Evolutionary rate and selection in SARS-CoV-2 during persistent infections
Скорость эволюции и отбор у SARS-CoV-2 при персистирующих инфекциях

School of molecular and theoretical biology

Sabancı Üniversitesi

Abstract

Persistent SARS-CoV-2 infections are when a viral lineage replicates within a single person for a long time (>25 days). The virus evolves in the process, and this evolution was suggested to be important in how new major SARS-CoV-2 variants originate. Here, we study this viral evolution using a published dataset of 547 persistent infections [1]. We estimate the rate of evolution at 5.6×10^{-4} substitutions per site per year, and show that this estimate is strongly dependent on the method of calculation. We also find that sequencing coverage tends to get lower as infection progresses (probably due to lower viral loads in recovering patients), which can confound evolutionary inference. Finally, we study individual variant trajectories across patients, and show that some of the nonsynonymous substitutions are favored by positive selection. The strength of this selection is correlated with that in between-host evolution, but the correlation is weak, suggesting that evolutionary forces differ. Together, our results point to within-patient selection as an important contributor to long-term evolution of SARS-CoV-2.

Персистентная инфекция SARS-CoV-2 - это такая инфекция, при которой одна линия вируса долго (>25 дней) реплицируется в одном человеке. При этом вирус эволюционирует, и этот процесс, как считается, важен для возникновения новых важных вариантов SARS-CoV-2. Мы исследовали эту эволюцию в опубликованном наборе данных по 547 персистентным инфекциям [1]. Мы оцениваем скорость эволюции вируса как 5.6×10^{-4} замен на сайт в год и показываем, что она сильно зависит от метода подсчёта. Мы также показываем, что покрытие сиквенсами снижается по ходу инфекции (вероятно, из-за снижения вирусной нагрузки в выздоравливающих пациентах), что может сбивать эволюционные оценки. Наконец, мы описываем индивидуальные эволюционные траектории вариантов при вирусной эволюции в пациентах и показываем, что некоторым несинонимичным заменам способствует положительный отбор. Сила отбора на замену коррелирует с полезностью замены в глобальной эволюции SARS-CoV-2, но эта корреляция слабая, что свидетельствует о различии действующих эволюционных сил. Наши результаты показывают, что отбор в персистентно инфицированных пациентах - важный фактор долгосрочной эволюции SARS-CoV-2.

Methods

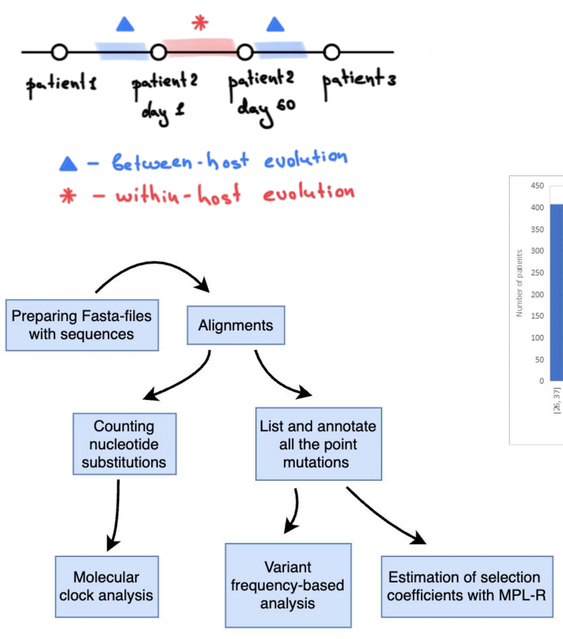


Figure 1. Analysis idea and workflow. MPL-R, see ref. [2]

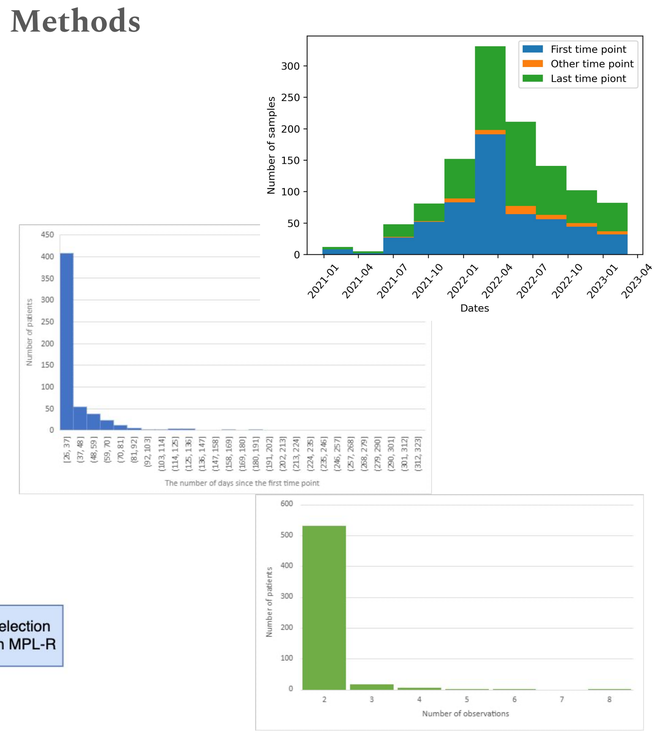


Figure 2. Number of samples per collection date (top), days between the first and the last time point (middle) and observations per patient (bottom)

Molecular clock analysis

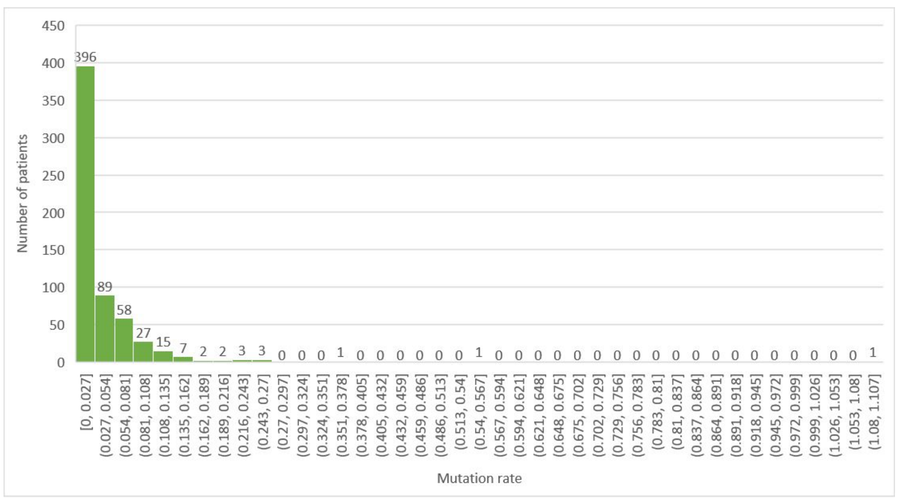
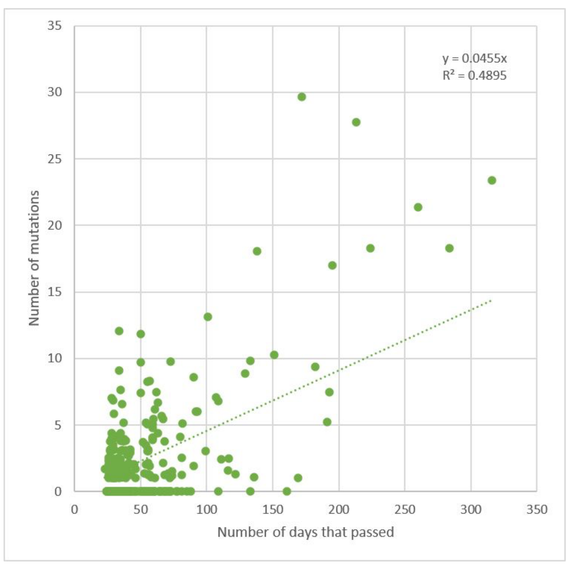


Figure 3. Numbers of hosts with each viral mutation rate



The rate of mutation we calculated based on consensus sequences is lower than the one presented in [1], which was calculated using variant frequencies. Between-host data from [1], which was also calculated using the consensus sequences, is closer to our mutation rate. This may be due to the amount of data that gets ignored when only consensus sequences are used; variant frequencies also allow the detection of rare polymorphisms, and not considering this data makes the mutation rate seem slower.

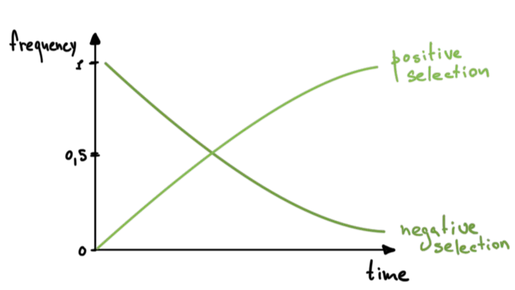
Figure 4. Number of mutations depending on sampling date (relative to the first sampling date)

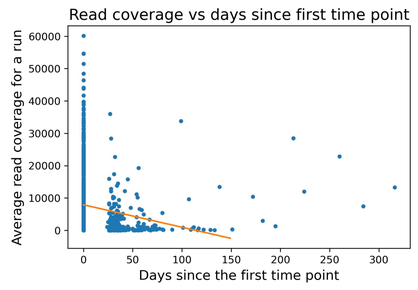
Number of mutations	Avg. of all μ^* from all patients	Avg. μ^* (total # of mutations / time)	Plot-based analysis (regression slope)	Within-host μ^* [1]	Between-host μ^* [1]
per day	0.03	0.03	0.05	0.07	0.02-0.05
per year	9.52	12.02	16.62	23.62	7.48-17.94
per year, per site	3.18×10^{-4}	4.02×10^{-4}	5.56×10^{-4}	7.9×10^{-4}	$(2.5\text{-}6.0) \times 10^{-4}$

Table 1. Viral mutation rates estimated with different methods

*Mutation rate

Variant frequency-based analysis





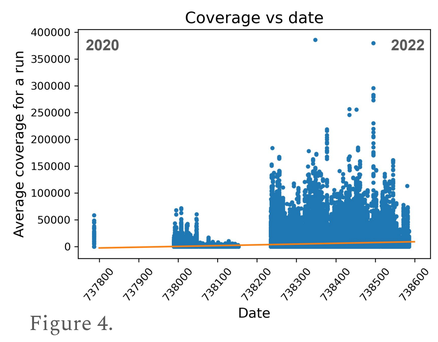


Figure 4.

➤ Read coverage decreases during infection

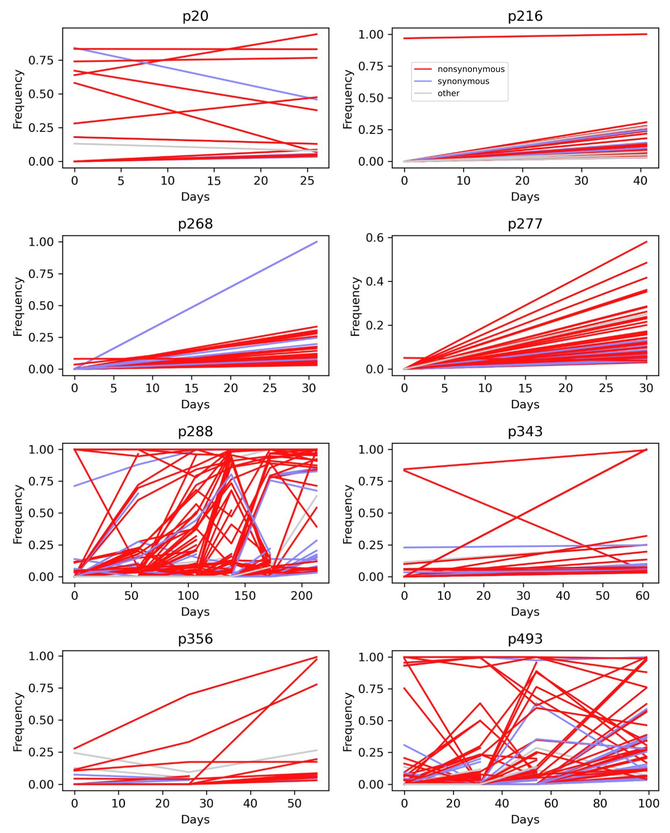


Figure 5. Variant frequency vs time. Each panel is a patient; each line is a mutation. Red, nonsynonymous; blue, synonymous; grey, other.

➤ Variant frequencies change a lot within a host
➤ Mutations are linked, complicating selection inference

Positive selection in spike protein
(17 patients with > 2 time-points)

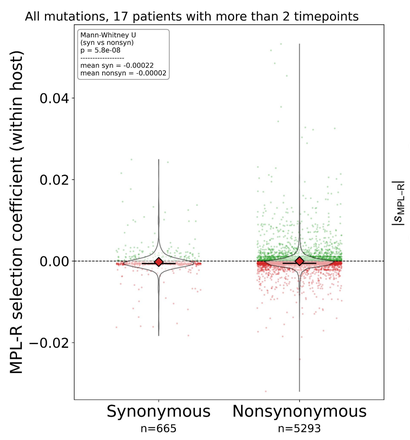


Figure 6. Distribution of MPL-R values for synonymous vs nonsynonymous variants

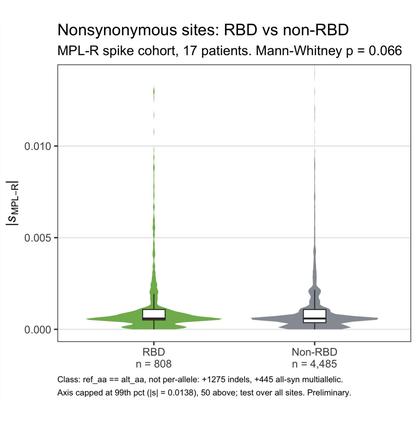


Figure 7. Distribution of MPL-R values for nonsynonymous variants in RBD vs non-RBD regions

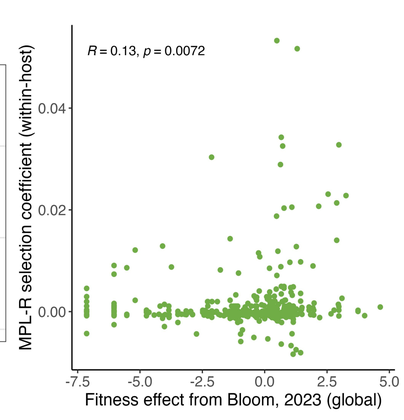


Figure 8. Within-host MPL-R selection coefficients vs fitness effects for the global population [3]

➤ Within-host MPL-R [2] selection coefficients are positively correlated with fitness effects estimated for the global population.
➤ Nonsynonymous variants are enriched in beneficial mutations (MPL-R > 0).
➤ On average, MPL-R coefficients for nonsynonymous variants are significantly higher than those for synonymous variants.
➤ Nonsynonymous variants from RBD do not appear to be enriched in beneficial mutations compared to non-RBD variants.

Fraction of mutations with MPL-R [2] selection coefficient > 0
$\chi^2 = 32.67$, f = 1, p = 1.09e-08

Mutation type	MPL-R > 0	MPL-R <= 0	Fraction > 0
Synonymous	157	508	**0.236**
Nonsynonymous	1837	3456	**0.347**

Table 2. Numbers of variants with positive and negative MPL-R estimates for different types of mutations.

Conclusions

➤ Within-patient mutation rate depends on the method of calculation. To compare within- and between-host evolution, substitution rates should be calculated in the same way. When both are calculated from consensus sequences, the rates are actually close: 5.6×10^{-4} within hosts vs. $(2.5\text{-}6.0) \times 10^{-4}$ between hosts [1].

➤ Sequencing coverage decreases over the course of infection, which can confound evolutionary estimates.

➤ Nonsynonymous mutations tend to increase in frequency more often than synonymous mutations, suggesting positive selection within hosts.

Future directions

➤ Use variant frequencies to calculate the mutation rate; only compare the mutation rates that were calculated in the same way.

➤ Extend the selection analysis to more patients and whole genome (beyond spike).

➤ Which mutations are positively selected? Are they the same as in inter-host transmission?

➤ Are these mutations epistatically linked?

References

[1] Ghafari, M. *et al.* SARS-CoV-2 genomic diversity and within-host evolution in individuals with persistent infection in the UK: an observational, longitudinal, population-based surveillance study. *The Lancet Microbe* 6, 101154 (2025).

[2] Abdullah, S. M. U. *et al.* Linkage-aware inference of fitness from short-read time-series genomic data. *Virus Evolution* 12, veag027 (2026).

[3] Bloom, J. D. & Neher, R. A. Fitness effects of mutations to SARS-CoV-2 proteins. *Virus Evolution* 9, vead055 (2024).