Multiple Introductions of SARS-CoV-2 Alpha and Delta Variants into White-Tailed Deer in Pennsylvania

Andrew D. Marques[1], Scott Sherrill-Mix[1], John K. Everett[1], Hriju Adhikari[1], Shantan Reddy[1], Julie C. Ellis[2], Haley Zeliff[3], Sabrina S. Greening[2], Carolyn C. Cannuscio[4, 5], Katherine M. Strelau[4,5], Ronald G. Collman[6], Brendan J. Kelly[7], Kyle G. Rodino[8], Frederic D. Bushman[1], Roderick B. Gagne[2], Eman Anis[3]

[1] Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

[2] Department of Pathobiology, Wildlife Futures Program, University of Pennsylvania School of Veterinary Medicine, New Bolton Center, Kennett Square, PA 19348

[3] Department of Pathobiology, University of Pennsylvania School of Veterinary Medicine, New Bolton Center, Kennett Square, PA 19348

[4] Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA 19104

[5] Department of Family Medicine and Community Health, University of Pennsylvania, Philadelphia, PA 19104

[6] Pulmonary, Allergy and Critical Care Division; Department of Medicine; University of Pennsylvania Perelman School of Medicine; Philadelphia, PA 19104

[7] Division of Infectious Diseases; Department of Medicine & Department of Biostatistics, Epidemiology, and Informatics; Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

[8]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

Correspondance

Eman Anis: eanis@vet.upenn.edu

Erick Gagne: rgagne@vet.upenn.edu

Frederic Bushman: bushman@pennmedicine.upenn.edu

Introductory paragraph (150 words max)

The SARS-CoV-2 pandemic likely began by viral spillover from animals to humans[1-3]; today multiple animal species are known to be susceptible to infection[4-8]. White-tailed deer, *Odocoileus virginianus* are infected in North America at substantial levels[9-11], and genomic data suggests that a variant in deer may have spilled back to humans[12,13]. Here we characterize SARS-CoV-2 in deer from Pennsylvania (PA) sampled during fall and winter 2021. Of 123 nasal swab samples analyzed by RT-qPCR, 20 (16.3%) were positive for SARS-CoV-2. Seven whole-genome sequences were obtained, together with six more partial spike sequences. These annotated as alpha and delta variants, the first reported observations of these lineages in deer, documenting multiple new jumps from humans to deer. The alpha lineage persisted in deer after its displacement by delta in humans, and deer-derived alpha variants diverged significantly from those in humans, consistent with a distinctive evolutionary trajectory in deer.

Main text

White-tailed deer were sampled throughout Pennsylvania from 10/2/2021 to 12/27/2021 with the goal of characterizing this potential new animal reservoir of SARS-CoV-2. Samples were tested for the presence of SARS-CoV-2 by qPCR, and the viral lineages present assessed using either viral whole genome sequencing after multiplex PCR, or, for samples with lower RNA amounts, sequencing of a single PCR amplicon encoding the spike receptor binding domain.

SARS-CoV-2 was detected by RT-qPCR in nasal swabs from 20 of 123 wild white-tailed deer sampled (16.3%; 95% CI: 10.4, 24.2) (described in Table S1 and S2). There was no significant difference in infection frequency associated with the cause of death (road kill 7/24, and hunter harvested 12/79; p-value = 0.14). There was no significant difference in infection rates between sexes (females 6/43, males 14/60; p-value = 0.46) and age groups (fawn 1/18, yearling 1/22, and adult 17/81; p-value = 0.093). No information was available on possible symptoms or disease for the animals sampled.

Virus-positive deer were identified in 10 of the 31 Pennsylvania counties surveyed (Figure 1); a Bayesian estimate of the infection rates is shown by the color scale. We estimate an average county has 9.9% positivity (95% CrI: 1.7-22%). Pike and Monroe counties were associated with higher apparent prevalence (Monroe: 7.5x higher odds (1.0-110x) of positivity; Pike: 9.3x higher odds (1.3-110x)). There was a significant difference in infection rates among regions (p-value = 0.0047), with the northeast region of PA showing a higher

proportion of positive deer than the southeast region (Fisher's exact test with FDR correction; p-value = 0.017).

Viral whole-genome sequencing was attempted on eight samples with relatively high viral RNA concentrations (cycle threshold values from RT-qPCR less than 30, a cut value determined from analysis of human specimens), and high-quality genome sequences were recovered from seven (Table S2). To confirm sample provenance, the non-viral sequence reads were analyzed for the deer samples and for human samples that were sequenced in parallel (Table S6). All deer samples yielded reads mapping to the deer genome, and few or none mapping to the human genome. In contrast, reads from human samples overwhelmingly mapped to the human genome (Figure S1), confirming the host species origin of our samples as deer.

Samples with lower SARS-CoV-2 RNA concentrations were amplified by a single nested PCR amplification targeting the spike coding region as described[14] and sequenced, yielding spike sequences from six additional samples. Spike sequences were then used for lineage assignment[15] (Table S2).

Deer whole genome sequences showed divergences relative to the original Wuhan SARS-CoV-2 strain and previously reported human and deer genome sequences (Figure 2 A). A full list of deer substitutions is in Tables S3. The complete deer genome sequences reported here annotated as either alpha or delta variants[16], the first reported identification of these lineages in wild white-tailed deer. Previously the alpha variant was shown to be able to infect white-tailed deer that were experimentally inoculated[11]. All the spike-only sequences

harbored polymorphisms that place them in the delta lineage and were inconsistent with other common lineages.

The types of base changes associated with substitutions was not uniform. Changes from C in the Wuhan reference to U were found to be the most frequent (Figure 2B). The frequency of C to U was even higher in deer than in humans (Pearson's Chi-squared test p=0.031, followed by a post hoc analysis based on residuals yielding p-value = 0.0010 for C to U substitutions), consistent with a previous study[13], and suggestive of host-specific mutation rates.

To identify any deer-specific sequence polymorphisms, the alpha and delta SARS-CoV-2 sequences from deer determined here were compared to viral genomes from humans and deer reported previously. All deer genome sequences available from GISAID (n=108) were downloaded (Table S4 and S5), corresponding to samples acquired from 9/28/2020 to 2/25/2021 in Iowa and Ohio and dominated by early pandemic lineages including B.1.2 and B.1.311[9,10]. Local human isolates were derived from our SARS-CoV-2 sequencing program monitoring the Delaware Valley, including Philadelphia, PA[17,18]. This revealed several substitutions that were highly enriched or invariant in deer isolates, but rare or absent in human isolates. These include three silent mutations in ORF1ab, C7303U, C9430U, and C20259U. Mutation C7303U was found in 86% of these PA deer and 29% of previously published deer, whereas it was found in 0.04% of Delaware Valley human subjects and 0.09% of global genomes reported by NextStrain. Mutation c9430t was found in 43% of PA deer and 56% of previously published deer, but 0.35% of Delaware Valley human subjects and

0.46% of global genomes. Mutation c20259t was found in 43% of PA deer and 19% of previously published deer, whereas it was absent in Delaware Valley human subjects and present in 0.12% of global genomes (data from 1/26/2022). The enrichment of these mutations suggests possible functional interaction with deer-specific factors, which could influence RNA synthesis, RNA folding, or protein binding.

We next analyzed the frequencies of synonymous and nonsynonymous substitutions in each genome to assess potential evolutionary pressures on SARS-CoV-2 in deer. The dN/dS ratios for the deer genomes did not diverge notably from matched viral variants in humans--the delta lineage in humans showed more evidence for diversifying selection than did alpha, and this was paralleled by sequences in deer. Strong diversifying selection was seen for the spike coding region in alpha and delta in both hosts, while neutral or purifying selection was more evident in most other coding regions (Supplementary Figure S2).

To assess relationships among deer and human-derived SARS-CoV-2 sequences, we used NextClade for sequence alignment and IQ-TREE to construct a maximum-likelihood phylogenetic tree for viral sequences from alpha (Figure 3A) and delta (Figure 3B) variants, in each case comparing PA deer to contemporaneous human-derived sequences from the same variant in the Delaware Valley region. We also compared the deer lineages observed longitudinally to all contemporaneous human lineages sampled in PA (Figure 3C, Table S7).

The two alpha sequences sampled from deer came from adjoining counties in northeastern PA. The isolates differed by 45 substitutions and were widely separated on the phylogenetic tree. A parsimonious explanation is that the two alpha lineages were introduced independently into deer and then diversified during transmission within deer. Only one out of 1,240 human genomes was as divergent or more divergent from the inferred alpha common ancestor as were the deer sequence, giving a probability of randomly drawing two such divergent lineages as $(1/1240)^2=6.5X10^{-7}$. Molecular clock analysis yielded data consistent with a model in which alpha jumped from humans to deer in April/May of 2021, during the alpha wave, and persisted in deer up until the sampling time six to seven months later in mid-November (Figure S3, Table S7 and S8).

For the delta variant (Figure 2B), five deer genomes annotated as delta, and all six spike sequences showed substitution patterns also consistent with delta (Table S3). Two deer genomes were assigned to the AY.103 clade, two genomes to AY.88, and one to AY.5. The delta lineages from deer were not as diverged from related human sequences as the alpha lineages, consistent with more recent introductions and less long-term circulation in deer. Molecular clock analysis suggested that the closely related AY.103 pair of genomes likely entered the deer population in mid-July 2021; the AY.5 lineage entered in mid-August 2021; and the AY.88 pair entered deer in late-September 2021. For the delta cases, the proposed jumps from deer to humans occurred during the height of the delta wave. The six samples sequenced as single spike amplicons also displayed mutations consistent with delta variants, and were collected at times

and locations similar to the fully sequenced delta variant genomes from deer. The small size of these amplicons prevents assignment to clade and identification of transmission events but confirms the extensive presence of delta in deer.

Thus, these data support five independent jumps of SARS-CoV-2 from humans to deer—one each for the two alpha sequences, and three for the delta sequences (Supplementary Figure S3). Other scenarios are also possible, involving more independent jumps and convergent evolution in deer, or fewer jumps and divergent evolution along trajectories matching trajectories in humans.

This study has several limitations. The sample size is modest, and whole genome sequence acquisition was limited by viral RNA concentrations in samples. Our sample size limits the interpretation of epidemiological findings (e.g., differences in infection rates between sample types) but has less influence on the primary findings of this work (e.g., divergence of the alpha lineage in deer, and evidence for multiple spillovers). The background human data is incomplete, with only an estimated 1.7% percent of all human cases in PA subjected to viral whole genome sequencing. Each evolutionary analysis of sequences requires assumptions on the structures of background populations that are not fully investigated. In addition, we were not able to obtain serum samples from the deer studied and so could not investigate immune responses.

In summary, a survey of SARS-CoV-2 in 123 deer in Pennsylvania over the fall-winter of 2021 showed 16% of the deer sampled to be positive. Prior surveys carried out over the fall and winter of 2020 also showed high point

prevalence of infection in Iowa (33%)[9] and Ohio (36%)[10], and later extensive infection at other locations [13,19]. We report the first examples of the alpha and delta lineages in wild white-tail deer, likely derived from roughly five independent spillovers from humans to deer. Given that there are estimated to be 30 deer per square mile in PA, and over a million deer total, this suggests an enormous number of spillovers and infected deer in the state[20,21]. Mechanisms of infection and transmission are incompletely understood. Studies of experimental infections show efficient transmission between deer, potentially involving close interactions such as touching noses and grooming[22]. The mechanism of efficient transmission from humans to deer remains obscure. Our findings of alpha persistence in deer after replacement of alpha by delta in humans, and the divergence seen between our deer and human alpha genomes, are all consistent with long-term persistence and spread of the alpha variant in deer. As yet there is no evidence for spill-back of the deer lineages identified here into humans; ongoing efforts to characterize human and deer SARS-CoV-2 lineages are valuable to maintain surveillance for such events.

Methods

Collection of samples from white-tailed deer

Samples were collected from hunter-harvested deer and injured deer that were euthanized by state game wardens. Nasal swabs were taken within hours of death, placed in phosphate buffered saline (PBS) and stored in commercial

refrigerators in field offices. Samples were shipped to the University of Pennsylvania within one week of collection and stored at -80°C until RNA extraction. Fisher's Exact tests were performed using R Statistical software (v4.0.5)[23] (R Core Development Team 2021) to test for differences in proportions of virus positive and negative deer by sex, age, and cause of death (i.e., hunter harvested, road-killed, or all other causes of death).

RNA extraction and detection of SARS-CoV-2 RNA by RT-qPCR

Nucleic acid was extracted with a QIAamp viral RNA mini kit (Qiagen) according to the manufacturer's instructions. The presence of SARS-CoV-2 RNA was assessed by a RT-qPCR targeting two regions of the viral nucleocapsid gene as described previously.[24]

SARS-CoV-2 whole genome sequencing

Viral genomes were sequenced using the POLAR protocol[25]. The sample RNA (5µl) was mixed with Random Hexamers (0.5 µl of 50 µM, Thermo Fisher), dNTPs Mix (0.5µl of 10mM each, Thermo Fisher), and nuclease-free water (1 µl). This mixture was incubated (5 minutes at 65ºC). Subsequently, reverse transcription was performed adding 6.5 µl from the previous reaction to SuperScript III Reverse Transcriptase (0.5 µl, Thermo Fisher), 5X First-Strand Buffer (2 µl, Thermo Fisher), DTT (0.5 µl of 0.1M, Thermo Fisher), and RNaseOut (0.5 µl, Thermo Fisher). This mixture was incubated (50 minutes at 42ºC, then 10 minutes at 70ºC). For cDNA amplification, the previous mixture

(2.5 µl) was added to Q5 Hot Start DNA Polymerase (0.25 µl, NEB), 5X Q5 Reaction Buffer (5 µl, NEB), dNTPs mix (0.5 µl of 10mM each, NEB), and pooled Artic-ncov2019 v4 primers (reaction set 1 used 4.0 µl of pooled primer set 1, reaction set 2 used 3.98 µl of pooled primer set 2, IDT) and water (volume reaction brought to 25 µl). The samples were brought to 98ºC for 30 seconds then cycled from 98ºC for 15 seconds to 65ºC for 5 minutes for 25 cycles, followed by 98ºC for 15 seconds and 65ºC for 5 minutes. The two reactions (one for each primer set) were pooled together. Tagmentation was completed with the Nextera XT Library Preparation Kit (Illumina). IDT for Illumina DNA/RNA UD Indexes were used for barcoding (Illumina). Quantification of DNA was performed using Quant-iT PicoGreen dsDNA quantitation assay kit (Invitrogen). The pooled libraries were quantified using the Qubit1X dsDNA HS Assay Kit (Invitrogen) and sequenced using the Illumina NextSeq platform.

SARS-CoV-2 spike amplicon sequencing

Amplicon sequencing was carried out essentially as described[14]. Briefly samples with genomic viral load deemed too low for whole genome sequencing, or samples that failed whole genome sequencing, underwent nested PCR targeting the spike's receptor binding domain sequence. Primers used are provided in Supplemental Table S9. PCR reaction 1 includes reverse transcription and initial DNA amplification using the Superscript IV One-Step RT-PCR System (Thermo Fisher Scientific,12594100). The 25 µl reaction consists of 5 µl of extracted viral RNA (extracted as described in the whole genome

sequencing section), 0.25 µl of Superscript IV, 12.5 µl of 2X Platinum SuperFi RT-PCR mix, 1.25 µl of each PCR 1 primers (10 µM), 4.75 µl molecular grade water. The cycle conditions are as follows: 25ºC for 2 minutes, 50ºC for 20 minutes, 95ºC for 2 minutes, 25 cycles of amplification (95ºC for 2 minutes, 55ºC for 30 seconds, 70ºC for 1 minute). PCR reaction 2 is a nested PCR and addition of sequencing adapters for the amplicon product from reaction 1. The 25 µl mixture contains 5 µl of amplicon product from reaction 1, 12.5 µl 2x Q5 hot start master mix (NEB, M0494S), 0.5 µl of 10mM dNTPs (NEB, N0447S), 1.25 µl of each PCR 2 primers (10 µM), and 7.5 µl water. The conditions are as follows: 95ºC for 2 minutes, 20 cycles of amplification (95ºC for 15 seconds, 55ºC for 30 seconds, 72ºC for 1 minute). PCR reaction 3 adds the unique dual indexes and clustering adapters. The 50 µl reaction contains 25 µl of amplicon product from PCR reaction 2, 0.5 µl of Phusion polymerase (NEB, M0530L), 10 µl of 5x Phusion buffer, 1 µl of 10mM dNTPs (NEB, N0447S), 5 µl of IDT for Illumina DNA/RNA UD Indexes set A (Illumina, 20027213), and 8.5 µl of water. The reaction conditions are as follows: 98ºC for 3 minutes, 7 cycles of amplification (95ºC for 15 seconds, 50ºC for 30 seconds, 72ºC for 30 seconds), and 72ºC for 7 minutes. 10 µl of the barcoded PCR 3 mixture was pooled together, AMPure purified, and quantified by Qubit1X dsDNA and TAPE station before being sequenced on an Illumina MiSeq instrument using a 600 cycle v3 standard flow cell 290/10/10/290 protocol (Illumina, MS-102-3003). After sequencing, BWA v0.7.17-r1188 was used to generate bam files, Samtools v1.10 (using htslib 1.10.2-3) was used to filter reads for lengths between 280 and 300 base pairs,

filter reads by quality using a PHRED threshold of 30, sort the reads, index the reads, and generate a pileup. Samples must meet a minimum mapping quality PHRED score of 30, have 95% 50-fold coverage for the 550bp amplicon target to pass additional filtering. Substitutions were called with a greater than 0.67 allele frequency for a given position.

Sequence Data Processing

Sequence data were processed as previously described[17]. BWA aligner tool (v0.7.17) was used to align viral sequences to the Wuhan reference sequence (NC_045512.2)[26]. Samtools package (v1.10) was used to filter alignments[27]. Variants were called using Pangolin lineage software (3.1.17 with the PangoLEARN 2021-12-06 release)[16,28]. NVSL's vSNP pipeline was also run on deer samples and results compared to those generated by our published pipeline; findings from both platforms deviated only for thresholds used to call low abundance substitutions. All variant calls remained the same between platforms.

Host Sequence Analysis

The proportion of host sequences was inferred using a sampling of raw reads from all samples on any sequencing batch performed with deer specimen. 1,000 raw reads for each sample were blasted[29-31] against a database constructed of a SARS-CoV-2 genome (NC_045512.2), human genome (GCF_000001405.39), feline genome (GCF_018350175.1), and white-tailed deer genome

(GCF_002102435.1). Reads were tallied for each genome to which they were closely matched.

Mutation Analysis

108 deer-derived SARS-CoV-2 genomes were downloaded from GISAID and compared with the 7 deer sequenced in this publication and the human dataset (Tables S4 to S6). The previously published genomes were downloaded on 12/29/2021. Mutations were called in reference to the Wuhan strain (NC_045512.2).

Bayesian analysis of county proportions

To account for the variable sampling between counties and potential similarities between neighboring counties, we estimated the underlying proportion of deer testing positive within each county of the m counties using a Bayesian conditional autoregressive model. The number of positive tests, y_i, out of n_i total tests within each county i was modeled as:

$$y_i \sim \text{Binomial}(p_i, n_i)$$

where p_i is the proportion of deer expected to test positive in that county and:

$$p_i = \text{logit}^{-1}(\alpha + \beta_i)$$

Here, α represents the average proportion positive for a county and the vector of differences from this average for each county, β, is distributed as a multivariate normal:

$$\beta \sim \text{Normal}_{\text{prec}}(0, \frac{1}{\sigma}(\text{D-}\theta\text{A}))$$

where D is a $m \times m$ matrix with 0s on the off diagonal and the diagonal element on each row i, $D_{i,i}$, equal to the number of counties that are adjacent to county i, A is a $m \times m$ adjacency matrix with element $A_{i,j}$ is 1 if county i neighbors county j and 0 otherwise and the diagonal set to 0 and $\text{Normal}_{\text{prec}}(x, y)$ is a multivariate normal distribution with means x and precision matrix y. For priors, θ was given a uniform prior between 0 and 1, $\sigma \sim \text{Gamma}(1,1)$ and $\alpha \sim \text{Normal}(-2,10)$.

Posterior probability distributions were estimated using Markov chain Monte Carlo sampling using Stan v2.21.0 [32].

County adjacency data was obtained from the US Census bureau (https://www.census.gov/geographies/reference-files/2010/geo/county-adjacency.html).

Analysis of dN/dS

Sequences were aligned using nextclade v1.14.0[33]. Coding sequences were extracted from the genomes ignoring insertions and treating ORF1a and ORF1b as separate entities. dN/dS was calculated from between individual gene sequences or the concatenation of all coding sequences for each genome against the ancestral Wuhan genome using the ape package [34] in R v4.1.3. Sequences with nonsense mutations and estimated dN/dS values less than 0 were discarded. Estimated dN/dS values >10 were censored to 10 and values <0.1 were censored to 0.1.

Time-Scale Bayesian Maximum Clade Credibility Tree

A phylogenetic approach using a time-scale Bayesian maximum clade credibility (MCC) tree was used to estimate the introduction of SARS-CoV-2 into PA WTD. An NCBI BLASTn search was used to identify the nearest human-derived SARS-CoV-2 sequences for each deer, querying sequences world-wide and not just from the Delaware Valley. If a deer's nearest hits included another deer, then both were analyzed in the same tree. Sequences without exact dates of collection were excluded. Sample sizes ranged from 50 to 79 nearest neighbor whole genome sequences. Each dataset was aligned using NextClade with Wuhan-Hu-1 as a reference (NC_045512.2). A Markov chain Monte Carlo (MCMC) method was used to generate time-scales Bayesian molecular clocks using BEAST v.1.10.4[35]. Several parameters were assessed to determine the tree best to model phylogenetic data using path sampling and stepping-stone sampling of marginal likelihood estimation (Supplemental Table S8). A substitution model described by Yang[36], 1996 with an uncorrelated relaxed lognormal clock using Bayesian Skyline tree prior over 100 million iterations with subsampling every 1,000 iterations was deemed to be the optimal conditions over the parameters tested. The BEAGLE 3 library was used to improve computational performance[37]. Tracer v.1.7.1 was used to visually assess convergence. 100,000 iterations were discarded as burn-in. TreeAnnotator v.1.10.4 was used to summarize the MCC tree. FigTree v.1.4.4. was used to visualize the tree.

Human subjects

Human sequences newly determined here were collected as follows. For most samples, the University of Pennsylvania Institutional Review Board (IRB) reviewed the research protocol and deemed the limited data elements extracted to be exempt from human subject research per 45 CFR 46.104, category 4 (IRB #848605). For hospitalized subjects, following informed consent (IRB protocol #823392), patients were sampled by collection of saliva, oropharyngeal and/or nasopharyngeal swabs, or endotracheal aspirates if intubated, as previously described[17]. Further samples were collected from asymptomatic subjects detected in a screening program at the Perelman School of Medicine at the University of Pennsylvania and symptomatic subjects tested throughout the PennMedicine clinical network under IRB protocols #843565 and #848608. Human samples were sequenced as described for deer samples.

Phylogenetic Analysis

Deer and human sequences are available at GenBank (Table S2 and Table S6). NextClade was used to align sequences to the Wuhan reference[33]. IQ-TREE (v1.6.12) was used to infer a tree using maximum-likelihood methods using 1,000 bootstrap replicates[38]. Visualization of the inferred tree was performed using FigTree (v.1.4.4)[39]. Only a small fraction of human cases from Pennsylvania have been analyzed by viral whole genome sequencing (estimated at about 1.7% comparing GISAID upload for PA to state data on infection rates);

thus, our background human data set for comparison to the deer samples is quite incomplete.

Materials

Key materials and resources are compiled in Table S9.

Data availability

Sequence accession numbers for deer-derived SARS-CoV-2 genomes can be found in Table S2 (OM570187-OM570193 & ON350842- ON350847). Accession numbers for human SARS-CoV-2 genomes can be found in Tables S6 and S7. Computer code is available at https://doi.org/10.5281/zenodo.4046252.

Acknowledgements

We are grateful to hunters and wildlife personnel who provided specimens, and to Laurie Zimmerman for artwork and help with manuscript preparation. We acknowledge help from staff of the Philadelphia Department of Public Health. This work was supported in part by the Penn University Research Foundation. SG is supported by the Robert J. Kleberg, Jr. and Helen C. Kleberg Foundation. Funding was provided by a contract award from the Centers for Disease Control and Prevention (CDC BAA 200-2021-10986 and 75D30121C11102/000HCVL1-2021-55232), philanthropic donations to the Penn Center for Research on Coronaviruses and Other Emerging Pathogens, and in part by NIH grant R61/33-HL137063 and AI140442 -supplement for SARS-CoV-

2. BJK is supported by NIH K23 AI 121485. Additional assistance was provided by the Penn Center for AIDS Research (P30-AI045008). This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N93021C00015. HZ's stipend to conduct this work was provided by a private donation by Tracy Holmes.

Contributions

Conceived and designed the experiments. A. D. M., J. C. E., R. G. C., B. J. K., K. G. R., F. D. B., R. B. G., E. A.

Performed the experiments A. D. M., H. A., S. R., H. A., B. G., E. A.

Analyzed the data A. D. M, S. S.-M., J. E., F. D. B., B. G., E. A., S. G.

Contributed materials/analysis tools J. C. E., H. Z., S. S. G., C. C., K. S., R. G. C., B. J. K., K. G. R., F. D. B., R. B. G., E. A.

Wrote the paper A. D. M., S. S.-M., J. C. E., H. Z., S. S. G., C. C., K. S., R. G. C., B. J. K., K. G. R., F. D. B., R. B. G., E. A.

Figure Legends

Figure 1. Map of Pennsylvania (PA), showing sampling sites and locations of SARS-CoV-2 positive deer. The counties comprising PA are outlined. The estimated proportion of positive samples is shown by the white-to-red color scale; counties that were not sampled are shown in grey. Total sample numbers and

the number positive are written on each county sampled. The deer samples sequenced were assigned to variants as indicated by the rectangles outside the map; variant type is color coded (teal for alpha/B1.1.7, purple/pink for delta/AY.#). The open boxes indicate sequences were available for the spike only.

Figure 2. Base substitutions detected in sequences of deer from Pennsylvania. In A) the proportions in humans sampled in PA is shown in blue along the top for alpha (upper) and delta (lower). Substitutions relative to the Wuhan strain references are shown by the colored boxes. The proportions in humans are shown by the blue shading in each box. Genes sampled are shown by the color code indicated along the bottom. Grey indicates lack of sampling. The bottom six rows of the delta samples indicate the spike-only amplicon sequences. B) Types of base substitutions away from the Wuhan reference detected in humans (right) and deer (left). The proportions are shown by the color code at the left.

Figure 3. Analysis of SARS-CoV-2 whole genome sequences from white-tailed deer in PA and comparison to local human-derived sequences. A) Phylogeny of contemporaneous human alpha lineage sequences from PA, collected from 1/6/2021 to 11/16/2021, with deer sequences marked (1240 human and 2 deer sequences). B) Phylogeny of human delta lineage sequences from PA, collected from 10/14/2021 to 11/28/2021 (spanning the period of genomes obtained from deer), with deer sequences marked (440 human and 5 deer sequences). Time-

resolved trees for each spillover are in Figure S3. C) Longitudinal comparison of deer variants and contemporary human variants. The bar plots show the progression of SARS-CoV-2 variants detected by genome sequencing in humans in eastern PA from 1/31/2021 to1/3/2022. The variants are color-coded according to the key to the right. The variants from white-tailed deer sequences are shown at the top of the figure, with the arrows showing the times of sampling. The deer isolates are color-coded on the deer icons as in the key; purple dots on deer icons indicate sequences identified as delta lineage by spike amplicon sequences but without whole genome sequences to further identify clade.

Supplementary Material

Supplementary Table 1. Table S1. SARS-CoV-2 prevalence estimates stratified by sex, age, sampling region, and cause of death.

Supplementary Table 2. Results of qPCR and sequencing analysis.

Supplementary Table 3. Deer SARS-CoV-2 substitutions from sequences in this study.

Supplementary Table 4. Single nucleotide polymorphisms from global whole genome sequencing of SARS-CoV-2 in deer.

Supplementary Table 5. GISAID Acknowledgements for deer sequences.

Supplementary Table 6. Human sequences used for Figure 3C.

Supplementary Table 7. Deer and human sequences used for the time-scale Bayesian maximum credibility tree.

Supplementary Table 8. Parameter tuning for time-scale Bayesian maximum clade credibility tree.

Supplementary Table 9. Key materials and resources.

Supplementary Figure 1. Checking sample tracking by analysis of nonviral sequences. Sequence reads from each SARS-CoV-2 sample were aligned to the white-tailed deer (GCF_002102435.1) and human (GCF_000001405.39) genomes, and nonviral reads enumerated. The numbers of deer and human reads are shown for each deer sample, denoted by the laboratory accession number VSP###. Results for 140 human samples, sequenced in the same sample batches, are shown to the far left. The proportion of reads aligning to the deer genome are shown in green, the average fraction aligning to the human genome is shown in orange.

Supplementary Figure 2. Results of quantification of dN/dS ratios. Values are shown in each panel from right to left dN/dS in deer, in human alpha, human delta, and all other lineages. Lineages in deer are color coded as in the human samples. The dN/dS values were calculated comparing each sequence to the reference Wuhan strain and are shown for the concatenation of all open reading frames (upper left), ORF1a (upper right), ORF 1b (lower left), and Spike (lower right).

Supplementary Figure 3. Time resolved phylogenies, comparing deer and human lineages. Deer isolates are compared to human isolates from an overlapping

geographic region. Trees are shown for individual lineages when relatively divergent from other deer isolates, or for pairs of deer isolates when similar in sequence.

References

1 Sarah, T. *et al.* Coronaviruses with a SARS-CoV-2-like receptor-binding domain allowing ACE2-mediated entry into human cells isolated from bats of Indochinese peninsula. *Nature Portfolio*, doi:10.21203/rs.3.rs-871965/v1 (2022).

2 Zhou, H. *et al.* Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. *Cell* **184**, 4380-4391 e4314, doi:10.1016/j.cell.2021.06.008 (2021).

3 Xiao, K. *et al.* Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* **583**, 286-289, doi:10.1038/s41586-020-2313-x (2020).

4 Oude Munnink, B. B. *et al.* Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* **371**, 172-177, doi:10.1126/science.abe5901 (2021).

5 Lu, L. *et al.* Adaptation, spread and transmission of SARS-CoV-2 in farmed minks and associated humans in the Netherlands. *Nat Commun* **12**, 6802, doi:10.1038/s41467-021-27096-9 (2021).

6 Oreshkova, N. *et al.* SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. *Euro Surveill* **25**, doi:10.2807/1560-7917.ES.2020.25.23.2001005 (2020).

7 Salajegheh Tazerji, S. *et al.* Transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) to animals: an updated review. *J Transl Med* **18**, 358, doi:10.1186/s12967-020-02534-2 (2020).

8 Murphy, H. L. & Ly, H. Understanding the prevalence of SARS-CoV-2 (COVID-19) exposure in companion, captive, wild, and farmed animals. *Virulence* **12**, 2777-2786, doi:10.1080/21505594.2021.1996519 (2021).

9 Kuchipudi, S. V. *et al.* Multiple spillovers from humans and onward transmission of SARS-CoV-2 in white-tailed deer. *Proc Natl Acad Sci U S A* **119**, doi:10.1073/pnas.2121644119 (2022).

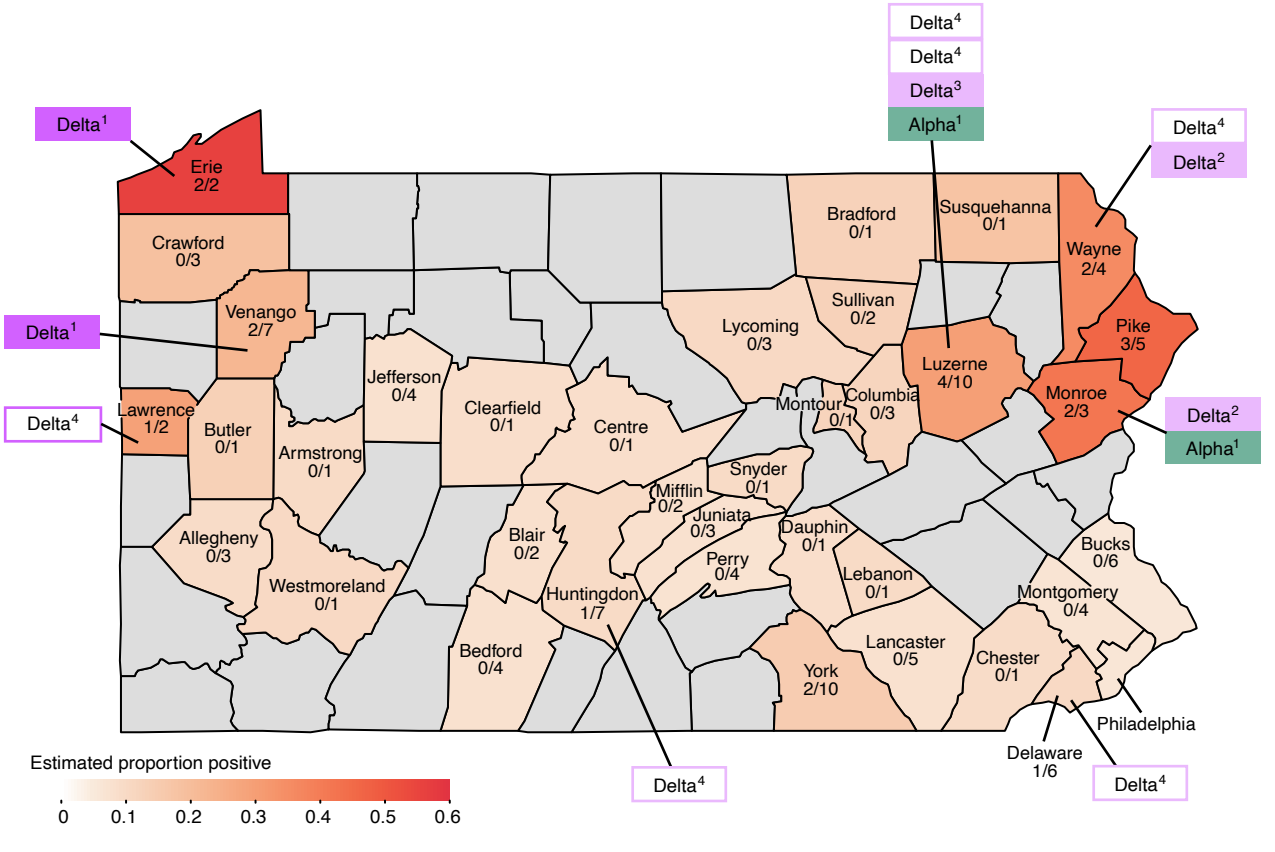
10 Hale, V. L. *et al.* SARS-CoV-2 infection in free-ranging white-tailed deer. *Nature*, doi:10.1038/s41586-021-04353-x (2021).

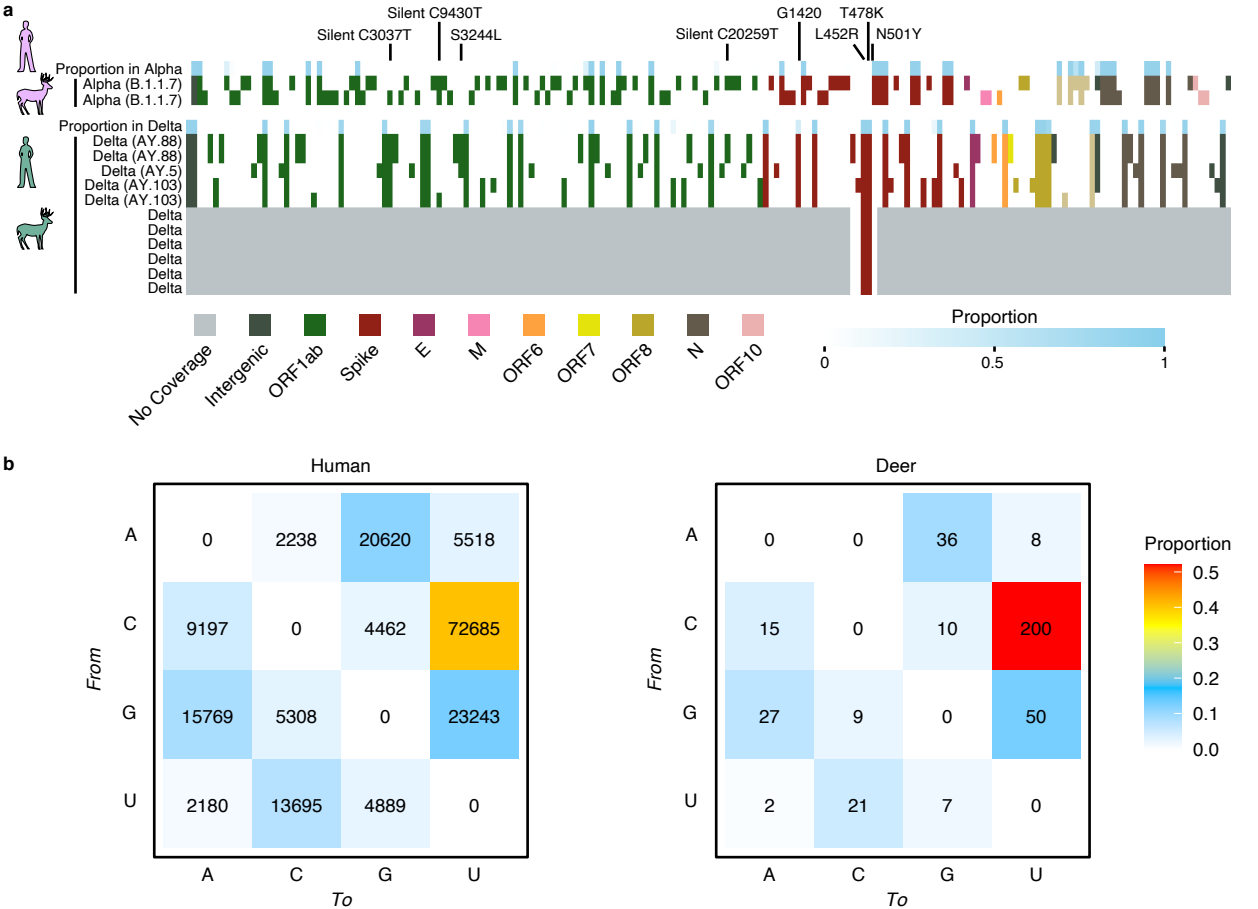
11 Cool, K. *et al.* Infection and transmission of ancestral SARS-CoV-2 and its alpha variant in pregnant white-tailed deer. *Emerg Microbes Infect* **11**, 95-112, doi:10.1080/22221751.2021.2012528 (2022).

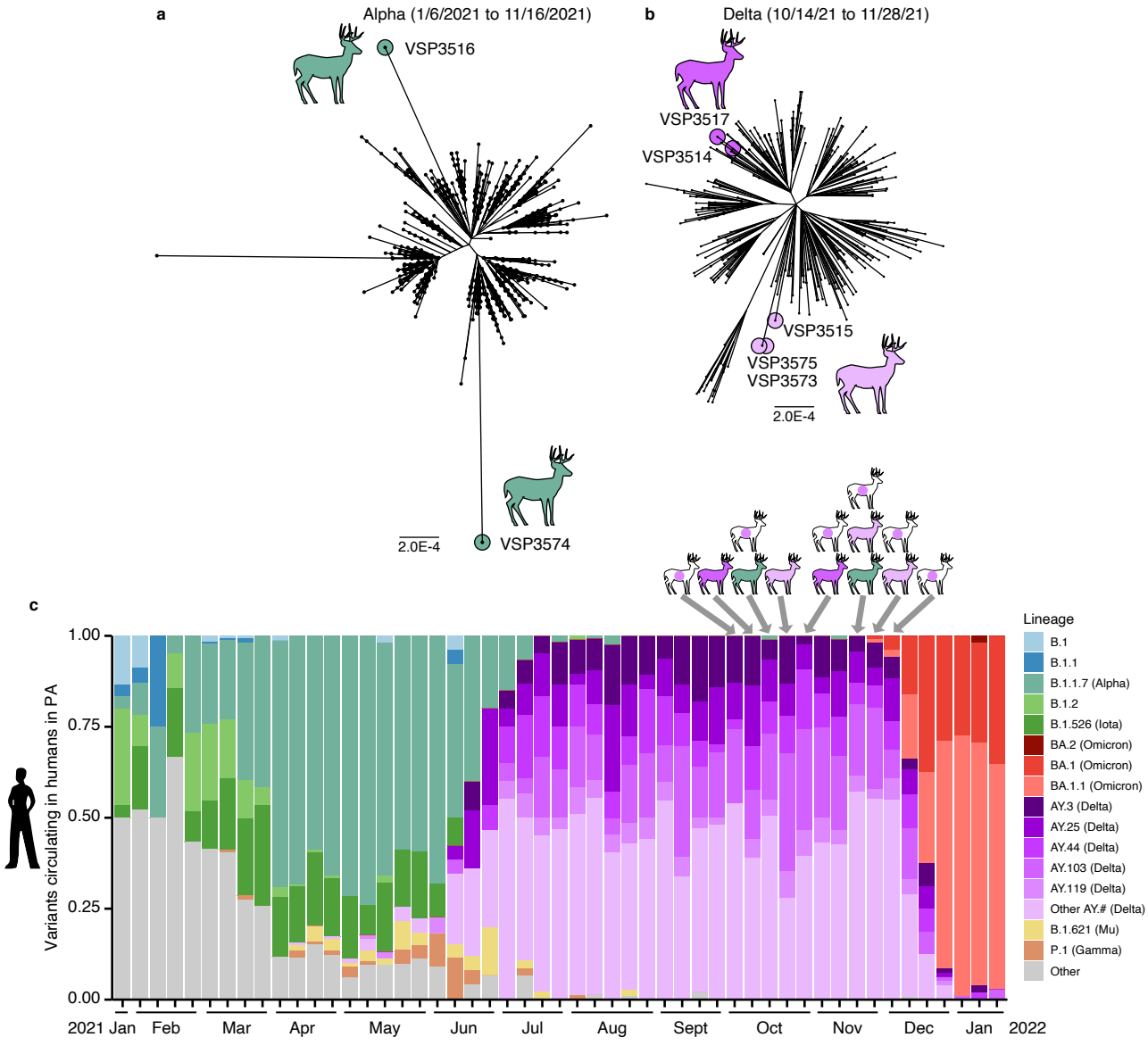
12 Bashor, L. *et al.* SARS-CoV-2 evolution in animals suggests mechanisms for rapid variant selection. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2105253118 (2021).

13 Pickering, B. *et al.* Highly divergent white-tailed deer SARS-CoV-2 with potential deer-to-human transmission. *bioRxiv*, 2022.2002.2022.481551, doi:10.1101/2022.02.22.481551 (2022).
14 Smyth, D. S. *et al.* Tracking cryptic SARS-CoV-2 lineages detected in NYC wastewater. *Nat Commun* **13**, 635, doi:10.1038/s41467-022-28246-3 (2022).
15 O'Toole, A., Pybus, O. G., Abram, M. E., Kelly, E. J. & Rambaut, A. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. *BMC Genomics* **23**, 121, doi:10.1186/s12864-022-08358-2 (2022).
16 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403-1407, doi:10.1038/s41564-020-0770-5 (2020).
17 Everett, J. *et al.* SARS-CoV-2 Genomic Variation in Space and Time in Hospitalized Patients in Philadelphia. *mBio* **12**, doi:10.1128/mBio.03456-20 (2021).
18 Marques, A. D. *et al.* SARS-CoV-2 Variants Associated with Vaccine Breakthrough in the Delaware Valley through Summer 2021. *mBio*, e0378821, doi:10.1128/mbio.03788-21 (2022).
19 Mallapaty, S. COVID is spreading in deer. What does that mean for the pandemic? *Nature* **604**, 612-615, doi:10.1038/d41586-022-01112-4 (2022).
20 National Deer Association Staff. An annual report on the status of white-tailed deer - the foundation of the hunting industry in North America. (National Deer Association, 2022).
21 Rosenberry, C. R., Fleegle, J. T. & Wallingford, B. D. Monitoring deer populations in Pennsylvania. *Bureau of Wildlife Management, Pennsylvania Game Commission* (2011).
22 Hirth, D. H. Social Behavior of White-Tailed Deer in Relation to Habitat. *Wildlife Monographs* **53**, 3-55, doi:https://www.jstor.org/stable/3830446 (1977).
23 R: A language and environment for statistical computing. R Foundation for Statistical Computing (2021).
24 Anis, E. *et al.* Evaluation of a real-time RT-PCR panel for detection of SARS-CoV-2 in bat guano. *J Vet Diagn Invest* **33**, 331-335, doi:10.1177/1040638721990333 (2021).
25 Hilaire, B. G. S. *et al.* A rapid, low cost, and highly sensitive SARS-CoV-2 diagnostic based on whole genome sequencing. doi:10.1101/2020.04.25.061499 (2020).
26 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
27 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
28 O'Toole, A. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol* **7**, veab064, doi:10.1093/ve/veab064 (2021).
29 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410, doi:https://doi.org/10.1016/S0022-2836(05)80360-2 (1990).

30 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402, doi:10.1093/nar/25.17.3389 (1997).
31 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, doi:10.1186/1471-2105-10-421 (2009).
32 Carpenter, B. *et al.* STAN: A Probabilisitic Programming Language. *J. Stat. Software* **76**, 1-32, doi:https://doi.org/10.18637/jss.v076.i01 (2017).
33 Aksamentov, I., Roemer, C., Hodcroft, E. B. & Neher, R. A. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software* **6**, 3773, doi:https://doi.org/10.21105/joss.03773 (2021).
34 Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528, doi:10.1093/bioinformatics/bty633 (2019).
35 Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* **4**, vey016, doi:10.1093/ve/vey016 (2018).
36 Yang, Z. Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data. *J Mol Evol* **42**, 587-596, doi:10.1007/BF02352289 (1996).
37 Ayres, D. L. *et al.* BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics. *Syst Biol* **68**, 1052-1061, doi:10.1093/sysbio/syz020 (2019).
38 Chernomor, O., von Haeseler, A. & Minh, B. Q. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology* **65**, 997-1008, doi:10.1093/sysbio/syw037 (2016).
39 Rambaut, A. *FigTree*, <http://tree.bio.ed.ac.uk/software/figtree/> (2007).


Delta[4]
Delta[4]
Delta[3]
Alpha[1]
Delta[1]
Delta[4]
Delta[2]
Erie
2/2
Crawford
0/3
Bradford
0/1
Susquehanna
0/1
Wayne
2/4
Venango
2/7
Delta[1]
Lycoming
0/3
Sullivan
0/2
Pike
3/5
Luzerne
4/10
Jefferson
0/4
Clearfield
0/1
Centre
0/1
Montour
0/1
Columbia
0/3
Monroe
2/3
Delta[2]
Alpha[1]
Delta[4]
Lawrence
1/2
Butler
0/1
Armstrong
0/1
Snyder
0/1
Mifflin
0/2
Juniata
0/3
Allegheny
0/3
Blair
0/2
Dauphin
0/1
Perry
0/4
Bucks
0/6
Westmoreland
0/1
Huntingdon
1/7
Lebanon
0/1
Montgomery
0/4
Bedford
0/4
Lancaster
0/5
Chester
0/1
York
2/10
Philadelphia
Delaware
1/6
Delta[4]
Delta[4]
Estimated proportion positive
0
0.1
0.2
0.3
0.4
0.5
0.6


a
Silent C9430T
Silent C3037T
S3244L
Silent C20259T
G1420
T478K
L452R
N501Y
Proportion in Alpha
Alpha (B.1.1.7)
Alpha (B.1.1.7)
Proportion in Delta
Delta (AY.88)
Delta (AY.88)
Delta (AY.5)
Delta (AY.103)
Delta (AY.103)
Delta
Delta
Delta
Delta
Delta
Delta
No Coverage
Intergenic
ORF1ab
Spike
E
M
ORF6
ORF7
ORF8
N
ORF10
Proportion
0
0.5
1
b
Human
From
To
A
C
G
U
0
2238
20620
5518
9197
0
4462
72685
15769
5308
0
23243
2180
13695
4889
0
Deer
0
0
36
8
15
0
10
200
27
9
0
50
2
21
7
0
Proportion
0.5
0.4
0.3
0.2
0.1
0.0


a
Alpha (1/6/2021 to 11/16/2021)
VSP3516
VSP3574
2.0E-4
b
Delta (10/14/21 to 11/28/21)
VSP3517
VSP3514
VSP3515
VSP3575
VSP3573
2.0E-4
c
Variants circulating in humans in PA
1.00
0.75
0.50
0.25
0.00
2021 Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sept
Oct
Nov
Dec
Jan
2022
Lineage
B.1
B.1.1
B.1.1.7 (Alpha)
B.1.2
B.1.526 (Iota)
BA.2 (Omicron)
BA.1 (Omicron)
BA.1.1 (Omicron)
AY.3 (Delta)
AY.25 (Delta)
AY.44 (Delta)
AY.103 (Delta)
AY.119 (Delta)
Other AY.# (Delta)
B.1.621 (Mu)
P.1 (Gamma)
Other