L-Shaped Distributions of the Relative Substitution Rates (c/μ) in SARS-COV-2 with or without Molecular Clocks, Challenging Mainstream Evolutionary Theories

Chun Wu[†1,2,*] *and Nicholas J. Paradis*[†1]

† These authors contributed equally

1 Department of Chemistry and Biochemistry, Rowan University, Glassboro, NJ 08028, USA

2 Department of Biological & Biomedical Sciences, Rowan University, Glassboro, NJ 08028, USA

* To whom correspondence should be addressed:

wuc@rowan.edu

Abstract

A definitive test to quantify fitness changes of mutations is required to end a continuing 50-year "neutralist-selectionist" debate in evolutionary biology. Our previous work introduced a substitution-mutation rate ratio c/µ test (c: substitution rate in Translated Region/TR or UnTranslated Region/UTR; µ: mutation rate) to quantify the selection pressure and thus the proportions of strictly neutral, nearly neutral, beneficial, and deleterious mutations in a genome. Intriguingly, both a L-shaped probability distribution of c/µ and molecular clock were observed for SARS-COV-2's genome. We found that the proportion of the different mutation types from the distribution is not consistent with the hypotheses of the three existing evolution theories (Kimura's Neutral Theory/KNT, Ohta's Nearly Neutral Theory/ONNT and the Selectionist Theory/ST), and a balance condition explains the molecular clock, thus we proposed a new theory named as Near-Neutral Balanced Selectionist Theory (NNBST). In this study, the c/µ analysis was extended beyond the genome to 26 TRs, 12 UTRs, and 10 TRSs (Transcriptional Regulatory Sequences) of SARS-COV-2. While L-shaped probability distributions of c/µ were observed for all of 49 segments, molecular clocks were observed for only 24 segments, supporting NNBST and Near-Neutral Unbalanced Selectionist Theory (NNUST) to explain the molecular evolution of 24/25 segments with/without molecular clocks. Thus, the Near-Neutral Selectionist Theory (NNST) integrates traditional neutral and selectionist theories to deepen our understanding of how mutation, selection, and genetic drift influence genomic evolution.

Author Summary

The “neutralist-selectionist” debate in molecular evolution has been unresolved for 50 years due to the three main theories of molecular evolution (Kimura’s Neutral Theory/KNT, Ohta’s Nearly-Neutral Theory/ONNT, Selectionist Theory/ST) disagreeing on the proportion of neutral mutations (KNT), nearly-neutral deleterious mutations (ONNT) and adaptive mutations (ST) within species. We recently developed a robust method, the c/μ relative substitution rate test, to quantify the proportion of each mutation type within >11K genomic sequences of SARS-COV-2 RNA virus. Our previous analysis revealed an L-shaped c/μ probability distribution and a constant substitution rate (e.g., molecular clock) for the SARS-COV-2 genome over 19 months, and the proportions of mutation types were inconsistent with those predicted by the three theories. We thus proposed the Near-Neutral Balanced Selectionist Theory (NNBST) to explain the molecular clock-feature and L-shaped probability distribution for SARS-COV-2. In this study, we extended this analysis to the 25 protein-coding gene segments and 24 non-protein-coding segments of SARS-COV-2. We observed that all 49 segments exhibited an L-shaped probability distribution and 24 out of the 49 segments exhibited a molecular clock, however the remaining 25 segments did not exhibit a molecular clock. We thus propose the Near-Neutral Unbalanced Selectionist Theory (NNUST) and NNBST to explain the segments without/with molecular clock features, respectively. We also coin the Near-Neutral Selectionist Theory (NNST) to combine traditional KNT, ONNT and ST to deepen our understanding of how mutation, selection, and genetic drift influence genomic evolution.

Introduction

All theories of molecular evolution (**Fig. 1A-B**) agree that most mutations are deleterious ($f^- \gg f_0, f_0'$ or f^+; f^-, f_0, f_0' and f^+ : the fraction of deleterious, strictly neutral, nearly neutral and beneficial mutations) and thus will be purified through evolution, but they all differ in their hypotheses on the proportion of the strictly neutral, nearly neutral and beneficial mutations that become fixed in a population (A.K.A. substitution).

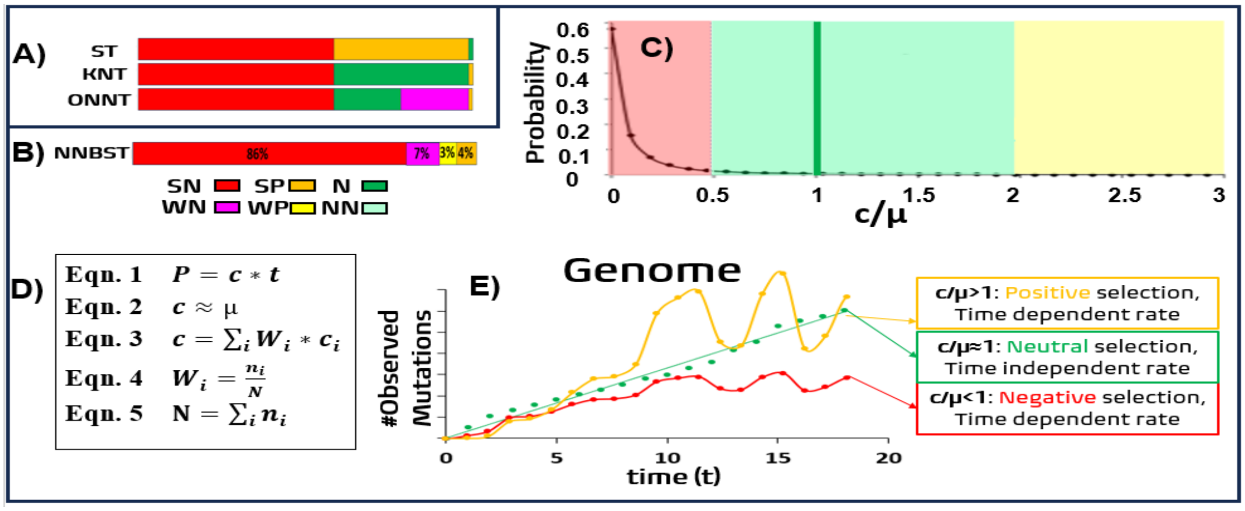


Figure 1. Comparison between the three mainstream evolution theories (A) and the newly proposed NNBST (B-E). A: Putative composition of the observed substitutions under different selection types for the three mainstream theories (ST, KNT and ONNT). **B:** The composition of SARS-COV-2 for NNBST. **SN:** strong negative selection (0 < c/µ < 0.5, red), **WN:** weak negative selection (0.5 < c/µ < 1.0, light red), **WP:** weak positive selection (1.0 < c/µ < 2.0, light yellow) and **SP:** strong positive selection (c/µ > 2.0, dark yellow). **NN**: WN + WP **C**: the L-shaped distribution of the relative substitution rate (c/µ) of SARS-COV-2. **D**: The five equations describing NNBST. **Equ. 1**: the integral substitution probability (**P**) per Nucleotide (NT) site of a genome at time *t* is the product of the substitution rate of the whole genome (c) with the time. **Equ. 2**: the substitution rate (c) is almost equal to the spontaneous mutation rate (µ) due to

effective neutrality. **Equ. 3**: the substitution rate of the whole genome (c) is the weighted sum of the substitution rate of each segment (c_i) with weight (W_i). **Equ. 4**: The weight (W_i) of each segment is determined by segment length (n_i) over the genomic length (N). **Equ. 5**: the length of a genome (N) is a summation of the length of all segments. **E**: effective neutrality across the genome despite varying selective pressures acting on different segments.

The Selectionist (Neo-Darwinian) Theory/ST[1, 2]states that most substitutions are beneficial mutations that lead to a fitness increase of the species (i.e. $f^+ \gg f_0 \approx 0$). In contrast, Kimura's Neutral Theory (KNT) [3, 4], [5, 6] argues that most substitutions are strictly neutral mutations that lead to no change in fitness (i.e. $f_0 \gg f^+ \approx 0$). Ohta's Nearly Neutral Theory/ONNT[7, 8] extends this perspective by acknowledging the importance of nearly neutral mutations that have slightly (deleterious) effects ($f_0^- \gg f_0^+ \approx 0$, $f'_0 > f_0 \gg f^+$; f_0^-,f_0^+: the fractions of slightly deleterious and slightly beneficial mutations) besides the strictly neutral mutations. Depending on the effective population size, environmental conditions, and selective pressures, these nearly neutral mutations can be fixed in the population with variable rates, likely leading to a time-dependent substitution rate. However, these theories could not be evaluated on the basis of empirical data if mutations could not be classified by their fitness changes into the different types and thus their proportion could not be quantified, thus leading to a continuing 50-year "neutralist-selectionist" debate[9] [10-12].

In our previous paper[13], we framed a replication-selection model (**Figure 2**) that phenomenologically quantifies selection pressure on any site in a genome including Translated

Region (TR) and UnTranslated Region (UTR) using sequence data.

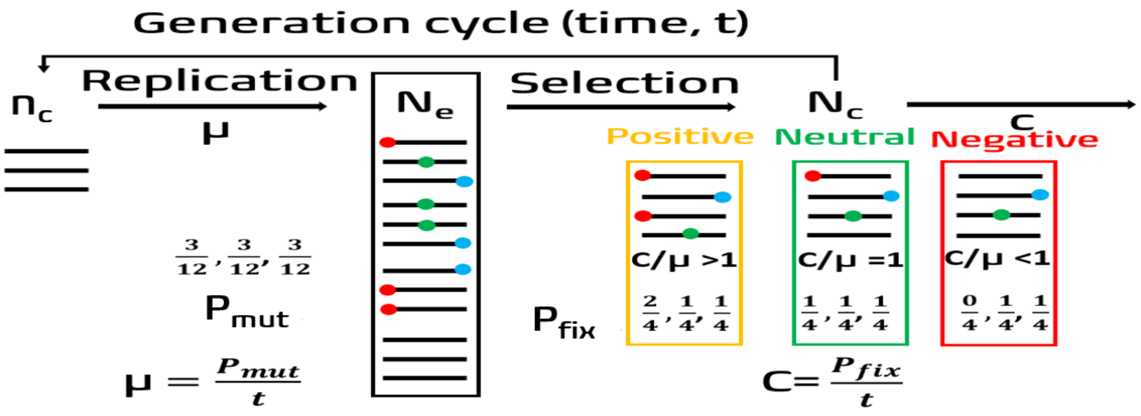

Generation cycle (time, t)
Replication
n_c
μ
N_e
Selection
N_c
c
Positive
Neutral
Negative
$\frac{3}{12}, \frac{3}{12}, \frac{3}{12}$
P_{mut}
$\mu = \frac{P_{mut}}{t}$
P_{fix}
c/μ >1
$\frac{2}{4}, \frac{1}{4}, \frac{1}{4}$
c/μ =1
$\frac{1}{4}, \frac{1}{4}, \frac{1}{4}$
c/μ <1
$\frac{0}{4}, \frac{1}{4}, \frac{1}{4}$
$C = \frac{P_{fix}}{t}$

Figure 2. Simplified replication-selection model for a virus population over time under positive selection, neutral selection and negative selection for the first nucleotide position (red circle). A detailed description has been placed in the methods section. Viral genomes are shown as black lines. Mutations are shown as red, green and blue circles.

Our model uses a mutation-selection frame that assumes the mutation step is independent from the selection step[14]. Our model considers the substitution rate (c) in the population as a product of the mutation rate (μ) with a substitution-mutation rate ratio (c/μ), reflecting neutral and natural selection effects. Therefore, this c/μ ratio (c: substitution rate in TR or UTR, μ: mutation rate) can serve as a test to detect fitness change of mutations in a genome (c/μ = 1, <1 and >1: strictly neutral, deleterious, and beneficial). The relative substitution rate c/μ for each nucleotide (NT) site in the genome of SARS-COV-2 (29,903 NTs) was obtained from its first 19-month genomic data (three independent sets of a total of 11,198 genome sequences from NextStrain[15]) to quantify the fitness change of mutations for the very first time (**Fig. 1B-C**). Intriguingly, both L-shaped discrete probability distributions of c/μ (**Fig. 1C**) and molecular clock (**Fig. 1E**) were observed for the whole genome. From the distribution, we found that the proportion of strictly neutral, nearly neutral, beneficial mutations are not consistent with the hypotheses of the three existing theories. We therefore proposed a novel theory that merges some aspects of ONNT and ST to explain the observed mutation composition for the genome (86% strongly negative, 7% weakly negative, 3% weakly positive, 4% strongly positive). We named this new theory as Near-Neutral Balanced Selectionist Theory (NNBST). NNBST not only accounts for the important role of nearly neutral mutations (7% weakly negative and 3% weakly positive), but it also recognizes the crucial role from the positive selection (3% weakly positive and 4% strongly positive) in determining the overall substitution rate (**Fig. 1B-E)**. NNBST challenges the core tenet of KNT and ONNT by suggesting that strictly neutral mutations are rare (0.5%, 1:20 between strictly neutral and near neutral mutations) due to natural selection, evidenced by the observation of an L-shaped distribution curve of c/μ rather than a Poisson distribution curve centered at strictly neutral ($c/\mu = 1$) for SARS-COV-2 (**Fig. 1C**).

In addition, a balancing act between beneficial ($c_i/\mu>1$) and deleterious mutations ($c_i/\mu<1$) in increasing and decreasing the substitution rate respectively leads to a molecular evolutionary clock (i.e. an effective neutrality or $c/\mu = 1$) across the pathogen's genome despite varying selective pressures acting on different genes (**Fig. 1D-E**). This balanced natural selection mechanism leads to a molecular clock to describe the linear increase of genomic variation over time using a simple set of equations (**Fig. 1D**). This mechanism does not require the original stronger assumptions that the fraction of beneficial mutations are negligible by KNT to derive the molecular clock.[6] If the substitution rates between beneficial and deleterious mutations are not balanced, a molecular clock will not be observed. We coined it Near-Neutral Unbalanced

Selectionist Theory (NNUST), which does not require the stronger assumption that that the fraction of near neutral mutations are negligible by ST. Combining NNBST with NNUST, a Near-Neutral Selectionist Theory (NNST) acknowledges the contribution of both near neutral and beneficial mutations in determining the overall substitution rate.[10]

In this study, the c/μ analysis was extended beyond the gnome to 26 TRs, 12 UTRs, and 10 TRSs (Transcriptional Regulatory Sequences) of SARS-COV-2. While discrete L-shaped probability distributions of c/μ were observed for all of 49 segments, molecular clocks (R^2>0.6000) were observed for only 25 segments. Therefore, we propose NNBST and NNUST to explain the molecular evolution of 25 and 24 segments with and without showing molecular clocks, respectively. Our findings advocate for the c/μ test as an effective test, potentially ending the longstanding debate in evolutionary biology with NNST.

Results

To validate c/μ values (per NT and per unit time) for 49 segments of SARS-COV-2 including 1 genome, 26 TRs, 12 UTRs and 10 TRSs, reported in our previous paper [13], that is based on position-based approach in which SRSRS (**Figure 4**) was used to get the rates (per 19 months per NT), their ratios, and the proportions of mutation types from cumulative PDRSR were calculated (**Figure 5**).

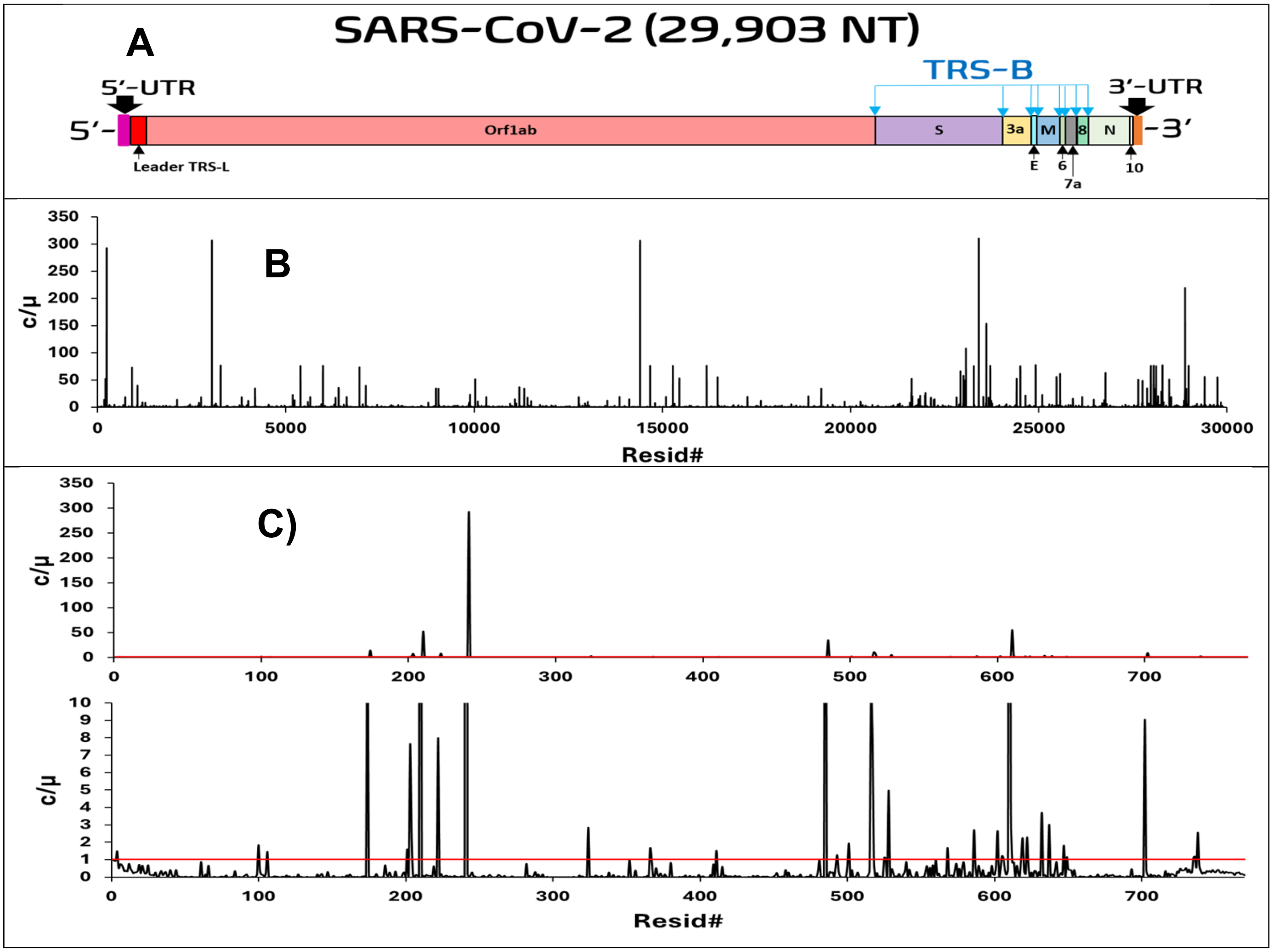


Figure 3. A) The structure of the SARS-CoV-2 genome, including major gene segments (Orf1ab, S, E, M and N), accessory genes (Orf3a, Orf6, Orf7a, Orf8 and Orf10), UTRs (Orf1ab 5'-UTR and Orf10 3'-UTR) and TRS (leader TRS-L and TRS-B). **B**) Position-based c/μ at each NT site in the SARS-COV-2 genome. **C**) Position-based c/μ at each NT site in All-UTR. The red line in **Figure 3C** represents neutral selection (c/μ=1). μ: the substitution rate of Orf1ab 5'UTR which is 5.46-fold of GSR.

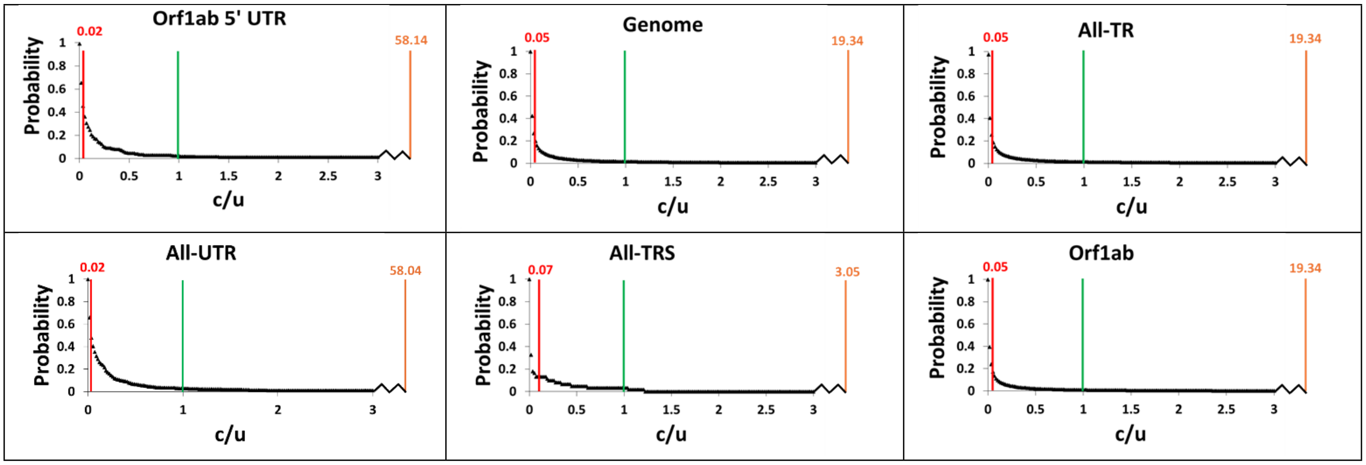


Figure 4. c/μ Cumulative probability distribution function of Orf1ab 5'UTR, genome, All-TR, All-UTR, All-TRS and Orf1ab segments demonstrating a high coefficient of determination, showing the abundance of sites under different selection types. The boundaries for weak negative (red line) to neutral selection (green line) to weak positive selection (orange line) and their determined c/μ positions are noted. Strong negative selection and strong positive selection would be to the left and right of the red and orange lines, respectively. A broken lined leading on the x-axis represents the distance between c/μ = 3.0 to the weak positive selection boundary. See **Figure S1** for the remaining segments demonstrating a high coefficient of determination.

These rates and ratios are also obtained in this study from the time-based approach in which temporal trend (timeline) in **Figure 5** is fitted by least square (y=m*x) to obtain the rates and the coefficient of determination (R^2), thus their ratios from the three independent genome sequence datasets used in our previous paper as described in the method section (**Table 1** and **Figure S1-S3**).

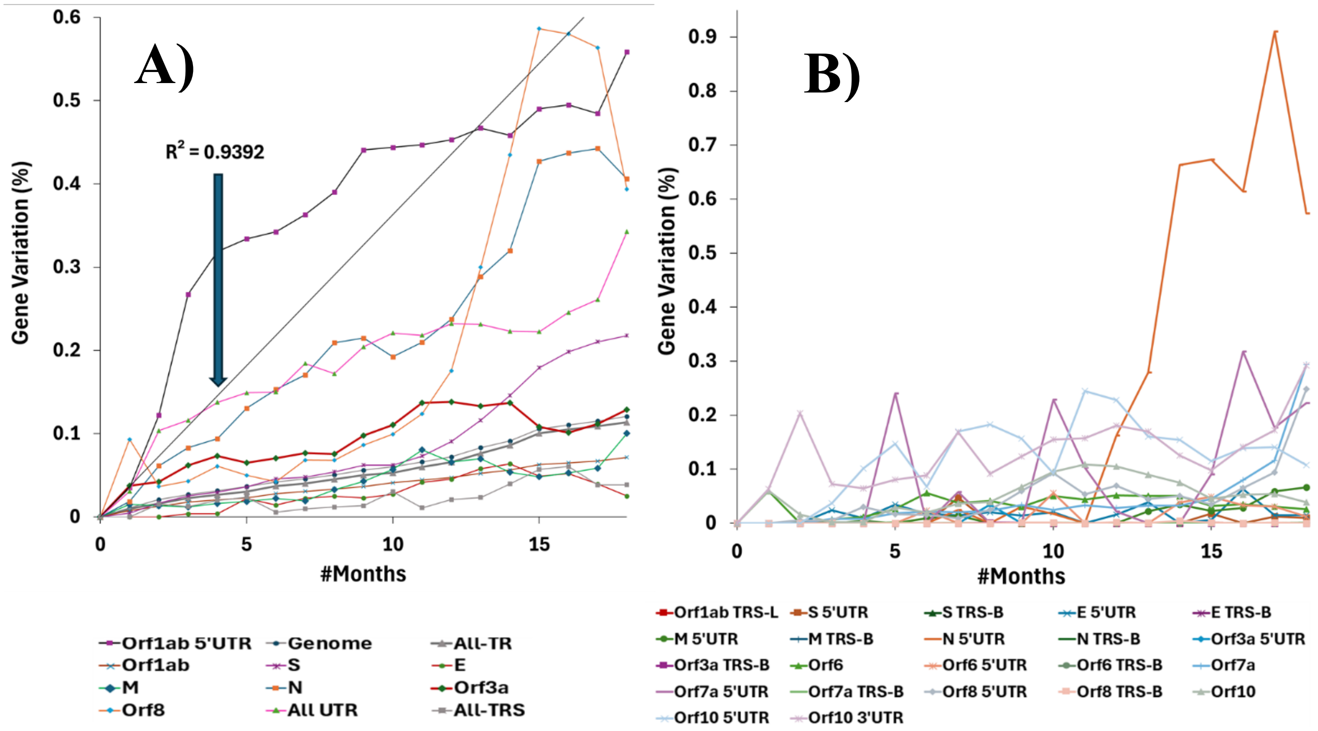


Figure 5. The total percent NT substitution variation (c) for segments exhibiting molecular clock (A) and non-molecular clock (B) features over evolution time, averaged over datasets A1A-A1C. The trendline and coefficient of determination for Orf1ab 5'UTR ($R^2 = 0.9392$) is defined as μ. See **Figures S2-S3** for the individual timelines for each segment.

Table 1. Time-based c/μ, R^2, c/μ and percent abundance at three selection type boundaries, the relative percent abundance of sites under four different selection types, the ratio of sites under weak negative to positive selection, the ratio of sites under near-neutral to neutral selection and evaluation of KNT, ONNT and ST for the genome and each segment exhibiting molecular clock features of SARS-COV-2 over 19 months.

Seg (NT length)	fo’ = c/μ	c R^2	WNB (c/μ)	NB (c/μ)	WPB (c/μ)	SN^a (%)	WN^a (%)	WP^a (%)	SP^a (%)	fo (%)	fo’: fo	WN: WP	WN: P	KNT	ONNT	ST
Orf1ab 5'UTR (265)	**1.00**	**0.9392**	0.02	1.01	58.14	34.72	63.40	1.89	0.00	0.13	795.00	N/A	33.60	3,5	NA	NA
Genome (29,903)	**0.18(H-)**	**0.9957**	0.05	1.02	19.34	84.21	14.28	1.33	0.19	0.03	576.70	5.24	9.43	3,4,5	N/A	NA
All-TR (29,133)	**0.17(H-)**	**0.9805**	0.05	1.01	19.34	85.11	13.44	1.25	0.20	0.03	586.56	5.54	9.31	3,4,5	NA	NA
Orf1ab (21,291)	**0.11(H-)**	**0.9854**	0.05	1.01	19.34	98.29	0.72	0.93	0.06	0.02	468.40	9.29	0.73	3,4,5	NA	NA
S (3,822)	**0.28(H-)**	**0.8711**	0.03	1.01	29.02	74.05	24.05	1.91	0.00	0.03	802.62	2.63	12.59	3,4,5	NA	NA
E (228)	**0.08(H-)**	**0.7201**	0.07	1.02	14.51	91.67	7.46	0.44	0.44	0.02	501.60	9.50	8.50	3,5	NA	NA
M (669)	**0.12(H-)**	**0.7902**	0.09	1.04	11.60	87.44	11.51	1.05	0.00	0.02	642.24	9.84	11.00	3,4,5	NA	NA
N (1,260)	**0.66(H-)**	**0.9497**	0.02	1.02	58.14	45.64	51.11	3.25	0.00	0.08	831.60	1.24	15.71	3,4,5	NA	NA
Orf3a (828)	**0.24(H-)**	**0.6116**	0.10	1.03	9.67	76.81	21.01	2.17	0.00	0.04	596.16	5.62	9.67	3,4,5	NA	NA
Orf8 (366)	**0.68(H-)**	**0.6135**	0.02	1.10	58.04	36.34	60.38	3.28	0.00	0.03	2613.24	8.99	18.41	3,4,5	NA	NA
Nsp1 (538)	**0.10(H-)**	**0.6735**	0.12	0.99	8.29	90.19	8.70	0.93	0.18	0.03	378.00	10.60	7.83	3,4,5	NA	NA
Nsp2 (1,912)	**0.12(H-)**	**0.8517**	0.12	1.01	8.29	88.56	10.14	1.31	0.00	0.02	524.98	9.52	7.76	3,5	NA	NA
Nsp3 (5,388)	**0.14(H-)**	**0.9801**	0.07	1.00	14.51	85.90	13.18	0.93	0.00	0.02	668.37	9.05	14.25	3,4,5	NA	NA
Nsp4 (1,498)	**0.09(H-)**	**0.6555**	0.12	1.01	8.29	91.67	7.33	0.93	0.07	0.02	540.00	10.93	7.33	3,4,5	NA	NA
Nsp6 (868)	**0.12(H-)**	**0.6016**	0.14	1.00	7.25	91.26	6.90	1.84	0.00	0.04	287.10	4.86	3.75	3,4,5	NA	NA
Nsp8 (592)	**0.04(H-)**	**0.8214**	0.19	1.08	5.28	95.46	3.87	0.67	0.00	0.04	114.05	1.00	5.75	3,5	NA	NA
Nsp9 (337)	**0.12(H-)**	**0.7971**	0.10	1.02	9.67	88.50	8.85	2.66	0.00	0.07	172.89	31.02	3.33	3,4,5	NA	NA
Nsp10 (415)	**0.03(H-)**	**0.7922**	0.28	1.17	3.63	96.64	3.12	0.24	0.00	0.01	406.58	3.75	12.99	3,5	NA	NA
Nsp11 (2,794)	**0.18(H-)**	**0.9169**	0.05	1.02	19.34	80.79	18.71	0.50	0.00	0.01	3355.20	5.99	37.34	3,4,5	NA	NA
Nsp12 (1,801)	**0.08(H-)**	**0.9507**	0.12	0.99	8.29	91.68	7.10	1.17	0.05	0.02	384.64	10.59	5.82	3,4,5	NA	NA
Nsp13 (1,579)	**0.07(H-)**	**0.6755**	0.14	1.06	7.25	92.73	6.39	0.82	0.06	0.03	265.61	8.10	7.21	3,5	NA	NA
Nsp15 (892)	**0.06(H-)**	**0.6361**	0.14	1.02	7.25	93.88	4.92	1.20	0.00	0.05	112.59	14.06	4.10	3,5	NA	NA
All-UTR (771)	**0.53(H-)**	**0.7167**	0.02	1.00	58.04	33.98	63.17	2.85	0.00	0.07	715.10	2.15	22.14	3,4,5	NA	NA
All-TRS (61)	**0.07(H-)**	**0.6299**	0.33	0.82	3.05	93.22	3.50	3.28	0.00	0.13	55.51	3.00	1.07	3,5	NA	NA

*c/μ calculated with reference to Orf1ab 5’UTR mutation rate (μ = 36.1E-03% substitutions/NT site/month; c = total substitutions/any site * year). Selection type is positive (+) if c/μ >1, negative (–) if c/μ <1 or near-neutral if c/μ ≈ 1. c/μ is classified with letters for high R^2 (H, R^2>0.6000), low R^2 (L, R^2 < 0.6000), positive c/μ (+) or negative c/μ (-) values. See **Table SX** for the average absolute substitution rate (c) value for the genome and each segment calculated from the three datasets.

fo’: Fraction of sites under near neutral selection is equivalent to the scaled c/μ value; **R^2**: Coefficient of determination; **WNB:** Weak negative boundary; **NB**: Neutral selection boundary; **WPB**: Weak positive selection boundary; **SN**: Percentage of strong negative selection; **WN**: Percentage of weak negative selection; **N**: Percentage of neutral selection; **WP**: Percentage of weak positive selection; **SP**:

Percentage of strong positive selection; **fo**: fraction of sites under neutral selection (obtained from scaled c/μ probability distribution); **fo’:fo**: Ratio of sites under near-neutral selection to neutral selection; **WN:P**: Ratio of sites under weak negative to positive selection. **WN:WP:** Ratio of sites under weak negative to weak positive ratio.

[b] **SN, WN, N, WP and SP** percent abundances calculated from scaled position-dependent c/μ values. **KNT**: Kimura’s Neutral Theory; **ONNT**: Ohta’s Nearly-Neutral Theory; **ST**: Selectionist Theory.

Violation descriptors for KNT, ONNT and ST: **Blank site**: No violation; **1**: %WP ≫ 0; **2:** %WN ≫ 0; **3**: %NN ≫ 0; **4**: %P ≫ 0; **5**: fo’ ≫ fo. **NA:** Not applicable.

[a] Relative percent abundance of selection types excluding SN selection. Values were calculated by taking the difference between absolute abundance values and dividing by the sum of all absolute abundances.

NA: Not applicable.

R^2 offers a statistical measure on how good molecular clock feature (i.e. time-independent constant rate) holds.

There are several earliest genome sequences such as Wuhan-Hu-1 and Wuhan IPBCAMS-WH-01 that can serve as the reference sequence for our substitution analysis. To quickly check if the substitution rate of the genome and the molecular clock feature change much when different reference sequences are used, we generated timelines using either as the reference sequence. The absolute rate changes slightly from 6.4E-3 to 7.2E-3 per NT per month and R^2 is changed from 0.9829 to 0.9912 when the reference sequence is changed from Wuhan-Hu-1 to Wuhan IPBCAMS-WH-01 (**Figure S1**). Therefore, its effect on the relative substitution rate should be even smaller, and thus the former sequence which was used in our previous study was still used as reference sequence in this study.

Molecular clock with R^2>0.6000 was observed for only 24 out of 49 segments including the genome, UTR, TR, 5'UTR, and most of genes and NSPs of SARS-COV-2. The raw time-based c and c R^2 values were tabulated (UTR/TRS: **Table S1; and** TR: **Table S2**). While the substitution timelines of Orf1ab 5'UTR, genome, All-TR, All-UTR, All-TRS and most of genes and NSPs of SARS-CoV-2 (24 segments in **Table 1**) have high coefficient of determination (R^2>0.6000), most of UTR/TRS segments, a few of genes and NSPs (25 segments in **Table 2**) have low coefficient of determination (R^2<0.6000)**.**

Table 2. Position-based c/μ, R^2, c/μ and percent abundance at three selection type boundaries, the relative percent abundance of sites under four different selection types, the ratio of sites under weak negative to positive selection, the ratio of sites under near-neutral to neutral selection and evaluation of KNT, ONNT and ST for each segment exhibiting non-molecular clock features of SARS-COV-2 over 19 months.

Seg (NT length)	**fo'= c/μ**	**c R^2**	**WNB (c/μ)**	**NB (c/μ)**	**WPB (c/μ)**	**SN[a] (%)**	**WN[a] (%)**	**WP[a] (%)**	**SP[a] (%)**	**fo (%)**	**fo': fo**	**WN: WP**	**WN: P**	**KNT**	**ONNT**	**ST**
Orf6 (186)	0.07(L-)	0.0000	0.07	1.16	14.51	92.43	5.96	1.61	0.00	0.03	260.40	7.67	7.67	NA	1,5	2,3
Orf7a (366)	0.26(L-)	0.4031	0.07	1.21	14.51	73.77	25.41	0.82	0.00	0.01	1934.92	15.99	15.99	NA	1,5	2,3
Orf10 (117)	0.11(L-)	0.3557	0.17	1.04	5.80	88.89	8.55	2.56	0.00	0.11	102.96	4.67	4.67	NA	1,5	2,3
Nsp5 (1,498)	0.07(L-)	0.5965	0.14	1.00	7.25	92.81	5.99	1.20	0.00	0.03	240.98	8.80	8.00	NA	1,4,5	2,3
Nsp7 (868)	0.05(L-)	0.3018	0.28	1.19	3.63	94.38	4.42	1.21	0.00	0.02	224.10	1.00	0.33	NA	1,4,5	2,3
Nsp14 (1,036)	0.06(L-)	0.4841	0.17	1.45	5.80	94.12	5.20	0.67	0.00	0.01	871.92	18.15	18.15	NA	1,5	2,3
Orf1ab TRS-L (7)	0.01(L-)	0.2699	N/A	1.01	N/A	100.00	0.00	0.00	0.00	0.00	N/A	N/A	N/A	NA	5	2,3
S 5'UTR (7)	0.01(L-)	0.0239	N/A	1.01	N/A	100.00	0.00	0.00	0.00	0.00	N/A	N/A	N/A	NA	5	2,3
S TRS-B (7)	0.01(L-)	-0.0075	N/A	1.01	N/A	100.00	0.00	0.00	0.00	0.00	N/A	N/A	N/A	NA	5	2,3
E 5'UTR (24)	0.04(L-)	0.0399	N/A	1.01	N/A	100.00	0.00	0.00	0.00	0.00	N/A	N/A	N/A	NA	5	2,3
E TRS-B (6)	0.00(L-)	N/A	N/A	1.01	N/A	100.00	0.00	0.00	0.00	0.00	N/A	N/A	N/A	NA	5	
M 5'UTR (50)	0.06(L-)	0.4925	N/A	1.20	N/A	100.00	0.00	0.00	0.00	0.06	102.00	N/A	N/A	NA	1,5	2,3
M TRS-B (7)	0.00(L-)	N/A	N/A	1.01	N/A	100.00	0.00	0.00	0.00	0.00	N/A	N/A	N/A	NA	5	
N 5'UTR (14)	0.77(L-)	0.5577	N/A	2.80	N/A	100.00	0.00	0.00	0.00	0.05	1412.18	N/A	0.00	NA	4,5	
N TRS-B (7)	0.03(L-)	0.2535	N/A	1.01	N/A	100.00	0.00	0.00	0.00	0.00	N/A	N/A	N/A	NA	5	2,3
Orf3a 5'UTR (8)	0.00(L-)	-0.0235	N/A	1.01	N/A	100.00	0.00	0.00	0.00	0.00	N/A	N/A	N/A	NA	5	2,3
Orf3a TRS-B (7)	0.02(L-)	-0.0402	N/A	1.01	N/A	100.00	0.00	0.00	0.00	0.00	N/A	N/A	N/A	NA	5	2,3
Orf6 5'UTR (10)	0.04(L-)	0.2871	0.24	1.01	4.15	90.00	10.00	0.00	0.00	0.00	N/A	N/A	N/A	NA	5	2,3
Orf6 TRS-B (6)	0.03(L-)	-0.2110	N/A	1.01	N/A	100.00	0.00	0.00	0.00	0.00	N/A	N/A	N/A	NA	5	2,3
Orf7a 5'UTR (6)	0.29(L-)	0.2692	0.34	0.82	2.90	66.67	16.67	16.67	0.00	1.28	22.62	2.00	2.00	NA	1,5	2,3
Orf7a TRS-B (7)	0.25(L-)	0.3129	0.28	0.82	3.63	71.43	14.29	14.29	0.00	1.10	22.75	2.00	2.00	NA	1,5	2,3
Orf8 5'UTR (134)	0.23(L-)	0.4862	0.07	1.08	14.51	79.10	20.15	0.75	0.00	0.09	246.56	10.00	10.00	NA	1,5	2,3
Orf8 TRS-B (7)	0.26(L-)	0.4698	0.34	0.82	2.90	71.43	14.29	14.29	0.00	0.62	41.86	1.00	1.00	NA	1,5	2,3
Orf10 5'UTR (24)	0.30(L-)	0.2323	0.19	1.90	5.28	70.83	25.00	4.17	0.00	0.06	504.00	5.00	5.00	NA	1,5	2,3
Orf10 3'UTR (229)	0.38(L-)	0.1415	0.17	1.02	5.80	62.88	32.31	4.80	0.00	0.44	87.02	1.90	1.73	NA	1,4,5	2,3

*c/μ calculated with reference to Orf1ab 5'UTR mutation rate (μ = 36.1E-03% substitutions/NT site/month; c = total substitutions/any site * year). Selection type is positive (+) if c/μ >1, negative (–) if c/μ <1 or near-neutral if c/μ ≈ 1. c/μ is classified with letters for high R^2 (H, R^2>0.6000), low R^2 (L, R^2 < 0.6000), positive c/μ (+) or negative c/μ (-) values. See **Table SX** for the average absolute substitution rate (c) value for the genome and each segment calculated from the three datasets.

fo': Fraction of sites under near neutral selection is equivalent to the scaled c/μ value; **R^2**: Coefficient of determination; **WNB:** Weak negative boundary; **NB**: Neutral selection boundary; **WPB**: Weak positive selection boundary; **SN**: Percentage of strong negative selection; **WN**: Percentage of weak negative selection; **N**: Percentage of neutral selection; **WP**: Percentage of weak positive selection; **SP**:

Percentage of strong positive selection; **fo**: fraction of sites under neutral selection (obtained from scaled c/μ probability distribution); **fo':fo**: Ratio of sites under near-neutral selection to neutral selection; **WN:P**: Ratio of sites under weak negative to positive selection. **WN:WP:** Ratio of sites under weak negative to weak positive ratio.

[b] **SN, WN, N, WP and SP** percent abundances calculated from scaled position-dependent c/μ values. **KNT**: Kimura's Neutral Theory; **ONNT**: Ohta's Nearly-Neutral Theory; **ST**: Selectionist Theory.

Violation descriptors for KNT, ONNT and ST: **Blank site**: No violation; **1**: %WP > ≫ 0; **2:** %WN ≫ 0; **3**: %NN ≫ 0; **4**: %P ≫ 0; **5**: fo' ≫ fo. **NA:** Not applicable.

[a] Relative percent abundance of selection types excluding SN selection. Values were calculated by taking the difference between absolute abundance values and dividing by the sum of all absolute abundances.

NA: Not applicable.

For the UTR, c R^2 values revealed only 3 out of 22 non-coding segments exhibited molecular clocks. Orf1ab 5'UTR (0.9392), All-UTR (0.7167) and All-TRS (0.6299) exhibited good molecular clocks, whereas the remaining UTRs and TRSs did not exhibit good molecular clocks (-0.2110 - 0.5577), with E TRS-B and M TRS-B being entirely conserved (c/μ=0) and thus no R^2 was generated (**Table S1).** For the TR, c R^2 values demonstrated 20 out of 25 coding segments showed good molecular clocks. Besides the genome (0.9957) and All-TR (0.9805), Orf1ab (0.9854), N (0.9497), S (0.8711), M (0.7902), E (0.7201), Orf8 (0.6135), Orf3a (0.6116), NSP1-4 (0.6555-0.9801), NSP6-13 (0.6755-0.9507) and NSP15 (0.6361) exhibit good molecular clocks, whereas NSP5 (0.5965), NSP14 (0.4841), Orf7a (0.4031), Orf10 (0.3557), NSP7 (0.3018) and Orf6 (0.0000) do not exhibit molecular clock features (**Table S2**).

The highest substitution rate of Orf1ab 5'UTR with a molecular clock was used as the Fundamental Genomic Mutation Rate (μ) of SARS-CoV-2. In our previous study[13], μ was its genomic substitution rate (GSR) derived from **equation 8a** ($c = \mu f'_0$). GSR was obtained from least square fitting of a linear line of the genomic substitutions over the time with a high correlation coefficient (R^2=0.9957) using **equation 1** ($P = c * t$). This ($\mu = c$) sets the fraction of near neutral mutation sites in the genome be one ($f'_0 = 1$). However, the substitution rates of some segments (e.g. Orf1ab 5'UTR, Orf8, N, all-UTR, S and Orf3a in decreasing order) also show a time-independent molecular clock feature (R^2>0.6000) with a much higher rate than GSR (e.g. 5.46-fold of GSR for Orf1ab 5'UTR). If GSR is still used as μ, the fraction of near neutral mutation sites in Orf1ab 5'UTR would go beyond one. To remove this kind of violation, the highest substitution rate of the segment satisfying molecular clock was set to be μ. In this study, the substitution rate of Orf1ab 5'UTR was used as μ.

Time-based and position-based c/μ are highly consistent for the 24 segments with molecular clock feature, also suggesting validity of molecular clock for these segments. The consistency of time-based and position-based c/μ values were assessed for UTR/TRS segments (**Table S3**) and the TR segments (**Table S4)** to verify the accuracy of the molecular clock assumption in the time-based approach. The consistency of both methods is dependent on the linearity (R^2) of c obtained from least square fitting: substitution rates exhibiting a good molecular clock (R^2>0.6000) should give very consistent c/μ values, whereas the absence of a molecular clock will produce larger deviations in the time-based and position-based c/μ as descripted in our method section. The absolute and percent differences between time-based and position-based c/μ methods were assessed to check the robustness of c/μ obtained from linear regression. For the UTR/TRS, larger differences were observed in segments with good molecular clocks (0.03-0.18, or 1.74%-8.11%) and without molecular clocks (0.00-0.71, or 0.00%-55.56%). For the TR, smaller differences were observed for segments exhibiting molecular clocks (0.00-0.07, or 0.00%-12.50%) and larger differences were observed in segments without molecular clocks (0.01-0.54, or 0.53%-38.57%). In summary, despite the larger percent differences of some segments (e.g., Orf6, Orf7a, Orf10, Orf1ab TRS-L, Orf3a TRS-B, etc.), 21 out of 25 coding segments and only All-UTR, All-TRS and Orf1ab 5'UTR exhibited a good molecular clock, indicating their selection type assignment by time-based c/μ is consistent with the position-based c/μ values. The remaining four coding segments (Orf7a, Orf10, NSP5 and NSP7) and each UTR and TRS should use the position-based c/μ for their accurate selection type assignment.

Discrete L-shaped probability distributions of c/μ were observed for all of 49 segments. Our previous paper [13] has shown that an L-shaped distribution curve rather than a Poisson distribution curve centered at ($\frac{c}{\mu} = 1$) was observed for all 49 segments (**Figures S6-S8** [13]). The c/μ of these discrete distributions was scaled down by 5.46-fold to reflect the new μ in this study (**Figure S4-S5**). To facilitate the assessment for different mutation types, the cumulative probability distribution function ($P(x) = \sum_{\frac{c}{u}} P(\frac{c}{u} \geq x)$) was also generated for all 49 segments (**Figure 5** and **Figure S6-S7**). Although the c/μ scaling does not change the fraction of conserved sites ($\frac{c}{\mu} = 0$), it does change the absolute fraction of strictly neutral, nearly neutral, beneficious, and deleterious mutation sites. The proportion of the strict neutral, nearly neutral, beneficial, deleterious mutations were determined from L-shaped PDRSR as described in the methods section (**Table 3-4 and Figure** 7).

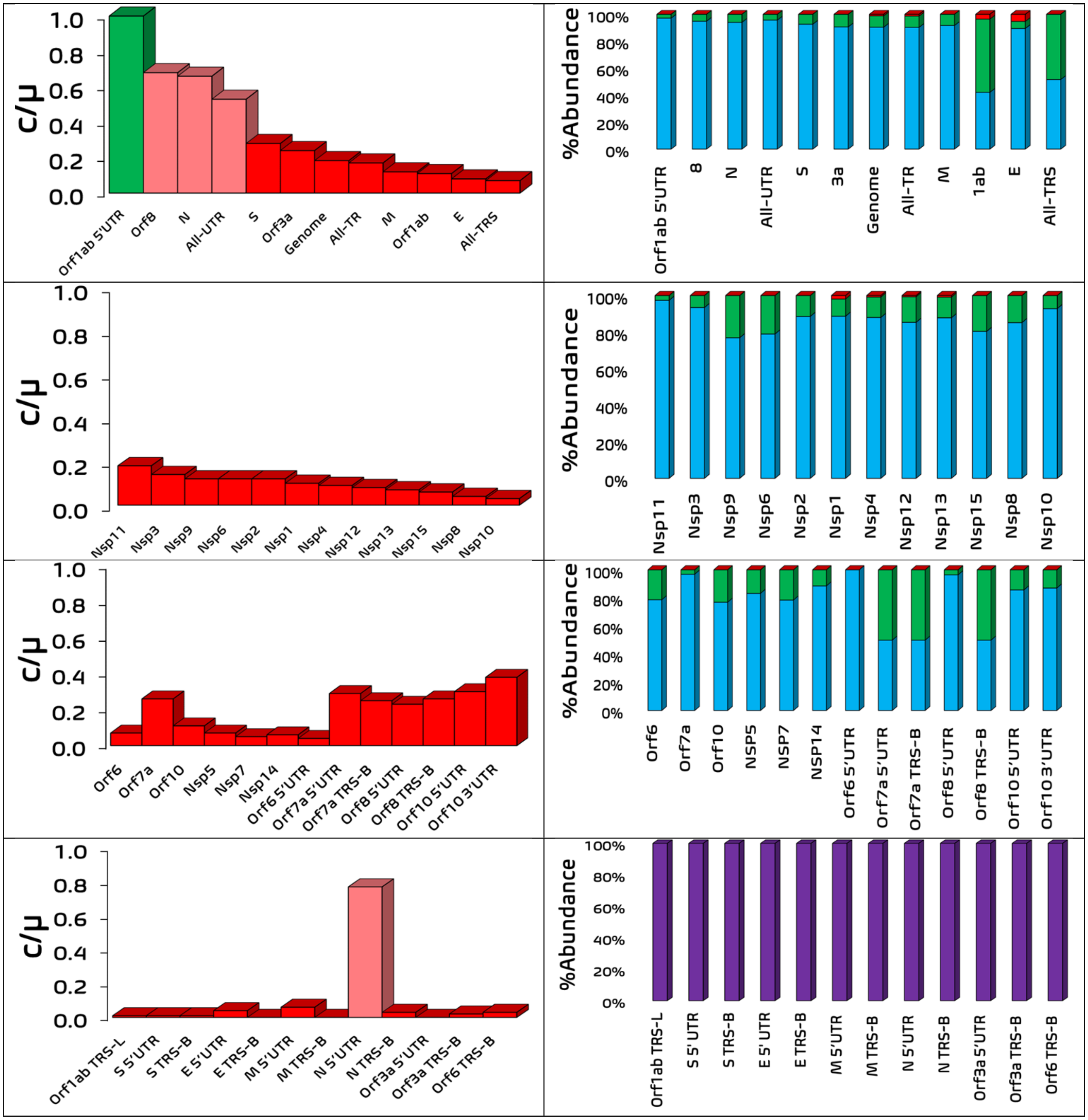


Figure 7. c/μ values (left column) and abundance of selection type (right column) for segments exhibiting molecular clock features (first and second rows) and non-molecular clock features (third and fourth rows). (Left column) Green, light red and dark red represent segments under neutral selection, weak negative selection and strong negative selection, respectively. (Right column) Blue, green, red and purple (where applicable) colors represent percentage of sites

under weak negative, weak positive, strong positive and strong negative selection, respectively. See **Figures S6-S7** for the graphs of segments decomposed into other selection types. See **Table S5** for tabulated values.

Negligible fraction of strictly neutral mutation sites ($f_0 \approx 0$) suggests strictly neutral mutations due to genetic drifting are very rare and non-neutral mutations due to natural selection are predominant. The discrete distribution allows us to precisely determine the fraction of strictly neutral mutations ($f_0 = P(\frac{c}{u} = 1)$). For the 24 segments with molecular clock, f_0 ranges from 0.03% of the genome to 1.64% of All-TRS (**Table 3**). For the 25 segments without molecular clock, f_0 ranges from 0.00% of 7 TRSs and 8 UTRs to 16.67% of Orf7a 5'UTR, 14.29% of Orf7a and Orf8 TRS-B, 0.85% of Orf10, 0.75% of Orf8 5', 0.54% of Orf6, 0.44% of Orf10 3'UTR, 0.40% of Nsp7, and 0.05% of Nsp5 (**Table 4**). Therefore, most of the mutations are non-neutral mutations due to natural selection.

The ratio of nearly neutral to strictly neutral mutations is greater than 2 up to 817 for 34 applicable segments, supporting that nearly neutral mutations play a more important role than strictly neutral mutations. This f'_0:f_0 ratio of the genome is ~577. For the 24 segments with molecular clock, this ratio spans from ~4 of All-TRS, to ~468 of Orf1ab, to ~587 of All-TR, and to ~817 of All-UTR (**Table 3**). For the 25 segments with molecular clock, this ratio is only available for 10 segments, ranging from ~2 of Orf7a 5'UTR, Orfa TRS-B and Orf8 TRS-B; to ~13 of Orf6, Orf10, and Nsp7; to ~31 of Orf8 5'UTR; to 87 of Orf10 3'UTR; to 129 of Nsp5; and 190 of Orf7a (**Table 4**).

The ratio of mutations under weak negative to weak positive selection is greater than 1 up to 31 for 36 applicable segments, suggesting that although weak negative sites are more abundant than weak positive sites, weak positives do exist. This WP:WN ratio of the genome is ~5. For the 24 segments with molecular clock, this ratio is available for 23 segments, spanning from ~1 of Nsp8 and N; ~2 of All-UTR; to ~3 of S and All-TRS; to 5-6 of All-TR, Orf3a, Nsp6, and Nsp11; to 9-11 of Orf1ab, E, M, Orf8, Nsp2-4, and Nsp12; to ~14 of Nsp15; and 31 of Nsp9 (**Table 3**). For the 25 segments with molecular clock, this ratio is only available for 12 segments, ranging from ~1-2 of Orf10 3'UTR, Orf8 TRS-B, Orf7a TRS-B, Orf7a 5'UTR, and Nsp7; to ~5-9 of Orf10 5'UTR, Orf6, Nsp5, Orf8 5'UTR and Orf10; to ~16-18 of Orf7a and Nsp14 (**Table 4**).

The ratio of mutations under weak negative to positive selection is greater than 1 up to 37 for all 37 applicable segments except for Orflab (0.73) and NSP7 (0.33) and Orf8 TRS-B (1.00), suggesting although more substitutions result from random fixation of slightly deleterious mutations, the beneficial substitutions due to positive selection are not so rare and could be more than slightly deleterious mutations for a few segments. This WN:P ratio of the genome is ~9. For the 24 segments with molecular clock, its spans from 0.73 of Orflab to 37.34 of Nsp11 with a mean of 11±9 (**Table 3**). For the 25 segments with molecular clock, this ratio is only available for 12 segments, ranging from 0.33 of Nsp7 to 18.15 with a mean of 6.4±5.9 (**Table 4**). This ratio is almost identical to its WN:WP ratio, because there is almost no SP site for many of them.

While NNST (union of NNBST and NNUST) are consistent with 6 empirical features of these 49 segments, KNT, ONNT and ST have many violations. The molecular clock feature (Yes or No), the abundance of weak positive sites (criteria 1: %WP ≫ 0), the abundance of weak negative sites (criteria 2: %WN ≫ 0), the abundance of nearly neutral sites (criteria 3: %NN ≫ 0), the abundance of positive sites (criteria 4: %P ≫ 0), and the ratio of nearly neutral to strictly neutral sites (criteria 5: fo'≫ fo) of 49 segments (**Table 3-4**) are used to evaluated the five evolutionary theories in **Table 3-4** (KNT, ONNT, ST, NNBST and NNUST). The summary of this evaluation is tabulated in **Table 5**.

Table 5. Evaluation of KNT, ONNT, ST, NNBST and NNUST for all segments. See **Table S9** for the evaluation at each segment. The number and percentage of violations calculated for applicable segments are presented.

Theory	NA	APP	1	2	3	4	5
KNT	25	24	0	0	24 (100%)	16 (66%)	24 (100%)
ONNT	24	25	13 (52%)	0	0	4 (16%)	15 (60%)
ST	24	25	0	22 (88%)	22 (88%)	0	0
NNBST	25	24	0	0	0	0	0
NNUST	24	25	0	0	0	0	0

*NA: Not applicable to explain molecular clock; APP: Applicable to explain molecular clock.

Violation descriptors: 1: %WP ≫ 0; 2: %WN ≫0; 3: %NN ≫0; 4: %P ≫0; 5: fo'≫ fo.

KNT, NNBST: Applicable if R^2>0.6000; ONNT, ST, NNUST: Applicable if R^2<0.6000 (L+).

KNT is not consistent with the lack of molecular clock for the 25 segments (**Table 4**). For the 24 segments with molecular clock, KNT is not consistent with criteria 3, 4 and 5 in 100%, 66%, and 100% cases, respectively (**Table 3&5**). In other words, KNT overestimates the abundance of strictly neutral sites (criteria 5), underestimates the abundance of near neutral sites (criteria 3) and positive sites (criteria 4). ONNT is not consistent with molecular clock for the 24 segments (**Table 3**). For the 25 segments without molecular clock, ONNT is not consistent with criteria 1, 4 and 5 in 52%, 16% and 60% cases, respectively (**Table 4 & 5**). In other words, ONNT overestimates the abundance of strictly neutral sites (criteria 5), underestimates the abundance of weak positive sites (criteria 1) and positive sites (criteria 4). ST is not consistent with molecular clock for the 24 segments (**Table 3**). For the 25 segments without molecular clock, ST is not consistent with criteria 2 and 3 in 88% and 88% cases, respectively (**Table 4 & 5**). In other words, ST underestimates the abundance of weak negative sites (criteria 2) and near neutral sites (criteria 3). While NNBST is not consistent with the lack of molecular clock for the 25 segments (**Table 4**), NNUST is not consistent with molecular clock for the 24 segments (**Table 3**). Without the strong assumptions in KNT, ONNT and ST, NNBST and NNUST have no violation on criteria 1-5.

Scaling up of μ likely increases WN:P and f'_0:f_0 for the genome. What if the true μ is higher than the substitution rate of Orf1ab 5'UTR (5.46-fold of GSR)? Given the cumulative probability distribution of the genome, the change of WN:P and f'_0:f_0 can be determined when μ is scaled up beyond the 5.46 scaling factor (μ'). The scaled-up μ' is determined from the cumulative probability of the original distribution with GSR as μ ($u' = \frac{1}{P(x)} = f'_0$), leading to determination of WNB and WPB; the factions of strictly neutral, near neutral, weak negative, weak positive; and the two ratios (**Figure 8**).

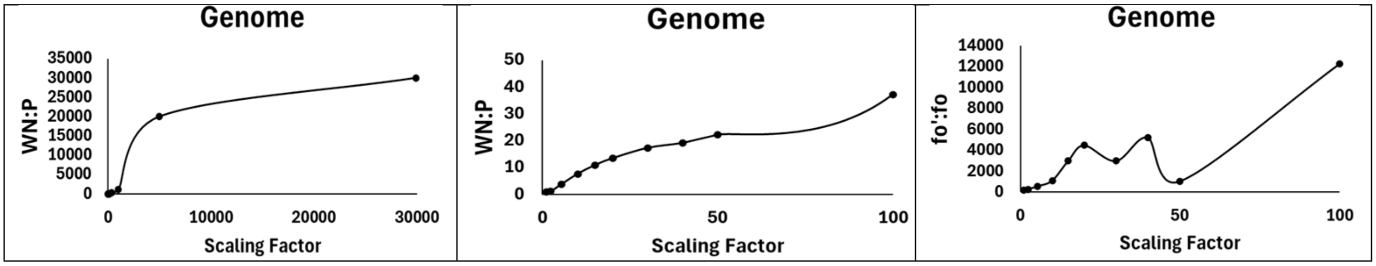


Figure 8. The impact of c/μ scaling factors on WN:P (left and middle) and fo':fo (right) parameters. See **Tables S7-S8** for tabulated parameters.

As μ' increases, the WN:P and f'_0:f_0 ratios of the gnome increase monotonically, being more consistent with NNBST than KNT and ST.

The balancing condition explains molecular clock feature of a polyprotein gene (Orf1ab) of SARS-CoV-2. The balance between the sites of higher substitution rate under more positive selection and the sites of lower rate under more negative selection explained the observed molecular clock at the genome level of SARS-CoV-2 in our previous study[13]. This mechanism also works at a segment level for its polyprotein gene (Orf1ab) including 15 NSPs. The monthly total NT percent variation of NSP-15 and Orf1ab was plotted together to show which NSP is under more positive or negative selection (**Figure 9**).

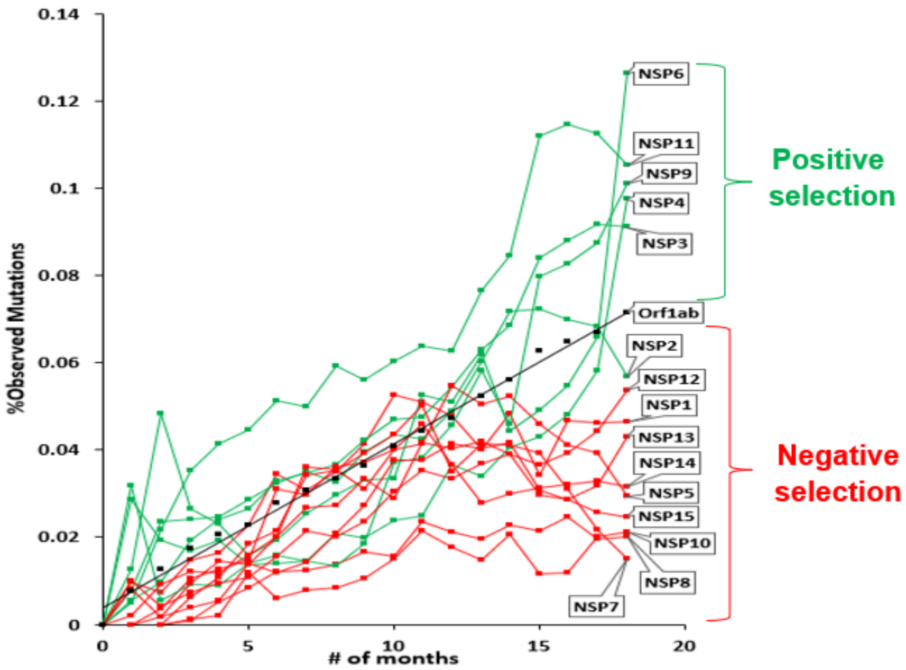


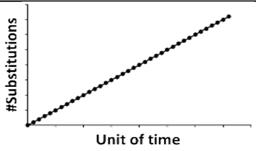
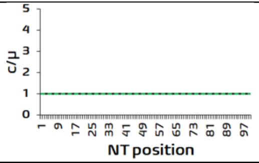
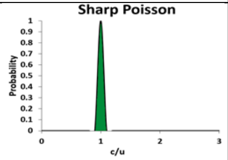
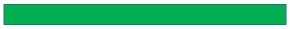
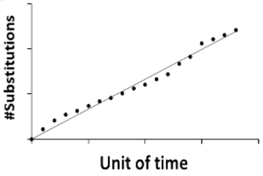
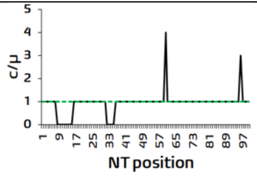
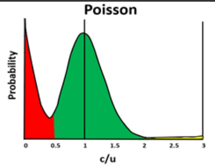
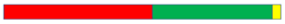
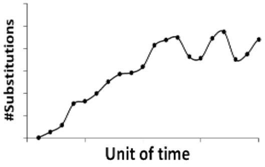
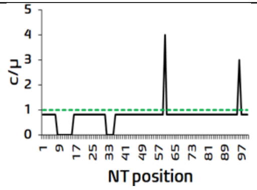
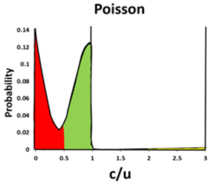
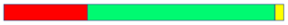
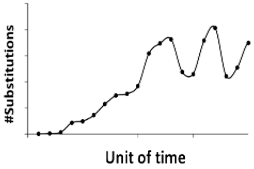
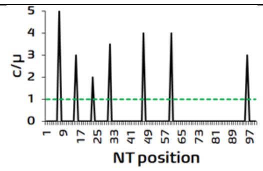
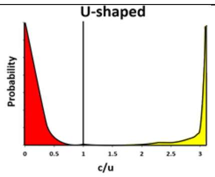
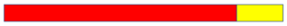
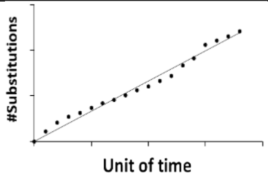
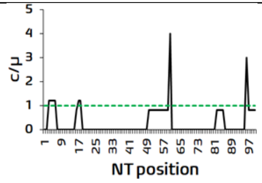
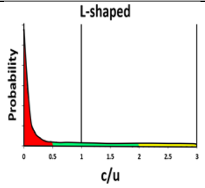
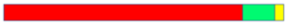
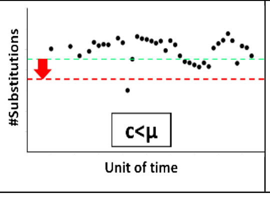
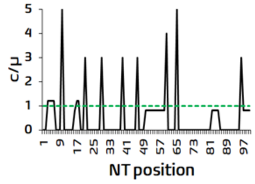
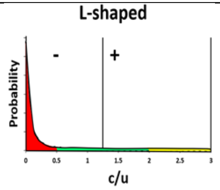
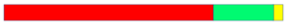
Figure 9. Demonstration of NNBST at the subgenetic level of Orf1ab and NSP1-15. Monthly percent total nucleotide substitution rate for Orf1ab and NSP1-15 genes over 19 months. NSP segments exhibiting faster or slower substitution rates relative to Orf1ab are under relative positive and negative selection with respect to Orf1ab (neutral selection).

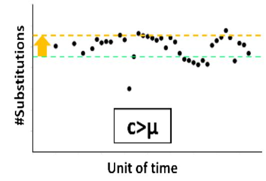
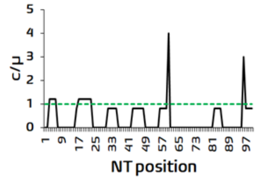
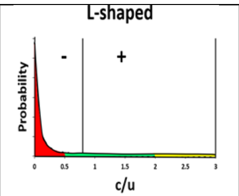
Due to the observed molecular clock feature of the Orf1ab SSR ($c/\mu = 0.11$ and $R^2 = 0.9854$), the faster and slower SSRs of NSP1-15 due to more positive and negative selection can balance out and produce the Orf1ab SSR, similarly as the individual SSRs produce the GSR. From this, NSP (2, 3, 6, 9 and 11) seem to undergo more positive selection ($c/\mu > 0.11$), whereas NSP (1, 4, 5, 7, 8, 10, 12, 13, 14 and 15) seem to undergo negative selection ($c/\mu < 0.11$). Hence, NNBST can explain the seemingly time-independent total NT substitution rate of Orf1ab.

Discussion

The c/μ test provides a powerful tool to resolve the "neutralist-selectionist" debate. If the assumption of a time-independent and site-independent mutation rate (μ) is true (**Figure 3**), its time-dependence (molecular clock or not) and site-dependence (SRSRS or PDRSR of c/μ) for a species will directly tell which molecular evolution theory (ST, ONNT, KNT, NNBST and NNUST) does the species follow (**Table 2**).

Table 2. Summary of two substitution rate models (GSRM and SSRM) and five theories (ST, KNT, ONNT, NNBST, NNUST) of molecular evolution in terms of molecular clock, SRSRS, PDRSR and percentage of sites under different selection types (strong negative/red, neutral/green, nearly neutral including (weak negative and weak positive)/light green, and strong positive/dark yellow).

SSRM	Theory	Molecular Clock	SRSRS	PDRSR	% of Selection types
N	GSRM				 $c = \mu, f_0 = 1$
Y	KNT				 $c = \mu f_0$ $f_0 = 1 - f^-$ $f^+ \approx 0, c^- \approx 0, f_0 \approx f'_0 \gg f^+$ c/μ<1
S	ONNT				 $c = \mu(f_0 + q(1 - f_0))$ $f^+ \approx 0, c^- \approx 0$ $f_0^- \gg f_0^+ \approx 0, f'_0 > f_0 \gg f^+$ c/μ < 1
	ST				 $c = c^+ f^+$ $f^+ = 1 - f^-$ $f'_0 \approx 0, c^- \approx 0, f^+ \gg f'_0$ c/μ>1 or c/μ<1
	NNBST				 $c = \mu f'_0$ $\lvert c^+ f^+ - \mu f'_0 \rvert = \lvert \mu f'_0 - c^- f^- \rvert \neq 0$ $f^+ > 0, f'_0 \gg f_0$, c/μ<1
	NNUST Negative Selection				 $c \neq \mu f'_0$ $\lvert c^+ f^+ - \mu f'_0 \rvert \neq \lvert \mu f'_0 - c^- f^- \rvert \neq 0$ $f'_0 > 0, f'_0 \gg f_0$, c/μ<1

	NNUST Positive Selection				🟥🟩🟨 $c \neq \mu f'_0$ $\lvert c^+f^+ - \mu f'_0 \rvert \neq \lvert \mu f'_0 - c^- f^- \rvert \neq 0$ $f'_0 > 0, f'_0 \gg f_0,$ c/μ>1

* GSRM (Genomic Substitution Rate Model); SSRM (Genomic Substitution Rate Model);Theory: ST (Selectionist theory); KNT (Kimura's Neutral Theory); ONNT (Ohta's Nearly Neutral Theory); NNBST (Near-Neutral Balanced Selectionist theory); NNUST (Near-Neutral Unbalanced Selectionist Theory; NNST (Near-Neutral Selectionist Theory): including NNBST and NNUST; $c = c^+f^+ + \mu f'_0 + c^-f^-$ and $1 = f^+ + f'_0 + f^-$ under SSRM/NNST: the total substitution rate c for a given genomic sequence, segment or collection of sites, can be described as the weighted sum of the substitution rates of sites under positive selection (c^+f^+), near neutral selection ($\mu f'_0$) and negative selection (c^-f^-); c : total substitution rate; $c^+, \mu,$ and c^-: Substitution rate of beneficial, near neutral, deleterious sites; $f^+,$ $f'_0,$ and f^-: the fraction of beneficial, near neutral and deleterious sites. f_0^- and f_0^+: Fraction of weak negative and weak positive; f_0: fraction of selectively neutral sites in KNT and ONNT (i.e. c/μ=1). q: time-dependent proportion of deleterious mutations that reach fixation; and it depends on the effective population size (N_e), environmental conditions and selective pressures; **Molecular clock present (Yes/No); SRSRS: Site relative rate spectra; PDRSR: Probability distribution of the relative substitution rate; SN: Strong negative selection (**c/μ < WN)**; WN: Weak negative selection (**wn ≤ c/μ < 1)**; N: Selectively Neutral selection (**c/μ = 1)**; WP: Weak positive selection (**1< c/μ ≤ wp)**; NN: Near Neutral selection (WN+WP): SP: Strong positive selection (**c/μ > wp).**

As for SARS-CoV-2, molecular clock was observed only for 24 out of 49 segments (**Figure S2-S3**), and an L-shaped PDRSR was observed for all 49 segments (c/μ_i) (**Figure S4-S5**), **First,** the dramatic change from the spike-shaped PDRSR of c/μ centered at 1 ($f_0 = 1$) under GSRM (**Figure 9A**) to the L-shaped PDRSR of c/μ (**Figure 9B)** show the power of natural selection rather than genetic drifting.

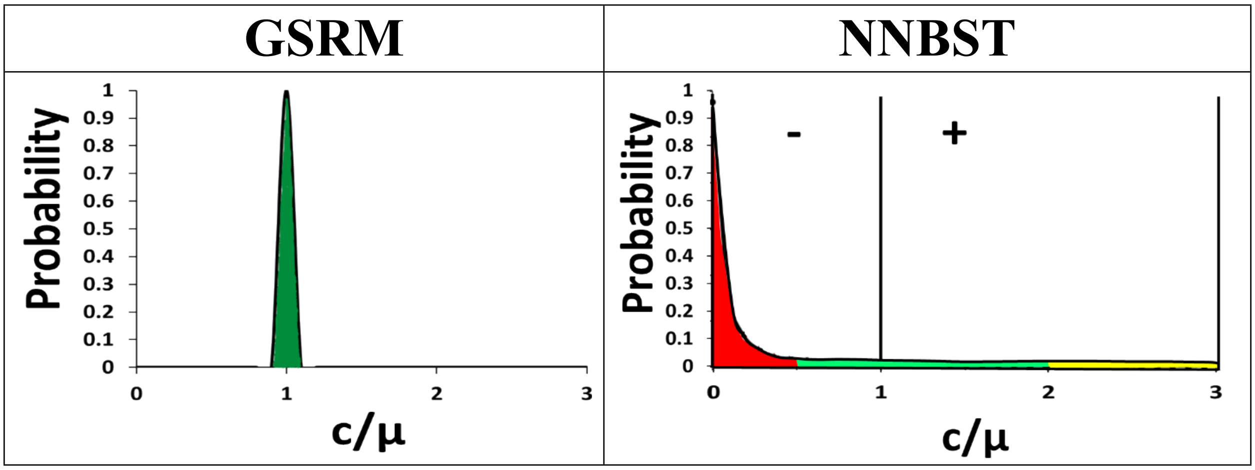


Figure 10. Comparison of the Poisson-shaped PDR of c/μ predicted by GSRM and the L-shaped PDR of c/μ obtained from the SARS-COV-2 genomic datasets explained by NNBST. Selection type is denoted by color (red: negative selection; green: neutral selection; yellow: positive selection).

Indeed, f_0 of this virus's genome is just 0.03% rather than 100% under GSRM. KNT and ONNT appear to overestimate the fraction of strict neutral mutations and thus the role of genetic drifting. This overestimate is clearly demonstrated by the high ratio ~577 of near neutral to strictly neutral sites of this virus's genome. Both natural selection and genetic drifting play important roles in the fixation of near neutral mutations. **Second**, although weak negative sites are typically more abundant than weak positive sites, weak positives do exist. WP:WN ratio of this virus's genome is ~5. The early version of ONNT with a shift model predicts only weak negative sites, the later version with a fixed model predicts both weak negative and weak positive sites[16]. Using the simplest form of the fixed model, Gillespie has shown that 50% of the substitutions that become fixed are deleterious and the other 50% are advantageous[17]. **Third**, although weak negative sites are typically more abundant than the positive sites for most of the segments, the exception for a few segments exists. While WN:P ratio of this virus's genome is ~9, that of Orflab (0.73) and NSP7 (0.33) is less than 1. While ST might underestimate the important role of genetic drifting in fixing slightly deleterious mutations, KNT and ONNT might underestimate the critical role of positive selection in a few segments and the negative selection in fixing slightly deleterious mutations. **Fourth,** the balance condition (**Equation 8a**) explains molecular clock. This molecular clock is due to nearly neutral mutations rather than strictly neutral mutations under the balance condition. The early version of ONNT predicts a population-size dependent substitution rate, since the lower purification efficiency on slightly deleterious mutations in a smaller population would lead to their higher substitution rate[16]. Yet, as the population size of SARS-COV-2 changes over seven magnitude (10^0-10^7), its GSR is still close to constant[13]. If the lower fixation efficiency on slightly beneficious mutations in the small population would lead to a lower substitution rate, then the higher rate from the slightly deleterious mutations will be cancelled out, and thus the population-size independent molecular clock can be observed. This mechanism is allowed in NNBST. **Fifth**, if the balance condition is not satisfied, then molecular clock will not be observed as predicted by NNUST. However, there is no good reason to support the positive selection must be balanced with the negative selection, thus a balanced selection should be less frequent than an unbalanced selection. In fact, lack of a constant molecular clock was observed in 25 of 32 RNA viruses [18]. Our unpublished data show that only ~5% of 2000 viruses show molecular clock. Put together, we proposed Near-Neutral Selectionist Theory (NNST) to combine ST with ONNT to acknowledge the important roles of both genetic drifting and natural selection in the molecular evolution. As to SARS-COV-2's genome with a scaled-up μ (5.45-fold of the original one[13]), 84.21%, 14.28%, 1.33% and 0.19% mutations are under strong negative, weak negative, weak positive and strong positive selection, respectively. Of the substitutions that fix in the population, 90.41% (14.28%), 8.41% (1.33%) and 1.19% (0.19%) are slightly deleterious, slightly advantageous and strong advantageous. While genetic drifting and the purifying selection of natural selection plays an important role in fixing 98.82% of near neutral mutations in the population, the positive selection of natural selection also plays an important role in fixing 9.6% of advantageous mutations.

Limitations of c/μ test and future directions are briefly discussed as follows. **1)**. Our assumption of time-independent site-independent global genomic mutation rate (μ) remains to be validated from the replication experiments. **2).** When ancestral sequences are not known, a phylogenetic tree must be inferred from a set of sequences and ancestral sequences with a root sequence must be reconstructed to get the reference for mutation counting. **3).** c must be sampled from a representative population of sufficient size; therefore, it might not capture the founder effect and genetic bottleneck conditions when population size is small. Accurate establishment of c should be obtained from a genomic dataset given **a)** sufficient sample size is obtained, **b)** sufficient evolution time has passed, and **c)** samples are representative of the effective population. Sufficient sample sizes increase the likelihood of representing the effective population by mapping the genetic variation accumulated over evolution time. Moreover, ensuring enough evolution time has passed allows mutations to become fixed in the effective population and the establishment of a molecular clock feature (if possible). Finally, a large, representative sample size can prevent the pitfall of the Founder Effect, which unique mutations and phenotypes associated with breakaway subpopulations may not represent the effective population. **4).** c/μ is the overall indirect fitness measurement for nucleic acid and protein sequence from experimental measurements, but it does not tell the mechanism of fitness change (e.g. selection coefficients) that can be used for predicting the molecular adaption.[14] While the c/μ values for a given segment and NT/AA site are useful for measuring the relative fitness change, additional analysis is required to probe the nature of such change for getting selection coefficients.[14] For instance, molecular modeling (e.g., molecular docking and dynamics simulations) can provide insight into mutation effects on protein/nucleic acid structure and drug binding interactions at the atomic level, and/or be experimentally characterized.

Conclusion

In this comprehensive analysis of the relative substitution rates (c/μ) across various segments of the SARS-COV-2 genome, we have demonstrated notable departures from traditional evolutionary models. By extending our investigation to 26 translated regions, 12 untranslated regions, and 10 transcriptional regulatory sequences, we observed L-shaped probability distributions of c/μ and variable presence of molecular clocks, supporting both our Near-Neutral Balanced Selectionist Theory (NNBST) and Near-Neutral Unbalanced Selectionist Theory (NNUST). Our findings significantly challenge the existing paradigms of evolutionary biology by demonstrating that neither strictly neutral mutations nor solely beneficial mutations dominate the genomic landscape. Instead, nearly neutral mutations including both slightly deleterious and slightly beneficious, influenced by both genetic drift and natural selection, predominate. This nuanced understanding suggests that the fixation of mutations is influenced by a balance of selective pressures rather than by simple stochastic events or direct selection alone. The discovery of L-shaped distributions across all studied segments, coupled with the variable presence of molecular clocks, underscores the dynamic nature of viral evolution and the need for a revised theoretical framework. The NNST (Near-Neutral Selectionist Theory) that we propose offers such a framework, combining elements of traditional neutral and selectionist theories to more accurately reflect the evolutionary processes observed in SARS-COV-2 and potentially other RNA viruses. Our research points to the necessity for a reevaluation of the "neutralist-selectionist" debate, advocating for a more integrated approach that considers the spectrum of mutation effects from deleterious to beneficial. This approach not only enhances our understanding of viral evolution but also may inform future therapeutic strategies by highlighting the evolutionary potential and constraints within viral genomes. As we continue to refine our theoretical models and expand our empirical datasets, it becomes increasingly clear that the molecular mechanisms governing evolution are far more complex and interconnected than previously assumed. The work presented here marks a significant step towards resolving long-standing debates in evolutionary biology and sets the stage for future investigations that will further elucidate the intricate dance of mutation, selection, and drift in shaping the genomes of living organisms.

Methods

A replication-selection model for describing molecular evolution of a virus, definition of the substitution-mutation rate ratio (c/μ) to quantify the fitness change due to mutations, use of c/μ probability distribution to define different evolution empirical models including Genomic Substitution Rate Model (GSRM), Segment Substitution Rate Model (SSRM); and evolutionary theories including ST, KNT, ONNT and NNBST, and the pipelines to calculate c/μ and its change over time, have been described in our previous paper[13]. In this paper, we concisely describe them (**Table 1 and 2);** and focus on deriving the general equations to define them.

Table 1. Definitions of Major symbols.

n_c/Nc	Census Population of a virus
Ne	Replicated Population of a virus before the selection
G	Number of genomes
t	Collecting date of a genome
Δt	$\Delta t=t_{max}-t_{min}$
m_{ij}	Substitution count matrix (Time, Position)
P_{mut}	Mutation probability
P_{fix}	Fixation/substitution* probability
μ	Number of mutations per nucleotide site per unit time/mutation rate
c	Number of substitutions per nucleotide site per unit time/substitution rate
c/μ	Substitution/mutation rate ratio/relative substitution rate
N	Number of nucleotide sites in the sequence
NT	Nucleotide
TR	Translated Region
UTR	UnTranslated Region
TRS	Transcriptional Regulatory Sequences
GSRM	Genomic Substitution Rate Model
SSRM	Segment Substitution Rate Model
ST	Selectionist (Neo-Darwinian) Theory
KNT	Kimura's Neutral Theory
ONNT	Ohta's Nearly Neutral Theory
NNST	Near-Neutral Selectionist Theory
NNBST	Near-Neutral Balanced Selectionist Theory
NNUST	Near-Neutral Unbalanced Selectionist Theory

The replication-selection model to describe molecular evolution under natural selection, the relative substitution rate (c/μ) to quantify the fitness change of mutations, the discrete c/μ probability distribution to define different models/theories of molecular evolution. A replication-selection model was introduced to describe the evolution of a virus in **Figure 2** of our previous paper[13]. An elaborated version has been generated (**Figure 2A**) and will be described here. The numbers used to represent different viral populations are arbitrary and are kept small here for simple calculations; more realistic conditions likely contain viral populations several orders of magnitude higher. A nascent virus census population containing n members (n_c=3) undergoes viral replication with a mutation rate (μ) that induces mutations at three different sites with equal probability (P_{mut} = 3/12, 3/12, 3/12) over the time period (t). After replication, a mutant population is produced (N_e = 12) and is subjected to either positive selection (yellow box), neutral selection (green box) or negative selection (red box) on the first NT position from the host (e.g., host immune response, host dies, vaccinations). The new population surviving selection is observed (N_c = 4) and bears fixed mutations with a probability for the three different mutants under positive selection (P_{fix} = 2/4, 1/4, 1/4), neutral selection (P_{fix} = 1/4, 1/4, 1/4) and negative selection (P_{fix} = 0/4, 1/4, 1/4) for the first NT position. The process is cyclic and continues for successive viral generations within the host and following infections with other hosts. It must be noted that the population N_c is used for genomic and mutation analyses, whereas population N_e is much harder to observe unless an in vitro experiment is conducted. Although if performed, μ could be experimentally determined and provide a more accurate c/μ per position and genomic segment.

One nuance to this process is that the total population count before and after selection can change, but the relative abundance of different genotypes should stay the same under neutral selection (i.e. an equal fixation probability ($1/N_e$) for each individual virus in the replicated population N_e).[4] For example, a hypothetical population after mutation (N_e = 12) contains 3 wild-type sequences (3/12 or 1/4), 3 red mutants, 3 green mutants and 3 blue mutants, corresponding to their relative abundance (1:1:1:1). After neutral selection, only 25% of the sequences survive ($N_c = 4$) of 1 wild-type, 1 red mutant, 1 green mutant and 1 blue mutant (**Fig. 2A** green box), however the relative abundance is the same (1:1:1:1). Therefore, $P_{mut} = P_{fix} = ¼$ for red mutation, thus, $\mu = c$ by dividing the same time period (t). This relationship is general so that when the rate of mutation generation (μ) is equal to the rate of substitution (c) in the population suggests neutral selection without a preference ($c/\mu=1$).[4] Only when neutral selection is well defined, then natural selection can be defined. If the substitution rate of the red mutation is higher than the rate of its generation ($c/\mu>1$), the mutation must be beneficial to increase the fitness of the virus (positive selection **Fig2A**, yellow box). If the substitution rate of red mutation is lower than the rate of its generation ($c/\mu<1$), the mutation must be deleterious (i.e. negative selection **Fig. 2A,** red box). Therefore, c/μ can quantify the fitness change of mutations at a specific NT site of a genome.

To define μ, two prior assumptions are made in the replication step of our replication-selection model (**Figure 3**).

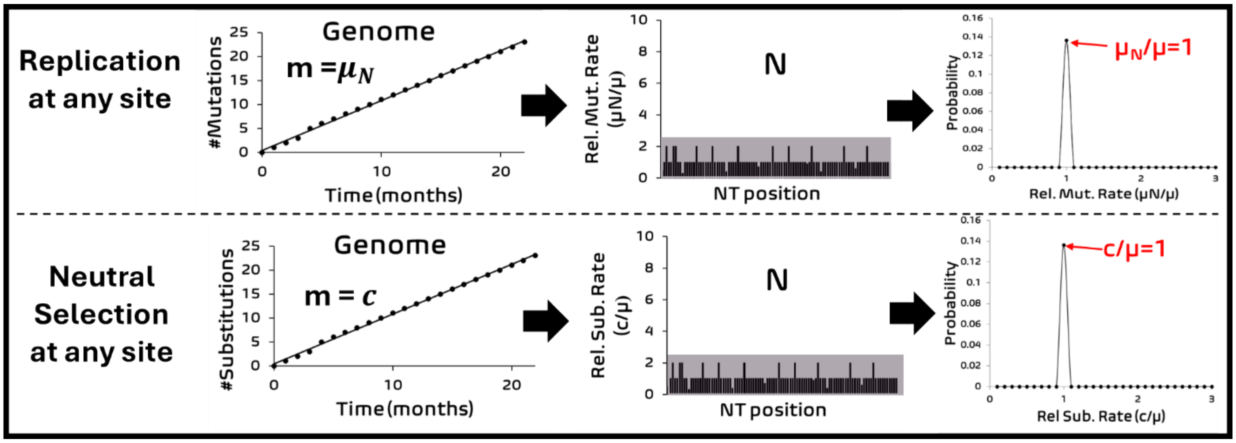


Figure 3. The viral genomic sequence under neutral selection. During replication, the viral polymerase induces copy errors at rate μ at each nucleotide site in the copied viral sequence strand. The total nucleotide sequence length (N) is the sum of all nucleotide sites in the translated and untranslated regions. Mutations are fixed into each site at rate c. The resulting probability distribution of the relative mutation rate (μ_N/μ) for a given site under neutral selection is described as a Poisson distribution curve.

For viral replication, 1) the viral replication machinery (e.g., polymerase with/without proofreading enzymes) has no preference for where it induces copy error and does so at any NT site with a phenomenological rate μ. 2) The polymerase structure does not significantly change from induced mutations; thus, the polymerase error rate (μ) does not significantly change over the measuring time period (i.e. time-independent constant leading to strict molecular clock in **Figure 3 left top panel**). Nonetheless, these two prior assumptions are shared among many mutation models[14, 19].

Key features of the discrete c/μ probability distribution define different models/theories of molecular evolution (**Table 2**). The c/μ of each NT site in a genome of c/μ (defined as Site-Relative Substitution Rate Spectrum (SRSRS)) spells out the selection pressure of the mutations at the NT site. SRSRS (**Figure 3 middle**) can be equivalently transformed into a discrete Probability Distribution of Relative Substitution Rate (PDRSR) for all NT sites of the genome including sampling errors (**Figure 3 right**). Assuming all NT sites undergo strictly neutral selection only (i.e. the fraction of strictly neutral sites is one, $f_0 = 1$; Genomic Substitution Rate Model/GSRM[13]; the Strict neutrality hypothesis [11]), the observed mutation rate at any NT site *j* (substitution rate, c_j) is equal to the spontaneous mutation rate in individual organisms ($c = \mu$ **Figure 3**) [3, 4]. Therefore, the substitution rate for any genetic segment, UTR, TRS, sub-genetic segment is a constant that is equal to the spontaneous mutation rate (μ), leading to a global molecular clock.[2] The temporal and site features outputted from the mutation model in the mutation step will be observed again under neutral selection **(Figure 3 Bottom Panel)**. Put together, empirical hallmarks of GSRM are the global molecular clock, discrete almost uniform distribution of c/μ at any site of the genome, and Poisson distribution centered at c/μ=1 with a broadened peak due to random sampling error (**Table 2**).

The Segment Substitution Rate Model was introduced to allow the time-dependent natural selection pressure acting on the genome, thus causing the deviation of the substitution rate of a specific segment or a NT site *j* from the constant mutation rate (i.e. $c_j/\mu \neq 1$). Various theories have been proposed to capture the relative contribution between genetic drifting forces (neutral selection) and natural selection forces in determining the evolutionary substitution rate.
In KNT (**Table 2**), while most NT sites are under negative selection ($f^- > f_0$; f^-: the fraction of deleterious sites), the sites under strictly neutral selection (c/μ=1) are much more abundant than the sites under positive selection (c/μ>1, $f^+ \approx 0, f_0 \gg f^+$; f^+: the fraction of beneficious sites) [20]. The hallmark of KNT is the molecular clock due to large number of neutral sites and a Poisson distribution of centered at c/μ=1 (**Table 2**). Along the same line of KNT, ONNT [7, 8] predicts while most NT sites are under negative selection ($f^- > f_0$), the sites under strictly neutral selection (c/μ=1) or nearly neutral selection (c/μ~1, f'_0:the fraction of near neutral sites) aremuch more abundant than the sites under positive selection ($f^+ \approx 0, f'_0 > f_0 \gg f^+$). Hallmarks of ONNT are population-size dependent substitution rate (***Eqn* 6**: $c = \mu(f_0 + q(1 - f_0))$; q:

time-dependent proportion of deleterious mutations that reach fixation, which is dependent on the effective population size (N_e), environmental conditions and selective pressures[11]) and an asymmetric distribution around c/μ=1 due to mostly slightly deleterious near neutral mutations (**Table 2**). In ST [2], NT positions largely exhibit natural selection ($c_j/\mu \neq 1$) and few sites under strictly neutral selection (**Table 2**). Hallmarks of ST are time-dependent substitution rate, and a bimodal distribution of c/μ showing one peak of negative selection (c/μ<1) and the other peak of positive selection (c/μ>1). In NNBST, hallmarks are L-shaped distribution of c/μ and molecular clock due to the cancelling out of the rate of change of the sites under more positive selection by that of the sites under more negative selection (defined as a balance condition that will be derived later). The balance condition is more general than the original assumptions in KNT that beneficial sites are so rare. In NNUST, hallmarks are L-shaped distribution of c/μ and lack of molecular clock due to the failure of the balance condition (**Table 2**). Again, NNUST is more general than ST without assuming that nearly neutral sites are so rare. If the average substitution rate is faster than the mutation rate, positive selection dominates the overall selection. Otherwise, the average substitution rate is slower than the mutation rate, thus negative selection dominates the overall selection.

While the balance condition leading to molecular clock in NNBST allows larger fraction of positive mutations than KNT does, the failure of the balance condition leading to lack of molecular clock allows a larger fraction of nearly neutral mutations than ST does. The contribution of each selection/mutation type towards the total substitution rate for a given genomic sequence, segment or collection of sites, can be described as the weighted sum of the substitution rates of sites under strong positive selection (c^+f^+), near neutral selection ($\mu f'_0$) and strong negative selection (c^-f^-) following NNST:

$$c = c^+f^+ + \mu f'_0 + c^-f^- \text{ (7a) and } 1 = f^+ + f'_0 + f^- \text{ (7b)}$$

Although the second neutral site term could be time-independent where the mutation rate (μ) is constant and f'_0 could be time-independent constant, the overall substitution rate is more likely to be time-dependent because of time-dependence of the first strong positive site term and the third strong negative site term. However, the total rate can be time-independent under the balance condition of NNBST where the increase of the rate due to the advantageous mutations (c^+f^+, $c^+ > \mu$) is exactly canceled by the decrease of the rate due to the deleterious mutations (c^-f^-, $c^- < \mu$):

$$c = \mu f'_0 \text{ (8a) and } |c^+f^+ - \mu f'_0| = |\mu f'_0 - c^-f^-| \neq 0 \text{ (8b)}$$

Thus, f'_0 ($f'_0 \leq 1$) and thus c is time-independent rate constant.

However, when the balance condition is not satisfied,

$$c = \mu f'_0 \text{ (9a) } |c^+f^+ - \mu f'_0| \neq |\mu f'_0 - c^-f^-| \neq 0 \quad \text{(9b)}$$

Equations 9a and 9b define NNUST, predicting a time-dependent f'_0 and c. In addition, the substitution rate c that can be higher than the mutation rate under effective positive selection ($f'_0 = c/\mu > 1$) or under effective negative selection ($f'_0 = c/\mu < 1$).

NNBST can be reduced to KNT, which defines the overall substitution rate as the product of the mutation rate with the fraction of strictly neutral mutations only by assuming the fraction of beneficial mutations and the substitution rate of deleterious mutations are negligible($f^+ \approx 0$ and $c^- \approx 0$) and strictly neutral sites comprises the majority of the near neutral mutations ($f'_0 \approx f_0$) :[6]

$$c = \mu f_0 \ \text{(10a) or} \ c = \mu\,(1 - f^-) \ \text{(10b)}$$

Conversely, NNUBT can be reduced to ST, which define the overall substitution would be the product of the substitution rate of strong positive sites (c^+) with its fraction (f^+) by assuming the fraction of neutral mutations and the substitution rate of deleterious mutations are negligible small ($f'_0 \approx 0$ and $c^- \approx 0$):

$$c = c^+ f^+ \ \text{or} \ c = c^+\,(1 - f^-) \ \text{(11)}$$

In other words, only beneficial mutations can be fixed in the population.

First, the balance condition in NNBST allows a non-zero fraction of positive selection sites than KNT so that molecular clock is not sufficient to prove $f_0 > f^+$. NNUST allows a more non-zero fraction of neutral selection sites than ST so that lack of molecular clock is not sufficient to prove $f^+ > f_0$.

Second, while **equation 8a** of NNBST or equation **10a** of KNT predicts that a molecular clock with variable f'_0 *or* f_0 of different genome segments under negative selection only and the maximum substitution rate is limited to the mutation rate ($c \rightarrow u$ when f'_0 *or* $f_0 \rightarrow 1$) due to lack of the positive selection; **equation 9a** of NNUST and **11** of ST predicts a time-dependent substitution rate due to time-dependent selections: ($c^+f^+ + c^-f^-$) in NNUST and (c^+f^+) in ST, leading to either an effective negative selection with a upper rate limit of u or an effective positive selection without a upper rate limit.

Position-based and time-based approaches for calculating substitution rates (c). Given the reference genome and a set of genomes for a species, the number of substitutions at each NT site can be counted for a given time interval after multiple sequence alignment using MAFFT.[21] To calculate a substitution rate (substitutions per NT per unit time, c) from a reference sequence of length N and an aligned sequence set of size G, a two-dimension (Time x Position) substitution count matrix (m_{ij}) at various collection dates and at different genome positions was obtained, but there are two ways for averaging substitution frequency across position and time: time-based and position-based approaches (**eq. 12**). While the time-based approach (eq. 12 left) begins by averaging substitution counts across positions to get temporal (timeline) trends, and then across time ($\Delta t = t_{max} - t_{min}$), the position-based approach (eq. 10 right) averages substitution counts across time to get SRS, and then across positions.

$$\frac{\sum_i \sum_j m_{ij}}{\Delta t * G * N} = \frac{\sum_j \sum_i m_{ij}}{N * \Delta t * G} \qquad \left(\frac{substitutions}{NT\ site\ *\ unit\ time}\right) \qquad (12)$$

Both methods offer distinct perspectives on substitution dynamics: while a timeline from the time-based approach emphasizes temporal trends (**Fig1 E)**, SRSRS form the position-approach highlights spatial patterns (**Fig1 C**). Both approaches should give identical results of an absolute rate, but we modified the time-based approach to allow a least square fitting (y=a*x, R^2) to timeline for getting the final rate. This rate should not be identical to the rates calculated from the position-based approach. This fitting implicitly assumes molecular clock of a constant rate (**Equations 1-5)**. If this assumption is true ($m_i = rate * t_i$), then the fitting could improve the accuracy of the rates by removing random "noise" in the sampling; and the obtained absolute true rate value differs from that from the position-based approach by a constant ($\frac{\sum_i t_i}{G * \Delta t}$). The scaling constant will be canceled when calculating the relative rates (c/μ), therefore both ways should give very similar values if the molecular clock assumption is true. The coefficient of determination (R^2) from the fitting can offer a statistical measure on how good the molecular clock assumption holds. If R^2 is too small ($R^2 \leq 0.6000$), the rates from the position-based approach should be used. Nonetheless, SRSRS is useful to identify adaptive mutations. SRSRS can be converted into PDRSR for each gene/UTR and the whole genome for characterization of different evolution theories (**Table 2**).

The proportion of the strictly neutral, nearly neutral, beneficial, deleterious mutations can be determined from PDRSR. The fraction of strictly neutral mutations ($\boldsymbol{f_0 = P(\frac{c}{u} = 1)}$) can be precisely obtained from PDRSR. The fraction of near neutral mutations was determined through equation 8a/9a ($\boldsymbol{f'_0 = \frac{c}{u}}$) for the segments with molecular clock and for the segments without molecular clock under an effective negative selection (c/μ<1), leading to determination of its weak negative boundary (WNB), its weak positive boundary (WPB=1/WNB); the fraction of mutations under strong negative (SN), weak negative (WN), weak positive (WP) and strong positive (SP) selection; and ratio of mutations under weak negative to positive selections (WN:P) and ratio of near neutral to strictly neutral mutations ($\boldsymbol{f'_0 : f_0}$). In our previous paper [13], WNB is arbitrarily set to ½ and so is WPB (2=1/WNB). Given **equation 8a/9a** ($c = \mu f'_0$), WNB can be determined from the fraction of nearly neutral mutations (f'_0) by finding $\frac{c}{u}$ such that its cumulative probability is equal to f'_0 ($P(x) = \sum_{\frac{c}{u}} P(\frac{c}{u} \geq x) = f'_0$) in this study. WPB is arbitrarily set to 1/WNB. Given the cumulative probability at conserved sites ($\boldsymbol{\frac{c}{u} = 0}$), WNB, strictly neutral ($\boldsymbol{\frac{c}{u} = 1}$)**,** and WPB**,** the fraction of mutations under strong negative **(**SN, $0 \leq \frac{c}{u} < WNB$)**,** weak negative (WN, $WNB \leq \frac{c}{u} < 1$), weak positive **(**WP, $1 \leq \frac{c}{u} < WPB$)**,** strong positive selection **(**SP, WPB $\leq \frac{c}{u} \ll \infty$) can be obtained from the cumulative probability function. While the fraction of near neutral (f'_0) mutation is equal to the sum of mutation under WN and WP selection; the fraction of positive (P) mutation is the sum of the mutations under WP and SP selection. The ratio of the mutations under weak negative to positive selection (WN:P) and the ratio of near neutral to strictly neutral mutations (f'_0:f_0) were calculated to validate the theories. f'_0, R^2, WNB, N, WPB, SN, WN, f_0, WP, SP, WN:P and f'_0:f_0 for the 24/25 segments with/without molecular clock are listed in **Table 3 and 4,** respectively.

Acknowledgements

C.W acknowledges the support by the New Jersey Health Foundation (PC 76-24) and the National Science Foundation under Grants NSF ACI-1429467/RUI-1904797, and ACESS/BIO230145. The Anton2 machine at the Pittsburgh Supercomputing Center (PSCA170090P) was generously made available by D. E. Shaw Research.

Author Contributions: Conceptualization and Equation Derivation, C.W.; MatLab scripts, C.W. and N.J.P.; Data preprocessing, N.J.P.; Plots and Tables, N.J.P. and C.W.; writing—original draft preparation, review and editing, C.W. and N.J.P.; funding acquisition, C.W.. All authors have read and agreed to the published version of the manuscript.

Reference

1. Clarke B. DARWINIAN EVOLUTION OF PROTEINS. Science. 1970;168(3934):1009-&. doi: 10.1126/science.168.3934.1009. PubMed PMID: WOS:A1970G307100038.

2. Bromham L, Penny D. The modern molecular clock. Nat Rev Genet. 2003;4(3):216-24. doi: 10.1038/nrg1020. PubMed PMID: WOS:000181511900014.

3. Gojobori T, Moriyama EN, Kimura M. Molecular clock of viral evolution, and the neutral theory. Proc Natl Acad Sci U S A. 1990;87(24):10015-8. doi: 10.1073/pnas.87.24.10015. PubMed PMID: WOS:A1990EN15900105; PubMed Central PMCID: PMCPMC55305.

4. Kimura M. Evolutionary rate at the molecular level. Nature. 1968;217:624-6. doi: 10.1038/217624a0

5. Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press; 1983.

6. Kimura M. THE NEUTRAL THEORY OF MOLECULAR EVOLUTION - A REVIEW OF RECENT-EVIDENCE. Japanese Journal of Genetics. 1991;66(4):367-86. doi: 10.1266/jjg.66.367. PubMed PMID: WOS:A1991GG94400001.

7. Ohta T. Near-neutrality in evolution of genes and gene regulation. Proc Natl Acad Sci U S A. 2002;99(25):16134-7. doi: 10.1073/pnas.252626899. PubMed PMID: WOS:000179783400057; PubMed Central PMCID: PMCPMC138577.

8. Ohta T. Slightly deleterious mutant substitutions in evolution. Nature. 1973;246(5428):96-8. doi: 10.1038/246096a0. PubMed PMID: WOS:A1973R229300031.

9. Moutinho AF, Bataillon T, Dutheil JY. Variation of the adaptive substitution rate between species and within genomes. Evolutionary Ecology. 2019;34(3):315-38. doi: 10.1007/s10682-019-10026-z.

10. Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D, et al. The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. Evolution. 2019;73(1):111-4. Epub 2018/11/22. doi: 10.1111/evo.13650. PubMed PMID: 30460993; PubMed Central PMCID: PMCPMC6496948.

11. de Jong MJ, van Oosterhout C, Hoelzel AR, Janke A. Moderating the neutralist-selectionist debate: exactly which propositions are we debating, and which arguments are valid? Biol Rev Camb Philos Soc. 2024;99(1):23-55. Epub 2023/08/25. doi: 10.1111/brv.13010. PubMed PMID: 37621151.

12. Kern AD, Hahn MW. The Neutral Theory in Light of Natural Selection. (1537-1719 (Electronic)).

13. Wu C, Paradis NJ, Lakernick PM, Hryb M. L-shaped distribution of the relative substitution rate (c/μ) observed for SARS-COV-2's genome, inconsistent with the selectionist theory, the neutral theory and the nearly neutral theory but a near-neutral balanced selection theory: Implication on “neutralist-selectionist” debate. Computers in Biology and Medicine. 2023;153:106522. doi: https://doi.org/10.1016/j.compbiomed.2022.106522.

14. Teufel AI, Ritchie AM, Wilke CO, Liberles DA. Using the Mutation-Selection Framework to Characterize Selection on Protein Sequences. Genes (Basel). 2018;9(8). Epub 2018/08/15. doi: 10.3390/genes9080409. PubMed PMID: 30104502; PubMed Central PMCID: PMCPMC6115872.

15. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121-3. doi: 10.1093/bioinformatics/bty407. PubMed PMID: WOS:000453453300023.

16. Ohta T, Gillespie JH. Development of neutral and nearly neutral theories. Theoretical Population Biology. 1996;49(2):128-42. doi: 10.1006/tpbi.1996.0007. PubMed PMID: WOS:A1996UG45900003.

17. Gillespie JH. ON OHTA HYPOTHESIS - MOST AMINO-ACID SUBSTITUTIONS ARE DELETERIOUS. J Mol Evol. 1995;40(1):64-9. doi: 10.1007/bf00166596. PubMed PMID: WOS:A1995PZ28900008.

18. Jenkins GM, Rambaut A, Pybus OG, Holmes EC. Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. J Mol Evol. 2002;54(2):156-65. doi: 10.1007/s00239-001-0064-3. PubMed PMID: WOS:000173522200003.

19. Kimura M, Crow JF. The number of alleles that can be maintained in a finite population. Genetics. 1964 49(4):725-38. doi: 10.1093/genetics/49.4.725.; PubMed Central PMCID: PMC1210609.

20. Duret L. Neutral theory: The null hypothesis of molecular evolution. Nature Education 2008;1(1):218.

21. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 2019;20(4):1160-6. doi: 10.1093/bib/bbx108. PubMed PMID: WOS:000493041400010; PubMed Central PMCID: PMCPMC6781576

Supporting information captions

Table S1. c total substitution rate slope and c R^2 values for each dataset and averaged over each dataset for all SARS-COV-2 UTR and TRS segments.

Table S2. Time-based c total genomic substitution rate slope and c R^2 values for each dataset and averaged over each dataset for all SARS-COV-2 coding proteins.

Table S3. Position-based c/μ (c/μ^a), time-based c/μ (c/μ^b) and their absolute difference values for the All-UTR, All-TRS and each UTR and TRS over 19 months, combined from Datasets A1A-A1C.

Table S4. Time-based c/μ (c/μ^a), Position-based c/μ (c/μ^b), their absolute and percent differences for the genome, All-UTR, All-TR and each coding segment of SARS-COV-2 over 19 months, combined from Datasets A1A-A1C.

Table S5. Abundance of sites under different selection types for segments exhibiting molecular clock feature. (Column 1): Relative abundance of sites under WN, WP and SP selection. (Column 2): Abundance of sites under NN and P selection. (Column 3): Abundance of sites under SN, NN and SP selection. (Column 4): Abundance of sites under WN and SP selection. (Column 5): Abundance of sites under N and P selection. See **Figures 7 and S6** for graphical representations.

Table S6. Abundance of sites under different selection types for segments exhibiting non-molecular clock feature. (Column 1): Relative abundance of sites under WN, WP and SP selection. (Column 2): Abundance of sites under NN and P selection. (Column 3): Abundance of sites under SN, NN and SP selection. (Column 4): Abundance of sites under WN and SP

selection. (Column 5): Abundance of sites under N and P selection. See **Figures 7 and S6** for graphical representations.

Table S7. Tabulated parameters for the WN:P to c/μ relationship in **Figure 8.**

Table S8. Tabulated parameters for the fo':fo to c/μ scaling factor relationship in **Figure 8.**

Figure S1. Percent genomic variation of sequences in set A1a with the references sequences Wuhan-Hu-1 (left) and Wuhan IPBCAMS-WH-01 2019 (right) set.

Figure S2. The percent total NT substitution rate for each TR, UTR and TRS segment over evolution time and averaged over the three combined datasets**.** See **Table S1-S2** for tabulated regression parameters.

Figure S3. The total percent NT variation for the genome, All-TR, All-UTR, All-TRS and each coding and non-coding gene segment over evolution time for each individual dataset. See **Table S1-S2** for tabulated regression parameters.

Figure S4. c/μ Empirical probability distribution function of 24 segments exhibiting molecular clock feature.

Figure S5. c/μ Empirical probability distribution function of 25 segments exhibiting non molecular clock feature.

Figure S6. c/μ Cumulative probability distribution function ($P(x) = \sum_{u}^{c} P(\frac{c}{u} \geq x)$) of 24 segments exhibiting molecular clock feature, showing the abundance of sites under different selection types. The boundaries for weak negative (red line) to neutral selection (green line) to weak positive selection (orange line) and their determined c/μ positions are noted. Strong negative selection and strong positive selection would be to the left and right of the red and orange lines, respectively. A broken lined leading on the x-axis represents the distance between c/µ = 3.0 to the weak positive selection boundary. See **Table 3** for tabulated c/µ boundaries and percent abundances for each selection type.

Figure S7. c/µ Cumulative probability distribution function ($P(x) = \sum_{u}^{c} P(\frac{c}{u} \geq x)$) of 25 segments not exhibiting molecular clock feature, showing the abundance of sites under different selection types. The boundaries for weak negative (red line) to neutral selection (green line) to weak positive selection (orange line) and their determined c/µ positions are noted. Strong negative selection and strong positive selection would be to the left and right of the red and orange lines, respectively. A broken lined leading on the x-axis represents the distance between c/µ = 3.0 to the weak positive selection boundary. See **Table 3** for tabulated c/µ boundaries and percent abundances for each selection type.

Figure S8. Percent selection types of each segment exhibiting molecular clock features decomposed into three categories for strong negative, near-neutral and strong positive selection (top row), weak negative and strong positive selection (middle row) and negative and positive selection (bottom row) using a c/µ scaling of 5.46. See **Table S5** for percent selection type values.

Figure S9. Percent selection types of each segment exhibiting non-molecular clock features decomposed into three categories for strong negative, near-neutral and strong positive selection (top row), weak negative and strong positive selection (middle row) and negative and positive selection (bottom row) using a c/µ scaling of 5.46. See **Table S6** for percent selection type values.