comprehensive insight into the mutation pattern of *Env* [6].

The binding of the virus and the CD4 receptor is the most obvious step during the HIV infection process. CD4 is a glycoprotein, “+” indicates positive, and CD4+ T cells

are an essential part of the human immune system. If CD4 cells are depleted, the body is vulnerable to infection. It makes the relationship among virus mutation, CD4+ count and viral load become an important research direction. We explore this relationship using dataset 3.

For participant 1, there are fifteen sampling time points—3, 14, 24, 34, 45, 51, 61, 66, 68, 77, 80, 87, 94, 98, and 105—whose unit is a month. There is more than one nucleotide sequence at each time point, and the length of the sequences at the same time point may be different. We first convert each sequence into a natural vector and then calculate NDDT. In this way, we obtain a mutation matrix and only consider the mutation sub-matrices of adjacent time points. Next, we take the average of the sub-matrix and obtain a mean mutation rate between time point D_1 and time point D_2 . For each participant, we calculate the mean mutation rate between time points D_1 and D_2 , and plot it at D_2 (Figure 4a).

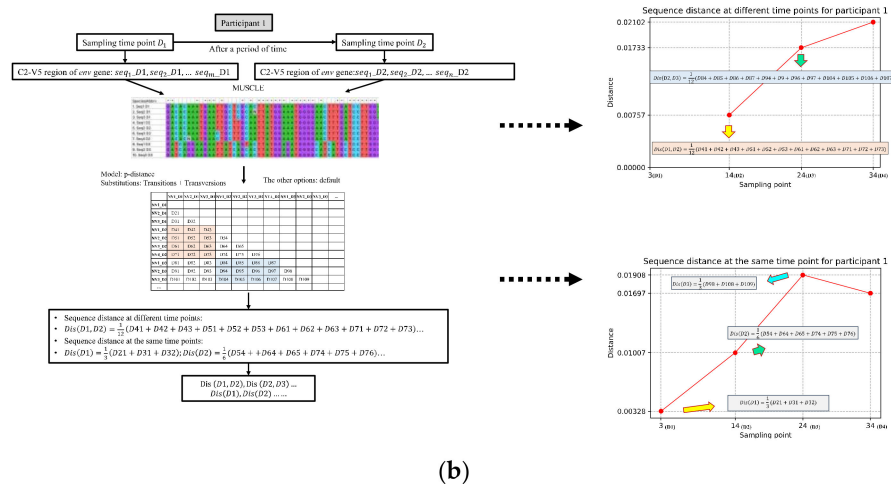
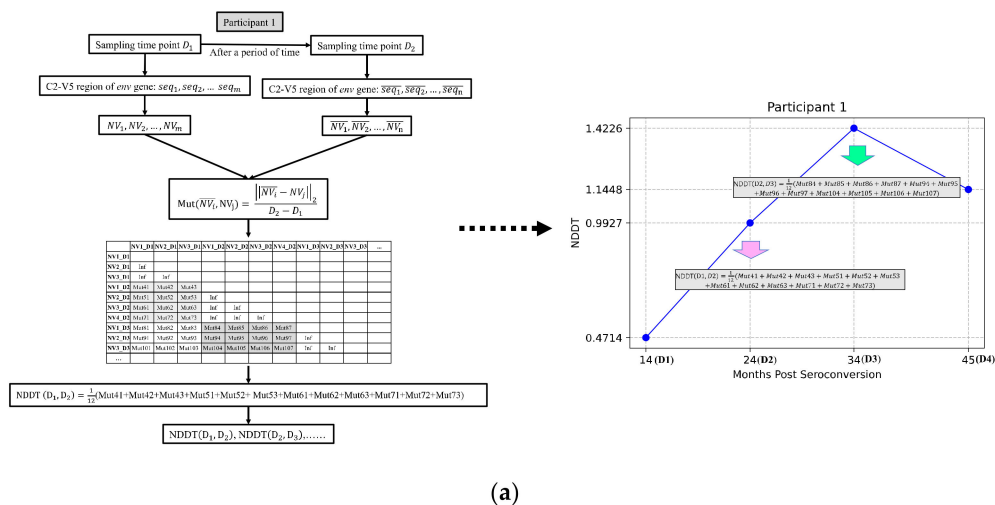


Figure 4. (a) Flowchart for computing the mean NDDT between two adjacent time points D_1 and D_2 ; (b) Flowchart for computing the viral divergence and diversity between two adjacent time points D_1 and D_2 .

Figures 5a and S31a–S38a show the variation trends of the C2-V5 region of the *Env* gene for the nine participants. Participants 4 and 10 are excluded for lack of specimens. The horizontal axis represents the time post seroconversion, and the three vertical axes represent the mean NDDT, CD4 count, and viral load, respectively. This provides a pattern in which viruses change over time. From the mutation trend plots (except participants 9 and 11), the NDDTs increase for several years post seroconversion, then reach a peak or several peaks, and appear to gradually slow down or decrease in the late stage of infection. The average

time to the best peak is 58 months post seroconversion for all participants. Before reaching the best peak, the NDDT for each individual increases linearly (mean $R^2 = 0.55$, and mean $slope = 0.07$, except participants 9 and 11 because their disease progressions were slow and their diseases were special cases); after the best peak, it declines linearly (mean $R^2 = 0.705$, and mean $slope = -0.9$, except participants 9 and 11). These facts suggest that the gene sequences relationship shows a progressive evolutionary trend, and the variation in the sequences is time-ordered, which is strong at the initial stage of infection, and then becomes weaker with the development of AIDS.

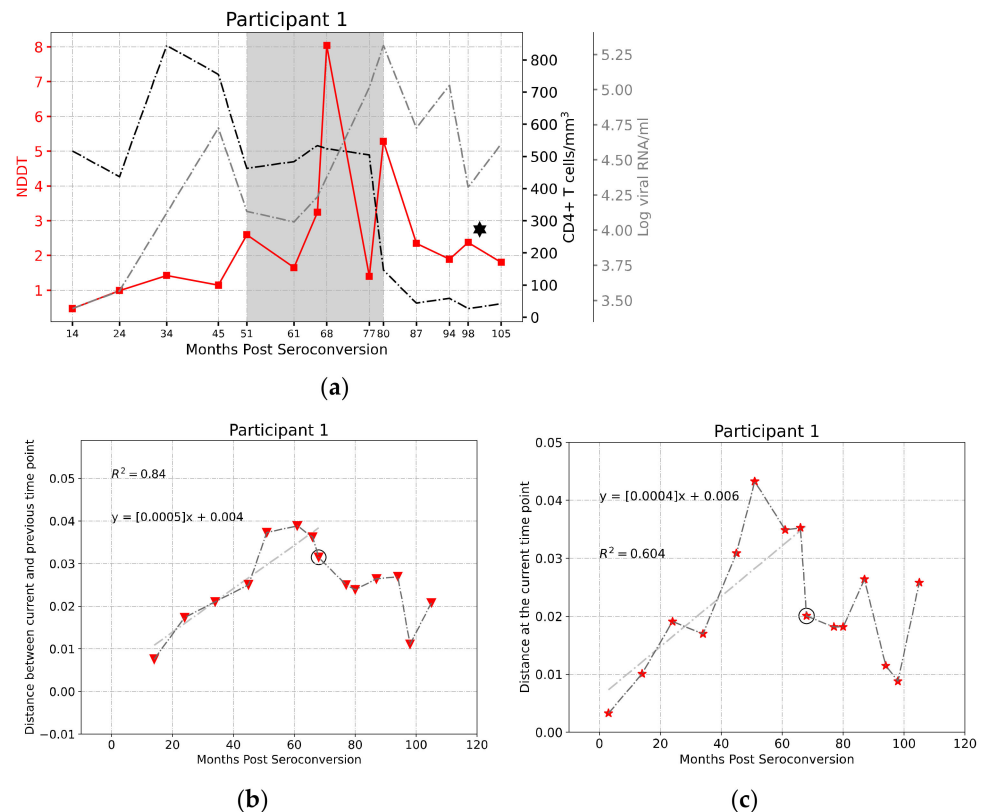


Figure 5. (a) The variation trends of C2-V5 region of HIV-1 *Env* gene for participant 1. The mutation progressions are shown with the filled blocks connected by the colorful line. The CD4+ T cell levels are shown with the black dotted line. The viral RNA levels are shown with the gray dotted line. Five participants (1, 3, 5, 6, 7) died after the last time point of analysis (marked with diamond), and their CD4+ T cell levels dropped to 200 cells/ μ L; (b) Viral divergence: Distance of sequences between the current time point and its previous time point; (c) Viral diversity: Distance of sequences at the current time point. The abscissa of the circle represents the time point of the mutation peak.

The above results are consistent with the changes in T-cell levels and viral counts. Participants 1 and 2 had high CD4+ T cell levels at the infection stage, and their NDDTs change very slowly. Participants 9 and 11 were the slowest disease progressors. The CD4 count of participant 9 was still over 200 cells/ μ L at the last study time point whose antiretroviral therapy began when the CD4 measure was 202 cells/ μ L. In the meanwhile, the mutation rate of participant 9 goes down during the initial 63 months (~5.5 years) post seroconversion, and upward after that until the last time point, which suggests that this participant was going through the disease stage. Participant 11 never received any antiretroviral therapy during all study time points, and had the highest protracted CD4 count decline. The cell count and viral load were stable for about seven years, and then were inevitably followed by a decline in CD4 and a rise in viral phenotypes in 70 months after seroconversion, which indicates that participant 11 entered the stage of disease progression.

Since it is widely accepted that results based on Multiple Sequence Alignment (MSA) are reliable, we calculate the genetic distance based on MSA at different time points or the same time point, and check whether the NDDT and the genetic distance are statistically indistinguishable, which would further support the interpretation that the NDDT and the mutation rate are strongly linked. Here, the mean genetic distance is used. The calculation flowchart in Figure 4b is described by taking participant 1 as an example, who has three sequences at time point D1, four sequences at time point D2, and three sequences at time point D3. MUSCLE is used to carry out multiple sequence alignment and the pairwise nucleotide distances among these 10 sequences are calculated. We choose the model as “p-distance”, the substitutions as “Transitions + Transversions”, and the other options by default. The above steps are implemented by MEGAX [39]. Then, a genetic distance matrix among these 10 sequences is obtained. Similarly, we compute the average distance of sequences at the current time point or the adjacent time points. Then, we can obtain insight into viral divergence and viral diversity. The viral divergence at a given time point is defined by the distance between the current time point and its previous time point, and the viral diversity at the given time point is estimated by the average genetic distance for pairwise sequences at the current time point.

Figure 5b and Figures S31b–S38b show the genetic distance changes of the virus sequences at different time points among all participants (viral divergence). The mean distance increases by about 0.36% per month for all participants when all longitudinal data are used (mean $R^2 = 0.583$). In fact, the distance increases highly linearly for the month until the NDDT reaches the highest peak (mean $R^2 = 0.876$, mean $slope = 0.0007$, except participants 9 and 11 who are slow disease progressors) and then decreases or is stabilized later in the late stage of infection and has a less linear correlation (mean $R^2 = 0.428$, except participants 9 and 11). Figure 5c and Figures S31c–S38c show the genetic distance changes of the virus sequences at the same time point (viral diversity). The linear relationships of the mean distance for the men alive at the last study time point (mean $R^2 = 0.687$ of participants 2, 8, 9, and 11; participants 1, 3, 5, 6, and 7 died at the last study time point) are stronger than that for all men (mean $R^2 = 0.563$, mean $slope = 0.0333\%$). Furthermore, the diversity shows a strong linear correlation before the NDDT reaches the best peak (mean $R^2 = 0.784$, mean $slope = 0.0006$, except participants 9 and 11) and a higher variability (mean $R^2 = 0.341$, except participant 9 and 11) afterwards.

To further verify the consistent pattern of virus mutation rate, viral divergence, and viral diversity, the *Env* genes of patients 1 and 3 in dataset 1 are selected for the same analysis. Patients 4 and 7 are excluded because they were suspected to superinfect and failed to amplify virus samples according to the original paper [21]. Other patients' data are too few to be analyzed. The variation trend and the sequence distance are shown in Figure S39. For patients 1 and 3, the sequence distance between the current time points and the previous time points increases linearly ($R^2 = 0.509$ of patient 1, $R^2 = 0.781$ of patient 3) before the mutation rate reaches the peak and then decreases. The sequence distance between the current time point and the initial time point also increases linearly ($R^2 = 0.967$ of patient 1, $R^2 = 0.999$ of patient 3). It is worth noting that we redefine viral divergence, which can reflect not only the divergence trend, but also the divergence speed.

The analyses in the previous part have demonstrated the efficiency of our NDDT definition, so we used it to analyze the mutation rate of the complete genomes in dataset 1. Figure S40 shows that the NDDT during all time periods ranges from 0 to 0.2 except for patient 1. The results indicate the general applicability of our approach.

Although the time points of mutation rate change could hardly be precisely determined, we can still assert that the NDDT, CD4 count, viral load, viral divergence, and diversity have close temporal relationships and the mutation trend patterns across different individuals are highly consistent. The conclusion combining the results of sequence distances at the same or different time points provides solid evidence that our mutation rate definition is reasonable.

3.4. Running Time Comparison of Alignment Method and Our Alignment-Free Method

We utilize MATLAB R2019b to calculate the natural vector, and the MUSCLE algorithm of MEGAX to carry out multiple sequence alignment. The two methods are only tested on the above three datasets. For the 69 complete HIV-1 genome sequences (dataset 1), the running time of the natural vector is 1.9 s, while the time of MSA is nearly 23 min. For the 1337 sequence segment regions (dataset 2), it only takes 0.9 s to compute all natural vectors, which is faster than MSAs of all sequence segment regions. For the 1032 C2-V5 region of *Env* gene sequences (dataset 3), the total running time of natural vector is about 8.3 s, which is much less than MSA (nearly 16 min). Our alignment-free method is much faster than the MSA-based method to gain reliable results because the MSA method needs much more memory and more time to process data.

4. Discussion

Evolution is a function of both mutations and selective pressures (positive and negative), eliminating unfit mutations and selecting for advantageous mutations. Mutation will lead to the substitution, translocation, deletion and insertion of bases in the sequence. To describe these changes, we define the nucleotide distribution difference over time from a statistical point of view. The definition of NDDT considers not only the local properties of mutation—that is, the variations in the number and position of each base over time—but also the global properties of mutation—that is, the nucleotide distribution difference between two time points. The rationality of the definition is verified by testing on three longitudinal classical datasets. Another advantage of our method of measuring sequence differences is its great practical significance. Firstly, the sequence alignment model and genetic distance calculation model to compute the variation of mutant sequences need to be determined, but different models may lead to different results. Secondly, the sequence alignment process is very time-consuming; the aligned sequences require lots of memory to store. Our method has overcome these disadvantages, characterizing the distributions of four nucleotides of longitudinal sequences naturally and effectively, and only needing to store 12-dimensional numerical vectors. Our novel mutation representation is promising for studying mutations in a large number of longitudinal virus data.

Many details of our paper are worth discussing. Firstly, we use l_2 - norm to measure the differences of the two natural vectors. There are many other measures, but l_2 - norm has obtained satisfying results in previous sequence comparison studies [19–21], so this measure is intuitively applied to our study. In future work, we will consider mutations from other perspectives, for example, percent identity, mutation bias, etc. Secondly, we use three classical HIV nucleotide sequence datasets. The limited number of HIV sequences with time points might be an issue. If more reliable data of different timescales, different individuals or different groups of individuals were added into the current datasets, the results would be more persuasive. Thirdly, our conclusions in Sections 3.1 and 3.2 are supported by prior studies in the literature, which is that the *Env* gene has the fastest evolution speed during the whole infection period [6,28]. The conclusion in Section 3.3 has been verified by the viral divergence and diversity, which have many broad applications [6,28–31], and the extended viral divergence definition can reflect not only the divergence trend, but also the divergence speed. Additionally, we found the nucleotide distribution difference over time has a close statistical relationship with viral divergence and diversity. Fourthly, we regard the sequence at the first sampling time point instead of the RefSeq (NC_001802) as the reference sequence, which can more valuably reveal the difference between the progressive sequences and the original sequence of the same individual. Fifthly, our study can be improved to compare HIV-1 subtype B to HIV-1 subtype C, or HIV-1 O group to HIV-1 M group, or perhaps other viruses with greater distances such as HCV. Sixthly, the advantage of our method for measuring nucleotide sequence differences can be applied to protein sequences; then, the natural vector is 60-dimensional because of the 20 amino acids.

The conclusion concerning mutation pattern is derived from a small number of patients; the problem of extensive applicability needs to be further explored. To start with,

patients have different disease progression rates because of differences in their immune abilities. Next, other genes evolve slowly, and this mutation pattern could be different from that of *Env*. Moreover, HIV keeps evolving in infected individuals. The virus may have different variations in the host at the same time, or even mutates back to the original form, so as to survive better. Therefore, the viral variation trend might change along with time. Furthermore, the choice of data may have a great impact on the results. Some patients were superinfected or failed to amplify virus samples with low initial viral content, or received antiretroviral treatment at the study stage, which could affect the results. Some patients in the three datasets did not receive any treatment during the study period. With the advancement of medical technology, HIV medication is becoming more and more prevalent; viruses evolve rapidly and adapt to survive in their host, which makes it difficult to obtain a consistent conclusion and find a general method to study the mutation rate of HIV. Whether and how the mutation rate is affected by treatment is worthy of investigation, but requires more data and goes beyond the scope of this paper.

5. Conclusions

We propose a novel description based on the natural vector to measure the distributions of the mutant sequence and original sequence, and test it on three HIV datasets. The comparison of mutation rate of the nine genes is in accordance with previous research results, indicating that the *Env* gene mutates the fastest. Further analysis shows that the nucleotide distribution of several gene segment regions (*Gag*, *Pol*, *Nef*, *Vpu*) changes greatly both in the early stage of infection and in the whole infection period. The consistent patterns of the mutation rate, viral divergence and viral diversity are strongly supported by statistical analyses. The viral mutation rate can be divided into two stages during the infection period: the NDDT increases along with time in a linear manner since the infection (mean $R^2 = 0.55$), experiences high diversity and reaches a peak, and then declines or levels off at the late stage of infection (mean $R^2 = 0.705$). Before reaching the peak, the average linearly increasing time of NDDT is 4.8 years (except participants 9 and 11). The definition of NDDT is enhanced by viral divergence and diversity: The distance between the current time points and the previous time points increases highly linearly before the mutation rate reaches the peak (mean $R^2 = 0.876$); then, the virus continues to diverge for a few months and slows down or is stabilized. The genetic distances at the same time point present a strong linear correlation before the mutation rate reaches the peak (mean $R^2 = 0.784$), and then plateau or decline with more variability (mean $R^2 = 0.341$). The above results provide a basis for the rationality of the proposed NDDT and imply that the new definition could be generalized to study the progression rate of other diseases as well.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/genes13020170/s1>, Figure S1. The average nucleotide distribution difference of the nine genes for patient 1 over all time periods, Figure S2. The average nucleotide distribution difference of the nine genes for patient 2 over all time periods, Figure S3. The average nucleotide distribution difference of the nine genes for patient 3 over all time periods, Figure S4. The average nucleotide distribution difference of the nine genes for patient 4 over all time periods, Figure S5. The average nucleotide distribution difference of the nine genes for patient 5 over all time periods, Figure S6. The average nucleotide distribution difference of the nine genes for patient 6 over all time periods, Figure S7. The average nucleotide distribution difference of the nine genes for patient 7 over all time periods, Figure S8. The average nucleotide distribution difference of the nine genes for patient 8 over all time periods, Figure S9. The average nucleotide distribution difference of the nine genes for patient 9 over all time periods, Figure S10. The average nucleotide distribution difference of the nine genes for patient 11 over all time periods, Figure S11. Single nucleotide mutation distributions and frequencies of sequences at 5 time points for patient 2, Figure S12. Single nucleotide mutation distributions and frequencies of sequences at 5 time points for patient 3, Figure S13. Single nucleotide mutation distributions and frequencies of sequences at 5 time points for patient 4, Figure S14. Single nucleotide mutation distributions and frequencies of sequences at 5 time points for patient 5, Figure S15. Single nucleotide mutation distributions and frequencies of sequences at

5 time points for patient 6, Figure S16. Single nucleotide mutation distributions and frequencies of sequences at 5 time points for patient 7, Figure S17. Single nucleotide mutation distributions and frequencies of sequences at 5 time points for patient 8, Figure S18. Single nucleotide mutation distributions and frequencies of sequences at 5 time points for patient 9, Figure S19. Single nucleotide mutation distributions and frequencies of sequences at 5 time points for patient 11, Figure S20. The nucleotide distribution difference of the eight segments for patient in Dataset 2 at each time periods, Figure S21. (a) The average nucleotide distribution difference of the eight sequence segments for patient 1 in Dataset 1 over all time periods. (b) Single nucleotide mutation distributions of the sequence segments at all time points. (c) Single nucleotide mutation frequencies of the segments at all time points, Figure S22. (a) The average nucleotide distribution difference of the eight sequence segments for patient 2 in Dataset 1 over all time periods. (b) Single nucleotide mutation distributions of the sequence segments at all time points. (c) Single nucleotide mutation frequencies of the segments at all time points, Figure S23. (a) The average nucleotide distribution difference of the eight sequence segments for patient 3 in Dataset 1 over all time periods. (b) Single nucleotide mutation distributions of the sequence segments at all time points. (c) Single nucleotide mutation frequencies of the segments at all time points, Figure S24. (a) The average nucleotide distribution difference of the eight sequence segments for patient 4 in Dataset 1 over all time periods. (b) Single nucleotide mutation distributions of the sequence segments at all time points. (c) Single nucleotide mutation frequencies of the segments at all time points, Figure S25. (a) The average nucleotide distribution difference of the eight sequence segments for patient 5 in Dataset 1 over all time periods. (b) Single nucleotide mutation distributions of the sequence segments at all time points. (c) Single nucleotide mutation frequencies of the segments at all time points, Figure S26. (a) The average nucleotide distribution difference of the eight sequence segments for patient 6 in Dataset 1 over all time periods. (b) Single nucleotide mutation distributions of the sequence segments at all time points. (c) Single nucleotide mutation frequencies of the segments at all time points, Figure S27. (a) The average nucleotide distribution difference of the eight sequence segments for patient 7 in Dataset 1 over all time periods. (b) Single nucleotide mutation distributions of the sequence segments at all time points. (c) Single nucleotide mutation frequencies of the segments at all time points, Figure S28. (a) The average nucleotide distribution difference of the eight sequence segments for patient 8 in Dataset 1 over all time periods. (b) Single nucleotide mutation distributions of the sequence segments at all time points. (c) Single nucleotide mutation frequencies of the segments at all time points, Figure S29. (a) The average nucleotide distribution difference of the eight sequence segments for patient 9 in Dataset 1 over all time periods. (b) Single nucleotide mutation distributions of the sequence segments at all time points. (c) Single nucleotide mutation frequencies of the segments at all time points, Figure S30. (a) The average nucleotide distribution difference of the eight sequence segments for patient 11 in Dataset 1 over all time periods. (b) Single nucleotide mutation distributions of the sequence segments at all time points. (c) Single nucleotide mutation frequencies of the segments at all time points, Figure S31. (a) The variation trends of C2-V5 region of HIV-1 env gene for participant 2. The mutation progressions are shown with the filled blocks connected by the colorful line. The CD4+ T cell levels are shown with the black dotted line. The viral RNA levels are with the gray dotted line. (b) Viral divergence: Distance of sequences between the current time point and its previous time point. (c) Viral diversity: Distance of sequences at the current time point, Figure S32. (a) The variation trends of C2-V5 region of HIV-1 env gene for participant 3. The mutation progressions are shown with the filled blocks connected by the colorful line. The CD4+ T cell levels are shown with the black dotted line. The viral RNA levels are with the gray dotted line. Participant 3 died after the last time point of analysis (marked with diamond) (b) Viral divergence: Distance of sequences between the current time point and its previous time point. (c) Viral diversity: Distance of sequences at the current time point, Figure S33. (a) The variation trends of C2-V5 region of HIV-1 env gene for participant 5. The mutation progressions are shown with the filled blocks connected by the colorful line. The CD4+ T cell levels are shown with the black dotted line. The viral RNA levels are with the gray dotted line. Participant 5 died after the last time point of analysis (marked with diamond) (b) Viral divergence: Distance of sequences between the current time point and its previous time point. (c) Viral diversity: Distance of sequences at the current time point, Figure S34. (a) The variation trends of C2-V5 region of HIV-1 env gene for participant 6. The mutation progressions are shown with the filled blocks connected by the colorful line. The CD4+ T cell levels are shown with the black dotted line. The viral RNA levels are with the gray dotted line. Participant 6 died after the last time point of analysis (marked with diamond) (b) Viral divergence: Distance of sequences between the

current time point and its previous time point. (c) Viral diversity: Distance of sequences at the current time point, Figure S35. (a) The variation trends of C2-V5 region of HIV-1 env gene for participant 7. The mutation progressions are shown with the filled blocks connected by the colorful line. The CD4+ T cell levels are shown with the black dotted line. The viral RNA levels are with the gray dotted line. Participant 7 died after the last time point of analysis (marked with diamond) (b) Viral divergence: Distance of sequences between the current time point and its previous time point. (c) Viral diversity: Distance of sequences at the current time point, Figure S36. (a) The variation trends of C2-V5 region of HIV-1 env gene for participant 8. The mutation progressions are shown with the filled blocks connected by the colorful line. The CD4+ T cell levels are shown with the black dotted line. The viral RNA levels are with the gray dotted line. (b) Viral divergence: Distance of sequences between the current time point and its previous time point. (c) Viral diversity: Distance of sequences at the current time point, Figure S37. (a) The variation trends of C2-V5 region of HIV-1 env gene for participant 9. The mutation progressions are shown with the filled blocks connected by the colorful line. The CD4+ T cell levels are shown with the black dotted line. The viral RNA levels are with the gray dotted line. (b) Viral divergence: Distance of sequences between the current time point and its previous time point. (c) Viral diversity: Distance of sequences at the current time point, Figure S38. (a) The variation trends of C2-V5 region of HIV-1 env gene for participant 11. The mutation progressions are shown with the filled blocks connected by the colorful line. The CD4+ T cell levels are shown with the black dotted line. The viral RNA levels are with the gray dotted line. (b) Viral divergence: Distance of sequences between the current time point and its previous time point. (c) Viral diversity: Distance of sequences at the current time point, Figure S39. (a) Mutation rate of Env gene sequence. (b) Sequence distance between the current time point and its previous time point. (c) Sequence distance between the current time point and the initial time point, Figure S40. Mutation rate analysis of complete genomes. Datasets S1 to S3.

Author Contributions: Conceptualization, S.S.-T.Y.; methodology, S.S.-T.Y. and J.Y.; software, N.S.; validation, N.S.; formal analysis, N.S.; investigation, N.S.; resources, N.S.; data curation, N.S.; writing—original draft preparation, N.S.; writing—review and editing, N.S., J.Y. and S.S.-T.Y.; visualization, N.S.; supervision, J.Y. and S.S.-T.Y.; project administration, S.S.-T.Y.; funding acquisition, S.S.-T.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by National Natural Science Foundation of China (NSFC) grant (12171275), Tsinghua University Spring Breeze Fund (2020Z99CFY044), Tsinghua University start-up fund, and Tsinghua University Education Foundation fund (042202008).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study can be downloaded from the public database, and are also available in Supplementary Materials.

Acknowledgments: We are grateful to the National Center for Theoretical Sciences (NCTS) for providing an excellent research environment while part of this research was done. We thank the researchers who sequenced and shared the nucleotide sequences of HIV-1 (<https://hiv.biozentrum.unibas.ch>, accessed on 27 June 2021; https://www.hiv.lanl.gov/content/sequence/HIV/SI_alignments/set10.html, accessed on 25 August 2021; https://www.hiv.lanl.gov/content/sequence/HIV/SI_alignments/datasets.html, accessed on 27 June 2021). We thank the reviewers for their insightful suggestions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Weiss, R. How Does HIV Cause AIDS? *Science* **1993**, *260*, 1273–1279. Available online: <http://www.jstor.org/stable/2881758> (accessed on 31 July 2021). [CrossRef]
2. Robertson, D.L. Recombination in aids viruses. *J. Mol. Evol.* **1995**, *40*, 249–259. [CrossRef]
3. Douek, D.C.; Roederer, M.; Koup, R.A. Emerging concepts in the immunopathogenesis of AIDS. *Annu. Rev. Med.* **2009**, *60*, 471–484. [CrossRef]
4. Gilbert, P.B.; McKeague, I.W.; Eisen, G.; Mullins, C.; Guéye-Ndiaye, A.; Mboup, S.; Kanki, P.J. Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal. *Stat. Med.* **2003**, *22*, 573–593. [CrossRef]

5. UNAIDS; WHO. AIDS Epidemic Update. Available online: https://data.unaids.org/pub/epislides/2007/2007_epiupdate_en.pdf (accessed on 8 June 2021).
6. Shankarappa, R.; Margolick, J.B.; Gange, S.J.; Rodrigo, A.G.; Upchurch, D.; Farzadegan, H.; Gupta, P.; Rinaldo, C.R.; Learn, G.H.; He, X.; et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **1999**, *73*, 10489–10502. [[CrossRef](#)] [[PubMed](#)]
7. Powell, M.; Benková, K.; Selinger, P.; Dogoši, M.; Luňáčková, I.K.; Koutníková, H.; Laštíková, J.; Roubíčková, A.; Špůrková, Z.; Lacrova, L.; et al. Opportunistic infections in HIV-infected patients differ strongly in frequencies and spectra between patients with low CD4 cell counts examined postmortem and compensated patients examined antemortem irrespective of the HAART Era. *PLoS ONE* **2016**, *11*, e0162704. [[CrossRef](#)]
8. Cunningham, A.L.; Donaghy, H.; Harman, A.N.; Kim, M.; Turville, S. Manipulation of dendritic cell function by viruses. *Curr. Opin. Microbiol.* **2010**, *13*, 524–529. [[CrossRef](#)] [[PubMed](#)]
9. Kwong, P.D.; Wyatt, R.; Robinson, J.; Sweet, R.W.; Sodroski, J.; Hendrickson, W.A. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **1998**, *393*, 648–659. [[CrossRef](#)] [[PubMed](#)]
10. Doitsh, G.; Galloway, N.L.K.; Geng, X.; Yang, Z.; Monroe, K.M.; Zepeda, O.; Hunt, P.W.; Hatano, H.; Sowinski, S.; Muñoz-Arias, I.; et al. Cell death by pyroptosis drives CD4 T-cell depletion in HIV-1 infection. *Nature* **2014**, *505*, 509–514. [[CrossRef](#)]
11. WHO. HIV/AIDS. Available online: <https://www.who.int/en/news-room/fact-sheets/detail/hiv-aids> (accessed on 8 June 2021).
12. CDC. About HIV. Available online: <https://www.cdc.gov/hiv/basics/whatishiv.html> (accessed on 8 June 2021).
13. Kumar, V.; Abbas, A.K.; Aster, J.C. *Robbins Basic Pathology*; Elsevier: Amsterdam, The Netherlands, 2012; p. 147.
14. Garg, H.; Mohl, J.; Joshi, A. HIV-1 induced bystander apoptosis. *Viruses* **2012**, *4*, 3020–3043. [[CrossRef](#)]
15. Reeves, J.D.; Doms, R.W. Human immunodeficiency virus type 2. *J. Gen. Virol.* **2002**, *83*, 1253–1265. [[CrossRef](#)]
16. Notredame, C. Recent Evolutions of multiple sequence alignment algorithms. *PLoS Comput. Biol.* **2007**, *3*, e123. [[CrossRef](#)]
17. Chatzou, M.; Magis, C.; Chang, J.-M.; Kemena, C.; Bussotti, G.; Erb, I.; Notredame, C. Multiple sequence alignment modeling: Methods and applications. *Brief. Bioinform.* **2016**, *17*, 1009–1023. [[CrossRef](#)]
18. Yin, C.; Chen, Y.; Yau, S.S.T. A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering. *J. Theor. Biol.* **2014**, *359*, 18–28. [[CrossRef](#)] [[PubMed](#)]
19. Dong, R.; Zhu, Z.; Yin, C.; He, R.L.; Yau, S.S.-T. A new method to cluster genomes based on cumulative Fourier power spectrum. *Gene* **2018**, *673*, 239–250. [[CrossRef](#)]
20. Pei, S.; Dong, R.; He, R.L.; Yau, S.S.-T. Large-scale genome comparison based on cumulative Fourier power and phase spectra: Central moment and covariance vector. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 982–994. [[CrossRef](#)]
21. Wen, J.; Chan, R.H.; Yau, S.-C.; He, R.L.; Yau, S.S. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene* **2014**, *546*, 25–34. [[CrossRef](#)]
22. Sun, N.; Dong, R.; Pei, S.; Yin, C.; Yau, S.S.-T.A. A new method based on coding sequence density to cluster bacteria. *J. Comput. Biol.* **2020**, *27*, 1688–1698. [[CrossRef](#)]
23. Yu, C.; Hernandez, T.; Zheng, H.; Yau, S.-C.; Huang, H.-H.; He, R.L.; Yang, J.; Yau, S.S.-T. Real time classification of viruses in 12 dimensions. *PLoS ONE* **2013**, *8*, e64328. [[CrossRef](#)] [[PubMed](#)]
24. Zielezinski, A.; Girgis, H.Z.; Bernard, G.; Leimeister, C.-A.; Tang, K.; Dencker, T.; Lau, A.K.; Röhling, S.; Choi, J.J.; Waterman, M.S.; et al. Benchmarking of alignment-free sequence comparison methods. *Genome Biol.* **2019**, *20*, 144. [[CrossRef](#)] [[PubMed](#)]
25. Li, Y.; Tian, K.; Yin, C.; He, R.L.; Yau, S.S.-T. Virus classification in 60-dimensional protein space. *Mol. Phylogenet. Evol.* **2016**, *99*, 53–62. [[CrossRef](#)] [[PubMed](#)]
26. Sun, N.; Pei, S.; He, L.; Yin, C.; He, R.L.; Yau, S.S.-T. Geometric construction of viral genome space and its applications. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4226–4234. [[CrossRef](#)] [[PubMed](#)]
27. Zhao, X.; Tian, K.; Yau, S.S.T. A new efficient method for analyzing fungi species using correlations between nucleotides. *BMC Evol. Biol.* **2018**, *18*, 200. [[CrossRef](#)] [[PubMed](#)]
28. Zanini, F.; Brodin, J.; Thebo, L.; Lanz, C.; Bratt, G.; Albert, J.; Neher, R.A. Population genomics of inpatient HIV-1 evolution. *eLife* **2015**, *4*, e11282. [[CrossRef](#)]
29. Liu, Y.; McNevin, J.; Cao, J.; Zhao, H.; Genowati, I.; Wong, K.; McLaughlin, S.; McSweyn, M.D.; Diem, K.; Stevens, C.E.; et al. Selection on the Human Immunodeficiency Virus Type 1 Proteome Following Primary Infection. *J. Virol.* **2006**, *80*, 9519–9529. [[CrossRef](#)]
30. Liu, Y.; McNevin, J.; Zhao, H.; Tebit, D.M.; Troyer, R.M.; McSweyn, M.; Ghosh, A.K.; Shriner, D.; Arts, E.J.; McElrath, M.J.; et al. Evolution of Human Immunodeficiency Virus Type 1 Cytotoxic T-Lymphocyte Epitopes: Fitness-Balanced Escape. *J. Virol.* **2007**, *81*, 12179–12188. [[CrossRef](#)] [[PubMed](#)]
31. Liu, Y.; McNevin, J.P.; Holte, S.; McElrath, M.J.; Mullins, J.I. Dynamics of Viral Evolution and CTL Responses in HIV-1 Infection. *PLoS ONE* **2011**, *6*, e15639. [[CrossRef](#)]
32. Leitner, T.; Kumar, S.; Albert, J. Tempo and mode of nucleotide substitutions in gag and Env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J. Virol.* **1997**, *71*, 4761–4770. [[CrossRef](#)]
33. Levy, J. Pathogenesis of human immunodeficiency virus infection. *Microbiol. Rev.* **1993**, *57*, 183–289. [[CrossRef](#)] [[PubMed](#)]
34. Kaslow, R.A.; Ostrow, D.G.; Detels, R.; Phair, J.P.; Polk, B.F.; Rinaldo, C.R., Jr. The multicenter aids cohort study: Rationale, organization, and selected characteristics of the participants. *Am. J. Epidemiol.* **2017**, *185*, 1148–1156. [[CrossRef](#)]

35. Yin, C. Genotyping coronavirus SARS-CoV-2: Methods and implications. *Genomics* **2020**, *112*, 3588–3596. [[CrossRef](#)]
36. Chen, J.; Gao, K.; Wang, R.; Wei, G. Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies. *Chem. Sci.* **2021**, *12*, 6929–6948. [[CrossRef](#)] [[PubMed](#)]
37. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [[CrossRef](#)]
38. Edgar, R.C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **2004**, *5*, 113. [[CrossRef](#)] [[PubMed](#)]
39. Stecher, G.; Tamura, K.; Kumar, S. Molecular evolutionary genetics analysis (MEGA) for macOS. *Mol. Biol. Evol.* **2020**, *37*, 1237–1239. [[CrossRef](#)] [[PubMed](#)]