



RESEARCH ARTICLE

# Phylogenetic signatures reveal multilevel selection and fitness costs in SARS-CoV-2 [version 1; peer review: 3 approved with reservations]

Vinicius Bonetti Franceschi , Erik Volz 

Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, England, W2 1PG, UK

**V1** First published: 19 Feb 2024, 9:85  
<https://doi.org/10.12688/wellcomeopenres.20704.1>  
Latest published: 24 Jul 2024, 9:85  
<https://doi.org/10.12688/wellcomeopenres.20704.2>

## Abstract

### Background

Large-scale sequencing of SARS-CoV-2 has enabled the study of viral evolution during the COVID-19 pandemic. Some viral mutations may be advantageous to viral replication within hosts but detrimental to transmission, thus carrying a transient fitness advantage. By affecting the number of descendants, persistence times and growth rates of associated clades, these mutations generate localised imbalance in phylogenies. Quantifying these features in closely-related clades with and without recurring mutations can elucidate the tradeoffs between within-host replication and between-host transmission.

### Methods

We implemented a novel phylogenetic clustering algorithm (mlscluster, <https://github.com/mrc-ide/mlscluster>) to systematically explore time-scaled phylogenies for mutations under transient/multilevel selection. We applied this method for a SARS-CoV-2 time-calibrated phylogeny with >1.2 million sequences from England, and characterised these recurrent mutations that may influence transmission fitness across PANGO-lineages and genomic regions using Poisson regressions and summary statistics.

### Results

We found no major differences across two epidemic stages (before and after Omicron), PANGO-lineages, and genomic regions. However, spike, nucleocapsid, and ORF3a were proportionally more enriched for

## Open Peer Review

**Approval Status** ? ? ?

1 2 3

### version 2

(revision)  
24 Jul 2024

### version 1

19 Feb 2024

?


[view](#)


?


[view](#)

?

[view](#)

1. **Richard Neher** , University of Basel, Basel, Switzerland

2. **Mahan Ghafari** , University of Oxford, Oxford, UK

3. **Ariane Weber** , Max Planck Institute of Geoanthropology, Jena, Thuringia, Germany

Any reports and responses or comments on the article can be found at the end of the article.

TFP-homoplasies than other proteins. We provide a catalog of SARS-CoV-2 sites under multilevel selection, which can guide experimental investigations within and beyond the spike protein.

## Conclusions

This study highlights the existence of important tradeoffs between within-host replication and between-host transmission shaping the fitness landscape of SARS-CoV-2.

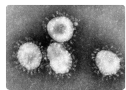
## Plain Language Summary

Viral mutations can potentially carry a transient advantage, being simultaneously favourable for replication within hosts (e.g. by evading host immune responses) and deleterious to transmission (e.g. by having reduced cell binding). To identify such mutations, called transmission fitness polymorphisms (TFPs), we developed a clustering algorithm entitled *mlscluster* that computes lineage-level statistics based on the number of descendants, persistence times, and growth rates of lineages in comparison with co-circulating lineages, which usually are different than expected in the presence of such TFPs. We then applied it to a representative SARS-CoV-2 time-scaled tree with >1 million whole-genome sequences from England.

Our statistical analysis suggested approximately constant levels of transient selection across waves driven by very distinct variants. It also showed that genomic regions of known functional significance such as spike, nucleocapsid, and ORF3a were enriched for TFPs. This is the first study to characterise SARS-CoV-2 recurrent mutations with complex fitness tradeoffs, highlighting the existence of important tradeoffs in selection between intrahost replication and inter-host transmission. It also provides target mutations for laboratory-based investigations of their impacts and mechanisms of interaction with human cells.

## Keywords

Molecular evolution, phylogenetic analysis, transmission fitness, natural selection, mutation, genetic clustering, within-host evolution, SARS-CoV-2



This article is included in the [Coronavirus \(COVID-19\)](#) collection.

**Corresponding author:** Vinicius Bonetti Franceschi ([v.franceschi@imperial.ac.uk](mailto:v.franceschi@imperial.ac.uk))

**Author roles:** **Bonetti Franceschi V:** Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Volz E:** Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Software, Supervision, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was supported by Wellcome (220885, <https://doi.org/10.35802/220885>; an Investigator Award in Science award to EV). VBF and EV also acknowledge funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/X020258/1), funded by the UK Medical Research Council (MRC). This UK-funded award is carried out in the frame of the Global Health EDCTP3 Joint Undertaking. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2024 Bonetti Franceschi V and Volz E. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Bonetti Franceschi V and Volz E. **Phylogenetic signatures reveal multilevel selection and fitness costs in SARS-CoV-2 [version 1; peer review: 3 approved with reservations]** Wellcome Open Research 2024, 9:85 <https://doi.org/10.12688/wellcomeopenres.20704.1>

**First published:** 19 Feb 2024, 9:85 <https://doi.org/10.12688/wellcomeopenres.20704.1>

## Introduction

It is generally held that, for most pathogens, the majority of polymorphic sites within a genome are selectively neutral or under weak selection<sup>1</sup>. However, some mutations may confer a large transient increase in fitness, being advantageous to viral replication within hosts but detrimental to transmission. For example, HIV-1 is subject to multilevel selection, evolving considerably faster within individuals than at the epidemic level<sup>2,3</sup>, and virus which is more highly diverged from the population consensus is less likely to be transmitted<sup>4</sup>.

Regarding SARS-CoV-2, the immense amount of genomic data collected during the COVID-19 pandemic has provided valuable insights about the competing forces influencing viral evolutionary dynamics<sup>5–10</sup>, but these data also presented novel challenges due to the scarcity of methods able to provide scalability using the power of big data streams while retaining fine-scale inferences (e. g. investigating fitness cost of individual mutations)<sup>5,11–13</sup>. Provided scalable analytic pipelines can be developed, data from densely sampled epidemics can enable the identification of recurrent mutations in different branches of the phylogenetic tree, which potentially arise convergently as a consequence of virus response to adaptive selective pressures within hosts.

A particular challenge has been to infer population structure and phenotypic differences (reflected by phylogenetic asymmetries and imbalances) from observed pathogen genealogies<sup>14</sup>. Even when clades are distantly related, they can present very similar distributions of coalescent times and branch lengths<sup>15</sup>, as well as the proportion of descendants, persistence time, and growth rates when compared with closely-related clades. Most importantly, mutations influencing virus transmission fitness are expected to affect the distribution of offspring<sup>5</sup>, consequently generating localised and quantifiable imbalance in time-scaled phylogenies. Therefore, the quantification and comparison of these parameters can indicate if similar evolutionary, demographic, or epidemiological processes are shaping viral evolution across different clades of a genealogy.

In molecular epidemiological studies, a set of particularly scalable approaches have been developed based on the calculation of phylogenetic clusters comprising two or more closely related samples. The frequency of phylogenetic clustering in a sample is sometimes considered a proxy for high transmission rate, especially in HIV datasets<sup>16–18</sup>, and can potentially indicate spread efficiency of a particular genotype (e.g. HIV drug resistance-associated mutations [DRAMs]). Intuitively and by extension, transmissibility and within-host evolution between variants can be considered a proxy for overall fitness<sup>19</sup>. Recently, a genetic clustering analysis of HIV-1 identified variants containing specific DRAMs in antiretroviral therapy (ART)-naïve transmission networks that reduce transmission fitness and suggested a negative correlation between lower frequencies of rare polymorphisms and fitness advantage<sup>18</sup>.

Currently, similar clustering analyses have not yielded major insights into negatively-selected variants in SARS-CoV-2 despite the collection of unprecedented numbers of whole-genome

sequences<sup>5</sup>. Furthermore, there is considerable scope to improve on distance-based genetic clustering methods because such approaches will potentially have poor specificity for variants that negatively influence fitness. During the past few years, positive and negative selection in SARS-CoV-2 have mainly been investigated using methods that rely on synonymous rate variation across sites/branches<sup>20,21</sup>, and results from these approaches on SARS-CoV-2 comprehensive datasets are available for comparison<sup>22</sup>. However, methodology to identify mutations that potentially have a transient fitness advantage is still lacking.

We developed a tree-based clustering algorithm, available as open-source R package `mlscluster` (<https://github.com/mrc-ide/mlscluster>)<sup>23</sup>, to identify potential transmission fitness polymorphisms (TFPs) by computing and comparing simple statistics from the offspring of recurring clade-defining mutations in a time-scaled phylogeny. This approach complements standard procedures based on synonymous rate variation across sites/branches by highlighting variants which likely have different and/or competing selective pressures within and between hosts. We demonstrate its applicability through the analysis of a representative >1.2 million SARS-CoV-2 genomic dataset from England, which indicated slightly higher TFP-homoplasy enrichment on B.1.1.7 and AY.4.\* lineages and across genomic regions of known functional significance such as spike (S), nucleocapsid (N), ORF3a, ORF7, and ORF8. By providing a comprehensive catalog of the main sites driving multilevel selective pressures throughout the SARS-CoV-2 genome, we also expand the understanding of SARS-CoV-2 fitness landscape outside the well-studied spike protein. Therefore, these results can guide experimental studies on the functional impact of specific mutations, especially the subset that is advantageous to within-host replication but detrimental to transmission.

## Methods

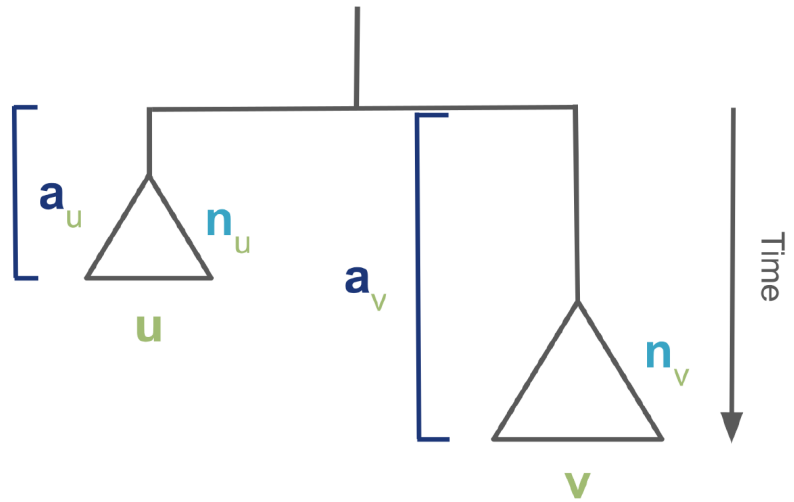
### Terminology and tree-based clustering statistics for detecting localised imbalance

We propose a phylogenetic tree-based clustering algorithm to systematically explore all nodes/clades in a time-scaled phylogenetic tree reconstructed from viral genomes. Assume two clades  $u$  (target clade) and  $v$  (comparator/sister) (Figure 1A) organised in a time-scaled tree  $t$  and sharing ancestry (*i.e.*, the same defining mutations). The size of each node ( $n$  in Figure 1) is defined as the number of descendant sequences arising from it until the leaves of the tree are reached. The persistence time (given by  $a$  in Figure 1, is defined as the difference between the maximum sample time of samples descended from that node and the estimated time of the most recent common ancestor [tMRCA] of the node). After computing these simple parameters for each clade, the target nodes  $u$  are then contrasted against their comparators  $v$ , which can be their sister clade (the clade sharing an immediate ancestor assuming bifurcating phylogenetic relationship) or against all other clades sharing the same immediate ancestor (in case of polytomies/multifurcations).

If considering a node  $u$  with sister clade  $v$ , we compute three statistics based on these local phylogenetic patterns: (i) the

A

Clades  
 Sizes (# descendants)  
 Persistence times



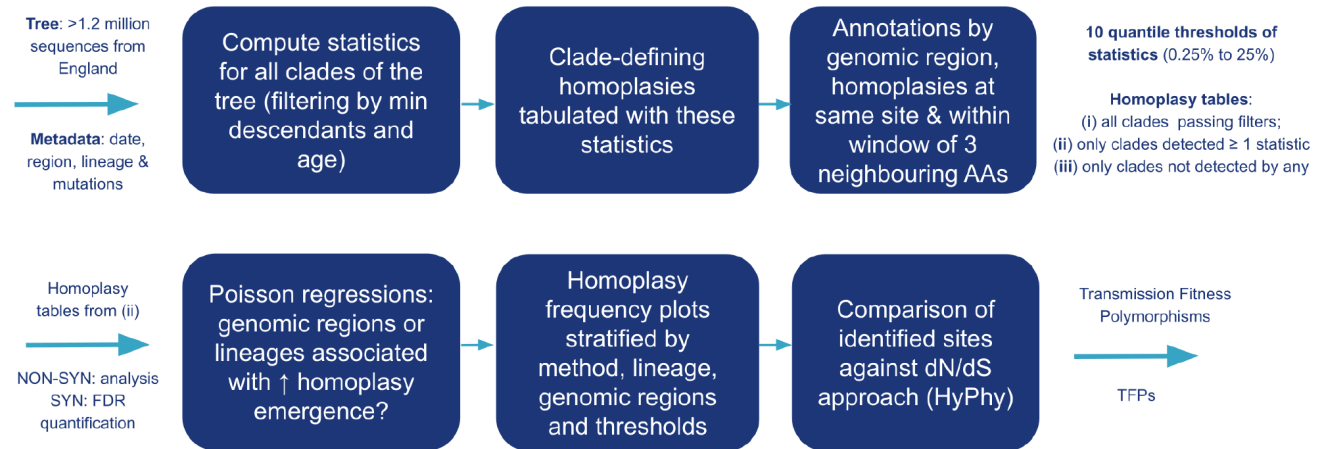
For a clade  $u$  and its sister  $v$ :

$$\text{Ratio of sizes} = \frac{n_u}{n_v}$$

$$\text{Ratio of persistence time} = \frac{a_u}{a_v}$$

$$\text{Logistic growth rate: } \log \left[ \frac{P(\text{clade}=u)}{1-P(\text{clade}=u)} \right] = \alpha + \beta(\text{time})$$

B



**Figure 1. Schematic view of the tree-based clustering algorithm implementation and analytic pipeline.** (A) Main notations, parameters, and respective statistic formulas that are computed by `m1scluster` (<https://github.com/mrc-ide/mlscluster>) for sister clades of the time-scaled phylogeny. (B) Analysis workflow with main steps from input data to TFP inference.

ratio  $S$  of the number of samples descended from  $u$  and  $v$ , denoted  $n_u$  and  $n_v$ , which we will also call the ‘clade size’ (Equation 1), (ii) the ratio  $T$  of persistence times denoted  $a_u$  and  $a_v$  (Equation 2), and (iii) the logistic growth rate. The latter is defined as the coefficient of a logistic regression having a response variable defining sampling a descendent of  $u$  versus  $v$  and the sample time as a predictor (Equation 3). The coefficient

of such a logistic regression quantifies the relative growth of the clade and is related to the selection coefficient<sup>24</sup>.

$$S_{uv} = \frac{n_u}{n_v} \tag{1}$$

$$T_{uv} = \frac{a_u}{a_v} \tag{2}$$

$$\log \left[ \frac{P(\text{clade}=\text{u})}{1 - P(\text{clade}=\text{u})} \right] = \alpha + \beta(\text{time}) \quad (3)$$

### Tree-based clustering algorithm implementation

The `mlscluster` method is implemented as an R package (<https://github.com/mrc-ide/mlscluster>)<sup>23,25</sup> that incorporates these multiple statistical methods for identifying especially convergently acquired mutations (homoplasies) that are detrimental for transmission (within a low quantile [e. g., 2%] of the probability distribution of at least one of the three statistics). These statistics applied to each clade are designed to be simple and computationally fast, making it possible to scan phylogenies with more than a million tips in hours using multiple CPU cores.

The clustering algorithm (Figure 1B) starts by receiving a rooted bi- or multifurcating time-scaled tree (e. g., estimated using `treedater`<sup>26</sup>, `treetime`<sup>27</sup> or `chronumetal`<sup>28</sup>) and associated metadata in a tabular format including sequence name, sample date, lineage, major lineage, and annotated mutations. The package then uses standard tree manipulation strategies implemented in the `ape` R package<sup>29</sup>, particularly postorder traversal to visit nodes and tips based on the two-column edge matrix from the “phylo” class. Given this efficient way to visit nodes of the tree and edge lengths, we can easily extract the parameters of interest (e. g., time of the most recent common ancestor of each node, descendant identifiers and quantities, clade ages, etc). Target nodes are extracted based on the following conditions: (i) minimum number of descendants (ii) maximum number of descendants, (iii) minimum cluster age (in years), (iv) minimum sampling date, and (v) maximum sampling date. Only nodes passing all those criteria are kept for analysis.

Subsequently, target nodes and comparator (sister) clade(s) are gathered together and ratio of sizes, ratio of persistence time and logistic growth rates are calculated as previously stated. Since every sequence should include a metadata column (e.g. precomputed by COG-UK consortium, see [Methods: Tree and metadata](#)) listing mutations from its genome, the clustering algorithm tool incorporates a function to identify defining polymorphisms in target nodes. The mutation must be present in >75% of sequences in that node (while absent or in a smaller fraction than this percentage in its comparator) to be considered as defining, although this cutoff value can be changed. After computing clade-defining mutations, these are all tabulated and those which happen more than once in different nodes are retrieved as homoplasies. To enhance inspection of results, homoplasies are annotated into (i) regions of interest (e.g. SARS-CoV-2 spike and nucleocapsid proteins), (ii) different mutations at the same site, and (iii) mutations within a 3 amino acid sliding window. There is also an additional sanity check for known sequencing artifacts<sup>30</sup> and for positively selected sites found by other analyses<sup>22</sup>. Then, based on cut points dividing the range of the probability distribution of each statistic into continuous intervals with equal probabilities (quantiles), cluster thresholds can be specified to retain

only clades potentially detrimental for transmission (default threshold of <1%) or carrying a positive fitness advantage (e.g., >99%). We intended to make the method flexible by creating a parameter that specifies in how many percentiles the statistic should be splitted (*default* = 1/100) and another to keep values below or above the cutoff point.

Different comma-separated detailed outputs are generated for each of the three statistics showing nodes (and defining homoplasies) contained in the chosen cluster threshold, as well as the intersection (nodes identified by the threshold of the three statistics) and union (clades associated with at least one statistic threshold). Additionally, for each of the three homoplasy-annotated categories, three files are generated with their frequencies, and clades of occurrence considering (a) all target nodes that passed minimal filtering conditions, (b) only clades detected by one or more statistic threshold, (c) nodes not detected by any cutoff. Finally, these outputs are joined into one *data.frame* to facilitate further statistical analyses.

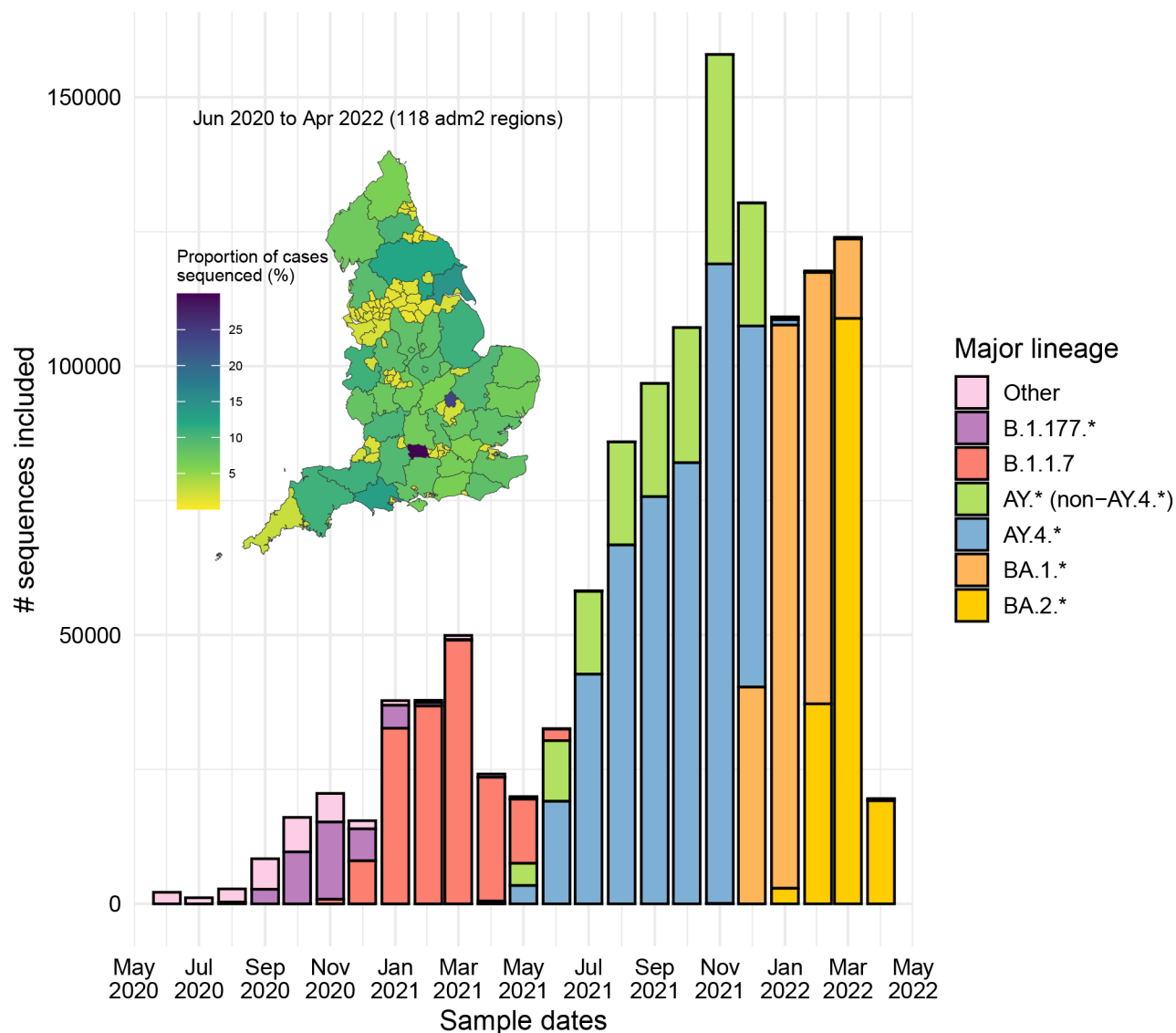
### Identifying potential TFPs using the `mlscluster` algorithm and a representative SARS-CoV-2 genomic dataset

**Tree and metadata.** A global SARS-CoV-2 maximum likelihood (ML) phylogenetic tree and associated metadata including adm2 regions following the Database of Global Administrative Areas (GADM) subdivisions, PANGO-lineages, and annotated synonymous and non-synonymous mutations were obtained from the COVID-19 Genomics UK Consortium (COG-UK). From the ML tree, we estimated a time-scaled phylogeny using `chronumetal` v0.0.53<sup>28</sup>. Only sequences from England were retained alongside the Wuhan/WH04/2020 (EPI\_ISL\_406801) reference sequence. In total, we included 1,275,669 sequences from all the 118 adm2 regions in England from June 01, 2020 to April 30, 2022, since they are associated with Pillar 2 representative community sampling efforts in the UK.

The proportion of cases sequenced for each region in England was computed by using UTLA and LTLA-level case counts obtained from the UK government website (<https://coronavirus.data.gov.uk/>, accessed on 27 April 2023) matched against GADM adm2 geographical regions contained in COG-UK metadata. Since adm2, LTLA and UTLA regions are not entirely compatible, we have not considered on sequence counts samples with ambiguous matches (33%) for the map representation (Figure 2).

### Statistical analysis for identifying genomic regions enriched for TFPs

We tested our approach using two COVID-19 pandemic time-periods: (i) from June 01, 2020 (including Wuhan/WH04/2020 reference sequence as root of the phylogeny) to November 15, 2022 (before Omicron BA.1.\* variant emergence) (ii) from June 01, 2020 to April 30, 2022 (considering Omicron BA.1.\* and up to Pillar 2 termination). For each period, 10 different thresholds (0.25, 0.5, 0.75, 1, 2, 3, 4, 5, 10, and 25%) of the clustering statistics are computed.



**Figure 2. Spatiotemporal distribution of the SARS-CoV-2 sequences from England included in this study during the investigated period (June 2020 to April 2022).** Main plot: Monthly-stratified frequency of the sequences stacked by major PANGO-lineage. Inset plot: Proportion of included sequenced cases across adm2 regions in England during the investigated period for 77% of the samples with unambiguous adm2-level assignments.

We also performed rigorous quality control to ensure our estimates were not biased by sequencing and base-calling artifacts. Firstly, we removed outlier sites highly enriched for homoplasies (above the 99% quantile of homoplasy frequencies for all thresholds), which we manually confirmed to be sequencing artifacts due to the high number of undetermined bases at respective sites in the alignment generating the phylogenetic tree. However, even after performing this approach, BA.1-defining mutations in the Receptor-binding Domain (RBD) were particularly identified as TFP-homoplasies for threshold=2% (Extended Data Figure S1A)<sup>31</sup>, which was an unexpected result. To further inspect this inconsistency, we selected eight BA.1-spike defining mutations that were in our top100 of most frequent homoplasies (S:S371L, S373P, S375F, G496S,

Q498R, N501Y, Y505H, N764K) and excluded all sequences (n=71,414, 5.6% of all sequences) that had undetermined bases (e.g. “NNN”) in their respective codons from the nucleotide alignment. As a result, these sites were not detectable anymore (Extended Data Figure S1B)<sup>31</sup> and we could confirm they were the result of sequencing artifacts, which generally occur due to systematic differences in sequencing protocols and primer selection over time<sup>32</sup> and in different laboratories.

Consequently, we decided to perform a more aggressive quality control (henceforth called alignment-aware artifact removal) that has the advantage of not relying on excluding sequences without perfect coverage. First, we ran the `mlscluster` algorithm (<https://github.com/mrc-ide/mlscluster>) without any

artifact removal. We then extracted every homoplastic site detected and used `seqtk v1.3-r106`<sup>33</sup> to create alignment files for each codon matching these sites in case of non-synonymous mutations and for each nucleotide in case of synonymous mutations. Afterwards, for each sequence, we added “X” (undetermined amino acid) for every site which had one or more non-ACTG character in respective codon positions, and “N” (undetermined nucleotide) for every non-ACTG synonymous site. Subsequently, we appended the “X” and “N” site annotations for all sequences within the existing metadata mutations column. We then incorporated a function named `.fix_sites` (<https://github.com/mrc-ide/mlscluster/blob/main/R/mlsclust.R#L314>) to deal with those highly uncertain sites within the `mlscluster` package. In summary, since the first step to extract a defining mutation for each target clade is to compute the frequency of each mutation within the target and the comparator clades, we used the proportion of the most frequent mutation at a given site and added up the frequency of the “X” or “N”, because it is most likely that the artifacts follow the majority. For example, if the target clade has the site S:G446 changing to S with frequency = 0.7 and to X with frequency = 0.3, we consider the S frequency = 1, and this mutation now has enough frequency (>75%) to be considered defining, which only would occur if the comparator has S:G446S at frequency <0.75%. In cases “X” and “N” are the most frequently mutated characters, the second-ranked amino acid or nucleotide at that site is added to these undetermined characters. For example, the target clade has the site S:N501 changing to X with frequency = 0.6 and to Y with frequency = 0.4, then the X frequency = 1. In such cases where either the target or comparator has a higher frequency of “X” or “N”, the sites are not considered defining. A mutation is only considered defining when (i) it has one of the 20 valid amino acids (or stop codon) or the four valid nucleotides, (ii) it has a >75% frequency on the target clade and simultaneously <75% on the comparator node. This approach removed not only the eight previously investigated BA-1-spike defining mutations but also other six S sites for threshold = 2%, and affected mostly the BA.1 lineage without major changes for other lineages. Therefore, we consider that results arising from this alignment-aware artifact removal method are more reliable than previously and report those throughout the paper.

We performed Poisson regressions using the `glm` function from the `stats` package<sup>34</sup> and having the frequency of homoplasies as response variable (Equation 4) to identify if any genomic regions and/or major PANGO-lineages were associated with increased TFP-homoplasy emergence for non-synonymous polymorphisms across different time periods and thresholds. A p-value  $\leq 0.05$  was considered statistically significant.

$$\begin{aligned} \text{Freq\_homopl}_i &\sim \text{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \alpha_{j[i]} \\ &+ \beta_{1,j[i]}(\text{major\_lineage}) \\ &+ \beta_{2,j[i]}(\text{genomic\_region}) \\ &+ \beta_3(\text{indep\_positive\_selection}) \end{aligned} \quad (4)$$

We assigned as major PANGO-lineages the following variants: B.1.177, Alpha (B.1.1.7), Delta AY.4 and sublineages (AY.4.\*), other Delta (AY.\* [non-AY.4.\*]), Omicron BA.1.\*, Omicron BA.2.\*, and Others (all other lineages excluding recombinants). These were main drivers of epidemic waves in the UK and around the globe. Genomic regions included all 15 non-structural proteins (NSPs) from ORF1ab (NSP1-10, 12-16), S, ORF3a, Envelope (E), Membrane (M), ORF6, ORF7a, ORF7b, ORF8, N, and ORF10. Moreover, regions of characterised functional significance including the N-terminal Domain (NTD), the RBD and the Furin-cleavage site (FCS) of S, as well as the Linker Domain of N<sup>35</sup> were considered as additional genomic regions. The other covariate was whether the sites was independently found under selection based on a HyPhy-based synonymous rate variation across sites/branches analysis<sup>22</sup>.

To further investigate the genomic regions enriched for TFP-homoplasies resulting from the Poisson regressions and to compare the sites identified as potentially under selection against results from the literature<sup>22</sup>, we generated different exploratory visualisations using `ggplot2`<sup>36</sup>. These were stratified by the method of detection (`mlscluster`, HyPhy, or both), major PANGO-lineages, genomic regions (including frequency normalisation by size), and cluster thresholds.

#### Codon-aware false discovery rate (FDR)

We used synonymous homoplasies for characterising the FDR of our approach under the assumption that synonymous sites would not provide a fitness advantage. Since these sites represent one-third of the genome and mutations tend to occur in the third codon position to preserve the encoded amino acid, this weighting needs to be taken into account when computing FDR. Firstly, we defined the percentage of erroneous TFP calls for each threshold  $t$  as:

$$\text{FDR}_t = \frac{Y}{i} \times 100 \quad (5)$$

where  $Y$  is the number of TFP calls specifically among the  $i$  polymorphic third codon position sites with > 100 mutated sequences at the given site (considering the analysed > 1.2 million genomes). This is also performed for each SARS-CoV-2 protein. Multiplication by 100 transforms the probability of erroneously calling a TFP into a percentage for easier interpretation.

Similarly, a separate error rate ( $\epsilon$  or codon-aware FDR) is also computed relative to the sites at first and second codon positions as follows:

$$\epsilon_t = \frac{\text{FDR}_t \times Z_t}{j} \quad (6)$$

where  $Z$  is the number of TFP calls at codon positions one and two, and  $j$  is the total number of polymorphic sites at first and second positions with > 100 mutated sequences at the given site.

Both calculations were performed for the two analysed time periods, with slightly smaller error rates for the timeframe before Omicron emergence.



## Results

We utilised our new approach (Figure 1)<sup>31</sup> to analyse 1,275,669 SARS-CoV-2 whole genome sequences from England sampled between June 2020 and April 2022. This time period encompasses: (i) a period from June to December 2020 dominated by A.\*, B.\* and B.1.177 lineages, (ii) a timeframe between January and May 2021 when Alpha (B.1.1.7) predominated, (iii) a wave from June to December 2021 characterised by rapid spread of Delta (AY.4.\* and other AY.\*), and (iv) the Omicron (BA.1 and BA.2) epidemic cycle from December 2021 to April 2022 (Figure 2). For clarity and due to the main patterns observed, the analytic period presented includes: (i) June 2020 to mid-November 2021 (pre-Omicron interval) and (ii) mid-November 2021 to April 2022 (including Omicron). Only data collected through community sampling (Pillar 2) were included to reduce bias towards more severe infections and avoid the inclusion of data that was collected for special purposes. The geographical representation of the data is similar across different regions of England, with a mean proportion of community cases having a sequence of 6.7% and median of 7.1% (Figure 2, inset plot).

### Lineages and genomic regions enriched with SARS-CoV-2 TFP-homoplasies

The presence of recurring synonymous polymorphisms classified as TFP-homoplasies allowed us to investigate the FDR for each genomic region as a function of the applied cluster thresholds. The frequency of synonymous mutations along the genome (Extended Data Figure S2)<sup>31</sup> and the FDR across genomic regions (Extended Data Figure S3)<sup>31</sup> support that sites only detected at thresholds  $\geq 10\%$  must be investigated with caution since they generally have associated FDRs  $\approx 40\%$  and  $\epsilon \approx 15\%$ , whereas cutoffs  $\leq 2\%$  retain an acceptable FDRs  $\approx 10\%$  and  $\epsilon \approx 2\%$ . Additionally, these more erroneous thresholds ( $\geq 10\%$ ) represent  $> 50\%$  of the identified TFP sites (Figure 3A).

Therefore, among the ten cluster thresholds ranging from the more strict (0.25%) to the more lenient (25%) values (Extended Data Text S1, Extended Data Figure S4)<sup>31</sup>, we report results with the 2% threshold and after performing rigorous quality control using an alignment-aware artifact removal method to represent sites under putative multilevel selection (see [Methods: Statistical analysis for identifying genomic regions enriched for TFPs](#)). With this threshold, the false discovery rate (FDR) and codon-aware FDR ( $\epsilon$ ) (see [Methods: Codon-aware false discovery rate \(FDR\)](#)) are respectively around 10% and 2.5% (Extended Data Figure S3)<sup>31</sup>. Results for sites identified with other thresholds are presented in the Extended Data<sup>31</sup>.

For the period predating Omicron BA.1.\* emergence, we found that B.1.1.7 was consistently the lineage with the highest coefficient for enrichment of TFP-homoplasies, reaching statistical significance for  $\geq 2\%$  thresholds. However, TFP enrichment was similar across lineages and not found for specific genomic regions (Extended Data File S1)<sup>31</sup>. Although not significantly different, TFP-homoplasies were slightly more abundant in the small linker domain<sup>35</sup> of the N protein, ORF3a, and ORF8 for this time period (Figure 3B and C).

All considered major lineages were associated with increased TFP-homoplasies emergence for  $\geq 1\%$  thresholds during the timeframe which includes Omicron BA.1.\* as the dominant variant. Although not consistent for different thresholds, the rank of lineage coefficients was Other lineages  $> B.1.1.7 > AY.4.* > BA.1.* > AY.*$  (non-AY.4.\*) for threshold = 2%. Once again, there were no statistically significant results at the 2% threshold regarding genomic regions (Extended Data File S2)<sup>31</sup>. However, N:linker domain, ORF3a, S:FCS, ORF7a, and ORF10 presented a higher number of TFPs per site (Figure 3D and E), which is relatively consistent with the preceding period.

Normalising homoplasies counts by the size of each genomic region (giving less weight to larger genomic regions) has a large influence in interpreting the relative rates of TFP acquisition. This is especially demonstrated by NSP3, which accrues more TFPs due to its size of 5835 nucleotides (Figure 3B and D), but when normalised has a similar distribution of TFPs compared to other genomic regions (Figure 3C and E).

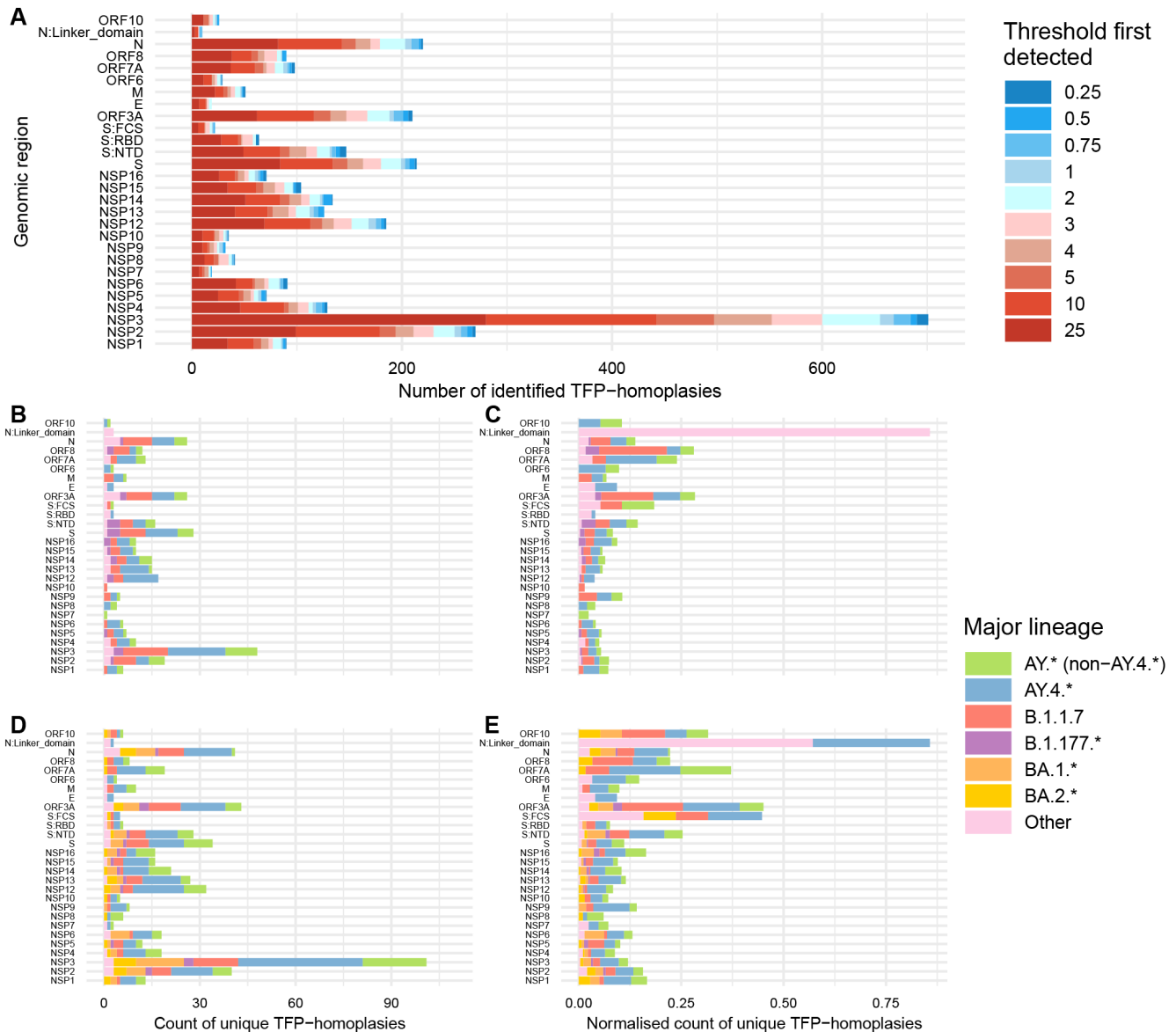
### TFPs along the SARS-CoV-2 genome and comparison with other approaches for detecting sites under selection

When comparing the 30 most frequent `mlscluster` TFP-homoplasies against the 30 mutations under positive selection detected by the HyPhy-based approach<sup>22</sup> for the cluster threshold = 2% and period before Omicron emergence, we detected three concordant sites (S:67, S:95, and S:484), 27 positively selected sites only detected by HyPhy, and 63 sites only detected by `mlscluster` statistics (Extended Data Figure S5)<sup>31</sup>. For the timeframe including Omicron, we identified two sites under selection concordantly between methods and with the previous period (S:67 and S:484), 28 discordant results, and 51 new potential TFPs across seven proteins/ORFs (Figure 4A).

The top 30 most frequent TFP-homoplasies across lineages shows that the B.1.1.7 ( $n=22$ ) and the AY.4.\* ( $n=21$ ) were similarly enriched for those highly-frequent TFP-homoplasies up to mid-November 2021 (Extended Data Figure S5)<sup>31</sup>, while AY.4.\* ( $n=20$ ) was notably the major lineage harbouring more TFPs (Figure 4B, Table 1) when considering Omicron BA.1.\*. Moreover, the analysis of lineage-specific top 30 TFP-homoplasies regardless of threshold shows that less than half of those are firstly detected on smaller cluster thresholds (up to 2%) (Extended Data Figures S6-S10)<sup>31</sup>.

When expanding to the top 100 TFP-homoplasies (Extended Data Table S1)<sup>31</sup>, 21 sites are located within the larger NSP3 which has 1945 amino acids, 14 are from ORF3a (275 sites), 12 from spike excluding NTD, RBD, and FCS (721 amino acids), nine from NTD (292 sites), four from FCS (38 amino acids), two from RBD (223 sites), and seven from the N protein (420 amino acids). Respectively, AY.4.\* and AY.\* account for 50 and 28 of these top 100 sites, followed by B.1.1.7 ( $n=24$ ).

A manual inspection of TFP-homoplasies with frequency  $\geq 5$  (Extended Data Table S1)<sup>31</sup> confirmed that they emerged independently in multiple lineages during the pandemic and



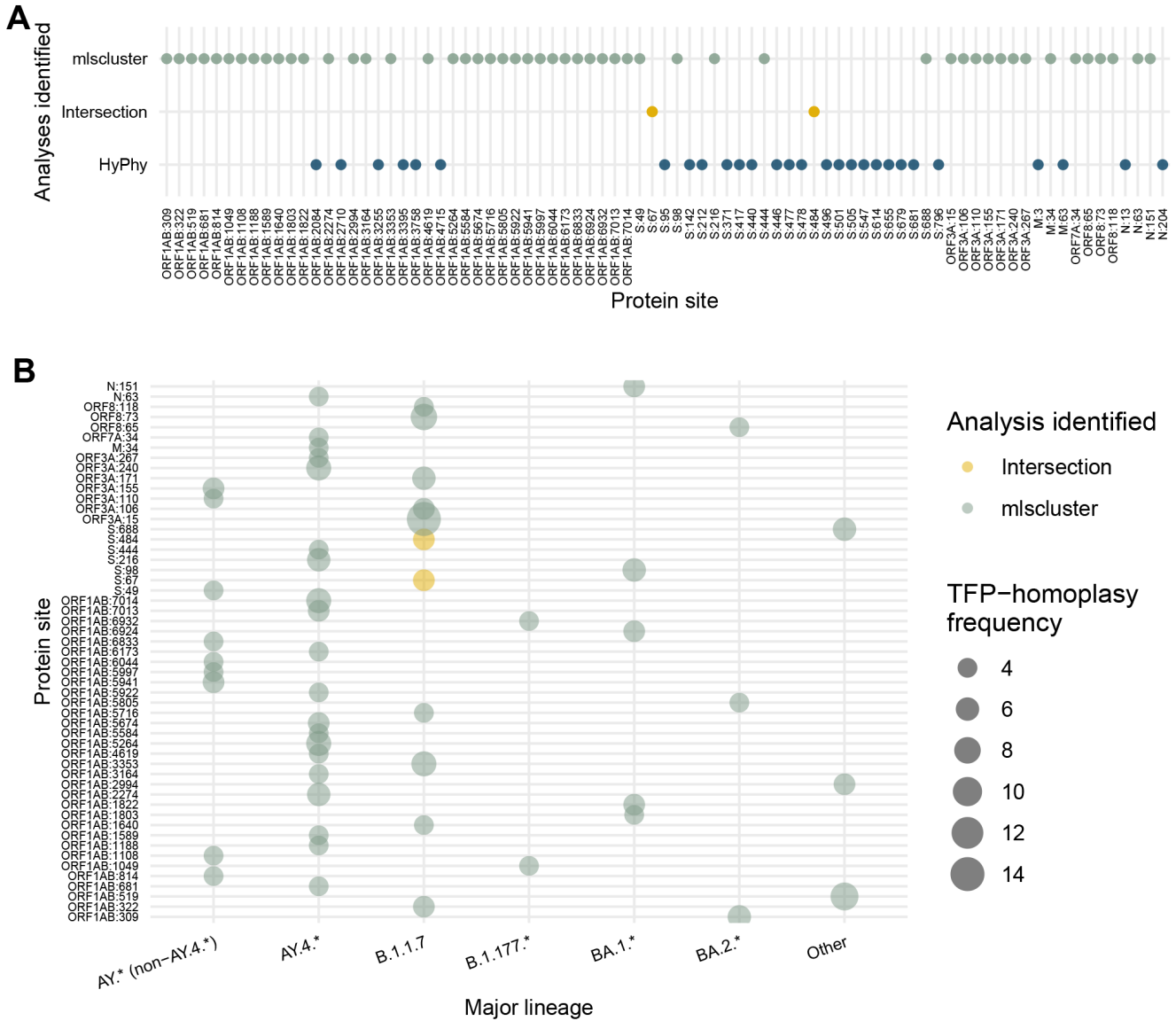
are predominantly found in extremely low (<1%) frequencies. This independent analysis provides additional evidence that their evolution is consistent with a transient selective pressure. Additionally, it shows that the impact of very few mutations outside the S protein have been characterised experimentally (Extended Data Table S1)<sup>31</sup>.

By focusing on individual proteins that harbour a major functional significance and higher normalised count of TFPs (S, N, ORF3a, ORF7a, and ORF8), we highlight relevant sites under multilevel selection for further experimental investigations.

These sites include S:A67V, S:S98F, S:L216F, S:E484K, S:A688V, N:P151L, ORF3a:L15F, ORF8:Y73C, etc. Additionally, these transient selective processes are more likely to be acting uniformly across each protein and not in specific hotspots (Figure 5, Extended Data Figure S11, Extended Data Table S1)<sup>31</sup>.

## Discussion

We have quantified transient selective forces acting on SARS-CoV-2 lineages and mutations through the calculation of three statistics (ratio of sizes, ratio of persistence time and



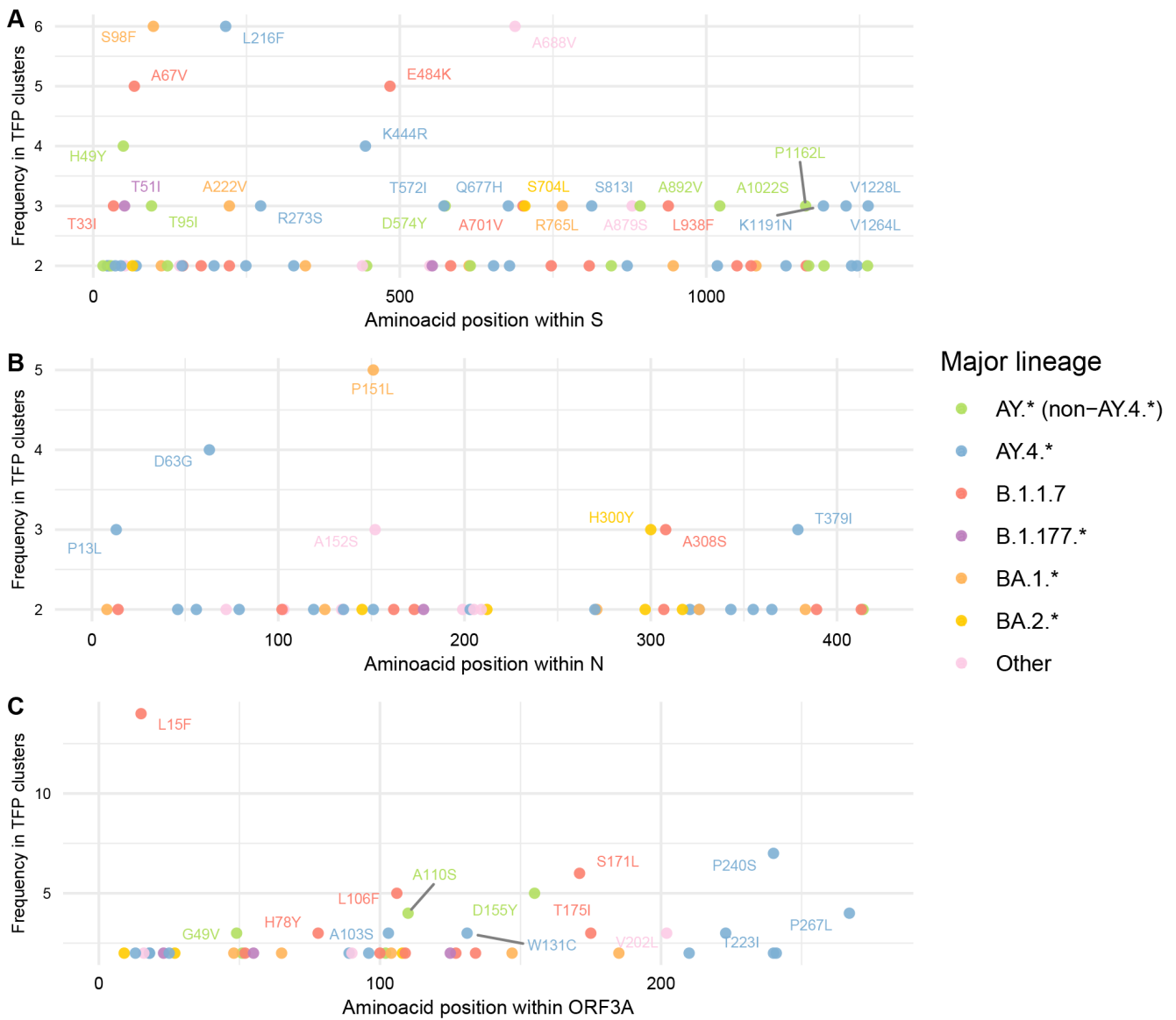
**Figure 4. TFP-homoplasy identification compared to sites identified under positive selection.** Sites are compared across different major lineages. **(A)** Comparison of top 30 identified sites under multilevel selection by our tree-based clustering approach for the whole-period (including Omicron) for cluster threshold = 2% against the HyPhy analysis<sup>22</sup>, also presenting concordant results (intersection) between both methods. **(B)** Bubble plot of TFP-homoplasy frequencies attributed to different major PANGO-lineages.

**Table 1. Top 30 TFP-homoplasies within the spike protein for the period between June 2020 and April 2022 (including Omicron) and cluster threshold = 2%.**

Homoplasy	Frequency	Major lineage	HyPhy	Genomic region	Amino acid length of protein
ORF3A:L15F	14	B.1.1.7	No	ORF3A	275
ORF1AB:G519S	9	Other	No	NSP2	638
ORF8:Y73C	8	B.1.1.7	No	ORF8	121
ORF1AB:H5264Y	7	AY.4.*	No	NSP12	932
ORF1AB:K3353R	7	B.1.1.7	No	NSP5	306
ORF1AB:R7014N	7	AY.4.*	No	NSP16	298
ORF3A:P240S	7	AY.4.*	No	ORF3A	275

Homoplasy	Frequency	Major lineage	HyPhy	Genomic region	Amino acid length of protein
ORF1AB:P309L	6	BA.2.*	No	NSP2	638
ORF1AB:T2274I	6	AY.4.*	No	NSP3	1945
ORF3A:S171L	6	B.1.1.7	No	ORF3A	275
<b>S:A688V</b>	6	Other	No	S:FCS	38
<b>S:L216F</b>	6	AY.4.*	No	S:NTD	292
<b>S:S98F</b>	6	BA.1.*	No	S:NTD	292
N:P151L	5	BA.1.*	No	N	413
ORF1AB:A2994V	5	Other	No	NSP4	500
ORF1AB:K322R	5	B.1.1.7	No	NSP2	638
ORF1AB:L6924F	5	BA.1.*	No	NSP16	298
ORF1AB:P7013L	5	AY.4.*	No	NSP16	298
ORF1AB:S5674L	5	AY.4.*	No	NSP13	601
ORF1AB:T1822I	5	BA.1.*	No	NSP3	1945
ORF1AB:T5941I	5	AY.* (non-AY.4.*)	No	NSP14	527
ORF3A:D155Y	5	AY.* (non-AY.4.*)	No	ORF3A	275
ORF3A:L106F	5	B.1.1.7	No	ORF3A	275
<b>S:A67V</b>	5	B.1.1.7	Yes	S:NTD	292
<b>S:E484K</b>	5	B.1.1.7	Yes	S:RBD	223
M:L34F	4	AY.4.*	No	M	222
N:D63G	4	AY.4.*	No	N	413
ORF1AB:A1049V	4	B.1.177.*	No	NSP3	1945
ORF1AB:A5922V	4	AY.4.*	No	NSP13	601
ORF1AB:A6044V	4	AY.* (non-AY.4.*)	No	NSP14	527
ORF1AB:D5584Y	4	AY.4.*	No	NSP13	601
ORF1AB:G6173V	4	AY.4.*	No	NSP14	527
ORF1AB:H1108Y	4	AY.* (non-AY.4.*)	No	NSP3	1945
ORF1AB:L681F	4	AY.4.*	No	NSP2	638
ORF1AB:M5997I	4	AY.* (non-AY.4.*)	No	NSP14	527
ORF1AB:P1640S	4	B.1.1.7	No	NSP3	1945
ORF1AB:P1803S	4	BA.1.*	No	NSP3	1945
ORF1AB:P4619L	4	AY.4.*	No	NSP12	932
ORF1AB:P6932S	4	B.1.177.*	No	NSP16	298
ORF1AB:R3164H	4	AY.4.*	No	NSP4	500
ORF1AB:R5716C	4	B.1.1.7	No	NSP13	601
ORF1AB:S1188L	4	AY.4.*	No	NSP3	1945
ORF1AB:T1589I	4	AY.4.*	No	NSP3	1945
ORF1AB:T5805M	4	BA.2.*	No	NSP13	601
ORF1AB:T6833I	4	AY.* (non-AY.4.*)	No	NSP16	298

Homoplasmy	Frequency	Major lineage	HyPhy	Genomic region	Amino acid length of protein
ORF1AB:T814I	4	AY.* (non-AY.4.*)	No	NSP2	638
ORF3A:A110S	4	AY.* (non-AY.4.*)	No	ORF3A	275
ORF3A:P267L	4	AY.4.*	No	ORF3A	275
ORF7A:P34L	4	AY.4.*	No	ORF7A	121
ORF8:A65V	4	BA.2.*	No	ORF8	121
ORF8:L118V	4	B.1.1.7	No	ORF8	121
<b>S:H49Y</b>	4	AY.* (non-AY.4.*)	No	S:NTD	292
<b>S:K444R</b>	4	AY.4.*	No	S:RBD	223



**Figure 5. Frequency of identified TFP-homoplasies alongside genomic regions with major functional significance and normalised counts for cluster threshold = 2% and period including Omicron. (A) Spike. (B) Nucleocapsid. (C) ORF3a. TFPs are coloured by major PANGO-lineage and annotated if frequency > 2.**

logistic growth rates between sister clades) and extraction of clades containing values of those statistics below small cluster threshold cutoffs. To mitigate the inclusion of spurious sites, we included only recurring clade-defining mutations (homoplasies) across cluster thresholds with low associated FDRs and excluded probable sequencing artifacts. To the best of our knowledge, this is the first attempt to identify SARS-CoV-2 polymorphisms that negatively influence transmission fitness while being beneficial for within-host replication. Our tree-based clustering approach provides a scalable way to analyse huge genomic datasets with >1 million sequences for multilevel selection while also accounting for shared ancestry.

Although the COVID-19 pandemic offered the opportunity to collate genomic datasets of unprecedented sizes, estimating the transmission fitness of individual polymorphisms in this context is challenging. In the early epidemic, inference of sites under positive selection was hampered by low sensitivity given the small genetic diversity of the virus. For example, a phylogenetic approach was developed to quantify imbalance in clades containing recurrent mutations<sup>5</sup>, and this approach found a lack of evidence for increased transmissibility from recurrent SARS-CoV-2 mutations. However, this approach only used ≈50,000 sequences up to July 2020 and has not considered persistence times and growth rates as measures of differential fitness across clades of the phylogeny. After the emergence of VOCs with elevated substitution rates, other attributes such as convergent evolution, sparse sampling, and vaccine-elicited immunity appeared as relevant confounding factors. Most importantly, the detection of positive selection does not necessarily imply enhanced transmissibility, and the effects of individual mutations on this trait will typically be modest<sup>6</sup>.

Genetic diversity in an infected individual is governed by repeated cycles of within-host (e.g. replication and immune escape pressures) and between-host processes (e.g. transmission bottlenecks), with the outcome of selection at each level having an effect on the other<sup>37</sup>. The rapid accumulation of mutations in individuals<sup>38,39</sup> with long-lasting chronic SARS-CoV-2 infection is hypothesised to contribute to the emergence of variants such as Alpha and Omicron<sup>40</sup>. Thus the interaction of within-host and between-host selective processes can occasionally have very large epidemic-level effects.

The inspection of the global and lineage distributions of highly frequent TFP-homoplasies confirmed that these mutations generally emerge independently in multiple lineages but remain quite rare, which is consistent with a simultaneous within-host advantage and between-host disadvantage. This systematic investigation also emphasises the scarcity of experimental studies to characterise the functional impact of mutations outside the spike protein.

Our approach identified modest differences in multilevel selection signals across two different epidemic phases, lineages and genomic regions in the UK. We hypothesised that transient

selective forces would become stronger after high-levels of convalescent and vaccine-induced immunity have been reached, but our results do not support this hypothesis. Our observation of approximately constant levels of transient selection across waves driven by extremely distinct variants may in part be driven by long-duration chronic infections which occur at low frequency and provide greater opportunities for accelerated within-host evolution favouring immune evasion. Our data did not include clinical covariates that would allow us to investigate the association of chronic infection or duration of infection and the presence of TFPs.

Sequencing of chronically-infected patients throughout multiple time points of their long-lasting infection provided external validation of our observed patterns. Nucleotide substitution rates were around twice as fast during chronic infections when compared with the global SARS-CoV-2 evolutionary rate<sup>41</sup>. Additionally, mutations identified in the top 100 most frequent TFP-homoplasies by our approach such as S:E484K<sup>41-44</sup>, S:T95I<sup>43,44</sup>, ORF8:Y73C<sup>42</sup>, ORF8:L118V<sup>41</sup>, ORF1ab:S944L<sup>41</sup>, and ORF1ab:T1543I<sup>41</sup> also emerged after days of chronic infection. Although usually associated with immune escape and increased ACE2 affinity, these recurrent mutations lack the capacity to enhance transmission<sup>43,44</sup> as demonstrated by their low epidemic-level frequency after multiple independent occurrences. Additionally, distinct viral populations appear to be residing in different niches (e.g. organs) of a patient's body<sup>44</sup> and an impaired immune system selects for mutations that confer intra-host replication and persistence (e.g. immune evasion) as opposed to general acute infections, in which mutations favouring inter-host transmission are a major target of selective pressures<sup>43</sup>.

Despite distance-based clustering in HIV networks having been extensively used as a proxy for transmissibility, this approach is generally based on a cutoff from pairwise distances separating sequences<sup>16</sup>. Consequently, poor specificity for variants negatively influencing fitness is evident (i) when a variant is isolated, occurring along a long branch not captured by distance threshold, (ii) when a variant is imported, and large genetic distances can reflect unsampled diversity in the country of origin or a rare recombination or hypermutation event, not necessarily reflecting a fitness cost. Our method addresses these limitations by incorporating the number of descendants, persistence times, and growth rates across sister clades with and without the mutations under investigation, and using independently-acquired substitutions to remove spurious relationships. Introducing these multiple sources of information can provide more accurate estimates, but also introduce biases. Primarily, our analysis is sensitive to sequencing artifacts. Although we used data from a highly standardised sequencing consortium (COG-UK), changes in primer sets after Omicron emergence<sup>32</sup>, as well as sequencing coverage and base-calling errors can potentially influence our conclusions, as demonstrated by our several quality control and artifact removal methods employed. A second caveat arises from the assumption of representative sampling. Although we utilised data from England during a period of proportional (to cases) community sampling to minimise this effect, the rate of sampling varied substantially

over time and further analyses are needed to investigate the impact of non-representative sequencing in our approach.

## Conclusions

We developed a method capable of identifying sites under multilevel selection from >1.2 million SARS-CoV-2 sequences using rigorous quality control, statistical tests, and control for false detection. The comprehensive catalog of TFPs identified here and especially abundant in S, N, ORF3a, ORF7, and ORF8 highlight the existence of important tradeoffs between within-host replication and between-host transmission of SARS-CoV-2 that may warrant further experimental investigation.

## Data availability

### Underlying data

Zenodo: Underlying and Extended data for: Phylogenetic signatures reveal multilevel selection and fitness costs in SARS-CoV-2. <https://doi.org/10.5281/zenodo.10522713><sup>31</sup>.

This project contains the following underlying data:

- ExtendedData.pdf supplementary text, figures, and table cited in the paper
- ExtendedData\_FileS1.txt - output of Poisson regressions across the ten employed cluster thresholds for the time period before Omicron BA.1.\* emergence (early June 2020 to mid-November 2021) that tested whether TFPs were enriched in particular genomic regions or major PANGO lineages.
- ExtendedData\_FileS2.txt - output of Poisson regressions across the ten employed cluster thresholds for the time period including Omicron BA.1.\* emergence (early June 2020 to the end of April 2022) that tested whether TFPs were enriched in particular genomic regions or major PANGO lineages.
- ExtendedData\_FileS3.zip - this ZIP file contains other four files:
  - sc2\_md\_curated\_WITH\_Xs\_Ns.rds - underlying preprocessed metadata file to use as input for `mlscluster` to reproduce the analysis.
  - sc2\_tre\_curated.rds - underlying preprocessed time-scaled phylogenetic tree file to use in combination with the metadata file as input for `mlscluster` to reproduce the analysis.

- res\_p2.rds - output of a time-consuming run of the `mlsclust` function (<https://github.com/mrc-ide/mlscluster/blob/main/R/mlsclust.R>) for the period excluding Omicron (June 2020 to mid-November 2021).
- res\_p3.rds - output of a time-consuming run of the `mlsclust` function for the period including Omicron (June 2020 to April 2022).

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) (CC-BY 4.0).

### Extended data

Analysis code available from: <https://github.com/vinibfranc/mls-cluster-experiments>

Archived source code at time of publication: <https://doi.org/10.5281/zenodo.10520060><sup>25</sup>

License: MIT

### Software availability

We implemented the inference methods described in this paper into a new R package entitled `mlscluster` Source code available from: <https://github.com/mrc-ide/mlscluster>

Archived source code at time of publication: <https://doi.org/10.5281/zenodo.10523291><sup>23</sup>

License: MIT

### Author contributions

Conceptualisation: EV; Data curation: VBF; Formal analysis: VBF; Investigation: VBF; Methodology: VBF, EV; Funding acquisition: EV; Project administration: EV; Resources: EV; Software: VBF, EV; Supervision: EV; Validation: VBF; Visualisation: VBF; Writing—original draft: VBF; Writing—review and editing: VBF, EV. All authors have read and agreed to the published version of the manuscript.

### Acknowledgements

The authors gratefully thank all members and contributors of the COG-UK consortium who are listed at <https://webarchive.nationalarchives.gov.uk/ukgwa/20230507102826/https://www.cogconsortium.uk/about/contributors/key-contributors/>, in addition to the Imperial College High Performance Computing Service ([doi.org/10.14469/hpc/2232](https://doi.org/10.14469/hpc/2232)).

## References

1. Kimura M: **The neutral theory of molecular evolution**. Cambridge University Press, 1983. [PubMed Abstract](#) | [Publisher Full Text](#)
2. Lemey P, Rambaut A, Pybus OG: **HIV evolutionary dynamics within and among hosts**. *Aids Rev*. 2006; **8**(3): 125–140. [PubMed Abstract](#)
3. Fraser C, Lythgoe K, Leventhal GE, *et al.*: **Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective**. *Science*. 2014; **343**(6177): 1243727. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Carlson JM, Schaefer M, Monaco DC, *et al.*: **HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck**. *Science*. 2014; **345**(6193): 1254031. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. van Dorp L, Richard D, Tan C, *et al.*: **No evidence for increased transmissibility**

- from recurrent mutations in sarscov-2. *Nat Commun.* 2020; **11**(1): 15986. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Volz E, Hill V, McCrone JT, et al.: **Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity.** *Cell.* 2021; **184**(1): 64–75.e11. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  7. Hodcroft EB, Zuber M, Nadeau S, et al.: **Spread of a SARS-CoV-2 variant through europe in the summer of 2020.** *Nature.* 2021; **595**(7869): 707–712. [PubMed Abstract](#) | [Publisher Full Text](#)
  8. Volz E, Mishra S, Chand M, et al.: **Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England.** *Nature.* 2021; **593**(7858): 266–269. [PubMed Abstract](#) | [Publisher Full Text](#)
  9. Kraemer M, Hill V, Ruis C, et al.: **Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence.** *Science.* 2021; **373**(6557): 889–895. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  10. Viana R, Moyo S, Amoako DG, et al.: **Rapid epidemic expansion of the SARS-CoV-2 omicron variant in southern africa.** *Nature.* 2022; **603**(7902): 679–686. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  11. Boyd O, Volz E: **tfpsscanner.** 2021. [Reference Source](#)
  12. Obermeyer F, Jankowiak M, Barkas N, et al.: **Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness.** *Science.* 2022; **376**(6599): 1327–1332. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  13. Jankowiak M, Obermeyer FH, Lemieux JE: **Inferring selection effects in SARS-CoV-2 with bayesian viral allele selection.** *PLoS Genet.* 2022; **18**(12): e1010540. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  14. Dearlove BL, Frost SD: **Measuring asymmetry in time-stamped phylogenies.** *PLoS Comput Biol.* 2015; **11**(7): e1004312. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  15. Volz EM, Carsten W, Grad YH, et al.: **Identification of hidden population structure in time-scaled phylogenies.** *Syst Biol.* 2020; **69**(5): 884–896. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  16. Kosakovsky Pond SL, Weaver S, Leigh Brown A, et al.: **HIV-TRACE (TRANSMISSION CLUSTER ENGINE): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens.** *Mol Biol Evol.* 2018; **35**(7): 1812–1819. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  17. Wertheim JO, Oster AM, Switzer WM, et al.: **Natural selection favoring more transmissible HIV detected in United States molecular transmission network.** *Nat Commun.* 2019; **10**(1): 5788. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  18. Wertheim JO, Oster AM, Johnson JA, et al.: **Transmission fitness of drug-resistant HIV revealed in a surveillance system transmission network.** *Virus Evol.* 2017; **3**(1): vex008. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  19. Kühnert D, Kouyos R, Shirreff G, et al.: **Quantifying the fitness cost of HIV-1 drug resistance mutations through phylodynamics.** *PLoS Pathog.* 2018; **14**(2): e1006895. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  20. Yang Z: **PAML 4: Phylogenetic Analysis by Maximum Likelihood.** *Mol Biol Evol.* 2007; **24**(8): 1586–1591. [PubMed Abstract](#) | [Publisher Full Text](#)
  21. Pond SLK, Frost SDW, Muse SV: **HyPhy: hypothesis testing using phylogenies.** *Bioinformatics.* 2005; **21**(5): 676–679. [PubMed Abstract](#) | [Publisher Full Text](#)
  22. Pond S: **Evidence of natural selection history operating on SARS-CoV-2 genomes.** 2022. [Reference Source](#)
  23. Franceschi V: **mrc-ide/miscluster: Publication archive.** 2024. <http://www.doi.org/10.5281/zenodo.10523291>
  24. Volz E: **Fitness, growth and transmissibility of SARS-CoV-2 genetic variants.** *Nat Rev Genet.* 2023; **24**(10): 724–734. [PubMed Abstract](#) | [Publisher Full Text](#)
  25. Franceschi V: **vinibfranc/miscluster-experiments: Publication archive.** 2024. <http://www.doi.org/10.5281/zenodo.10520060>
  26. Volz EM, Frost SDW: **Scalable relaxed clock phylogenetic dating.** *Virus Evol.* 2017; **3**(2): vex025. [PubMed Abstract](#) | [Publisher Full Text](#)
  27. Sagulenko P, Puller V, Neher RA: **Treetime: Maximum-likelihood phylodynamic analysis.** *Virus Evol.* 2018; **4**(1): vex042. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  28. Sanderson T: **Chronumental: time tree estimation from very large phylogenies.** *bioRxiv.* 2021. [PubMed Abstract](#) | [Publisher Full Text](#)
  29. Paradis E, Schliep K: **ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.** *Bioinformatics.* 2019; **35**(3): 526–528. [PubMed Abstract](#) | [Publisher Full Text](#)
  30. Maio ND, Walker C, Borges R, et al.: **Masking strategies for SARS-CoV-2 alignments.** *Virological.* 2020. [Reference Source](#)
  31. Bonetti Franceschi V, Volz E: **Underlying and extended data for Phylogenetic signatures reveal multilevel selection and fitness costs in SARS-CoV-2.** 2024. <http://www.doi.org/10.5281/zenodo.10522713>
  32. Davis JJ, Long SW, Christensen PA, et al.: **Analysis of the ARTIC Version 3 and Version 4 SARS-CoV-2 Primers and Their Impact on the Detection of the G142D Amino Acid Substitution in the Spike Protein.** *Microbiol Spectr.* 2021; **9**(3): e0180321. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  33. Li H: **seqtk: Toolkit for processing sequences in fasta/q formats.** 2023. [Reference Source](#)
  34. R Core Team: **R: A Language and Environment for Statistical Computing.** R Foundation for Statistical Computing, Vienna, Austria, 2022. [Reference Source](#)
  35. Syed AM, Taha TY, Tabata T, et al.: **Rapid assessment of SARS-CoV-2-evolved variants using virus-like particles.** *Science.* 2021; **374**(6575): 1626–1632. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  36. Wickham H: **ggplot2: Elegant Graphics for Data Analysis.** Springer-Verlag New York, 2016. [Reference Source](#)
  37. Mideo N, Alizon S, Day T: **Linking within-and between-host dynamics in the evolutionary epidemiology of infectious diseases.** *Trends Ecol Evol.* 2008; **23**(9): 511–517. [PubMed Abstract](#) | [Publisher Full Text](#)
  38. Kemp SA, Collier DA, Datir RP, et al.: **SARS-CoV-2 evolution during treatment of chronic infection.** *Nature.* 2021; **592**(7853): 277–282. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  39. Avanzato VA, Matson MJ, Seifert SN, et al.: **Case study: prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised individual with cancer.** *Cell.* 2020; **183**(7): 1901–1912.e9. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  40. Markov PV, Ghafari M, Beer M, et al.: **The evolution of SARS-CoV-2.** *Nat Rev Microbiol.* 2023; **21**(6): 361–379. [PubMed Abstract](#) | [Publisher Full Text](#)
  41. Chaguza C, Hahn AM, Petrone ME, et al.: **Accelerated SARS-CoV-2 intrahost evolution leading to distinct genotypes during chronic infection.** *Cell Rep Med.* 2023; **4**(2): 100943. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  42. Sonnleitner ST, Prelog M, Sonnleitner S, et al.: **Cumulative SARS-CoV-2 mutations and corresponding changes in immunity in an immunocompromised patient indicate viral evolution within the host.** *Nat Commun.* 2022; **13**(1): 2560. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  43. Wilkinson SAJ, Richter A, Casey A, et al.: **Recurrent SARS-CoV-2 mutations in immunodeficient patients.** *Virus Evol.* 2022; **8**(2): veac050. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  44. Harari S, Tahor M, Rutsinsky N, et al.: **Drivers of adaptive evolution during chronic SARS-CoV-2 infections.** *Nat Med.* 2022; **28**(7): 1501–1508. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)



# Open Peer Review

Current Peer Review Status: ? ? ?

---

## Version 1

Reviewer Report 17 April 2024

<https://doi.org/10.21956/wellcomeopenres.22912.r77882>

© 2024 Weber A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Ariane Weber** 

Max Planck Institute of Geoanthropology, Jena, Thuringia, Germany

### Summary

In “Phylogenetic signatures reveal multilevel selection and fitness costs in SARS-CoV-2” Bonetti Franceschi and Volz present a tree-based method, *m/scluster*, to identify homoplasies that are associated with the most unbalanced parts of the respective tree and apply it to SARS-CoV-2 genome sequences sampled in England between June 2020 and April 2022. The method includes a Poisson regression model to test if the identified homoplasies are enriched in specific genome regions, between different time intervals or in major PANGO lineages.

### General opinion

The method represents a fast and scalable way of extracting patterns of transmission heterogeneity caused by individual mutations from phylogenetic trees. It is thereby of topical relevance, as exemplified by the empirical analysis included in the article. Although the authors scrutinise their results by the calculation of a False Discovery Rate, my main concern relates to the lack of a more thorough evaluation of the performance of the method (see major comments). I deem this of particular relevance with regards to the interpretation of the identified homoplasies as transmission fitness polymorphisms providing evidence of multilevel selection.

All data necessary for and produced by the empirical analysis is publicly shared in a well-readable format, as well is all software. The article itself is sensibly structured and accompanied by visualisations for all important elements, *i.e.* methodology and results. In my opinion, the phrasing is lacking in clarity in some instances and would require rephrasing for easier readability (reason for response “Partly” to question “Is the work clearly and accurately presented and does it cite the current literature?”, see also minor comments). A more detailed discussion of the limitations of the method would improve the contextualisation, especially some, to me rather far-fetched, claims would need qualifying (see minor comments).

### Major comments

1. To my understanding, the methodological result of this study is a list of mutations that have occurred multiple times in the phylogeny in clades that are characterised by relatively small

values of one or all of the test statistics (size, persistence, growth). By arguing that the recurrence is evidence for a within-host transmission advantage and the clade characteristics for a between-host disadvantage, these homoplasies are interpreted as transmission fitness polymorphisms (TFP) and evidence for multilevel selection in SARS-CoV-2. However, especially given the small absolute frequency of reoccurrence of most identified TFPs, this interpretive step is not straightforward and a bridge between result and interpretation is missing. I would therefore recommend the addition of either a validation study on either synthetic or empirical data (providing evidence that the identified mutations are really the result of a transient transmission advantage) or of a very well-constructed argument supported by appropriate literature. (This is the reason why I chose the response "Partly" to "Are the conclusions drawn adequately supported by the results?")

2. Which target node conditions were used for the SARS-CoV-2 tree? How are the choices justified and how sensitive are the results to changes in these parameters? I do see in the code that the default values are `min_descendants = 10`, `max_descendants = 20*10^3`, `min_cluster_age_yrs = 1/12`, `min_date = max_date = NULL`, but I cannot find information regarding it in the manuscript.
3. Why is caution only recommended starting from thresholds associated with an FDR of around 40%?
4. If sister clades both qualify as target nodes, is the same (but reversed) comparison done twice and counted in the empirical distribution?
5. Are some parts of the tree counted multiple times, *e.g.* if a parent and child node both qualify as target nodes with the same defining mutation?
6. As demographic factors will still influence the clade characteristics, even in the presence of multilevel selection, how many independent samples are necessary, *i.e.* how often would the mutations have to be observed in the tree, to draw statistically reliable conclusions (see also Supplementary Figure S17 in van Drop et al. 2020)? What is empirically observed for the reported homoplasies?
7. Do the identified homoplasies also appear in parts of the tree that are not characterised as imbalanced?
8. How are multiple defining mutations in one clade handled?
9. I would recommend expanding the Discussion in the following two points:
  - More thorough discussion of the limitations that are already listed (how does the method deal with the challenges described in the second paragraph of the Discussion) and additional ones that apply (*e.g.*, phylogenetic uncertainty and reliable tree inference as potentially time-consuming preprocessing step, statistical power from only few recurrences)
  - Explanation for and discussion of the pronounced differences observed between the `mlscluster` and `HyPhy` results.
10. I would recommend a more careful phrasing of the possible explanation of the observed

TFP-homplasy pattern by accelerated within-host rates, as I see no ground justifying a connection presented in the study and too few references showing evidence of the connection and rate acceleration (see also preprint [1]).

### Minor comments

#### 1. Abstract:

- “We applied this method **for** a SARS-CoV-2 time-calibrated phylogeny..” --> “We applied this method **to** a SARS-CoV-2 time-calibrated phylogeny..”
- To my understanding, the conclusion statement is not supported by the study results (see also major comment 1), as they do not allow to draw direct conclusions about within-host vs. between-host evolution. I would recommend rephrasing this.

#### 2. Plain Language Summary:

- I would recommend using “clade” instead of “lineage” for consistency.
- “...highlighting the existence of important tradeoffs in selection between intrahost replication and inter-host transmission”: I would recommend rephrasing this or providing evidence that the identified homplasies are necessarily the product of multilevel selection (see major comment 1)

#### 3. Introduction:

- A literature reference is missing for the statement “In molecular epidemiological studies, a set of particularly scalable approaches have been developed based on the calculation of phylogenetic clusters comprising two or more closely related samples.” Additionally, I would suggest replacing “calculation” with, for example, “identification”, “detection” or similar.
- Either a literature reference or a clearer argument is missing for the statement “Furthermore, there is considerable scope to improve on distance-based genetic clustering methods because such approaches will potentially have poor specificity for variants that negatively influence fitness.”
- In the sentence following the one cited in the previous point the reference to SARS-CoV-2 is duplicated.
- A formal definition of “transmission fitness polymorphism” is missing.
- “We demonstrated its applicability through the analysis of a representative >1.2 million SARS-CoV-2 genomic data set from England...” should, *e.g.*, rather be phrased as “We demonstrated its applicability through the analysis of a representative SARS-CoV-2 data set comprising >1.2 million genome sequences from England...”
- “By providing a comprehensive catalog of the main sites driving multilevel selective pressures throughout the SARS-CoV-2 genome, we also expand the understanding of [**the**] SARS-CoV-2 fitness landscape outside the well-studied spike protein”. I would recommend rephrasing of this sentence, as (i) the identified sites are not the drivers of the selective pressures, rather the result of them and (ii) see major comment 1.

#### 4. Methods

- Usage of “node” and “clade” is inconsistent. I would recommend primarily using only one of the two terms and, if necessary, write “node” only to refer to the actual internal node and “clade” when referring to the whole subtree arising from said “node”.
- “Assume two clades u (target clade) and v (comparator/sister) organised in a time-scaled tree t and sharing ancestry (*i.e.* the same defining mutations).” --> “Assume two clades u (target clade) and v (comparator/sister) organised in a time-scaled tree t and sharing **full** ancestry (*i.e.* the same defining mutations).”

- “The persistence time (given by  $a$  in Figure 2), is defined as..” --> “The persistence time (given by  $a$  in Figure 2), is defined as..” (closing bracket at wrong position)
- “...which can be their sister clade (the clade sharing an immediate ancestor assuming bifurcating phylogenetic relationship) or against all other clades...” --> “...which can be their sister clade (the clade sharing an immediate ancestor assuming bifurcating phylogenetic relationship) or all other clades...”
- Unclear if ‘clade size’ refers to  $S_{uv}$  or  $n_u$
- I would recommend explicitly stating that in this study it is assumed that a homoplasy that arises in a subtree for which the statistics fall below a low quantile of the empirical distribution (see sentence “... for identifying especially convergently acquired mutations (homoplasies) that are detrimental for transmission (within a low quantile of the probability distribution of at least one of the three statistics”).
- The observed distribution of each of the three statistics in the tree is formally not a probability distribution. I would recommend replacing this terminology.
- “Given this efficient way to visit nodes of the tree and edge lengths, we can easily extract the parameters of interest (e.g. the time of the most recent common ancestor of each node, ...” --> “Given this efficient way to visit nodes of the tree and edge, we can easily extract the parameters of interest (e.g. the time of the most recent common ancestor of each **clade**, ...”
- “Target nodes are extracted based on the following conditions ...” The following needs rephrasing, as it currently reads as if only the node with the smallest number of descendants etc. is kept.
- Definition of “sharing ancestry” as same defining mutations between sister clades and “defining mutation” in one sister clade is contradicting.
- Which homoplasy annotations are used later in the text?
- “We intended to make the method flexible by creating a parameter that specifies in how many percentiles the statistic should be splitted...”: Isn’t it rather a parameter that specifies which percentile to use (e.g. 2%)?
- I find the terminology “cluster threshold” confusing, as the output of the method is not primarily a clustering of parts of the tree into groups, but rather the identification of recurring mutations based on relative differences between any sister clades (passing the filtering conditions).
- A literature reference is missing for the ML tree inference.
- “We tested our approach using two COVID-19 pandemic time-periods: (i) from June 01, 2020 (including Wuhan/WH04/2020 reference sequence as root of the phylogeny) to November 15, 2022 (before Omicron BA.1.\* variant emergence)..” --> “We tested our approach using two COVID-19 pandemic time-periods: (i) from June 01, 2020 (including Wuhan/WH04/2020 reference sequence as root of the phylogeny) to November 15, **2021** (before Omicron BA.1.\* variant emergence)..”
- The sentence “For each period, 10 different thresholds (...) of the clustering statistics are computed” needs rephrasing, e.g. “For each period, 10 different thresholds (...) of the clustering statistics are considered”. I would also suggest adding “%” after each threshold value.
- Alignment-aware artifact removal: Is it correct that this is only necessary if occurrences of “X” and “N” at each genome position are not randomly distributed over the tree?
- Poisson regression: definition of  $i$  and  $j$  is missing

- Definition of Y as TFP calls among the I polymorphic third codon position sites with >100 mutated sequences: Why 100?
- “Multiplication by 100 transforms the probability of erroneously calling a TFP into a percentage for easier interpretation”: The FDR does not represent the probability of erroneously calling a TFP, I would recommend rephrasing this.
- Closing bracket missing: “(see Methods: Statistical analysis for identifying genomic regions enriched for TFP).”
- “This independent analysis provides additional evidence that their evolution is consistent with a transient selective pressure.”: Why is this the case? To my understanding, the manual inspection of the relatively frequent TFP homoplasies mainly shows that the method worked as expected here (i.e. the identified homoplasies are recurrent and in ‘small’ clades)

#### 5. Discussion

- “We have quantified transient selective forces acting on SARS-CoV-2 lineages and mutations...”: I would recommend rephrasing this, as the study results do not represent a quantification of the selective forces.
- “After the emergence of VOCs with elevated substitution rates, other attributes...”: I would recommend rephrasing this, since it reads as if VOCs have elevated evolutionary rates.
- “Genetic diversity in an infected individual is governed by repeated cycles...” --> “Genetic diversity in infected individuals is governed by repeated cycles...”
- “Additionally, mutations identified in the top 100 most frequent TFP-homoplasies [...] also emerged after days of chronic infection.”: Since they emerged within days, were they shown to only emerge in chronically infected individuals and not others?

#### 6. Conclusions:

- “We developed a method capable of identifying sites under multilevel selection...”: I would recommend being more careful with the wording here, see major comment 1.

### References

1. Översti S, Gaul E, Jensen B, Kühnert D: Phylogenetic meta-analysis of chronic SARS-CoV-2 infections in immunocompromised patients shows no evidence of elevated evolutionary rates. *bioRxiv*. 2023. [Publisher Full Text](#)

#### **Is the work clearly and accurately presented and does it cite the current literature?**

Partly

#### **Is the study design appropriate and is the work technically sound?**

Yes

#### **Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

#### **If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

#### **Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Pathogen phylodynamics

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 15 Jul 2024

**Vinicius Bonetti Franceschi**

Thank you very much for your time to review the paper. We apologise for the delay in getting back to you. This was due to other concurring projects and our tentative to address the high number of comments raised by the three reviewers in the best possible way. We appreciate your very detailed and thoughtful comments, and believe that they helped to improve the readability and contextualisation of the paper. Most importantly, we tried to make the wording less definitive about the sites actually being TFPs and explain that these will be used as input for more rigorous coalescent-based modelling in the near future to confirm or not their multilevel selection effects. Please see below the response to the specific points:

### **Major comments**

**1. To my understanding, the methodological result of this study is a list of mutations that have occurred multiple times in the phylogeny in clades that are characterised by relatively small values of one or all of the test statistics (size, persistence, growth). By arguing that the recurrence is evidence for a within-host transmission advantage and the clade characteristics for a between-host disadvantage, these homoplasies are interpreted as transmission fitness polymorphisms (TFP) and evidence for multilevel selection in SARS-CoV-2. However, especially given the small absolute frequency of reoccurrence of most identified TFPs, this interpretive step is not straightforward and a bridge between result and interpretation is missing. I would therefore recommend the addition of either a validation study on either synthetic or empirical data (providing evidence that the identified mutations are really the result of a transient transmission advantage) or of a very well-constructed argument supported by appropriate literature. (This is the reason why I chose the response "Partly" to "Are the conclusions drawn adequately supported by the results?")**

Response: Thank you very much for your comment. We tried to make it clearer now throughout the paper that the objective of mlscluster is to detect CANDIDATE sites under multilevel selection to feed into more complex coalescent-based models that we are still developing. These models will account for and have at least two selection coefficients (within and between-host). Therefore, mlscluster narrows down our search by making us

start from the sites that are more likely to be TFPs and not considering every single site in the genome, which would probably be prohibitive. We are currently working on such methods and simulations to compare the accuracy of mlscluster vs these coalescent-based models, and will present such benchmarks in future work. We agree that the small absolute frequency of reoccurrence of most identified TFPs makes the interpretation weaker. But given that comprehensive analyses of SARS-CoV-2 fitness effects (doi: [10.1093/ve/veae026](https://doi.org/10.1093/ve/veae026)) point to a nearly neutral effect of synonymous mutations along the SARS-CoV-2 genome, we believe that our FDR of ~10% (that uses TFP detection in synonymous sites as a proxy for false detection) for threshold = 2% is reasonable enough for these sites to be considered under putative multilevel selection. These, in turn, will need further coalescent-based modelling or experimental studies to confirm the existence of multilevel selection effects. We included such explanation at the end of the 'Discussion'. Location of change in manuscript: R3#1 (Abstract, Plain Language Summary, Introduction, Discussion, Conclusions)

**2. Which target node conditions were used for the SARS-CoV-2 tree? How are the choices justified and how sensitive are the results to changes in these parameters? I do see in the code that the default values are `min_descendants = 10`, `max_descendants = 20*10^3`, `min_cluster_age_yrs = 1/12`, `min_date = max_date = NULL`, but I cannot find information regarding it in the manuscript.**

Response: The default values of (i) 10, (ii)  $20 \times 10^3$ , and (iii) 1/12 (1 month) were chosen to avoid ratios being taken from small or unreliable clade sizes/persistence times, an unrealistically high number of viral generations, and very short timeframes. These values were selected based on previous experience with UK SARS-COV-2 analyses using a similar software developed by our research group (doi: [10.1016/j.ebiom.2023.104939](https://doi.org/10.1016/j.ebiom.2023.104939)).

Unfortunately, a systematic evaluation of how these parameters would influence the results was not performed, but is planned for future applications of this method. We have included such clarifications in the revised manuscript. Location of change in manuscript: R3#2

**3. Why is caution only recommended starting from thresholds associated with an FDR of around 40%?**

Response: We agree that this statement was not rigorous enough and changed it accordingly to recommend caution mainly starting at threshold = 5% that has an FDR ~ 20%. Location of change in manuscript: R3#3

**4. If sister clades both qualify as target nodes, is the same (but reversed) comparison done twice and counted in the empirical distribution?**

Response: Yes, reversed comparisons between sister clades are included in our empirical distribution without any controls to avoid counting them twice. While it is true that the ratios are reciprocals and a log-transformation could be used to make them symmetric, our objective is to aggregate transmission fitness for all target nodes (and their resulting defining mutations) against their sisters. This means we are interested in capturing all comparisons, regardless of whether they are reversed. Excluding one of the two ratios or transforming them would not align with our goal of aggregating of transmission fitness. Therefore, we include both ratios to ensure that all potential relationships are considered in our analysis.

**5. Are some parts of the tree counted multiple times, e.g. if a parent and child node both qualify as target nodes with the same defining mutation?**

Response: A child clade will have the mutations of the parent and their own. We only consider a defining mutation if it does happen in the target clade in a percentage > 75% across its tips while occurring in < 75% of the tips in the sister clade. Therefore, since the target and sister will carry the mutations of their parent, extracting the difference will remove such mutations and only keep the ones that are really defining, therefore not counting the parent mutations over and over. One of the reasons we do the alignment-aware artifact removal is because some tips in a clade could not have some mutations found in their parent called due to lack of sequencing coverage at that position, and that could make us call mutations that are not actually defining multiple times.

**6. As demographic factors will still influence the clade characteristics, even in the presence of multilevel selection, how many independent samples are necessary, i.e. how often would the mutations have to be observed in the tree, to draw statistically reliable conclusions (see also Supplementary Figure S14 in van Drop et al. 2020)? What is empirically observed for the reported homoplasies?**

Response: Sample density will certainly impact the sensitivity of the method to detect circulating TFPs. By definition, these variants will have low persistence and growth, and therefore will only be detectable when a sufficiently high sample density is achieved. A detailed (quantitative) understanding of detection thresholds will need to be carried out in future work, but in the current paper we have added some discussion of these issues. Location of change in manuscript: R3#6

**7. Do the identified homoplasies also appear in parts of the tree that are not characterised as imbalanced?**

Response: Thank you for such a great question. Yes, I compared the frequencies of all non-synonymous homoplasies found in the tree (n=4692) against the detected TFP-homoplasies (cluster threshold=2%, n=543). This means that only 11.57% of all non-synonymous homoplasies are also detected as TFP-homoplasies. The median difference in frequency for homoplasies detected at both (as general homoplasy [i.e. non-TFP] and as TFP) is 4, mean=8.52, and IQR=13. Non-synonymous TFP-homoplasy median frequency is 2, mean=2.45, and IQR=0.5, while non-synonymous general homoplasy median frequency is 3, mean=6.73, and IQR=4. So this preliminary investigation suggests that the identified homoplasies also appear in parts of the tree not characterised as imbalanced (in this case above the 2% cluster threshold) roughly a median of 4 times, but the variation can be quite large. We will not include such results in the manuscript, but will think of ways of formally quantifying such differences in future work.

**8. How are multiple defining mutations in one clade handled?**

Response: The defining mutations of each clade are computed and the associated statistics (ratio of sizes, ratio of persistence time, and logistic growth rate) of the clade are attached to them, regardless of the clade having none (polytomy), one or multiple defining mutations. In the final data.frame, the mutations are 'individualised' to find the ones that happen independently in different parts of the tree that fall below the specified cluster threshold.



**9. I would recommend expanding the Discussion in the following two points: 9.1. More thorough discussion of the limitations that are already listed (how does the method deal with the challenges described in the second paragraph of the Discussion) and additional ones that apply (e.g., phylogenetic uncertainty and reliable tree inference as potentially time-consuming preprocessing step, statistical power from only few recurrences)**

Response: Thank you very much for your recommendation. We expanded the existing discussion to address the limitations already listed, although sampling biases is more thoroughly discussed by the end of the discussion and changes in immunological landscape were not considered. The additional points were also incorporated into the discussion. Location of change in manuscript: R3#9.1

**9.2. Explanation for and discussion of the pronounced differences observed between the mlscluster and HyPhy results.**

Response: Thank you. This was also pointed out by Reviewer#2. The minimal overlap suggests that these methods are capturing different features / selective pressures, as expected. We included this clarification in the 'results' section. Location of change in manuscript: R3#9.1

**10. I would recommend a more careful phrasing of the possible explanation of the observed TFP-homplasy pattern by accelerated within-host rates, as I see no ground justifying a connection presented in the study and too few references showing evidence of the connection and rate acceleration (see also preprint [1]).**

Response: Thank you very much for the recommendation. We agree this point deserves further discussion and a more careful phrasing, and included a statement in the discussion citing the suggested paper, which indeed does a great systematic investigation of the within-host evolutionary rates in chronically infected SARS-CoV-2 patients. Location of change in manuscript: R3#10

### **Minor comments**

**Abstract: 11.1. "We applied this method for a SARS-CoV-2 time-calibrated phylogeny.." --> "We applied this method to a SARS-CoV-2 time-calibrated phylogeny.."**

Response: Fixed. Location of change in manuscript: R3#11.1

**11.2. To my understanding, the conclusion statement is not supported by the study results (see also major comment 1), as they do not allow to draw direct conclusions about within-host vs. between-host evolution. I would recommend rephrasing this.**

Response: Thank you very much, we agree with the recommendation and rephrased it to make it clear that these sites are not mechanistically (using e.g. realistic simulations) proven to be under multilevel selection. Location of change in manuscript: R3#11.2 **Plain**

### **Language**

**Summary: 12.1. I would recommend using "clade" instead of "lineage" for consistency.**

Response: Thank you for the suggestion. We agree using "clade" is better for consistency and adjusted accordingly. We also rephrased the following to avoid repetition and provide additional clarification: "growth rates of clades carrying a specific mutation in comparison

with their immediate sisters without the mutation". Location of change in manuscript: R3#12.1

**12.2. "...highlighting the existence of important tradeoffs in selection between intrahost replication and inter-host transmission": I would recommend rephrasing this or providing evidence that the identified homoplasies are necessarily the product of multilevel selection (see major comment 1)**

Response: Similar to 11.2. Location of change in manuscript: R3#12.2

**Introduction: 13.1. A literature reference is missing for the statement "In molecular epidemiological studies, a set of particularly scalable approaches have been developed based on the calculation of phylogenetic clusters comprising two or more closely related samples." Additionally, I would suggest replacing "calculation" with, for example, "identification", "detection" or similar.**

Response: Thank you for noticing this. We included a reference to the most widely used distance-based method (HIV-TRACE) and to a paper that evaluates the limitations of this and other similar methods. We also replaced "calculation" with "detection". Location of change in manuscript: R3#13.1

**13.2. Either a literature reference or a clearer argument is missing for the statement "Furthermore, there is considerable scope to improve on distance-based genetic clustering methods because such approaches will potentially have poor specificity for variants that negatively influence fitness."**

Response: We included a clarification on why these approaches are expected to present poor specificity for variants that negatively influence fitness. They were demonstrated to be systematically biased to detect variation in sampling rates instead of transmission rates. Location of change in manuscript: R3#13.2

**13.3. In the sentence following the one cited in the previous point the reference to SARS-CoV-2 is duplicated.**

Response: We are not sure we understand this point. Actually, in this sentence: "During the past few years, positive and negative selection in SARS-CoV-2 have mainly been investigated using methods that rely on synonymous rate variation across sites/branches (26, 27) , and results from these approaches on SARS-CoV-2 comprehensive datasets are available for comparison (28)" reference 27 refers to the HyPhy method description paper and reference 28 to one of its applications to SARS-CoV-2 by the same author.

**13.4. A formal definition of "transmission fitness polymorphism" is missing.**

Response: We agree it was not defined clearly, so we included a definition in the 'introduction' section. Location of change in manuscript: R3#13.4

**13.5. "We demonstrated its applicability through the analysis of a representative >1.2 million SARS-CoV-2 genomic data set from England..." should, e.g., rather be phrased as "We demonstrated its applicability through the analysis of a representative SARS-CoV-2 data set comprising >1.2 million genome sequences from England..."**

Response: This sentence has been changed accordingly. Location of change in manuscript: R3#13.5

**13.6. “By providing a comprehensive catalog of the main sites driving multilevel selective pressures throughout the SARS-CoV-2 genome, we also expand the understanding of [the] SARS-CoV-2 fitness landscape outside the well-studied spike protein”. I would recommend rephrasing of this sentence, as (i) the identified sites are not the drivers of the selective pressures, rather the result of them and (ii) see major comment 1.**

Response: We have rephrased it as "potentially resulting from multilevel selective pressures". Location of change in manuscript: R3#13.6

**Methods: 14.1. Usage of “node” and “clade” is inconsistent. I would recommend primarily using only one of the two terms and, if necessary, write “node” only to refer to the actual internal node and “clade” when referring to the whole subtree arising from said “node”.**

Response: We agree using only one of the terms avoids confusion, and therefore chose "clade". Then, all instances of "node" were replaced by "clade" and some minor edits were made to avoid repetition of "clade" too many times. Location of change in manuscript: R3#14.1 (in the first replacement) and throughout the manuscript

**14.2. “Assume two clades u (target clade) and v (comparator/sister) organised in a time-scaled tree t and sharing ancestry (i.e. the same defining mutations).” --> “Assume two clades u (target clade) and v (comparator/sister) organised in a time-scaled tree t and sharing full ancestry (i.e. the same defining mutations).”**

Response: I have added "full" as suggested, and also added the clarification that the ancestry is up to the point they diverge and therefore present their own exclusive defining mutations. Location of change in manuscript: R3#14.2

**14.3. “The persistence time (given by a in Figure 2), is defined as..” --> “The persistence time (given by a in Figure 2), is defined as..” (closing bracket at wrong position)**

Response: Fixed. Location of change in manuscript: R3#14.3

**14.4. “...which can be their sister clade (the clade sharing an immediate ancestor assuming bifurcating phylogenetic relationship) or against all other clades...” --> “...which can be their sister clade (the clade sharing an immediate ancestor assuming bifurcating phylogenetic relationship) or all other clades...”**

Response: Fixed. Location of change in manuscript: R3#14.4

**14.5. Unclear if ‘clade size’ refers to  $S_{uv}$  or  $n_u$**

Response: Thanks for noticing this. It referred to  $n_u$  and  $n_v$  and not to the ratio  $S_{uv}$ . We tried to rephrase this to ensure the meaning is not dubious. Location of change in manuscript: R3#14.5

**14.6. I would recommend explicitly stating that in this study it is assumed that a homoplasy that arises in a subtree for which the statistics fall below a low quantile of the empirical distribution (see sentence “... for identifying especially convergently acquired mutations (homoplasies) that are detrimental for transmission (within a low quantile of the probability distribution of at least one of the three statistics”).**

Response: We agree that the suggested change clarify the analytic methods used, making it easier to understand and technically more accurate. It is now incorporated into the manuscript. Location of change in manuscript: R3#14.6

**14.7. The observed distribution of each of the three statistics in the tree is formally not a probability distribution. I would recommend replacing this terminology.**

Response: We agree and have fixed this throughout the manuscript. Location of change in manuscript: R3#14.7

**14.8. “Given this efficient way to visit nodes of the tree and edge lengths, we can easily extract the parameters of interest (e.g. the time of the most recent common ancestor of each node, ...” --> “Given this efficient way to visit nodes of the tree and edge, we can easily extract the parameters of interest (e.g. the time of the most recent common ancestor of each clade, ...”**

Response: Fixed. Location of change in manuscript: R3#14.8

**14.9. “Target nodes are extracted based on the following conditions ...” The following needs rephrasing, as it currently reads as if only the node with the smallest number of descendants etc. is kept.**

Response: Thank you very much for noticing this. We have rephrased accordingly and believe the sentence implies the intended meaning now. Location of change in manuscript: R3#14.9

**14.10. Definition of “sharing ancestry” as same defining mutations between sister clades and “defining mutation” in one sister clade is contradicting.**

Response: We are not sure we completely understand this statement, but we believe this was accordingly addressed together with 14.2 above. The main idea is to compare clades carrying a specific defining mutation against their immediate sisters without the mutation. Location of change in manuscript: R3#14.10

**14.11. Which homoplasy annotations are used later in the text?**

Response: Homoplasies are annotated in the text and figures using the standard protein:{ancestral\_aminoacid}{aminoacid\_coordinate}{mutated\_aminoacid} notation to make them consistent and comparable with other studies. The specific annotations (e.g. regions of interest including RBD and NTD of spike) are mentioned when appropriate to illustrate important results, in table 1 ('Genomic region' column), and in the detailed CSV outputs of the package.

**14.12. “We intended to make the method flexible by creating a parameter that specifies in how many percentiles the statistic should be splitted...”: Isn’t it rather a parameter that specifies which percentile to use (e.g. 2%)?**

Response: 'quantile\_choice' parameter specifies in how many percentiles the statistics should be splitted. For example, 'quantile\_choice=1/100' (default) splits from 1% to 100% in intervals of 1%, 1/400 splits from 0.25% to 100% in intervals of 0.25%. The later allows more strict "cluster thresholds" to be used (e.g. 0.25%). Each statistic has its own parameter to set the percentile to use ('quantile\_threshold\_ratio\_sizes', 'quantile\_threshold\_ratio\_persist\_time', and 'quantile\_threshold\_logit\_growth'). Again, this

option just makes the method more flexible, but it is recommended to consider the same threshold for all three statistics (which was the only scenario we tested). These options are properly documented in the R package, available using: `help("run_diff_thresholds")`.

**14.13. I find the terminology “cluster threshold” confusing, as the output of the method is not primarily a clustering of parts of the tree into groups, but rather the identification of recurring mutations based on relative differences between any sister clades (passing the filtering conditions).**

Response: Thank you very much for raising this terminology opinion. We respect it and agree it is a bit confusing, but preferred to use this instead of more technical terms such as "quantile threshold", "empirical distribution threshold", etc. It was a topic of discussion during the paper preparation and both of us agreed "cluster threshold" would not be a perfect name but better than other terms we could think of.

**14.14. A literature reference is missing for the ML tree inference.**

Response: Thank you for noticing this. The citation to FastTreeMP, UShER, and phylopipe is now incorporated into the revised manuscript. Location of change in manuscript: R3#14.14

**14.15. “We tested our approach using two COVID-19 pandemic time-periods: (i) from June 01, 2020 (including Wuhan/WH04/2020 reference sequence as root of the phylogeny) to November 15, 2022 (before Omicron BA.1.\* variant emergence)..” --> “We tested our approach using two COVID-19 pandemic time-periods: (i) from June 01, 2020 (including Wuhan/WH04/2020 reference sequence as root of the phylogeny) to November 15, 2021 (before Omicron BA.1.\* variant emergence)..”**

Response: Fixed. Location of change in manuscript: R3#14.15

**14.16. The sentence “For each period, 10 different thresholds (...) of the clustering statistics are computed” needs rephrasing, e.g. “For each period, 10 different thresholds (...) of the clustering statistics are considered”. I would also suggest adding “%” after each threshold value.**

Response: Fixed. Location of change in manuscript: R3#14.16

**14.17. Alignment-aware artifact removal: Is it correct that this is only necessary if occurrences of “X” and “N” at each genome position are not randomly distributed over the tree?**

Response: We think it should be performed regardless of whether their occurrence is randomly distributed over the tree or not. Since the method relies on extracting homoplasies along the tree, it will be highly prone to artifacts as demonstrated by our Omicron case scenario, in which we found a massive enrichment for TFPs (due to the known Omicron sequencing dropout issues) when not performing such sanity checks.

**14.18. Poisson regression: definition of i and j is missing**

Response: These are already defined. "(...) i polymorphic third codon position sites (...)" (count of polymorphic sites at third codon position) and "(...) j is the total number of polymorphic sites at first and second positions (...)".

**14.19. Definition of Y as TFP calls among the I polymorphic third codon position sites**

**with >100 mutated sequences: Why 100?**

Response: This was selected after retrieving the number of polymorphic synonymous sites at each codon position for a couple of thresholds of mutated sequences (ranging from 10 to 1000). N=100 represents ~0.1% of the ~1.7 million sequences analysed, so a threshold as high as 1000 or 500 was excluding too many legitimate polymorphic sites and a threshold as low as 10 or 50 could include artifacts or spurious polymorphic sites. We included a summarised version of this explanation in the revised manuscript. Location of change in manuscript: R3#14.19

**14.20. "Multiplication by 100 transforms the probability of erroneously calling a TFP into a percentage for easier interpretation": The FDR does not represent the probability of erroneously calling a TFP, I would recommend rephrasing this.**

Response: Thanks for noticing this. You are correct. We rephrased it as "proportion of erroneous TFP calls". Location of change in manuscript: R3#14.20

**14.21. Closing bracket missing: "(see Methods: Statistical analysis for identifying genomic regions enriched for TFP)."**

Response: Fixed. Location of change in manuscript: R3#14.21

**14.22. "This independent analysis provides additional evidence that their evolution is consistent with a transient selective pressure.": Why is this the case? To my understanding, the manual inspection of the relatively frequent TFP homoplasies mainly shows that the method worked as expected here (i.e. the identified homoplasies are recurrent and in 'small' clades)**

Response: Thank you for the suggestion. We agree and rephrased it to highlight only that it shows that the method works as expected and provided additional explanation ("indicated by the recurrence of such mutations and their appearance in clades where the size, longevity or growth rate of the target clade is much smaller when compared to its sister(s).") Location of change in manuscript: R3#14.22

**Discussion: 15.1. "We have quantified transient selective forces acting on SARS-CoV-2 lineages and mutations...": I would recommend rephrasing this, as the study results do not represent a quantification of the selective forces.**

Response: Agreed. We replaced "quantified" with "presented a tree-based clustering method to investigate". Location of change in manuscript: R3#15.1

**15.2. "After the emergence of VOCs with elevated substitution rates, other attributes...": I would recommend rephrasing this, since it reads as if VOCs have elevated evolutionary rates.**

Response: Thank you, we rephrased this for clarity. Location of change in manuscript: R3#15.2

**15.3. "Genetic diversity in an infected individual is goverened by repeated cycles..." --> "Genetic diversity in infected individuals is goverened by repeated cycles..."**

Response: Fixed. Location of change in manuscript: R3#15.3

**15.4. "Additionally, mutations identified in the top 100 most frequent TFP-homoplasies**

**[...] also emerged after days of chronic infection.”: Since they emerged within days, were they shown to only emerge in chronically infected individuals and not others?**

Response: The studies cited investigated their occurrence specifically in chronically infected patients, and to the best of our knowledge these mutations were not demonstrated to be particularly prevalent at the population level (acute infections) in several lineages throughout the pandemic.

**Conclusions: 16. “We developed a method capable of identifying sites under multilevel selection...”: I would recommend being more careful with the wording here, see major comment 1.”**

Response: Agreed. We replaced "capable" with "designed to identify candidate sites (...)". Location of change in manuscript: R3#16

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 11 April 2024

<https://doi.org/10.21956/wellcomeopenres.22912.r76365>

© 2024 Ghafari M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Mahan Ghafari** 

University of Oxford, Oxford, UK

Franceschi and Volz have developed a method for identifying mutations that have a transient selective advantage at the within-host level but are disadvantageous at the between-host level. Their method utilises three statistics based on the size, persistence time, and logistic growth rates of clades harbouring the target mutations. Their analysis of SARS-CoV-2 is an interesting one and their methodology appears sound. One external validation of their method is the identification of recurrent mutations found during chronic infections, including mutations outside Spike. I have a few comments:

**Precedence:**

The authors' claim of being the first to identify mutations with complex fitness trade-offs may benefit from further clarification and context regarding previous research in this area. However, I think their approach to identifying such mutations is particularly useful for large-scale datasets. For instance, the work by Harari et al., which the authors also cited, found trade-offs between immune evasion (and viral replication) and transmissibility. They showed that certain recurrent within-host mutations in chronic infections do not appear at the between-host level, pointing to the trade-off between immune escape and transmissibility as the possible explanation.

More recently, Ghafari et al. [1] demonstrated that certain mutations, such as T1638I in ORF1ab, are recurrent in persistent infections but are mildly deleterious at the between-host level, while

many other recurrent mutations in persistent infections are also highly advantageous at the between-host level.

**Methods:**

The assumption that artifacts ("X") would necessarily follow the majority allele at a given position is not immediately clear to me. Could the authors clarify what this assumption is based on or verify from base frequency files whether "X" is indeed the majority nucleotide in at least some cases?

A potential limitation of this approach is its focus on consensus sequences. A de novo mutation reaching over 50% frequency within a host takes time, and given the tight transmission bottleneck, minority variant transmission is unlikely. Thus, at least some beneficial within-host mutations might be missed.

It would be interesting to know if the authors have explored context-dependent selective advantage/disadvantage of TFPs, for example, whether certain mutations are only detrimental at the between-host level if they appear in one major lineage and not others.

**Results:**

It is still unclear to me whether the set cluster thresholds chosen for the analyses correspond to all three considered statistics, some of them, or at least one.

The significance of observing B.1.1.7 and AY.4.\* enriched for TFP-homoplasies is not well explained. Should this be interpreted as these lineages being more likely to give rise to beneficial within-host mutations that are deleterious at the between-host level? If so, the biological explanation for such lineage-dependent effects needs some clarification.

Given that the study period extends only until mid-2022, it would be worth clarifying whether any second-generation BA.2 lineages, as well as BA.5 or XBB, were included in the analysis or not for interested readers.

The discordance between sites identified as under positive selection using mlcluster and the HyPhy-based approach is noteworthy. The minimal overlap suggests these methods are capturing different features. Some comments/clarification helps here.

**Discussion:**

Bloom and Neher's [2] recent investigation into the between-host fitness effect of SARS-CoV-2 mutations across all major lineages pre- and post-Omicron uses the number of independent appearances of a mutation on a global phylogeny as a measure of mutation fitness effects (not cluster size or their persistence times). Their approach also seems to work particularly well in capturing deleterious or nearly neutral mutations at the between-host level. I would be curious to see some discussion on how the methods compare with the author's approach in this paper.

The discussion on how this approach could be used to investigate the association between chronic infection, duration of infection, and the presence of TFPs is an interesting one. Given Ghafari et al.'s [1] findings that many mutations during persistent infections are also beneficial at the between-host level, would the authors expect that their approach cannot capture mutations that are both beneficial at the within- and between-host level?



The authors have explained other caveats and advantages of their approach well in the context of existing literature.

**Minor Points:**

- TFP-homoplasies should be defined before being mentioned in the abstract.
- The terms Upper and Lower Tier Local Areas need definitions before their abbreviations are used.
- Consider changing "... up to July 2020 and has not considered persistence times" to "did not consider persistence times".
- There seems to be an accidental text break in the following paragraph "... ORF1ab:T1543I also emerged after days of chronic infection" which seems unnecessary.
- On figure 4, it would help to indicate which mutations belong to the intersection category (yellow). Are these S484 and S98?
- The sentence "...it shows that the impact of very few mutations outside the S protein have been characterised experimentally" could be rephrased for clarity. The results highlight potential functionally important mutations outside the Spike protein that warrant further investigation rather than directly showing lack of experimental characterisation.

**References**

1. Ghafari M, Hall M, Golubchik T, Ayoubkhani D, et al.: Prevalence of persistent SARS-CoV-2 in a large community surveillance study. *Nature*. 2024; **626** (8001): 1094-1101 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Bloom JD, Neher RA: Fitness effects of mutations to SARS-CoV-2 proteins. *Virus Evol*. 2023; **9** (2): vead055 [PubMed Abstract](#) | [Publisher Full Text](#)

**Is the work clearly and accurately presented and does it cite the current literature?**

Partly

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Virus evolution, phylogenetics, and epidemiology.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 15 Jul 2024

**Vinicius Bonetti Franceschi**

Thank you very much for your time to review the paper. We apologise for the delay in getting back to you. This was due to other concurring projects and our tentative to address the high number of comments raised by the three reviewers in the best possible way. We appreciate your detailed and interesting comments, and believe that they helped to improve the contextualisation of the paper. Please see below the response to the specific points:

**Precedence:**

**1. The authors' claim of being the first to identify mutations with complex fitness trade-offs may benefit from further clarification and context regarding previous research in this area. However, I think their approach to identifying such mutations is particularly useful for large-scale datasets. For instance, the work by Harari et al., which the authors also cited, found trade-offs between immune evasion (and viral replication) and transmissibility. They showed that certain recurrent within-host mutations in chronic infections do not appear at the between-host level, pointing to the trade-off between immune escape and transmissibility as the possible explanation.**

Response: Thank you very much for your comment. We agree the first version of the paper did an inadequate job of putting our results in the appropriate context of existing work. We also included the suggested discussion of the work by Harari that corroborates our findings. Location of change in manuscript: R2#1

**2. More recently, Ghafari et al. [1] demonstrated that certain mutations, such as T1638I in ORF1ab, are recurrent in persistent infections but are mildly deleterious at the between-host level, while many other recurrent mutations in persistent infections are also highly advantageous at the between-host level.**

Response: Thank you very much for suggesting this reference. We included a statement in the 'discussion' citing such mutations that appear to be deleterious at the between-host level and that the majority of recurrent mutations in persistent infections you found in this large community study tend to be beneficial at the between-host level. Location of change in manuscript: R2#2

**Methods:**

**3. The assumption that artifacts ("X") would necessarily follow the majority allele at a given position is not immediately clear to me. Could the authors clarify what this assumption is based on or verify from base frequency files whether "X" is indeed the majority nucleotide in at least some cases? A potential limitation of this approach is**

**its focus on consensus sequences. A de novo mutation reaching over 50% frequency within a host takes time, and given the tight transmission bottleneck, minority variant transmission is unlikely. Thus, at least some beneficial within-host mutations might be missed.**

Response: Thank you for pointing this out. We understand it was probably not accurately described before. What we actually did was to compute the proportion of the most frequent mutation at a given site across all sequences within the clade and add up the frequency of the "X" or "N" (undetermined bases) at that site (across all sequences in that same clade). This assumption is based on the complete shared ancestry across sequences in that clade. In cases where "X" and "N" are the most frequently mutated characters at the given clade, we discard them to avoid the inclusion of artifacts. We added such clarification in 'Methods: Statistical analysis for identifying genomic regions enriched for TFPs ' and hope this is clear now. We agree the focus on consensus sequences brings limitations such as the one cited, so we included it in the 'discussion' with an appropriate reference. Location of change in manuscript: R2#3 (methods and discussion)

**4. It would be interesting to know if the authors have explored context-dependent selective advantage/disadvantage of TFPs, for example, whether certain mutations are only detrimental at the between-host level if they appear in one major lineage and not others.**

Response: Thank you very much for such an interesting question. As pointed out by Reviewer 3, the statistical power to detect such relationships would be very limited given the lower frequency of TFP-homoplasies. We have not investigated these context-dependent relationships in depth because we think any conclusions from that would be highly speculative given these limitations. We included a statement in the 'discussion' (paragraph 3) mentioning this. We found some TFPs to occur in more than one major lineage and others to happen only in one, so probably the answer is yes, there are lineage-specific differences. Hopefully, analysis of genomic data from other viruses where multilevel selection is more common and some realistic coalescent-based simulations we are currently working on will help us to better understand these relationships. Location of change in manuscript: R2#4

### **Results:**

**5. It is still unclear to me whether the set cluster thresholds chosen for the analyses correspond to all three considered statistics, some of them, or at least one.**

Response: We agree this was not very clear. We incorporated this sentence: "homoplasies occurring in a subtree for which at least one of the three statistics fall below a low quantile [e. g., 2%] of the empirical distribution" into the 'Methods: Tree-based clustering algorithm implementation' section to clarify that the analysis is performed for homoplasies and associated clades detected by at least one of the three statistics. Location of change in manuscript: R2#5

**6. The significance of observing B.1.1.7 and AY.4.\* enriched for TFP-homoplasies is not well explained. Should this be interpreted as these lineages being more likely to give rise to beneficial within-host mutations that are deleterious at the between-host level? If so, the biological explanation for such lineage-dependent effects needs some**

**clarification.**

Response: We do not intend to make any claims about the biological propensity of these variants to generate TFPs, but merely to make the observation which may generate new hypotheses. It is possible that this observation is driven by surveillance effects, since there was relatively deep sampling of B.1.1.7 and AY.4 that may have enabled the detection of more candidate TFPs. Location of change in manuscript: R2#6

**7. Given that the study period extends only until mid-2022, it would be worth clarifying whether any second-generation BA.2 lineages, as well as BA.5 or XBB, were included in the analysis or not for interested readers.**

Response: We agree it is important to emphasise what lineages were considered. Unfortunately, after Pillar 2 termination in the UK (around April 2022), sampling was quite biased towards hospitalised cases. It is important to have a well-defined sampling frame over the duration of the study, so we chose not to consider sequences taken after this time (April 2022). We included a statement in the beginning of 'results' to clarify that such lineages (second-generation BA.2, BA.5, XBB, etc) were not included in the analysis. Location of change in manuscript: R2#7

**8. The discordance between sites identified as under positive selection using mslcluster and the HyPhy-based approach is noteworthy. The minimal overlap suggests these methods are capturing different features. Some comments/clarification helps here.**

Response: Thank you very much for pointing this out. We agree it is important to explicitly say that this minimal overlap suggests that these methods are capturing different selective pressures. We included this clarification in the 'results' section. Location of change in manuscript: R2#8

**Discussion:****9. Bloom and Neher's [2] recent investigation into the between-host fitness effect of SARS-CoV-2 mutations across all major lineages pre- and post-Omicron uses the number of independent appearances of a mutation on a global phylogeny as a measure of mutation fitness effects (not cluster size or their persistence times). Their approach also seems to work particularly well in capturing deleterious or nearly neutral mutations at the between-host level. I would be curious to see some discussion on how the methods compare with the author's approach in this paper.**

Response: Thank you very much for the interest in such comparison. As also suggested by Reviewer#1, we added the comparison of our identified TFPs against the fitness effects estimated by Bloom & Neher and some discussion on how the methods compare and why they would lead to different results. Location of change in manuscript: R2#9 (Methods, Results, and Discussion)

**10. The discussion on how this approach could be used to investigate the association between chronic infection, duration of infection, and the presence of TFPs is an interesting one. Given Ghafari et al.'s [1] findings that many mutations during persistent infections are also beneficial at the between-host level, would the authors expect that their approach cannot capture mutations that are both beneficial at the**

**within- and between-host level?**

Response: Since our approach is especially designed to detect beneficial mutations at the within-host level, we do not expect that it directly captures mutations that are, at the same time, favourable for between-host replication. We included such statement in the discussion. Location of change in manuscript: R2#10

**11. The authors have explained other caveats and advantages of their approach well in the context of existing literature.**

Response: Thank you for your positive feedback!

**Minor:****12. TFP-homoplasies should be defined before being mentioned in the abstract.**

Response: Thank you very much for noticing this. It is fixed. Location of change in manuscript: R2#12

**13. The terms Upper and Lower Tier Local Areas need definitions before their abbreviations are used.**

Response: Thank you very much. It is addressed. Location of change in manuscript: R2#13

**14. Consider changing "... up to July 2020 and has not considered persistence times" to "did not consider persistence times".**

Response: Fixed. Location of change in manuscript: R2#14

**15. There seems to be an accidental text break in the following paragraph "... ORF1ab:T1543I also emerged after days of chronic infection" which seems unnecessary.**

Response: Thank you very much for pointing this out. We have tried to fix it but there still seems to be a formatting issue there, so left a comment for the editorial team. Location of change in manuscript: R2#15

**16. On figure 4, it would help to indicate which mutations belong to the intersection category (yellow). Are these S484 and S98?**

Response: These are S:A67V and S:E484K. Unfortunately, the available HyPhy analysis only contained the identified sites and not the actual replacements under positive selection. Consequently, the actual mutations and further annotations are presented in Table 1. We added this clarification in the legend of Figure 4. We believe the yellow color is sufficient to distinguish these sites in the 'intersection' category and these sites are mentioned in the 'TFPs along the SARS-CoV-2 genome and low concordance'. Location of change in manuscript: R2#16

**17. The sentence "...it shows that the impact of very few mutations outside the S protein have been characterised experimentally" could be rephrased for clarity. The results highlight potential functionally important mutations outside the Spike protein that warrant further investigation rather than directly showing lack of experimental characterisation.**

Response: Thank you for your suggestion. We have rephrased the sentence for clarity. We

believe this revision more accurately conveys the intended meaning. Location of change in manuscript: R2#17

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 28 March 2024

<https://doi.org/10.21956/wellcomeopenres.22912.r76372>

© 2024 Neher R. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Richard Neher** 

University of Basel, Basel, Switzerland

Francheschi and Volz present a method to investigate the effects of specific mutations on viral fitness by comparing clades that carry a specific mutation to their immediate siblings without the mutation. The method relies on comparing the total size, the 'longevity', and the relative growth of the clades.

The paper is rather hard to read. It is, for example, not said explicitly that the method is specifically looking at homoplasies and aggregating information across many occurrences of the same mutation. Instead the discussion is kept general (target/comparator) when it would help a lot if the canonical scenario was explained. There are several other ways in which the readability of this paper could be improved.

The specific claim that the method allows to identify multi-level selection as opposed to simply aggregate transmission fitness is not well supported.

Specific points:

Introduction:

- the first paragraph is oddly neutralist. For SARS-CoV-2 or the HA segment of influenza A, adaptation is common and we witness a repeated pattern of variant replacement. While it might still be technically true that the 'majority' of polymorphisms is neutral or weakly selected, this introduction strikes me as strange. Furthermore, discussing transient fitness advantage as an exception to the neutral majority is particularly weird given the rest of the article is about SC2 with frequent replacements driven by adaptive evolution. A useful reference in this context could be Kistler and Bedford [1]

- The motivation with multi-level selection in HIV is a bit strange and does not have much to do with what is discussed in the manuscript. It is unclear that different within/between host evolutionary rates in HIV are directly related to multi-level selection through preferential

transmission of ancestral variants, or whether they are the result of frequent reversions of host-specific adaptations, for example T-cell escape. Influenza virus would be a much better model for the SC2 analysis done here than HIV. Strelkova and Laessig [2] might be useful context.

- the 2nd and 3rd paragraph of the introduction are rather cryptic.

- the idea of fitness inference from the shape of phylogenies was introduced much earlier, for example [3]

- the most comprehensive estimates of (mostly deleterious) fitness effects of mutations in SC2 to date were probably published in [4]

Methods:

- throughout the manuscript, it remains unclear whether identification of sites with fitness effects is the primary result, or enrichment of such sites in major variants or regions. This makes the manuscript hard to follow.

Results

- the paragraphs in the second column of page 9 summarize counts and observations without putting them into context and it is unclear what the reader is supposed to take away from this discussion.

- I believe the functional significance of ORF7 and 8 is still not very well understood.

Discussion:

- it remains unclear to me why the approach presented here specifically reveals 'multi-level selection'. The approach picks up signals of clade growth and thus aggregate transmission rate in the population. The fact that most mutations are rare isn't evidence for a within-host/between-host trade-off. This is a natural consequence of a very large densely sampled pathogen with rapid exponentially growing outbreaks.

- the results would be more convincing if they would be quantitatively compared to inferences from other analyses or deep mutational scanning data.

## References

1. Kistler KE, Bedford T: An atlas of continuous adaptive evolution in endemic human viruses. *Cell Host Microbe*. 2023; **31** (11): 1898-1909.e3 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Strelkova N, Lässig M: Clonal interference in the evolution of influenza. *Genetics*. 2012; **192** (2): 671-82 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Neher RA, Russell CA, Shraiman BI: Predicting evolution from the shape of genealogical trees. *Elife*. 2014; **3**. [PubMed Abstract](#) | [Publisher Full Text](#)
4. Bloom JD, Neher RA: Fitness effects of mutations to SARS-CoV-2 proteins. *Virus Evol*. 2023; **9** (2): vead055 [PubMed Abstract](#) | [Publisher Full Text](#)

**Is the work clearly and accurately presented and does it cite the current literature?**

No

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Viral evolution.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 15 Jul 2024

**Vinicius Bonetti Franceschi**

Thank you very much for your time to review the paper. We apologise for the delay in getting back to you. This was due to other concurring projects and our tentative to address the high number of comments raised by the three reviewers in the best possible way. Thank you especially for suggesting the comparison against DMS and computational fitness effects. We believe such comparisons made our discussion richer, especially by pointing out how the methods are different and why they lead to different results. Most importantly, we tried to make the wording less definitive about the sites actually being TFPs and explain that these will be used as input for more rigorous coalescent-based modelling in the near future to confirm/disconfirm their multilevel selection effects. Please see below the response to the specific points:

**1. Franceschi and Volz present a method to investigate the effects of specific mutations on viral fitness by comparing clades that carry a specific mutation to their immediate siblings without the mutation. The method relies on comparing the total size, the 'longevity', and the relative growth of the clades.**

Response: Thank you for this suggestion. We agree that this is indeed a good way of summarising the method. We incorporated "clades carrying a specific mutation in comparison with their immediate sisters without the mutation" into the "Plain Language Summary" to make it clearer. Location of change in manuscript: R1#1

**2. The paper is rather hard to read. It is, for example, not said explicitly that the method is specifically looking at homoplasies and aggregating information across**



**many occurrences of the same mutation. Instead the discussion is kept general (target/comparator) when it would help a lot if the canonical scenario was explained. There are several other ways in which the readability of this paper could be improved.**

Response: We appreciate your comment. We made it explicit now in the 'Introduction' that the method is looking at homoplasies and aggregating summary statistics across many occurrences of the same mutation and tried to explain the canonical scenario further. As it was not pointed out by the other reviewers as a major issue, we will not perform additional major changes to improve the readability of the paper, but of course will take this suggestion into consideration for future work. Location of change in manuscript: R1#2 (Introduction and Methods: Tree-based clustering algorithm implementation)

**3. The specific claim that the method allows to identify multi-level selection as opposed to simply aggregate transmission fitness is not well supported.**

Response: See point #12 for a more detailed response (as both comments 3 and 12 are quite similar).

#### **Introduction:**

**4. The first paragraph is oddly neutralist. For SARS-CoV-2 or the HA segment of influenza A, adaptation is common and we witness a repeated pattern of variant replacement. While it might still be technically true that the 'majority' of polymorphisms is neutral or weakly selected, this introduction strikes me as strange. Furthermore, discussing transient fitness advantage as an exception to the neutral majority is particularly weird given the rest of the article is about SC2 with frequent replacements driven by adaptive evolution. A useful reference in this context could be Kistler and Bedford [1]**

Response: There is of course a broad spectrum of selective effects and it was not our intention to imply that all variation is neutral, although it is certainly the case that the distribution of selective effects observed in circulating virus is concentrated strongly around zero. We have revised the discussion to be more balanced. Location of change in manuscript: R1#4

**5. The motivation with multi-level selection in HIV is a bit strange and does not have much to do with what is discussed in the manuscript. It is unclear that different within/between host evolutionary rates in HIV are directly related to multi-level selection through preferential transmission of ancestral variants, or whether they are the result of frequent reversions of host-specific adaptations, for example T-cell escape. Influenza virus would be a much better model for the SC2 analysis done here than HIV. Strelkova and Laessig [2] might be useful context.**

Response: We agree HIV-1 evolution is quite diverse when compared to SARS-CoV-2 or Influenza, but would like to keep HIV-1 as the most extreme (and probably best) example of documented multilevel selection. We are well aware that diverse mechanisms (variable immune selection, store-and-retrieve etc.) can generate recurrent evolutionary patterns in HIV-1, and HIV-1 is nevertheless a source of good examples of the effects we are looking for. We incorporated the potential mechanisms of HIV-1 multilevel selection as mentioned and appropriate references. We also included Influenza as the example of a virus that evolves more similarly to SARS-CoV-2 as suggested. Location of change in manuscript: R1#5

**6. The 2nd and 3rd paragraph of the introduction are rather cryptic.**

Response: We tried to improve the readability of these two paragraphs, and hope it is easier to follow now. Location of change in manuscript: R1#6

**7. The idea of fitness inference from the shape of phylogenies was introduced much earlier, for example [3]**

Response: Thank you very much for pointing this out. We were aware of the method, and it has been influential to our thinking, and we cited this paper in the revised version of the manuscript. Note however that the design and objectives of [3] are very different than our proposed method for identifying multi-level selection effects. Location of change in manuscript: R1#7

**8. The most comprehensive estimates of (mostly deleterious) fitness effects of mutations in SC2 to date were probably published in [4]**

Response: Thank you very much for bringing this to our attention. Note, however, that the objective of our analysis is somewhat different than those carried out by [4] and others. We removed the sentence that our paper was the first to do such an analysis and performed comparison against the other suggested methods (see especially points #12 and #13 for more details). Location of change in manuscript: R1#8

**Methods:****9. Throughout the manuscript, it remains unclear whether identification of sites with fitness effects is the primary result, or enrichment of such sites in major variants or regions. This makes the manuscript hard to follow."**

Response: Thank you very much for the feedback. The primary purpose of the paper is to characterise sites with potential multilevel fitness effects, but the analysis of enrichment in major lineages and genomic regions comes together to put those findings into context (i.e. the findings would not be realistic if not considering the context of lineage and genomic region the TFPs appear). So basically these findings are both quite important. The results section is separated in 'Lineages and genomic regions enriched with SARS-CoV-2 TFP-homoplasies' and 'TFPs along the SARS-CoV-2 genome and low concordance with positively selected sites' to provide this didactic separation among these result 'categories'.

**Results****10. The paragraphs in the second column of page 9 summarize counts and observations without putting them into context and it is unclear what the reader is supposed to take away from this discussion.**

Response: Thank you for the feedback. We agree some of the paragraphs needed contextualisation and added them when relevant (e.g. explanation of HyPhy minimal overlap, adding counts normalised per site instead of absolute counts, clearly stating genomic regions and major lineages in top-ranked TFP-homoplasies). Location of change in manuscript: R1#10

**11. I believe the functional significance of ORF7 and 8 is still not very well understood.**

Response: We agree and therefore rephrased the parts of the manuscript that made such a statement about ORF7a and ORF8 (Introduction and results section). Location of change in manuscript: R1#11

### **Discussion:**

**12. It remains unclear to me why the approach presented here specifically reveals 'multi-level selection'. The approach picks up signals of clade growth and thus aggregate transmission rate in the population. The fact that most mutations are rare isn't evidence for a within-host/between-host trade-off. This is a natural consequence of a very large densely sampled pathogen with rapid exponentially growing outbreaks.** Response: There seems to be a misunderstanding that the proposed methodology would only identify rare polymorphisms, the frequency of which is, of course, very sensitive to population dynamics. On the contrary, the scanning methodology is intended to shortlist all mutations that meet several criteria (recurrence, low persistence, low logistic growth). Population dynamics ("rapid exponentially growing outbreaks") is certainly an inadequate explanation for a variant meeting all three criteria. We tried to make it clearer now throughout the paper that the objective of mlscluster is to detect CANDIDATE sites under multilevel selection to feed into more complex coalescent-based models that we are still working on. These models will account for and have at least two selection coefficients (within and between-host). Therefore, mlscluster narrows down our search by making us start from the sites that are more likely to be TFPs and not considering every single site in the genome, which would probably be prohibitive. We believe that by considering homoplasies (independent occurrences of a mutation) in a subtree for which at least one of the three clade growth-derived statistics (ratio of sizes, ratio of persistence time, and logistic growth rate) fall below a low quantile (2%) of the empirical distribution, we are not just aggregating transmission rate in the population, but picking up these candidate TFPs that are detrimental to transmission because they lead, multiple times, to fewer offspring, shorter lifespan, or smaller growth rate. Note as well that we have characterised false discovery rates using 3rd codon position, and it is unlikely given the ~10% FDR that most of these outcomes are observed by chance. We believe that a greater certainty will be reached when we provide quantitative estimates of multilevel selection using more rigorous coalescent-based modelling in the near future. Location of change in manuscript: R1#12 (Abstract, Plain Language Summary, Introduction, Discussion, Conclusions)

**13. The results would be more convincing if they would be quantitatively compared to inferences from other analyses or deep mutational scanning data.**

Response: Thank you very much for the suggestion, although please note that these methods are estimating fundamentally different quantities. We compared our approach against DMS results and comprehensive Bloom & Neher fitness estimates (from point 8). Our main conclusion is that the sites we identified under multilevel selection (TFPs) tend to have fitness effects that skew towards zero. First, it is important to clarify that our method is identifying sites potentially under multilevel selection, while your approach (Bloom & Neher) is quantifying the magnitude of one level of selection. A single parameter is never sufficient to describe a site under multiple levels of selection. Future coalescent-based models we are developing — and will use as input the candidate TFPs given by mlscluster —, will incorporate two selection coefficients (between and within-host) for each site. We believe

these findings are the result of the fact that your approach assumes fixed (unchanged) effects. It would, therefore, average out these actual slightly positive and negative multilevel effects because only considering one level of selection. Location of change in manuscript: R1#13 (Methods, Results, and Discussion)

**Competing Interests:** No competing interests were disclosed.

---