Within-host genetic diversity of SARS-CoV-2 across animal species

Authors: Sana Naderi[1], Selena M. Sagan[1,2]*, B. Jesse Shapiro[1,3]*

1. Department of Microbiology & Immunology, McGill University, Montreal, QC, Canada
2. Department of Microbiology & Immunology, The University of British Columbia, Vancouver, BC, Canada
3. McGill Genome Centre, Montreal, QC, Canada

*for correspondence: jesse.shapiro@mcgill.ca; selena.sagan@ubc.ca

Abstract

Infectious disease transmission to different host species makes eradication very challenging and expands the diversity of evolutionary trajectories taken by the pathogen. Since the beginning of the ongoing COVID-19 pandemic, SARS-CoV-2 has been transmitted from humans to many different animal species, and viral variants of concern could potentially evolve in a non-human animal. Previously, using available whole genome consensus sequences of SARS-CoV-2 from four commonly sampled animals (mink, deer, cat, and dog) we inferred similar numbers of transmission events from humans to each animal species but a relatively high number of transmission events from mink back to humans (Naderi et al., 2023). Using a genome-wide association study (GWAS), we identified 26 single nucleotide variants (SNVs) that tend to occur in deer – more than any other animal – suggesting a high rate of viral adaptation to deer. Here we quantify intra-host SARS-CoV-2 across animal species and show that deer harbor more intra-host SNVs (iSNVs) than other animals, providing a larger pool of genetic diversity for natural selection to act upon. Within-host diversity is particularly high in deer lymph nodes compared to nasopharyngeal samples, suggesting tissue-specific differences in viral population sizes or selective pressures. Neither mixed infections involving more than one viral lineage nor large changes in the strength of selection are likely to explain the higher intra-host diversity within deer. Rather, deer are more likely to contain larger viral population sizes, to be infected for longer periods of time, or to be systematically sampled at later stages of infections. Combined with extensive deer-to-deer transmission, the high levels of within-deer viral diversity help explain the apparent rapid adaptation of SARS-CoV-2 to deer.

Introduction

Viral genetic diversity, at the scale of an outbreak or pandemic, ultimately originates within individual hosts as all mutations that spread between hosts must arise from the population of replicating viruses within an infected individual. Natural selection can act on this genetic diversity both within an individual infection and upon transmission to a new host. In acute infections, selection may have limited time to act or have limited genetic diversity to act upon (Lythgoe et al., 2021). Therefore the strongest selective filter may occur upon transmission to new hosts (Xue et al., 2020, N'Guessan et al., 2023). In some cases, such as chronically infected patients, viral lineages that are selectively favored within a host can be transmitted to other hosts (Gonzalez-Reiche et al., 2023). As a result, within-host viral diversity may foreshadow the success of viral mutations and lineages at a global scale (Xue et al., 2017, Harari et al., 2022).

Since the beginning of the COVID-19 pandemic, SARS-CoV-2 has been transmitted from humans to many different animal species. Persistence of the virus in animal reservoirs can open new evolutionary trajectories and allow for the selection of novel mutations. This highlights the need to study within-host diversity of SARS-CoV-2 in animal species, as a complement to the recent analyses of consensus-level genomic diversity (Naderi et al., 2023, Tan et al., 2022, Pickering et al., 2022). Consensus sequences represent the 'average' or 'majority rule' genome from a potentially genetically diverse population of viruses within a host. The consensus sequence provides a proxy for the genetic variation most likely to have been transmitted from host to host, while ignoring low-frequency mutations more likely to have arisen within a host (Andersen, Shapiro et al., 2015).

Several studies have quantified and analyzed the level of intra-host diversity of SARS-CoV-2 in humans. Intra-host diversity of SARS-CoV-2 is generally low in acute human infections (Valesano et al., 2021, Lythgoe et al., 2021, Braun and Moreno et al., 2021), but much higher in chronic infections, during which mutations can be selected within a host and potentially transmitted to new hosts (Gonzalez-Reiche et al., 2023). In acute infections, the median number of intrahost single nucleotide variants (iSNVs) at measurable frequency has been estimated to be between one and eight (Lythgoe et al., 2021, Valesano et al., 2021), with most samples having fewer than three (Lythgoe et al., 2021) iSNVs. The expected number of nucleotide differences between any two genomes within an average human infection has been estimated to be 0.83 (Tonkin-Hill et al., 2021). A measure of selective constraint at the protein level, the ratio of

nonsynonymous to synonymous substitutions rates (dN/dS), was similar between consensus sequences (transmitted between hosts) and moderate frequency iSNVs, suggesting similar levels of selection acting within and between hosts (Tonkin-Hill et al., 2021). By contrast, chronic infections, often in immunocompromised individuals (Avanzato et al., 2020, Choi et al., 2020), allow both the time and modest selective pressure (an immune system that is weak enough to allow prolonged viral replication, but strong enough to impose selection for immune evasion) to favor adaptive evolution (Clark et al., 2021, Choi et al., 2020). Many of the mutations selected within chronic SARS-CoV-2 infections are identical to those present in globally circulating variants of concern, strongly suggesting that successfully transmitting variants might arise from chronic infections (Harari et al., 2022).

While previous studies quantified within-host diversity of SARS-CoV-2 in humans, the diversity within animal hosts remains unexplored. In this study, we analyzed short-read sequencing data from SARS-CoV-2 sampled from cats, dogs, mink, and deer available on NCBI (National Library of Medicine (US)) as of September 2022 (Naderi et al., 2023). For comparison, we matched the animal-derived sequences to an equivalent number of human-derived sequences by their geographical location and time of sample collection. We found that deer have higher intra-host diversity than other animals, consistent with longer infections and potentially providing a larger pool of genetic diversity upon which natural selection can act.

Methods

Data. Raw Illumina sequencing reads from SARS-CoV-2 whole-genomes were downloaded from the NCBI Short Read Archive (National Library of Medicine (US)) for cat, dog, mink, and deer hosts as of September 2022. Samples with incomplete metadata were discarded. The dataset consists of 16 samples from cats, 10 from dogs, 66 from minks, and 81 from deer. Most of these sequences came from natural infections, with the exception of six dog samples and 12 cat samples derived from laboratory infection experiments. We searched the SRA database in June 2023 for human-derived short read sequences collected in the same geographical location as each group of animal host sequences, and matched them to animal host sequences based on their month of collection. For 65 out of the 173 starting animal samples, human matches fell within the same month, for 45 the human matches were 1-2 months away, for 46, 3-7 months away, and for three samples 10 months away. Among the deer samples, 14 sequences from Ontario had best matches to human-derived samples with incomplete data (no paired-end

reads) that were therefore excluded. We proceeded with 60 human-derived samples (similar to the average sample size from other species), for a total of 233 samples across all species. The SRA identifiers for these samples and their respective metadata are listed in **Supplementary Table 1**.

Read mapping and variant calling. Reads in FASTQ format were mapped to the Wuhan-Hu-1 reference genome (MN908947.3) using the GenPipes 4.4.5 CoVSeq pipeline (Bourgey et al., 2019). Within this pipeline, intrahost single nucleotide variants (iSNVs) are called with Freebayes 1.3.4 (Garrison and Marth et al., 2012), generating Variant Call Format (VCF) files that were used for subsequent analyses. In all subsequent analyses, only samples with a minimum average depth of 50x, and those with at least 15,000 nucleotide sites with minimum of 50x coverage were retained.

Calculating population genetic metrics. From each VCF file, we calculated the total number of iSNVs in the sample and the nucleotide diversity (the average number of pairwise nucleotide differences, π). To minimize the impact of sequencing errors, nucleotide positions covered by fewer than 50 mapped reads, or had a within-sample minor allele frequency of less than 5% were excluded from the calculations. We calculated π, the average number of pairwise differences, as follows:

$$\pi = \sum_{i=1}^{n} p_i(1 - p_i) \qquad \text{Equation 1}$$

Where *n* is the number of SNVs and p_i is the frequency of an SNV at position *i*, following Tonkin Hill et al., 2021. Corresponding values for each sample are listed in **Supplementary Table 2**.

Richness and Shannon diversity. To assess within-host lineage diversity, we used the demixing algorithm implemented in Freyja (Karthikeyan et al., 2022) to infer mixtures of SARS-CoV-2 lineages assigned based on the Pango nomenclature, and their relative abundances in each sample, with *richness* defined as the number of unique lineages detected in each individual. Lab-infected samples are known to have been infected experimentally with only one lineage, allowing us to establish a minimum lineage frequency threshold below which only false positive lineage calls are expected. We found that all 18 lab-infected samples were inferred by Freyja to contain only one lineage as long as we included lineages at a frequency of 0.025 or

more; therefore we used this minimum frequency threshold to remove likely false-positive lineage calls from natural infection samples. We then calculated the Shannon diversity index, *H,* as follows:

$$H = -\sum p_i.log(p_i) \qquad \text{Equation 2}$$

Where p_i is the relative abundance of Pango lineage *i* within a sample. Corresponding values for each sample are listed in **Supplementary Table 2**.

Generalized Linear Mixed Models. All regression models were fitted using the *glmer* function in the *lme4* package (Bates D et al., 2015) in R (R Core Team, 2023).

The variables in the models were: host species (with 'human' set as the reference level), type of infection (natural vs. lab infection, with natural set as the reference), tissue isolation source (lymph node v.s nasopharyngeal samples, with nasopharyngeal set as reference), sample mean depth of coverage, sample breadth of coverage (number of bases with at least 50x coverage), and assay type (whole genome shotgun vs. amplicon sequencing).

Assuming richness and the number of iSNVs follow a Poisson distribution, we fitted a Poisson regression model for these two variables. The number of maximum iterations for all models was set to 2 x 10^5. Initially, for both richness and the number of iSNVs, host species, type of infection, and tissue isolation source were considered fixed effects as they were variables of interest; depth and breadth of coverage, and assay type were considered random effects. In the model for the number of iSNVs, this arrangement of variables yielded a near singular fit warning, which was resolved by removing assay type from the random effects. This modified arrangement had the optimum Akaike information criterion (AIC) compared to other models (**Supplementary Table 3-1**). The initial arrangement of variables in the model for richness also resulted in a near singular fit warning. The near singularity issue was resolved by setting breadth of coverage to a fixed effect, yielding a model with the optimum AIC compared to other arrangements (**Supplementary Table 3-2**). Therefore, taking AIC and singularity of fit into account, the final model for iSNVs included host species, type of infection, and tissue isolation source as fixed effects, and breadth and depth of coverage as random effects. For richness, the final model included host species, type of infection, tissue isolation source, and breadth of coverage as fixed effects, and depth of coverage and assay type as random effects.

Wilcoxon rank-sum test for the differences in π and Shannon diversity. We could not identify a distribution that provided a good fit to the data for the average number of pairwise differences (π) and Shannon diversity index; therefore we did not apply GLMMs to these data. Instead, we used a non-parametric test, the Wilcoxon rank-sum test, to test for differences in π and the Shannon diversity index between humans and each animal species. For comparison, and also to validate the GLMMs described above, the Wilcoxon test was also performed on the number of iSNVs and richness data. The test was run using the *wilcox.test* function in the R stats package (R Core Team, 2023).

pN/pS. We used the genetic distance-based method by Nei and Gojobori (Nei, M., and T. Gojobori. 1986) to calculate pN/pS, the ratio of nonsynonymous to synonymous polymorphism rates, within each sample. Other than the rare cases of mixed infections (detected by Freyja above), all sequences within a sample share a common ancestor at the time of infection (Lythgoe et al., 2021). Over this relatively short evolutionary time, we considered it rare for a codon to contain more than one mutation event. For each sample, we calculated pN/pS as follows:

$$\frac{pN}{pS} = \frac{\frac{N_{nsub}}{N_{nsites}}}{\frac{N_{ssub}}{N_{ssites}}} \qquad \text{Equation 3}$$

Where N_{nsub} and N_{ssub} are the number of observed nonsynonymous and synonymous iSNVs within the sample, N_{nsites} and N_{ssites} are the number of nonsynonymous and synonymous sites, respectively.

In all calculations, the Wuhan-Hu-1 genome (MN908947.3) was used as reference. pN/pS was calculated both genome-wide, and for the S gene alone. We used the Wilcoxon rank sum test to compare pN/pS values genome-wide and in the S gene for each animal species versus humans.

Detecting animal-specific mutations within-host. Previously, we used a genome-wide association study (GWAS) to identify mutations in the SARS-CoV-2 genome associated with different animal species compared to humans. For each of the previously identified 26 deer-associated and three mink-associated mutations (Naderi et al., 2023) we counted the number of times they appear as “fixed” or “ambiguous” (iSNVs) in the “consensus tag” column

of the VCF files generated from deer and mink samples. We then counted the same quantities for every site in the SARS-CoV-2 that was not associated with either of these animals (non-GWAS hits). The categorization was done based on the Consensus Tag for each mutation in the VCF file generated by Freebayes (Garrison and Marth et al., 2012). The counts for each animal were tallied to compare GWAS and non-GWAS sites. The detailed breakdown of the counts for each GWAS mutation is listed in **Supplementary Table 4**.

Results

Deer contain relatively high within-host SARS-CoV-2 diversity

To compare within-host diversity of SARS-CoV-2 across animal species, we first counted the number of iSNVs within samples taken from cats, dogs, deer, mink, or humans. The median number of two iSNVS within human samples in our dataset was similar to previous estimates from acute human infections (Lythgoe et al., 2021) and also similar to mink (**Figure 1a**). By contrast, deer had a significantly higher median of 18 iSNVs per sample (Wilcoxon rank sum test, $p < 0.001$) compared to humans, and this difference remains when comparing only nasopharyngeal samples ($p = 0.023$). To test the statistical significance of these differences while controlling for sequencing coverage and covariates such as lab or natural infection and the tissue that was sampled, we used generalized linear mixed models (GLMMs; Methods). Our goal was to determine how the number of iSNVs varied across animal species compared to humans as a reference. We fitted a model with species, type of infection, and tissue isolation source as fixed effects, and depth and breadth of coverage as random effects (**Supplementary Table 3-1**). Host species and sample isolation source were both significant effects in this model (**Table 1**). Deer and mink had significantly different numbers of iSNVs compared to humans, with a rate ratio of 1.78 and 0.58, respectively, indicating that the number of iSNVs is 78% higher in deer, and 42% lower in mink than in humans. The low diversity in mink may be driven by an excess of samples with no detectable within-host diversity (**Figure 1a**), possibly due to sampling very early in the course of infection, as might occur during prospective sampling of an outbreak.

Lymph node tissue also had a significant effect, with a rate ratio of 3.68, indicating that the number of iSNVs is 2.68 times higher in lymph node tissue compared to nasopharyngeal samples (**Table 1**). Lymph node tissue was only sampled from deer, which partly explains the high number of iSNVs observed in deer (**Figure 1a**). However, the GLMM still identified a

significantly higher number of iSNVs in deer even after controlling for the effect of lymph node sampling and after removing lymph node tissue samples from the dataset entirely **(Table 1, Supplementary table 3-1)**. Deer and mink remained significant in the model applied to nasopharyngeal samples only, with the sign and relative magnitude of their effect size unchanged. This confirms that deer harbor more iSNVs than other animals, and mink fewer, even when only nasopharyngeal samples are considered.

	Model including lymph-node tissue samples			Model excluding lymph-node tissue samples		
Predictors (Fixed effects)	**Estimate**	**Rate ratio**	**P-value**	**Estimate**	**Rate ratio**	**P-value**
cat	-0.8100	NA	> 0.05	-0.7948	NA	> 0.05
dog	0.2910	NA	> 0.05	0.3623	NA	> 0.05
mink	-0.5416	0.5818	0.00405	-0.5590	0.5718	0.0140
deer	0.5810	1.7878	0.01109	0.6321	1.8815	0.0250
Isolation source (Reference: Nasopharyngeal swab)						
Lymph node tissue	1.3036	3.6825	1.95E-09	NA	NA	NA
Type of infection (Reference: Natural infection)						
Lab infected	0.8550	NA	> 0.05	0.9826	NA	> 0.05

Table 1: Results of the most parsimonious GLMM for the number of iSNVs, including and excluding lymph node tissue samples. Significant effects ($P < 0.05$) are highlighted in pink. See **Supplementary Table 3-1** for full model details and alternative less parsimonious models (based on AIC).

Next, we compared π, the average number of pairwise nucleotide differences within a sample. Unlike the number of iSNVs, which treats variants at all frequencies equally, π is maximized when variants approach 50% frequency and is less sensitive to low-frequency variants. We found that deer have significantly higher values of π compared to humans (Wilcoxon rank sum test, $p < 0.001$), and the difference remains significant even after excluding lymph tissue samples ($p = 0.021$). While dogs also had significantly higher π values, this result relied on a relatively small sample of dogs **(Figure 1b)**.

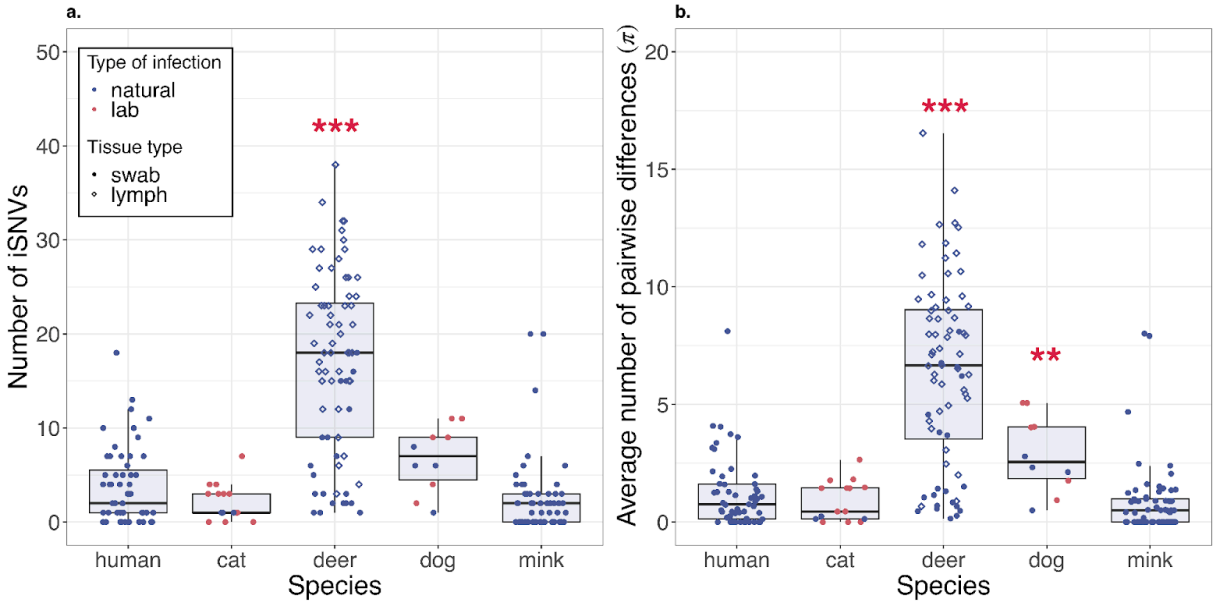

Figure 1. High within-host diversity in deer. Boxplots display (a) the number of iSNVs and, (b) the average number of pairwise differences (π) across animal species. Raw data points are shown with random horizontal jitter. Blue points are from natural infections and experimental lab infections in pink. Nasopharyngeal samples are shown with closed circles; lymph node tissue samples are shown with open diamonds. The boxes display the interquartile range, with the midline marking the median, and whiskers extending up to 1.5x of the interquartile range, or data extremums. Datapoints beyond whiskers are outliers according to the interquartile criterion. The statistical significance of the Wilcoxon rank sum test comparing each animal species with humans is marked with asterisks: *$p<0.0125$; **$p<0.001$, ***$p<0.0001$. Only p-values significant after Bonferroni correction for multiple tests are shown ($p < 0.05/4 = 0.0125$).

Mixed infections are rarely detected

Within-host genetic diversity can emerge via *de novo* mutations occurring during infection or from mixed infections of lineages that diverged before the animal was infected. To determine whether mixed infections could have contributed to the observed within-host diversity, we inferred the presence and relative abundance of distinct lineages of SARS-CoV-2 within each sample (Methods).

Despite eliminating low frequency lineages whose frequency fell below our minimum threshold of 0.025 (based on lab-infected samples), multiple viral lineages were inferred in 68 out of the 233 samples. The lineages inferred within a sample were almost always closely related, suggesting false-positive coinfections that could plausibly have arisen instead by *de novo*

mutation. This is particularly likely if natural infections typically last longer, or are sampled later, than experimental lab infections. Of the 68 samples inferred to contain more than one distinct lineage, only one (from a human sample) contained two distinct WHO-defined variants of concern, Alpha and Gamma (**Supplementary Table 5**). Seven other samples contained mixtures of Delta and multiple closely-related B.1.x lineages. In the remaining 60 samples, all identified lineages were closely-related B.1.x descendants (**Supplementary Table 5**). The inferred co-infection infections should therefore be interpreted with caution.

With these caveats in mind, we found more lineages in deer (median of two lineages per sample) compared to humans, with a median of one lineage per sample (**Figure 2a**; Wilcoxon rank sum test, $p < 0.001$). However, after controlling for effects of coverage, infection type, and tissue type, deer did not have higher lineage richness than humans (**Methods**). The model fit for lineage richness included species, sample isolation source, type of infection and breadth of coverage as fixed effects and depth of coverage and assay type as random effects (**Methods, Table 2**). In this model, breadth of coverage was a significant effect with a rate ratio of 0.15, indicating that samples with a lower breadth of coverage tend to have a higher number of inferred lineages, highlighting the importance of controlling for coverage effects in these analyses. Cat, dog, and mink all had significantly lower lineage richness compared with humans. In deer, lymph node tissue samples had significantly higher richness than nasopharyngeal samples (Wilcoxon rank sum test, $p < 0.001$), which could explain the overall higher number of lineages observed overall (**Figure 2a**). Similar to the number of iSNVs, to assess potential effects of collinearity between tissue isolation source and species, we re-ran the best model for lineage richness on the dataset after removing lymph node tissue samples **(Table 2, supplementary table 3-2)**. The primary conclusions remain unchanged, with deer not significant in the model, and cat, dog, and mink significant with negative effect sizes, indicating lower richness in these animals.

The Shannon diversity, a metric that accounts for both lineage number and evenness, was also highest in deer and lowest in dog and cat (**Figure 2b**). Shannon diversity did not follow a distribution amenable to a GLMM, so we only applied nonparametric tests that did not account for coverage and other sample effects. Together, these results suggest that, although within-host lineage diversity is somewhat higher in deer than humans, this effect may be explained by the higher number of lineages present in lymph nodes compared to nasopharyngeal samples. The higher within-host genetic diversity observed in deer, even among nasopharyngeal samples

(**Figure 1**), is therefore not easily explained by mixed infections of multiple SARS-CoV-2 lineages.

	Model including lymph-node tissue samples			Model excluding lymph-node tissue samples		
Predictors (Fixed effects)	**Estimate**	**Rate ratio**	**P-value**	**Estimate**	**Rate ratio**	**P-value**
Species (Reference: human)						
cat	-1.3561	0.2577	0.00951	-1.38011	0.2515	0.00960
dog	-1.3084	0.2702	0.00321	-1.34320	0.2610	0.00225
mink	-1.0709	0.3427	3.41E-07	-1.07778	0.3403	1.41E-07
deer	-0.1144	NA	> 0.05	-0.08527	NA	> 0.05
Isolation source (Reference: Nasopharyngeal swab)						
Lymph node tissue	0.8788	2.4080	8.26E-05	NA	NA	NA
Type of infection (Reference: Natural infection)						
Lab infected	0.9651	NA	> 0.05	1.01541	2.76050	0.04687
Breadth of coverage	-1.8793	0.1527	1.11E-07	-2.35462	0.0949	1.72E-07

Table 2: Results of the most parsimonious GLMM for lineage richness, including and excluding lymph node tissue samples. Significant effects ($P < 0.05$) are highlighted in pink. See **Supplementary Table 3-2** for full model details and alternative less parsimonious models (based on AIC).

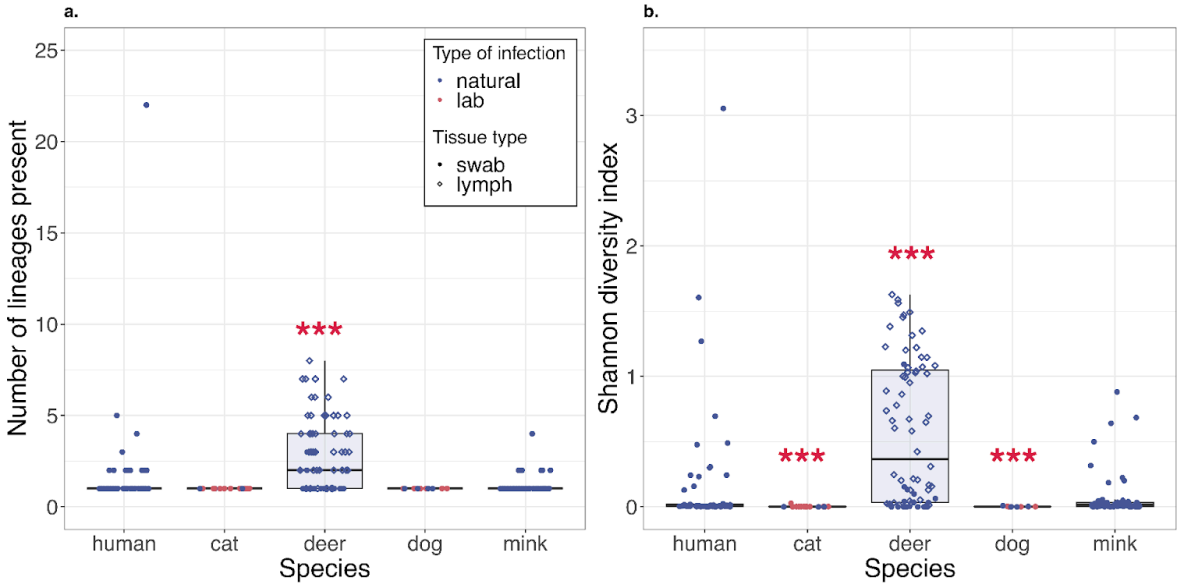

Figure 2. Multiple-lineage infections are rare across species. Boxplots display (a) the number of inferred lineages present and (b) the Shannon diversity index across species. Blue points are from natural infections and experimental lab infections in pink. Nasopharyngeal samples are shown with circles; lymph node tissue samples are shown with diamonds. The boxes display the interquartile range, with the midline marking the median, and whiskers extending up to 1.5x of the interquartile range, or data extremums. Datapoints beyond whiskers are outliers according to the interquartile criterion. The statistical significance of the Wilcoxon rank sum test comparing each animal species with humans is marked with asterisks: *p<0.0125; **p<0.001, ***p<0.0001.

Tests for selection on within-host diversity

To test for species-specific differences in the efficacy of natural selection acting on within-host diversity, we compared the ratio of nonsynonymous and synonymous polymorphism rates (pN/pS). In the absence of efficient selection against nonsynonymous mutations, which tend to be deleterious, we expect a pN/pS ratio close to one. By contrast, pN/pS ratios below one point to purifying selection against non-synonymous mutations, while pN/pS > 1 is interpreted as positive selection on adaptive protein changes. We observe a genome-wide pN/pS ratio below one (range of 0.4-0.7) for all species (**Figure 3**), consistent with the predominance of purifying selection within hosts across the genome. However, none of the animals have pN/pS significantly different from humans (Wilcoxon rank-sum test, $P > 0.05$ after Bonferroni correction). Genome-wide pN/pS ratios were significantly higher (Wilcoxon rank-sum test, $P < 0.01$) in deer lymph node samples (mean 0.87) compared to deer nasopharyngeal swab samples (mean 0.49). This could be due to higher rates of positive selection in lymph nodes, resulting in

many samples with pN/pS > 1 (**Figure 3**), relaxed purifying selection in lymph nodes, or some combination of these and other factors. Other than these tissue-specific differences, we find no evidence for systematic differences in the efficacy of selection on protein coding sequences among animal species.

The Spike protein is a major target of the immune system and might be under distinct, and possibly species-specific, selective pressure compared to other genes. We found that pN/pS values in Spike are close to one in all species (**Figure 3**). This could be due to weak purifying selection on Spike, or a combination of purifying and positive selection. Another possibility is that our pN/pS estimates are skewed towards one because of a numerical correction to avoid division by zero. Due to this correction, samples with zero iSNVs are assigned pN/pS = 1. With this caveat in mind, we found significantly higher pN/pS in mink than in humans in the S gene (**Figure 3**). This could be due to shorter sampled infection periods in mink, resulting in less time for selection to purge deleterious nonsynonymous mutations compared to humans. However, because genome-wide pN/pS does not differ between humans and mink, this explanation seems unlikely. Instead, species-specific selective pressures on Spike might explain the differences in pN/pS.

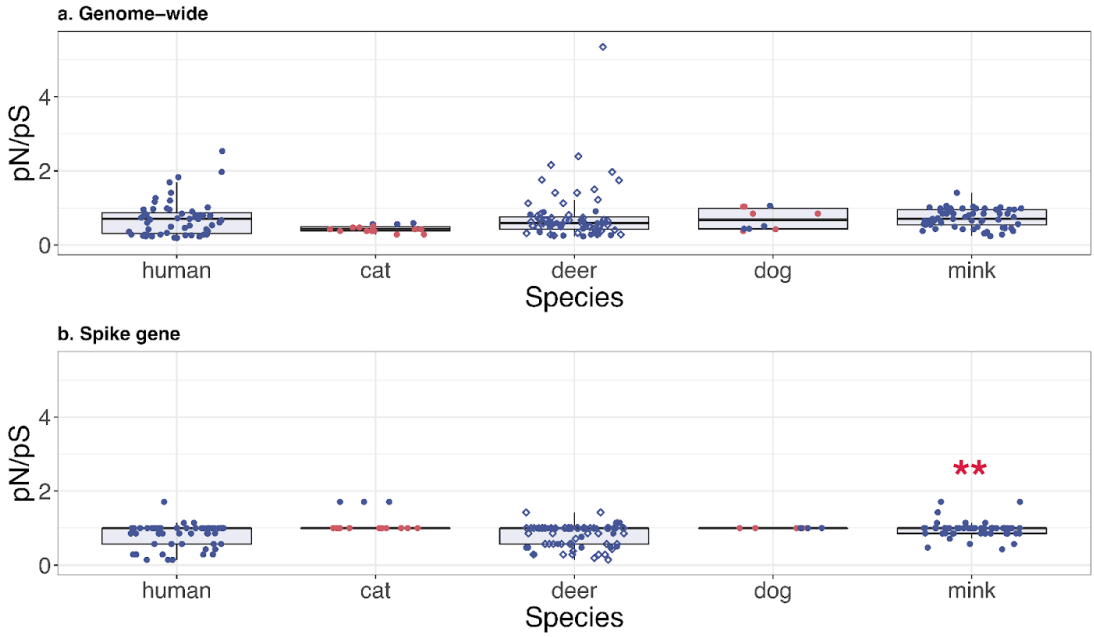

Figure 3. Low variation in within-host selection on protein-coding (pN/pS) across animal species. Boxplots show pN/pS ratios calculated for iSNVs (a) genome-wide and (b) in the S gene. Plots are grouped by host species on the x-axis. Blue points are from natural infections and experimental lab

infections are in pink. Nasopharyngeal samples are shown with circles; lymph node tissue samples are shown with diamonds. The boxes display the interquartile range, with the midline marking the median, and whiskers extending up to 1.5x of the interquartile range, or data extremums. Datapoints beyond whiskers are outliers according to the interquartile criterion. The statistical significance of the Wilcoxon rank sum test comparing each animal species with humans is marked with asterisks: *p<0.0125; **p<0.001, ***p<0.0001.

Through genome-wide association studies (GWAS), we previously identified three significantly mink-associated and 26 deer-associated SNVs (Naderi et al., 2023). These SNVs are putative species-specific adaptations that should therefore be under species-specific positive selection. We therefore expect deer-associated GWAS hits to be fixed (at 100% frequency within-host) or polymorphic at intermediate frequencies (iSNVs) within deer, representing complete selective sweeps and transmitted variants, or sweeps in progress, respectively. Consistent with this expectation, we found that deer-associated GWAS hits are more likely to be found as both fixed and polymorphic within deer samples compared to SNVs that were not GWAS hits, and compared to mink-associated GWAS hits which are not expected to be adaptive in deer (**Figure 4a-b**). Similarly, mink-associated GWAS hits were both more likely to be fixed and more likely to be polymorphic within mink samples than other categories of SNVs (**Figure 4c-d**). These results both validate the GWAS results on a sample of animal sequences that is independent from those used in the original study (Naderi et al., 2023) and also indicate that GWAS hits tend to be under selection within individual animals. Together with the pN/pS results, our observations support the idea that natural selection is occurring within animals, even if the efficacy of selection does not vary widely across animal species.

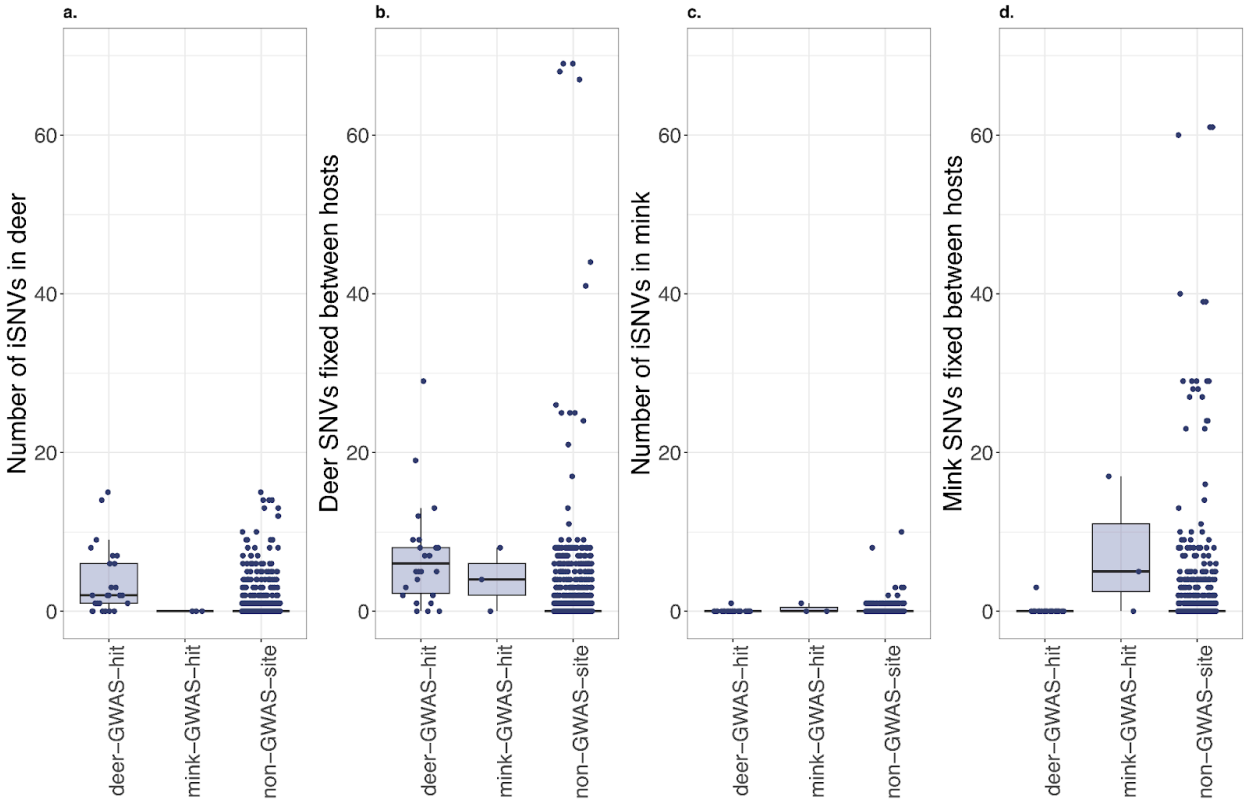

Figure 4. Deer- and mink-associated GWAS hits are more likely to be found as polymorphic (iSNVs) within hosts or fixed SNVs in their respective host-species compared with other sites on the genome. The boxes display the interquartile range, with the midline marking the median, and whiskers extending up to 1.5x of the interquartile range, or data extremums. Datapoints beyond whiskers are outliers according to the interquartile criterion.

Discussion

In our previous GWAS, we identified 26 putative species-specific mutations in deer, and only three in mink (Naderi et al., 2023). This suggests a greater amount of adaptive evolution in deer compared to mink, which could be explained by longer deer-to-deer transmission chains, as deer may be in less frequent contact with humans than farmed mink (Naderi et al., 2023, Lu et al., 2021). The high incidence of SARS-CoV-2 in deer – over 20% positivity rate in New York State – are also consistent with frequent deer-to-deer transmission (Feng et al., 2023). Longer deer-to-deer transmission chains, uninterrupted by passage through humans, could allow deer to accumulate deer-specific mutations that are rarely seen in humans. Here, our analysis of within-host diversity across animal species revealed a higher level of polymorphism within individual deer compared with humans.

411 Why do deer harbor more within-host diversity than other animals? Both measures of
412 within-host diversity, the number of iSNVs and the pairwise nucleotide identity, π, were
413 elevated in deer relative to humans. Based on these results, we favor the hypothesis that deer
414 infections are systematically longer-lasting – or sampled at later stages of infection – compared
415 to humans. Persistent infections potentially provide more raw genetic variation for natural
416 selection to act upon within deer, which could explain the numerous putative deer-specific
417 adaptive mutations (*i.e.* GWAS hits). These observations are also consistent with the long-term
418 accelerated evolutionary rate of SARS-CoV-2 among deer (McBride et al., 2023).
419
420 Chronic infections of SARS-CoV-2 in immunocompromised humans have been extensively
421 studied, indicating accelerated and potentially adaptive evolution within these patients (Borges
422 et al., 2021). Comparing patterns of evolution in chronic infections in humans with global
423 transmission chains suggests high similarity between chronic infections and lineage-defining
424 mutations of variants of concern (Harari et al., 2022). Similarly, chronic infections could be a
425 viable explanation for accelerated evolution in deer (McBride et al., 2023) and the higher
426 within-host diversity observed in our data. Another possibility is that deer are no more likely to
427 experience chronic infections than humans, but that deer are systematically sampled at later
428 stages of infection. This explanation is plausible if humans actively seek medical care (and
429 associated sampling) shortly after (or even before) symptom onset, whereas deer are sampled
430 opportunistically at random time points during infection. Distinguishing between these
431 possibilities is challenging given the available data. Mixed infections by distinct viral lineages do
432 not easily explain the higher diversity in deer – at least when comparing nasopharyngeal
433 samples only and controlling for confounders like sequencing effort. Deer lymph node tissues do
434 show evidence for more frequent mixed infections, although this is not comparable across other
435 animal species which lacked lymph node samples in our data.
436
437 Human chronic infections and deer infections with SARS-CoV-2 bear relevant similarities and
438 differences. In both cases, viral evolution is rapid and likely adaptive, supported by the
439 similarities of human chronic infection-derived mutations and those found in global variants of
440 concern (Harari et al., 2022), and by the high number of GWAS hits in deer (Naderi et al.,
441 2023). Here we showed that both iSNVs and fixed SNVs in deer are enriched in deer-associated
442 GWAS hits, suggesting that these mutations are under selection within hosts. While it has been
443 suggested that variants of concern such as Omicron could have arisen in animal hosts (Wei et
444 al., 2021), there remains no strong evidence that this has occurred to date. It is more plausible

445 that a virus highly adapted to humans would arise from selection within a chronically infected
446 human host than within an animal host, due to divergent species-specific selective pressures. On
447 the other hand, it may be easier than previously thought for a population to evolve from one
448 fitness peak (*e.g.* deer-adapted) to another (*e.g.* human-adapted) peak (Papkou et al., 2023).
449 Viral evolution in non-human animals therefore merits continued surveillance because
450 animal-specific adaptation does not preclude future adaptation to humans.
451
452 We have focused this discussion on deer, the species with the strongest evidence for
453 species-specific adaptation and with the highest levels of within-host diversity, but we also made
454 some other observations worth noting. First, lymph nodes, although only sampled from deer,
455 contain more within-host diversity than nasopharyngeal samples. This is consistent with
456 previous findings of compartmentalization of infection dynamics in different tissue types (Ke et
457 al., 2022). It could also indicate a larger viral population size capable of harboring greater
458 genetic diversity, more distinct viral lineages, or differences in the stage of infection. As
459 evidenced by the significant differences in genome-wide pN/pS, selective pressures may differ
460 between viral populations sampled from lymph nodes and nasopharyngeal swabs. Second, the
461 low within-host diversity in cats and dogs is consistent with relatively short infections in these
462 animals, resulting in less time for purifying selection to remove deleterious mutations in Spike.
463 However, the small sample size of cats and dogs prevents us from drawing firm conclusions.
464 Mink, on the other hand, were well-sampled and contained low within-host viral diversity,
465 consistent with infections that were acute or sampled at early stages – which may be more likely
466 in the context of early asymptomatic sampling during COVID-19 outbreaks on mink farms (Lu et
467 al., 2021). Rates of mixed infections are also rare in mink, which could be explained by
468 outbreaks of a single viral lineage. Mink also tend to have higher within-host pN/pS ratios in the
469 Spike gene than humans, consistent with distinct species-specific selective pressures on Spike.
470 The Spike gene also contained all three mink-associated GWAS hits in our previous study
471 (Naderi et al., 2023), suggesting that mink-specific selective pressures may be targeted on Spike,
472 both within and between hosts, in contrast to more genome-wide adaptation in deer. Together,
473 our results highlight the need for further sampling of diverse animal species – and the analysis
474 not just of consensus sequences, but of raw sequencing reads that can provide insight into
475 within-host evolution.

Acknowledgments. We are grateful to Sally Otto and Art Poon for providing constructive feedback on an earlier draft of the manuscript.

Code availability. All code used for the analyses is available at https://github.com/Saannah/animalSARS-within-host

Funding. This work was supported by a grant from the Canadian Institutes for Health Research (CIHR) Coronavirus Variants Rapid Response Network (CoVaRR-Net; ARR-175622).

Supplementary Tables:

Supplementary table 1: Table of NCBI SRA identifiers for samples originally downloaded, and their respective metadata.

Supplementary table 2: Table of the metrics of interest for each sample, post quality control filtering.

Supplementary table 3-1: Table of different GLMMs fit to the number of iSNVs along with their convergence and AIC values.

Supplementary table 3-2: Table of different GLMMs fit to richness along with their convergence and AIC values.

Supplementary table 4: Table of the occurrence of previously identified GWAS hits in each sample.

Supplementary table 5: Table of samples with richness larger than one, and the breakdown of lineages called for each by Freyja.

Bibliography

Andersen, K. G., Shapiro, B. J., Matranga, C. B., Sealfon, R., Lin, A. E., Moses, L. M., ... & Sabeti, P. C. (2015). Clinical sequencing uncovers origins and evolution of Lassa virus. Cell, 162(4), 738-750.

Avanzato, V. A., Matson, M. J., Seifert, S. N., Pryce, R., Williamson, B. N., Anzick, S. L., ... & Munster, V. J. (2020). Case study: prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised individual with cancer. Cell, 183(7), 1901-1912.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Borges, V., Isidro, J., Cunha, M., Cochicho, D., Martins, L., Banha, L., ... & Gomes, J. P. (2021). Long-term evolution of SARS-CoV-2 in an immunocompromised patient with non-Hodgkin lymphoma. Msphere, 6(4), e00244-21.

Bourgey, M., Dali, R., Eveleigh, R., Chen, K. C., Letourneau, L., Fillon, J., ... & Bourque, G. (2019). GenPipes: an open-source framework for distributed and scalable genomic analyses. Gigascience, 8(6), giz037.

Braun, K. M., Moreno, G. K., Wagner, C., Accola, M. A., Rehrauer, W. M., Baker, D. A., ... & Moncla, L. H. (2021). Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. PLoS pathogens, 17(8), e1009849.

Choi, B., Choudhary, M. C., Regan, J., Sparks, J. A., Padera, R. F., Qiu, X., ... & Li, J. Z. (2020). Persistence and evolution of SARS-CoV-2 in an immunocompromised host. New England Journal of Medicine, 383(23), 2291-2293.

Clark, S. A., Clark, L. E., Pan, J., Coscia, A., McKay, L. G., Shankar, S., ... & Abraham, J. (2021). SARS-CoV-2 evolution in an immunocompromised host reveals shared neutralization escape mechanisms. Cell, 184(10), 2605-2617.

Feng, A., Bevins, S., Chandler, J., DeLiberto, T. J., Ghai, R., Lantz, K., ... & Wan, X. F. (2023). Transmission of SARS-CoV-2 in free-ranging white-tailed deer in the United States. Nature Communications, 14(1), 4078.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907.

Gonzalez-Reiche, A. S., Alshammary, H., Schaefer, S., Patel, G., Polanco, J., Carreño, J. M., ... & van Bakel, H. (2023). Sequential intrahost evolution and onward transmission of SARS-CoV-2 variants. Nature Communications, 14(1), 3235.

549 Harari, S., Tahor, M., Rutsinsky, N., Meijer, S., Miller, D., Henig, O., ... & Stern, A. (2022).
550 Drivers of adaptive evolution during chronic SARS-CoV-2 infections. Nature Medicine, 28(7),
551 1501-1508.
552
553 Karthikeyan, S., Levy, J. I., De Hoff, P., Humphrey, G., Birmingham, A., Jepsen, K., ... & Knight,
554 R. (2022). Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission.
555 Nature, 609(7925), 101-108.
556
557 Ke, R., Martinez, P. P., Smith, R. L., Gibson, L. L., Mirza, A., Conte, M., ... & Brooke, C. B.
558 (2022). Daily longitudinal sampling of SARS-CoV-2 infection reveals substantial heterogeneity
559 in infectiousness. Nature Microbiology, 7(5), 640-652.
560
561 Lu, L., Sikkema, R. S., Velkers, F. C., Nieuwenhuijse, D. F., Fischer, E. A., Meijer, P. A., ... &
562 Koopmans, M. P. (2021). Adaptation, spread and transmission of SARS-CoV-2 in farmed minks
563 and associated humans in the Netherlands. Nature Communications, 12(1), 6802.
564
565 Lythgoe, K. A., Hall, M., Ferretti, L., de Cesare, M., MacIntyre-Cockett, G., Trebes, A., ... &
566 Golubchik, T. (2021). SARS-CoV-2 within-host diversity and transmission. Science, 372(6539),
567 eabg0821.
568
569 McBride, D. S., Garushyants, S. K., Franks, J., Magee, A. F., Overend, S. H., Huey, D., ... &
570 Bowman, A. S. (2023). Accelerated evolution of SARS-CoV-2 in free-ranging white-tailed deer.
571 Nature Communications, 14(1), 5105.
572
573 Naderi, S., Chen, P. E., Murall, C. L., Poujol, R., Kraemer, S., Pickering, B. S., ... & Shapiro, B. J.
574 (2023). Zooanthroponotic transmission of SARS-CoV-2 and host-specific viral mutations
575 revealed by genome-wide phylogenetic analysis. Elife, 12, e83685.
576
577 Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and
578 nonsynonymous nucleotide substitutions. Molecular Biology and Evolution, 3(5), 418-426.
579
580 NCBI: National Library of Medicine (US), National Center for Biotechnology Information; 2009
581 - [cited 2024/01/20]. Available from: https://www.ncbi.nlm.nih.gov/sra/
582
583 N'Guessan, A., Brito, I. L., Serohijos, A. W., & Shapiro, B. J. (2021). Mobile gene sequence
584 evolution within individual human gut microbiomes is better explained by gene-specific than
585 host-specific selective pressures. Genome Biology and Evolution, 13(8), evab142.
586
587 Papkou, A., Garcia-Pastor, L., Escudero, J. A., & Wagner, A. (2023). A rugged yet easily
588 navigable fitness landscape. Science, 382(6673), eadh3860.
589
590 Pickering, B., Lung, O., Maguire, F., Kruczkiewicz, P., Kotwa, J. D., Buchanan, T., ... & Bowman,
591 J. (2022). Divergent SARS-CoV-2 variant emerges in white-tailed deer with deer-to-human
592 transmission. Nature Microbiology, 7(12), 2011-2024.
593

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Tan, C. C., Lam, S. D., Richard, D., Owen, C. J., Berchtold, D., Orengo, C., ... & Balloux, F. (2022). Transmission of SARS-CoV-2 from humans to animals and potential host adaptation. Nature Communications, 13(1), 2988.

Tonkin-Hill, G., Martincorena, I., Amato, R., Lawson, A. R., Gerstung, M., Johnston, I., ... & Wellcome Sanger Institute COVID-19 Surveillance Team. (2021). Patterns of within-host genetic diversity in SARS-CoV-2. Elife, 10, e66857.

Valesano, A. L., Rumfelt, K. E., Dimcheff, D. E., Blair, C. N., Fitzsimmons, W. J., Petrie, J. G., ... & Lauring, A. S. (2021). Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. PLoS Pathogens, 17(4), e1009499.

Wei, C., Shan, K. J., Wang, W., Zhang, S., Huan, Q., & Qian, W. (2021). Evidence for a mouse origin of the SARS-CoV-2 Omicron variant. Journal of Genetics and Genomics, 48(12), 1111-1121.

Xue, K. S., & Bloom, J. D. (2020). Linking influenza virus evolution within and between human hosts. Virus Evolution, 6(1), veaa010.

Xue, K. S., Stevens-Ayers, T., Campbell, A. P., Englund, J. A., Pergam, S. A., Boeckh, M., & Bloom, J. D. (2017). Parallel evolution of influenza across multiple spatiotemporal scales. Elife, 6, e26875.